ISSI 2013

SCIENTOMETRICS

14th International Society of Scientometrics and Informetrics Conference

15th–19th July 2013

Vienna, Austria

PROCEEDINGS Volume I

PROCEEDINGS OF
ISSI 2013
Vienna

VOLUME 1

14th International Society of
Scientometrics and Informetrics Conference

Vienna, Austria
15th to 20th July 2013

Editors

Juan Gorraiz, Edgar Schiebel, Christian Gumpenberger, Marianne Hörlesberger, Henk Moed

Sponsors

ASIS&T, USA
Elsevier B.V.
EBSCO Information Services, USA
Federal Ministry for Science and Research, Austria
Federal Ministry for Transport, Innovation and Technology, Austria
Information Assistant, Verein für Informationsmanagement, Vienna
ORCID, Inc.
Science-Metrix/R&D Reports
Swets Information Services
Thomson Reuters
ZSI - Centre for Social Innovation, Vienna

Printed by Facultas Verlags- und Buchhandels AG,
Stolbergasse 26, A-1050 Wien

ISBN: 978-3-200-03135-7
ISSN: 2175-1935

ORGANISATION AND COMMITTEES

Conference Chairs

Juan Gorraiz
Edgar Schiebel

Programme Chairs

Christian Gumpenberger
Marianne Hörlesberger
Henk Moed

Poster Chairs

Jacqueline Leta
Wolfgang Mayer

COMMITTEES

Local Committee

Doctoral Forum Chairs:
Ivana Roche and Christian Schlögl

Local Organising Committee:
Ulrike Felt
Peer Vries
Wolfgang Claudius Müller
Martin Fieder
Johannes Sorz
Bernard Wallner
Wolfgang Mayer
Martin Wieland
Ambros Wernisch
Manuela Kienegger
Manuela Korber
Beatrix Wepner
Maria-Elisabeth Züger
Beatrice Rath
Silvia Steinbrunner

Scientific Committee

Jonathan Adams
Isidro Aguillo
Per Ahlgren
Isola Ajiferuke
Dag W Aksnes
Jens-Peter Andersen
Eric Archambault
Clément Arsenault
Joaquin Azagra-Caro
Tomas Baiget
Rafael Ball
Judit Bar-Ilan
Tomaz Bartol
Aparna Basu
Guntram Bauer
Donald Beaver
Nicola Bellis
Sada Bihari-Sahu
Johan Bollen
Andrea Bonaccorsi
Maria Bordons
Katy Börner
Lutz Bornmann
Hamid Bouabid
Kevin W. Boyack
Barry Bozeman
Tibor Braun
Quentin Burrell
Guillaume Cabanac
Alvaro Cabezas
Juan Miguel Campanario
David Campbell
Neeraj Chaurasia
Dar-Zen Chen
Chaomei Chen
Alexander Chervyakov
Rodrigo Costas
Grégoire Côté
Blaise Cronin
Hans-Dieter Daniel
Cinzia Daraio
Hamid Darvish
Prabir Dastidar
Koenraad Debackere
Gernot Deinzer
Ying Ding
Sandhya Diwakar
Leo Egghe
Tim Engels
Martin Fieder
Claire François
Jonathan Furner
Antonio Garcia
Aldo Geuna
Elea R. Giménez Toledo
Yves Gingras
Wolfgang Glänzel
Isabel Gomez
Alicia Gomez
Juan Gorraiz
Jiancheng Guan
Christian Gumpenberger
Raf Guns
Stefanie Haustein
Sybille Hinze
Michael Hofer
Marianne Hörlesberger
Stefan Hornbostel
Xiaojun Hu
Mu-Hsuan Huang
Sven Hug
Masatsura Igami
Peter Ingwersen
Ludmila Ivancheva
Siladitya Jana
Margriet Jansz
Evaristo Jimenez-Contreras
Milos Jovanovic
Yuya Kajikawa
Sylvan Katz
Dick Klavans
Manuel Krauskopf
Hildrun Kretschmer
J P S Kumaravel
Vincent Lariviere
Birger Larsen
Karl-Heinz Leitner
Benedetto Lepori
Jacqueline Leta
Jonathan Levitt
Loet Leydesdorff
Yang Li Ying
Liming Liang
Judith Licea
Deming Lin
Yuxian Liu
Szu-Chia Lo
Carmen Lopez Illesca
Bob Losee
Terttu Luukkonen
Marc Luwel
Domenico Maisano
Valentina Markusova
Werner Marx
Wolfgang Mayer
Kate McCain
Eustache Megnigbeto
Lokman Meho
Raul Mendez-Vasquez
Henk Moed
Alexis Michel Mugabushaka
Rogerio Mugnaini
Wolfgang Müller

Francis Narin
Anton Nederhof
Pentti Nieminen
Ed Noyons
Michael Ochsner
H. Peter Ohly
Carlos Olmeda-Gómez
José Luis Ortega
Maria Antonia Ovalle-Perandones
Manfred Paier
Andres Pandiella
Antonio Perianes-Rodríguez
Bluma C. Peritz
Olle Persson
Fernanda Peset
Alan Porter
Anastassios Pouris
Gangan Prathap
Junping Qiu
Luc Quoniam
Ismael Rafols
Ravichandra Rao
Emanuela Reale
Steve Reding
Ralph Reimann
John Rigby
Nicolás Robinson
Ivana Roche
Jürgen Roth
Ronald Rousseau
Jane Russell
Victor Rybachuk
Bibhuti Sahoo
Ulf Sandström
Shivappa Sangam
Elias Sanz
Kate Sapir
Andrea Scharnhorst
Edgar Schiebel
Christian Schloegl
Ulrich Schmoch
Jesper Schneider
Antoine Schoen
Torben Schubert
András Schubert
Philip Shapira
Robert Shelton
Gunnar Sivertsen
Stig Slipersæter
Henry Small
Johannes Sorz
Divya Srivastava
Marie Strahle
Cassidy Sugimoto
Yuan Sun
Mike Thelwall
Bart Thijs
Robert Tijssen
Yasar Tonta
Daniel Torres-Salinas
Metin Tunc
Peter van Den Besselaar
Nees Jan Van Eck
Thed Van Leeuwen
Bart Van Looy
Anthony Van Raan
Benjamin Vargas-Quesada
Liwen Vaughan
Peter Vinkler
Martijn Visser
Bernard Wallner
Ludo Waltman
Xianwen Wang
Beatrix Wepner
Ambros Wernisch
Howard White
Jos Winnink
Matthias Winterhager
Dietmar Wolfram
Yishan Wu
Erjia Yan
Ying Ye
Lin Zhang
Yajuan Zhao
Dangzhi Zhao
Michel Zitt
Alesia Zuccala

PREFACE

The 14th International Society of Scientometrics and Informetrics Conference takes place at the University of Vienna 15-19 July 2013 and is jointly organised by the University of Vienna and the Austrian Institute of Technology (AIT) under the auspices of ISSI – the International Society for Scientometrics and Informetrics.

This conference provides an international open forum for scientists, research managers, authorities and information professionals to debate the current status and advancements of informetric and scientometric theories, concepts and indicators. In addition to the traditional evaluative focus, participants will discuss practical applications in related fields such as library and information science, history of science, philosophy of science, R&D-management, etc.

This conference raises particularly the issues of new metrics (usage metrics and altmetrics) as complement to the classical citation metrics and opens the floor to discuss manifold aspects: what can really be measured with them as proxies, which could turn out to be adequate and robust indicators, and finally which reliable data sources are available to retrieve them?

The importance of this topic is underpinned by two plenary sessions. In the first one keynote speaker Johan Bollen provides an overview of social network services and analyses. In the second one old metrics are contrasted with new ones in short introductions by experts (Henk Moed, Juan Gorraiz, Victor Henning) and followed by a panel discussion with representatives from research, research management and information industry, who will shed light on the pros and cons of these indicators from their specific point of view.

The third plenary session deals with an evergreen as much as cumbersome topic, namely the methodological and ethical problems of individual-level evaluative bibliometrics. Wolfgang Glänzel and Paul Wouters will present "10 things one must not do with individual-level bibliometrics" followed by "10 things one can do with individual-level bibliometrics", both commented by Henk Moed and Gunnar Sivertsen.

The ISSI conference is certainly one of the world's largest international conferences devoted to this field, as is illustrated by the large number of 338 submissions received this year. 912 authors are affiliated to organisations located in 42 countries from all over the world. The top three contributing countries are China (149), Spain (129) and the USA (101). Chile, Cuba, Malaysia, Sri Lanka and Ukraine are represented by at least one author, too.

All contributions were evaluated by at least three reviewers of the International and Local Committees. Thereof only 145 (107 full papers and 38 research in progress papers) could be accepted for oral presentations. 36 sessions run in parallel thrice a day in groups of four covering the gamut from "citation analysis" to "open access". In addition, 107 posters are shown in two dedicated poster sessions.
All oral presentations and posters can be found in the conference proceedings.

Moreover, four tutorials either deal with several mapping tools (like e.g. "Sci2" and "Citespace") or address the unification issue of organizations, whereas four pre-conference workshops focus on information retrieval, topic extraction methods, standards for classifications, and bibliometric analysis for funding agencies. The pre-conference day is complemented by a doctoral forum.

By organising the 14th International Conference in Vienna we hope not only to extend the tradition of the ISSI conferences as one of the most important international meeting points for the scientometric and bibliometric community, but also to promote the respective on-going activities in Austria.

Our thanks go to the ISSI board for their trust and their constant support, all the contributors for their submissions, the members of the Local and International Committee for their reviewing effort as well as the sponsors for their generous financial support.
We are particulary grateful for the engagement of Heike Faustmann, Alfred Kerschenbauer, Nikolaus Ortner, Johannes Sorz, Silvia Steinbrunner, and Maria-Elisabeth Züger.

Last but not least each conference should also be a feast for all senses. Every endeavour has been made to not only put together an outstanding scientific programme, but also to organize interesting and diverse social events, which will allow you to embrace the beauty and cultural richness of Vienna and its surroundings.

We wish you a great time at the 14th International Society of Scientometrics and Informetrics Conference!

Juan Gorraiz, Edgar Schiebel, Christian Gumpenberger,
Marianne Hörlesberger, and Henk Moed

INDEX

[1] Research in progress paper

KEYNOTE

SOCIAL NETWORK ANALYSIS

Johan Bollen

jbollen@indiana.edu
School of Informatics and Computing, Indiana University

Abstract

Online social networking services play an increasingly important role in the private and public lives of hundreds of millions of individuals, capturing the most minute details of their whereabouts, thoughts, opinions, feelings, and activities, in real-time. Advances in social network analysis and natural language processing have enabled computational social science which leverages computational methods and large-scale data to develop models of individual and collective behavior to explain and predict a variety of economy, financial, and social phenomena.

In this keynote I provide an overview of the ability of large groups of people to collectively produce information that is dynamic, complex, and adaptive. In addition to explicit information, text analysis algorithm can be used to extract indicators of social mood and sentiment from social media data. Researchers have used these techniques to gauge "national happiness" as well as consumer sentiment towards particular brands and products. Perhaps most tantalizing, evidence has been found that online social mood and sentiment may yield predictive information with regards to a variety of socio-economic phenomena, such as movie box office receipts, product adoption rates, elections, public health, and even stock market fluctuations. With respect to the latter, I will outline our own research on the subject of stock market prediction from large-scale Twitter and Google Trends data, and discuss recent efforts to leverage social media data to study scientific communication.

ORAL PRESENTATIONS

ACADEMIC CAREER STRUCTURES – HISTORICAL OVERVIEW GERMANY 1850-2013

Cathelijn J.F. Waaijer[1]

[1] *c.j.f.waaijer@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University, PO box 905, 2300 AX Leiden, The Netherlands

Abstract
Long selection periods and small initial chances of achieving a successful lifelong career characterize the current academic career system. In this study we investigate how this system has developed. We first introduce a conceptual framework of academic careers of which the components can be used to characterize academic positions. We use this framework to trace the historical development of the academic career system in Germany from the early 19th century until now. We chose Germany as it was the leading country in science from 1800-1933. Professorships used to be the only official academic positions in Germany. Gradually, however, academic positions below the professorship emerged, or rather, became more formalized. First only positions directly below professorships developed into official academic positions, but later the informal PhD and second degree (*Habilitation*) student positions were formalized into research assistantships. This has led to a decrease in the share of professorships in official figures. At first glance, this decrease implies that the chances of young people starting on a PhD obtaining an eventual professorship must have decreased. Further lines of work will include investigating numbers of PhD students and *Habilitation* pursuers without official positions, in order to determine whether chances have indeed decreased.

Conference Topic
Modeling the Science System, Science Dynamics, and Complex System Science (Topic 11)

Introduction
Currently, when university graduates seek to pursue a scientific career, they have to go through a long probationary period. Only a small fraction will actually become tenured staff at academic institutions. Graduates start their careers by working as 'apprentices', first as PhD students, later as postdoctoral researchers. During this period they are typically on scholarships or employed temporary contracts. After this period postdoctoral researchers can be hired on 'tenure track' positions, which give the researchers the prospect of obtaining a permanent position if successful (Dooris & Guidos 2006). This means university graduates aspiring to an academic career can be employed on temporary contracts for a total of ten to twenty years: first three to five or more years as a PhD student, then for one or more years as postdoctoral researchers, and finally for five to seven years

on a tenure track, often at different scientific organizations (Nerad & Cerny 1999).
Researchers and policymakers identify problems regarding current academic careers, such as the small number of PhD students eventually becoming tenured staff at academic institutions and the long probationary period with a high level of consensus (Waaijer 2013). A common sentiment, for example, is that is has become more difficult to obtain a professorship because the relative number of professorships compared to the number of PhD positions has decreased, i.e., the academic career pyramid has become steeper. Several studies suggest that the period until tenure has been increasing in academia over the past two decades (Bridges to Independence: Fostering the Independence of New Investigators in Biomedical Research 2005, van Balen & van den Besselaar 2007).
However, these studies only focus on recent decades. To fully understand why the academic career system is as it is, we need to look back further and trace how it has developed to determine whether the pyramidal academic career structure has indeed become steeper, making it more difficult to achieve a career in academic research (i.e., obtain a professoriate), and whether probationary periods have become longer.
First, we need to introduce a conceptual framework of academic research careers and the different aspects that shape these careers. Second, we will investigate more thoroughly how specific characteristics of the current academic career system developed in Germany, which was the most important country in science and technology in the late 19th and the beginning of the 20th century. Third, we will conclude this paper by outlining future lines of investigation.

Conceptual framework

A useful definition of career was given by Baruch and Rosenstein: "a process of development of the employee along a path of experience and jobs in one or more organizations" (Baruch & Rosenstein 1992). The traditional view of careers is that of vertical movement through a rigid, well-defined system within one organization, but over the past decades, as careers themselves have become more fluid, career models that are more dynamic and multidirectional have been proposed, both with regards to the position on the career ladder and between different organizations (Baruch 2004, Peiperl & Baruch 1997). In contrast to careers in many sectors, careers in academia are usually still quite linear with regard to positions – one typically enters at a young age, works as an "apprentice" and tries to move up on the career ladder. At the same time most researchers who do not succeed in moving up leave academia to work in another sector. On the other hand, inter-organizational mobility is quite high; especially when transitioning from the PhD to the postdoctoral phase researchers are expected to change institutions and preferably even work abroad (Enders & Kaulisch 2006, Ackers 2008).
As a point of departure, Figure 1 gives a highly stylized scheme we have designed of the archetypical academic career system in the United States. The width of the

arrows represents the fraction of researchers transitioning from one phase to the next.

Figure 1. Archetypical academic career in the United States. The width of the arrows represents the percentage of researchers moving from one position to a higher one.

This archetypical US academic career scheme, however, is only an hierarchical depiction of five positions currently found in the US and with associated characteristics specific to the present day and the US university system. As such, it is too specific to model positions of all leading scientific countries and different historical time periods. It disregards differences between fields, institutions and individuals. For example, researchers could skip the postdoctoral phase because it is customary in their field to be hired directly as an assistant professor, because that is the policy of the specific institution, or because they are seen as so talented their institutions want to bind them by offering a post as an assistant professor directly. Furthermore, the scheme omits some very important characteristics of academic careers, such as when tenure is granted, when one is allowed to supervise students, when one is allowed to pursue an own line of research, etcetera. Therefore, we also introduce a more elaborate conceptual model of academic careers that does incorporate these aspects (Fig. 2).

This model makes explicit four important aspects of the academic career: how scientists perform research, the extent to which they have to attract funding, the control they have over their scientific activities and over resources, and their terms of employment, all broken down in multiple characteristics. In addition, we have sketched a rough estimate of how we expect these characteristics to progress during a typical research career (e.g., when is the maximum scientific production reached with regards to control over personal resources?). The framework is based on the presumption that scientific production is the main determinant of scientists' ability to obtain grants or other forms of funding. The level of funding determines their control of both monetary and personal resources, and choice of research lines. Increased control of resources, in turn, leads to a difference in research performance and in type of research activities: less hands-on research and more supervisory. Simultaneously, terms of employment improve for researchers higher on the career ladder. As stated, we have sketched the evolvement of different characteristics (i.e., the placement of the boxed variables) for an academic career in Figure 2, but the placement is likely to be different for different career systems. Thus, the characteristics outlined in our model and can be used to classify an academic position on various scales (e.g., salary, tenure, degree of independence, degree of supervision received or given etc.). In this way, academic career systems of different countries, institutes, or scientific fields can be characterized using a multidimensional scale. Some characteristics are quite

easily measured or estimated, whereas data on other characteristics are more difficult to obtain. We will use a few characteristics, such as an estimation of progress in the scientific career, supervision of students and researchers, and whether an academic position is paid to characterize changes in academic positions in Germany from the 19th century until now.

The case of Germany

Germany is one of the leading countries in science and technology (S&T), spending 2.82% of its gross domestic product (GDP) on research and development (R&D) in 2009 (OECD 2009), having the fourth largest number of scientific publications based on the total number of citable publications in the Web of Science database in 2011 (own calculations), and having 39 universities in the top-500 research universities as measured by the Leiden 2011/2012 ranking (Centre for Science and Technology Studies 2012). Thus, merely because of its importance in contemporary science the German academic career system is already interesting to study.

But in addition, the German academic career system is interesting because of its historical development and especially the influence it has had on academic career systems across the globe. The concept of the research university originates from Prussia, from linguist, philosopher and government official Wilhelm von Humboldt. Before the 19th century, teaching students was the primary focus of universities. However, Humboldt introduced a model of higher education with the unity of research and teaching at its core. Humboldt's idea was that students should not merely study existing knowledge, but should perform scientific work themselves, under the supervision of the academic staff. This led to a shift in focus of universities disseminating existing scientific knowledge to staff and students working to increase scientific knowledge (Clark 1993). The Humboldtian university model in which research was central would be adopted by several countries and is the model on which current research universities are based.

Possibly due to the focus on science at German universities, Germany became the leading country in chemistry, physics and medicine in the 19th century and the first third of the 20th century, until the national socialist takeover in 1933. An example of the German dominance in science is that between 1901 and 1931 Germany was leading in the number of Nobel prizes for the sciences – a total of 32. Like the United States in the late 20th century, Germany was the country for foreign scientists to move to in order to be trained at the most prestigious institutes. It seems likely that its proven success and influence on foreign scientists made the academic career system of Germany a standard for other countries during this period.

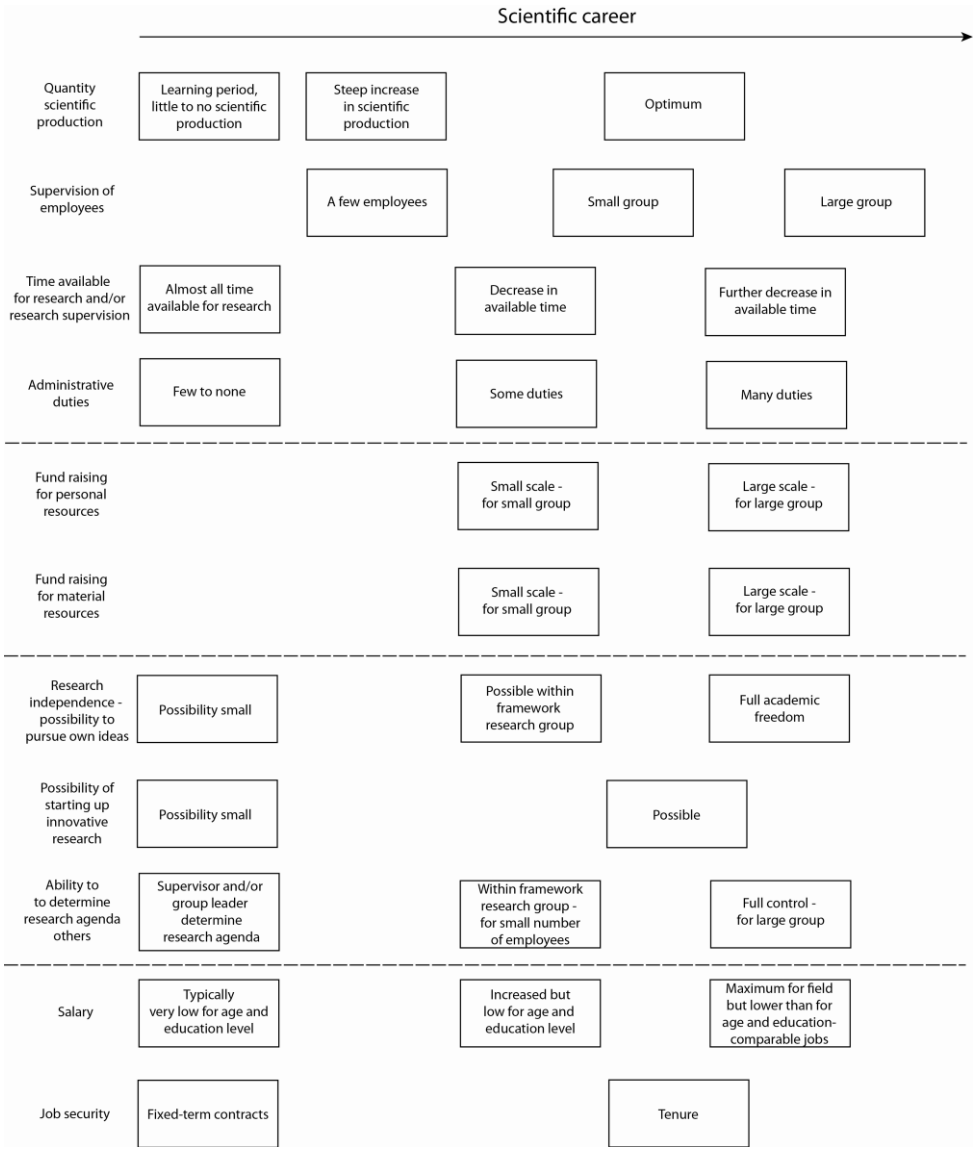


Figure 2. Conceptual framework of academic career with its various characteristics. Horizontally the progress in the academic career is depicted; the further to the right, the higher the status of the researcher.

Contemporary academic career system

The current German academic career system[1] can be qualified as quite hierarchic: full professors determine the research agenda of their groups. These groups are, apart from the full professors, made up by early-career scientists (*Nachwuchswissenschaftler*; literally translated "offspring scientists"), who are considered to be training to obtain professorships themselves (Kreckel 2008). An

academic career typically starts as a PhD student (*Doktorand*)[2], for whom a variety of positions and remunerations exist. In the natural sciences, PhD students usually have a position as a research or teaching assistant (*wissenschaftliche Hilfskraft*) or research affiliate (*wissenschaftlicher Mitarbeiter*). The difference is that assistants are typically employed part-time and expected to work on their PhD for the remainder of the time, while affiliates are full-time employed. In the humanities and social sciences, on the other hand, the percentage of PhD students on a scholarship or even without any financial allowance is much higher (Fräßdorf, Kaulisch & Hornbostel 2012). After the PhD, researchers attempt to obtain positions as (postdoctoral) research affiliates (like PhD students also called *wissenschaftliche Mitarbeiter*, but sometimes described as *wissenschaftliche Assistente*, although the latter designation is becoming less common). At the end of the 20th century, one needed to work as a research affiliate for a relatively long period of time in order to write a 'second dissertation': the *Habilitation*. With the *Habilitation*, a scientist could be 'called' to a university as a full professor.
Figure 3 shows that the current German academic career system is characterized by a relatively low number of professorships (just over 10% of the total number of academic positions), whereas other positions are much more abundant. More than 70% of all scientific employees hold positions as research affiliates/assistants.

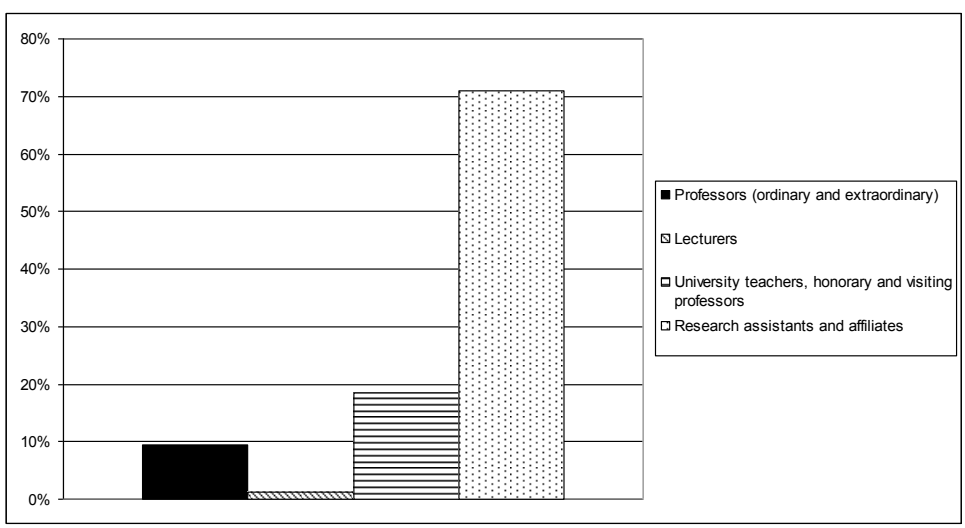


Figure 3. Distribution of academic positions at German universities in 2010, grouped by professors - lecturers - university teachers, honorary and visiting professors - research assistants and affiliates. Source: *Statistisches Bundesamt* 2010.

Thus, the German academic career system is characterized by a relatively small 'top' of scientists, which is only made up by full professors. A few other research positions (combined in the Lecturer group) fill the gap on the career ladder

between full professors and research affiliates, but the group of scientists in these positions is very small. Finally, almost 20% of all jobs in German higher education are positions focused on teaching (combined in the University teachers group); although these academics may also perform research, their main task is teaching.

Historical development

In this section we describe the development of the German academic career system and illustrate it with changes in the (relative) numbers of different positions. For this purpose, we combine quantitative data from various studies on changes in academic careers with qualitative background data. The quantitative studies employ different methodologies, especially in how they define groups of researchers. Various reasons for these differences exist: the focus of the study can be different (e.g., focus on professors only or a broader scope including other positions), the existence of positions (thus, the introduction of junior professorships is a recent phenomenon), and of course methodological choices (for example, are ordinary and extraordinary professors lumped together or left as two separate groups?). We focus on 'the big picture' and sketch developments in academic career systems with such data as are available.

Academic positions 19th century until mid 20th century

In the 19th century, only three official academic positions existed: the ordinary professorship, the extraordinary professorship and the private lectureship. Before a man could fill an official academic position (the appointment of the first female scientist to an ordinary professorship would only occur in 1926), he needed to obtain first a PhD and after that a *Habilitation* based on original research to give him the right of lecturing at a university (Latin: *venia legendi*). The work associated with both theses was unpaid and persons working towards these theses were not considered to be academics just yet (von Ferber 1956, Weber 1946).
The lowest formal position one could fill was the one of private lecturer (*Privatdozent*) (Ben-David & Zloczower 1961, Weber 1946). Note, however, that private lecturers were not remunerated for their work by their employers, the universities, but were rather paid on a 'freelance' basis by collecting lecture fees directly from their students:

"(…) he gives a course of lectures without receiving any salary other than the lecture fees of his students" (Weber 1946).

On top of the hierarchy were (and still are) the professors: the ordinary professor (*ordentlicher öffentlicher Professor* or *Ordinarius*) and extraordinary professor (*außerordentlicher Professor* or alternatively, *Extraordinarius*). The difference between the two positions is that the ordinary professor holds a professorial chair in a broad subject, whereas an extraordinary professor does not hold a chair and typically works on a narrower subject. A private lecturer could be promoted to

extraordinary professor or directly to ordinary professor. An extraordinary professor could fill that position for his entire career, but could also later be promoted to ordinary professor. As these customs show, ordinary professorships were considered to be higher positions than extraordinary professorships (Bock 1972 pg. 120, von Ferber 1956 pg. 107).

From around 1930 a distinction was made between permanent (*planmäßige*) and non-permanent (*außerplanmäßige*) extraordinary professorships. The difference between the two was that permanent extraordinary professorships were positions continuously in place that were filled by another scientist when the extraordinary professor left the position, while appointments of non-permanent extraordinary professors were made on an *ad hoc* basis. The latter appointments were typically made in the case of researchers who had performed excellent research and had been habilitated, but had not been able to obtain the position of ordinary professor. As extraordinary professor holding a non-permanent post, a researcher could supervise doctoral theses and conduct research in their subject relatively autonomously, without the need for the university to retain a permanent (extraordinary) professorial post in the subject. For clarity, permanence here refers to the permanence or non-permanence of the positions itself; the appointed extraordinary professors did obtain a permanent (tenured) job regardless of whether they filled permanent or non-permanent positions.

Another category of employees were research assistants (*wissenschaftliche Assistente*). We now associate these assistantships with positions for scientists at the beginning of their careers, aiming for a professorship themselves (called research assistants or affiliates). However, originally they were introduced as literal assistants to professors. As described above, in the Humboldtian university model education by experimental work was considered vital for students. An assistant's function consisted of aiding experimental demonstrations by professors during lectures and of facilitating and performing experimental work for their professors. In addition, supporting personnel such as librarians and museum curators were also given an 'assistant' status. Over time, a so-called 'assistant career' according to age, years of service, and merit developed, with tiered positions called 1st, 2nd and 3rd research assistant (Bock 1972, pg. 121). Assistantships could become "probation extensions" ("*Bewährungsaufstiege"*), and assistants could hold their position for a long period while sometimes unofficially functioning as departmental heads (Bock 1972 pg. 127).

To our knowledge the first study on numbers of academics is the detailed study by sociologist and economist Christian von Ferber, who investigated the German academic career system as it developed from 1864 until 1953 (von Ferber 1956). He counted the numbers of different types of professors and lecturers at eleven points in time. Unfortunately, his data do not include numbers of persons working to obtain a PhD or a *Habilitation*, probably because positions that would now be called research affiliate or research assistant were not considered to be academic positions during that period. Nevertheless, his study reveals some very interesting developments in the distribution of academic positions.

Most notably, he shows that the proportion of ordinary professors has steadily decreased from half of all academic employees in 1864 to approximately 25% by 1953 (Fig. 4). From 1864 until 1910 there was an increase in the relative number of private lecturers, but after this period the number decreased again to below the original 1864 proportion. The line showing the proportion of extraordinary professors is uneven as well; the results suggest the relative number of extraordinary professors remained stable until 1920, but had increased dramatically in 1931. A possible explanation for these results could be that non-permanent extraordinary professorships, which are grouped with permanent extraordinary professorships, were introduced around 1930. After World War II, the proportion had decreased again. The position with the largest increase in relative numbers was the position of university teachers. This group, however, is a fairly heterogeneous group consisting of university teachers paid on a contractual basis, part-time professors and research candidates[3], so it is difficult to say definitively which subgroup or subgroups contributed to the increase.
When broadly looking at the data, the main trend (already noted by Ben-David and Zloczower (1961)) is that the relative number of ordinary professorships declined, while other positions below the professorship (private lecturers, extraordinary professors and a heterogeneous group of [part-time] university teachers and researchers) grew. Our analysis shows the different positions took turns in "filling the gap" left by the relative decrease in ordinary professorships.

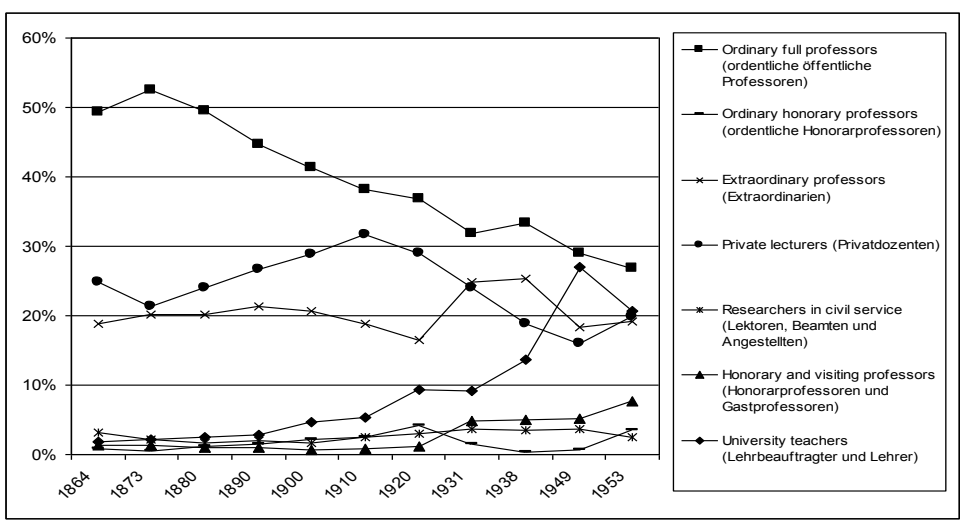


Figure 4. Distribution of academic positions at German universities 1864-1953. For 1953 West Germany only. Source: von Ferber 1956 pp. 195, 210.

The conclusion that positions underneath the professorship became more prevalent is supported by von Ferber's analysis of the relative numbers of private lecturers and non-permanent extraordinary professors in comparison to the

numbers of the ordinary professors and extraordinary professors on permanent positions. The analysis shows that the relative number of private lecturers and non-permanent extraordinary professors increased dramatically over time, especially in medicine and the natural sciences (Fig. 5).

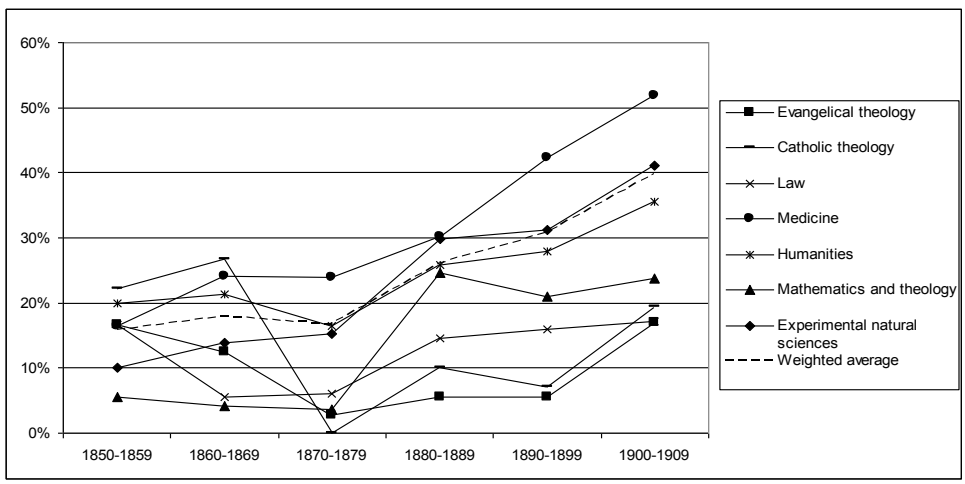


Figure 5. Private lecturers and non-permanent extraordinary professors as a percentage of all higher academic positions per subject area. Source: von Ferber, tables 8 and 9, pg. 81.

Academic positions mid 20th century until now

For investigating the distribution of academic positions from around 1950 until now we turn to publications on this subject by the German Statistical Office (*Statistisches Bundesamt*), which were published from 1952 (Statistisches Bundesamt 1953, 1966, 1969, 1976, 1982, 1992, 2004 & 2011). These publications not only publish the number of professors, but also of other university employees. This means data on non-professorial scientific employees such as research affiliates and assistants can now be incorporated into our analysis. A limitation of these publications is that their classification of academics has changed over the years due to the fact that designations and job content of positions have changed, which makes it difficult to track the development of specific positions. Therefore, we show differences in the academic career system by showing the development of four broad groups of positions: professorships combining research and teaching (both ordinary and extraordinary professorships), positions just below the professorship (lecturers), positions with a focus on teaching that are often part-time (university teachers, honorary professors and visiting professors), and the lower academic positions (research affiliates/research assistants).

Like in many other Western countries, Germany saw a huge increase in the number of enrolled students at universities and vocational colleges following World War II (Enders 1996). In the 1950s the educational burden due to the

massification of higher education had become so large for the assistants that they could not dedicate enough time to research to pursue a *Habilitation*. As the growing numbers of students also meant an increase in resources within the university system, more assistants could be hired. Whereas it had been common for a professor to have one assistant at his disposition, in the 1960s he was able to hire two or even three (Boch 1972 pg. 194-195).
This development can be observed in the relative number of professorships (both ordinary and extraordinary, which declined even further from close to 30% to approximately 10%, whereas the percentage of affiliates/assistants rose from about 40% to more than 70% (Fig. 6). In addition, the relative number of lecturers also declined.

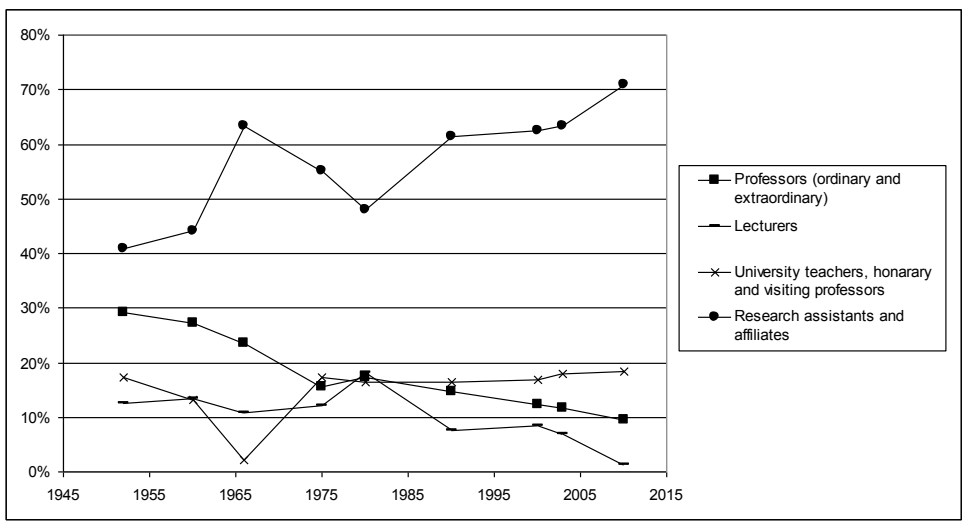


Figure 6. Distribution of academic positions 1952-2010. Source: *Statistisches Bundesam*t, 1953, 1966, 1969, 1976, 1982, 1992, 2004 & 2011. For 1953-1980 West Germany only.

Interestingly, the terms of employment for academics seem to have changed throughout the years. In the beginning of the 19th century only professors were paid by the universities, and private lecturers were only compensated by receiving lecture fees from students, as described previously. Later private lecturers would become paid staff as well, as would research assistants. Terms of employment have changed for research assistants and affiliates, though. In publications from 1980 and 1990, the German Statistical Office differentiated between affiliates hired on permanent and non-permanent contracts. Their data show that whereas in 1980 the distribution between the groups was about 50-50, in 1990 more than 70% of research affiliates were employed on temporary contracts (Statistisches Bundesamt 1982 & 1992).

Conclusion

In conclusion, the primary official academic position used to consist only of the professoriate. However, in the 19th century the position of private lecturer emerged, and these private lecturers were considered to be a pool from which the best could be picked to fill professorships. Looking at the relative numbers of different types of professors, it becomes apparent that already during the 19th century there was a trend towards a relative increase of non-permanent extraordinary professors and private lecturers compared to ordinary and extraordinary professors occupying permanent posts. During this period, aspiring scientists working towards a *Habilitation* did not occupy official positions yet. This changed with the emergence of larger laboratories, where research assistants would contribute to experimental science. The 1960s saw a disproportionate growth in the relative number of research affiliates/assistants, and them becoming the source for the new generation of professors. Nowadays, also PhD students often hold formal research assistantships or affiliateships, although they are still considered to be "trainee scientists".

Our data also suggest that the current "probationary" periods as research assistant or research affiliate have replaced a period in which trainee scientists were quite often employed on a permanent contract. However, in an even earlier period, trainees were still considered students when working towards a PhD dissertation or *Habilitation*. Arguably, the current situation constitutes an improvement in working conditions for people we would now call PhD students and postdocs over the 19th century, but is probably a disimprovement compared to the situation of research affiliates from 1960-1980.

Using our model of contemporary scientific careers (presented in Fig. 2) we can conclude that in the 19th century official academic positions in Germany were only found in what we would now call the higher echelons of academic career ladder and that over time positions below the professoriate became more proper academic positions of their own. Terms of employment such as salary and tenure have changed as well, as at first only professors were paid, while with the formalization of positions, researchers in lower positions were compensated as well. Finally, our results suggest the relative number of researchers on temporary contracts has increased.

Further lines of investigation

The results from our literature study suggest it has indeed become relatively more difficult to obtain a professorship in Germany over the course of the past two centuries, as the relative number of professorships (and even lectureships) has decreased compared to lower academic positions, such as research assistantships or affiliateships. However, as our research also shows, such positions may have existed informally before and only become formalized over time. As a further line of investigation we will look into the number of successful PhD defences and *Habilitationen in the same period* because it will give an estimate of the pool of researchers competing for higher academic positions.

In order to understand the development of the German academic career system even more completely, it would be interesting to know more about some of the other aspects of scientific careers discussed in our conceptual model in Figure 2, such as which control over material and personal resources researchers had at different career stages. In addition, the timing within the framework, i.e., time spent at different career stages, is of interest so we can find out whether probationary periods as trainee scientists have been lengthened.

Acknowledgements

We would like to thank Thed N. van Leeuwen (CWTS) for supplying calculations of total research output, and Ingrid Urlichs, Gertruda Huber (both German Federal Statistical Office) and Petra Langhein (German Council of Science and Humanities) for mailing archival statistics.
Cornelis A. van Bochove, Anthony F.J. van Raan, Inge C.M. van der Weijden (all CWTS) and Rosalie Belder (Rathenau Institute) are gratefully acknowledged for fruitful discussion of the study and manuscript. Finally, we thank three anonymous reviewers for their comments.

Endnotes

1. In this section we describe the academic career system of German universities, where most research takes place. In Germany, however, institutes of scientific societies, e.g., the *Max-Planck-Gesellschaft* and the *Fraunhofer-Gesellschaft* (Kreckel 2008) also play an important role in research. But as their career systems are based on those of universities, we will not describe them separately.
2. The correct German translation is *Doktorandin* for a female PhD student. For brevity we will from now on only use the male forms of German terms throughout the text.
3. The original German term was "*Kandidaten der wissenschaftlichen Forschung*". We are not completely sure what this term encompasses, but we presume these candidates are researchers striving for a *Habilitation*.

References

Ackers, L. (2008). Internationalisation, Mobility and Metrics: A New Form of Indirect Discrimination? *Minerva*, 4:411-35

Balen, B. van & Besselaar, P. van den (2007). *Universitaire onderzoeksloopbanen: een verkenning van problemen en oplossingen*. Rathenau Instituut, Den Haag.

Baruch, Y. (2004). Transforming careers: from linear to multidirectional career paths. *Career development international*, 9:58-73.

Baruch, Y. & Rosenstein, E. (1992). Career planning and managing in high tech organizations. *International Journal of Human Resource Management*, 3:477-96.

Ben-David, J. & Zloczower, A. (1961). The Idea of the University and the Academic Market Place. *European Journal of Sociology*, 2:303-314

Bock, K.D. (1972). *Strukturgeschichte der Assistentur: Personalgefüge, Wert- und Zielvorstellungen in der deutschen Universität des 19. und 20. Jahrhunderts*. Düsseldorf: Bertelsmann Universitätsverlag.

Centre for Science and Technology Studies (2012). *Leiden Ranking 2011/2012*. http://www.leidenranking.com/ranking.aspx. Retrieved 26 September 2012

Clark, B.R. (1993). *The Research Foundations of Graduate Education: Germany, Britain, France, United States, Japan*. Berkeley: University of California Press.

Dooris, M.J. & Guidos, M. (2006). *Tenure achievement rates at research universities*. In Annual Forum of the Association for Institutional Research. Chicago.

Enders, J. (1996). *Die wissenschaftlichen Mitarbeiter: Ausbildung, Beschäftigung, und Karriere der Nachwuchswissenschaftler und Mittelbauangehörigen an den Universitäten*. Frankfurt: Campusverlag.

Enders, J. & Kaulisch, M. (2006). *The Binding and Unbinding of Academic Careers* in *The Formative Years of Scholars*, Teichler, U. (ed.). London: Portland Press

Ferber, C. von (1956). *Die Entwicklung des Lehrkörpers der deutschen Universitäten und Hochschulen 1864-1954*, vol. III, edited by H. Plessner. Göttingen: Vandenhoeck & Ruprecht.

Fräßdorf, A., Kaulisch, M. & Hornbostel, S. (2012). Armut und Ausbeutung? Die Finanzierungs- und Beschäftigungssituation von Promovierenden. *Forschung & Lehre*, 622-624.

Kreckel, R.; Burkhardt, A.; Lenhardt, G.; Pasternack, P. & Stock, M. (2008). *Zwischen Promotion und Professur: das wissenschaftliche Personal in Deutschland im Vergleich mit Frankreich, Großbritannien, USA, Schweden, den Niederlanden, Österreich und der Schweiz*, edited by R. Kreckel. Leipzig: Akademische Verlagsanstalt.

National Research Council (2005). *Bridges to Independence: Fostering the Independence of New Investigators in Biomedical Research*.

Nerad, M. & Cerny, J. (1999). Postdoctoral patterns, career advancement, and problems. *Science*, 285:1533-1535.

OECD (2009). *MSTI Main Science and Technology Indicators*. http://stats.oecd.org/Index.aspx?DataSetCode=MSTI_PUB. Retrieved 26 September 2012

Peiperl, M. & Baruch, Y. (1997). Back to square zero: The post-corporate career. *Organizational Dynamics*, 25:7-22.

Statistisches Bundesamt (1953). *Statistische Berichte: Die Lehrpersonen und das wissenschaftliche Hilfspersonal an den wissenschaftlichen Hochschulen des Bundesgebietes und West-Berlin im Wintersemester 1952/53*. Stuttgart: Kohlhammer Verlag.

Statistisches Bundesamt (1966). *Bevölkerung und Kultur: Hochschullehrer und sonstiges wissenschaftliches Personal an den Wissenschaftlichen Hochschulen 1960*. Stuttgart: Kohlhammer Verlag.
Statistisches Bundesamt (1969). *Bevölkerung und Kultur: Hochschullehrer und sonstiges wissenschaftliches Personal an Wissenschaftlichen und Pädagogischen Hochschulen 1966*. Stuttgart: Kohlhammer Verlag.
Statistisches Bundesamt (1976). *Bevölkerung und Kultur: Personal an Hochschulen 1975*. Stuttgart: Kohlhammer Verlag.
Statistisches Bundesamt (1982). *Bildung und Kultur: Personal an Hochschulen 1980*. Stuttgart: Kohlhammer Verlag.
Statistisches Bundesamt (1992). *Bildung und Kultur: Personal an Hochschulen 1990*. Stuttgart: Metzler-Poeschel.
Statistisches Bundesamt (2004). *Bildung und Kultur: Personal an Hochschulen 2003*. Stuttgart: Metzler-Poeschel.
Statistisches Bundesamt (2011). *Bildung und Kultur: Personal an Hochschulen 2010*. Stuttgart: Metzler-Poeschel.
Waaijer, C.J.F. (in press). Careers in science: Policy issues according to *Nature* and *Science* editorials. *Scientometrics*.
Weber, M. (1946). *Science as a vocation* (pp. 129-156). Edited by H. H. Gerth and C. Wright Mills. From Max Weber: Essays in Sociology, New York: Oxford University Press.

ACADEMIC RESEARCH PERFORMANCE EVALUATION IN BUSINESS AND MANAGEMENT USING JOURNAL QUALITY CITING METHODOLOGIES

Evangelia A. E. C. Lipitakis[1] and John C. Mingers[2]

[1]*eael2@kent.ac.uk*
Kent Business School, University of Kent**,** Canterbury, Kent CT2 7PE, England

[2]*j.mingers@kent.ac.uk*
Kent Business School, University of Kent**,** Canterbury, Kent CT2 7PE, England

Abstract
The quality of the journals of the received citations of a set of publications for evaluating the quality performance in the case of individual researchers, research groups and academic departments is investigated. An adaptive model incorporating and examining variables, such as the quality of journals of the citing articles within the set of publications is considered. This hybrid bibliometric methodology, as an alternative methodology to citations counts, examines the quality of the journals of the citing articles that they have been published in and evaluates the quality of the received citations (citing articles) in the set of publications. The considered Journal Quality Citing index is used in combination with predetermined evaluation weight parameters in order to produce an efficient research quality evaluation methodology.
The new academic research quality methodology has been tested in three leading UK business schools in the fields of business, economics, management, OR and management science. The obtained numerical results indicate that the new research quality methodology can be also used in large scale academic research quality cases. The proposed Journal Quality Citing methodology can be considered as a quality performance evaluation approach by using efficient research journal quality ranking indicators based on weighted parameters.

Keywords
bibliometric indicators, hybrid bibliometric methodologies, journal quality citing index, journal quality ranking, research quality evaluation, quantitative methods

Conference Topic
Scientometrics Indicators (Topic 1)

Introduction
Various advanced bibliometric indicators have been used for measuring the productivity and impact of research at several academic levels, such as at the level of individual researchers, research groups and university departments (Van Raan,

2003; Hirsch, 2005; Mingers & Walsham, 2008; Waltman et al., 2011; Lipitakis, 2013). In this research work, we introduce a new research quality evaluation methodology that examines the number of citations of a publication as well as its source of publication. More specifically, this new methodology evaluates a publication by taking into account the following variables: (i) the total number of publications (productivity), (ii) the number of citations a paper has received (impact) and (iii) the quality of the journal that the citing article has been published in (citing quality factor).

Traditionally, citation based metrics have only considered the number of citation a paper has received and not the quality of the citations. All citations are treated equally regardless of whether the citations have been received by papers that have been published in high quality journals or low quality journals. This practice does not consider that citations from top journals should perhaps count for more than citations from poor quality journals.

We consider the following two basic questions, *"How can we measure the quality of a journal?"* and *"In what way can we incorporate the quality of the citation?"* One basic assumption of bibliometric analysis is that scientists with important and original material endeavour to publish their results vigorously in the open international journal literature. Although journal articles differ widely in importance, it has been noticed that authors in most cases seek to publish in the better and, if possible, the best journals (Glanzel, 1996; van Raan, 2004). Therefore one should consider whether citations from top journals are worth more than from citations from lower quality journals. In this article we answer the first question by considering a class of well know journal quality indicators. We test their efficiency at measuring the quality of a journal and determine how well the different journal quality indicators correlate with each other in order to decide which is the most suitable for our research study. In the second stage, we answer the second question by proposing an innovative bibliometric methodology. This methodology is based on weighted parameters affected by the journal quality for the evaluation of each given citation. Our research study aims to provide a comprehensive and complete study of research performance evaluation. We achieved our aim by using a quantitative method (citation count) that incorporates the quality of the journals of the citing papers in a hybrid methodology, based on existing and novel research performance quality assessments. Characteristic model problems and numerical results are also presented.

Citation Methodologies and Journal Ranking Indicators: A Synoptic Literature Review

The practical measure of publications is the share of the citation index (CI)-covered publications in the total research output. The CI refers to the following citation indices: Science Citation Index, Social Science Citation Index, Arts and Humanities Citation Index and the specialty citation indexes (such as CompMath, Biotechnology, Neuroscience, Material Science, Biochemistry and Biophysics) published by the Institute for Scientific Information (ISI/Thomson Scientific).

The citation analysis has been efficiently used by several academic and research institutions mainly for research policy making, visualization of scholarly networks, monitoring scientific developments, promotions, tenure, salary raise and grants decisions, etc. Several bibliometric indicators that have been used by various online citation databanks as complementary quality performance measures in the process of ranking scientific journals have been presented. The citation information can be used for academic research journal ranking by following several citation methodologies, such as (i) the direct citation data (Doyle et al., 1995), (ii) the citations indicators for journal ranking (Yu, 2005), (iii) a combination of peer review and citation studies (Kelly et al., 2009).

New methodologies for the evaluation of research quality performance are mainly focused on the efficient use of advanced bibliometric indicators and scientometric information (data), efficient adapted peer reviewing methods and certain hybrid methodologies for identifying excellent researchers of all types (Nederhof, 2006). Furthermore, new classes of hybrid methodologies incorporating both qualitative and quantitative advantageous elements are being developed. More specifically, certain quantitative elements are present in the class of qualitative methods, while qualitative elements appear also in the quantitative methods. The derivation of new and extended (modified) methodologies for measuring the research output quality of an individual researcher and/or a research unit in science and social sciences is a challenging research topic currently under investigation. Several bibliometric studies have focused more on the citation impact of a journal rather than that of the published research paper (Nederhof, 2006).

Various recent studies have focused on new methods of assessing scholarly influence based on journal ranking indicators. For example, a recent study examined the use of the Hirsch-type indices for the evaluation of the scholarly influence of Information Systems (IS) researchers (Truex et al., 2008). Another study assessed the impact of a set of IS journals, publications and researchers using a weighted citations count on authors and institutions where a publication with less authors receives more weight than a publication with more authors (Lowry, 2007). The presentation of a method, using advanced statistical methods, is based on cumulative n^{th} citation distributions on a publications ranking classification scheme has also been considered (Egghe, 2007) . Additionally, a study on the complementary Hirsch type index, the h_m, for the comparison of journals within the same subject field (Molinary and Molinary, 2008) has also been presented.

A class of journal ranking indicators on a large data set of journals in the wider area of business and management has been considered in our research study. These indicators measuring the quality of journals include the following: the journal impact factor (Garfield, 1972; Glanzel, 1996), the journal impact factor 5 years (Truex et al., 2008), the Immediacy index (Garfield, 1999; Glanzel and Moed, 2002), the Eigenfactor indicator (Bergstrom et al. 2008), the Hirsch index (h-index) (Hirsch, 2005; Mingers, 2008, Rousseau, 2008). An alternative bibliometric indicator for evaluating collections of publications is the so-called

Citations per Paper (CPP), while the Association of Business Schools (ABS) academic journal quality guide provides journal rankings for journals in the wider area of business and management (ABS, 2010). A series of statistical analyses using SPSS software to investigate which is the most efficient journal ranking indicator for the application of our proposed new methodology for research quality evaluation based on journal ranking indicators has been performed (Lipitakis, 2013).

In this article we propose a new methodology for research quality evaluation based on journal ranking indicators that assess the quality of the journals of the received citations of a publication. In the first stage of our research study we consider a class of well known journal ranking indicators for the measurement of journal quality. We test how efficiently they measure the quality of journals using a large data set of over 1,000 journals. We investigate the similarities and differences, how they relate with each other, compare the results and discover which journal ranking indicator is more suitable for the application of our proposed methodology.

Data Collection and Journal Ranking Methodology

The considered data collection includes a total of 1,151 journals in the wider area of business and management. The selected 1,151 journals that we decided to investigate consisted of all the journals included in both the ABS Journal Quality Guide 2010 and the Harzing Journal Quality List 2011 (Harzing, 2011). For each journal in our data set we have collected the numerical values of eight well known journal ranking indicators. The eight journal ranking indicators were: the total number of citations (TC), the citations per publications index (CPP), the 2 year journal impact factor (IF2), the 5 year journal impact factor (IF5), the immediacy index (II), the eigenfactor score (ES), the h-index and the ratings of the ABS journal ranking quality guide (ABS). All the journal ranking indicators, except the ABS journal quality ratings, are calculated and published by ISI Thomson Reuters and can be found in the Journal Citations Reports (JCR) section of the online citation database Web of Science (http://webofknowledge.com, accessed 05/11/12). For each journal, we recorded the corresponding numerical values of the eight journal ranking indicators available by WoS. The data collection was performed manually. It was a time consuming task and a lot of effort was made in order to secure the validity of the data.

The Web of Science online citation database has been used for the numerical values of the journal ranking indicators and the selected time period of data collection of the TC, CPP and h-index was 2000-2010. The selected starting year of the data collection for the IF, IF5, II, ES was 2010. For the data collection of the ABS journal quality rankings of the journals in our dataset we used the Association of Business Schools journal quality guide, 2010, version 4 (ABS, 2010).

The data collection of the 8 journal ranking indicators concerning TC, CPP, HI, IF2, IF5, II, ES and ABS Journal Quality Ratings, has been performed in a recent

research study (Lipitakis, 2013). A series of statistical analyses using the collected data, including related tables and graphs such as correlations, related scee plot, component/rotated component matrices, component plot in rotated space of the 8 journal ranking indicators, has been recently presented (Lipitakis, 2013). The proposed journal ranking indicator is the journal impact factor 5 years, based on the results of the corresponding statistical analysis (Lipitakis, 2013). Next, we propose a generalized version of a new research evaluation methodology using weighted parameters based on research journal quality ranking indicators.

The Journal Quality Citing Methodology

In the framework of our new approach to research quality performance evaluation of a research group or academic departments, we propose the following bibliometric methodology using weighted parameters based on research journal quality ranking indicators. The Journal Quality Citing (JQC) index calculates the weighted citations of a publication, incorporating in its algorithm the quality of the journals of the citing articles. The Journal Quality Citing approach is an alternative research quality evaluation methodology to citation counts. The JQC index suggests the use of a weighted parameter that will act as a quality evaluation parameter of the journal of the citing article of a publication. The generalized theoretical framework of our research evaluation approach, which uses weighted parameters based on research journal quality ranking indicators, is presented in the following text.

Following this quantitative approach, we evaluate each received citation of a set of publications. We note that the sum of the citations will be affected by the corresponding research journal's quality weight ε_j, in such a way we can obtain the Journal Quality Citing (JQC) index that can be defined as follows:

$$\text{JQC index} = \sum_{i=1}^{\max NCD} \left(\sum_{j=1}^{\max NCG} \varepsilon_j . c_{i,j} \right), \quad j=1,2,\ldots, \text{maxNCG} \text{ and } i=1,2,\ldots, \text{maxNCD} \tag{1}$$

Table 1. Definition of the Journal Quality Citing index (numbers of cited and citing papers and their upper bounds).

i	Number of *cited* papers (Examined research output)
j	Number of *citing* papers (Received citations)
maxNCD	Number of total *cited* papers (Total examined research output)
maxNCG	Number of total *citing* papers (Total number of received citations)

Note that in equation (1) the term $c_{i,j}$ corresponds to the citations received by the publication/paper-i of an individual researcher, with i=1,2,…, maxNCD, where maxNCD is the max number of papers of individual researcher. The first index *i* of $c_{i,j}$ denotes the number of publications *i* of the individual researcher, while the

second index j denotes the number of citing papers *j* of the other researchers that have cited in their papers the above paper *i* of the individual researcher.

JQC Methodology: The Case of the 5 year Impact Factor as weighted parameter

In this section we propose and apply a modified version of the Journal Quality Citing index which normalizes for subject field and time, based on our selected weighted parameter; the 5 year impact factor. The proposed methodology can be used as an alternative quantitative research quality evaluation approach for the assessment of the research output of academic departments to citation counts. In the following sections, we test the proposed new methodology using the research output of three leading UK business schools.

The purpose of the JQC indicator is to investigate the citations a paper has received and evaluate these by allocating to each citation a different weight, according to the quality of the journal they have been published in. So far, the traditional citation is a metric that reflects the impact of a paper, by counting how many times the given paper has been used in another researcher's work. The JQC indicator weighs citations according to the quality of the journal they have been published in. Its aim is to evaluate the number of publications (productivity), the number of the citations a publication has received (impact) and the received citations by assessing the quality of the journal of the citing articles the so-called *citing quality factor* (Lipitakis, 2013).

In the following we propose a modification to the JQC index presented in equation (1) that uses the 5 year journal impact factor as a weighted parameter. Additionally, we explain why these modifications are necessary for the 5 year Impact Factor as the weighted parameter ε_j.

The proposed modified Journal Quality Citing index is the following:

$$\text{JQC index} = \sum_{i=1}^{\max NCD} \left(\sum_{j=1}^{\max NCG} \varepsilon_{IF5j} \cdot c_{i,j} / \text{Field Mean } \varepsilon_{IF5i} \right), \qquad (2)$$

where $j=1,2,\ldots,$ maxNCG and $i=1,2,\ldots,$maxNCD, with

ε_{IF5} is the numerical value of IF5 of the journal (as found in WoS- JCR) and Field Mean IF5 is the

Mean number of the IF5 of all the journals within a WoS subject field.

The proposed JQC index presented in equation (2) includes two major modifications compared to the general theoretical approach presented in the previous section; the 5 year journal impact factor as our selected weighted parameter ε and the variable Field Mean $\varepsilon_{IF5,i}$. The first modification is that we have substituted the weighted parameter ε with the 5 year journal impact factor. The 5 year journal impact factor acts as a quality evaluation parameter of the journal of each citing article. The second modification Field Mean $\varepsilon_{IF5,i}$ is the mean numerical value of the 5 year impact factor of a journal in a given WoS field. It can be calculated by the sum of the 5 year journal impact factors of all available journals divided by the number of all journals with a 5 year impact

factor. It should be noted that not all journals are provided with a 5 year journal impact factor by WoS. In our calculations of the Field Mean $\varepsilon_{IF5,i}$ we have divided by the number of journals that actually have a corresponding 5 year journal impact factor. To calculate the JQC indicator the sum of the 5 year impact factors of the citing journals of a publication *i* is divided by the average 5 year impact factor in the given field of the publication *i* to obtain field normalized results

The JQC indicator gives more weight to citations of papers that have been published in high quality journals, compared to citations that have been published in low quality journals. Therefore, the JQC indicator produces weighted citations and allows us to compare these with the actual citations of the same set of publications. If the number of the weighted citations of the JQC index is smaller than the number of the actual citations of the publication, this means that the majority of the citations have been published in low quality journals. If the JQC index is larger, it means that the majority of the citations have been published in high quality journals. If we assume that high quality journals publish significant and original scientific research then the weighted citations JCR indicator reflects both the productivity and impact of a set of publications in a given field.

Next, we state some clarifications concerning the Journal Quality Citing methodology and JQC indicator. The Journal Quality Citing methodology proposes a quality performance evaluation approach, using weighted parameters based on research journal quality ranking indicators. It is not confined to the use of the 5 year journal impact factor as the only weighted parameter, but can be used with any alternative efficient journal quality ranking indicator. A further extension of the methodology, namely, examining larger datasets of academic research output and alternative weighted parameters is currently under investigation by the author. Furthermore, in our research study we consider only one publication type; journal articles. When we refer to 'publication(s)' or 'research output', we mean journal article(s). This mainly because in this research study we investigate the use of 5 year impact factor, as a journal ranking weighted parameter, which is currently only available for journals by JCR/WoS. A more comprehensive extension of the JQC methodology that includes a classification of various publication types (such as books, book chapters, conference proceedings, reports, working papers, etc) is currently under investigation by the authors.

Certain important issues concerning the usage of WoS online citation for the numerical experimentation (Van Raan, 2003; Mahdi, D'Este & Neely, 2008; HEFCE, 2008; Mingers & Lipitakis, 2010; Waltman et al., 2011; Mingers & Lipitakis, 2013) and the impact factor as a measure of quality for the evaluation of journals (Moed & van Leeuwen, 1995; Todorov & Glanzel, 1988) have been presented in various related publications.

Data Collection and JQC Methodology

For the numerical experimentation of the Journal Quality Citing methodology and the calculation of the Journal Quality Citing indicator we have used the research output of three leading UK business schools, namely the Cambridge Judge

Business School (JBS), the Liverpool Management School (LMS) and the Kent Business School (KBS) in the fields of Business, Economics, Management and O/R & Management Science. We present the data collection and methodology for our numerical experimentation of the application of the Journal Quality Citing methodology for the time period 2001-2008. For our research study we have considered only the journal articles that were available in the online citation database Web of Science (WoS).
In the following table the research output of the three UK business schools in its various forms can be shown.

Table 2. Overview of the research output of the three UK business schools

2001-2008	*JBS*	*LMS*	*KBS*	*TOTAL*
Total Research Output (various publications types)	1,681	1,191	1,025	3,897
Total Journal Articles	679	593	473	1,745
Total Journal Articles found in WoS	341	266	314	921
Total Journal Articles found in WoS classified as BEMO/R*	240	108	158	506
Total Number of WoS *Citations* in BEMO/R* Journal Articles	3,683	908	1,496	6,087
Total Number of (different) *Citing Journals* of BEMO/R* Journal Articles	783	304	456	1,543

*by BEMO/R, we mean the subject areas of Business, Economics, Management and O/R & Management Science.

In the framework of the application of the Journal Quality Citing methodology we considered the research output of the three business schools, which consisted of 506 journal articles in the area of BEMO/R and their 6,087 corresponding citations as found in WoS for the time period 2001-2008. The data collection was performed manually. This was a very time consuming task due to the volume of data and our efforts to secure the validity of our data.
Then we proceeded by allocating the available numerical value of the IF5 to the corresponding journals of the 6,087 citing publications. For the journals that did not have an IF5 numerical value in WoS, we allocated the *Field Median IF5*. The Field Median IF5 is the median value of the IF5 of all the journals in the journal list, for a given subject area (BEMO/R). Following this approach, we obtained 6,087 citing publications and the corresponding IF5 for their journals. Next, we proceeded by calculating the *Field Mean IF5* for each one of the BEMO/R subject categories. The Field Mean IF5 is used as the denominator of the Journal Quality Citing indicator and is calculated by dividing the sum of the total numerical values of all available IF5 by the total number of the journals with an available IF5, in a given subject area journal list. Note that in the case of the cited publications with more than one WoS fields, the Field Median IF5 and the Field Mean IF5 of a publication i, can be computed by considering the harmonic average (Lipitakis, 2013).

Finally, we calculated the Journal Quality Citing Indicator for the research output of the three UK business schools for the time period 2001-2008 in the fields of BEMO/R.

Numerical Experimentation and Results

In this section we will present the indicative numerical experimentation and obtained results of the application of the Journal Quality Citing methodology in three leading UK business schools in the subject areas of business, economics, management and O/R & management science for the time period 2001-2008. In Tables 3 and 4 we can see the results of the calculated Field Mean IF5 for all four subject fields and publications with overlapping fields. The JCR journal lists in certain fields are currently somewhat limited. For instance, the O/R & management science area includes 75 journals (from which 63 had an IF5).
The WoS subject area classification and journal coverage in the social sciences area can be considered poor, usually less than 50%, in comparison to other citation databases such as Google Scholar (GS), (Moed, 2005; Mingers & Lipitakis, 2010) but at this point GS does not include a subject area classification. If the Field Mean IF5 and Field Median IF5 could be calculated from a larger journal subject classification environment perhaps they could produce more comprehensive and indicative results.

Table 3. Field Median IF5 for BEMO/R WoS subject fields

WoS Fields	*Field Median IF5*
Business	2.26
Economics	1.25
Management	2.32
O/R & Management Science	1.36

Table 4. Field Mean IF5 for BEMO/R WoS subject fields and overlapping fields

WoS Fields	*Field Mean IF5*
Business	2.70
Economics	1.59
Management	2.93
O/R & Management Science	1.58
Business/Economics	2.00
Business/Management	2.81
Economics/Management	2.06
Business/Economics/Management	2.24
Management/ O/R & Management Science	2.05

In Table 5 we can see a part of the obtained results by the application of the Journal Quality Citing methodology and the weighted citations of the Journal

Quality Citing index. The number of publications reflects the productivity of each department in the BEMO/R fields.
If we compare the weighted citations of the JQC index with the actual number of citations, we can see that they have a different numerical value. That happens because the JQC index has incorporated in its algorithm the quality of the journals of the citing articles and translated the quality of the journals into a number of citations that are weighted by journal quality. In the case that the number of the weighted citations of a set of publications is larger than the number of actual citations that means that, overall, the set of publications has received citations that have been published in prestigious and high quality journals in a given subject area. If the number of the weighted citations of a set of publications is smaller than the number of actual citations, it means that the set of publications has received citations that have been published in lower quality journals. The number of weighted cpp is calculated by dividing the total number of weighted citations by the number of publications. We can clearly see that the weighted cpp of JBS and KBS have increased and that LMS weighted cpp has decreased when compared with the unweighted cpp.

Table 5. A comparison between the actual citations and weighted citations and the cpp index and JQC (weighted) cpp index.

	Publications	***WoS Citations*** *(unweighted citations)*	***CPP*** *(unweighted cpp)*	***JQC index*** *(weighted citations)*	***JCQ CPP*** *(weighted cpp)*
2001-2008					
JBS	240	3,683	15.35	3,718	15.49
LMS	108	908	8.41	793	7.34
KBS	158	1,496	9.47	1,524	9.65

Our results presented in Table 5 reveal that for the considered time period 2001-2008, the weighted citations for the research output of JBS are higher than the actual number of citations. That suggests that JBS research output tends to be cited in papers that are published in better quality journals than the average publication. The same applies for KBS which has a larger number of weighted citations than the actual citations. LMS has lower weighted citations than the actual citations. LMS is the department with the least publications and citations in the area of BEMO/R while at the same time is quite active in Health Care Sciences and other related medical areas and has spend a share of its research output in subject areas that we have not included in this study. However the results show that its research output tends to be cited in average/lower quality journals.
Finally, it would be useful to compare the results of this study with the results of the 2008 UK Research Assessment Exercise (RAE). According to the 2008 RAE assessments, each university department was evaluated on a 4-point scale, where grade four is “world leading quality” and grade one “national” quality (RAE,

2008). The RAE 2008 results consist of the Business and Management Studies Unit of Assessment (UoA) that includes 100 selected submissions per institution in the fields of business, economics, management, management science and any other field or subfield aligned to business and management, for the time period 2008-2001 (RAE, 2008). The data we have used for the application of the JCQ methodology comprise the research output in the WoS areas of BEMO/R for the time period 2001-2008. Our sample is not identical to that submitted to the RAE 2008, but it is directly comparable because it examines published research output of same subject area and time period, within a larger size of set of publications of the examined HEIs. The results showed that JBS scored 3.05, KBS scored 2.50 and LMS scored 2.45. These results agree with the results of the application of the Journal Quality Citing methodology, for the time period 2001-2008, so based on a small sample of the application of the methodology, we can say that the Journal Quality Citing index produced similar results with to the RAE peer review assessment in terms of ranking the business schools.

Table 5. A comparison between the results of RAE 2008, cpp index and weighted cpp of the Journal Quality Citation methodology 2001-2008

	RAE scores 2001-2008	CPP (unweighted cpp)	JCQ CPP (weighted cpp)
JBS	3.05	15.35	15.49
KBS	2.50	9.47	9.65
LMS	2.45	8.41	7.34

Conclusions

The purpose of the proposed methodology is to examine the quality of the received citations of a set of publications and evaluate the quality of the received citations according to the quality of the journal they have been published in. This is a research quality performance methodology alternative to citation counts. The considered JQC index is used in combination with predetermined evaluation weight parameters in order to produce an efficient research quality evaluation methodology. The proposed academic research quality methodology has been tested in three leading UK business schools in the fields of business, economics, management and O/R & management science. The obtained numerical results have indicated that the new research quality methodology can be efficiently used in large scale academic research quality cases.

The proposed approach is a research quality performance methodology alternative to citation counts. The differences between the research outputs of the three UK business schools have been examined. We have applied the JQC indicator that has weighted the research output according to the quality of the journal they have been published in. We have used the 5 year journal impact factor in order to give more weight to the citing articles that have been published in journal with higher 5 year journal impact factors. The results for field and time have been normalized. The obtained results showed that this methodology magnifies the existing

differences between the schools and discriminates better the research outputs in the given fields. Although the applicability has been demonstrate, we state that further numerical experimentation is needed in order to demonstrate the full advantages of the proposed methodology. More specifically, we need to include more data in our further numerical experimentations to examine a greater dispersion between different academic environments. Furthermore, we state that in future work we are planning to consider the use of wider, moving time periods for the more efficient application of the methodology and comparison of results.
At this stage the proposed Journal Quality Citing methodology investigates the quality of the journal of the received citations of a set of publications for the evaluation quality performance evaluation in the case of an academic department or a researcher. An extensive modified model that incorporates and examines more variables such as the quality of the journals of the publications of the academic/researcher's research output, as well as publications and journals of the cited references of the publications is currently under investigation. Furthermore an extended classification that includes other publication types (such as books, conference papers, reports, working papers, etc.) is considered.

References

ABS (2010). Academic journal quality guide (Association of Business Schools), Version 4, http://www.associationofbusinessschools.org/

Bergstrom, C., West J.D. and Wiseman M.A. (2008). The Eigenfactor Metrics. *The Journal for Neuroscience*, 28, 11433-11434

Doyle J. and Arthurs A. (1995). Judging the quality of research in business schools. *Omega, International Journal of Management Science*, 23, 257–270

Egghe L. (2007). Probabilities for encountering genius, basic, ordinary or insignificant papers based on the cumulative n^{th} citation distribution, *Scientometrics*, 70, 167-181

Gadfield E. (1972). Citation analysis as a tool in journal evaluation – journals can be ranked by frequency and impact of citations for science policy studies, *Science*, 178, 471–479

Gadfield E. (1999). Journal Impact Factor: A Brief Review, *Canadian Medical Association Journal,* 161, 979-980

Glanzel W. (1996). A bibliometric approach to social sciences, national research performance in 6 selected social science areas 1990-1992, *Scientometrics,* 35, 291-307

Glanzel W. and Moed H.F. (2002). Journal impact measures in bibliometric research, *Scientometrics,* 53, 171-193

Harzing A. (2011). Journal Quality List, Forty-first edition, http://www.harzing.com

HEFCE (2008). Counting what is measured or measuring what counts. In: HEFCE.

Hirsch J. (2005). An index to quantify an individual's scientific research output, *Proceedings of The National Academy of Sciences of the U.S.A.*, 102:46, 16569-16572

Lipitakis Ev. A.E.C. (2013). The use of bibliometric methods in evaluation research performance in Business and Management: A study of three UK Business Schools, Doctoral Thesis, KBS, University of Kent, Canterbury, England

Lowry P.B., Karuga G.G. and Richardson V.J. (2007). Assessing leading Institutions, Faculty and Articles in premier Information Systems Research Journals, *Communications of the Association for Information Systems*, 20, 142-203

Madhi S., D' Este P. and Neely A. (2008). Citation counts: are they good predictors of RAE scores? In. London: AIM Research

Mingers J. (2008). Measuring the research contribution of management academics using the Hirsch-index, *Journal of Operational Research Society*, 60 (8), 1143-1153

Mingers J. and Harzing A.W. (2007). Ranking journals in business and management: a statistical analysis of the Harzing data set, *European Journal of Information Systems,* 16, 303-316

Mingers J. and Lipitakis Ev. A.E.C. (2010): Counting the Citations: A Comparison of Web of Science and Google Scholar in the Field of Business and Management, *Scientometrics*, 85, 613-625

Mingers J. and Lipitakis Ev. A.E.C. (2013). Evaluating a Department's Research: Testing the Leiden Methodology in Business and Management, *Information Processing and Management* (to appear)

Mingers J. and Walsham G. (2010). Towards ethical information systems: The contribution of discourse ethics, *MIS Quarterly*, 34 (4), 833-854

Moed H.F. and Van Leeuwen T.N. (1995). Improving the accuracy of the Institute for Scientific information's journal impact factors, *Journal of the American Society for Information Science*, 46, 461-467

Moed H.F.. Citation Analysis in Research Evaluation, Dordrecht (Netherlands), *Springer*, 2005

Molinary J.F. and Molinary A. (2008). A new methodology for ranking scientific institutions, *Scientometrics*, 75, 163-174

Nederhof A.J. (2006). Bibliometric monitoring of research performance in the social sciences and humanities, *Scientometrics*, 66, 81-100.

RAE (2008). Research Assessment Exercise, http://www.rae.ac.uk/, (accessed, 16/10/12)

Rousseau R. (2008). Journal evaluation by environmental and resource economists: a survey, *Scientometrics,* 77, 223-233

Todorov R. and Glanzel W. (1988). Journal Citation Measures: A Concise Review, *Journal of Information Science*, 14, 47-56

Truex III D.P., Cuellar M.J. and Takeda H. (2008). Assessing scholarly influence: Proposing new metrics, *ICIS 2008 Proceedings*, 1-19

Van Raan A.F.J. (2003). The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments, *Technology Assessment – Theory and Practice*, 1, 20–29.
Van Raan A.F.J. (2004). Measuring Science: Capita selecta of current main issues, in Moed-Glanzel (eds): Handbook of quantitative science and technology research group, *Kluwer Academic*, 19-50
Waltman L., Eck N.J., Leeuwen T.N., Visser M.S. and Van Raan A.F. (2011). Towards a new crown indicator: an empirical analysis, *Scientometrics,* 87, 467-481

ACCESS TO UNIVERSITIES' PUBLIC KNOWLEDGE: WHO'S MORE REGIONALIST?

Manuel Acosta[1] · Joaquín M. Azagra-Caro[2] · Daniel Coronado[1]

[1] *Facultad de Ciencias Económicas y Empresariales, Universidad de Cádiz, c/Duque de Nájera, 8, 11002 Cádiz, Spain*

[2] *jazagra@ingenio.upv.es*
INGENIO (CSIC-UPV), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
European Commission, Joint Research Centre (JRC)-Institute for Prospective Technological Studies (IPTS), Edificio Expo, C/Inca Garcilaso 3, E-41092 Sevilla, Spain

Abstract
Patent citations are widely used indicators of knowledge flows. One originality of this paper is to track not patent-to-patent or paper-to-patent citations as usual but university-to-firms' patent citations. Another one is not to explain citations as a function of distance between cited and cited regions but to explain regional and non-regional citations as a function of the characteristics of knowledge supply and demand in the region –a complementary approach to the geography of knowledge flows. Using a dataset of European Union regions in years 1997-2007, we find that fostering university R&D capacity enlarges the attractiveness of the local university knowledge base for firms in the region. However, it has a trade-off, since firms will take less resource to university knowledge produced elsewhere. It is possible to compensate this through increases in local business absorptive capacity, which will enable firms to access university knowledge outside the region.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 6) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction
Codified university knowledge such as patenting and scientific publications may have an influence on innovation in regions because of the flow of technological knowledge between universities and firms. This flow of knowledge can take place through a variety of interaction channels between academics and firms (by reading the patent and/or a scientific paper, or via direct conversation or informal meetings with the academic inventor or researcher, through the hiring of graduate or doctorate students, etc.). However, sometimes there is a mismatch between the university-codified knowledge produced in the region and the firms' acquisition of that knowledge. This paper explores the causes explaining why firms use the inward regional university knowledge and why they acquire that knowledge elsewhere outside the region.

Our interest for this topic is motivated for several facts. First, the regional focus for analyzing the acquisition of knowledge from universities is suitable given the growing role of policies at regional level to achieve the European Research Area (ERA). The program to develop the ERA is primarily a partnership between the European Commission and the member states; but the Commission, the Council and the Committee of the Regions all see a role for the regions in the ERA, as a result of a greater involvement of the regions in research and innovation policies (Charles et al., 2009). Second, some regions generate scientific and technological knowledge in their universities, but sometimes regions producing that codified knowledge are unable to fully absorb it or exploit it (Caragliu and Nijkamp, 2012). Third, despite the importance of knowing what explains the acquisition of university knowledge outside or inside the region for regional policy, only a few recent papers have analyzed this topic. For example, Acosta et al. (2011b), study the outside dimension of research collaboration patterns; Abramo et al. (2010) addresses both dimensions for a single country; and Azagra (2012) takes a large number of countries and years to analyze the national patterns of accessing public knowledge. None of this previous research centers on a regional perspective for EU27.

Particularly, two groups of hypotheses are tested about the role of absorptive capacity for academic knowledge, and the importance of the regional presence of regional scientific and technological opportunities on the firms' acquisition of university knowledge. For this purpose we draw on a regional sample of around 6,000 university references (both patents and papers) contained in 4,000 firms' patents across EU27 regions for 1997-2007. The econometric results show a significant role of the university opportunities to increase the acquisition of inward university knowledge, while the firm absorptive capacity is not relevant in explaining the use of knowledge by the firms located in the same region where the knowledge is produced. However, the outward acquisition of knowledge is positively explained for the absorptive capacity and negatively for the regional opportunities for spillovers.

The paper is organized as follows. Section 2 reviews the relevant literature and establishes the hypotheses. Section 3 discusses the empirical framework. Section 4 explains the data and provides summary statistics. Section IV presents the empirical results. We briefly summarize the conclusions, policy implications, and discuss future research in the final Section.

Literature review and hypotheses

This paper has a regional focus, but the proposal of hypotheses describing the causes of the regional acquisition of university knowledge requires a discussion at firm level. In this respect, this review starts by including some ideas about the open innovation paradigm that helps to classify the process of acquisition of university knowledge and to explain why firms engage in acquiring external knowledge. Afterwards this literature is linked with the empirical background on the geographical dimension of knowledge sourcing, which discusses the role of

proximity in the process of absorbing knowledge. Finally, we take into account the supply side perspective by referring to some papers stressing the relevance of the availability and characteristics of university knowledge for the process of acquisition of knowledge by firms to take place.

The process of incorporating new knowledge into firms from other institutions such as universities has been recently discussed in the frame of the open innovation paradigm. According to the open innovation model, firms incorporate external as well as internal ideas, and internal and external paths to market, as they look to advance their technology (Chesbrough, 2003, 2006). Since Chesbrough's seminal work, a considerable number of papers have analysed the open innovation process at various levels, including at firm, industry and region levels (see van de Vrande et al., 2009 for a review), and new trends and directions have been identified (see, for example, Gassmann et al., 2010). This literature provides an analytical framework to explain the process of acquisition of knowledge by firms.

The open innovation ideas assume acquiring knowledge from different sources. Dahlander and Gann (2010) developed an analytical framework by structuring the process of open innovation in two dimensions: inbound/outbound (see also Chesbrough, 2006, Gassmann and Enkel, 2004) and pecuniary/non-pecuniary. Inbound open innovation is an outside-inwards process and involves opening the innovation process to knowledge exploration. External knowledge exploration refers to the acquisition of knowledge from external sources. By contrast, outbound open innovation is an inside-outwards process and includes opening the innovation process to knowledge exploitation. Open innovation is then a broad concept encompassing different dimensions and it is useful to classify the type of acquisition of knowledge addressed in this paper. According to this literature, the firms' acquisition of knowledge from university outputs such as patents open to public and scientific papers is a kind of inbound and non-pecuniary process of innovation. From a spatial perspective, regions exhibit similar patterns to firms; innovative success might depend on the appropriate combination of knowledge inputs from local and regional as well as national and global sources of knowledge (Kratke, 2010); moreover as pointed by Cooke et al., (2000) and Cooke (2005), it is impossible to discuss innovation processes and policies without reference to the interactions of local–regional, national and global actors and institutions.

The empirical evidence on businesses' external knowledge sourcing through university spillovers has revealed two facts: First, there is a geographical dimension in the external process of knowledge acquisition from universities. The relevant role of distance has been tested largely by a long list of empirical papers on university spillovers (e.g. Anselin et al. 1997, 2000; Feldman and Florida 1994; Fischer and Varga 2003; Jaffe 1989; Varga 1998). The main finding of these studies is that knowledge spillovers from universities are localized and contribute to higher rates of corporate patents or innovations in geographically bound areas. Moreover, knowledge spillovers are usually "confined largely to the

region in which the research takes place" (Hewitt-Dundas, 2011). Second, spillovers from neighbouring sources of knowledge inside the region or other ways of acquisition of knowledge outside the region do not occur automatically. A certain degree of "absorptive capacity" (Cohen and Levinthal, 1990) is necessary; that is, firms must have the ability to recognise the value of new, external information, assimilate it, and apply it" (Cohen and Levinthal, 1990). This means that factors hampering the open innovation process such as culture, modes of organization, bureaucratic elements, lack of resources, etc. (van de Vrande et al. 2009) would be encompassed in the broad concept of absorptive capacity. Using the terminology of the open innovation paradigm, absorptive capacity is "a pre-condition for organising inbound open innovation activities" (Spithoven, 2011).

In the light of the above arguments, the open innovation paradigm suggests that firms incorporate external as well as internal ideas to advance their technology. These ideas include knowledge from external institutions such as universities inside and outside the region where the firm is located, but a certain degree of absorptive capacity for university knowledge seems to be one of the main requirements for firms to absorb university knowledge through spillovers.

As pointed out above, one of the main findings of the empirical university spillover literature is that distance is a relevant factor for explaining the use (by firms) of academic knowledge produced in the same area or region where firms are located. However, several papers suggest that knowledge sourcing occurs at a variety of different spatial scales such as supra-regional and global connections that might be equally important to those in the region in order to get access to external knowledge sources (Arndt and Sternberg, 2000; Kaufmann and Todtling, 2001; Bathelt et al., 2004). Davenport (2005) reports some research that has analyzed how many firms do not acquire their knowledge from within geographically proximate areas, concluding that there are some factors that may work against geographically proximate knowledge-acquisition activities such as the role of foreign firms and multi-nationals, or firms working on some specific kind of technologies. Boschma (2005) argues that although geographical proximity facilitates interaction and cooperation for acquisition of knowledge, it is neither a prerequisite nor a sufficient condition for interactive learning to take place; other forms of proximity may frequently substitute for geographical proximity. Cargliu and Nijkamp (2012) recently explore the relationship between outward knowledge spillovers (measured as total factor productivity) and regional absorptive capacity for a sample of European regions,. Their result show that lower regional absorptive capacity increases knowledge spillovers towards surrounding areas, hampering the regions' capability to decode and efficiently exploit new knowledge, both locally produced and originating from outside. One of the main reasons explaining why some firms relies on proximity rather than in long distance sources of knowledge seems to be the grade of absorptive capacity: when firms' absorptive capacity is low, geographically proximate collaborations may be their only option. In contrast, high absorptive capacity enabling firms to

collaborate for innovation at greater geographical distance (Drejer and Vinding, 2007; De Jong and Freel, 2010).
This literature suggests two important conclusions: first, distance is not an obstacle for many firms with high absorptive capacity to acquire knowledge from other regions. Second, the acquisition of knowledge from surrounding areas is easier for firms with lower absorptive capacity. This discussion leads to the following two hypotheses. Both hypotheses concern the influence of the absorptive capacity on the use of university knowledge produced inside and outside the region:

Hypothesis 1: The acquisition of codified knowledge in form of patents and papers produced by universities inside the region is negatively related to the absorptive capacity for academic knowledge of firms in the region.

Hypothesis 2: The acquisition of codified knowledge in form of patents and papers produced by universities outside the region is positively related to the absorptive capacity for academic knowledge of firms in the region.

The above hypotheses concern the firm capacity to acquire university knowledge, but academic knowledge is a flow; we need to take into account the other party in the game: universities. The question is to what extent the availability, quality or characteristics of the knowledge produced in universities stimulate or hinder the acquisition of inward and outward regional academic knowledge? In this respect, some empirical research has stressed the role of universities to encourage the flow of knowledge between universities and firms at regional level. Audrestch and Feldman (1996) find a positive relationship between ‘‘local university research funding’’ and ‘‘local industry value-added’’ at the state level. Their results indicate the relative economic importance of new knowledge to the location and concentration of industrial production. Zucker et al. (2002) relate the input “number of local research stars” to the output “number of new local biotech firms” and examine the variance in this relationship across geographic space at the economic region level. They find that the number of local stars and their collaborators is a strong predictor of the geographic distribution of US biotech firms in 1990. Branstetter (2001) identifies a positive relationship between ‘‘scientific publications from the University of California’’ and ‘‘patents that cite those papers’’, also at the state level. In another more recent paper Branstetter (2005) points out that the more rapid growth in the intensity with which U.S. patents cite academic science suggests a response to new technological opportunities created by academic research.
Other related literature on firm formation/location also suggests the importance of the characteristics of the academic knowledge for the spillovers to take place in the region. For example, Audretsch et al. (2004) focused on whether knowledge spillovers are homogeneous with respect to different scientific fields. They found that firms’ locational-decision is shaped not only by the output of universities (for

instance, students and research), but also by the nature of that output (that is, the specialized nature of scientific knowledge). Audretsch and Lehmann (2005) concluded that universities in regions with greater knowledge capacity and higher knowledge output also generate a larger number of technology start-ups. Several empirical papers in different spatial contexts point to the potential positive relationship between local university R&D expenditures and the number of newly created high technology firms (e.g. Harhoff, 1999 for Germany; Woodward et al., 2006 for US; Abramovsky et al., 2007, provide evidence on the extent business sector R&D activity is located near high quality university research departments in Great Britain; Acosta et al. 2011a found a significant relationship between some university outputs and new firm formation for the case of Spain).
According to this literature, we expect that a territorial environment with a well-established university presence increases the opportunities for the companies to access and absorb relevant new scientific knowledge more easily, in comparison with other companies located in regions with weak university capacities. At the same time, firms in regions with low technological and scientific opportunities will acquire academic knowledge elsewhere outside the region. This reasoning leads to the following two hypotheses:

Hypothesis 3: The acquisition of codified knowledge in form of patents and papers produced by universities inside the region is positively related to the university capacity to produce scientific and technological knowledge in the region.

Hypothesis 4: The acquisition of codified knowledge in form of patents and papers produced by universities outside the region is negatively related to the university capacity to produce scientific and technological knowledge in the region.

Model and variables

The basic model for testing our hypotheses relates the acquisition of university knowledge (UKA) by firms in a region to two main explanatory factors: the absorptive capacity (AC) and the availability of university knowledge in the region (U).
The regional function is given in general form as:

$$UKA_{it} = f(AC_{it}, U_{it}) \quad \text{for } i = 1, 2, ..., N$$

Where the subscripts "i" and "t" refer to region i and time t, respectively. We may call this equation the University Knowledge Acquisition Function (UKAF), and it concerns the activity in which firms in a region capture knowledge from inward and outward regional university knowledge (university knowledge produced in universities located in the region or elsewhere). To fully explain the knowledge acquisition we have extended this function in two ways:

- The model should control for the technological specialization and regional technological size. Although -to our knowledge- there is not empirical research on the effects of technological diversification (or specialization) on the acquisition of university knowledge, regions specialized in high technology might rely on external knowledge rather than on regional internal knowledge. For example, some authors (E.g. Klevorick, 1995, Acosta and Coronado, 2003, Laursen and Salter, 2004) suggest that in some industrial sectors, the relationship between universities and industrial innovation appears to be a tight one, such as in biotechnology, while in others such as textiles it appears to be weaker. On the other hand, European regions differ in their size. To avoid spurious correlation the model must control by the technological size of inward outward knowledge (using for example the size of the patent portfolio in each region).
- Regions are grouped in countries and consequently some correlation is expected across regions of the same country. For example, national innovative measures, incentives -or more general firms' policies- influencing the regions of the whole country. The presence of higher-order hierarchical structures with different characteristics (regions are grouped in countries) point to the multilevel nature of the factors influencing the acquisition of university knowledge.

We may reformulate the initial model by including these additional factors in an extended UKAF:

$$UKA_{git} = f(AC_{git-2}, U_{git-2}, S_{git-2}, Z_{git-2}, e_{gt}, u_{git}) \text{ for } i = 1,2,...,N \;\; g = 1,2,...,G$$

Where g indexes the group or cluster. S controls for the technological specialization of the region and Z for its size. e is an unobserved cluster-effect capturing the regional influences of the group (country) on the regional acquisition of inward and outward knowledge and u is the idiosyncratic error. Finally, the empirical estimations also include some dummies to capture temporal fixed effects. All the explanatory variables consider a two-year lag.[2]
The following paragraphs explain how we have measured our variables.
Dependent variables. We consider two dependent variables in two separate models:

- The acquisition or use of inward regional university knowledge is captured by the number of citations in firms' patents to universities located in the same region where the firm is established.
- The acquisition or use of outward regional university knowledge is captured by the number of citations in firms' patents to universities located outside the region where the firm is established.

[2] Two, three or even five-year lags between dependent and independent variables are usually taken into account in the patent literature, but in this case the specification of lag structures should not be an important concern because the explanatory variables are supposed to be stable over the years.

Independent variables:

- Absorptive capacity (AC). The empirical literature on absorptive capacity has to a large extent limited itself to the amount of R&D expenditures or presence of an R&D unit as a measure of absorptive capacity both at firm and at regional level. Other popular indicators of absorptive capacity include human resources, and networks. In this paper we use R&D efforts as a viable proxy of absorptive capacity (firms' R&D as percentage of GDP -gross domestic product-). The original paper by Cohen and Levinthal (1990) used firm-based R&D data as proxies for absorptive capacity in the empirical section of their paper. Subsequent extensive evidence has used firm R&D to analyse the firms' capability to access knowledge from external sources (e.g. seminal papers such as Kim, 1997, and Kodama, 1995, stressed the crucial role of a firm's internal R&D in determining its ability for the acquisition and assimilation of external knowledge).
- Presence in the region of university technological opportunities (U). We capture the capacity of universities to produce quality patents in each region the regional 'Higher Education R&D' expenditure as percentage of regional GDP. This is a resource variable to proxy for the strength of the university system to produce outputs. We expect that greater effort in university R&D should lead to more university outputs that could increase the opportunities for firms to acquire and exploit this knowledge.
- To control for the regional specialization (S) we calculate a similar measure to the revealed technological advantage index (Soete and Wyatt, 1983): TAI= $\frac{P_{ij} / \sum_{s=1}^{S} P_{is}}{\sum_{i=1}^{N} P_{is} / \sum_{i=1}^{N} \sum_{s=1}^{S} P_{is}}$, where $P_{is} / \sum_{s=1}^{S} P_{is}$ is the number of patents of region i in sector j over the number of patents of region i in all sectors; $\sum_{s=1}^{N} P_{is} / \sum_{i=1}^{N} \sum_{s=1}^{S} P_{is}$ is the number of patents of all regions in sector s over the total number of patents. To construct the index we use eight sections of the International Patent Classification (IPC) (see the bottom of Table 2).
- To control for the size of the region (Z) we include the number of firms' patents in each region. This variable prevent from obtaining spurious relationships (as regions with more patents are expected to have more citations).

For estimating the models, we apply a conditional fixed and random effects negative binomial estimator in which we assume that units (regions) are positively correlated within clusters (countries). Then, the econometric estimations are in the framework of the cluster count data models. The decision to use a two-level hierarchical analysis (regions clusters in countries) has two main objectives: (a) to evaluate the unobserved heterogeneity—along with the fixed effects—of the regional acquisition of knowledge; the inclusion of random effects in the model considers that there is natural heterogeneity across regions of the same country;

(b) to correctly estimate the confidence intervals, taking into account the intra regional correlation of regions in of the same country. Failures to take into account the clustering of data result in serious biases (see, for example, Moulton, 1990; Antweiler, 2001; Wooldridge, 2003, 2006).
To summarize, the empirical base models are as follows.

- A negative binomial model with a hierarchical data structure (regions grouped into countries) for analyzing the acquisition of inward regional knowledge.
- A negative binomial model with a hierarchical data structure (regions grouped into countries) for the acquisition of outward regional knowledge.

The previous paragraphs describe the base specifications. However, taking into account the structure of our sample, the nature of the data, and other considerations such as the number of zeros in the sample, we have considered additional models:

- A negative binomial model and a zero inflated negative binomial model with a pooled data structure and clustered robust standard errors (the clusters are countries) for the acquisition of inward regional knowledge (Table 4)
- A negative binomial model and a zero inflated negative binomial model with a pooled data structure and clustered robust standard errors (the clusters are countries) for the acquisition of outward regional knowledge (Table 4)

Data

The data collection process was designed by the Institute for Prospective Technological Studies (IPTS) in 2009. An international consortium of researchers from the University of Newcastle, Incentim and the Centre for Science and Technology Studies (CWTS) were responsible for implementing the data collection. Figure 1 may help visualising data construction. The EPO Worldwide Patent Statistical Database (PATSTAT) database was used to compile a dataset of 228.594 direct EPO patents applied for in the period 1997-2007. The team then identified 10,307 patents with university references, i.e. citations to patents applied for by universities or to WoS scientific articles, signed by authors with a single university affiliation. Actually, this single-university affiliation criterion is the main limitation of the database, due to resource constraints, and implies that both the number of patents with references and the share of papers within university references are underestimated.

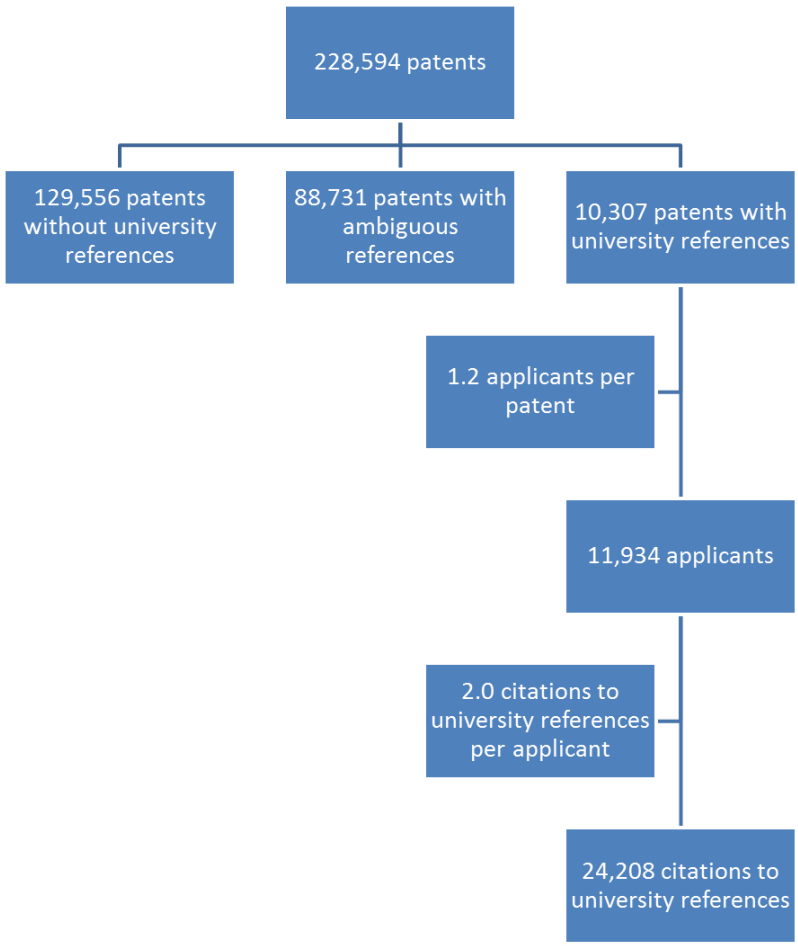


Figure 1. University references in direct EPO patents, 1997-2007

Each patent had an average of 1.2 applicants, resulting in a total of around 12,000 applicants.; and each applicant cited an average of 2 university references, so the starting number of citation to university references was slightly over 24,000. In order to match the NUTs II region of the citing applicant and the cited university, we excluded citations by non-EU27 applicants and a few EU27 applicants without regional information (Figure 2). In order to test our hypotheses, we excluded applicants other than firms, resulting into a total of some 13,000 citations. For these, we could check whether there was a match between applicant region and region of a citation from a university: 2 percent produced a positive match.

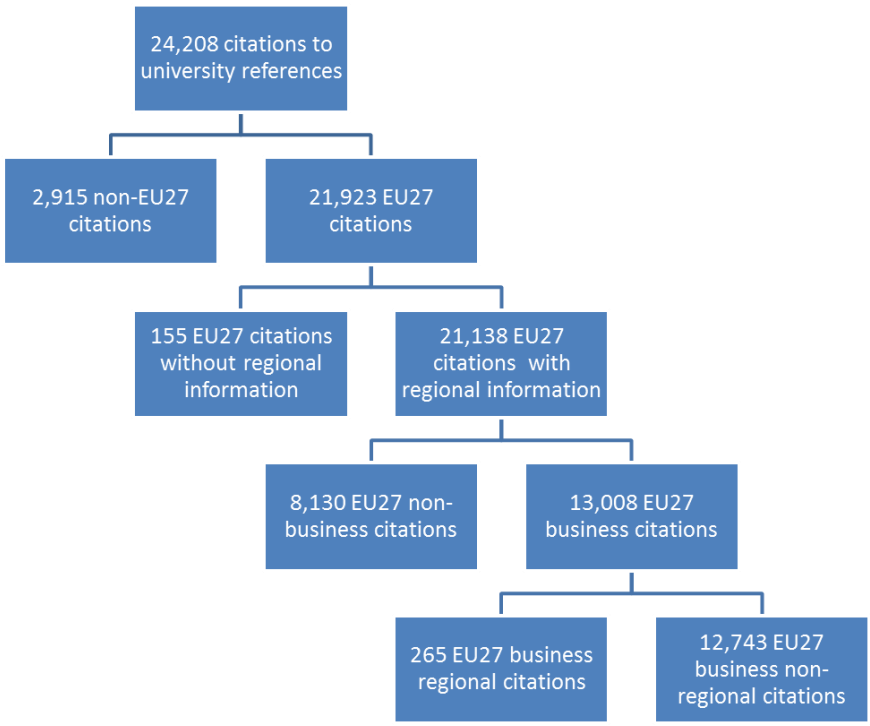


Figure 2. Citations to university references in direct EPO patents, 1997-2007

We aggregated patent and citation counts per region and year to produce a panel that was linkable to Eurostat regional R&D statistics. This results in a sample of 2,365 observations (Figure 3); however, there are 1,181 observations in which there is not any patent belonging to firms. The consequence is that we finally count on fewer observations. The estimated models in the next Section include firm and university R&D intensity as explanatory variables. As there are many missing data for these variables at regional level, this results in a new reduction in the number of observation to 503 for 22 countries in the UE27 from 1997 to 2007. The number of patents drops to around 4,000 and that of citations to universities to around 6,000, of which a 2 percent are still regional citations.

We mentioned in section 3 that the nature of the data suggests the specification of grouped and pooled models. Tables 1 and 2 show the descriptive statistics for each type of model. Note that the use of the fixed effects estimator requires that countries with only one observation is omitted; that's why there is a different number of observations depending on the type of model (Figure 3).
The two dependent variables show a remarkable different behaviour. In the case in which we have 464 observations, the acquisition of university knowledge from the region (inward) by firms takes into account 388 observations with zero citations, and 76 observations with one or more citations (Table 1). In models

with 499 regions, the outward acquisition of knowledge by firms has only five observations with zero citations and 494 with one or more citations (Table 2).

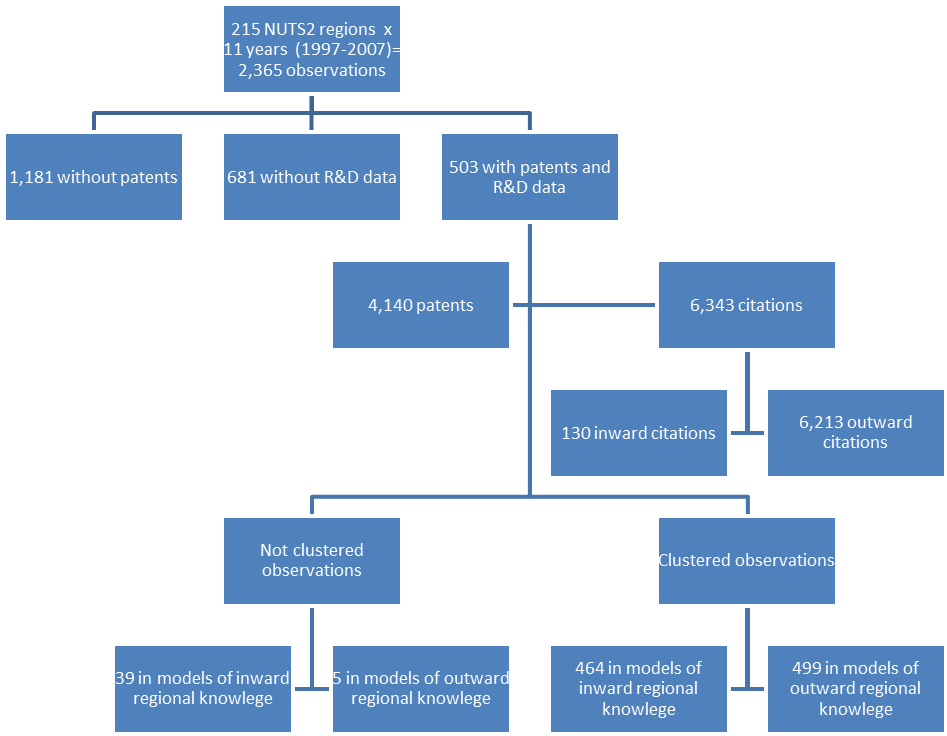


Figure 3. The panel

Table 1

Descriptive Statistics
464 observations

	Mean	*Std. Dev.*	*Min*	*Max*
Acq. Inward reg. know	0.280	0.763	0	6
A=Firms' R&D/GDP	1.135	0.890	0.04	6.83
U=Universities' R&D/GDP	0.395	0.205	0.01	1.30
Numberpatents	8.933	17.515	1	151
speA (1)	0.931	0.690	0	3.83
speB	0.684	0.960	0	7.42
speC	0.693	0.595	0	2.17
speD	0.313	1.504	0	22.19
speE	0.294	1.320	0	17.20
speF	0.505	1.211	0	8.57
speG	0.598	0.618	0	3.94
speH	0.447	0.738	0	5.15

Table 2

	Descriptive Statistics *499 observations*			
	Mean	*Std. Dev.*	*Min*	*Max*
Acq. Outward reg. know	12.790	26.366	0	243
A=Firms' R&D/GDP	1.136	0.902	0.04	6.83
U=Universities' R&D/GDP	0.398	0.225	0	1.32
Numberpatents	8.531	16.988	1	151
speA (1)	0.917	0.698	0	3.83
speB	0.698	1.002	0	7.42
speC	0.693	0.597	0	2.17
speD	0.291	1.452	0	22.19
speE	0.308	1.385	0	17.20
speF	0.513	1.231	0	8.57
speG	0.581	0.610	0	3.94
speH	0.444	0.733	0	5.15

Figure 4 shows that the number of citations has remained quite stable through time. It has oscillated around almost a horizontal line in the case of both inward and outward citations during the period of observation. Actually, the share of regional over total citations has also moved around the average of 2 percent without clear upward or downward patterns.

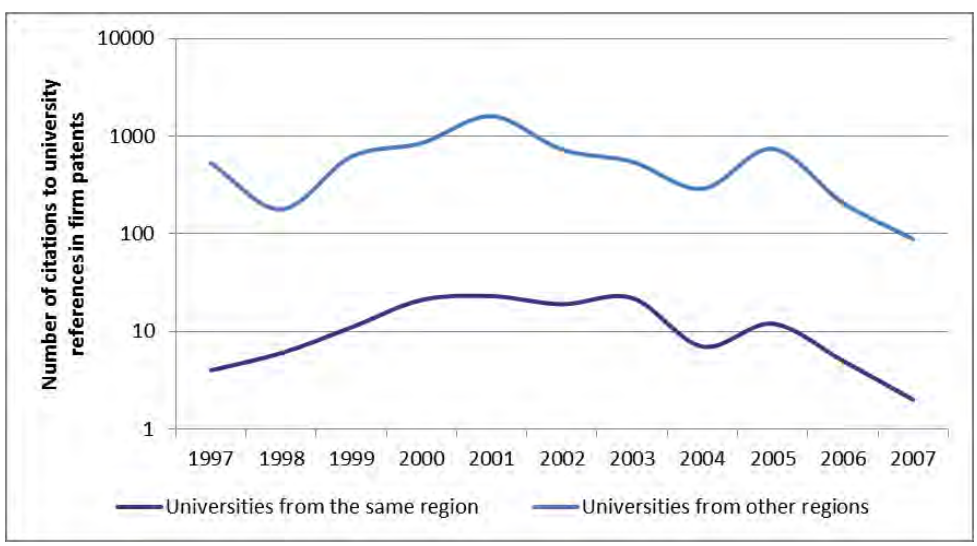


Figure 4. Stability on the evolution of firm citations to university references

On the contrary, Figure 5 illustrates that cross-sectional variation is apparently more important. If we compare the top ten regions in number of inward versus outward citations (upper and lower parts of the figure, respectively), only three appear in both rankings: Île de France, London and Berlin. The rest are different, suggesting that the processes of university knowledge acquisition depend on

varied factors according to the inward or outward nature of the flow. It is also an empirical validation of the interest of the topic, raised in the introduction.

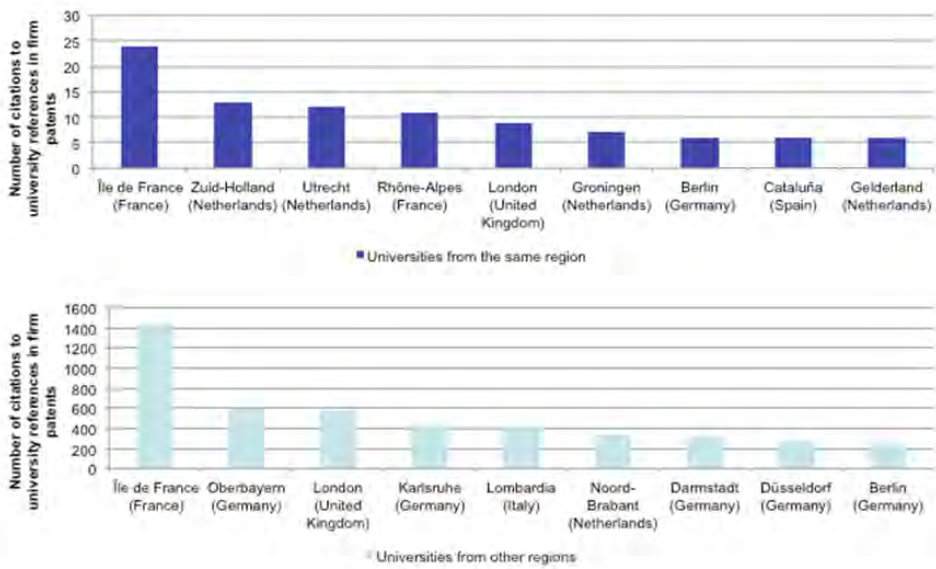


Figure 5. Cross-regional variation in firm citations to university references: top regions in number of citations

Econometric results

Baseline results

This section presents the results for both analysis (inward and outward acquisition of knowledge) and taking into account two different structures of the data (hierarchical and pooled):

Firstly, Table 3, Columns 1-2 and 4-5, show the estimated models for the acquisition of inward and outward knowledge following the hierarchical data structure (applying a fixed and random effects estimator for each one). In order to compare the results of different estimators, we have used the same number of observations (464 for the inward knowledge acquisition and 499 for the outward).

Secondly, Table 3, Columns 3 and 6, show the pooled models for the same number of observations. Given the nature of the dependent variable, we provide the ZINB estimation when the dependent variable is the acquisition of inward knowledge (which has many zeros), and a NB when the dependent variable is the acquisition of outward knowledge (these are the preferred models according to the Vuong statistic).

Results about variables affecting inward university knowledge are taken from Column 3 because likelihood ratio test suggest models with pooled data (Column 3) are preferred to models with hierarchical structure (Columns 1-2). Column 3

shows that the absorptive capacity of firms in the region does not play any role in determining the use of scientific and technological university knowledge generated in the same region of the firm's location. There is no evidence in favour of Hypothesis 1.

Columns 4-5 show that the firms' absorptive capacity of the region determines the use of outward university knowledge (grouped data preferred to pooled data according to LR test). That is, regions with greater effort in private R&D have a greater absorption of scientific and technological university knowledge from outside the region (from other countries or other regions in the same country). Hence, Hypothesis 2 is confirmed.

Concerning the influence of the university capacity of the region to produce spillovers, Column 3 shows that the use of scientific and technological university knowledge by firms from the same region is positively related to the university capacity of the region. This means that the greater the R&D effort in the universities of the region, the larger the use of scientific and technological knowledge from the own regional universities, i.e. the evidence supports Hypothesis 3.

Columns 4-5 give us the opportunity to contrast the effect of university capacity of the region on the acquisition of outward university knowledge. University capacity of the region is negatively related with de acquisition of university knowledge from outside the region by private firms, and consequently there is evidence in favor of Hypothesis 4.

Table 3

Dependent Variable: UKA (University knowledge Acquisition)												
	I. Acquisition of inward regional knowledge						II. Acquisition of outward regional knowledge					
	Negative binomial models for grouped data				ZINB model for pooled data		Negative binomial models for grouped data				NB model for pooled data	
	1 FE		2 RE		3 Robust Std Err Adjusted (country)		4 FE		5 RE		6 Robust Std Err Adjusted (country)	
Cons	-18.715		-21.740		-16.595	**	-1.156	**	-1.216	**	-0.523	**
A=Firms' R&D/GDP	-0.347	*	-0.340	*	-0.291		0.078	**	0.088	**	0.049	
U=Universities' R&D/GDP	2.460	**	2.265	**	2.137	**	-0.330	**	-0.258	*	0.138	
Numbpatents	0.017	**	0.018	**	0.016	**	0.022	**	0.021	**	0.040	**
speA (1)	0.742	**	0.866	**	1.595	**	0.459	**	0.474	**	0.484	**
speB	0.290		0.292		0.282	**	0.161	**	0.163	**	0.131	**
speC	1.255	**	1.190	**	-0.042		0.872	**	0.874	**	0.888	**
speD	-0.042		-0.044		0.190		0.014		0.017		0.041	**
speE	0.142		0.147		-0.072		0.021		0.023		0.019	
speF	0.267		0.195		0.265		0.080	**	0.079	**	0.089	*

speG	0.433		0.363		0.315		0.506	**	0.524	**	0.527	**
speH	0.578	**	0.503	**	-0.011		0.311	**	0.312	**	0.283	**
Ln_r			3.122						2.464			
Ln_s			2.160						3.306			
Inflation model (logit)												
Cons					1.583							
speA (1)					1.134							
speB					-0.270							
speC					-2.849	**						
speD					0.289							
speE					-0.703							
speF					0.295							
speG					0.515	*						
speH					-1.657							
Number of obs	464		464		464		499		499		499	
Number of groups	9		9		9		18		18		18	
Wald chi2	115.20	**	122.66	**			2746.73	**	2823.93	**		
Loglikelihood	-201.35		-230.51		-220.41		-1334.04		-1417.03		-1314.75	
LR Test Panel vs Pooled			1.63						57.44	**		

Notes:
(1) IPC Sections to construct the specialization indexes (spe): A Human Necessities; B Performing Operations; Transporting; C Chemistry; Metallurgy; D Textiles; Paper; E Fixed Constructions; F — Mechanical Engineering; Lighting; Heating; Weapons; Blasting: G Physics; H Electricity.
- **, * denote that the coefficients are statistically different from zero at the 5% and 10% levels, respectively.
- All models include year dummies for 1997 to 2007.
- VIF suggests no signs of multicollinearity.
- Likelihood ratio test favours Poisson against NB in Models 3 and 6
- Vuong statistics favours ZINB against NB in Model 3 and NB against ZINB in Model 6.

Robustness check

The fixed effects panel models shown so far are computable only for the 464 and 499 observations used in the previous section. However, in the rest of the models, using the same number of observations is an imposition to facilitate comparison. As robustness check, we have estimated the same specifications as in previous section but without restrictions in the number of observations for each model. The advantage of not imposing any restriction is that we can count on more data for the estimations; however, the comparisons for selecting models are now more difficult. The number of observations increases to 503 in the random effects models, ZINB and NB. Table 4 provides the descriptive statistics.

Table 4

Descriptive Statistics
503 observations

	Mean	*Std. Dev.*	*Min*	*Max*
Acq. Inward reg. know	0.258	0.737	0	6
Acq. Outward reg. know	12.704	26.278	0	243
A=Firms' R&D/GDP	1.128	0.903	0.02	6.83
U=Universities' R&D/GDP	0.396	0.225	0	1.32
Numberpatents	8.473	16.933	1	151
speA (1)	0.917	0.711	0	3.83
speB	0.698	1.003	0	7.42
speC	0.693	0.600	0	2.17
speD	0.412	3.113	0	62.12
speE	0.305	1.379	0	17.20
speF	0.509	1.227	0	8.57
speG	0.577	0.610	0	3.94
speH	0.442	0.732	0	5.15

For these 503 observations the preferred model for inward UKA is ZINB with pooled data structure (presented in Table 5, Column 3). The preferred model for outward UKA is NB with hierarchical structure (presented in Table 5, Column 6). These new estimations, which have not been forced to use the same number of observation, confirm the previous results; the hypotheses rejected and non-rejected are just the same as in Section 5.1.

Table 5

Dependent Variable: UKA (University knowledge Acquisition)												
	I. Acquisition of inward regional knowledge						II. Acquisition of outward regional knowledge					
	Negative binomial models for grouped data				ZINB model for pooled data		Negative binomial models for grouped data				NB model for pooled data	
	1 FE		2 RE		3 Robust Std Err Adjusted (country)		4 FE		5 RE		6 Robust Std Err Adjusted (country)	
cons	-18.715		-21.893		-16.987	**	-1.156	**	-1.217	**	-0.527	**
A=Firms' R&D/GDP	-0.347	*	-0.421	**	-0.311		0.078	**	0.091	**	0.057	
U=Universities' R&D/GDP	2.460	**	1.973	**	1.943	**	-0.330	**	-0.259	*	0.132	
Numbpatents	0.017	**	0.018	**	0.015	**	0.022	**	0.021	**	0.039	**
speA (1)	0.742	**	0.850	**	1.774	**	0.459	**	0.469	**	0.478	**
speB	0.290		0.304	*	0.334	**	0.161	**	0.163	**	0.128	**
speC	1.255	**	1.195	**	0.204		0.872	**	0.873	**	0.885	**
speD	-0.042		-0.031		0.188		0.014		0.005		0.007	
speE	0.142		0.132		-0.089		0.021		0.022		0.018	
speF	0.267		0.170		0.331		0.080	**	0.083	**	0.095	**
speG	0.433		0.425	*	0.428		0.506	**	0.522	**	0.522	**
speH	0.578	**	0.545	**	0.052		0.311	**	0.314	**	0.285	**

Ln_r			2.556						2.411			
Ln_s			1.488						3.210			
Inflation model (logit)												
cons					0.964							
speA (1)					1.254							
speB					-0.160							
speC					-2.249	**						
speD					0.198							
speE					-0.545							
speF					0.462							
speG					0.451							
speH					-1.472							
Number of obs	464		503		503		499		503		503	
Number of groups	9		22		22		18		22		22	
Wald chi2	115.20	**	122.40	**			2746.73	**	2832.37	**		
Loglikelihood	-201.35		-237.10		-227.67		-1334.04		-1425.57		-1323.28	
LR Test Panel vs Pooled			3.28	**					58.84	**		

Notes:
(1) IPC Sections to construct the specialization indexes (spe): A Human Necessities; B Performing Operations; Transporting; C Chemistry; Metallurgy; D Textiles; Paper; E Fixed Constructions; F — Mechanical Engineering; Lighting; Heating; Weapons; Blasting: G Physics; H Electricity.
- **, * denote that the coefficients are statistically different from zero at the 5% and 10% levels, respectively.
- All models include year dummies for 1997 to 2007.
- VIF suggests no signs of multicollinearity.
- Likelihood ratio test favours Poisson against NB in Models 3 and 6.
- Vuong statistics favours ZINB against NB in Model 3 and NB against ZINB in Model 6.

Conclusions

In this paper we argue that the knowledge that firms in a region can acquire from university spillovers is a function of both the absorptive capacity of the firms developed by investing in knowledge, and the opportunities for university spillover. To test our hypotheses we put forward an external knowledge acquisition function which explains the factors affecting the regional inward and outward acquisition of university knowledge by firms.

Our models yield to reject hypothesis 1. Hypotheses 2, 3 and 4 are not rejected. According to these findings, absorptive capacity is not relevant in explaining the acquisition of inward scientific and technological university knowledge; however, regional absorptive capacity plays a relevant positive effect in the acquisition of outward university knowledge. Regarding the other relevant variable in the models, university opportunities for spillovers in the region have a positive effect on the acquisition of local knowledge by firms from the same region, and a negative influence in the acquisition of outward university regional knowledge.

These findings have some relevant policy implications. Considering the objective of policy makers, we can divide implications into two types:

- If the objective of regional government is encouraging the use of university knowledge produced in the region (by firms established in the region), our results suggest that the only way is the stimulation of the supply side, that is the investment in university scientific and technological knowledge to produce regional opportunities. However, this has a trade-off: it also decreases the acquisition by firms of university knowledge produced outside the region. Hence, it opens the risk of lock-in effects by closing regions to external knowledge.
- If the objective is improving the competitiveness of local firms (in the sense that they could understand and incorporate university knowledge from elsewhere), our results suggest that absorptive capacity is the variable to spur. In addition, it has a dual role, since it compensates the negative effect of high university R&D capacity on outward knowledge acquisition.

Future research would include increasing the number of cited university references in order to break down the data by type of cited literature (patent or non-patent literature) or origin of the citation (examiner or applicant). For the time being, the number of regional citations is too scarce to produce meaningful results. Another line would be to face the traditional geographic approach to patent citations –the role of distance– versus this paper's approach –the role of regional borders– and ask which one matters more: distance or borders. Adding more measures of firms absorptive capacity and university supply capacity would be enriching, but would require previous research about how they can be defined at regional level that is outside the scope of this paper. It would be also worth investigating whether having engaged into actual cooperation with universities shapes citation patterns. Replicating the analysis at NUTs III level would be potentially interesting, but regions at that level have less margin for implementing their own policies, the number of regional citations would be lower and R&D statistics less available. Finally, a complementary approach should retrieve information from full-text rather than front-page citations, but this would require much manual work and be enormously costly for such a large sample.

Acknowledgements

This research was initiated with the framework of ERAWATCH, a joint initiative of the European Commission's Directorate General for Research and the Joint Research Centre-Institute for Prospective Technological Studies (IPTS). The views expressed in this article are those of the author and do not necessarily reflect those of the European Commission (EC). Neither the EC nor anyone acting on behalf of the EC is responsible for the use that might be made of the information. Joaquín M. Azagra-Caro is grateful to René van Bavel and Xabier Goenaga for their support, and to the international consortium that produced the database, including Henry Etzkowitz, Marina Ranga and members of Incentim and CWTS, led, respectively, by Bart Van Looy and Robert J.W. Tijssen. The research has continued with funding from project 2091 of the Polytechnic University of Valencia and project 201010I004 of the CSIC. A previous version of the paper was presented at the 2013 Technology Transfer Society Annual

Conference, 2012, and the authors acknowledge helpful comments from the audience.

References

Abramo, G., D'Angelo, C.A., & Solazzi, M. (2010). Assessing public–private research collaboration: is it possible to compare university performance? Scientometrics, 84, 173–197.

Abramovsky, L., Harrison, R. Simpson, H., 2007. University Research and the Location of Business R&D. Economic Journal 117, C114-41

Acosta, M., Coronado, D., 2003. Science-technology flows in Spanish regions: An analysis of scientific citations in patents. Research Policy 32, 1783-1803.

Acosta, M., Coronado, D., Ferrándiz, E., León, M., 2011b. Factors affecting inter-regional academic scientific collaboration within Europe: the role of economic distance. Scientometrics 87, 63-74.

Acosta, M., Coronado, D., Flores, E., 2011a. University spillovers and new business location in high-technology sectors: Spanish evidence. Small Business Economics 36, 365-376.

Anselin, L., Varga, A., Acs, Z. J., 2000. Geographic and sectoral characteristics of academic knowledge externalities. Papers in Regional Science 794, 435-443.

Anselin, L., Varga, A., Acs, Z., 1997. Local geographic spillovers between university research and high technology innovations. Journal of Urban Economics, 423, pp. 422-448.

Antweiler, W., 2001. Nested random effects estimation in unbalanced panel data. Journal of Econometrics 101, 295–313.

Arndt, O.; Sternberg, R., 2000. Do manufacturing firms profit from intraregional innovation linkages? An empirical answer, European Planning Studies 84, 465–485.

Audretsch, D. B., Lehmann, E. E., 2005. Does the Knowledge Spillover Theory of Entrepreneurship Hold for Regions? Research Policy 34, 1191–1202.

Audretsch, D., Feldman, M. P., 1996. R&D spillovers and the geography of innovation and production. American Economic Review 86, 630-640.

Audretsch, D., Lehmann, E., Warning, S., 2004. University Spillovers: Does the Kind of Science Matter? Industry and Innovation 11, 193–206.

Azagra, J. M., 2012. Access to universities' public knowledge: who's more nationalist? Scientometrics, DOI 10.1007/s11192-012-0629-5

Bathelt, H., Malmberg, A., Maskell, P., 2004. Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography 28, 31–56

Boschma, R.A., 2005. Proximity and Innovation: A Critical Assessment. Regional Studies 39, 61–74

Branstetter, L. G., 2001. Are knowledge spillovers international or intranational in scope? Microeconometric evidence from the U.S. and Japan. Journal of International Economomics 531 53–79.

Branstetter, L. G.; Yoshiaki, O., 2005. Is Academic Science Driving a Surge in Industrial Innovation? Evidence from Patent Citations. Department of Social and Decision Sciences. Paper 48. http://repository.cmu.edu/sds/48

Caragliua, A.; Nijkamp, P., 2012. The impact of regional absorptive capacity on spatial knowledge spillovers: the Cohen and Levinthal model revisited. Applied Economics 44, 1363-1374.

Charles, D.; Damianova, Z.; Maroulis, N., 2009. Contribution of policies at the regional level to the realisation of the European Research Area. Prepared by ERAWATCH NETWORK ASBL.

Chesbrough, H., 2003. Open innovation: the new imperative for creation and profiting from technology. Boston, MA: Harvard Business School Press.

Chesbrough, H., Vanhaverbeke,W., andWest, J., (eds.), 2006. Open innovation: researching a new paradigm. Oxford: Oxford University Press.

Cohen, W., Levinthal, D., 1990. Absorptive Capacity: A New Perspective on Learning and Innovation. Administrative Science Quarterly 35, 128–152.

Cooke, P., 2005 Regionally asymmetric knowledge capabilities and open innovation. Exploring 'Globalisation 2'—A new model of industry organisation. Research Policy 34, 1128–1149.

Cooke, P., Boekholt, P., T¨odtling, F., 2000. The Governance of Innovation in Europe. Pinter, London.

Dahlander, L. Gann, D.M., 2010. How open is innovation? Research policy 39, 699-709.

Davenport, S., 2005. Exploring the role of proximity in SME knowledge-acquisition. Research policy 34 , 683-701.

de Jong J.P.J., Freel M., 2010. Absorptive capacity and the reach of collaboration in high technology small firms. Research Policy 39, 47–54.

Drejer, I., Vinding, L. A., 2007. Searching near and far: Determinants of innovative firms' propensity to collaborate across geographical distance. Industry and Innovation 143, 259–275.

Feldman, M., Florida, R., 1994. The Geographic Sources of Innovation: Technological Infrastructure and Product Innovation in the United States. Annals of the Association of American Geographers 84, 210–229.

Fischer, M., Varga, A., 2003. Spatial knowledge spillovers and university research: Evidence from Austria. Annals of Regional Science 372, 303-322.

Gassmann, O. Enkel, E., 2004 Towards a theory of open innovation: three core process archetypes. R&D Management Conference, July 6–9, Lisbon, Portugal.

Gassmann, O., Enkel, E. Chesbrough, H., 2010. The future of open innovation.R&D Management 40, 213-221.

Harhoff, D., 1999. Firm Formation and Regional Spillovers—Evidence from Germany. Economics of Innovation & New Technology 8, 27–55.

Hewitt-Dundas, N., 2011. The role of proximity in university-business cooperation for innovationJournal of Technology Transfer DOI 10.1007/s10961-011-9229-4.

Kaufmann, A., Todtling, F., 2001. Science–industry interaction in the process of innovation: The importance of boundary-crossing systems. Research Policy 30, 791–804.
Kim, L., 1997. Imitation to innovation: the dynamics of Korea's technological learning. Boston, MA: Harvard Business School Press.
Klevorick, A. K., Levin, R., Nelson, R., Winter, S., 1995. On the sources and significance of inter-industry differences in technological opportunities. Research Policy 24, 342–349.
Kodama, F., 1995. Emerging patterns of innovation: sources of Japan's technological edge. Boston, MA: Harvard Business School Press.
Kratke, S. , 2010. Regional knowledge networks. A network analysis approach to the interlinking of knowledge resources. European Urban and Regional Studies 171, 83–97.
Laursen, K., Salter, A., 2004. Searching high and low: what types of firms use universities as a source of innovation? Research Policy 33, 1201-1215.
Lichtenthaler, U., 2011. Open Innovation: Past Research, Current Debates, and Future Directions. Academy of Management Perspectives, February, 75-93.
Moulton, B. R., 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. Review of Economics and Statistics 72, 334–338.
Soete, L.G., Wyatt, S.M.E., 1983. The Use of Foreign Patenting as an Internationally Comparable Science and Technology Output Indicator. Scientometrics 5, 31–54.
Spithoven A., Clarysse B., Knockaert, M. 2011. Building absorptive capacity to organise inbound open innovation in traditional industries. Technovation 31, 10–21.
van de Vrande, V., deJong, J. P. J., Vanhaverbekec, W. de Rochemontd, M., 2009. Open innovation in SMEs:Trends, motives and management challenges. Technovation 29, 423–437.
Woodward, D., Figueiredo, O., Guimaraes, P., 2006. Beyond the Silicon Valley: University R&D and High-Technology Location. Journal of Urban Economics 60, 15–32.
Wooldridge, J. M., 2003. Cluster-sample methods in applied econometrics. American Economic Review 93, 133–138.
Wooldridge, J.M., 2006. Cluster-sample Methods in Applied Econometrics: An Extended Analysis. Department of Economics, Michigan State University Mimeo: Michigan.
Zucker, L., Darby, M., Armstrong, J., 2002. Commercializing Knowledge: University Science, Knowledge Capture, and Firm Performance in Biotechnology. Management Science 48, 138-153.

ADVANTAGES OF EVALUATING JOURNALS THROUGH ACCA - LIS JOURNALS (RIP)

D. Gnana Bharathi

dgbharathi@yahoo.com
Central Leather Research Institute, Chennai – 600020 (India)

Abstract
Aggregated Citations of Cited Articles (ACCA) is an empirical method for evaluating research journals. The method is bases on four important characteristics, viz. determination of a denominator based on articles identified by at least one citation, weightage to the volume of publication, five years of citations - including publication year and the ratio of citations to the individual year of publication. LIS journals are characterized by receiving a peak level of citations beyond the Impact Factor range, fluctuation in the number of articles published every year, the presence of uncited articles, etc. In the present article ACCA applies to evaluation of LIS journals and compared with the IF, 5 year IF, SNIP, SJR and Eigen Factor values.

Conference Topic
Scientometrics Indicator (Topic 1) and Science Policy and Research evaluation (topic 3)

Introduction

Number of articles published by the journals, per annum, varies widely with some journals publishing tens of articles whereas other journals publishing thousands of articles. Citations of the scientific journals differ from one journal to other with peak levels of citations, which may come in first, second, third, fourth, fifth or years after a decade (Egghe & Rousseau, 1990; Klavans & Boyack, 2007; Vieira & Gomes, 2009). Averaging of citations per article without giving due to consideration for the volume of publication result in unfair comparison. The efforts put forward in collecting, organizing, evaluating, compiling and publishing large number of articles keeping in interest of the readers without compromising the quality has to be rewarded. Publication of the number of articles by the given journals do not remain as fixed and may considerably vary on year to year basis. Bulk averaging of all the citations to all the articles or citable articles published in different years may either increase or decrease the ranking of the journals. Selection of what can be considered as a citable article is controversial as evaluation of whether an article qualifies as citable or not depends on person to person.

Library and Information Science journals (LIS journals) involve in research and development in library science as well as scientometric analyses of publications in various subjects by the journals and researchers. The peak citations of LIS

journals come in third, fourth, fifth or even later years of publication. The traditional evaluation methods such as IF covers citations only of second and third years of publication. Articles which may not be considered as a qualitative original research output but as informative one such as editorial, news, etc. are sometimes included as citable articles. This occurs to journals of almost all the subject categories, including LIS journals. Therefore, including these articles as citable articles will affect the ranking of the journals by increasing the denominator in the IF calculation.

Aggregated Citations of Cited Articles (ACCA) is based on empirical analysis of the number of journals representing various characteristics (Bharathi, 2011). ACCA considers all the aforementioned parameters into consideration and gives an unbiased evaluation method based on most of the characteristics of the journals in publications and citations. The article discusses the ranking of LIS journals through ACCA, and the ranking values are compared with the Impact Factor (IF) and 5 years IF (Huh, 2011; Jacso, 2001), SNIP (Moed, 2010), SJR (González-Pereira, Guerrero-Bote, & Moya-Anegón, 2010) and Eigen Factor (Factor, 2011).

Selection of journals

Fifteen LIS journals are selected for the study. These are the journals that are ranked by IF ranging from 0.25 to 4.45. The subject category as "Library Information Science journals" in WoS is considered as LIS journals. These journals are characterized by publication of the fluctuating number of articles per annum, number of citations, peak year of citations, etc. The journals are: *Mis Quart, J Am Med Inform Assn, J InfTechnol, J Strategic InfSyst, J Manage Inform Syst, Inform Manage-Amster, Online Inform Rev, J Am SocInfSci Tec, Annu Rev Inform Sci, Scientometrics, Eur J Inform Syst, Inform Process Manag, Int J GeogrInfSci, J InfSci and Inform TechnolLibr.*

As a reference six journals of two categories are selected. Journals *Nature, Science* and *Plos One*are are selected as multidisciplinary journals. Similarly, *CA A Cancer Journal for Clinicians (CA Cancer), New England Journal of Medicine (New Engl Med)* and *Lancet* are selected as medical journals.

Aggregated Citations of Cited Articles

ACCA is calculated on citations in the fifth year for five preceding years of publication by the journal. The method uses citations received by the journal and cited articles of the journal as the parameters. Number of citations in the evaluation years, from five years of publication are used as a numerator. Five years of cited articles, those articles that received at least one citation, are used as a denominator. The citations per cited article is multiplied by log square value of cited articles, which divided by 10, to give weightage to the size volume of the valuable articles. The value is further divided by five in order to make the aggregation of the citations per article on the yearly basis. *N* is to control a

fluctuation in those journals with the absence of citations during the years under consideration. The formula for ACCA is as follows:

$$ACCA = \frac{1}{(5+N)} \times \sum_{y=1}^{5} \left\lfloor \left(\frac{C_y}{A_y} \times \left\{ 1 + \frac{\log(A_y)^2}{10} \right\} \right) \right\rfloor$$

C_y is the number of citations in ACCA year from year, y
A_y is the number of cited articles from the publication year, y
N *is* the number of publication years that failed to get any citations in the ACCA year
Y is the year.

Articles published by the of LIS Journals

Number of articles published by LIS greatly differs from one journal to other. Of the LIS journals studied number of articles published by them has increased for almost all the journals published from 2001 to 2006, excepting the *J Am Med Inform Assn,* which decreased the number of articles from 506 in 2001 to 91 in 2006. This growth for a number of articles shows confidence in LIS publishers' confidence in publishing more articles. This also corresponds to increase in citations to these journals. The number of articles published by the LIS journals is given to the legends of the Fig 1 & 2 within the brackets.

ACCA gives weightage to the volume of publication by the journals. ACCA considers only the cited articles of a journal as active articles in determining the volume of publication. This is in contrast with IF where the denominator is citable articles which is determined by the specialists of the WoS database systems. There are various claims that the process of identifying what is citable is biased. Cited articles, on the other hand, is determined by receiving at least one citation by the article, irrespective of merits of the article. This overcomes the bias in selecting the active article. This also ensures the articles that contribute to the citations accountable, by including in the denominator of the ACCA calculation.

For calculation of ACCA for the n^{th} year, citations in the n^{th} year is calculated by citations from 1,2,3,4 and 5 previous years of publication, which is divided respectively, by the number of cited articles in the corresponding year 1, 2, 3, 4 and 5. This contrasts with the IF where all citations in the n^{th} years are divided by the all the citable articles. As the LIS journals do not publish an fixed number of articles every year, the cumulative ratio of citations per articles may increase or decrease the ranking of the journals depending on the number of citable articles. ACCA overcomes this fluctuation in a number of articles published or cited articles in its case, by the opting individual ratio of citations per cited article of that year.

Citation Characteristics of LIS Journals

The peak level or modal value of citations for LIS journals comes few years after the publication (Fig 1). Citations received in first, second and third years of publication are superseded by citations received in later years. Only two journals received peak citations in the third year of publication, one in 2001 (*J InfSci*) and other in 2006 (*Inform Process Manag*). As IF calculation takes into consideration of just second and third years of citations, most of the citations, including the modal citations, received by LIS journals do not fall into IF limitations.

As far the 2001 publication of the LIS journals most of the modal citations come in 7th and 8th year of publication. Only one journal, *MIS Quarterly*, which is characterized by the growth in citations year after the years, received peak citations in 11th year of publication. All other journals show decreasing in citations by the year 2011. For the LIS journals published in the year 2006, modal citations come in 4th, 5th and 6th year of publication. There are possibilities that modal citations may change in later years as only six years are available for the citations of 2006 publications. However, one cannot wait indefinitely to find the modal citations for every journal. A period up to 5 or 6 years can give LIS journal sufficient time to express its citations behaviors.

Discussion

Citation characteristics of LIS journals are different from the perspective of highly cited journals, which receive early citations (Fig 1 & 2). The consistency of citations by the LIS journals is not taken into consideration of the IF or other ranking systems. Even well cited journals do not get their due, if the peak citations come in later years.

Fig 1 clearly shows that LIS journals receive sustainable citations as per the citations of 2001 publications. The citations of LIS journals continue to grow over the year and stabilize after fifth or sixth year of publication. On the other hand, the reference journals, excepting CA Cancer, receive modal citations in third or fourth year. CA Cancer receive modal citations in the second year, and its third year of citations is just second to the modal value. Therefore, CA Cancer scores high IF.

For 2006, publications, LIS journals show an increase in citations year after the year. Reference journals, on the other hand, appear to be a plateau after third year of publication. CA Cancer show an exception as its citation decrease dramatically after the peak citations in the second year. The citation trends clearly show that LIS journals as well the reference journals can benefit from 5 year IF or other methods.

Fig 2 shows five years of publications starting from 2007 to 2011 and their corresponding citations in 2011. Publication is classified as total articles, as given

in the WoS, citable articles are those articles that are classified as article, review and proceeding paper in the WoS. Cited articles, those articles that received at least one citation, is also given in the figure. Further, Total citations and citations from the citable articles of WoS is also incorporated into the Fig 2.

In some of the LIS journals such as J Am Soc Inf Sci Tech, Scientometrics, Int Geogr Inf Sci, J Inf Sci, Inform Tech Libr, Online Inform Rev and Int J Coop Inf Syst number of citable articles so far not received even one citations. Therefore, ranking by IF or 5 year IF the result in the increased denominator thus reducing the IF and 5 year IF of these journals.

On the other hand, citable articles of the J Am Med Inform Assn are lower than the cited articles. This means number of citations from those articles such as editorials without increasing the denominator. Most of the reference journals also receive these free citations, but New Engl J Med and Lancet receive a large number of articles contributing to the citation, without increasing the denominator as seen in the difference between citations from citable articles and total citations in Fig 2.

ACCA ranks LIS and reference journals as per their citations and publication characteristics (Table 1). The table also compares ACCA with IF, 5 year IF, SNIP, SJR and Eigen Factor. Therefore, ranking of LIS and all the journals through ACCA would give unbiased and reasonable ranking of the journals.

Acknowledgement

The author acknowledges the Director, Central Leather Research Institute, Chennai for the permission to publish the article.

References

Bharathi, D. G. (2011). Methodology for the evaluation of scientific journals: Aggregated Citations of Cited Articles. *Scientometrics, 86*(3), 563-574.

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics : quantitative methods in library, documentation and information science*: Elsevier Science Publishers.

Eigen Factor. (2011). Eigenfactor.org. from http://www.eigenfactor.org/

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics, 4*(3), 379-391.

Huh, S. (2011). Citation analysis of The Korean Journal of Internal Medicine from KoMCI, Web of Science, and Scopus. *The Korean journal of internal medicine, 26*(1).

Jacso, P. (2001). A deficiency in the algorithm for calculating the impact factor of scholarly journals: The journal impact factor. *Cortex, 37*(4), 590-594.

Klavans, R., & Boyack, K. W. (2007). *Is there a convergent structure of science? A comparison of maps using the ISI and scopus databases*.
Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*(3), 265-277.
Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. *Scientometrics, 81*(2).

Table 1. Impact Factor, 5 year Impact Factor, SNIP, SJR, Eigen Factor and ACCA of LIS and reference journals

Journal Title	ISSN	Impact Factor (2011)	5-year IF (2011)	SNIP (2011)	SJR (2011)	Eigen Factor (2011)	ACCA (2011)
LIS Journals							
Mis Quart	0276-7783	4.447	7.497	3.885	5.138	0.009769	7.02
J Am Med Inform Assn	1067-5027	3.609	4.329	2.510	2.275	0.013606	4.89
J Inf Technol	0268-3962	2.321	3.000	1.147	0.704	0.002678	1.64
J Strategic Inf Syst	0963-8687	1.457	2.000	1.966	1.041	0.000996	2.43
J Manage Inform Syst	0742-1222	1.423	2.945	1.984	1.466	0.004057	2.62
Inform Manage-Amster	0378-7206	2.214	3.796	2.904	2.217	-	4.07
Online Inform Rev	1468-4527	0.939	1.246	1.053	0.751	0.001538	1.79
J Am SocInfSci Tec	1532-2882	2.081	2.113	1.975	1.517	0.013111	3.34
Annu Rev Inform Sci	0066-4200	2.955	2.984	3.357	-	0.001424	3.62
Scientometrics	0138-9130	1.966	2.443	1.407	1.387	0.010084	3.35
Eur J Inform Syst	0960-085X	1.500	2.218	1.688	1.432	0.003038	2.26
Inform Process Manag	0306-4573	1.119	1.443	2.709	1.197	-	1.96
Int J GeogrInfSci	1365-8816	1.472	1.848	1.964	0.846	0.004051	2.35
J Inf Sci	0165-5515	1.299	1.686	1.778	0.885	0.002738	2.18
Inform Technol Libr	0730-9295	0.250	0.398	1.280	0.655	0.000289	0.54
Medical Journals							
Ca-Cancer J Clin	0007-9235	101.780	67.410	41.082	24.976	0.044999	47.91
New Engl J Med	0028-4793	53.298	50.075	14.971	9.740	0.663830	26.72
Lancet	0140-6736	38.278	33.797	6.197	5.917	0.360954	16.39
Multidiscipl. Journals							
Nature	0028-0836	36.280	36.235	8.647	14.548	1.655240	33.41
Science	0036-8075	31.201	32.452	8.064	11.187	1.41162	27.88
P NatlAcadSci USA	0027-8424	9.681	10.472	2.582	5.350	1.60168	19.08

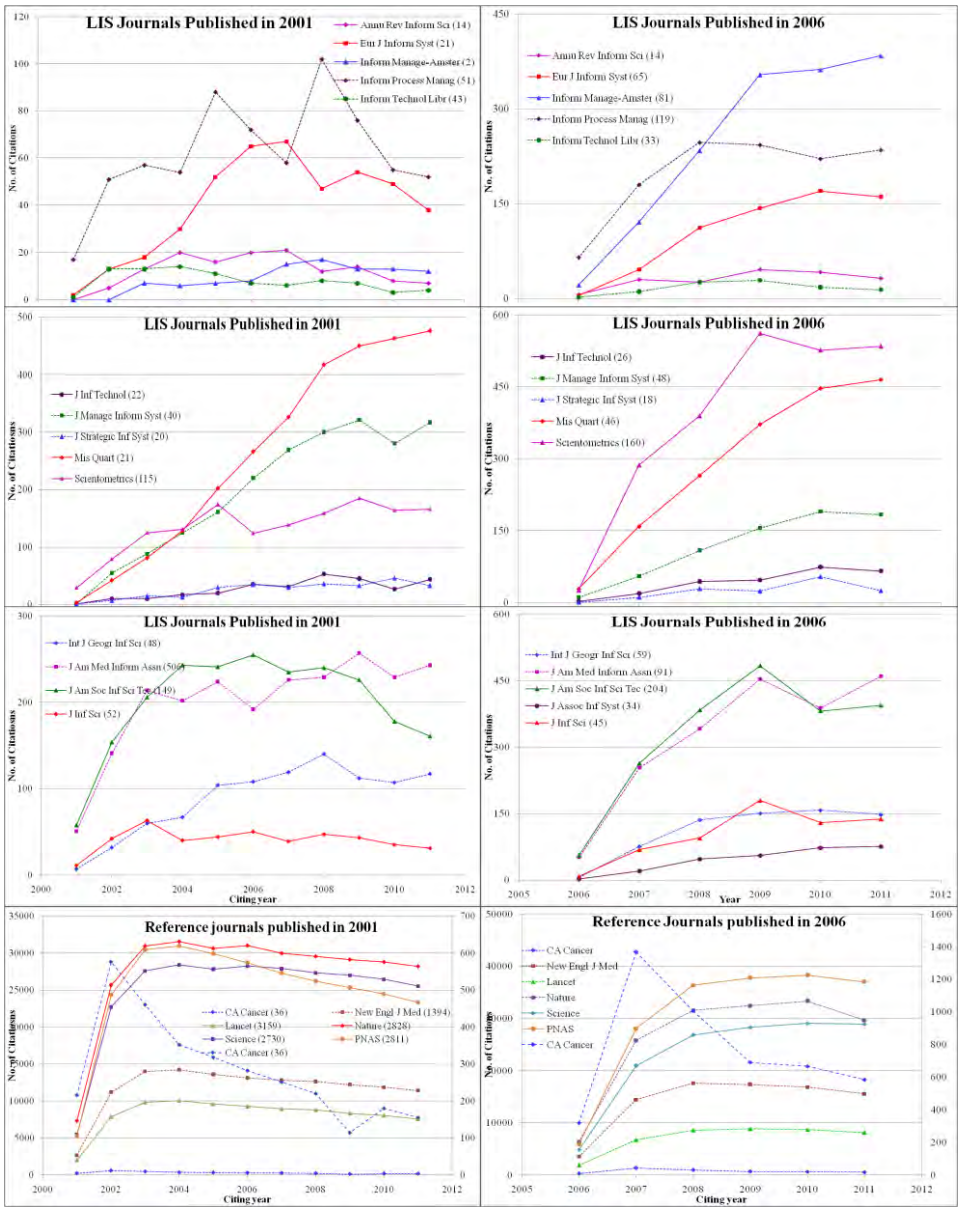


Fig. 1. Citation characteristics of LIS and Reference journals published in 2001 and 2006 (CA Cancer is also shown in the secondary scale for clarity).

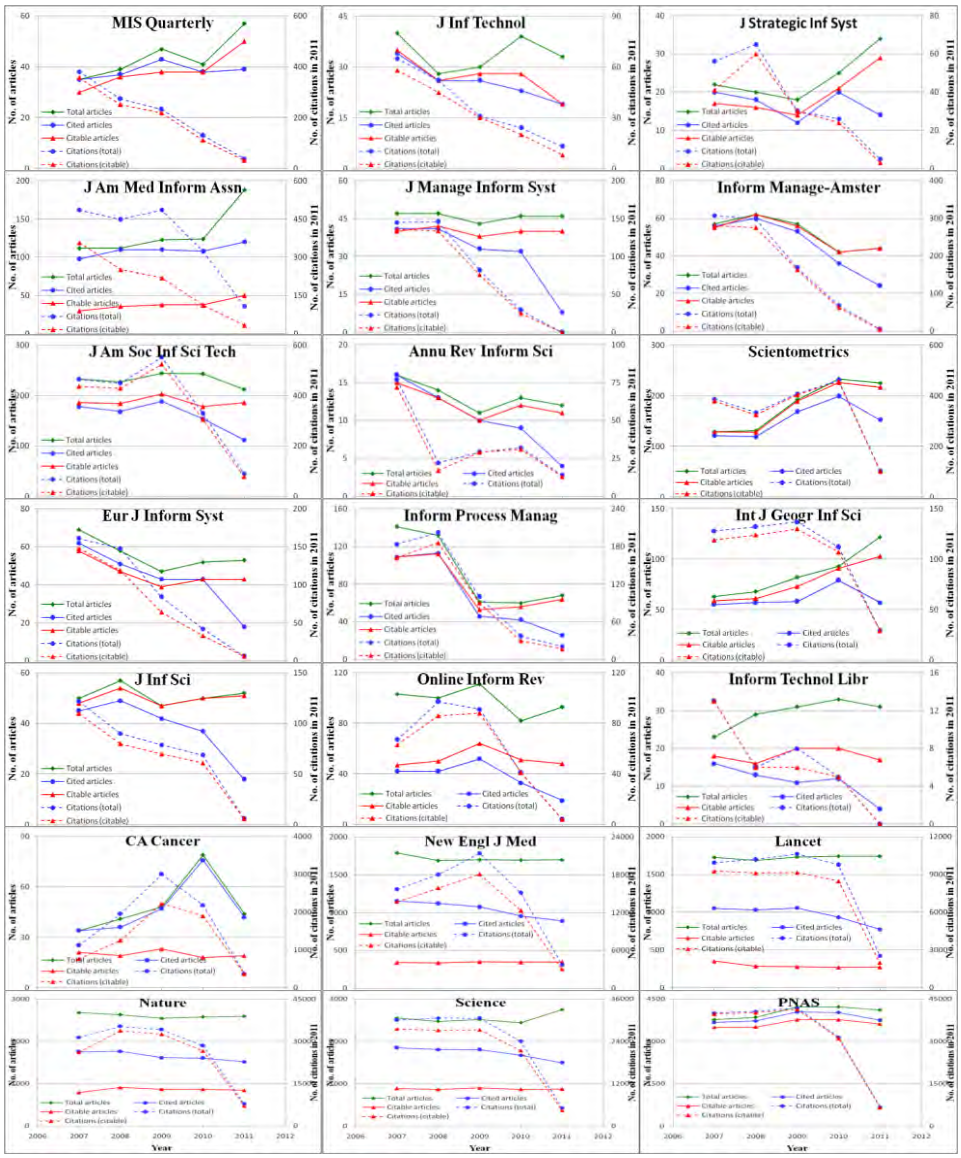


Fig 2. Total articles, citable articles and cited articles published from 2007 to 2011 and total citations and citations from citable articles in the year 2011 for LIS and reference journals.

ANALYSIS OF JOURNAL IMPACT FACTOR RESEARCH IN TIME: DEVELOPMENT OF A SPECIALTY?

Thed van Leeuwen and Paul Wouters

[1] *Leeuwen@cwts.nl*
Centre for Science & Technology Studies (CWTS), Leiden University, Leiden, Wassenaarseweg 62a PO Box 905, 2300 AX Leiden, the Netherlands

[2] *p.f.wouters@cwts.leidenuniv.nl*
Centre for Science & Technology Studies (CWTS), Leiden University, Leiden, Wassenaarseweg 62a PO Box 905, 2300 AX Leiden, the Netherlands

Abstract
In this paper we present the results of an analysis that describes the research centred on journal impact factors. The purpose of the analysis is to make a start of studying part of the field of quantitative science studies that relates to the most famous bibliometric indicator around, and see what characteristics apply to the research on journal impact factors. In this paper we start with a general description of the research, from the perspective of the journals used, the fields in which research on journal impact factors appeared, and the countries that contribute to the research on journal impact factors. Finally the paper presents a first attempt to describe the coherence of the field, which will form the basis for next steps in this line of research on journal impact factors.

Introduction
One of the most widely used bibliometric indicators is the Journal Impact Factor (JIF). This indicator is a relatively simple measure, is easily available, and relates to scientific journals which are the main currency in the natural sciences and biomedicine. The bibliometrics community mainly studied the methodological issues related to JIFs and other journal impact measures, such as EigenFactor (Bergstrom et al, 2008), Audience Factor (Zitt & Small, 2008), SNIP (Moed, 2010). Some confusion has been created as Impact factor started to become a generic term in itself, when talking about bibliometric measures, and the way these are applied. However, this is not the correct use of the term, which only relates to the impact measurement of journals, as designed by Garfied (Garfield, 1972). In that light, JIF is sometimes understood as a way to 'predict' the chance of being cited. Many studies outside the bibliometrics community examined the possibilities of the application of JIFs in management of research, journals and journal publishing in a less critical way, or simply reported on the value of the Journal Impact Factor for their own journal. This literature is an indication of the growing awareness and relevance of this bibliometric indicator for science and science management.

In this study we will describe the development of the research related to Journal Impact Factor from 1981 onwards, until 2010. The focus will be on development of output related to Journal Impact Factor, looking at the cognitive and geographical origin of the output on JIFs. Co-occurrence analysis of title and abstract words is used to see how the publications in the research on JIFs are related. This paper is a first step in a line of research, in which we want to follow the development of the research on journal impact factors, in order to see whether we can speak of the development of a scholarly specialty.

Data collection

We collected from the Web of Science all publications that contain the words "Impact factor" in their title or abstract. This search was conducted in November 2012, and resulted in a set of 2,855 publications of various document types. This set of publications was combined with the in-house version of the Web of Science at CWTS, a bibliometric version of the original Web of Science database. This resulted in a set of 2,467 publications, which was present in both versions of the WoS database. Main reason for the difference is the gap between the periods covered in both sets, where the CWTS version was up to date for analytical purposes to 2011. A detailed analysis of the contents of the publications resulted in the deletion of 367 publications with another topic from journal impact factors[3]. The resulting dataset contained 2,100 publications in WoS.

Methods and indicators

Below we will analyze the disciplinary embedding of the publications selected in the WoS database according to the so-called Journal Subject Categories. As the data collected for the study are collected irrespective of the field to which the publications belong, the set contains publications from a variety of subfields. Information on geographical origin was extracted from the addresses attached to the publications selected. We only looked at country names attached to publications, and counted a country name only once when it appears on a publication.

In this study we use the VOS Viewer methodology, through which structures between publications are identified on the basis of the co-occurrence of title and abstract words (van Eck & Waltman, 2010). We start the analysis of the data collected from 1996 onwards, the year in which WoS publications structurally started having abstracts in the database. This availability of this type of data in our set prescribes that we have 15 years of publication data, which we will analyze according to three equally long periods of five years (1996-2000, 2001-2005, and 2006-2010).

[3] "Impact Factor" is a term or combination of terms that is not only used in relation to journals and scientific publishing, but also relates to engineering disciplines (e.g., in relation to the construction of bridges, and forces working on the steel construction) and in biomedicine (where in pharmacological research impact factors are used to indicate influences on drug treatment effectiveness).

In the part on the disciplinary and geographical background, we introduce a standard bibliometric impact indicator, namely MNCS, the field normalized mean citation score, in order to give an impression to what extent publications in the research on JIFs are more or less influential and visible in the fields to which they belong (Waltman et al, 2011)

Results

The results are presented in Figure 1, which shows that after an initial small increase in number of publications started to grow in since 1994. From this year onwards, the increase was larger every year, and compared to the overall increase of the output in the WoS, the output of the research on JIFs seems to be growing at a faster pace. And although the trends seem to be somewhat influenced by coverage policy effects (the sharp increases in output numbers in 2004 and 2007), the output on JIFs keep growing relative to global output trends, indicating that research on journal impact measures and journal impact factor is booming.

In Figure 2, the trend shown in Figure 2 is broken down into various document types in the WoS database. Normal articles do account for the largest share of the output, as can be expected. Remarkably, the document type editorial covers nearly 25% of all publications on JIFs. Editorials are apparently a way to discuss JIFs. Moreover, we hypothesize that editorials function as a way to make public the value of the journal impact factor of their journals. The other document types play a relatively modest role.

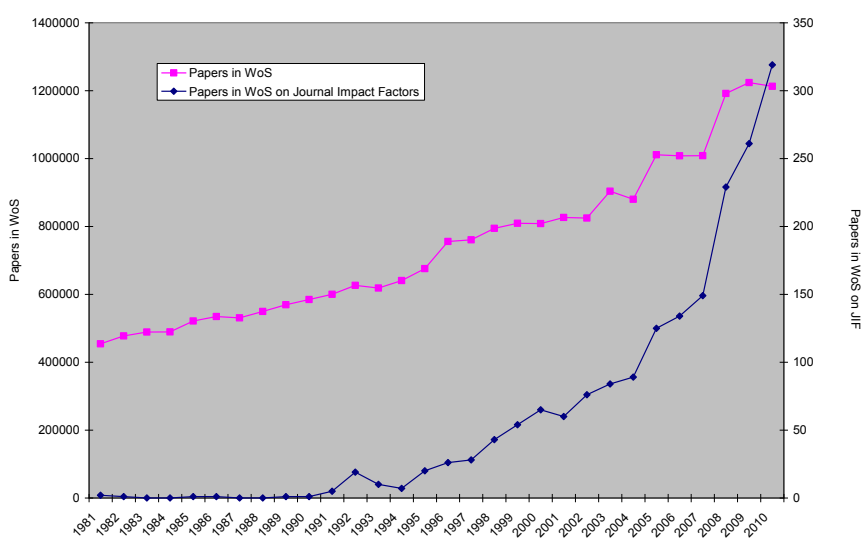


Figure 1: Trend analysis of output in WoS on JIFs, '81-'11

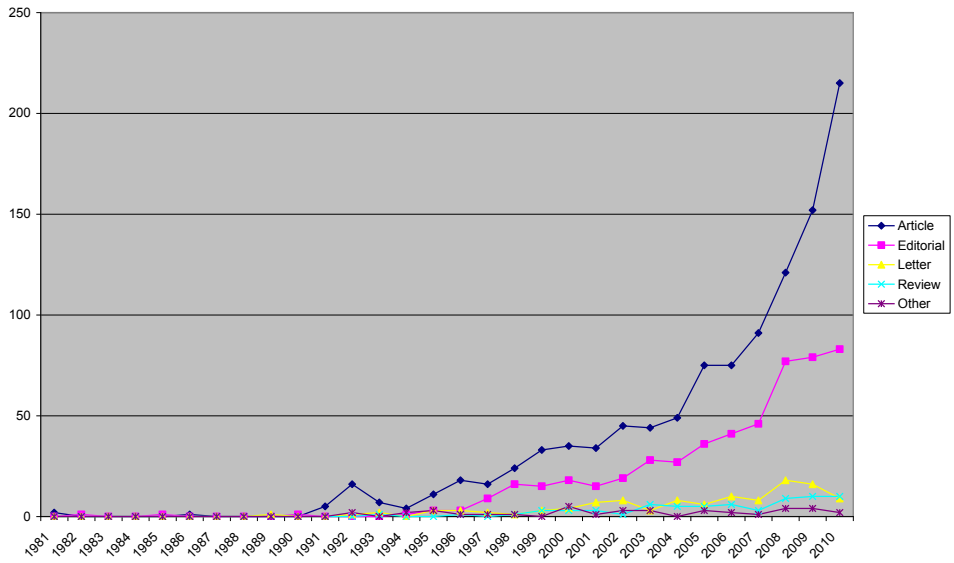


Figure 2: Trend analysis of output in WoS on JIFs across various document types, '81-'11

In Table 1 we present the output on JIFs in the period 1996 to 2010 by journal. We show only the 25 journals that appeared most frequently in the period 2006-2010, as these are strongly contributing to the overall increase of output in the research on JIFs. Among the 25 journals that publish most frequent on JIFs in 2006-2010, we find *Scientometrics* as the top ranking journal, with 80 publications on JIFs, while the journal published on this topic also in the two previous periods. A new journal on this topic is the *Journal of Informetrics*, with 24 publications in the research on JIFs in the period 2006-2010. Other journals among the top-15 most frequently publishing journals that published in every period of the analysis are: *Current Science*, *Learned Publishing*, *Revista Medica de Chile*, and *JAMA*. For eight journals we observe output in the two last periods, while the other journals started publishing on JIFs only in the period 2006-2010. Research and publishing on JIFs is becoming more popular from 2006 onwards.

In Table 2 the journals are shown that contained publications on JIFs in every period of our analysis. The first five journals were mentioned in the discussion of Table 1, but remarkably enough the sixth journal, the *Journal of Documentation,* one of the main library and information science journals, displays a decreasing number of publications on JIFs. Another remarkable fact is the relative large number among these 22 journals of Spanish background, six in total. An explanation for this phenomenon may be the strong development of the field of library and information science in Spain. In the manual selection process we noted a relatively large number of publications from Germany and German

language publications, on JIFs. However, this did not result in a very strong visibility of one particular journal from Germany or in the German language among those high ranking of frequently publishing journals (as presented in Tables 1 and 2, respectively).

Table 1: Output numbers across journals in WoS on JIFs, '96-'10 (based on most frequently occurring journals in '06-'10)

#	*Journal*	*P 96-00*	*P 01-05*	*P 06-10*
1	SCIENTOMETRICS	25	39	80
2	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY		9	29
3	JOURNAL OF INFORMETRICS	.	.	24
4	CURRENT SCIENCE	3	6	12
5	PLOS ONE	.	.	11
6	ARCHIVUM IMMUNOLOGIAE ET THERAPIAE EXPERIMENTALIS	.	.	10
7	EPIDEMIOLOGY	.	.	9
8	LEARNED PUBLISHING	1	1	9
9	ONLINE INFORMATION REVIEW	.	1	8
10	ARCHIVES OF ENVIRONMENTAL & OCCUPATIONAL HEALTH	.	.	7
11	CORTEX	.	3	7
12	INDUSTRIAL HEALTH	.	.	7
13	INTERNATIONAL JOURNAL OF CARDIOLOGY	.	3	7
14	RESEARCH EVALUATION	.	4	7
15	JOURNAL OF CLINICAL EPIDEMIOLOGY	.	3	6
16	REVISTA MEDICA DE CHILE	1	1	6
17	CHIRURG	1	.	5
18	CLINICS	.	.	5
19	ELECTRONIC LIBRARY	.	.	5
20	JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	1	5	5
21	JOURNAL OF CHILD NEUROLOGY	.	2	5
22	JOURNAL OF THE MEDICAL LIBRARY ASSOCIATION	.	5	5
23	REVISTA DE NEUROLOGIA	.	.	5
25	SAO PAULO MEDICAL JOURNAL	.	.	5
25	SCIENCE	.	.	5

In Table 3, we present the distribution of main contributing countries to the research around JIFs. The countries are shown according to the order of numbers of publications in the period 2006-2010. The USA takes a first position in the research on JIFs. Rather surprising is, and it was mentioned before in the analysis on journals in the research on JIFs, is the position of Spain. Although the share of the output of Spain decreases, the absolute numbers increases strongly, and equally interesting, the citation impact of these publications increases as well. An explanation for this can be found in the fact that in Spain Library and Information Sciences is a well-developed discipline at universities all over the country, contrary to many other European countries, in combination with the coverage of Spanish language journals in which applications of JIF, and in particular in biomedicine, are published. China appears in the second period of our analysis, increasing its output in research on JIFs even more in the last period. Most

countries seem to contribute to the boost in output as we have seen before. And although based on somewhat lower numbers of publications, the citation impact of the Dutch papers on JIFs stands out.

Table 2: Output numbers across journals in WoS on JIFs, '96-'10 (based on regular occurrences across three periods)

#	*Journal*	*P 96-00*	*P 01-05*	*P 06-10*
1	SCIENTOMETRICS	25	39	80
2	CURRENT SCIENCE	3	6	12
3	LEARNED PUBLISHING	1	1	9
4	REVISTA MEDICA DE CHILE	1	1	6
5	JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	1	5	5
6	JOURNAL OF DOCUMENTATION	6	5	4
7	QUIMICA NOVA	1	1	4
8	ANESTHESIA AND ANALGESIA	2	1	3
9	BRITISH MEDICAL JOURNAL	2	1	3
10	CROATIAN MEDICAL JOURNAL	2	4	3
11	INFORMATION PROCESSING & MANAGEMENT	1	3	3
12	NATURE	5	1	3
13	OCCUPATIONAL AND ENVIRONMENTAL MEDICINE	1	1	3
14	ORAL ONCOLOGY	1	1	3
15	BRAZILIAN JOURNAL OF MEDICAL AND BIOLOGICAL RESEARCH	1	2	2
16	JOURNAL OF INFORMATION SCIENCE	5	6	2
17	ACTAS ESPANOLAS DE PSIQUIATRIA	1	3	1
18	CANADIAN MEDICAL ASSOCIATION JOURNAL	3	3	1
19	CARDIOVASCULAR RESEARCH	3	4	1
20	MEDICINA CLINICA	7	3	1
21	REVISTA CLINICA ESPANOLA	1	2	1
22	REVISTA ESPANOLA DE ENFERMEDADES DIGESTIVAS	1	2	1

Table 3: Output numbers across countries in WoS on JIFs, '96-'10

	96-00			*01-05*			*06-10*		
	P	*%*	*MNCS*	*P*	*%*	*MNCS*	*P*	*%*	*MNCS*
USA	27	18.62	1.77	52	15.85	2.02	169	19.43	1.78
SPAIN	21	14.48	0.99	45	13.72	1.44	87	10.00	1.71
AUSTRALIA	2	1.38	0.71	10	3.05	1.08	56	6.44	2.25
GREAT BRITAIN	7	4.83	1.71	26	7.93	1.30	56	6.44	1.15
GERMANY	19	13.10	0.64	28	8.54	0.78	49	5.63	0.77
CHINA	.	.	.	8	2.44	1.10	44	5.06	1.09
BRAZIL	3	2.07	0.23	6	1.83	0.71	39	4.48	0.58
NETHERLANDS	7	4.83	6.17	10	3.05	2.26	33	3.79	4.18
FRANCE	12	8.28	1.18	12	3.66	0.62	29	3.33	1.90
ITALY	6	4.14	0.96	20	6.10	0.99	29	3.33	1.54
CANADA	1	0.69	0.00	10	3.05	2.86	24	2.76	2.33
GREECE	1	0.69	4.41	13	3.96	1.17	23	2.64	1.94
INDIA	9	6.21	2.06	9	2.74	1.16	22	2.53	0.94
BELGIUM	2	1.38	2.92	11	3.35	3.00	18	2.07	2.32
DENMARK	4	2.76	4.12	9	2.74	2.45	9	2.74	2.45

In Figure 3, we compare the output of the countries most active in the research on JIFs with their total contribution to global science in the period 2006-2010. Please note that shares are taken among this group only, so the global shares presented here are not the actual contributions, these might be somewhat smaller due to the exclusion of some countries from this analysis. Moreover, these scores contain all document types, since editorials seem to be of importance in the research on JIFs. The countries are presented in the order of their contribution to global science. So we expect the USA and China, together with large science producing European countries such as Great Britain, Germany and France to be on top of the figure. However, in all these cases, their contribution to the research on JIFs is lower compared to their contribution to global science. Spain has a more than double as high contribution to the research on JIFs, just as Denmark and Greece. Other countries that are over represented compared to their national share on global science are Australia, the Netherlands, and Brazil.

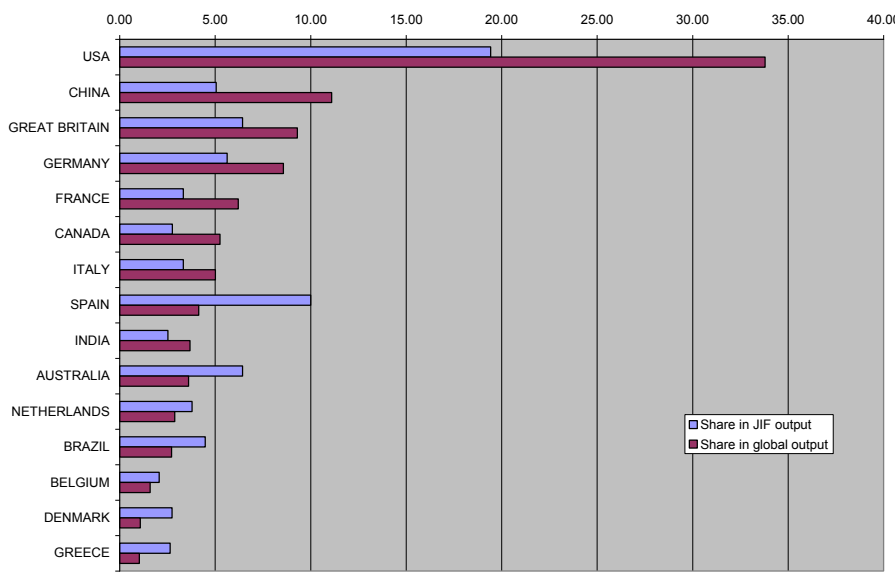


Figure 3: Comparing output numbers on JIFs and global science, '06-'10.

In Table 4, we present the disciplinary background of the journals publishing on JIFs. The social sciences field ***Library & Information Science*** plays the most important role here. Many journals classified under this heading in the WoS are also labelled as ***Computer Science, interdisciplinary applications*** (which explains why that field is so strongly visible in all three periods). The second largest field when it comes to publications on the topic of JIFs is ***Medicine, general & internal***. This field contains, next to the well known general medicine journals such as *New England Journal of Medicine*, *British Medical Journal*, *JAMA*, and *The Lancet*, many local medicine journals, many of which publish

occasionally on JIFs. We see this as evidence for the popularity of journal based impact factors in these fields.
Another remarkable phenomenon in the overview of the disciplinary composition of the research on JIFs is the fact that these publications tend to have high impact, and more particularly, those publications in rather peripheral fields, as seen from the core of the research on JIFs, still seem to generate high impact scores, while these publications also appear in journals with a quite high impact standing in their respective fields (not shown in the table). The main reason for this high impact position of JIF publications is the fact that these appeal to two types of audiences: the core bibliometrics community, conducting research on the methodological and application dimension of the indicator, as well as the 'user' audience of JIFs, in which various applications of JIF are analyzed.

Table 4: Output numbers across fields in WoS on JIFs, '96-'10

	96-00			*01-05*			*06-10*		
	P	*%*	*MNCS*	*P*	*%*	*MNCS*	*P*	*%*	*MNCS*
*INFORM SC&LIBR	28.00	19.44	3.50	55.50	18.62	2.97	145.80	19.36	3.23
MEDICINE,GEN&INT	21.50	14.93	1.05	33.00	11.07	1.27	51.17	6.80	1.00
COMP SC,INT APPL	12.33	8.56	2.34	19.00	6.38	1.64	43.22	5.74	1.53
SURGERY	1.00	0.69	1.28	9.83	3.30	0.73	35.33	4.69	0.84
MULTIDISCIPL SC	7.00	4.86	4.04	6.00	2.01	4.57	27.00	3.59	2.50
COMP SC,INFO SYS	6.83	4.75	2.56	13.33	4.47	2.01	25.13	3.34	3.78
PUBL ENV OCC HLT	2.00	1.39	0.72	10.00	3.36	0.31	19.00	2.52	1.16
BIOLOGY	0.50	0.35	0.99	4.33	1.45	0.54	16.33	2.17	1.33
*PSYCHOL,MULTID	0.25	0.17	0.00	10.00	3.36	1.13	16.33	2.17	1.47
NURSING	.	.	.	2.00	0.67	2.00	13.83	1.84	1.73

Next, we want to focus on the way the publications in the research on JIFs are inter-related on the basis of terms (title and abstract words), and how these terms co-occur on the publications in the research on JIFs. For this we used the VOS Viewer methodology.
In Figure 4 we present the publications in the period 2006-2010. The words plotted in the graph show a dense network, in which we distinguish three different clusters, which are of nearly equally large size /volume of words, and density. On the lower left (in red), we observe the cluster that contains the core of the library and information science and evaluation related topics. The second cluster (in green) contains both elements of scientific publishing as well as terms form biomedicine, while the lower-right part of the figure contains the third cluster (in blue). This cluster contains mainly elements of a geographical nature.

In Figure 5 we present the density map from the VOS Viewer methodology. The map, and particularly the colour-coding, indicates the density of words, and their relationship with neighbouring words in the map (blue indicates low density, and red indicates a high density). We can see that the cluster that we described in Figure 4 as the core of library and information science, is the most active area in research on JIFs (the red area left).

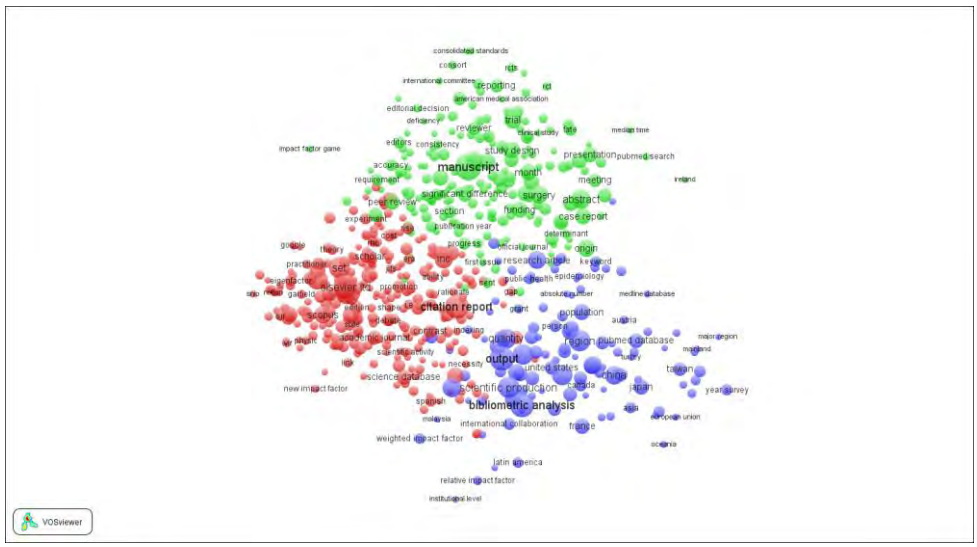


Figure 4: Term map of title and abstract words in output on JIFs, '06-'10 (based on VOS Viewer)

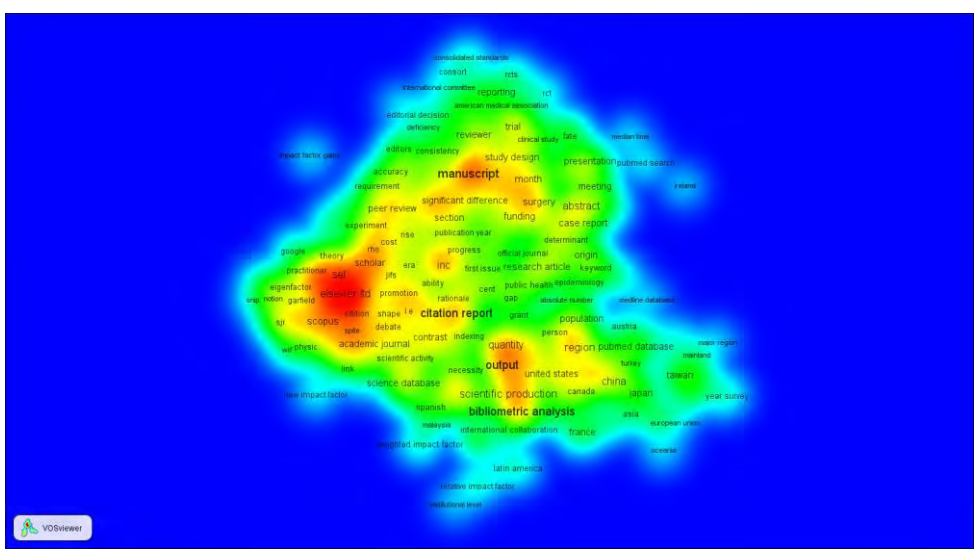


Figure 5: Density map of title and abstract words in output on JIFs, '06-'10 (based on VOS Viewer)

Conclusions, Discussion and Future Research

In this study we first focused on a description of the characteristics of the research on JIFs. We observed a strong growth in output in the research on JIFs, even stronger than the overall output growth in WoS. Next, editorials play an important role in the publishing on JIFs, as these do cover some 25% of the output in the research published on JIFs. We concluded that some countries contribute particularly strongly to research on JIFs, such as Spain and Australia. For these countries we observe relatively larger contributions to the research on JIFs, compared to their overall contribution to science.

The initial selection of the publications in this study taught us that we can distinguish three different types of publications on JIFs: publications from the field of library and information science that forms the core of the research on the topic (e.g., the comparison of JIF with newly developed journal impact measures); a set of publications in other fields that relate to the popularity of the indicator in research management (e.g., publications that report on the value of the JIF, or proposes usage in a policy context); and finally research papers on the controversies around JIFs (these can be of a methodological or a more policy oriented nature).

The VOS Viewer graphs in the paper suggest a strong coherence of the research on JIFs. However, future research based on citation relations might help understanding the development of the research on the topic in more detail. Does research on JIFs demonstrate the characteristics of an emerging specialty or can we explain the observed coherence in other ways? We are also interested in the question of replicablility and redundancies in the literature. Is this area demonstrating scientific progress by building up a more advanced body of knowledge or do we rather witness a cyclical process in which older findings are regularly repeated?[4] And how can we characterize the social network underpinning the body of literature? Do we see a fragmented adhocracy or rather a distributed community?

Acknowledgments

The authors want to express their thanks to Ludo Waltman, for his instruction on the VOS Viewer usage for this study.

[4] The issue of the cyclic nature or redundancy in reporting research came to our attention by three publications on the composition of the JIF, and more in particular on the nominator and de denominator. These publications (Rossner et al, 2007, McVeigh & Mann, 2009, Hubbard & McVeigh, 2011) seem to report on the exact same issues as were reported in two publications from the 1990's (Moed & van Leeuwen, 1995, Moed & van Leeuwen, 1996)

References

Bergstrom, C.T., D. Jevin, and M.A. Wiseman (2008) The Eigenfactor ™ Metrics. Journal of Neuroscience, 28, 11433-11434

Garfield, E. (1972), Citation analysis as a tool in Journal Evaluation, Science, 178, 471-479.

Hubbard, SC & ME McVeigh (2011) Casting a Wide net: the Journal Impact Factor numerator, *Learned Publishing*, 4, 133-137.

Mc Veigh, ME &SJ Mann, (2009) The Journal Impact factor denominator. Defining citable (counted) items, *JAMA*, 302, 1107-1109

Moed, HF & TN van Leeuwen (1995) Improving the accuracy of Institute for Scientific Information's JIFs, *Journal of the American Society for Information Science (JASIS)* 46, 461-467.

Moed, HF & TN van Leeuwen (1996) Impact factors can mislead *Nature* 381, 186.

Moed, H.F. (2010) Measuring contextual citation impact of scientific journals. Journal of Informetrics, 4, 265-277.

Rossner, M, H. van Epps, and E Hill. (2007) Show me the data. Journal of Cell Biology, vol 179 (6), 1091-1092

Van Eck, NJ, & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84, 523-538

Waltman, L, NJ van Eck, TN van Leeuwen, MS Visser, & AFJ van Raan (2011) Towards a new crown indicator: Some theoretical considerations, *Journal of Informetrics*,, 5, 37-47

Zitt, M. and H. Small (2008) Modifying the journal impact factor by fractional citation weighting: The audience factor. Journal of the American Society of Information Science Technology, 59, 1856-1860

THE ANALYSIS OF RESEARCH THEMES OF OPEN ACCESS IN CHINA: IN THE PERSPECTIVE OF STRATEGIC DIAGRAM (RIP)

Zhao Rongying[1] and Wu Shengnan[2]

[1]*Zhaory@whu.edu.cn* [2]*vivian19870220@163.com*
Center for Studies of Information Resources of Wuhan University,
Research Center for Science Evaluation of Wuhan University,
School of Information Management, Wuhan University, Wuhan, China 430072

Abstract:
Using Chinese National Knowledge Infrastructure (CNKI) as the data resource, this paper searched some papers about Open Access. Some VBA programs were developed to generate the co-word matrix, compute the E-index value of keywords as well as the density and centrality of thematic clusters. Callon's clustering method was also used to generate keywords clusters. Then, co-word analysis method and strategic diagrams were utilized to detect the main research themes as well as explore the development situation and status of these research themes. Based on this, some conclusions were got in the end.

Conference Topic
Open Access and Scientometrics (Topic 10).

Introduction

Open Access (OA) is one of the most popular publishing systems that the academic information can be shared freely. Under the condition of open access, any researcher can access the academic achievements without the limitation of time, place and money (Li and Liu2004). In recent years, open access received more and more attention of the international community. Compared to the traditional commercial publishing system, Open Access has an unparalleled advantage in encouraging academic exchange. Increasingly, scientific research has been carried out with open access in China. Shen and Gao (2011) led the way in studies of Chinese research situation of open access from the year of 2003 to 2009. Subsequent research focused on Chinese research status of open access during other period (Wang2005; Chen and Zhu 2008). However, these researches mentioned above mostly adopted the traditional methods, such as word frequency count, to reveal the research status of open access. There were significant limitations in revealing the development of research themes in the field of open access. In order to overcome the shortcomings of previous studies, this paper advances the techniques to study Chinese research situation of open access more completely.

Data and Methods

Data

This study chose Chinese National Knowledge Infrastructure (CNKI: http://www.cnki.net) as the data source. CNKI, regarded as "China's largest full-text database", covers the widest range of academic journals published in China. We retrieved the related articles open access field with the keyword "open access", and the time span was not limited. There were 1447 documents, each documents contains author name, title, affiliation, keywords etc. After deleting duplicate records and comments, we finally obtain data sample of 1364 articles.

A total of 1364 related articles and 5095 keywords were collected as the co-word analysis sample. Taking into account the frequency of most keywords is low, which results in the low co- occurrence frequency. We chose the top 50 keywords(word frequency>=10) as the research subjects, which are shown in table 1.

Table 1. The top 50 high frequency keywords

keyword	*frequency*	*keyword*	*frequency*	*keyword*	*frequency*
Open access	1126	Government information resource	21	Open resource	16
library	148	Information service	21	Information open	12
institutional repository	116	Academic resource	20	Research situation	12
OA journal	114	Information resource construction	20	American	12
High school library	99	Self-archiving	20	DOAJ	12
Academic exchange	75	Institutional repositories	19	Electronic journal	12
Information resource	58	countermeasure	19	OA publishing	12
Academic journal	48	Digital library	18	Information sharing space	11
copyright	34	Resource sharing	17	Literature resource	11
Scientific journal	33	influence	17	Web resource	11
OA resource	32	intellectual property rights	16	Digital resource	11
Academic publishing	31	policy	16	Database	11
Journal	26	Information exchange	16	Open journal	11
Publishing model	26	Quality Control	15	Publishing	11
OA repositories	25	Development	14	Preprint	10
Resource construction	24	repository	13	online publishing	10
Academic information	23			Citation analysis	10

Methodology

Co-word analysis

Co-word analysis, counting and analyzing the co-occurrences of keywords in articles on the given subject, could provide an immediate picture of the actual content of research topics (Callon etal. 1991; Ding etal. 2001). In co-word analysis, once a research area is selected, keywords are extracted from the related journal articles or other publications; and then, a matrix based on the keyword co-occurrence will be built. The value of the cell in matrix represents the co-word frequency of two words. The higher co-occurrence frequency of two keywords means the more correlative they are. Finally, the original matrix is transformed into a correlation matrix using specific correlation coefficient for the further analysis.

Clustering method

In this study, the clustering algorithm was learned from Callon(1991).E-index is the core indicator which this algorithm used, and formula is shown as below:

$$Eij=（Cij/Ci）*（Cij/Cj）=（Cij）^{2}/Ci*Cj。$$

In this formula, Ci /Cj represents frequency of word i/j in the data sample, and Cij represents the co-occurrence frequency of word i and word j. Generally speaking, the E-index value is between 0 and 1. The greater its value is, the larger co-occurrence frequency of the words is.

In Callon's clustering algorithm, each cluster contained less than 10 keywords. What's more, We regarded a couple of keywords whose E-index value is the largest in this cluster as the core content of this cluster. Moreover, specific steps of the cluster algorithm were as follows:

a. A couple of keywords whose E-index is the largest in the co-word matrix was chosen as the main content of the first cluster;

b. Ordering the E-index between chosen keywords and the other keywords, and then we chose ten keywords according to the descending sort order;

c. When got the first cluster, we must delete the rows and columns where keywords of the first cluster stayed to ensure that those keywords are not added to the other cluster;

d. Repeating the above steps until all the keywords whose value of E-index is not zero were added in clusters. Meanwhile, even if there were some keywords in the co-word matrix, the clustering progress was end. Because the E-index value of those words is zero, which means there were not co-occurrence relationship among those words.

Strategic Diagram

The strategic diagram was proposed by Law (1988). The strategic diagram divides these clusters into four quadrants. In the strategic diagram, x-axis stands for degree centrality representing the strength of interaction among research fields. The high degree centrality means that research field may tend to lie in an essential and center position. y-axis stands for density representing the internal relation in a

specific research field. From the perspective of research field, density represents the capability to maintain and develop itself (Law et al. 1988). In this study, the density of each cluster was calculated by two steps. We summed up the co-word frequency in each cluster first, and then calculated their averages. What's more, degree centrality of each cluster was calculated though the sum of co-word frequency between keywords in this cluster and keywords in the other clusters.

Result and Discussion

Clustering

In this study, a program in VBA was developed to calculate the times that two keywords appeared together in the same article. Subsequently, we achieved a co-occurrence matrix called symmetric matrix. The data in diagonal cells were the frequency of the top 50 keywords and the values of non-diagonal cell were the co-occurrence frequency. And then, this original matrix was transformed into a E-index's correlation matrix to indicate the similarity and dissimilarity of each keyword pair.

According to the clustering algorithm mentioned above, These 50 keywords of OA field in China were divided into seven clusters named Cluster1 to Cluster7. Each cluster stands for a research theme of OA field in China. As mentioned above, a pair of keywords whose E-index is the largest in this cluster was regarded as the core content of this cluster. So the content of each research theme of OA field in China was as shown below:

Topic 1: OA of the government's information resource
Topic 2: Influence of OA over the information sharing and scholarly communication
Topic 3: OA journal and OA repositories
Topic 4: Quality evaluation of OA journals
Topic 5: Development strategy of OA
Topic 6: OA Publishing of academic journals
Topic 7: Institutional repositories' building strategy

Analysis of research themes by Strategic Diagram

In this study, we analyzed the research themes by Strategic Diagram. Density and degree centrality of each cluster were calculated on the basis of the method mentioned above. The origin is (2.433819, 42.99014), the average of degree centrality and density. In order to make the origin of Strategic Diagram to be (0,0), density and degree centrality of each cluster were re-calculated, as shown in table 2.

Table 2. Density and degree centrality of each cluster

cluster	Density	Centrality	*Normalized* Density	*Normalized* Centrality
1	1.69	18.9	-0.75	-24.09
2	10.5	112.11	8.07	69.12
3	1.58	33.4	-0.86	-9.59
4	1.00	24.71	-1.43	-18.28
5	1.38	58.00	-1.05	15.01
6	0.39	24.56	-2.04	-18.43
7	0.50	29.25	-1.93	-13.74

A strategic diagram was generated as Fig. 1 according to Table 2. The strategic diagram divided these seven clusters into four quadrants. Research topics with high degree centrality and density in quadrant I are well-developed and the core of the field. Research topics in quadrant II are not central but well-developed. Research topics in quadrant III are both marginal and neglected. Research topics in quadrant IV are central in the network but undeveloped (Callon etal.1991).

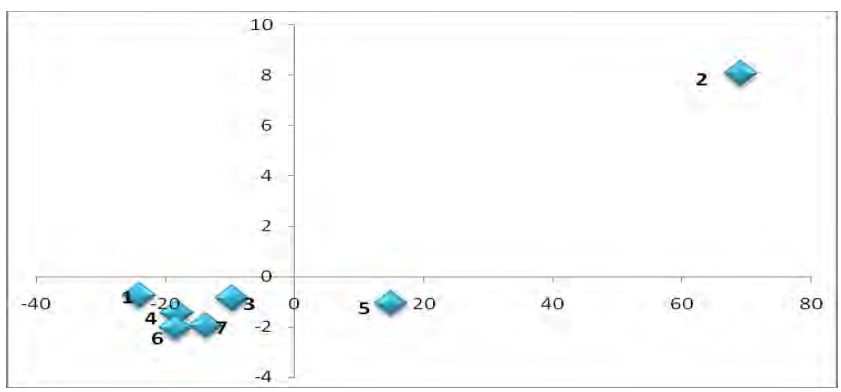


Figure1. The strategic diagram of seven clusters

From the overall distribution of the research themes in strategic diagram, we can see that there were only two topics on the right side of y-axis. In the other word, most of research themes of OA field in China were not centralized, so most research on OA in China now were marginal and undeveloped. The study of OA in China was at the beginning stage. What's more, we also noticed that more than 80 percent of research themes of OA field in China were under the x-axis. That's to say, the degree of most OA topics is low, which indicates that most of OA research themes in China were not well-developed. On the basis of literature research, we found that the research of OA in China mainly focused on two aspects: the basic concept of OA and the domestic and foreign development in OA research. All these research played an important role in promoting Chinese research in OA field. However, it also revealed that Chinese research was mainly

in theoretical stage of shallow level. As for empirical research, Chinese research was seldom involved. All this showed that Chinese research on Open Access was at the beginning stage, the research of Open Access were not well-developed and mature. This discovery brought into correspondence with the result displayed in strategic diagram.

As mentioned above, the strategic diagram divided these seven clusters into four quadrants. Subsequently, we introduced the current status and trend of research themes more clearly through the quadrant analysis.

(1) Cluster in Quadrant Ⅰ is cluster 2. This cluster's degree centrality and density were both high. High density indicates that these clusters are of high internal correlation, and the research topics in clusters have been well-developed and tend to be mature in China. High degree centrality indicates that the cluster is widely connected with other clusters. Research theme in this cluster is the core content of Open access in China. As we all know, the aim of Open Access is to resolve the crisis of scholarly communication. It can well explain why research theme in cluster 2 became the research focus of Chinese researcher as well as the core content of Open access in China.

(2) Cluster 1,3,4,6 and 7 located in Quadrant Ⅲ. The low density and degree centrality reveal that the research topics in these clusters were marginal and undeveloped in China. The reason for these clusters' low density and degree centrality is that there was a significant bottleneck for the research of these topics. In the future study, the research of these topics may break through the vase neck, and become the research hotspots in OA field. Possibly, the research of these topics may not break through the vase neck, and these research topics in OA field were still at the marginal level in China.

(3) Cluster5 located in quadrant IV with high degree centrality but low density. This phenomenon illustrated that the research topics in these clusters are the cores but undeveloped in OA field in China. On the whole, this research topic was with great potential for development. However, because of the lower capability to maintain and develop itself, the research of this theme was usually not as steady as we think. If the research topic in cluster 5 wanted to become research trend, it needed to be further studied.

Conclusion

The purpose of this study is to explore the research situation of Open Access in China. Co-word analysis, and subsequent Callon's clustering method were used to discover the research themes of OA field in China. The results of this study identified seven research themes of Open Access in China.

Strategic diagram method was also used to accurately estimate the research status of each theme in OA field in China. On the basis of the overall display of the strategic diagram, it can be said that less mature but more marginal research was the current situation of OA research in China. A quadrant analysis was also made to indicate the current status of each theme in detail. The results showed that Major research topics of OA field have formed in China, but there are more

smaller and isolated research topics. It could be said that the research topics in Cluster2 and 5 are the cores of OA field in China. The well-developed and core research theme of OA field in China are fewer, such as Influence of open access over the information sharing and scholarly communication property (in Cluster2). In this study, the conclusion was draw from the above analysis. In the Open Access field, there are more theoretical researches, but fewer technological and practical researches. Influence of open access over the information sharing and scholarly communication (in Cluster2) and development strategy of open access (in Cluster5) are the significant research topic in China. Recently, OA journal, OA repositories and OA Publishing were being paid more and more attention to. Therefore, the whole research of OA in China should integrate theory (as basis), technology (as support) and practical researches (as core).

Acknowledgments

This paper is supported by Major Program of National Social Science Foundation in China (11&ZD152).

References

Callon, M etal(1991). Co-word analysis as a tool for describing the network of interaction between basic and technological research. Scientometrics, 22(1):155-205.

Chen hongqin, Zhu Ning(2008). Bibliometrics Analysis of the Literature on Open Access from 2003 to 2007 in China. Information Science (in China),26(9):1317-1320.

Law, J., Bauin, S., Courtial, J. P., & Whittaker, J. (1988). Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification. Scientometrics, 14(3), 251–264.

Li Wu, Liu Ziheng (2004). A new scholar publishing model: Open Access publishing model. Journal of Library Science in China,30(154) : 66-69.

Shen Chen, Gao Zhimin(2011). Visualizing Map of the Researches on Open Access from 2003 to 2009 in China: Based on Journals from CSSCI. Library and Information Service (in China), 55(24):61-65.

Wang Yuncai(2005). The review of Open Access at home and abroad. Library and Information Service (in China), (6):40-45.

ANALYSIS OF THE WEB OF SCIENCE FUNDING ACKNOWLEDGEMENT INFORMATION FOR THE DESIGN OF INDICATORS ON 'EXTERNAL FUNDING ATTRACTION'

Alfredo Yegros-Yegros and Rodrigo Costas

{a.yegros, rcostas}@cwts.leidenuniv.nl
CWTS-Centre for Science and Technology Studies, Leiden University, PO Box 905
2300 AX Leiden (the Netherlands)

Abstract
Indicators based on 'funding attraction' provide information on the degree of successfulness and capacity in acquiring (or attracting) external research funding. Bibliometric databases on their side are starting to collect and provide information on the funding that scientific authors acknowledge in their publications. This study analyses the extent to which these acknowledgements reflect the actual funding of research units, with the aim of exploring the possibilities of developing indicators that can inform on the capacity and degree of research units to 'attract' funding for their research.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction
Research income is normally considered in research evaluation systems as a performance indicator. Among the different sources of public funding for scientific research probably the most important in all the countries and across all disciplines is the government funding, which is transferred directly to the universities or public research institutions but it is also distributed in an indirect way by means of competitive calls for funding research projects. Other sources of research funding include supra-national funding bodies (e.g. the European Framework Programmes), private companies and other type of institutions like foundations, charities and other for non-profit organizations.
While the direct governmental funding is considered an internal 'core' funding of the research institution, both the indirect governmental funding (through research councils and other public funding bodies) and contract research are considered external sources of funding.
Indicators based on funding attraction usually focus these external sources (i.e. funding that researchers have to acquire from outside their own institution). This

"external" character it is somehow reflected in the terms used in the literature on funding attraction indicators, for instance, grant income, third-party funding or external research funding.

The rationale behind indicators based on external research funding is that the application for research grants usually entails a peer review process in a competitive environment (Rigby, 2011) in which only the best proposals in terms of quality and promising results are awarded. In this sense, it is considered to be a kind of reflection of some research quality and therefore it is sometimes used as a performance indicator.

However, some authors have raised also warnings on the usefulness of the external research income as indicators of research performance. Gillett (1991) stresses that research income constitutes an input measure and therefore it does not provide any information on the research produced with the help of this financial support. Others point some limitations that contribute to decrease its validity as indicator of research performance (Hornbostel, 2001; Laudel, 2005, 2006), or even its potential negative effects in terms on research efficiency (Schmoch & Schubert, 2009).

Leaving aside the debate on whether the external research funding constitutes a suitable indicator on scientific research performance, we consider that the acquisition of external funds is part of the scientific endeavour, and it reflects the capacity of researchers to capture/attract financial resources and their competitiveness (Garcia & Sanz-Menendez, 2005). However, the information related to research funding acquired by researchers or research units is not always easily accessible.

The recent inclusion of research funding acknowledgements in the web of science[5] represents a new challenge to the possibility of creating new indicators of 'funding attraction' based on this new information. In principle, it can be considered that funding acknowledgements in scientific publications could be a good way of measuring the capacity of attracting external research funding given that this is the type of funding that is normally acknowledged by authors in contrast to the internal funding (Wang & Shapira, 2011). Indeed, our results indicate that the share of publications acknowledging funding is related to the actual relative external funding in research units.

On the other hand, some indicators have been already proposed on the basis of the funding acknowledgements provided by the Web of Science, for instance in order to measure the internationalization level of funding agencies (van der Besselaar, et al, 2012) and also trying to measure the gain of citations for publications that have any journal peer review acknowledgment[6] (Costas & van Leeuwen, 2012). However, to our knowledge this source has never been used to create indicators on the capacity of capturing/attracting funding resources by research units. Part of

[5] http://wokinfo.com/products_tools/multidisciplinary/webofscience/fundingsearch/.

[6] When the authors acknowledge funding, the Web of Science include the whole text of the 'acknowledgments' section in the paper so that other kind of acknowledgements can be also identified.

the reason for this lack of new indicators is that they still require exploring the level of presence, coverage and accuracy of this information recently incorporated in bibliometric databases.
The main objective of this study is to determine to what extent acknowledgements in scientific publications can be used as a proxy for external research funding. To do so we first analyze how accurately acknowledgements reflect the actual funding acquired by research units, then we identify which factors contribute to the presence of funding acknowledgements in scientific publications and finally we will introduce a normalized indicator of external funding attraction and also foresee what problems and challenges (both technical and conceptual) should be solved in the future in order to be able to produce valid, robust and reliable indicators considering this source.

Data & Methods

Relationship between funding acknowledgements and external research funding
In order to analyze the relationship between research income and funding acknowledgements we compare the publications by 13 Dutch universities in 2010 with the share of publications acknowledging any kind of funding. We consider a delay of one year between the income and the publication of results in scientific journals so that we analyse the presence of funding acknowledgement in the 25,988 publications of Dutch universities in 2011.
The actual funding that each university acquired in 2010 is expressed in terms of percentages according to the type of source[7]:

- Direct funding: share of funding received directly by each university from the Dutch government.
- Indirect funding: share of indirect governmental funding that universities receive from the Netherlands Organization for Scientific Research (NWO) and the Royal Netherlands Academy of Arts and Sciences (KNAW).
- Contract research: funding received from non-governmental institutions (e.g. private companies).

The type of funding which is acknowledged is either the indirect funding or the contract research (we also refer to these types of funding as third party funding), so that we expect to find some correlation between the share of papers acknowledging funding and the share of external funding acquired by each university.
The universities included in the analysis are: Delft University of Technology (TUD); Eindhoven University of Technology (TU/e); Erasmus University Rotterdam (EUR); Leiden University (LEI); Maastricht University (UM);

[7] Chiong Meza, C. (2012) Universities in the Netherlands: facts and figures 2012. Rathenau Instituut: Den Haag.

Radboud University Nijmegen (RU); The University of Groningen (RUG); The University of Amsterdam (UvA); The University of Twente (UT); Tilburg University (TiU); Utrecht University (UU); VU University Amsterdam (VU); Wageningen University and Research Centre (WUR). This analysis was restricted to these universities due to data availability on funding.

Factors influencing funding acknowledgments: regression analysis

All the articles, letters and reviews (all the remaining document types are excluded) published by Canada (N=174,357), Germany (N=280,000), Netherlands (N=99,054) and Spain (N=146,094) during the period 2009-2011 were extracted from the Thomson Reuters' Web of Science have been included in this analysis.

Given that we analyze the factors influencing the presence or absence of funding acknowledgements (binary dependent variable), we have performed a logistic regression analysis.

A paper has been considered to be funded if it contains a *Funding Text* (FT)[8]. As factor influencing the presence of funding acknowledgements, we consider several variables classified into four groups: scientific collaboration, scientific areas, level of the journal and other characteristics of the papers.

Scientific collaboration refers to joint research efforts that finally ended up with the publication of results in scientific journals. Combining the information on the number of institutes and countries involved in the publications we created three dummy variables: we consider that there is no collaboration when the paper has been published by a single institution (*Collab_none*), there is national collaboration when in the publication participate more than one institutions but all located in the same country (*Collab_nat.*) and we consider that the paper has been published in international collaboration when at least two different countries participate (*Collab_internat.*). On the other hand, we include two binary variables, one for USA (*Collab_USA*) and another for China (*Collab_China*), to analyze the extent to which the scientific collaboration with specific countries contributes to the publication of more papers including funding acknowledgements. The reason for including the variables for these two specific countries is that China is the country with the highest share of papers acknowledging funding (cf. Costas & van Leeuwen, 2012) in the database while USA is the main collaborator for the four countries included in the analysis.

The statistical analysis also controls for scientific area of the publications, to do so we classified all the papers into five research areas[9], creating a dummy variable for each field: Biomedical and health sciences (*Biom_hlth*); Life and earth sciences (*Life_Earth*); Mathematics and computer science (*Maths_Comp*);

[8] The Web of Science also include two other fields related to the funding acknowledgement: the Funding Body (GB) and the Grant Number (GN), but this information is always extracted from the Funding Text (FT)

[9] We assigned each Web of Science category to one field. The articles of the category 'multidisciplinary' were re-classified on the basis of their references.

Natural sciences and engineering (*Nat_Eng*); Social sciences and humanities (*SS_Hum*).
Field-normalized journal impact (MNJS) is the citation score of the journals in which researchers of each country publish in comparison to the international level in the field and we use this variable to measure the level of the journal in which the paper was published.
As other characteristics of the paper we consider the publication year and the document type. We created three dummy variables for the three years included in the analysis (*2009, 2010 and 2011*) and another three dummy variables for the document types (*article, letter and review)*.
We selected as reference variables: the field of Biomedical and health sciences (*Biom_hlth*), *Collab_none*, *2009* and *Reviews*.
The values of the correlation matrices and the variance inflation factors (VIF) are available upon request. VIF values indicate the absence of multicollinearity problem in the data as they as far lower than 10.

Indicator on external research funding based on funding acknowledgements
Based in the presence of funding acknowledgements in publications, what this indicator measures is if the external funding acknowledged by a given research group or unit is higher, lower or similar to the acknowledgment of research funding by similar research groups or units. By similar we mean groups that have the same chances of acquiring external funding (i.e. belong to the same country and to the same scientific field).
The indicator is calculated at the level of paper following a similar approach as the Mean Normalize Citation Score (MNCS) (Waltman et al, 2011). However, instead of citations this indicator is calculated on the basis of funding acknowledgements and calculated in the framework of a given country, so that we will refer to this indicator as the Mean Normalized Funding Acknowledgement (MNFA).
In order to normalize by field (within the same country), we divide the binary value of each publication (1 if it contains a funding acknowledgement/ 0 if there is no funding acknowledgement) by the expected value of funding acknowledgements for that particular publication. The expected value of funding acknowledgement is calculated as the average number of publications including funding acknowledgements in the field in which the publication was published. The normalization also takes into account the publication year and the document type.
A value of 1 for the MNFA would mean that the number of publications including funding acknowledgements equals the country average of funding acknowledgements. We will illustrate the indicator using the data on Dutch universities described above.

Results

Relationship between research income and funding acknowledgements

In order to compare to what extent the funding acknowledgements in the WoS reflect the actual external funding acquired by researchers we focus on 13 Dutch universities. Figure 1 shows the relationship between the share of external funding acquired by each university in 2010 (i.e. the sum of indirect governmental funding and contract research) and the share of papers published by these universities which include a funding acknowledgement.

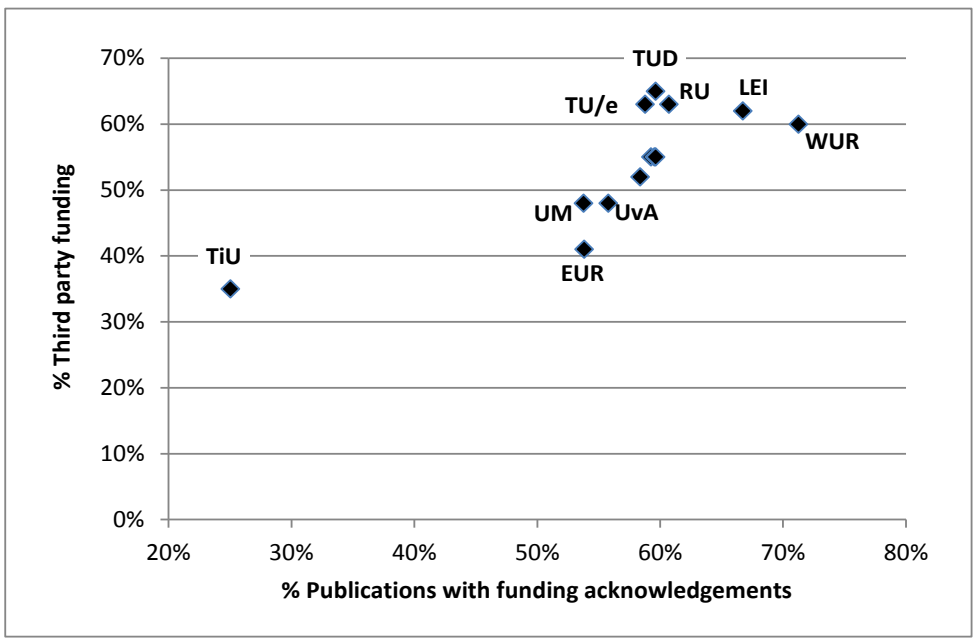


Figure 1. Relationship external funding and funding acknowledgements

These two variables show a positive correlation but the relationship between these variables is not perfect (r = 0.79). Universities with the highest share of external funding are not necessarily those with the highest share of publications acknowledging funding, like some technical universities. However, for other universities such as Maastricht University (UM), the University of Amsterdam (UvA) and especially the Erasmus University of Rotterdam (EUR) the share of publications acknowledging funding is higher than the share of external funding. These differences might be motivated by several factors, among others: propensity to publish results in scientific journals, scientific profile of universities or differences in acknowledgement patterns between disciplines.

Factors influencing the acknowledgement of funding

In this section we analyse which factors contribute to the presence of funding acknowledgements in scientific publications, focussing in the scientific outputs of four countries: Canada, Germany, Netherlands and Spain.

Table 1 reports the descriptive statistics according to the country of publication. For the four countries at least 50% of the papers published in the period 2009-2011 acknowledge funding. Most of the papers in these countries are produced in collaboration between two or more institutions, and usually at least one of them is a foreign institution. In this regard, the case of Netherlands is especially remarkable as publishes 52.8% of the papers in international collaboration.

USA is the main collaborator for the four countries included in the analysis; however the scientific collaboration with USA is especially high in the case of Canada (22.6%) maybe favoured by its geographic proximity. Also Canada shows the highest share of papers published in collaboration with China (4.7%).

Table 1. Descriptive statistics

	Canada	Germany	Netherlands	Spain
Funding report				
Funding	59.2	53.8	50	61.7
Non-funding	40.8	46.2	50	38.3
Collaboration				
Collab_none	36.1	32.5	27.8	35
Collab_nat.	18.2	18.5	19.4	22.9
Collab_internat.	45.7	49.0	52.8	42.1
Collab_USA	22.6	14.6	16	11.4
Collab_China	4.7	2.8	2.4	1.2
Scientific area				
Biomedical	45.3	43.3	53.3	37.7
Life	22.5	18.6	18.5	24.0
Math	10.6	7.6	7.2	11.2
Natural	22.5	34.7	20.1	32.0
Social	14.9	8.4	16.2	10.9
Publication year				
2009	32.5	32.1	31.4	31.2
2010	33.3	33.2	33.6	32.9
2011	34.3	34.7	35.0	35.8
Document type				
Article	91.2	91.9	89.1	91.0
Review	2.3	2.0	3.3	3.9
Letter	6.5	6.1	7.6	5.1
Level of the journal				
MNJS				
Min.	0.000	0.000	0.000	0.000
Max.	48.17	31.08	31.08	19.33
Mean	1.175	1.125	1.283	1.054
St. Dev.	1.118	1.156	1.195	0.969

According to the scientific areas, it is possible to see some similarities between countries as in all cases the highest shares belong to the area of Biomedical and health sciences (Netherlands presents the highest share in this area, 53.3%). The second most important area in terms of share of publications is Natural sciences and engineering (where Germany stands out with 34.7% of publications) followed by Life and earth sciences (Spain present the highest share, 24%). The lower shares are in the areas of Social sciences and humanities and Mathematics and computer science.

Table 2. Logistic regression results explaining the funding acknowledgements in scientific papers

	Canada	Germany	Netherlands	Spain
	β	*β*	*β*	*β*
Collaboration				
Collab_nat.[a]	0.313***	0.384***	0.296***	0.380***
Collab_internat.[a]	0.214***	0.944***	0.830***	0.514***
Collab_China	0.709***	0.633***	0.622***	0.382***
Collab_USA	0.188***	0.252***	0.270***	0.149***
Scientific area[c]				
Life	0.638***	0.625***	0.465***	0.866***
Math	-0.283***	-0.233***	-0.288***	0.627***
Natural	0.307***	0.456***	0.333***	0.726***
Social	-2.282***	-2.141***	-2.095***	-2.345***
Publication year[b]				
2010	0.360***	0.348***	0.343***	0.392***
2011	0.541***	0.498***	0.511***	0.562***
Document type[c]				
Article	0.519***	0.361***	0.548***	0.206***
Letter	-7.420***	-6.174***	-6.683***	-7.036***
Journal				
MNJS	0.357***	0.691***	0.382***	0.792***
Constant	-0.822	-1.838	-1.606	-1.298
N Obs.	174,357	280,000	99,054	146,094
Chi-square (d.f.)	42737.305***	72156.6***	25859.474***	48407.812***
Nagelkerke R^2	0.293	0.304	0.306	0.383
% Correct predictions	72%	71%	70.2%	77.1%

*, ** and *** indicate statistical significance at 10%, 5% and 1%

[a] The reference category is Collab_none (i.e. those papers published just by one institution); [b] The reference category is 2009; [c] The reference category is Review

The results of the logistic regression for each of the four countries included in this analysis are summarised Table 2, which includes the coefficient (β) and the level of statistical significance. All the variables included in the analyses are related to the fact that scientific publications acknowledge funding. Compared to those

papers published by a single institution, those published in collaboration at the national or international level are more likely to include funding acknowledgments. Also the collaboration with specific countries like USA or China contributes to the presence of funding acknowledgements.
The likelihood of including funding acknowledgements also depends on the scientific discipline. Papers in the areas of Life and earth sciences and Natural sciences and engineering are more likely to include funding than papers in the area of Biomedical and health sciences.
On the other hand, in the areas of Social Sciences and Humanities and Mathematics and computer science (except in Spain) are less likely to include funding acknowledgments compared to Biomedical and health sciences.

Papers published in 2010 and 2011 are also more likely to include funding acknowledgements than those published in 2009, which might suggest a relative increase over time of the type of funding usually acknowledged (i.e. third part funding). Regarding the document type, compared to reviews, articles are more likely to acknowledge funding while letters are less likely to include this kind of acknowledgements. The level of the journal in which a research article is published also contributes to increase the probabilities for a paper to include funding acknowledgement.

Indicator on external research funding
As described in the methodology, the indicator we propose (Mean Normalized Funding Acknowledgements-MNFA) is calculated considering the scientific outputs of a single country so that we compare researchers with the same possibilities of acquiring external research funding available in that country. To compare between different scientific areas we also normalize by field, dividing the binary value of each publication (1 if it contains a funding acknowledgement/ 0 if there is no funding acknowledgement) by the expected value of funding acknowledgements for that particular publication (average number of publications including funding acknowledgements in the field in which the publication was published). Table 3 shows an example of the calculation of the MNFA indicator for a group of 4 publications, all of them coming from the same country[10]:

In this example, the value 0.98 indicates that the research group is slightly below the average in its country in terms of articles published including funding acknowledgements.
We have calculated the values of the MNFA for the Dutch universities considering the articles and reviews published in 2009, 2010 and 2011 (Table 4). According to the MNFA, Tilburg University is the only university below the country average while University of Twente presents the highest value (although

[10] It is important to mention that like the MNCS, the MNFA also normalizes by document type and publication year.

is not the university with the highest share of publication with funding, being this Leiden University).

Table 3. Example of the calculation of the MNFA

Publication	FA	Field	Expected FA	NFA
A	1	X	0.45	2.22
B	1	Y	0.59	1.70
C	0	Y	0.59	0
D	0	Y	0.59	0
MNFA = (2.22+1.70+0+0)/4 = 0.98				

Table 4. MNFA for Dutch universities

University	P	FA	%FA	MNFA
Delft University of Technology	22,232	12,755	57.4%	1.09
Eindhoven University of Technology	16,775	10,146	60.5%	1.10
Erasmus University Rotterdam	104,271	66,683	64.0%	1.17
Leiden University	85,830	59,543	69.4%	1.15
Radboud University Nijmegen	105,564	75,093	71.1%	1.16
University of Amsterdam	114,784	74,420	64.8%	1.09
University of Groningen	84,987	53,315	62.7%	1.11
Maastricht University	55,077	32,148	58.4%	1.16
Tilburg University	8,874	2,292	25.8%	0.96
University of Twente	15,122	9,182	60.7%	1.20
Utrecht University	109,231	71,649	65.6%	1.13
VU University Amsterdam	86061	55246	1.19	1.19
Wageningen University and Research Centre	36351	25391	1.18	1.18

Discussion and conclusions

Indicators based on external funding attraction are expected to provide information on the degree of successfulness in acquiring external funding resources. In this study we focus on the funding acknowledgement information recently incorporated to the Web of Science in order to explore the possibility of using this information to create indicators on 'funding attraction'. Our contribution is threefold. First, we show the extent to which funding acknowledgements in scientific publications could reflect the actual external research funding of research institutions. Second, we determine which factors contribute to the presence of funding acknowledgements in scientific publications and finally we introduce a new normalized indicator which could be potentially used to measure, with some cautions, the level of external funding in research units.

We show that funding acknowledgements in publications reflect the relative importance of external funding of research units at least when the analysis is

performed at the meso-level (e.g. university level). However, some elements could contribute to weaken the relationship between external research funding and the presence of funding acknowledgements. For instance, when authors awarded with research grants do not acknowledge their funding or when contract research plays an important role in the external research funding, which might contribute to publish fewer publications and therefore lead to an underrepresentation in terms of funding acknowledgements. Indeed, research funded by private companies might be kept in secret or made public through other kind of document such as patents (Bolli & Somogyi, 2011; Wang & Shapira, 2011).

The regression analysis shows the distinct effect of the considered variables on the probabilities of including funding acknowledgements in scientific publications. Our results indicate that research collaboration is one of the elements which contribute to increase the probabilities of having funding acknowledgements for publications. This aspect prevents to claim that what we are measuring with the proposed indicator is exclusively related to the external research funding acquired by a research unit, given that it is not possible to verify if the funding was obtained by the research unit itself or by their collaborators. Thus, what the MNFA indicator measures is not only related to the direct acquisition of research funding by the units but also to the capacity of working with collaborators who themselves have funding (what at some degree could be also considered as a mild type of 'attraction').

The isolation of the funding brought by collaborators seems not to be possible at this moment according to the way in which authors acknowledge funding (not always each author indicate his/her funding acknowledgment individually) and how it is reflected in bibliographic databases (even if the information is available, there is not a direct link between the author and his/her individual funding acknowledgment).

Further research would be required to introduce other interesting aspects about external funding, as the identification of funding bodies to consider the 'prestige' of these bodies or the possibility of comparing researchers from different countries having different funding schemes and thus different possibilities of acquiring external funding for scientific research. However other important elements as the amount awarded would remain unknown as it is not usually included in the funding acknowledgements.

References

Bolli, T., & Somogyi, F. (2011) Do competitively acquired funds induce universities to increase productivity? Research Policy, 40(1), 136-147.

Chiong Meza, C. (2012) Universities in the Netherlands: facts and figures 2012. Rathenau Instituut: Den Haag.

Costas, R. & van Leeuwen, T. (2012) Approaching the "reward triangle": General analysis of the presence of funding acknowledgements and "peer review communication" scientific publications. *Journal of the American Society for Information Science and Technology*. 63(8), 1647-1661.

Costas, R. & van Leeuwen, T. (2012). New indicators based on the 'Funding acknowledgement' information in the Web of Science: analysis of the effect of peer review over the impact of scientific journals. 17th International Conference on Science and Technology Indicators. Quebec, Canada, September 5-8 2012. pp. 193-205

Garcia, C.E. & Sanz-Menendez, L. (2005) Competition for funding as an indicador of research competitiveness. *Scientometrics*, 64(3), 271-300

Gillett, R. (1991) Pitfalls in assessing research performance by grant income. *Scientometrics*, 22(2), 253-263

Hornbostel, S. (2001). Third party funding of German universities. An indicator of research activity? *Scientometrics*, 50, 523-537.

Laudel, G. (2005) Is external research funding a valid indicator for research performance?. *Research Evaluation*, 14(1), 27-34

Laudel, G. (2006) The 'quality myth': promoting and hindering conditions for acquiring research funds. Higher Education, 52, 375-403.

Rigby, J. (2011). Systematic grant and funding body acknowledgement data for publications: new dimensions and new controversies for research policy and evaluation. *Research Evaluation*, 20(5), 365-375

Schmoch, U., Schubert, T. (2009) Sustainability of incentives for excellent research – The German case. *Scientometrics,* 81(1), 195-218.

van der Besselaar, P., Inzeit, A; Reale, M. (2012) Measuring Internationalization of Funding Agencies. 17th International Conference on Science and Technology Indicators. Quebec, Canada, September 5-8 2012. pp. 121-130

Wang, J., Shapira, P., (2011) Funding acknowledgment analysis: an enhanced tool to investigate research sponsorship impacts: the case of nanotechnology. *Scientometrics,* 87(3), 563-586.

Waltman L., Eck N.J., Leeuwen T.N., Visser M.S., Raan A.F.J. (2011) Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*. 5(1), 37–47.

ANALYZING THE CITATION CHARACTERISTICS OF BOOKS: EDITED BOOKS, BOOK SERIES AND TYPES OF PUBLISHERS IN THE BOOK CITATION INDEX

Daniel Torres-Salinas[1], Nicolás Robinson-García[2], Álvaro Cabezas-Clavijo[3], Evaristo Jiménez-Contreras[2]

[1] *torressalinas@gmail.com*
EC3: Evaluación de la Ciencia y de la Comunicación Científica, Centro de Investigación Biomédica Aplicada, Universidad de Navarra (Spain)

[2] *{elrobin, evaristo}@ugr.es, acabezasclavijo@gmail.com*[3]
EC3: Evaluación de la Ciencia y de la Comunicación Científica, Departamento de Información y Documentación, Universidad de Granada (Spain)

Abstract
This paper presents a first approach to analyzing the factors that determine the citation characteristics of books. For this we use the Thomson Reuters' Book Citation Index, a novel multidisciplinary database launched in 2010 which offers bibliometric data of books. We analyze three possible factors which are considered to affect the citation impact of books: the presence of editors, the inclusion in series and the type of publisher. Also, we focus on highly cited books to see if these factors may affect them as well. We considered as highly cited books, those in the top 5% of the most highly cited ones of the database. We define these three aspects and we present the results for four major scientific areas in order to identify field-based differences (Science, Engineering & Technology, Social Sciences and Arts & Humanities). Finally we conclude observing that differences were noted for edited books and types of publishers. Although books included in series showed higher impact in two areas.

Conference Topic
Scientometrics Indicators - Relevance to Science and Technology, Social Sciences and Humanities (Topic 1), Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2).

1. Introduction
One of the basic outcomes from the field of bibliometrics and citation analysis is the characterization of document types and field-based impact which allow fair comparisons and a better understanding on the citation patterns of researchers (Bar-Ilan, 2008). These studies are of great relevance within the field as they put into context the impact of research as well as certain 'anomalies' such as, for instance, the higher impact of reviews over research papers (Archambault & Larivière, 2009), the impact of research collaboration (Lambiotte & Panzarasa,

2009) or the importance of monographs within the fields of the Humanities (Hicks, 2004). In this sense, the role played by citation indexes in general and the ones developed by Eugene Garfield and carried by Thomson Reuters in particular, have being of key importance for the development of such analyses (Garfield, 2009). However, these citations indexes are mainly devoted to scientific journals, neglecting other communication channels such as monographs. Hence and despite the many attempts made to analyze their impact (e.g., Torres-Salinas & Moed, 2009; White et al., 2009; Linmans, 2010; Kousha, Thelwall & Rezaie, 2011), little is known on the characterization of books' citation patterns.

Many studies have tried to characterize the citation patterns of books. However, these studies are normally based on small data sets based on specific disciplines. For instance, Cronin, Snyder & Atkins (1997) compared a list of top influential authors derived from journals citations with one derived from books in Sociology, concluding that these two types of publications reflect two complementary pieces of a fragmented picture. Tang (2008) takes a step further and deepens on the citation characteristics of a sample of 750 monographs in the fields of Religion, History, Psychology, Economics, Mathematics and Physics, finding significant differences when compared with the findings in the literature regarding citation in journal articles. Georgas & Cullars (2005) adopt a different approach and analyze the citation characteristics of the Linguistics literature in order to conclude if the habits of the researchers of this field are more closely related to the Social Sciences than to the Humanities. In general, the conclusions of these studies must always be taken with caution as they cannot be extended to all scientific fields.

But this scenario may change radically with the launch of the Thomson Reuters' Book Citation Index (henceforth BKCI) which provides large sets of bibliometric data regarding monographs. This database was launched in October 2010 as a greatly delayed answer to Eugene Garfield's request, who stated: 'Undoubtedly, the creation of a Book Citation Index is a major challenge for the future and would be an expected by-product of the new electronic media' (Garfield, 1996). At the time of its launch, it indexed 29618 books and 379082 book chapters covering a time period from 2005 to the present (Torres-Salinas et al., 2012). However, it now covers a time-span from 2003. According to Testa (2010), the BKCI follows a rigorous selection process in which the main criteria are the following: currency of the publications, complete bibliographic information for all cited references, English language is desirable and the implementation of a peer review process. To date, only two studies have been found analyzing the internal characteristics of the BKCI (Leydesdorff & Felt, 2012; Torres-Salinas et al., 2012). These types of seminal studies dissecting the coverage, caveats and limitations are considered of great regard as they serve to validate the accuracy and reliability of sources for bibliometric purposes.

In this context, we present an analysis of the citation characteristics of books relying on the data provided by the BKCI. Specifically, this study aims to analyze if the following factors may influence the citation patterns of the four main macro-areas of the scientific knowledge:

1) Edited books vs. Non-edited books. There is a perception that edited books usually have a greater impact than non edited books. To what extent is this true? Are there differences by field?
2) Series books vs. Non-series books. The prestige or impact derived from the collection in which the book is included is considered in certain areas as an evidence of the quality of books. Is there any empirical evidence on such claim?
3) Type of publishers. Is the publishers' prestige related with books impact? Which publishers have more impact, university presses, comercial publishers or other academic publishers?

2. Material and methods

This section is structured as follows. First we describe the data retrieval and processing procedures, indicating the normalization problems encountered and how these were solved. Also, we define the areas under study and how these were constructed, basing our methodology for this on previous studies and offering an overview of the distribution of books by areas in the BKCI. Then, in subsection 2.2, we define the variables analyzed and we describe the methodology followed as well as the statistical analysis undertaken in order to pursue the goals of the study.

2.1. Data retrieval and processing, and definition of areas

Records indexed as 'book chapters' and as a 'book' according to the Book Citation Index were downloaded in May 2012. We selected the 2005-2011 study period. The chosen time period is based on the availability of the data at the time of the retrieval. Then, data was included into a relational database created for this purposes. During data processing, publisher names were normalized as many had variants that differed as a function of the location of their head offices in each country. For instance, Springer uses variants such as Springer-Verlag Wien, Springer-Verlag Tokyo, Springer Publishing Co, among others. Next, we unified the citations received by books adding citations received by book chapters. The reason for doing so relies on the way the BKCI is designed, as it considers as separate citations received by a book and by a book chapter included in it. In this study we considered as citations to books, the sum of those received by their book chapters as well as those received by the books.

It is necessary to mention that a fixed citation window was included, which means that older books have a greater chance to get cited than the rest. Also, we must indicate that citations included in the BKCI come from all the citation indexes

provided by Thomson Reuters (SCI, SSCI and A&HCI) and not only the BKCI. Once the total citation of books was established we excluded Annual Reviews, which includes a total of 234 records as this publisher does not have books but journals, as indicated by Torres-Salinas et al. (2013). Hence the final books sample analyzed was of 28634 books.

In order to provide the reader with a general overview, we decided to cluster all subject categories of the BKCI (249) into four macro areas: Arts & Humanities (HUM), Science (SCI), Social Sciences (SOC) and Engineering & Technology (ENG). Aggregating subject categories is a classical perspective followed in many bibliometric studies when adopting a macro-level approach (Moed, 2005; Leydesdorff & Rafols, 2009). These aggregations are needed in order to provide the reader with an overview of the whole database. This way we minimized possibilities of overlapping for records assigned to more than one subject category. Also, we consider that such areas are easily identifiable by the reader as they establish an analogy with the other Thomson Reuters' citation indexes (Science Citation Index, Social Science Citation Index and Arts & Humanities Citation Index). With the exception of Sciences, which due to the heterogeneity of such a broad area, it was divided into two areas: Science and Engineering & Technology. In table 1 we show the distribution of the sample of books analyzed through the four disciplines.

Table 1. Distribution of books analyzed in this study by areas as well as total and average citations received according to the Book Citation Index. 2005-2011.

Discipline	Total Books	% Books from the BKCI	Total Citations	Average Citations
ENGINEERING & TECHNOLOGY	3871	14%	34705	8,97
ARTS & HUMANITIES	8251	29	52224	6,33
SCIENCE	9682	34%	241230	24,91
SOCIAL SCIENCE	10637	37%	99943	9,40
Total Books without duplicates	28634	100%	392429	13,70

2.3. Definition of variables and indicators

Now, we define and describe the three variables analyzed to characterize books' citations: presence of editors, inclusion of books in a series and type of publisher.

Presence of editors. In order to analyze edited and non-edited books we considered as the former those which were indexed as such according the Book Editor (ED) field provided by the BKCI. We considered non-edited books those which had no information in this field. For instance, the book entitled 'Power Laws in the Information Production Process: Lotkaian Informetrics' which is single-authored by L. Egghe has no information in the ED field, therefore it is

considered a non-edited book. On the contrary, the book 'Web 2.0 and Libraries: Impacts, Technologies and Trends' is edited by D. Parkes and G. Walton and has contributions from different authors, therefore it is considered and edited book.

Inclusion in a series or collection. In order to analyze the inclusion of books in a series or a collection we used the field defined in the BKCI as Series (SE), tagging as such those records which contained information in this field and as non series, those which did not. We identified a total of 3374 different series in the BKCI. The series with a higher number of books indexed in the BKCI for each field are: 'Studies in Computational Intelligence' published by Springer (243 books) for Engineering & Technology, 'New Middle Ages' by Palgrage (49 books) in Arts & Humanities; 'Methods in Molecular Biology' by Humana Press Inc (232 books) in Science, and 'Chandos Information Professional Series' by Chandos (118) in Social Sciences.

Type of publisher. In order to define the type of publisher, first we normalized them according to the name variants described above. As a result of such normalization process, 280 publishers were identified in the BKCI. Then, these publishers were distributed across the three following categories:

- University Press. Defined as any publisher belonging to a University such as the Imperial College Press or Duke University Press.

- Non-University Academic Publisher. Publishers belonging or related to organizations such as research institutions, scientific societies or any other type of entity not linked to universities such as the Royal Society of Chemistry or the Technical Research Centre Finland.

- Commercial Publisher. Publishers considered in this group are those not related to universities or any other scientific entity but to firms with profit motive such as Routledge or Elsevier.

Finally, we characterized the factors that determine books' citations using different statistical descriptive indicators. The statistical analysis of data was carried out with SPSS v 20.0.0. As patterns of citations were not normally distributed, non-parametric tests were also used to derive levels of statistical significance. These tests were applied for the comparison of means (Mann–Whitney and Kruskal-Wallis tests) between the different factors analyzed at a 0.05 significance level. Furthermore, we analyzed the characteristics of the Highly Cited Books (henceforth HBC), that is, the 5% share most highly cited according to these three variables. 1534 books were identified as HBC.

3. Results

In this section we offer the results of our analysis on the impact of books in the BKCI depending on according to three variables: presence of editors, inclusion in series and type of publisher. This section is structured accordingly to these variables.

3.1 Edited vs. Non-edited books

In Table 2 we offer an overview of the sample of books analyzed according to the presence of editors. At large, from the total sample (ALL), 12646 books (44%) have been edited while 15988 books (56%) are not. Edited books have a significantly higher citation rate than those which are non-edited, as shown by the average and the median values. This occurs in the four areas studied. The most significant differences are found in the field of Science where edited books have an average of 35.51 citations per book in opposition to 10.16 citations per book for non-edited books. Also, edited books reach higher citation values as indicated by the standard deviation and median values. To a lesser extent, this situation also occurs in the Social Science and Engineering & Technology fields. The lowest differences between edited and non-edited books are found in the field of Arts & Humanities, where edited books have a citation average of 7.61, while non-edited books have an average of 5.81. Differences in citations were statistically significant in all disciplines (CI=95%, p<0, 05) with the median values of edited books much higher than the ones for non-edited books.

Table 2. Citation and statistical indicators, and percentage of Highly Cited Books. Edited vs. Non-edited books. 2005-2011

Discipline		Nr of Books	% of Books	% HCB	Citation Average	Standard Desv.	Median
ALL **	Edited Books	12646	44%	65%	21.81	± 99.35	5.00
	Non Edited Books	15988	56%	35%	7.16	± 7.61	2.00
ENG **	Edited Books	1841	48%	66%	12.00	± 24.59	4.00
	Non Edited Books	2030	52%	34%	6.21	± 15.82	1.00
HUM **	Edited Books	2384	29%	42%	7.61	± 15.26	3.00
	Non Edited Books	5867	71%	58%	5.81	± 14.45	2.00
SCI **	Edited Books	5658	58%	90%	35.41	± 145.45	7.00
	Non Edited Books	4024	42%	10%	10.16	± 35.96	2.00
SOC **	Edited Books	4254	40%	57%	12.0	± 29.24	4.00
	Non Edited Books	6383	60%	43%	7.66	± 24.35	2.00

** Non Parametric Test for comparing means: Mann-Whitney: CI=95%; P<0,05

Regarding the 1534 HBC, 65% of them were edited while 35% were non-edited. This general pattern also takes place in three of the areas analyzed, especially in the field of Science with 90% of the HBC being edited books, followed by Engineering & Technology (65%) and Social Sciences (57%). The only exception is found in Arts & Humanities where the percentage of edited HCB is lower than the one for non-edited with a 42% of the total share of HBC. However, when interpreting the data for HCB presented in Table 2, one must read it taking into account their total share. For instance, in Engineering & Technology there is a higher share of edited books. In the case of the Arts & Humanities, edited books represent only 28% of the total share, however, the share for HCB edited is of 42% of the total HCB in this field. This means that HCB are more commonly edited than non-edited books.

3.2 Inclusion in series vs. non-inclusion in series

There are a total of 17789 books included in a series (62% of the total share) while books not included in series are 10845 (38%) (Table 3). The distribution of books in series varies according to the area. Science and Engineering & Technology are the fields with the highest shares, especially the latter where books in series represent 71%. In regard with the citation average and median values of books included in series, also these two areas and especially Science are the ones which show the most significant differences. On the contrary, there are no significant differences in the Social Sciences, and the median values for both; included and non-included books, are 3.00. The only exception noted is Arts & Humanities, where non-included in series books have a higher citation average and median values than those included in series. Differences in citations were statistically significant in Science and Engineering & Technology (CI=95%, $p<0,05$) with the median values of books included in series much higher than the ones of those not included. In the Social Sciences there are no differences (CI=95%, $p >0,05$) and in Humanities differences in citations were statistically significant (CI=95%, $p<0,05$) for books not included in series.

If we focus on HBC included in series, we observe that 65% of the overall most cited books belonged to a series or collection. Also, books included in series are better represented in the area of Engineering & Technology, while in Science the proportion is of 50%. In Social Sciences and Arts & Humanities, the HCB with greater presence are the ones not included in series. In the four areas analyzed there are no significant differences on the overall distribution of books included and non-included in series as well as on the distribution of HCB. Therefore, while 71% of the books in Engineering & Technology are included in series, 67% of the HCB of this area are also included in series, following a similar distribution.

Table 3. Citation and statistical indicators, and percentage of Highly Cited Books. Included in series vs. Non-included in series books. 2005-2011

Discipline		Nr of Books	% of Books	% HCB	Citation Average	Standard Desv.	Median
ALL **	Series Books	17789	62%	65%	15.62	± 45.68	
	Non Series Books	10845	38%	35%	10.98	± 95.38	
ENG **	Series Books	2746	71%	67%	10.06	± 28.25	3.00
	Non Series Books	1125	29%	33%	7.62	± 23.50	2.00
HUM **	Series Books	4585	56%	46%	5.91	± 14.43	2.00
	Non Series Books	3666	44%	54%	6.86	± 15.04	3.00
SCI **	Series Books	6349	66%	51%	29.63	± 69.19	5.00
	Non Series Books	3333	34%	49%	15.93	± 169.37	2.00
SOC	Series Books	5854	55%	43%	9.1	± 27.10	3.00
	Non Series Books	4783	45%	57%	9.75	± 25.76	3.00

** Non Parametric Test for comparing means: Mann-Whitney: CI=95%; P<0,05

3.3 Type of publisher

Overall, 83% of the books included in the BKCI belong to commercial publishers, followed by far by university presses (14%) and non-university academic publishers (3%) (Table 4). This distribution varies substantially depending on the area. In Engineering & Technology the presence of commercial publishers is even higher (97%), while in the Arts & Humanities and the Social Sciences, the university presses have a higher presence (27% and 15% respectively).

When analyzing the citation average and median values in general, we observe that there is a common pattern for all areas: the commercial publishers are the ones with the lowest citation averages and the university presses are the ones with the highest figures. The highest difference is noted in the Arts & Humanities, where the latter show a citation average of 11.62 while the former have values of 4.32. The same occurs with the Social Sciences, where university presses have a citation average of 20.40 on opposition to commercial publishers, with 7.33. These differences are also significant for Engineering & Technology and Science, although not to the same extent. Differences in citations were statistically significant in all disciplines (CI=95%, p<0, 05) with the median values for university presses much higher than for the other two types, commercial and academic publishers.

Table 4. Citation and statistical indicators, and percentage of Highly Cited Books. Included in series vs. Non-included in series books. 2005-2011

Discipline	Type	Nr of Books	% of Books	% HCB	Citation Average	Standar Desv.	Median
ALL **	Academic Non Univ	919	3%	4%	23.90	53.40	5.00
	Comercial Publisher	23843	83%	66%	12.36	39.67	2.00
	University Press	3872	14%	30%	20.22	156.60	7.00
ENG **	Academic Non Univ	72	2%	4%	16.22	± 28.45	5.00
	Comercial Publisher	3726	97%	93%	8.62	± 18.90	2.00
	University Press	57	1%	3%	23.96	± 65.73	7.00
HUM **	Academic Non Univ	51	1%	1%	5.33	± 9.55	2.00
	Comercial Publisher	5906	71%	38%	4.32	± 9.86	2.00
	University Press	2270	28%	61%	11.62	± 22.20	6.00
SCI **	Academic Non Univ	696	7%	9%	28.26	± 59.75	6.00
	Comercial Publisher	8375	87%	85%	22.81	± 61.23	3.00
	University Press	517	6%	6%	50.88	± 421.40	9.00
SOC **	Academic Non Univ	181	2%	2%	10.71	± 24.24	3.00
	Comercial Publisher	8816	83%	58%	7.33	± 21.13	2.00
	University Press	1626	15%	40%	20.40	± 44.17	8.00

** Non Parametric Test for comparing means: Kruskal-Wallis: CI=95%; P<0,05

Regarding the distribution of HCB according to the types of publisher, the most significant event is the high representation of HCB among books from university presses, almost always higher than the other two types of publishers for all areas. For instance, in general (ALL) university presses represent 14% of the total share. However, when focusing on HCB, they represent 30%. This phenomenon is especially relevant in two of the four areas under study. Thus, in Arts & Humanities 28% of the total share are published by university presses, but 61% of the HCB are from this type of publisher. The same happens in the Social Sciences, where they represent 15% of the total share but have 40% of the total HCB.

4. Discussion and concluding remarks

This paper analyzes the citation characteristics of books according to three variables: the presence of editors, their inclusion in series and the type of publisher. For this, we used a sample of 28634 books indexed in the Book Citation Index for four macro-areas during the 2005-2011 time period. We must indicate that the Book Citation Index is a novel database by Thomson Reuters

which opens new possibilities for analyzing the citation phenomenon in books as it happens with journals, where such characteristics have been already thoroughly analyzed (see e.g., Peritz, 1981; Aksnes, 2003). This study is therefore, one of the first ones analyzing the factors that determine citations of books using such a large dataset. In Table 5 we show the main findings of this study.

Table 5. Highlights of the main findings of this study analyzing the factors which determine the citation of books in four major gareas. Data: Book Citation Index. 2005-2011

	ENG	HUM	SCI	SOC
CHARACTERISTICS OF THE BKCI COVERAGE				
Edited Vs Non Edited	Edited and non-edited books are equally distributed	There are more non-edited books than edited (71%)	Edited and non-edited books are not equally distributed	There are more non-edited books than edited (60%)
Series Vs Non Series	Most books are included in series (71%)	Books included and not included in series are equally distributed	Most books are included in series (66%)	Books included and not included in series are not equally distributed
Type of Publisher	Most books are from commercial publishers (97%)	Most books are from commercial publishers (71%)	Most books are from commercial publishers (87%)	Most books are from commercial publishers (83%)
CITATION CHARACTERISTICS OF BOOKS INCLUDED IN THE BKCI				
Edited Vs Non Edited	Edited books are more cited	Edited books are more cited	Edited books are more cited	Edited books are more cited
Series Vs Non Series	Books included in series are more cited	Books not included in series are more cited	Books included in series are more cited	There are no citation differences
Type of Publisher	Books from university presses are more cited	Books from university presses are more cited	Books from university presses are more cited	Books from university presses are more cited

These are the following:

1) There are more non-edited books in the Arts & Humanities and Social Sciences fields which reflect that single-authored books are more frequent and therefore, could be better considered rather than collective works in which various authors contribute in individual chapters in a more similar fashion to that of journal publications. However, in all areas edited books have a greater impact than non-edited books. This may be due to the effects of collaboration with a more diversified content and therefore, more probabilities of being cited.

2) Regarding their inclusion in series, these are more frequent in the fields of Engineering & Technology as well as Science, while in the Arts & Humanities and Social Sciences, the distribution of books is more homogeneous. However, the impact of books according to their inclusion in series varies depending on the field. Therefore, in Engineering & Technology and Science, books included in series are more cited than those which are not, although not in such as significant way as in other cases. In the Arts & Humanities the books not included in series are the ones with higher impact, but these differences are not determinant. In the case of Social Sciences, there are almost no differences regarding this variable.

3) Considering the type of publisher, most of the books indexed in the BKCI belong to commercial publishers, especially in the fields of Engineering & Technology and Science. Though the distribution is similar in the Arts & Humanities as well as the Social Sciences, the university presses are better represented in these cases. However, there is a common phenomenon which occurs across all areas: books published by university presses have a significantly higher impact than the rest. At this point we must take this statement with caution, as after supervising these publishers, we find out that these university presses included in the BKCI are considered of huge prestige such as Cambridge UP, Princeton UP and University of California P. That is, the high impact books published in university presses have may rely on a better selection of books and topics on which to publish than that followed by commercial publishers such as Elsevier, Routledge or Palgrave, for instance.

Finally, we must point out that the results offered in this analysis inherit the shortcomings of the database from which the data was retrieved. The BKCI is an on-going project which still show significant limitations. Some of these may affect the results presented such as a bias towards English language publications (96% of its books are written in this language and 75% of the publishers come from the United Kingdom or the United States) and a great concentration of publishers. For example, Springer, Palgrave and Routledge accumulate by themselves 50% of the total database. Therefore these findings must be read taking into consideration such issues. However, the large data set used may be a significant step towards a better comprehension of the citation characteristics of books.

Acknowledgments

Nicolás Robinson-García is currently supported by a FPU grant from the Spanish Ministerio de Economía y Competitividad of the Spanish government.

References

Aksnes, D.W: (2003). Characteristics of highly cited papers. *Research Evaluation*, 12(3), 159-170.

Archambault, E. & Larivière, V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics*, 79(3), 639-653.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century — A review. *Journal of Informetrics*, 2(1), 1-52.

Cronin, B., Snyder, H. & Atkins, H. (1997). Comparative citation rankings of authors in monographic and journal literature: a study of sociology. *Journal of Documentation*, 53(3), 263-273.

Georgas, H. & Cullars, J. (2005). A citation study of the characteristics of the Linguistics Literature. *College & Research Libraries*, 66(6), 496-516.

Garfield, E. (1996). Citation indexes for retrieval and research evaluation. *Consensus Conference on the Theory and Practice of Research Assessment.* 7 October. Available at http://www.garfield.library.upenn.edu/papers/ciretreseval-capri.html

Garfield, E. (2009). Five decades of citation indexing. *International Workshop for Scientometrics.* Beijing, China. Available at http://garfield.library.upenn.edu/papers/beijingchina2009.html

Hicks, D. (2004). The four literatures of social science. In: Moed, H.F., Glänzel, W. and Schmoch, U. (Eds.). *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems.* Kluwer Academic Publishers, Netherlands, pp. 473-496.

Kousha, K., Thelwall, M. & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, 62(11), 2147-2164.

Lambiotte, R. & Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, 3(3), 180-190.

Leydesdorff, L. & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, 60(2), 348-362.

Leydesdorff, L. & Felt, U. (2012). Edited volumes, monographs and book chapters in the Book Citation Index. *Journal of Scientometric Research*, 1(1), 28-34.

Linmans, A.J.M. (2010).Why with bibliometrics the Humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library findings and productivity measures. *Scientometrics*, 83(2), 337-354.

Moed, H.F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Tang, R. (2008). Citation characteristics and intellectual acceptance of scholarly monographs. *College & Research Libraries*, 69(4), 356-369.

Peritz, B.C. (1981). Citation characteristics in Library Science: Some further results from a bibliometric survey. *Library Research*, 3(1), 47-65

Testa, J. (2010). The book selection process for the Book Citation Index in Web of Science. Available at http://wokinfo.com/media/pdf/BKCI-SelectionEssay_web.pdf

Torres-Salinas, D. & Moed, H.F. (2009). Library Catalog Analysis as a tool in studies of social sciences and humanities: An exploratory study on published book titles in Economics. *Journal of Informetrics*, 3(1), 9-26.

Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E. & Delgado López-Cózar, E. (2012). Towards a 'Book Publishers Citation Reports'. First approach using the 'Book Citation Index'. *Revista Española de Documentación Científica*, 35(4), 615-620.

Torres-Salinas, D., Rodríguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J. & García, J.A. (2013). Mapping citation patterns of book chapters in the Book Citation Index. *Journal of Informetrics*, in press.

White, H., Boell, S.K., Yu, H. Davis, M., Wilson, C.S. & Cole, F.T.H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, 60(6), 1083-1096.

THE APPLICATION OF CITATION-BASED PERFORMANCE CLASSES TO THE DISCIPLINARY AND MULTIDISCIPLINARY ASSESSMENT IN NATIONAL COMPARISON

Wolfgang Glänzel[1,2]

[1] *Wolfgang.Glanzel@kuleuven.be*
Centre for R&D Monitoring and Dept. MSI, KU Leuven (Belgium)

[2] *glanzw@helka.iif.hu*
Dept. Science Policy & Scientometrics, Library of the Hungarian Academy of Sciences, Budapest (Hungary)

Abstract
The analysis of the high end of citation distributions represented by its tail provides important supplementary information on the citation profile of the unit under study. In a previous study by Glänzel (2012), a parameter-free solution providing four performance classes has been proposed. Unlike in methods based on pre-set percentiles, this method is not sensitive to ties and ensures needless integration of measures of outstanding and even extreme performance into the standard tools of scientometric performance assessment. The applicability of this method is demonstrated for both subject analysis at the large scale and the combination of different subjects.

Conference Topic
Scientometrics Indicators: Criticism and new developments (Topic 1) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction
One of the objectives of a previous study (Glänzel, 2012) was to analyse to what extent the tail of scientometric distributions are in line with the 'head' and 'trunk' forming the major part of the distribution and if in how far 'outliers' might be responsible for possible deviations. Two important observations are relevant in this context. Unlike in many other fields, where outliers can simply be discarded as being exceptions, in bibliometrics extreme values represent the high end of research performance and therefore deserve special attention. The second observation refers to empirical evidence concerning specific tail properties of citation distributions. Glänzel & Schubert (1988a) have shown that the often extremely long tail cannot be explained by the underlying distribution model. While extreme performance in publication activity was in keeping with the parameters estimated on the basis of the underlying distribution model, in the case of citation impact, the tail proved to be distinctly heavier than estimated on the

basis of the head and trunk of the empirical distribution, which, in turn, usually represents 95% (or even more) of all observations. This effect was observed even if a Paretian distribution model was assumed. This property could be confirmed in the above-mentioned study by Glänzel (2012). One solution proposed in the study was to use tail indices as a supplement to traditional citation-based performance indicators, such as the share of uncited papers and the mean citation rate. The analysis of the tail, which was based on ordered or ranked observations, can practically be uncoupled from the overwhelming rest of the empirical distribution. Most studies of the tail of scientometric distributions proceed from a Pareto model. The estimation of the tail parameter can directly be obtained from subsets of order statistics and are mostly based on the Renyi's representation (Rényi, 1953). Versions of Hill's estimator (Hill, 1975) and estimators based on so-called quantile-quantile plots (Kratz & Resnik, 1995; Beirlant et al., 2004) are the most commonly used statistics. It has been shown that these estimators are consistent and asymptotically normally distributed. This property allows to construct confidence intervals for tail parameters. The practicability of quantile plotting in scientometrics and using the Pareto tail parameter for the assessment of individual research performance has been proposed, for instance, by Beirlant et al. (2007). Nevertheless, the estimation of the tail index remains rather problematic since most methods are still sensitive to the cut-off point for the tail. Since already minute changes of the tail parameter might have consequences in an evaluative context, the recommendation in the study by Glänzel (2012) was to favour a parameter-free solution for the assessment of outstanding performance. This might also help avoid parameter conflicts resulting from estimating parameters on the basis of head and trunk of the distributions, on one hand, and from their tail, on the other hand. Therefore, a "reduction" of the original citation distribution to performance classes on the basis of *Characteristic Scores and Scales* (CSS) introduced by Glänzel & Schubert (1988b) was proposed as an alternative parameter-free solution. Taking into account that citation standards considerably differ in the various disciplines, the method was developed for *individual subjects*. The classes obtained from this method can be applied to the comparative analysis of the citation-impact profiles of given units amongst themselves as well as with the reference standard in the given subject. It has been stressed that the calculation of a "single" indicator over these classes is not suitable as this would reduce the gained added value and thus destroy the advantages of the method. However, it has also been mentioned that the application to combinations of different disciplines might indeed be possible. Besides the demonstration of the applicability of the proposed method to individual subjects on a large scale, its application to the combination of different subjects will be the main objective of the present study.

A parameter-free solution using Characteristic Scores and Scales (CSS)

An alternative to the tail analysis supplementing standard indicators is the "reduction" of the original citation distribution to a distribution over some

essential performance classes including one or more particular classes corresponding to the high end of performance, i.e., to the tail of the original distribution. A solution using six classes has already been suggested by Leydesdorff et al. (2011). According to their model, a pre-set set of six rank percentages is calculated on the basis of the reference distribution. Individual observations are then scored according to the percentage the publications in question belong to. Two particular problems arise from this approach, namely the arbitrariness of pre-set percentiles and the ties in both the reference distribution and the observations.

Another solution has recently been suggested by Adams et al. (2007). The proposed classification proceeds from the mean citation rate on the basis of the world standard. The lowest class is formed by uncited papers. Other performance classes are then formed by setting thresholds at one quarter and one half of the standard for the lower performance classes and the double and quadruple of the standard for the higher classes, respectively. This procedure can be continued by extending the geometrics series based on positive and negative powers of 2. This method avoids the problem of ties but still uses preset threshold. In what follows, a self-adjusting method will be presented. The thresholds subdividing the population and samples into different performance classes are produced by the method itself and only depend on the underlying citation distribution. The sole arbitrarily chosen value is then the number of performance classes.

Characteristic Scores and Scales (CSS)

A self-adjusting solution can be based on the method of Characteristic Scores and Scales (CSS) proposed by Glänzel & Schubert (1988b). Characteristic scores are obtained from iteratively truncating a distribution according to conditional mean values from the low end up to the high end. In particular, the scores b_k ($k > 0$) are obtained from iteratively truncating samples at their mean value and recalculating the mean of the truncated sample until the procedure is stopped or no new scores are obtained. Instead of the verbal description given here, an exact mathematical description can be found, e.g., in the study by Glänzel & Schubert (1988b).

First put $b_0 = 0$. b_1 is then defined as the mean of the original sample. The procedure is usually stopped at $k = 3$ since the number of papers remaining in the subsequent truncated sample might otherwise become too small. The k-th class is defined by the pair of threshold values $[b_{k-1}, b_k)$ with $k > 0$. The last and highest class is defined by the interval $[b_k, \infty)$, with usually $k = 3$. The number of papers belonging to any class is obtained from those papers, the citation rate of which falls into the corresponding half-open interval. This definition solves the problem of ties since all papers can uniquely be assigned to one single class. In earlier studies the resulting four classes were called *poorly* cited (if less cited than average), *fairly* (if cited above average but received less citations than b_2), *remarkably* cited (if received at least b_2 but less than b_3 citations) and *outstandingly* cited (if more frequently cited than b_3). In the present study 'Class k' (k =1, 2, 3, 4) is used instead for the sake of convenience. The robustness of

scales and classes has already been analysed and reported, for instance, by Glänzel in 2007. In addition, one important property should be pointed out here, particularly,

$$b_k / b_1 \approx \sum_{i=0}^{k-1} \left(\frac{\alpha}{\alpha - 1} \right)^i ,$$

provided the underlying distribution is of Pareto-type and α is its tail parameter. According to this property, the ratios of the *k*-th and the first score form a geometrics series. As all location parameters, characteristic scores, too, are very sensitive to the subject field and the citation window. b_1 is, by definition, the mean value of the empirical citation distribution; all other scores are conditional means that depend on this initial value. This property is also reflected by the above approximate formula. Therefore, characteristic scores should not be used for comparison across subject areas.

Another property refers to the distribution of papers over the classes. The studies by Glänzel (2007; 2012) give empirical evidence that, in contrast to the b_k scores, this distribution over classes is strikingly stable with respect to the underlying subject field, the publication year as well as the citation window. This property makes the method useful for longitudinal and multi-disciplinary studies. Classes 1 and 2 represent "head" and "trunk" of the underlying citation distribution over individual papers. Usually, this refers to 90% or a slightly larger share of all papers. The upper two classes, representing nearly 10% of all papers, stand for the highly cited part of publications. Class 4, finally, covers the top 2%–3% of the corresponding population or sample and forms the most interesting category. It also contains possible outliers that have, however, no further effect on the outcomes as merely their assignment to the class but not their actual value counts. The following subsection will provide an introduction into the application of the method.

Application of Characteristic Scores and Scales in comparative studies

After these introductory methodological remarks, the assessment of the citation impact according to performance classes will be explained in detail. This will be done in two steps. In the first step, the application to topics and disciplines is explained, thereafter the application to a combination of disciplines or even to all fields combined will be described. In the latter case a special procedure is necessary since simply forming four classes on the basis of the citation distribution in all fields combined would bias the results in favour of the life-sciences and to the detriment of mathematics and engineering sciences.

Disciplinary analysis

For the disciplinary analysis, first a brief summary of the procedure described in the already mentioned study (Glänzel, 2012) is given. Again, preferably four

classes should be used. First the b_k (k = 1, 2, 3) thresholds are calculated from the world total in the discipline or topic under study. These scores are used to define the reference standard, which is based on the four classes $[b_{k-1}, b_k)$, k = 1, 2, 3 and $[b_3, \infty)$. For the demonstration 20 out of the 60 subfields in the sciences according to the Leuven-Budapest classification scheme (see Glänzel & Schubert, 2003) have been selected. Furthermore, two publication years have been chosen, 2007 with a five-year citation window (2007–2011) and 2009 with the three-year citation window 2009–2011. All journal publications indexed as article, letter, proceedings paper or review in the 2007 and 2009 volumes of Thomson Reuters' Web of Science (WoS) have been selected and processed.

Table 1. Characteristic scores of publications in 2007 and 2009 for 20 selected subfields according to the Leuven-Budapest scheme [Data sourced from Thomson Reuters Web of Knowledge]

*Subfield**	*2007 (5-year citation window)*			*2009 (3-year citation window)*		
	b_1	b_2	b_3	b_1	b_2	b_3
A2	6.43	13.80	21.97	2.68	6.01	10.68
B1	16.75	39.24	79.61	8.21	19.96	38.24
B2	23.05	58.33	116.72	11.34	28.96	56.28
C1	9.37	22.04	40.48	5.13	12.37	21.68
C3	11.22	24.68	42.04	5.84	12.24	20.83
C6	8.21	23.67	51.24	4.56	12.71	26.50
E1	5.04	14.75	29.83	2.37	6.64	12.60
E2	4.71	11.90	21.97	2.27	6.15	11.54
E3	6.57	17.82	34.00	4.19	11.19	21.10
G1	15.55	38.35	74.51	8.75	20.82	39.17
H1	5.21	14.36	29.83	2.41	6.66	12.88
I1	13.52	34.87	69.24	6.01	15.92	29.58
I5	16.24	41.52	84.74	7.96	19.26	39.49
M6	11.50	28.31	51.81	5.27	13.51	24.88
N1	15.28	35.38	64.73	7.18	16.92	29.77
P4	7.25	17.71	32.75	3.09	8.12	15.13
P6	7.27	20.05	43.89	4.30	12.15	26.54
R2	10.60	23.99	42.54	4.82	10.64	18.37
R4	11.42	26.19	48.62	5.49	12.65	22.50
Z3	12.80	29.48	54.96	6.36	15.25	28.88

* *Legend*: A2: plant & soil science & technology; B1: biochemistry/biophysics/molecular biology; B2: cell biology; C1: analytical, inorganic & nuclear chemistry; C3: organic & medicinal chemistry; C6: materials science; E1: computer science/information technology; E2: electrical & electronic engineering; E3: energy & fuels; G1: astronomy & astrophysics; H1: applied mathematics; I1: cardiovascular & respiratory medicine; I5: immunology; M6: psychiatry & neurology; N1: neurosciences & psychopharmacology; P4: mathematical & theoretical physics; P6: physics of solids; R2: biomaterials & bioengineering; R4: pharmacology & toxicology; Z3: microbiology

As expected, both subject and citation window have a strong effect on the actual values of the characteristic scores b_k. The lowest value has been found in A2 (plant & soil science & technology) in 2009 on the basis of a 3-year citations windows, while the highest one was observed in B2 (cell biology) in 2007 with a 5-year citation window. Increasing the citation window changed all b_k values: For the used combination of publication year and citation window, this resulted in roughly doubling the corresponding values with respect to the shorter window. The b_k values for the two WoS volumes are presented in Table 1.

Table 2. CSS-class shares of publications in 2007 and 2009 for 20 selected subfields according to the Leuven-Budapest scheme [Data sourced from Thomson Reuters Web of Knowledge]

*Subfield**	*2007 (5-year citation window)*				*2009 (3-year citation window)*			
	Class 1	*Class 2*	*Class 3*	*Class 4*	*Class 1*	*Class 2*	*Class 3*	*Class 4*
A2	65.2%	22.6%	8.1%	4.2%	63.3%	26.0%	7.1%	3.6%
B1	69.4%	22.5%	6.0%	2.1%	70.6%	21.0%	6.3%	2.2%
B2	72.0%	20.2%	5.6%	2.2%	71.6%	20.1%	5.8%	2.4%
C1	68.2%	22.5%	6.6%	2.7%	69.2%	21.3%	6.4%	3.0%
C3	67.4%	22.2%	7.5%	3.0%	63.6%	24.9%	7.7%	3.9%
C6	73.5%	19.5%	5.3%	1.8%	71.6%	20.5%	5.8%	2.1%
E1	73.7%	18.8%	5.5%	2.0%	71.4%	19.9%	6.2%	2.4%
E2	68.2%	21.7%	7.0%	3.1%	70.8%	20.9%	5.7%	2.5%
E3	70.7%	20.2%	6.3%	2.9%	70.9%	20.6%	6.1%	2.4%
G1	70.1%	21.4%	6.3%	2.2%	68.1%	22.4%	7.2%	2.4%
H1	72.3%	20.3%	5.4%	1.9%	71.0%	20.4%	6.2%	2.4%
I1	70.2%	21.3%	6.2%	2.3%	71.2%	20.0%	6.1%	2.7%
I5	71.9%	20.4%	5.4%	2.2%	68.7%	22.8%	6.1%	2.3%
M6	68.9%	21.6%	6.5%	3.0%	69.9%	20.9%	6.3%	2.9%
N1	69.1%	21.7%	6.4%	2.8%	69.1%	21.1%	6.8%	3.0%
P4	69.6%	21.2%	6.7%	2.4%	71.2%	20.8%	5.7%	2.3%
P6	72.4%	20.7%	5.3%	1.7%	72.8%	20.4%	5.2%	1.6%
R2	72.4%	20.7%	5.3%	1.7%	64.7%	23.7%	7.8%	3.8%
R4	68.4%	22.5%	6.4%	2.7%	67.3%	22.5%	7.1%	3.0%
Z3	68.2%	22.3%	6.8%	2.6%	69.3%	22.1%	6.2%	2.5%

* *Legend*: A2: plant & soil science & technology; B1: biochemistry/biophysics/molecular biology; B2: cell biology; C1: analytical, inorganic & nuclear chemistry; C3: organic & medicinal chemistry; C6: materials science; E1: computer science/information technology; E2: electrical & electronic engineering; E3: energy & fuels; G1: astronomy & astrophysics; H1: applied mathematics; I1: cardiovascular & respiratory medicine; I5: immunology; M6: psychiatry & neurology; N1: neurosciences & psychopharmacology; P4: mathematical & theoretical physics; P6: physics of solids; R2: biomaterials & bioengineering; R4: pharmacology & toxicology; Z3: microbiology

By contrast, the citation classes defined by the characteristic scores are by and large insensitive to both the length of the citation window and the underlying subject. Table 2 gives the corresponding values for the same subfields as above. The share of papers cited less frequently than the average (Class 1) amounts to roughly 70%, the share of those categorised to Class 2 to about 21% and the in the highest two classes one finds 6%–7% and 2%–3% of all publications, respectively. This coincides with the observations made by Glänzel (2007) on the basis of the 1980 volume of the *Science Citation Index* (SCI) and a 21-year citation window.

The comparison of national citation impact with the world standard can readily be done by using the above classes $[b_{k-1}, b_k)$, $k = 1, 2, 3$ and $[b_3, \infty)$ as the respective subject standard. The comparison of the distribution over classes provides a more detailed picture, notable on the high end of the performance range, than the comparison of the means and the shares of uncited papers alone. The calculation of the corresponding scores for each individual country is not necessary. The share of a given country's (or any other unit's) publications found in the four performance classes of the reference population can be compared with the world standard as shown in Table 2 or with those of other countries (or other units). Note that the unit under study (and all other benchmark units as well) must be part of the reference population. If a unit under study were the true mirror of the entire population, its share in all four classes would be identical with the reference standard. Any deviation from this standard indicates a specific profile. The unit's profile might be more or less *skewed* with higher or lower shares in the lower classes, respectively, and more or less *polarised* according as the lower/higher share of lower-class papers is compensated by a higher/lower share of upper-class papers. Such cases have been reported by Glänzel (2012) for the *Scientometrics* sample, where China had a more skewed profile than the reference standard, Belgium had a less skewed profile and the profile of the USA was somewhat less polarised than the reference standard.

In the following, the method will be explained on the basis of a discipline in the life sciences. In particular, the subfield 'cardiovascular & respiratory medicine' (I1) has been chosen. The country Belgium is used as the example unit and the publication year is 2007. 55 out of 561 papers with at least one Belgium (co-)author have received at least 35 but less than 70 citations each (cf. Table 1). These 9.8% of all Belgian papers are considered remarkably cited (Class 3). 26 papers have been cited at least 70 times each. Thus 4.6% of Belgian papers in the subfield cardiovascular & respiratory medicine are outstandingly cited (Class 4). The share of papers (38.5%) in the three Classes 2, 3 and 4 exceeds the reference standard of 29.8%. Consequently, the remaining class of poorly cited papers (Class 1) contains less papers than expected on the basis of the world standard.

The indicators for the world's 20 most active countries in this subfield are presented in Table 3. The comparison among the individual countries can be interpreted analogously. The "reduced" distribution with four classes provides a quantified overview of citation impact with respect to the world standard while it

keeps the peculiarities of the shape and skewness of the original citation distribution.

Table 3. National shares of publications in the reference CSS classes in 2007 and 2009 for subfield I1 according to the Leuven-Budapest scheme (in alphabetic order) [Data sourced from Thomson Reuters Web of Knowledge]

Country	*2007 (5-year citation window)*				*2009 (3-year citation window)*			
	Class 1	*Class 2*	*Class 3*	*Class 4*	*Class 1*	*Class 2*	*Class 3*	*Class 4*
BEL	61.5%	24.1%	9.8%	4.6%	61.2%	24.0%	9.3%	5.5%
BRA	73.5%	19.8%	4.7%	2.0%	87.0%	8.8%	3.1%	1.2%
CAN	61.8%	25.3%	9.2%	3.7%	59.4%	26.6%	8.7%	5.3%
CHE	60.8%	25.2%	10.7%	3.3%	61.7%	23.0%	9.5%	5.8%
CHN	68.7%	24.4%	5.6%	1.3%	72.8%	21.0%	4.8%	1.4%
DEU	62.5%	24.5%	8.9%	4.1%	63.0%	23.7%	8.6%	4.7%
ESP	73.8%	17.8%	5.2%	3.2%	72.9%	17.0%	6.7%	3.4%
FRA	71.3%	17.8%	7.5%	3.4%	66.4%	20.9%	7.9%	4.8%
GBR	61.0%	26.2%	8.5%	4.3%	62.1%	24.0%	8.9%	5.0%
GRC	74.8%	19.4%	4.2%	1.6%	75.6%	17.8%	4.6%	2.0%
ITA	70.8%	20.0%	6.3%	3.0%	66.9%	21.7%	7.3%	4.0%
JPN	73.2%	19.9%	5.3%	1.5%	71.6%	21.3%	5.2%	1.8%
KOR	74.2%	18.2%	5.2%	2.3%	65.4%	25.1%	7.6%	1.9%
NLD	56.4%	28.9%	9.9%	4.8%	57.7%	28.0%	9.9%	4.4%
POL	71.4%	20.6%	4.2%	3.8%	82.4%	10.3%	3.8%	3.5%
SWE	59.1%	27.7%	10.0%	3.2%	60.2%	24.3%	9.9%	5.6%
TUR	92.7%	6.3%	0.9%	0.0%	93.8%	4.7%	1.1%	0.4%
TWN	78.6%	17.4%	2.6%	1.4%	76.4%	16.8%	5.0%	1.7%
USA	61.0%	26.4%	9.0%	3.6%	61.8%	25.0%	8.9%	4.3%
Total	70.2%	21.3%	6.2%	2.3%	71.2%	20.0%	6.1%	2.7%

The distributions over the four "performance" classes provide more detailed insight than traditional citation indicators. Clearly, Italy's distribution in this subfield reflects a more favourable situation than that of Japan in both years and Turkey has the least favourable one in the country set. The question arises of what indicators could possibly be built on the basis of these shares. Glänzel (2012) has argued that no combination or composite indicator *over classes* should be built. Except for smoothening the effect of outliers, such indicators would not provide more information than properly calculated elementary statistics. It has been stressed that, on the other hand, a combination *over subjects* is, in principle, possible, provided of course that document assignment to performance classes can be "disambiguated" in case of multiple subject assignment. In any case, classes should be determined for each individual subject first, and appropriate shares should be combined on the basis of the unit's publication counts in the

corresponding classes afterwards. Also the choice of the level of aggregation of the underlying subject is crucial. If subject areas are too broad, the high end of the citation distribution is formed by papers in subjects that have, in general, a high standard but theoretical or technology-oriented topics would scarcely appear in the upper classes. If, on the other hand, subjects are too narrow then the number of papers is not sufficient to form stable classes, or, in other words, the upper classes remain (nearly) empty for most units. The above 60 subfields seem to form a stable groundwork for both national and institutional assessment. In the next subsection the combination of subjects will be discussed.

CSS in all fields combined

One precondition for the application of CSS to broad science fields or to all fields combined is the unique assignment of publications to performance classes. The following example describes this problem. Assume, for instance, that a paper is assigned to two subjects, here denoted by S1 and S2. According to possibly different citation standards in the two subjects, the paper is then assigned, for instance, to Class 3 in subject S1 and to Class 4 in S2 because its citation rate does not exceed b_3 in S1 but it is greater than the corresponding threshold b_3 in S2. A direct combination can, therefore, not provide any acceptable solution. A proper subject-based fractionation must be applied such that each publication is gauged against only one individual threshold value. As argued in the study by Glänzel et al. (2009) one important consequence of multiple assignments is the necessity of fractionation by subjects and thus of calculating proper weights for the corresponding individual subject-expected citation rates. Furthermore, it was stressed that the weighting of fractional data is correct only if the sum of the individual field expectations over all publications in the system equals the citation total of the database in the combination of these fields. This will result in an 'implicit' classification without calculating any *common* thresholds b_k. Again, the procedure is based on an iteration, where the first step is identical with the procedure of calculating subfield-expected citation rates. A first fractionation is applied when the citation means of subfields is determined. This is done on the basis of the respective number of subfields to which a publication is assigned. Both publications and citations are fractionated. The second one follows when individual expectations are calculated for each paper. This expectation is then the mean value of the fractionated subfield standards. In the following step of the iteration, all papers, that have received less citations than their individual expectation, are removed. The above procedure is repeated on the remaining set. This is done three times in total to obtain the *individual* characteristics scores b_k^* ($k = 1, 2, 3$) for each publications. All papers can now uniquely be assigned to one of the four classes. It should be mentioned in passing that, if the underlying paper set comprises only publications from one single subfield and fractionation is not required, the results will be identical with those described in the previous subsection. It is straightforward that, in this case, the individual thresholds are identical with the common characteristic scores.

Table 4. Distribution of publications over major fields in 2007 and 2009 according to the Leuven-Budapest scheme
[Data sourced from Thomson Reuters Web of Knowledge]

Field	*2007 (5-year citations)*		*2009 (3-year citations)*	
	WoS	*Class 4*	*WoS*	*Class 4*
A	7.0%	8.2%	7.5%	8.5%
B	10.1%	10.1%	9.3%	9.3%
C	20.2%	19.8%	20.0%	21.7%
E	11.2%	8.5%	11.8%	9.1%
G	5.7%	6.9%	5.8%	6.7%
H	4.5%	4.1%	5.0%	4.1%
I	12.2%	11.0%	12.0%	10.5%
M	18.4%	18.3%	18.7%	18.3%
N	5.7%	6.8%	5.6%	6.7%
P	15.0%	13.6%	14.3%	13.2%
R	7.2%	6.4%	7.2%	6.8%
Z	10.3%	9.6%	10.0%	9.8%

* *Legend:* A: Agriculture & environment; B: Biosciences (General, cellular & subcellular biology; genetics); C: Chemistry; E: Engineering; G: Geosciences & space sciences; H: Mathematics I: Clinical and experimental medicine I (General & internal medicine); M: Clinical and experimental medicine II (Non-internal medicine specialties); N: Neuroscience & behavior; P: Physics; R: Biomedical research; Z: Biology (Organismic & supraorganismic level)

One important validity aspect of this method is the appropriate subject distribution in all performance classes, notably in the highest one since this reflects outstanding performance. Thus the question arises of whether all subject fields are proportionally represented in what is considered the high end of the citation distribution. Table 4 gives the distribution of papers over major fields according to the Leuven-Budapest scheme and the field distribution of papers assigned to Class 4 in 2007 and 2009. The same citation windows as above have been used here as well. Some deviation from the complete WoS representation can be observed in both years but this deviation should not be considered a serious bias. The patterns in Table 4 are strikingly stable over time although different citation windows have been applied. All subjects can, therefore, be considered adequately represented among highly cited publications.
The distribution of papers over classes reflects the same stability as already found in the disciplinary analysis in the previous subsection (cf. Table 2). The CSS procedure in all fields combined resulted in the following distribution for the two selected WoS volumes.

- 2007 (5-year citations): 69.8% (Class 1), 21.5% (Class 2), 6.3% (Class 3), 2.4% (Class 4)
- 2009 (3-year citations): 69.7% (Class 1), 21.4% (Class 2), 6.4% (Class 3), 2.5% (Class 4)

Figure 1 gives a graphic presentation of the world standard and the national shares in the upper three classes in 2007 for the 30 most active countries in 2007 and 2009. Among these countries, Belgium, Denmark, the Netherlands and Switzerland have the highest shares in the upper three CSS classes with more than 40% each. Norway, Sweden, UK and USA, with slightly lower values, have a similar profile. This, of course, corresponds to the lowest share of "poorly" cited papers (Class 1) since, by definition, the content of the four classes adds up to 100%. Besides, a similar share of Class 1 papers does not imply the same distribution over the upper classes. France and Poland in 'cardiovascular & respiratory medicine' (I1) in 2007 might serve just as an example (see Table 3). Even very similar shares of Class 2 papers might go with different distributions over the two other upper classes as the comparison of the country pairs Belgium-Sweden, Finland-USA and Brazil-China in all fields combined (2007) convincingly illustrates (cf. Figure 1).

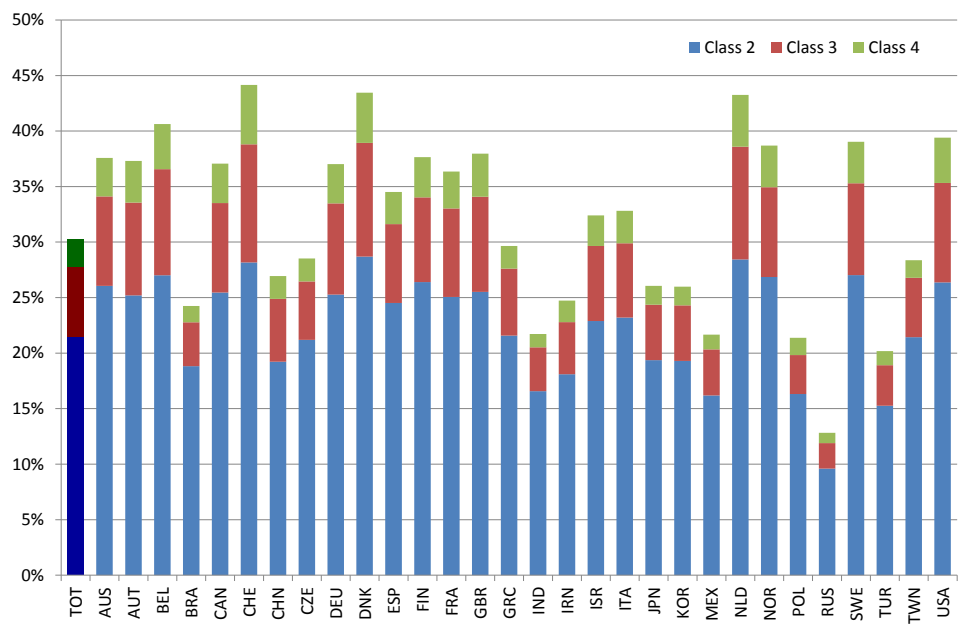


Figure 1. The world standard (left-most column) and national shares of publications (in alphabetic order) in the upper three CSS classes in all fields combined in 2007 (5-year citation window) [Data sourced from Thomson Reuters Web of Knowledge]

The same presentation for the WoS volume 2009 on the basis of a three-year citation window can be found in Figure 2. The reference standard is practically unchanged with respect to the 2007 volume with the five-year citation window. Nevertheless, a certain polarisation can be observed. UK, Italy and Switzerland (with growing shares in the upper three CSS classes), and Poland, Iran and Brazil (with decreasing shares in these classes) are the most concerned countries in this selection.

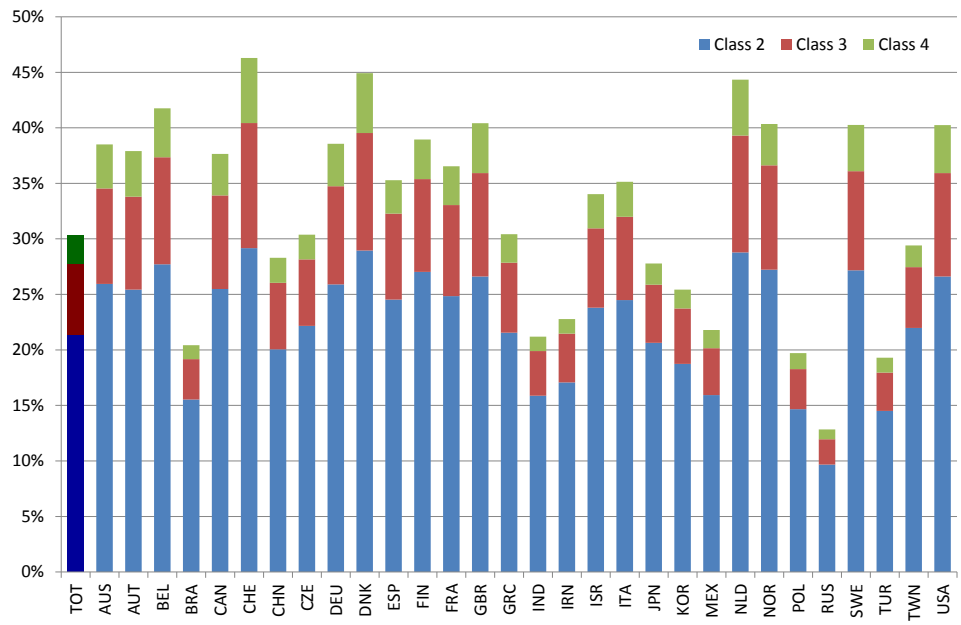


Figure 2. The world standard (left-most column) and national shares of publications (in alphabetic order) in the upper three CSS classes in all fields combined in 2009 (3-year citation window) [Data sourced from Thomson Reuters Web of Knowledge]

Belgium, Denmark, the Netherlands and Switzerland are the four countries with the highest standard and the lowest share of Class 1 papers in 2009 as well, and are again closely followed by the Norway, Sweden, UK and the US. The profile of Russia reflects the least favourable situation, but is along with that of Mexico and France the most stable one in the selection.

The possibility of the identification of individual highly-cited papers (Class 4 publications) forms a further added value of this method.

Finally it should be mentioned, that in contrast to the "subject disambiguation" in the calculation of citation thresholds, assignment to addresses is not unique. Note that, for instance, a paper in Class 4 is counted as highly cited for both Germany and France, whenever it has co-authors from the two countries.

Discussion and conclusions

The analysis of the high end of scientific distributions is one of the most difficult and challenging issues in evaluative scientometrics. This is, of course, not merely a mathematical issue as it is always difficult to draw a sharp borderline between "very good" and "outstanding". Also the effect of outliers, i.e., of observations that might bias or even distort statistics, impressively shown by Waltman et al. (2012), is not typically a bibliometric issue. So-called censored data or data distorting extreme values of a distribution are known in several fields, for instance, in insurance mathematics (e.g., Matthys et al., 2004). In the proposed CSS-based method the effect of outliers is limited as the influence of individual observation on the total is marginal and observation for the units under study are represented by classes instead of individual values.

Self-adjusting classes, such as those based on CSS, allow the definition of proper performance classes without any pre-set thresholds. This is certainly one of the main advantages of the proposed method. Another one is the needless integration of measures of outstanding performance into the assessment tools of standard performance. The method of "implicit" subject fractionation can also be used in the context of other publication and citation indicators, whenever the issue of multiple subject assignment needs to be resolved.

References

Adams, J., Gurney, K. & Marshall, S. (2007), Profiling citation impact: A new methodology. *Scientometrics*, 72(2), 325–344.

Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J.L. (2004). *Statistics of extremes: Theory and application.* Wiley.

Beirlant, J., Glänzel, W., Carbonez, A. & Leemans, H., Scoring research output using statistical quantile plotting. *Journal of Informetrics*, 1(3), 2007, 185–192.

Glänzel, W. & Schubert, A. (1988a), *Theoretical and empirical studies of the tail of scientometric distributions*. In: L. Egghe, R. Rousseau (Eds.), Informetrics 87/88, Elsevier Science Publisher B.V., 75–83.

Glänzel, W. & Schubert, A. (1988b), Characteristic Scores and Scales in assessing citation impact. *Journal of Information Science*, 14(2), 123–127.

Glänzel, W. & Schubert, A. (2003), A new classification scheme of science fields and subfields designed for bibliometric evaluation purposes. *Scientometrics*, 56(3), 357–367.

Glänzel, W. (2007), Characteristic scores and scales. A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, 1(1), 92–102.

Glänzel, W., Schubert, A., Thijs, B. & Debackere, K., Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. Scientometrics, 78(1), 2009, 165–188.

Glänzel, W. (2012), *High-end performance or outlier? Evaluating the tail of scientometric distribution*. In: H.N. Choi, H.S. Kim, K.R. Noh, S.H. Lee, H.J. Kang, H. Kretschmer (eds), Proceedings of the 8th International Conference on Webometrics, Informetrics and Scientometrics (WIS) & 13th COLLNET Meeting, 23-26 October 2012, Seoul, Korea, KISTI, 42–52. (A previous version accessible as MTA TTO working paper #2012/4 at: http://www.mtakszi.hu/kszi_aktak/doc/ ksziaktak_1204.pdf)

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3, 1163–1174.

Kratz, M. & Resnick, S. (1996), The qq-estimator of the index of regular variation. *Communications in Statistics: Stochastic Models*, 12, 699–724.

Leydesdorff, L., Bornmann, L., Mutz, R. & Opthof, T. (2011), Turning the tables on citation analysis one more time: principles for comparing sets of documents. *JASIST*, 62(7), 1370-1381.

Matthys, G., Delafosse, E., Guillou, A. & Beirlant, J. (2004), Estimating catastrophic quantile levels for heavy-tailed distributions. *Insurance Mathematics & Economics* 34(3), 517–537.

Renyi, A. (1953). On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungaricae*, 4(3-4), 191–231.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., van Eck, N.J., van Leeuwen, Th.N., van Raan, A.F.J., Visser, M.S. & Wouters, P., (2012), *The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation*. In: Eric Archambault, Yves Gingras, & Vincent Lariviere (Eds.), Proceedings of STI 2012 Montreal (17th International Conference on Science and Technology Indicators), Volume II, 791–802.

APPROACH TO IDENTIFY SCI COVERED PUBLICATIONS WITHIN NON-PATENT REFERENCES IN PATENTS

Masashi Shirabe[1]

[1] *shirabe.m.aa@m.titech.ac.jp*
Tokyo Institute of Technology, Oookayama 2-12-1 S6-5, Meguro, Tokyo 152-8550 (Japan)

Abstract
In order to evaluate approaches for identifying SCI covered publications within NPRs, I introduce a simple method that uses two key indicators, recall and precision, to evaluate the relevance of information retrieval systems. There are two primary reasons that conventional methods of evaluating matching results are insufficient: there is nothing in place to evaluate accuracy, and there is a direct dependence on the intermediate outcome. The proposed approach consists of five main steps: 1) data collection, 2) creation of supervised data and test data, 3) selection and execution of matching algorithms, 4) evaluation of algorithms and optimization of their combinations, and 5) evaluation of optimized combinations. A comparison of the proposed and conventional methods show that the proposed approach works well; its results (i.e., 99% precision and 69% recall) are better than the target implicitly set in a previous study (Tomizawa, 2008). In that sense, the proposed approach is quite promising.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5).

Introduction
Although several previous studies (e.g., Meyer, 2000a, 2000b) have pointed out the theoretical and methodological problems in non-patent references (NPRs) in patents, they are frequently used, especially in scientific citations, as valuable data for studying interactions between science and technology (e.g., Narin, Hamilton & Olivastro, 1997; Tamada, 2010). The rationale behind these problems could be described from various angles (Verbeek et al., 2000d). It is important to note that approaches to analyzing knowledge interaction between science and technology by using science citation data "enable a high quality analysis of knowledge exchange linkages (flows) between S&T and allows one to touch upon the degree of diffusion of science into technology (Verbeek et al., 2000b)."
Theoretical and methodological discussions about scientific citations of patents as a source for analyzing S&T interactions are both interesting and important, but in the present paper I step back from this issue and focus on a practical problem relevant to such citations. Specifically, I present an approach to identify SCI publications based on NPRs in U.S. patents and evaluate its effectiveness.

The identification of SCI publications essentially means matching items within NPRs and documents in the SCI database. This task is complex and resource consuming because of the immense number of NPRs (more than 1.1 million in U.S. patents granted in 2009 alone) and also the number of documents in the SCI database. Moreover, NPR data are neither formatted nor indexed, and the dataset of records contains various errors, from spelling mistakes to a lack of required information such as publication year and journal title. In some cases, one reference can contain two or more publications.

Fortunately, the SCI database is formatted and reliable to some extent, and we can access it through an online retrieval system. Therefore, if we wanted to identify a small number of NPR records, we could complete the task easily and in a relatively short period of time (e.g., several minutes per record). Presumably, the accuracy is high enough.

However, if we wanted to complete a comprehensive identification of NPR records manually, the task would become nightmarish because of the astonishingly large number of records to be processed. It is therefore necessary to automatize matching tasks. Unfortunately, such automatization is very difficult. Due to the unformatted, unindexed, and mistake-filled nature of the data, automated matching has serious problems with its accuracy. What this means is that we face a tradeoff between accuracy and cost.

One might suggest, "Why don't you simply purchase a dataset of matched linkages between patents and science publications?" Indeed, a few vendors have built such databases and provide commercial access to them. However, cost is an issue. When a given research design requires only a limited range of linkage data, it might be affordable. For example, if we require a small amount of data to analyze the top 500 influential patents and their scientific linkages (Tomizawa, Hayashi, Yamashita & Kondo, 2005), or are studying the interaction between science and a specific technology (e.g., amorphous silicon solar cells), the cost would be relatively low. However, if we were interested in the overall characteristics of S&T interactions (e.g., the difference in strength of scientific linkages by fields and their time series variation), the cost would likely go well beyond our budget because this type of research requires millions of pieces of linkage data. In addition to price, accuracy is another issue. Presumably, the accuracy of commercial databases listing scientific citations in patents is high enough, but to the best of my knowledge typical vendors do not provide enough information about accuracy. To overcome these cost and accuracy issues, we need to develop a method of matching scientific citations in patents and science publications.

Some research has already been done on this topic. For example, in a pioneering work in 1997, Narin et al. identified an increase of scientific linkages (e.g., the average number of scientific citations in patents) in the latter half of 1980s and beyond by using matched linkages between SCI and USPTO data. Moreover, they pointed out that a number of science publications cited by patents are generated from research projects supported by government research funding. This suggests

the importance of government research funding for promoting interactions between science and technology. Regrettably, they made little mention of how precisely they matched references in U.S. patents and SCI bibliographic records or of the success number and/or rate of their matching procedures.

The conventional approach to matching scientific citations in patents and science publications can be divided into two phases: extraction and matching. In the first phase, a set of journal references is extracted from all NPRs by using keywords characteristic of science and non-science publications. In the second phase, by focusing on match-keys (e.g., the family name of the leading author), each NPR in the journal reference set is compared with science publications stored in a publication database. If the NPR and a science publication share a certain number of match-keys, the two are judged as matched. Two previous studies specified match-keys as leading author, publication year, volume, and starting page (Verbeek et al., 2002c) and as ending page and partial information of title (Tomizawa, 2008).

Table 1. Percentage of journal documents and SCI journal document in NPRs of US patents

	Patent registered period	*% of journal docs*	*% of SCI journal docs*
Narin and Noma (1985)	1978–80	48%	37%
Van Vianen et al. (1990)*	1982–85	56%	46%
Tomizawa (2008)	1995–2005	55%**	n/a
Shirabe (2008)	2001–2005	n/a	49%***
Callaert et al. (2012)****	n/a	50%	n/a

*Dutch patents **Identified by ipIQ ***JCR journals ****US, EPO, PCT patents

The first phase, which can be seen as the retrieval of science publications within NPRs, has been the subject of several studies. Table 1 shows some of the relevant results. Although we have to keep in mind that these percentages were not evaluated directly, the numbers are quite consistent. In this (admittedly weak) sense, the results of automatized extraction of journal documents from NPRs in patents are somewhat reliable.

In contrast, the second phase, which is the matching of scientific citations with publications, has not received much attention. Only a handful of studies have described any concrete methods or evaluated their matching results. This is partly because evaluating the matching results has a paradoxical characteristic, as Verbeek et al. (2002c) described: for a successful evaluation, correct matching results are required. To avoid this paradox, it is necessary to introduce a sampling test.

Verbeek et al. (2002c) used an analysis of randomly sampled NPRs in 10,000 USPTO patents to show that their matching method could identify (i.e., retrieve) 927 SCI publications out of 2,653 successfully parsed NPRs. Moreover, in consideration of their window of analysis and the coverage of the SCI (1972–

1996) used, they estimated that at least 1,287 citations are indeed covered by the SCI. They evaluated the success rate of their matching (and parsing) approach by retrievability (hit) ratios, which is defined as the ratio of successfully retrieved items to items to be retrieved.

This evaluation of matching results is not necessarily adequate, for two reasons. First, "successfully retrieved" does not necessarily mean "successfully identified" in the strict sense. As their approach judges items shared {lead author name} and two of the three match-keys ({publication year}, {volume}, and {starting page}) with SCI publications as "retrieved" (i.e., "identified"), there can be misidentifications, especially in case of lead authors with common names. Although such misidentifications might be few, they should be considered in the evaluation of matching results. Second, this evaluation of the success rate depends heavily upon the results of their parsing NPRs. Thus, even if the same matching approach was adopted, the evaluation result could change depending on the parsing approach used. This is not an adequate characteristic for evaluation. Although one might insist that matching and parsing as well as the extraction of journal references from NPRs should be evaluated as a whole, it makes even less sense that the success rate can change with the same number of "successfully identified" items. Therefore, the development of methods to match SCI publications and NPRs first needs an adequate method to evaluate the matching results.

In a weak sense, the approach of Verbeek et al. (2002c) is regarded to "successfully" identify 927 SCI publications out of 10,000 randomly sampled NPRs, and it remains unclear how precise this identification is. Meanwhile, Tomizawa (2008) reported that the retrievability ratio of his matching approach is estimated to be at least 60% based on tentative results of matching NPRs and Scopus publications. The retrievability ratio is better than that of Verbeek et al. (2000c), although they should not be compared directly due to differences in the publication database and patent publication periods. As 55% of NPRs are estimated to be retrievable, his tentative estimation suggests that 33% (i.e., 55% × 60%) of NPRs might be matched to publications stored in a scientific literature database. Although the precision of these matching results is unclear (just like Verbeek et al.'s (2000c)), this figure can at least provide a target for the development of matching methods.

Based on the above discussions, I propose an approach to match SCI publications within NPRs in U.S. patents.

Method

The proposed approach consists of five main steps: 1) data collection, 2) the creation of supervised data and test data, 3) selection of matching algorithms and their executions, 4) evaluation of algorithms and selection of matching algorithms by optimizing their combinations for the supervised data, and 5) evaluation of the optimized combinations of matching algorithms. I shall explain these steps one by one. However, first, I propose a method for evaluating matching results.

Evaluation of matching results

For evaluating matching results, I apply an information retrieval perspective. As Kita, K., Tsuda, K., and Shishibori, M. (2002) explain, the effectiveness of information retrieval systems can be evaluated in terms of the relevance, pertinence, and usefulness of the retrieved results. The task of identifying SCI publications within NPRs can be regarded as using an information retrieval system for the manual identification of which items in the SCI database to use. That is to say, this system produces outputs (retrieved documents = SCI publications) from inputs (retrieval phrases = NPRs). Moreover, both the pertinence and usefulness of the retrieval system fully depend on the relevance of the retrieved documents in the context of this study. That is, only the relevance needs to be known to evaluate the identification of SCI publications within NPRs. Typically, the relevance of retrieval systems is evaluated using two indicators: recall and precision rates (Kita et al., 2002). They are defined as follows (see also Figure 1).

recall = A/(A+C)
= (number of adequate (correct) documents within retrieved documents) / (number of adequate documents within all documents)

precision = A/(A+B)
= (number of adequate documents within retrieved documents) / (number of retrieved documents)

Documents	Adequate	Inadequate
Retrieved	A	B
Not retrieved	C	

Figure 1. Precision and recall of information retrieval systems.

The recall and precision indicate the coverage and accuracy of the matching results, respectively. In other words, recall and precision are relevant to Type 2 and Type 1 errors, respectively. Obviously, these indicators are independent of the intermediate outcomes of matching and depend only on the matching results. In this sense, such a combination can be an adequate indicator for evaluating the matching results. It is on this basis that I explain the five steps of the proposed approach.

Step 1: Data collection

For the purpose of this research, I used NPRs from all the U.S. utility patents registered between 2000 and 2009, which can be downloaded from the USPTO

Web site. I also used a set of WoS data on a hard disk; that is, we bought access rights from Thomson Reuters and obtained data from 2011 for our research project. The present study is part of the project. The records contained in the set were stored in the WoS database between 1992 and 2011, so the database years start in 1992 and end in 2011.
I should point out two properties of this WoS dataset. First, each record contains most of bibliographic data of the Web version of WoS but no abstract. This means it contains the title, author(s), affiliation(s), journal title (full title and abbreviations), its volume and issue, its beginning page and ending page, and so forth. Second, the Web version of WoS is updated and corrected regularly and retroactively. This means that the dataset we used is slightly different from the Web version and contains a few more errors.

Step 2: Creation of supervised data and test data

To create the supervision and test data, we randomly sampled 2,000 NPRs for each register year between 2000 and 2009 from an entire set of NPRs in U.S. utility patents by using pseudorandom numbers produced by a script language, Perl. That is, a total of 20,000 NPRs were extracted from targeted NPRs. Next, 1,000 NPRs for each register year were randomly allocated to the respective supervised and test data sets.

Table 2. Number of U.S. utility patents, NPRs, and matched SCI publications.

Registered year	*2000*	*2001*	*2002*	*2003*	*2004*
No. of patents	157,496	166,038	163,518	169,035	164,291
No. of NPRs	466,056	519,743	549,741	585,150	557,524
No. of matched SCI publications (out of 2,000 samples)	1042	1032	1040	1012	983
% of SCI publications in NPRs	52.1%	51.6%	52.0%	50.6%	49.2%
Registered year	*2005*	*2006*	*2007*	*2008*	*2009*
No. of patents	143,806	173,770	157,283	157,772	167,349
No. of NPRs	557,780	851,232	868,929	936,926	1,139,407
No. of SCI matched publications (out of 2,000 samples)	932	922	895	844	835
% of SCI publications in NPRs	46.6%	46.1%	44.8%	42.2%	41.8%

We then manually matched the NPRs to SCI publications contained in the Web version of WoS (database years: 1965–present) by using its own retrieval system. This task was executed in 2011 and 2012. Although it is ideal to determine matched records by matching up the outputs of independent sources, due to a lack of resources we could only check the results of outsourced matching works. This at least ensured that the matched results were double-checked and therefore sufficiently accurate. Table 2 shows the results of manual matching and some basic indicators of the U.S. utility patents registered.

Because the number of NPRs has grown more rapidly than the references to SCI publications, the share of SCI references has gradually declined. However, SCI references per patent have been growing. The average in the sample from 2001 is 1.54 while that in 2009, 2.84, is almost double. In addition, it is worth mentioning that this share of matched SCI publications is fairly consistent with a previous work (Shirabe, 2008) in spite of differences in identification methods.

Step 3: Selection of matching algorithms and their executions

Inspired by previous studies (Verbeek et al., 2000c; Tomizawa, 2008), I use five match-keys ({leading author's family name}, {volume}, {publication year}, {beginning page}, and {title}) in matching the algorithms. I also use {journal title}. All the matching algorisms are coded in Perl, and its regular expression engine is fully utilized.

Before applying each algorism, some of the records in sampled NPRs are excluded because they are judged to refer to (foreign) patents, gene database records, books, and so forth. Such records are excluded by using regular expressions like /\bep \d* \d*/i, /application no\./i, and /chapter \d/i. Of course, these excluded records are not excluded from the evaluation of the matching results.

Among the above match-keys, {volume}, {publication year}, {beginning page}, and {title} are parsed for NPRs by using regular expressions. However, this parsing algorism is implemented in a rather ambiguous way. That is, each record could have two or more candidates for those four keys. I introduced this algorism design because even a good parsing algorism makes mistakes to some extent, and those mistakes sometimes exclude SCI references to be matched in later processes. In other words, the main purpose of parsing here is not to extract matching candidates implicitly but to narrow the scope of the matching. Without such reduction, we would not be able to complete the matching within a reasonable amount of time (Tomizawa, 2008).

Moreover, with regard to {volume}, {publication year}, and {beginning page}, I use not only the results of parsing but also figures that appeared in references in some algorithms. That is to say, in some algorithms, NPRs containing "1993" are judged in principle to be candidates to match with SCI publications published in 1993. This less constrained matching might be expected to minimize the effect of parsing errors. We still need to test whether the algorithm works as designed, though.

To use {title} as a match-key, I adopt a unique method of reducing information. I chose to do this partly because identification of titles from unformatted texts like NPRs can result in many errors and partly because such a task is difficult to implement even with insufficient accuracy. Therefore, I use combinations of three initial characters of three consecutive words in NPR texts as a key. This means that each NPR record has quite a long list of such combinations in most cases. This system of labeling narrows the scope of matching drastically yet still seems to retain sufficient information contained in titles. Needless to say, the latter detail

(retaining sufficient information) should be judged based on the evaluation of matching results.

To match journal titles, I use their abbreviations. For example, “International Journal of Solids and Structures” is abbreviated to “Int. J. Solids Struct.” in ISO abbreviation. By using such abbreviations, if an NPR is matched to /int.*\bj.*\bsolids.*\b.*struct/i in matching by regular expression, the NPR is judged to contain the journal title (i.e., “International Journal of Solids and Structures”). As previous studies (e.g., Verbeek et al., 2000c) suggest, matching by journal title often results in misidentifications due to misspellings, the variety of abbreviations, and so forth. Combinations with other keys might reduce such misidentifications to a tolerable extent. Even if not, we just have to exclude algorithms using journal titles in the next step. Thus, here I use {journal title} as a match-key.

Based on the combinations of match-keys explained above and the less constrained pattern matching for {volume}, {publication year}, and {beginning page}, I coded 98 different algorithms to match SCI publications and NPRs (while considering computation time for matching) and then applied them to the set of supervised data. In these algorithms, two items are judged as matched only if one SCI item is matched for a source NPR item.

Step 4: Evaluation of algorithms and optimization of combinations of algorithms

First, the 98 algorithms are evaluated in terms of recall and precision by using the supervised data set. Then, combinations of algorithms are evaluated by using the same dataset. As shown later in the Findings section of this paper, each algorithm differs from one another in terms of recall and precision. In addition, each algorithm might have its own specialized patterns of NPRs. Therefore, to enhance recall and precision, combinations of algorithms are taken into account.

These combinations of algorithms are created as follows. (1) The number of algorithms to be combined is determined as 2, 3 and 4, as a larger number might result in over-fitting and because of the limited computing time for evaluating all the combinations. (2) A rule for combining algorithms is determined. Although the majority rule could be a candidate, here I use the first-come-first-out rule, which outperforms the majority rule. First-come-first-out means that the retrieved documents of the first algorithm are always selected first and that those of the second are selected from among the others. Thus, 98×97, $98 \times 97 \times 96$ and $98 \times 97 \times 96 \times 95$ combinations are taken into account. (3) These combinations are evaluated in terms of recall and precision by using the same supervised data. (4) In accordance with the previous evaluation, the best combination of algorithms for each percentage of precision (from 90 to 95%) used as the minimum precision are selected (i.e., optimization of the supervised data).

Step 5: Evaluation of the optimized combinations of matching algorithms

In this step, we check whether the optimized combinations of matching algorithms perform up to expectation. To do so, the combinations are evaluated in

terms of recall and precision by using the test data set. Since this set is randomly sampled and independent from the set of supervised data used in the previous optimization, the results of this evaluation are reliable to an extent.

Findings

Figures 2 and 3 summarize the result of step 3: the evaluation of 98 algorithms in terms of precision and recall.

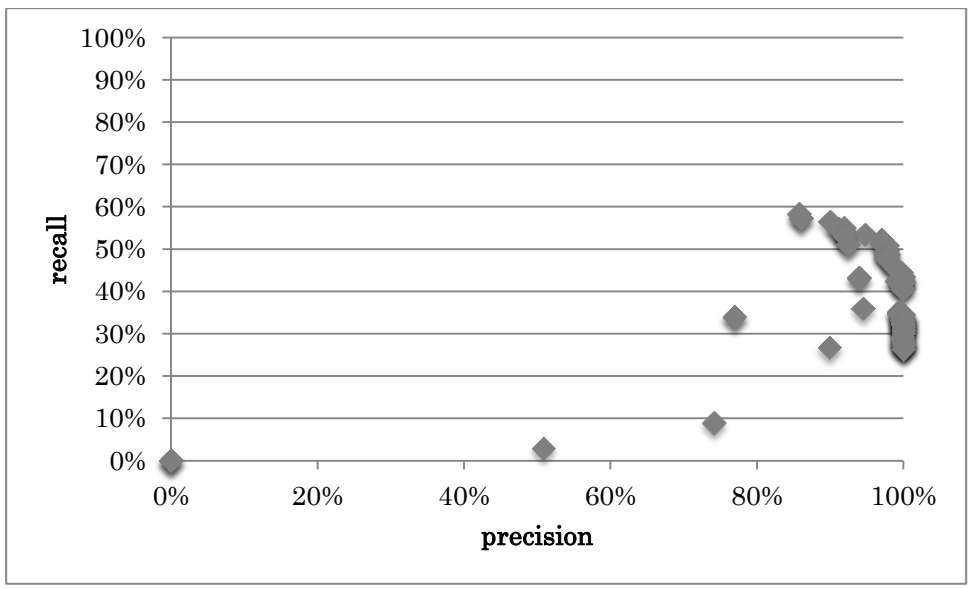


Figure 2. Evaluation of 98 algorithms.

As shown in Figure 2, the algorithms analyzed here vary in terms of recall and precision: precision ranges from 0% to 100% and recall ranges from 0% to 57%. However, there is an overall correlation between recall and precision ($R = 0.51$, $p < 0.05$), which may suggest that there are certain factors at play in determining effective algorithms. Meanwhile, focusing on the frontline of effective algorithms reveals a different perspective: there might be a tradeoff between recall and precision among the best algorithms (see Figure 3). As matching algorithms to be available for practical use are located only along this frontline, it is difficult to automatize the task of identifying SCI publications within NPRs.

One of the most precise algorithms had 100.0% precision and 26.3% recall. Only if an NPR shares all the match-keys with an SCI publication would the algorithm judge the two as matched. That is to say, virtually only identical publication/reference pairs could pass this test. However, there used to be a few matching "errors" even in the algorithm, because there were about 1% of errors in the WoS dataset, confirmed by its vendor, as well as the rare but certainly existing

errors in manual matching. Thus, in a practical sense, there is a sort of upper limitations in precision.

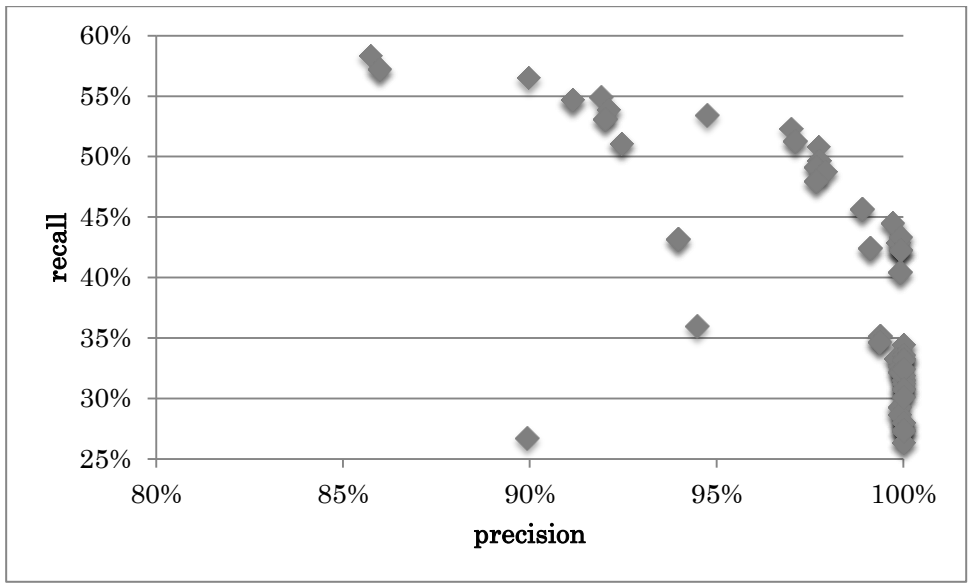


Figure 3. Evaluation of 98 algorithms (partially enlarged).

Table 3 and 4 show the result of step 4: optimized combinations of algorithms and their recall/precision.

Table 3. Results of optimization: Combinations of algorithms to indicate the maximum recall in each level of precision.

Minimum	*Number of algorithms =2*			*Number of algorithms =3*		
precision	*Combination*	*Precision*	*Recall*	*Combination*	*Precision*	*Recall*
99%	(No. 67, No. 2)	99.3%	61.4%	(No.34, No. 67, No. 79)	99.4%	68.5%
98%	(No. 53, No. 2)	98.7%	62.5%	(No. 34, No. 2, No. 87)	98.1%	69.3%
97%	(No. 40, No. 2)	97.3%	69.2%	(No. 34, No. 40, No. 2)	97.5%	72.5%
96%	same as above			same as above		
95%	(No. 2, No. 32)	95.7%	70.3%	(No. 34, No. 2, No. 32)	96.0	73.0

As shown in Table 3 and 4, with the decrease of the minimum precision, the combined algorithms have shown the increase of recall. It becomes saturated, though. Number of algorithms combined also has positive effects in terms of recall and precision, but such effects become saturated too.

Table 5 and 6 show the evaluation of optimized combinations using the test data set. Performances for the test data were slightly worse than those for the supervised data (i.e., "learning data") in terms of both recall and precision. However, the difference was too small to be regarded as the result of overfitting. Among the combinations in Table 4 and 5, the combination of No. 24, No.33, No. 2, No. 67 had the best performances. Their precision and recall rates were also very stable, which suggests their robustness. For readers' information, Table 7 provides the breakdown of these combinations.

Table 4. Results of optimization: Combinations of algorithms to indicate the maximum recall in each level of precision (contd.).

Minimum precision	*Number of algorithms =4*		
	Combination	*Precision*	*Recall*
99%	(No. 24, No.33, No. 2, No. 67)	99.2%	69.1%
98%	(No. 34, No. 45, No. 67, No. 2)	98.0%	72.1%
97%	(No. 24, No. 34, No. 40, No. 2)	97.6%	73.1%
96%	(No. 17, No. 2, No. 34, No. 32)	96.1%	73.3%
95%	(No. 24, No. 34, No. 40, No. 23)	95.4%	73.4%

Table 5. Evaluation of optimized combinations.

Minimum	*Number of algorithms =2*			*Number of algorithms =3*		
precision	*Combination*	*Precision*	*Recall*	*Combination*	*Precision*	*Recall*
99%	(No. 67, No. 2)	99.3%	60.3%	(No.34, No. 67, No. 79)	99.3%	67.8%
98%	(No. 53, No. 2)	98.8%	61.8%	(No. 34, No. 2, No. 87)	97.2%	68.6%
97%	(No. 40, No. 2)	96.3%	67.7%	(No. 34, No. 40, No. 2)	96.5%	71.6%
96%	same as above			same as above		
95%	(No. 2, No. 32)	94.5	69.4	(No. 34, No. 2, No. 32)	94.6%	72.4%

Table 6. Evaluation of optimized combinations (contd.).

Minimum precision	*Number of algorithms =4*		
	Combination	*Precision*	*Recall*
99%	(No. 24, No.33, No. 2, No. 67)	99.1%	69.0%
98%	(No. 34, No. 45, No. 67, No. 2)	97.5%	71.2%
97%	(No. 24, No. 34, No. 40, No. 2)	96.5%	72.2%
96%	(No. 17, No. 2, No. 34, No. 32)	94.8%	72.7%
95%	(No. 24, No. 34, No. 40, No. 23)	94.5%	72.4%

Although there are differences among publication databases (WoS or Scopus), the target for matching approaches mentioned in the introduction (i.e., that 33% of NPRs are presumably matched to SCI publications) can be calculated as 63.3% of recall in 2001 (i.e., 33/52.1). This indicates that the result of the combined algorithms is better than roughly estimated target.

Discussion and conclusion

In this paper, I introduced a standard method to evaluate the effectiveness of information retrieval systems for evaluating methods to identify SCI publications within NPRs. I then proposed an approach for identifying SCI publications. Evaluation of the approach by the method has shown that the proposed approach works well and its results seem at least comparable to the target of a previous study set. In that sense, the proposed approach has good potential.

However, in terms of using the matching results for practical purposes, we cannot say the recall is high enough, especially for research evaluation. More effective matching, especially for the recall, is needed.

For improving the recall, several approaches could be considered. One of the most promising is to use titles as match-keys more efficiently. Tomizawa (2008) introduced an efficient way to reduce the amount of information contained in titles and then use the information to identify Scopus publications within NPRs. This method presumably has a good performance. The development of such ways to reduce information contained in NPRs effectively but efficiently promises to be a key direction for future research. Another promising approach is to use probability matching on match-keys with a statistically optimized threshold rather than to use on/off matching as the present work has done. In general, probability matching works well on matching tasks with lots of ambiguities like the present task, thus the approach might be promising as well.

Table 7. Breakdown of outstanding combinations of algorithms.

Algorithm	Match-keys	Items used in less constrained pattern matching
No. 2	{leading author's family name}, {volume}, {publication year}, {journal title},{beginning page}	
No. 24	{leading author's family name}, {volume}, {beginning page},{titles}	
No. 33	{journal title},{titles}	publication year
No. 67	{volume}, {titles}	publication year

Acknowledgments

This research is partly supported by JSPS KAKENHI (Grant Number: 22500848 and 23500304) and JST/RISTEX research funding program "Science of Science, Technology and Innovation Policy". The author wishes to acknowledge the anonymous reviewer #3 for his/her detailed and helpful comments to the manuscript.

References

Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., & Thijs, B. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics*, 69, 3–20.

Callaert, J., Grouwels, J. & Van Looy, B. (2012). Delineating the scientific footprint in technology: Identifying scientific publications within non-patent references. *Scientometrics*, 91, 383-398.

Kita, K., Tsuda, K. & Shishibori, M. (2002). Information Retrieval Algorithms. Tokyo: Kyoritsu Press. (in Japanese).

Meyer, M. (2000a). Does science push technology? Patents citing scientific literature. *Research Policy*, 29, 409–434.

Meyer, M. (2000b). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, 49, 93–123.

Narin, F., Hamilton, K. & Olivastro, D. (1997). The Increasing Linkage between U.S. Technology and Public Science. *Research Policy*, 26, 317-330.

Narin F. & Noma, E. (1985). Is technology becoming science? *Scientometrics*. 7. pp. 369-381.

Shirabe, M. (2008). Analysis of Linkages between Patents and Scientific Articles: Focusing on Fields of Science. In *NISTEP REPORT No.111* Part 2 (pp. 1-7). NISTEP, Tokyo. (in Japanese).

Tamada, S. (2010). *Industry-University Cooperative Innovation*. Hyogo: Kwansei Gakuin University Press. (in Japanese).

Tomizawa, H., Hayashi, T., Yamashita, Y. & Kondo, M. (2005). Quantitative analysis of science publications referred in influential patents. In *Proceedings of the 20th annual meeting of JSSPRM*. pp. 228-231.

Tomizawa, H. (2008). Matching Science Citation data to Bibliographical Data. In *NISTEP REPORT No.111* Part 2 (pp. 1-7). NISTEP, Tokyo. (in Japanese).

Van Vianen, B., Moed H. & Van Raan A. (1990). An exploration of the science base of recent technology. *Research Policy*, 19, 61-81.

Verbeek,A., Debackere, K., Luwel, M., Andries, P., Zimmermann, E. & Deleus, F. (2002a). Linking science to technology: Using bibliographic references in patents to build linkage schemes. *Scientometrics*, 54, 399-420.

Verbeek, A., Andries, P., Callaert, J., Debackere, K., Luwel, M. & Veugelers, R. (2002b). *Linking Science to Technology - Bibliographic References in Patents,* Volume 1, Report to the European Commission (EUR 20492/1). Brussels: EC.

Verbeek, A., Andries, P., Callaert, J., Debackere, K., Luwel, M. & Veugelers, R. (2002c). *Linking Science to Technology - Bibliographic References in Patents,* Volume 2, Report to the European Commission (EUR 20492/2). Brussels: EC.

Verbeek, A., Andries, P., Callaert, J., Debackere, K., Luwel, M. & Veugelers, R. (2002d). *Linking Science to Technology - Bibliographic References in Patents,* Volume 3, Report to the European Commission (EUR 20492/3). Brussels: EC.

ARE CITATIONS A COMPLETE MEASURE FOR THE IMPACT OF E-RESEARCH INFRASTRUCTURES?

Jonkers Koen[1], Derrick Gemma Elizabeth[1], Lopez Illescas Camen[2], Van den Besselaar, Peter[3]

koen.jonkers@csic.es
[1] CSIC Institute for Public Goods and Policies, Department of Science and Innovation dynamics, C/Albasanz 26-28, 28033 Madrid, Spain
[2] SCImago group. Department of Information Science. University of Extremadura, Badajoz, Spain.
[3] Vrije University van Amsterdam, Department of Organization, Science and network institute, Amsterdam, The Netherlands

Abstract
This micro-level study explores the extent citation analysis provides an accurate and representative assessment of the use and impact of bioinformatics databases. The case study suggest that there is a relation between number of visits and number of citations. The second finding is that citation analysis underestimates acknowledged use by between 5 and 30% for most of the databases and applications studied. The paper discusses the implications of the findings for various aspects of impact measurement.

Conference Topic
Topic 2: Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability

Introduction
This paper explores to what extent citation analysis provides an accurate and complete assessment of the usage of e-research infrastructures in the research underlying published scientific articles. One of the reasons that measuring impact is generally based on citations, may be the mere existence of large, accessible databases such as WoS and Scopus. This is in addition to the preference evaluators have for measures that are "countable". The extent to which citations fully reflect the usage of knowledge claims by other scientists, however, is disputed. A number of alternative metrics, including citations in patents and social media statistics, have been promoted as ways to assess the broader impact of research, among many others e.g. De Jong et al (2011). However, for measuring scholarly impact of research, citation based indicators are still the dominant approach.
Recently, measuring impact of research infrastructures has been put on the agenda. The scholarly use and impact of research technologies, as of scientific knowledge claims, could be assessed through citation analysis. For many

scientific innovations, especially in the case of research infrastructures citations may no longer be a sufficient way with which to represent 'impact', as the user community may be very diverse. Where citations can help to measure scholarly use as a component of an infrastructure's impact, there are a number of alternatives that complement the measurement of its visibility and influence, such as the log-files that measure the visits to the website of the infrastructure. Considering the importance of research instruments in biotechnological innovation processes (e.g.: Senker, 1995) a full assessment of the impact of e-research infrastructures should also include an analysis of the references in patents. Nevertheless, citations may be a relevant representation of the use and impact of research infrastructures.

This article aims to investigate firstly to what extent that is the case: to what extent do citations to the original articles that introduce a research infrastructure provide an accurate representation of use and impact? If so, the intensity of use (measured in number of visits to the URLs of the infrastructures' domains) is systematically related to the citations to the articles in which these research infrastructures were introduced. Citations would therefore be a strong indicator of usage.

Apart from citations, papers may include in-text references to the research infrastructure. Therefore, the second aim of the paper is to investigate whether citations are an adequate representation of these in-text references to used e-research technologies. In other words, we investigate how much of the acknowledged use of research technologies is neglected when using only citation counts, while not considering the in-text references. Both questions will be explored, using research databases with biological info hosted by ExPASy.

Theoretical background: Why citations?

Two main bodies of theory underlie the use of citation analysis for the assessment of research output. The normative theory of citations states that researchers cite documents that are relevant to their topic, and that provide useful background for their research. By citing they acknowledge an intellectual debt (Bornmann & Daniel, 2006). Cronin (1984) argues that citations perform a scholarly communication function between texts in line with the normative theory of citations, and according Martin and Irvin, citations can indicate a measure of reward for past work or scientific status (Martin & Irvine, 1983).

The second theory, whilst not mutually exclusive to the first, emphasises that citations to documents are not free from personal bias or social pressures. Therefore the "social-constructive theory of citations" states that citing is a social process, and as such citations are used as an aid for persuasion(Gilbert, 1977; Cozzens, 1989).

The social constructivist theories provide some explanations for why people would add additional citations, beyond those that could be expected on the basis of the normative view of citations. In an age in which citation analysis is becoming an increasingly prominent feature of research evaluation, authors are

inclined to cite in an attempt to raise the visibility of their own work or that of their colleagues, with or without the implicit expectation that this favour will be returned. Unlike previous contributions, this paper is not concerned with these additional citations but with the phenomenon that authors may *not cite* certain knowledge claims even if they explicitly state their usage.

One potential explanation for this is that the origin of knowledge claims can be lost over time as new (arguably improved) claims emerge. The original knowledge claims may be absorbed into the common knowledge of a research discipline or even of the general public (Martin & Irvine, 1983). Researchers who use the knowledge claim may either not be aware of the existence of a citable item or consider it superfluous. Forgetting is another obvious motivation for not including a citation, as is the consideration that the knowledge claim in question does not merit a citation. Finally the possibility exists that alternative forms of acknowledgements besides citations are being used.

The motivation to include a reference can differ from author to author and from reference to reference. It is therefore probably too simplistic to think within just the two theories discussed in this section. In fact, it may be impossible to develop a convincing 'theory of citations' (Weingart, 2005), as citing behaviour and citations as indicators for impact and quality may actually be two unrelated issues. The more aggregated, the more citation counts may be detached from citing behaviour and the more useful they may be for investigating impact. Despite the highlighted limitations, there are several characteristics of citations that contribute to our understanding of what they actually represent, and these can be used to determine when it is appropriate to apply citation analysis and when a suitable alternative or complement is required.

Data and Methodology

Not all types of knowledge claims receive, on average, an equal amount of citations (Martin & Irvine, 1983). Reviews, for example, tend to receive more citations than articles (Asknes, 2005; Moed et al, 1995). Peritz (1983) showed that methodological papers in sociology were more frequently cited when compared to non-methodology papers. There are grounds to expect this is the case in the life sciences as well. A famous example is one of the most cited articles of all times (*Protein measurement with the folin phenol reagent*). Published in 1951 and with 299,133 "WoS citations" in Dec 2012, the article outlines a commonly used method in biochemistry to determine protein concentrations (The Lowry method) (Lowry et al, 1951; Garfield, 1998). The databases and applications on which this study focuses, are research tools which are used by many life scientists. The papers introducing them therefore have the potential to receive a high number of citations as well.

The databases and applications analysed in this project are hosted by the Expert Protein Analysis Server, ExPASy, developed and maintained by the Swiss Institute of Bioinformatics. They are used by life scientists to analyze and interpret among other the genetic and protein sequence information they

encounter in their research. These databases form an interesting example with which to consider how the knowledge claims which are entailed in research technologies are transmitted within the scientific community. The databases under study are PROSITE, Swiss-2dPAGE, HAMAP and ENZYME. We have selected these databases because they are only accessible through the ExPASy server, in contrast to some of the other (ExPASy) databases which can be accessed through multiple servers[11]. This makes counting of visits feasible when one has access to the original log files.

PROSITE is a protein database (Sigrist et al, 2012). It consists of entries describing protein families, domains and functional sites as well as amino acid patterns, signatures, and profiles in them. The SWISS-2DPAGE database assembles data on proteins identified on various 2-D and 1-D PAGE maps. Each SWISS-2DPAGE entry contains textual and image data on one protein, including mapping procedures, physiological and pathological information, experimental data and bibliographical references (Hoogland et al, 2004). HAMAP is a system, based on manual protein annotation that identifies and semi-automatically annotates proteins that are part of well-conserved families or subfamilies: the HAMAP families. HAMAP is based on manually created family rules and is applied to bacterial, archaeal and plastid-encoded proteins, which are contained in the database under study (Lima et al, 2009). ENZYME is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided (Bairoch, 2000).

The four databases differ somewhat from each other. Two (PROSITE and SWISS-2dPAGE) contain a great amount of data, generated by researchers worldwide, and collected and maintained by researchers from (a.o.) the Swiss Institute of Bioinformatics. The other two (HAMAP and ENZYME) contain a set of rules which are used to classify information in other protein sequence databases.

This paper aims to analyse firstly the extent to which citations to original articles provide an accurate representation of the usage of the databases with biological information hosted by ExPASy. We expect that the usage intensity (measured in number of visits to URL domains) is systematically related to the frequency of citations to the articles in which these research technologies are introduced: i.e.

[11] There may be some exceptions to this in the form of ExPASy mirror servers at some universities in several European countries, China, Australia, and Japan. The size of the weblogs of these mirror servers, however, is dwarfed by the size of the main server of ExPASy. These mirror servers were especially important in the times before quick internet facilitated easy access to the server based in Switzerland. In any case it is unlikely that the inclusion of the weblog data from these mirror servers would have made a difference in the distribution of the number of visits to the four databases. In contrast to the study by Jonkers et al (2012) the weblog data for the different directories used in this study was not cleaned by removal of visits from robots, web-crawlers etc. This may account for a substantial share of the reported web-traffic.

citation is a strong indicator of usage. In other words, we expect that the ratio of use (measured as visits to the site) and citations is about the same for the four infrastructures.
Secondly we aim to explore the extent to which citations are an adequate representation of the in-text references to e-research technologies - in this case, databases with biological information hosted by ExPASy. In other words, we want to explore if and how much of the acknowledged use of these research technologies is neglected when measuring citations alone, and whether this differs between the four infrastructures. We expect that the number of references to the articles introducing these databases in general is roughly similar to the references to these technologies made in the text.

Table 1 Source publications

PROSITE	SWISS-2Dpage	HAMAP	ENZYME
Sigrist CJA_2010_Nucleic Acids Res		limaetal_2009_nucleic acid res	Bairoch_2000_nucleic_acid_res
Falquet L_2002_Nucleic Acids Res	Hooglandetal_2004_proteomics	Gattiker A_2003_computa biol chem	Bairoch_1999_nucleic_acid_res
Sigristetal_2002_briefings bioinformatics_Scopus	Hooglandetal_2000_NAR		Bairoch_1996_nucleic_acid_res
De Castro E_2006_Nucleic Acids Res	Hooglandetal_1999_NAR		Bairoch_1994_nucleic_acid_res
Hulo N_2006_Nucleic Acids Res	Hooglandetal_1999_electrophoresis		Bairoch_1993_nucleic_acid_res
Hoffman K_1999_Nucleic Acids Rest	Tonellaetal_1998_electrophoresis		
Sigrist CJA_2005_Bioinformatics	Hooglandetal_1998_NAR		
Hulo N_2008_Nucleic Acids Res	Appeletal_1996_NAR		
Hulo N_2004_Nucleic Acids Res	Sanchezetal_1996_electrophoresis		
Bairoch A_1997_Nucleic Acids Res_1 AND Bairoch A_1997_Nucleic Acids Res_2	Pasqualietal_1996_electrophoresis		
Bairoch A_1996_Nucleic Acids Res	Appeletal_1996_electrophoresis		
Bairoch A_1994_Nucleic Acids Res	Sanchezetal_1995_electrophoresis		
Bairoch A_1993_Nucleic Acids Res	Appeletal_1994_NAR		
Bairoch A_1992_Nucleic Acids Res	Appeletal_1993_electrophoresis		
Bairoch A_1991_Nucleic Acids Res			

To answer both research questions, measures are needed of the frequency with which researchers use a database and the frequency with which they cite it. The first type of data consists of usage data of the databases, which is based on the

number of visitors which each of the directories that gives access to these databases receive. For the analysis of the ExPASy server weblog (Jonkers et al, 2012) use is made of the free software Funnel Web Analyzer developed by QUEST (2010). This data allows for the construction of an indicator of the number of visitors of these databases in the time period 2003-2008, which is used as a proxy for usage intensity.

The researchers responsible for establishing the biological databases under study request users to refer in their publications to one of a number of references mentioned on their website. Over the years, the responsible researchers have published articles with updates of and extensions to the databases. We use all the articles in order to cover all relevant references. For HAMAP we found two core references, for SWISS-2DPAGE thirteen, for PROSITE fifteen and for ENZYME five core references (see table 1).

Using the bibliometric databases[12] Scopus[13] and SCI[14], we retrieved all papers citing these articles in the period 2000-2011 (time of download June 2012). Both databases provide powerful analytical tools for citation analysis and although "*Scopus* is a database with criteria similar to those of *Thomson Reuters*, not only in the development of the collection but also in its coverage on the world level" (Moya-Anegón et al., 2007, p. 76), each database still shows differences in terms of collection policy. The *WoS* list of indexed journals is shorter than that of *Scopus*, while the time period covered by *WoS* is longer. Cited references in a large number of sources indexed in *Scopus* do not go back further than 1996. The implications of these two apparently different policies (depth versus breadth) are analysed by several information scientists (Fingerman, 2006; Ball & Tunger, 2006). This paper is mainly based on Scopus, because of its better coverage of Science Direct journals. This is relevant for our analysis, as we want to use a specific tool for full text analysis, which will be discussed below.

The number of in-text references to the infrastructures was analysed using the software "section search" of NEXTBIO (2012) offered through the SCIVERSE platform. This program analyses full texts of articles contained in the Science direct database (mainly journals owned by Elsevier) for the sections: *Title*, *Abstract*, *Introduction*, *Methods*, *Results*, *Discussion*, *Summary* and *Captions*. It

[12] Since both databases are available on the market, the number of papers comparing them from a scientometric perspective has been growing (e.g. López-Illescas et al., 2008; Gorraiz & Schlögl, 2007; Jacso, 2006).

[13] Scopus covers over 19,500 titles from more than 5,000 publishers worldwide. It includes coverage of 18,500 peer-reviewed journals and over 4.9 million conference papers, 400 trade publications and 350 book series. It provides 100 % coverage of Medline. On May 1, 2012, it contained about 47 million records, 70% with abstracts, of which 26 million records going back to 1996. [Scopus, 2012. http://www.scopus.com]

[14] *Thomson Reuters' Web of Science* covers over 12,000 research journals worldwide and provides access to "the *Science Citation Index* (1900-present), *Social Sciences Citation Index* (1956-present), *Arts & Humanities Citation Index* (1975-present), Index Chemicus (1993-present), andwww.thomsonscientific.com/products/ccr (1986-present), plus archives 1840 - 1985 from INPI." [Thomson Reuters, 2012. http://thomsonreuters.com].

does not cover the bibliography.[15] This search yields the list of articles and reviews in which one (or more) of the databases was mentioned in the text by the authors. As will be clear to the reader a search for the keyword "enzyme" will yield a large number of false positives as this word is not only used to refer to this database but also to a specific, and often researched, type of protein. Also a search for "enzyme database" yields false positives, as several other enzyme databases exist that are found through such a search.

Since NEXTBIO only analyses Science Direct journals, we refined our citation analysis. To do so we collected the smaller set of references made in Science Direct life science journals. We controlled whether all Science Direct[16] journals identified were covered in Scopus, and this proved to be the case, confirming the expectation that Scopus includes all Science Direct journals. This implies that the citation counting in Scopus covers all journals included in the NEXTBIO analysis in addition to potential references in journals not included in the Science Direct database. The next step was to compare the number of publications in which the authors refer to one of the databases in the full text with the citations of the source articles found in Scopus.

By comparing the citations made in Science Direct journals to the articles found through NEXTBIO's "section search" disregarding those that are also found through the citation analysis (M), an assessment of the extent that citation analysis leads to an underestimation of acknowledged use was made, using the following formula:

$$U\,(\%) = \left(1 - \frac{C}{C+M}\right) * 100\% \qquad (1)$$

U refers to underestimation (%); *C* refers to the number of citing Science Direct articles; and *M* refers to the number of articles mentioning the database in Science Direct journals (minus the publications also appearing in C). As the citation behaviour of authors publishing in Science Direct journals was expected to be similar to those of authors publishing in other journals, the expected total number of citations if all acknowledged reports of usage would have been reflected in citations, can be inferred.

The databases that will be presented in table 2 and 3 were selected because they are accessed only through the ExPASy server and could therefore serve to show the potential use of weblog analyses. To explore the usefulness of the proposed methodology further an additional 36 bioinformatic applications hosted on the ExPASy server were studied (Annex 1 provides a short description of each of the applications). Some of the applications to which the ExPASy server provides access (e.g. MARCOIL, pROC, PRATT, TMPred, TCS, T-Coffee, TagIdent, Swiss-PdbViewer, SwissParam, RAxML, PepPepSearch, PaxDb, OpenStructure,

[15] Reviews are included in addition to articles and for this reasons they were also included in our citation analysis.

[16] The Science Direct database contains over 2500 journals (primarily owned by Elsevier). Links on the following webpage provide information on coverage.

neXtProt, MyHits, MassSearch) were developed by other organisations, but they have also been analysed because they are hosted on the ExPASy server as well. Only thirteen of these 36 applications can be studied because of the limitations of the proposed approach. These thirteen are, apart from HAMAP and Swiss-2DPage: Msight; MIAPEGelDB; MALDIPepQuant; Make2D-DB II; HCD/CID Spectra merger; GlycosuiteDB; OpenStructure; MyHits; tagident; SwissParam; MARCOIL.

Results

We introduced an alternative measure for database use (see also Jonkers et al, 2012; Duin et al 2012,), which is independent of the academic literature. Table two shows that as expected the database which shows the highest usage intensity (in terms of the number of visits in the period 2003-2008) is also the database which is cited most frequently (PROSITE). Due to the small sample size we cannot do correlation analysis. But the data fit in the expected pattern, and the number of unique visitors is ten (HAMAP and Swiss 2DPAGE) to around thirty (PROSITE and Enzyme) times higher than the number of citations. More details about the existence and nature (linear or not) of the relationship cannot be derived from the available data.

Table 2 Citations (2003-2009) and visits (2003-2008)

	PROSITE	HAMAP	SWISS-2DPAGE	ENZYME
Citations in Scopus	2225	79	239	248
Visits	71890	914	3081	9194
Visits / citations	32	12	12.9	37
Log10 visits / log10 citations	1.45	1.56	1.1.49	1.66

Table 3 results data collection: citations and text mentions of the databases (2000-2011)

	PROSITE	HAMAP	SWISS-2DPAGE
Citations by articles/reviews all Scopus	4634	102	575
Citations in SD journals in Scopus	1000	16	52
Mentions in full text (minus references) of SD articles	1730	7	29
Mentions in full text without formal reference in Scopus	X	2	20
Total mentions + cites in SD journals in Scopus	X	18	72
Underrepresentation	X	11.1%	27.8%
Expected number of cites and mentions in entire Scopus	X	113	735

X: data not available

Table 3 presents a) the number of citations which were made to the source articles in which the four databases were introduced in Scopus between 2000 and 2011, b) in Science Direct Journals in Scopus in the same period. The table also includes the number of publications (articles and reviews from Science Direct journals) found through the full text section searches. It was expected that most of these mentions of acknowledged use would be found in the methods section, but this is certainly not exclusively so.

The second part of the analysis shows that the rate of underestimation found in the case of two of the four databases was 11.1% and 27.8% respectively. This indicates a) a substantial under-estimation of acknowledged use of e-research technologies through citation analyses and b) a considerable variation in the extent to which this underestimation occurs.

We find that 11 articles/reviews in Science Direct journals mention the HAMAP database in their full text. One of these is one of the original source articles, which leaves 10 after its exclusion. 7 of these have been published before 2012 and we decided to exclude this last year. The reason for doing so is that the online versions of the bibliometric databases used did not provide stable results for this year when measurements were made in the summer of 2012. Another motivation was that records for 2012 would not be complete as measurements were made before the end of this year. The total number of articles/reviews found in Scopus which cite one of the two source articles of HAMAP is 110, 102 of which were made in the years before 2012. Sixteen of these citations are made in Science Direct journals. Five of the ten articles which refer to the HAMAP database in the full text, do not cite either of the two HAMAP source articles. When excluding 2012, this is two out of seven. Some eighteen articles in Science Direct journals either cite one of the source articles of the HAMAP database, or mention it in the text. The total number of citations to the source articles in Science Direct journals is sixteen. Hence only a small underestimation of around 11% is found. As it is expected that citing behavior in other journals included in Scopus is similar to Elsevier journals, it is expected that there are around 113 articles/reviews which either cite HAMAP or refer to it in the text in the Scopus database.

A similar approach is followed to analyze the results from the citation and full text search for acknowledged use of the database SWISS-2DPAGE. 575 articles/reviews are found in Scopus which refer to one of the thirteen source articles. NEXTBIO finds 52 results in which Swiss-2DPage is found in the text (+ two false positives). 20 of these NEXTBIO results do not include a formal reference included in Scopus. The estimate for underestimation here is thus substantially higher at around 27.8%. Since authors publishing in Science Direct journals are expected to cite in a similar way as authors publishing in non-Science Direct Scopus journals, a total of 735 articles/reviews is expected to be present in the Scopus database that either cite the source articles of SWISS-2DPAGE, or mention the use of it in the text.

Considering the relatively large rate of underestimation of “acknowledged use” through formal citations, a manual analysis was performed of the articles that

mentioned Swiss 2DPAGE but did not cite any of the thirteen source articles. One expectation was that - as this database collects, stores and provides access to the empirical results of other studies - these 'non-citing' articles would refer to the underlying source articles instead. This, however, was not the case. Instead of including a formal reference, thirteen of these articles provided a URL to the Swiss-2DPAGE site. Two articles could not be accessed. Only five mentioned Swiss 2DPage in the text, while not presenting any acknowledgement (citation or URL) to their readers. [17]

Table 4 Underestimation of acknowledge usage by citation analysis for other ExPASy applications (2000-2011)

	Scopus cites	C	NEXTBIO	M	C+M	U (%)	U_1%
Quickmod	4	0	0	0	0	x	
MSight	81	12	5	3	15	20	4
MIAPEGelDB	7	1	0	0	1	0	0
MALDIPepQuant	5	2	0	0	2	0	0
Make2D-DB II	15	3	2	0	3	0	0
HCD/CID spectra merger	38	7	0	0	7	0	0
GlycoSuiteDB	120	17	4	1	18	6	1
FindPept *	45	13	31	28 (26)	41 (39)	68*(66)	
FindMod *	182	39	30	25 (23)	64 (62)	39* (37)	
PeptideMass*	175	59	91	59 (54)	118 (114)	50* (48)	
MARCOIL	101	20	12	1	21	5	1
T-coffee	2706	820	x	x	x	x	
tagident	16	0	26	26	26	100	61
Swiss-PdbViewer	5910		x	x	x	x	
SwissParam	2	1	0	0	1	0	0
RAxML	902	167	x	x	x	x	
PaxDb	0	0	0	0	0	x	
OpenStructure	3	0	0	0	0	0	0
MyHits	20	6	18	18	24	75	47

M = articles containing NEXTBIO in text references but no SD citations; C = Scopus cites included in Science Direct; *As mentioned in the methodological section the analysis for these four applications is incomplete and the real percentage of underestimation is therefore expected to be considerably lower.

Unfortunately the NEXTBIO software has some limitations, which makes it impossible to do the same analysis for the more popular PROSITE database. In contrast to the small numbers of articles in which HAMAP or SWISS-2Dpage

[17] For authors using bibliometric data it may be interesting that of the 518 SD publications that were found through NEXTBIO to mentioning the use of the Scopus databases in their full text, only 12 included the URL (though in some the URL may have been in the reference list).

were mentioned, a total of 1730 publications (in Science Direct journals) were found that mention PROSITE somewhere in the full text (minus the references). Unfortunately the software only shows a limited number of around 776 of these 1730 bibliographic references. It was therefore not possible to repeat the analysis conducted for the other databases. In total, the source articles in which the PROSITE database was introduced, received 4643 from publications included in Scopus. 1000 and 661 of these were made in Elsevier journals.

For Peptidecutter, Peppepsearch, NextProt and Masssearch an appropriate source article could not be identified. Some applications also had to be excluded such as compute pi/MW, sulfonator, myristoylator, blast, biochemical pathways, allall, pROC, PRATT and TCS because they gave too many unrelated hits due to name ambiguity similar to the "Enzyme database". Multiident received 153 Scopus citations and 28 SD citations. Given these numbers one would have expected a considerable number of in-text references as found through NEXTBIO. However none were found – though with the alternative spelling "multi-ident" one in-text reference was identified as well as five unrelated articles as the name was not sufficiently unambiguous. For this reason this application was also excluded from table 4.

The four applications Findpept, FindMod, PeptideMass and Peptidecutter have, apart from in the article analysed, also been introduced in a book chapter. The URLs giving access to these applications suggest this book chapter as a potential reference. This chapter, which is not included in Scopus and could therefore not be studied, received over 1400 Google scholar citations. Part of these citations is likely to have come from Scopus SD journals. This suggests that a considerable number of the articles with an in-text reference as found through NEXTBIO which were not found to have a corresponding SD citation may have included citations to the book chapter. While they are mentioned in the table, these results are therefore not considered reliable. For the applications Findpept, FindMod, and PeptideMass an alternative M was created through a manual search of the reference lists of these alternatives. Where a reference to the book chapter was found this was deducted from the original M and presented between brackets in the table. The rate of underrepresentation remains high, but would have been lower if it would have been possible to assess the number of Scopus cites (C) to the book chapter. As was the case for Swiss-2DPage part of this under-representation is caused because authors refer to the URL rather than including a formal citation.

Some applications such as Swiss-model, RaXML, Swiss-PDBviewer and T-coffee were too popular to be studied through this approach as was the case for the PROSITE database. They received 7707; 2706, 5910 and 902 Scopus citations respectively, but the in-text references yielded by NEXTBIO could not be analysed in detail. For the applications Glycanmass, Glycomod, GPSDB, PLcarber; protscale; protparam the suggested reference is the same general article. This article received 924 citations in Scopus and 204 citations in Science direct journals. However, some of the applications yielded too many NEXTBIO in text

references so that this "group" of applications could not be studied either as was the case for PROSITE. The reason why they are included in the table is that this helps to make an assessment of the relative share of SD citations in the total Scopus citation coverage in this field.
Some applications such as Pax-DB, OpenStructure, Quickmod, MIAPEgelDB and MALDIPepQuant and HCD/CID spectramerger, do not yield any in text references through the NextBio search. As a consequence the estimated rate of under-representation of acknowledged use is zero. One potential explanation is that some of these applications were introduced very recently and there has not yet been much time to cite these in either the references of articles or in the text. This reasoning lies behind the exclusion of PaxDb of which the source article was published in 2012, which is after the period in which the citations were measured.
The rate of underestimation of the other applications studied was, 5% for Marcoil, 6% for GlycoSuiteDB to around 20% for MSight. The underestimation of the acknowledge use of both MyHits (75%) and Tagident (100%) is high in comparison to the other applications as well as the databases studied in table 4. In the case of Tagident all citations were made in non SD journals. While there appears a strong underestimation of acknowledged use in the case of this application, in reality it can never be 100%, as the source article is referred to in non-SD journals. For this reason we adapt our indicator somewhat to provide a lower bandwith of the estimated underestimation (U_l). For this we take instead of "C" the number of Scopus citations. In the case of Tagident U_l is 61 %, in the case of Myhits it is 47 %, indicating that the underestimation of Tagident lies between at least 61 and 100% and the underestimation of Myhits lies between Myhits lies between at least 47 and 75%. According to this /(very conservative) estimate of underestimation the lower boundaries of the underestimation of HAMAP and Swiss-2Dpage would be 2 and 3 %.

Discussion and Conclusion

While citations appear systematically related with usage measured through unique visitors, it is not yet clear how these indicators are related. We find that a considerable share of the acknowledged use in research articles is not captured by citation analyses. The degree of underestimation varies between the databases and applications studied.
Both observations raise some concern over the accuracy, completeness and suitability of the sole use of citation analyses for measuring the impact of e-research infrastructures. This concern also potentially extends to other types of knowledge claims. The observed variations may be explained using existing citation theories. Publications that have already received a large number of citations may be more citable than those cited less, a derivation of the Matthew effect (Merton, 1995). Conversely, if the technology has become ubiquitous, researchers may consider that they no longer need to cite knowledge claims which have become "common knowledge". This echoes an argument made in Martin & Irvine (1983). A combination of these explanations might be used to explain the

observed relation between usage (as measured through weblog analysis) and citations. Neither of these explanations, however, can explain the variation in the rates of underestimation of acknowledged use through citation analysis between applications. It does appear from table 4 that the underestimation of young applications which have not yet received a substantial number of citations tends to be zero.

A more in depth exploration of the instances of acknowledged use that were not reflected in citations for the case of the Swiss 2Dpage database, revealed that in a large share of these instances, the authors had referred to the URL that provided access to the application in either the reference list or inside the text. This type of acknowledgement is more difficult to analyse than formal citations, but it may nonetheless be a common way for researchers to refer to electronic databases and applications.

The approaches highlighted in this paper: 1) "web usage statistics derived from the analysis of web logs", 2) "citation analyses" and 3) "the analysis of in-text references to specific research infrastructures" do not provide a complete insight in the actual scholarly usage of e-research infrastructures and their impact. Not all usage will be acknowledged by researchers in the reference list or as in-text reference. Furthermore, researchers may also be using technologies without being fully aware of it. A discussion of the HAMAP database studied in this paper will serve to explain this. It is important to realize that there is a difference between 1) first order users, who make direct use of, for example, the HAMAP rule book and 2) second order users who, while not making use of the rule book or HAMAP database, do make use of the information of HAMAP annotated proteins contained in other protein databases. When referring to usage, this paper only referred to the first order users. However it is important to realize that the actual use and impact of such technologies may be extended beyond its direct use.

This is one of the first articles that introduce an (exploratory) comparative analysis of in text reference analysis and citation analysis. The main part of the analysis is limited to journals included in the Science Direct database. It is clear that the proposed approach to the analysis of in-text references through the use of NEXTBIO has its limitations: especially with reference to name-ambiguity and popular applications. The second limitation can probably be solved relatively easily through alternative approaches to the analysis of in-text referencing. The first limitation is more difficult to solve. In the case of Tagident the underestimation appears to be 100 %. This is not an accurate reflection of reality since citations have been made in non-SD journals. This suggests a weakness of the proposed approach when dealing with applications which had still received only a very small number of citations in the period measured. In-text reference analysis which is not restricted to SD journals will not face this problem.

Analysts have argued that it is somehow "unfair" to compare citations to reviews with those to theoretical or empirical papers. Some may argue that this argument can be extended to publications introducing new methods, research instruments or research infrastructures. Normalisation is often used to account for differences in

the average frequency of citation to different document types (Moed et al, 1995, Rehn & Kronman, 2008). Due to the structure of the bibliometric databases methodological papers, papers introducing research instruments or research infrastructures are normally not identified as such. Therefore they are also not normally subjected to such normalisations. Furthermore a complete theoretical justification for assigning a different value to citations received by different document types is still lacking. The differential underestimation of "acknowledged use" via citation measurement might provide part of such a justification if the rate of under-acknowledgement differs systematically between types of knowledge claims. In this paper an indication is found that citation analysis underestimates the acknowledged use of some types of knowledge claims (in this case biological databases). Further analysis of the varying degree of underestimation of different knowledge claim types could provide a way forward to a more complete justification for both citation normalisation and/or the use of alternative metrics in assessing the impact of different knowledge claim types. As highlighted in a recent Nature materials editorial (2012), the merit of the latter should be evaluated with care for: "Not everything that can be counted counts and not everything that counts can be counted". This oft used and paraphrased quote is sometimes attributed to Cameron (1963), but often also to Albert Einstein's blackboard writing.

Acknowledgements

The Spanish Ministry of Economics and Competitiveness funded the project of which this paper forms part through the grant: CSO2011-23508. The research of the third author was supported by the Juan de la Cierva (JDC)-MICINN program of the same Ministry. SIB Swiss Institute of Bioinformatics allowed for the use of the server web log data used for part of this analysis. We would also like to thank Felix de Moya Anegón for introducing us to the NEXTBIO application "section search" and Isidro F Aguillo for advice on the use of Quest's Funnelweb software. Researchers at the Centre for Science and Technology Studies of Leiden University (NL) provided stimulating ideas in discussions during a research stay of one of the authors. The usual disclaimer applies.

References

Ball, R., Tunger, D. (2006) Science indicators revisited - Science Citation Index versus Scopus: A bibliometric comparison of both citation databases, *Journal Information Services and Use*, 26: 293-301

Bairoch A. (2000) *The ENZYME database in 2000*, *Nucleic Acids Res* 28:304-305.

Bornmann, L. & Daniel, HD (2008) What do citation counts measure? A review of studies on citing behavior, *Journal of Documentation,* 64 (1): 45-80

Cameron, WB. (1963) *Informal Sociology: A Casual Introduction to Sociological Thinking*, New York, Random House

Cozzens, SE (1989) What do Citations count? The rhetoric-first model, *Scientometrics*, 15 (5-6): 437-447

Cronin, B. (1984) *The Citation Process. The Role and Significance of Citations in Scientific Communication*, Taylor Graham, Oxford.

De Jong, S., van Arensbergen, P. Daemen, F., van der Meulen, B., van den Besselaar, P. (2011) Evaluating research in its context: an approach and two cases. *Research Evaluation* 20 (1): 61-72.

De Solla Price, D. (1976) A General Theory of Bibliometric and Other Cumulative Advantage Processes, *Journal of the American Society for Information Science*, 27 (5-6): 292-306 1976

Duin, D., King, D., van den Besselaar, P. (2012) Identifying Audiences of E-Infrastructures - Tools for Measuring Impact, *PLOS ONE*, 7(12)

Editorial (2012) Alternative metrics, *Nature Materials,* 11, 907

Fingerman, S. (2006) Web of Science and Scopus: Current Features and Capabilities, *Issues in Science and Technology Librarianship*, DOI:10.5062/F4G44N7B

Garfield, E. (1998) Random thoughts on citationology, its theory and practice, *Scientometrics*, 43 (1): 69-76

Gilbert, N. (1977) Referencing as Persuasion, *Social Studies of Science*, 7 (1)

Gorraiz, J., & Schlögl, C. (2007) Comparison of two counting houses in the field of pharmacology and pharmacy. In *Proceedings of the international conference of the international society for scientometrics and informetrics,* 11: 854–855.

Hoogland C., Mostaguir K., Sanchez J.-C., Hochstrasser D.F., Appel R.D. (2004) SWISS-2DPAGE, ten years later. *Proteomics*, 4(8), 2352-2356.

Jacso, P. (2006) Evaluation of citation enhanced scholarly databases. *Journal of Information Processing and Management*, *48*(12), 763–774.

Jonkers, K., De Moya Anegon, F., Aguillo, F. (2012) Measuring the use of research infrastructures as an indicator of research activity, *Journal of the American Society of Information Science and Technology*, 63 (7): 1374–1382

Lima, T., Auchincloss, AH., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L., Bairoch, A. (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot, *Nucl. Acids Res.* 37 (suppl 1): D471-D478. doi: 10.1093/nar/gkn661

López-Illescas, C., Moya-Anegón, F., Moed, HF. (2008) Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of Informetrics*, *2 304–316*

Lowry, OH., Rosebrough, NJ, Farr, AL, Randall, RJ. (1951) Protein Measurement with the Folin Phenol Reagent, *Journal of Biological Chemistry*, 193:265-275 .

Martin, BR., Irvine, J. (1983) Assessing basic research: Some partial indicators of scientific progress in radio astronomy, *Research Policy*, 12 (2): 61–90

Merton, RK (1995) The Thomas Theorem and the Matthew Effect, *Social Forces*, 74 (2)

Moed, H.F.; Colledge, L.; Reedijk, J.;Moya-Anegón, F.; Guerrero-Bote, V.; Plume, A.; Amin, M. (2012**)** Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way**.** *Scientometrics*, 92 (2) 367-376**.**

Moed, H.F., De Bruin, RE., Van Leeuwen, TN (1995) New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications, *Scientometrics* 33 (3): 381-422

Moya-Anegon, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A. (2007) Coverage analysis of Scopus: A journal metric approach. *Scientometrics*, 73(1), 53–78.

NEXTBIO (2012) section search, retrieved from http://www.applications.sciverse.com/action/appDetail/293416

Quest*. (*2010*). Funnel Web Analyzer®—overview. Retrieved from http://www.quest.com/funnel-web-analyzer/index.asp*

Rehn, C., & Kronman, U. (2008). *Bibliometric handbook for Karolinska Institutet V1.05* http://ki.se/content/1/c6/01/79/31/bibliometric_handbook_karolinska_institutet_v_1.05.pdf.

Senker, J. (1995) Tacit knowledge and Models of Innovation, Industrial and Corporate Change,

Scopus (2012) http://www.scopus.com

Science direct journal coverage (2012) http://www.info.sciverse.com/sciencedirect/content/journals/titles

Sigrist C.J.A., de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. (2012) *New and continuing developments at PROSITE*, *Nucleic Acids Res.* doi: 10.1093/nar/gks1067

Thomson Reuters (2012) http://thomsonreuters.com

Van Raan, A. (2005) Measuring Science, in Moed, HF., Glänzel, W., Schmoch, U., *Handbook of Quantitative Science and Technology Studies,* Dordrecht: Springer

Weingart, P. (2005) Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1): 117-131

ARE LARGER EFFECT SIZES IN EXPERIMENTAL STUDIES GOOD PREDICTORS OF HIGHER CITATION RATES? A BAYESIAN EXAMINATION.

Jesper W. Schneider[1] and Dorte Henriksen[2]

[1] *jws@cfa.au.dk*
Danish Centre for Studies in Research and Research Policy, Department of Political Science and Government, Aarhus University, Bartholins Álle 7, DK-8000 Aarhus C (Denmark)

[2] *dh@cfa.au.dk*
Danish Centre for Studies in Research and Research Policy, Department of Political Science and Government, Aarhus University, Bartholins Álle 7, DK-8000 Aarhus C (Denmark)

Abstract
Effect sizes are perhaps the most important quantitative information in statistical inferential studies. Recently, the hypothesis that rational citation behaviour in general ought to give credit to studies that successfully apply a treatment and detect greater effects, resulting in such studies being cited more frequently among comparable studies. Hence, it is predicted that larger effect sizes increases study relative citation rates.
Two recent studies in biology provide contradictory results on this hypothesis. The present study investigates the same hypothesis but in different research areas and with a more credible model selection procedure.
Using meta-analyses, we identify comparable individual experimental studies (n=259) from five different research specialties. Effect sizes are compared to the citation rates of the individual studies and impact factors for the journals where the studies are published. Contrary to the previous findings, and in fact most studies in scientometrics, we examine the hypothesis with a Bayesian model selection procedure. This is advantageous, as we thereby are able to quantify the statistical evidence for both hypotheses, H_0 and H_1. This is not possible in classical statistical inference, though the implicit inferential decision made by most researchers when they fail to reject H_0 is to accept it. This is a flawed logic. Given uniform priors for the two hypotheses, the result from the present data set is posterior odds of 13/4 to 1 in favor of the null models examined. Consequently, the study give positive evidence to the claim made by Lortie et al. (forthcoming) that effect sizes do not predict citation rates and are as such poor proxies for the quantitative merit of a given experimental treatment.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Sociological and Philosophical Issues and Applications (Topic 13)

Introduction

In a forthcoming study, Lortie et al. (doi: 10.1007/s11192-012-0822-6) hypothesize that if citation behavior is supposed to be rationale, articles that report larger biological effect sizes from successful treatments in ecology and evolutionary biology studies, should generally also have higher relative citation rates. The hypothesis is apparently not supported by their data and the conclusion is that citations are a poor proxy for quantitative merit of a given treatment in ecology and evolutionary biology. A similar hypothesis, also from a sample of ecology studies, is investigated by Barto and Rillig (2012). Contrary, to Lortie et al., Barto and Rillig (2012) do identify a positive relationship between effect size and citation rates. The importance of effect size in reference behavior was also previously indicated in a survey by Shadish et al. (1995). The hypothesis is interesting, though not without problems, and warrants replication for other research areas. In this study, we examine the hypothesis in five different research specialties (in psychiatry, clinical psychology, brain research, psychotherapy and educational research) in order to further examine if and how the magnitude of effect sizes and citation rates are related.

The general hypothesis has some merit. It seems reasonable to assume that rational reference behavior in quantitative experimental domains would entail that in specialized research areas, studies that, *ceteris paribus*, demonstrate larger effect sizes will also generally be more cited. The notion of empirical science being cumulative warrants such an assumption. Also, higher impact journals should on average publish studies with larger effect sizes if they do indeed differentiate for stronger evidence (Song, Eastwood & Gilbody, 2000). At face value, effect sizes are very important in quantitative studies that rely on statistical inference. It is well known that statistical significance tests are flawed, seriously misused and misinterpreted (e.g., Berkson, 1942; Oakes, 1986; Cohen, 1994; Nickerson, 2000; Kline, 2004; and for a scientometric perspective Schneider, 2013). *P* values cannot quantity the importance of a result, but effect sizes with confidence limits can (e.g., Goodman, 1999a; Goodman, 2008; Ellis, 2010; Cumming, 2012). Reporting effect sizes are also important for meta-analytic purposes. The latter basically serves as a formal tool of evidence, where effect sizes from comparable studies are evaluated statistically. Notice, the latter is certainly not without its problems (Berk & Freedman, 2003; Berk, 2007).

However, a straightforward relation between the magnitudes of effect sizes and reference behavior is doubtful. The question is whether effect sizes alone are sufficient to warrant a reference. For example, often large effect sizes (relatively to the phenomenon studied) are reported in the earliest studies within a domain (Barto & Rillig, 2012). Often such findings cannot be replicated and the subsequent effect sizes become more moderate. Also, samples size and quality of the study design are crucial elements in relation to effect sizes and their reference potential. Large effect sizes from a non-experimental study with a relative small sample size are generally considered less robust and causally inferior and thus

have *de facto* lower evidence. Consequently, other rational epistemic factors may be of more importance to the citing author when he or she decides to reference an experimental study. Clearly, references are given (or not given) for a whole number of reasons, some rational and sound, others haphazardly or perfunctory, and still others suspicious, self-promoting and political, and citations are perceived differently among researchers (for overviews see for example Bornmann & Daniel, 2008; Aksnes & Rip, 2009). At the same time, numerous studies have tried to identify citation predictors for articles in restricted settings (van Dalen & Henkens, 2005; Stremersch, Verniers & Verhoef, 2007; Mingers & Xu, 2010). Common for many of these studies is that their model specification and subsequent fitting procedure is done on the same data set. Further, a preponderance of the proxy variables specified and tested seems to be easily quantifiable document attributes from the bibliographic records retrieved from a citation database. Hence, what we are left with is an ordinal knowledge base about the potential influence - on average - upon citations to articles from indicators such as journal status, document type, number of authors and similar proxies. The meaning and validity of the proxies are seldom discussed. The influence of more cognitive aspects of documents, i.e., the content that ought to stimulate reactions from peers, whether positive, neutral or negative, is not well established quantitatively. Clearly more effort is needed to analyze cognitive and epistemic patterns relating to reference behaviors. In that respect, effect sizes in quantitative experimental studies are interesting. The aim of experimental studies is to investigate treatment effects and the most important quantitative entity when reporting the results is the estimated effect size (standardized or non-standardized) and its margin of error.

Consequently, this study further examines the hypothesis that somehow effect sizes ought to influence citation rates and show a relation to journal impact factors. Contrary to other studies, we take a Bayesian approach, where we provide statistical evidence for the hypotheses investigated. The next section explains the methods and materials used, including our Bayesian perspective; subsequently we report on our results, and end with a discussion of the results.

Methods and materials

We basically follow the same data collection strategy as Lortie et al. Our aim in this study is to explore whether Lortie et al.'s claims are discernible in other domains, or alternatively, to find support for the claims by Barto and Rillig (2012). Hence, we have not set-up a strict data collection procedure for a specific domain. An initial search was conducted in Thompson Reuters' *Web of Science*© (WoS) with various forms of the term 'meta-analysis'. The result was restricted to meta-analyses published from 2003-2012. From the large set of meta-analyses identified (≈ 26.000), five was chosen based on the following inclusion criteria 1) a random selection procedure selected five different WoS subject categories; 2) 25 meta-analyses were randomly selected within each of the categories; 3) these meta-analyses were scanned to see if they reported individual standardized effect

sizes as well as sample sizes for the studies analyzed. Among those meta-analyses eligible, only the ones where all studies analyzed were experiments with random procedures was selected in order to have some control of the study design quality. Among these, one meta-analysis was chosen randomly for each of the five categories resulting in 259 individual effect sizes (i.e., the five selected meta-analyses are Willcutt et al., 2005; Sommer et al., 2008; Beck et al., 2012; Furtak et al., 2012; Oldham et al., 2012). Effect sizes from the individual studies (e.g., Glass'Δ, Hedges'*g* and Pearson's *r*) were transformed to one scale Cohen's *d*. In studies where multi-effect sizes were reported only the largest reported effect size was included, effectively favoring the hypothesis investigated. Citation statistics for each of the 259 studies were obtained from WoS, as well as 5-year Impact Factors (JIF) from journals where the studies were published. JIFs were calculated so that they matched the years immediately after publication of the study.

Though random elements are used in the selection process, the overall sampling frame cannot be considered a probability sample. However, the sampling frame ensures that we can analyze fairly homogenous studies across domains. Firstly, the choice of meta-analyses as pointers to individual studies ensures that we identify a restricted set of articles that presumably study the same phenomenon often with similar approaches (i.e., the meta-analysis has already enforced strict inclusion criteria); secondly, reporting of standardized effect sizes entail a common scale so that comparison of effect sizes across domains is possible. The requirement that sample sizes should be reported (i.e., not just shown as confidence limits) enable us to control for sample size when predicting the potential influence of effect sizes on citation rates (i.e., large sample size, *ceteris paribus*, produce more stable effect sizes).

Contrary to Lortie et al., as data are continuous, we apply simple OLS as our primary models to explore the hypothesized positive linear trend between effect sizes and citation rates of individual studies, as well as impact factors at the journal level, i.e., larger effect sizes tend to be published in higher impact journals (individual article citation rates and 5-year JIFs are log-transformed). To mimic Lortie et al., we also specified Poisson models. It may be reasonable to model citation rates as counts, despite the fact that data are continuous, since $y \geq 0$ (e.g., Wooldridge, 2002). Even so, the GLM models provide the same interpretations as the logged-*y* OLS-models, but with less convincing diagnostics.

The individual studies (n = 259) are collapsed into one sample, as sensitivity analyses revealed no discernable effects relevant for this study. For example, field normalization with logarithm-based citation z-scores (Lundberg, 2007) does not alter the pattern of relations compared to simple mean annual citation scores when all studies are collapsed. Likewise, there was no discernible difference when using mean annual citation scores for all years versus a 5-year period after publication. What is of importance is whether higher effect sizes tend to

influence citation rates, minor differences in general citation activity between domains does not affect this aim.

Bayesian hypothesis testing

Contrary to most studies in scientometrics and the social sciences, we take a Bayesian approach to statistical evidence. Inference by *p* values, in the frequentist amalgam "null hypothesis significance testing" (NHST), is nearly ubiquitous despite longstanding serious criticisms concerning its logical flaws, rote use, misunderstandings and misuses (see some good introductory references in the introduction section out of literally hundreds). Two critical issues are important for this study. NHST does not allow researchers to state evidence *for* the null hypothesis (e.g., Hubbard & Lindsay, 2008), nevertheless, this is more or less the implicit inferential decision made by most researchers when they fail to reject H_0; as an example, Lortie et al., base their claims of no effect on the failure to reject H_0. Further, it has been clearly demonstrated that *p* values themselves overstate the evidence against the null hypothesis, i.e., a rejection of H_0, especially in the *p*-interval from .05 to .01 (Jeffreys, 1961; Berger & Sellke, 1987; Goodman, 1999b; Sellke, Bayarri & Berger, 2001).

Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995) have been advocated as superior to *p* values for assessing statistical evidence in data (Edwards, Lindman & Savage, 1963; Raftery, 1995; Wagenmakers, 2007; Rouder et al., 2009). We entirely concur with this claim. The Bayes factor computes the probability of the observed data under H_0 *vis-a-vis* H_1. Notice, in contrast to the frequentist *p* value, the Bayes factor allows researchers to quantify evidence in favor of H_0. In the Bayesian model selection procedure, the ultimate objective is to compute a probability reflecting which model is more likely to be correct, on the basis of the obtained data and the core concept is Bayes' theorem.

We use Bayes factors as the model selection procedure in this study. The two models examined are H_1, that effect sizes predict citation rates, against H_0 of no or a minuscule relation. Although the Bayes factor is conceptually straightforward, its use is not widespread in the social sciences. Bayesian models require specification of priors. Like, NHST it is uncomplicated to calculate $p(D|H_0)$, however, H_1 does not specify one particular a priori value for the effect in question. Rather, H_1 is associated with a distribution of possible effect sizes, and the value of the Bayes factor depends on the nature of that distribution. Therefore, exact computation of the Bayes factor quickly becomes complex, involving integration over the space of possible effect sizes using procedures such as Markov chain Monte Carlo methods. This is complicated and no general commercial software package enables Bayesian modeling.

In this study we apply a more practical alternative suggested by Raftery (1995) and Wagenmakers (2007), where we approximate the Bayes factor using the Bayesian Information Criterion (BIC) (Raftery, 1995). BIC is often used to quantify the goodness of fit of a model to data, accounting for the number of free

parameters in the model. BIC is easy to compute and for some models, popular statistical computer programs already provide the raw BIC numbers, so that in order to perform an approximate Bayesian hypothesis test, one only needs to transform BIC values for two competing models, H_0 and H_1, to posterior probabilities (for details, see e.g., Raftery, 1995; Glover & Dixon, 2004; Wagenmakers, 2007).

Some assumptions and limitations are in order. Obviously, the Bayes factor is sensitive to the shape of the prior distribution, but the use of BIC does not require the researcher to specify his or her own prior distribution. This is appealing, but also the main drawback of using BIC. BIC implicitly assumes the *unit information prior* (Kass & Wasserman, 1995) and it has been argued that this prior is too wide, resulting in a decrease of the prior predictive probability of H_1, and therefore makes H_0 appear more plausible than it actually is. In this sense, the BIC estimate of the Bayesian posterior probabilities should be considered somewhat conservative with respect to providing evidence for the alternative hypothesis (Raftery, 1999). Thus, the drawback of the BIC is that it does not incorporate substantive information into its implicit prior distribution; the virtue of the BIC is that the specification of the prior distribution is completely automatic. Another limitation of BIC is that its approximation ignores the functional form of the model parameters, focusing exclusively on the number of free parameters. A full-blown Bayesian analysis is sensitive to the functional form of the parameters because it averages the likelihood across the entire parameter space. Although the issue of functional form is important, it is much more important in complicated nonlinear models than it is in standard linear statistical models.

The Bayes factor plays a crucial role in establishing the relative evidential support for H_0 and H_1. The Bayes factor (BF) can be estimated using the following transformation of the difference in BIC values for two competing models:

$$\mathbf{BF} \approx \frac{\boldsymbol{p}\mathbf{BIC}(\mathbf{D}|\mathbf{H_0})}{\boldsymbol{p}\mathbf{BIC}(\mathbf{D}|\mathbf{H_1})} = \boldsymbol{e}^{(\Delta \mathbf{BIC}/\mathbf{2})} \quad (1)$$

where $\Delta BIC = BIC(H_1) - BIC(H_0)$. The resulting estimate of the Bayes factor yields the odds favoring the null hypothesis, relative to the alternative hypothesis. BF can then be converted to the posterior probability that the data favor the null hypothesis as follows (assuming equal priors):

$$p\mathrm{BIC}(H_0|D) = \frac{\mathbf{BF}}{\mathbf{BF+1}} \quad (2)$$

With only two competing models, the posterior probability that the data favor the alternative hypothesis is just the complement of Equation 2:

$$p\mathrm{BIC}(H_1|D) = 1 - p\mathrm{BIC}(H_0|D) \quad (3)$$

In the present study we use total sum of squares and the sum of squares of the error term to derive $BIC(H_1)$ and $BIC(H_0)$ (e.g., Wagenmakers, 2007). Finally,

based on Jeffreys' (1961) rules of thumb for interpreting Bayes factors, Raftery (1995) has provided descriptive terms for strength of evidence as follows: pBIC(H_i|D) .50-.75 (weak), .75-.95 (positive), .95-.99 (strong), and >.99 (very strong).

Results

First we present some figures that explore the data set. Figure 1 below shows the distribution of standardized mean effect sizes for the individual studies in the five meta-analyses (MA1, MA2, MA3, MA4 and MA5).

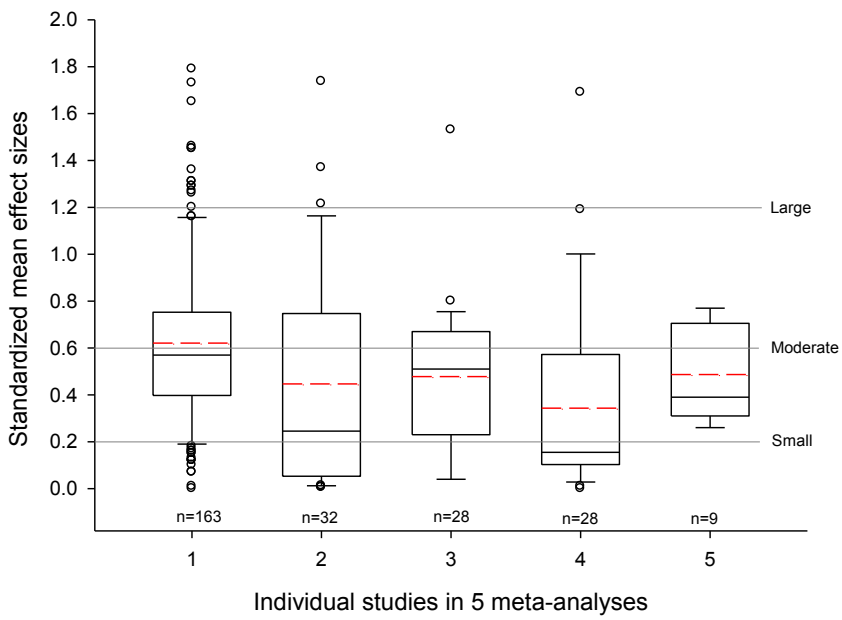


Figure 1. Box plot of standardized mean effect sizes in individual studies reported in 5 meta-analyses (n=259). Solid lines in boxes show median effect sizes and dotted lines average effect sizes for studies included in the meta-analyses.

If we apply Cohen's reference categories for interpreting effect sizes (Cohen, 1988), we can see that median effect sizes for all but one meta-analysis (MA4) can be considered small, whereas MA4's is trivial. We also see that MA1, with its large n come closest to a Gaussian distribution, whereas the other meta-analyses show considerable skewness. Three meta-analyses have rather long whiskers (at the high end) and four meta-analyses have outliers, which corresponds to large effect sizes.

Figure 2 below shows the distribution of citation scores for the five meta-analyses. MA1 and MA4 have the largest median "mean annual citation scores", 6.5 and 5.1 respectively. The other three meta-analyses have considerable lower citation activity. All distributions are skewed, but skewness for MA1 is considerably lower than the others (except MA5, but *n* here is only 9 and scarcely robust). We see some marked outliers in MA2 and MA4; the outlier in MA4 is an article published in *Nature* with an annual mean citation score of 38.2.

Figure 3 below is a plot of mean annual citation scores for individual studies, as well as journal impact factors, as a function of the magnitude of effect sizes. It is clear that the concentration of observations is in the reference categories trivial (*n*=52), small (*n*=98) and medium (*n*=91). The outlier on the border to the large category is the *Nature* article (high annual citation scores and obviously a high JIF).

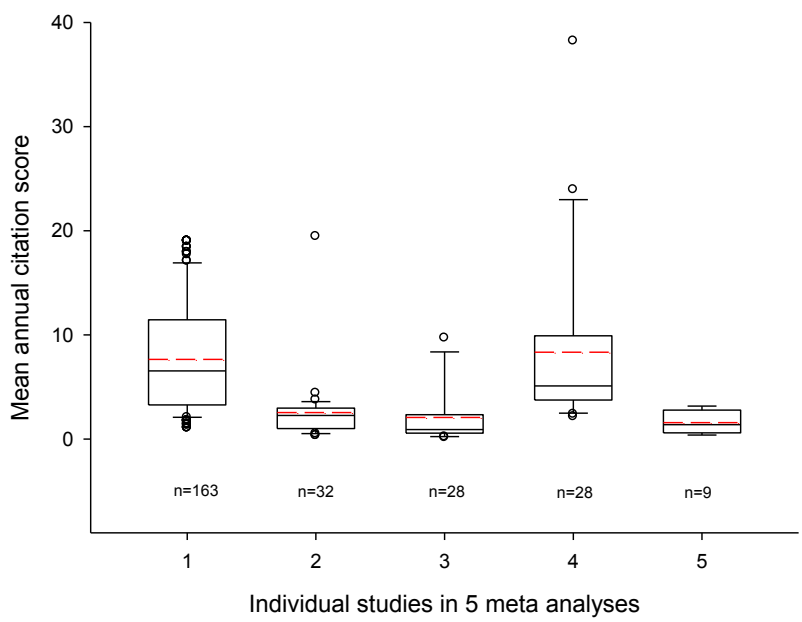


Figure 2. Box plot of mean annual citation scores for individual studies reported in 5 meta-analyses (*n*=259). Solid lines in boxes show median citation scores and dotted lines average citation scores for the studies included in the meta-analysis.

Finally, we group the individual experimental studies according to their reference category as defined by Cohen (1988) and plot this against mean annual citation scores as illustrated in Figure 4 above. This box plot reveals almost identical central tendencies in citation activity across the four reference groups. If the

predicted hypothesis of a linear trend was true, then the boxes should be staggered so that the trivial box was at the bottom, followed by the small and medium boxes, ending with the large box at the top. Clearly this is not so, hence we can expect support for the null hypothesis.

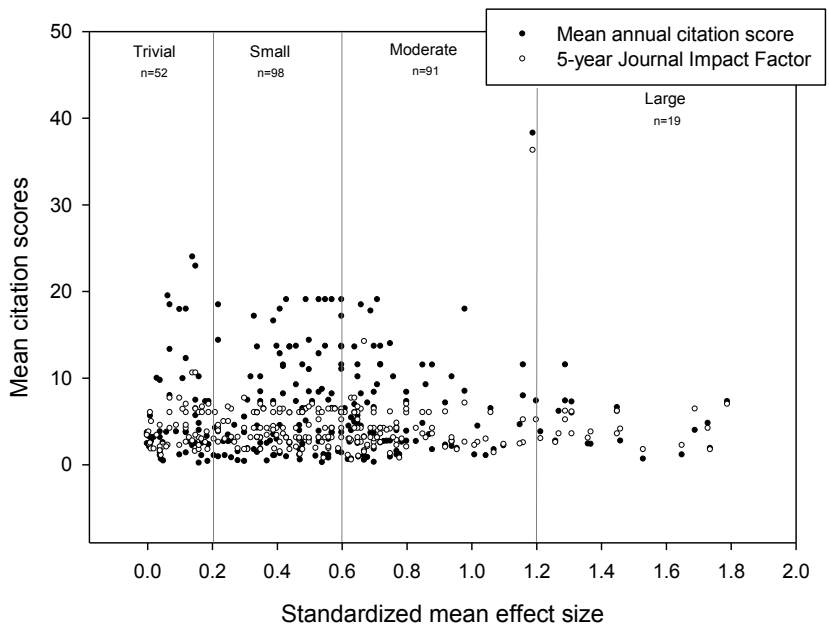


Figure 3. Plot of mean citation scores (mean annual citation scores for individual studies and journal impact factors) as a function effect sizes (N = 259). Vertical grey lines show Cohen's reference categories for interpretation of effect sizes and *n* indicate the number of studies in each reference.

Table 1. Scatter matrix of relationships between effect size, mean citation score, journal impact factor and sample size.

	Mean annual citation Score	5-year journal impact factor	Sample size
Effect size	-0.001	0.072	-0.296
Mean annual citation Score		0.456	0.253
5-year journal impact factor			0.065

In Table 1 below, we report Pearson correlation coefficients between effect sizes, mean annual citation scores, 5-year journal impact factors and sample sizes; rank correlations give similar correlations.
As one would suspect from inspecting Figure 3 and 4, there are close to no linear relation between effect sizes and citation rates for individual articles or JIFs from the journals where these articles are published. However, there is a negative relation between effect size and sample size. The relation is moderate and not surprising. To some extent larger sample sizes in studies result in relatively lower effect sizes. Large sample sizes reduce variability and thus provide more stable effect sizes. Citation rates and JIFs are also moderately correlated, though this is uninteresting in this context. What is more important is that citation rates and sample size have a small correlation. This may indicate that to some vague degree citing authors are aware of the importance of sample size for the robustness of results.

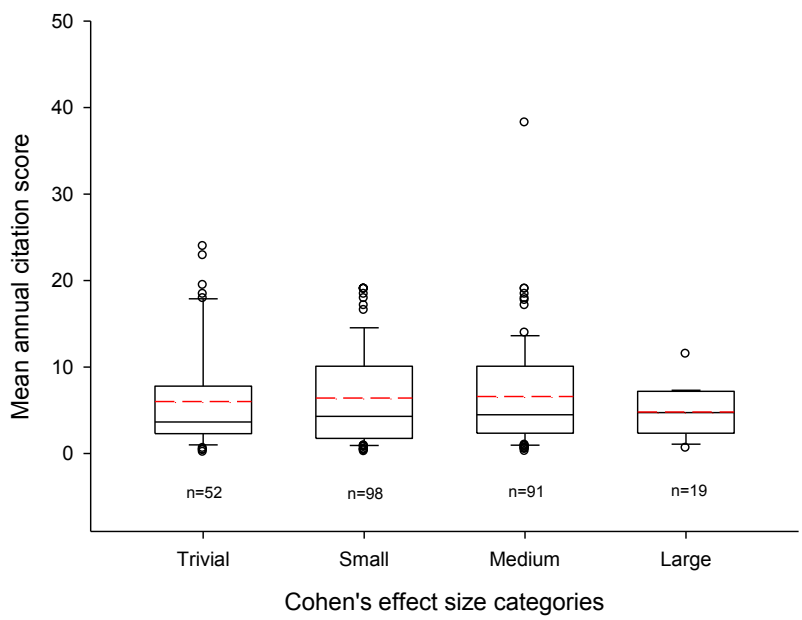


Figure 4. Box plot of mean annual citation scores distributed according to Cohen's reference categories for interpretation of effect sizes. Solid lines in boxes show median citation scores and dotted lines average citation scores for the studies included in the reference category.

Like Lortie et al., we use a simple model with one predictor (effect size). However, unlike Lortie et al. we apply Bayes factors to assess the evidence for the two competing hypotheses. Unlike *p* values and NHST, we are therefore able to

quantify the evidence *for* H_0. The exploratory data analysis has already indicated that we should expect a slope close to zero (H_0). Model 1, where effect sizes should predict log-transformed mean citation scores, results in a Bayes factor of 12.9, which, with equal priors, gives a posterior probability for the null model of $p(H_0|D)$ =.93 versus $p(H_1|D)$ =.07 for the linear model. The result qualifies in Raftery's (1995) descriptive terms as positive evidence in favor of the null hypothesis. Model 2, where effect size should predict log-transformed journal impact factors, results in a Bayes factor of 13.7, which, with equal priors, also gives a posterior probability for the null model of $p(H_0|D)$ =.93 versus $p(H_1|D)$ =.07 for the linear model, also qualifying as positive evidence in favor of the null hypothesis. The result is clear, with the given equal priors, we have positive evidence, approximately 13/4 to 1, that the data are clearly most probable under the null model.

As the data analyses suggest, a model specification where sample size and journal impact factor act as controls, brings nothing. Likewise, controlling for potential differences between studies, brings nothing. Consequently, effect size is no predictor of citation rates in the present data set. Of curiosity, a model where sample size is a predictor of log-transformed citation rates yields a Bayes factor of 2.5 in favor of H_0 and posterior probabilities of $p(H_0|D)$ =.72 versus $p(H_1|D)$ =.28. The *F*-test for the model is .0558; some would declare this statistical significant at the 10% level, others will have a hard time explaining why .0499 means a statistical significant model, whereas .0558 does not. But all will fail to appreciate that the evidence against the null hypothesis is only .28 and that the odds in fact favors H_0. This is an example of the Lindley paradox (Lindley, 1957) where *p* values overstate the evidence against H_0.

Discussion

The present study supports the overall claims by Lortie et al. that the effect size of a given study in general does not directly predict its subsequent citation rate, and at an aggregated level, populations of effect sizes associated with journals did not predict the impact of journals. The findings therefore suggest that for the present data set, across five research domains, citing authors do not generally use effect sizes of a given study directly when they find primary motivation for citing an experimental study. Other epistemic factors play a role, one of them may be sample size. Considering, the numerous factors and motivations suggested that may influence citing behavior, it is perhaps not surprising that effect size alone is no good predictor of citation rates. Other studies have for example shown that in some fields studies were cited more often when results were statistically significant (Kjaergard & Gluud, 2002; Leimu & Koricheva, 2005; Etter & Stapleton, 2009). A fact Lortie et al. also stress from their findings. Nevertheless, the *de facto* zero correlation and the clear positive quantitative evidence supporting H_0 are surprising. Usually in the social sciences, we can detect the "crud factor", i.e., that "everything is related to everything else", with some reasonable samples size (Meehl, 1990). In the case of effect sizes and

citation rates in the present data set this is apparently not the case and this is surprising.

Contrary to Lortie et al., the present study is able to present numerical evidence *for* the null hypothesis (as well as the alternative hypothesis). Given equal priors, both null hypotheses are supported with approximate odds of 13/4 to 1. We find the Bayes factor superior to *p* values for assessing statistical evidence and as such our result is important. While an inspection of Lortie et al.'s Figure 1 may indicate their claim of no relation, their ritual inferential procedure is faulty. *P* values are conditional probabilities of the data given the null hypothesis and they cannot provide support for a null hypothesis, as Fisher himself pointed out (Fisher, 1934). Given that contradictory claims were present in the literature, we find it reasonable to commence our exploration with a uniform prior. Two apparent Bayesian opportunities arise from these results, further studies with uniform priors that can confirm the present findings and/or a move to a full-blown Bayesian analysis where the current finding can be used to inform the priors, meaning that H_0 should have a higher prior probability compared to H_1 and a spectrum of different priors should then be analyzed.

In the current data set, we need to investigate what may be the primary reasons for citing authors to give references to the highly cited articles, now that effect sizes apparently seems not to be a principal reason, even though they could be, given their epistemic importance in experimental studies. The issue is essential because it touches upon the question of citations' relation to research quality (aka importance). If a citation network depicts the temporal and cumulative nature of science, it is reasonably to imagine that the highly cited articles in the network, for a large part, are important nodes, where importance *also* embraces the explicit quantitative statements about the phenomenon under study such as effect sizes. If this generally seems not to be the case in the social, behavioral and medical sciences, we clearly need to examine what then characterizes these fields as cumulative when it comes to citations. Instead of positing novel far-fetched models to investigate potential citation predictors among these highly cited articles in this study, citation context analysis may be more fruitful for this restricted purpose.

Finally, it is important to examine whether in the long run, meta-analyses, with their aggregated effect sizes, will eventually be more cited on average compared to the individual studies they set out to evaluate. This may not be a foregone conclusion, as inclusion criteria, comparability of studies and aggregated effect sizes are controversial issues in the debate about meta-analyses and the purported evidence they claim.

References

Aksnes, D. W., & Rip, A. (2009). Researchers' perceptions of citations. *Research Policy, 38*(6), 895-905.

Barto, E. K., & Rillig, M. C. (2012). Dissemination biases in ecology: Effect sizes matter more than quality. *Oikos, 121*(2), 228-235.

Beck, N. N., Johannsen, M., Stoving, R. K., Mehlsen, M., & Zachariae, R. (2012). Do postoperative psychotherapeutic interventions and support groups influence weight loss following bariatric surgery? A systematic review and meta-analysis of randomized and nonrandomized trials. *Obesity Surgery, 22*(11), 1790-1797.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis - the irreconcilability of p-values and evidence. *Journal of the American Statistical Association, 82*(397), 112-122.

Berk, R. A. (2007). Statistical inference and meta-analysis. *Journal of Experimental Criminology, 3*(3), 247-270.

Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg & S. Cohen (Eds.), *Punishment and social control* (pp. 235-254). New York: Walter de Gruyter.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*(219), 325-335.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

Ellis, P. D. (2010). *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press.

Etter, J. F., & Stapleton, J. (2009). Citations to trials of nicotine replacement therapy were biased toward positive results and high-impact-factor journals. *Journal of Clinical Epidemiology, 62*(8), 831-837.

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). London: Oliver & Boyd.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research, 82*(3), 300-329.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review, 11*(5), 791-806.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine, 130*(12), 995-1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The bayes factor. *Annals of Internal Medicine, 130*(12), 1005-1013.

Goodman, S. N. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology, 45*(3), 135-140.

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology, 18*(1), 69-88.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795.

Kass, R. E., & Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association, 90*(431), 928-934.

Kjaergard, L. L., & Gluud, C. (2002). Citation bias of hepato-biliary randomized clinical trials. *Journal of Clinical Epidemiology, 55*(4), 407-410.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution, 20*(1), 28-32.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*(1-2), 187-192.

Lortie, C., Aarssen, L., Budden, A., & Leimu, R. Do citations and impact factors relate to the real numbers in publications? A case study of citation rates, impact, and effect sizes in ecology and evolutionary biology. *Scientometrics*, 1-8.

Lundberg, J. (2007). Lifting the crown-citation z-score. *Journal of Informetrics, 1*(2), 145-154.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often interpretable. *Psychological Reports, 66*(1), 195-244.

Mingers, J., & Xu, F. (2010). The drivers of citations in management science journals. *European Journal of Operational Research, 205*(2), 422-430.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Oldham, M., Kellett, S., Miles, E., & Sheeran, P. (2012). Interventions to increase attendance at psychotherapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology, 80*(5), 928-939.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology 1995, Vol 25, 25*, 111-163.

Raftery, A. E. (1999). Bayes factors and bic - comment on "a critique of the bayesian information criterion for model selection". *Sociological Methods & Research, 27*(3), 411-427.

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237.

Schneider, J. W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics, 7*(1), 50-62.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of rho values for testing precise null hypotheses. *The American Statistician, 55*, 62 - 71.

Shadish, W. R., Tolliver, D., Gray, M., & Sengupta, S. K. (1995). Author judgments about works they cite - 3 studies from psychology journals. *Social Studies of Science, 25*(3), 477-498.

Sommer, I. E., Aleman, A., Somers, M., Boks, M. P., & Kahn, R. S. (2008). Sex differences in handedness, asymmetry of the planum temporale and functional language lateralization. *Brain Research, 1206*, 76-88.

Song, F., Eastwood, A. J., & Gilbody, S. (2000). Publication and related biases *Health Technological Assessments* (Vol. 4, pp. 1-115).

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing, 71*(3), 171–193.

van Dalen, H. P., & Henkens, K. (2005). Signals in science - on the importance of signaling in gaining attention in science. *Scientometrics, 64*(2), 209-233.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804.

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry, 57*(11), 1336-1346.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Boston, MA: MIT Press.

ARE THERE INTER-GENDER DIFFERENCES IN THE PRESENCE OF AUTHORS, COLLABORATION PATTERNS AND IMPACT? (RIP)

Elba Mauleón[1] and María Bordons[2]

[1] *elba114@hotmail.com*
Department of Management, University of Bologna, Bologna (Italy)

[2] *maria.bordons@cchs.csic.es*
IEDCYT, Centre for Human and Social Sciences (CCHS), Spanish National Research Council (CSIC),
Madrid (Spain)

Abstract
This paper analyses the presence of men and women as authors and editorial board members in a selection of international scientific journals from different fields and explore potential inter-gender differences in collaboration practices and impact of research. Female presence is lower than male presence in authorship and editorial board membership in all fields and the share of women in editorial boards is lower than as authors of articles. Our results suggest there are differences in the collaboration practices of scientists by gender, since the share of women in internationally co-authored articles is lower than expected in all fields –except in Clinical Medicine-. Although large inter-gender differences in citation rates are not observed, women are under-represented in highly cited articles in most of the fields.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction
The under-representation of women in science is a matter of great current concern. The proportion of women is lower than that of men in many fields and decreases as we move up in the hierarchical structure of higher education and research institutes. Different policy initiatives have been undertaken at the national and supra-national level to promote female participation in science, not only due to equity reasons but also because our society needs to take advantage of all potential talent to increase the productivity and innovative capacity of countries. In this context, the development of studies about the situation of women in science and the collection of sex-disaggregated indicators to track the evolution over time is recommended to monitor progress (She Figures, 2009).

Since scientific journals play a crucial role in science, the study of the presence of male and female scientists in scientific journals as authors of publications and as members of editorial boards constitute relevant topics to be addressed. The study of authorship allows us to analyse the contribution of men and women to the knowledge base, while their participation as editorial board members or editors ("gate-keepers" of science) (Crane, 1967) can be understood as a sign of their scientific reputation in the field (Robinson et al, 1998). Previous studies have analysed the presence of women in the editorial board of journals (e.g. Amrein et al., 2011) or as authors of papers (e.g.Torres-Salinas et al., 2011) while the overall study of both authorship and editorial board membership is less common in the literature (Robinson 1998; Mauleón et al., 2013).

Objectives

The main objective of this study is to assess the presence of men and women as authors and editorial board members in a selection of international scientific journals and to explore potential differences in their trend to collaborate and in the impact of their research.

Main questions addressed include: What is the presence of men and women as authors of papers? Are there differences by field? What is the composition by gender of journal editorial boards? Is it a reflection of the existing community of scientists in each field? Are there gender differences in trend of scientists to collaborate? Are there gender differences in the citations received by papers?

Table 1. Journals analysed by field

Field	*Journals*	*Publication country*
Biology	MARINE BIOLOGY	GER
Biomedicine	DEVELOPMENT	UK
	J PHYSIOL-LONDON	UK
Chemistry	J AM OIL CHEM SOC	USA
	J SCI FOOD AGR	UK
Clinical Medicine	EUROP HEART J	UK
	LANCET	UK
Economics	EUR ECON REV	NETH
	REV ECON STAT	USA
Information Science	SCIENTOMETRICS	NETH
	JASIST	USA
Materials Science	CEMENT & CONCRETE	UK
	METAL MATER TRANS	USA
Mathematics	ANNALS OF STATISTICS	USA
	BIOMETRIKA	UK
Psychology	PERS INDIV DIFFER	UK

Methodology

This paper analyses authorship and editorial board membership in 16 international journals covered by the Web of Science (WoS) database in 2008 and selected as reference journals in their corresponding fields according to their specialisation

profile and impact factor value (JCR). Only citable items are considered ("articles").

The information about the composition of the editorial boards was obtained from the official journal website, from the print edition of the journals or from the journal editor (if the former options failed). For the identification of the sex of authors and editorial board members the following procedures were used: a) sex was inferred from the name of the authors when they had well-known names whose sex assignation was clear; b) through the automatic sending of electronic mail to authors asking for their sex; c) through a search of web pages, either personal or institutional.

The following indicators have been calculated.

- *Percentage of women* and *percentage of men in editorial boards.*
- *Gender gap in editorial boards* (W/M ratio): percentage of women divided by the percentage of men in a given editorial board. This ratio is below 1 when the percentage of women is lower than that of men, and above 1 in the opposite situation. A ratio of 1 indicates gender parity.
- *Presence by gender in articles*: female presence is the share of women in the total number of authors who sign articles in a given field. Male presence is calculated accordingly. These indicators are calculated taking into account the total number of author occurrences (authorships) and not unique authors.
- *Participation by gender in articles*: percentage distribution of articles in three different types: a) articles authored only by women; b) articles authored only by men; and c) articles authored by cross-gender teams.
- *Collaborative trends by gender*. Collaborative patterns of men and women are compared a) by number of authors, where male and female presence in single and multi-authored articles are compared; b) by type of collaboration, where male and female presence is studied in three sets of articles: one-centre articles, nationally collaborative articles (2 or more centres from the same country) and internationally collaborative articles (at least two different countries).

  Female presence index by number of authors. Two different indexes are calculated. The female presence index in single-authored articles is the percentage of female authors in single-authored articles divided by the percentage of female authors in total articles. An index >1 indicates that women are over-represented in this set of articles, while an index <1 is a sign of women under-representation. The female presence index in multi-authored articles is calculated in the same manner.

  Female presence index by type of collaboration. Three different indexes are calculated, by dividing the percentage of female authors in articles with one centre (or in nationally co-authored articles or in internationally

co-authored articles) between the percentage of female authors in total articles.

- *Citations by gender*. Five-year citation counts (from 2008 to 2012) are assigned to men and women on a fractional count basis. For example, in the case of a article with ten citations and five authors of whom four are males and one female, eight citations would be assigned for males and two for females. The total citation count for each sex group is divided by the fractional total contribution to obtain the mean citations per article for each sex (Lewison and Markusova, 2011).

Results

This study analyses the editorial boards of 16 scientific journals comprising a total of 832 members in 2008 (average size 52). These journals include a total of 3,186 articles totalling 14,764 authorships.

a) Editorial boards

The presence of women ranges form 9% in Mathematics to 31% in Information Science (Table 2).

Table 2. Presence of men and women in editorial boards (2008)

Field	*Editorial board members*		*Gender gap*
	%Men	*%Women*	
Biology	80.00	20.00	0.25
Biomedicine	79.30	20.70	0.26
Chemistry	79.80	20.20	0.25
Clinical Medicine	83.00	17.00	0.20
Economics	88.00	12.00	0.14
Information Science	69.40	30.60	0.44
Materials Science	85.40	14.60	0.17
Mathematics	90.80	9.20	0.10
Psychology	85.70	14.30	0.17

Table 3. Participation of authors by gender (2008)

Field	*No. Authors*	*No. Articles*	*Articles by gender of authors*		
			Only men %	*Only women %*	*Cross-gender %*
Biology	987	258	24.03	7.75	68.22
Biomedicine	3700	741	23.35	2.97	73.68
Chemistry	1882	451	23.28	7.54	69.18
Clinical Medicine	4840	552	24.28	0.54	75.18
Economics	268	114	71.93	4.39	23.68
Information Science	637	280	53.93	13.57	32.50
Materials Science	1217	346	63.01	0.87	36.13
Mathematics	363	157	67.52	4.46	28.03
Psychology	870	287	31.36	6.97	61.67

b) Authorship: participation and presence

More than 2/3 of the articles are signed by cross-gender teams in five out of the 9 fields analysed, which are mainly hard sciences fields. However, in social science fields, Mathematics and Materials Science, more than half of the articles are signed only by men (Table 3).

The percentage of female authors range from 14% in Materials Science and Economics to 39% in Psychology (Figure 1).

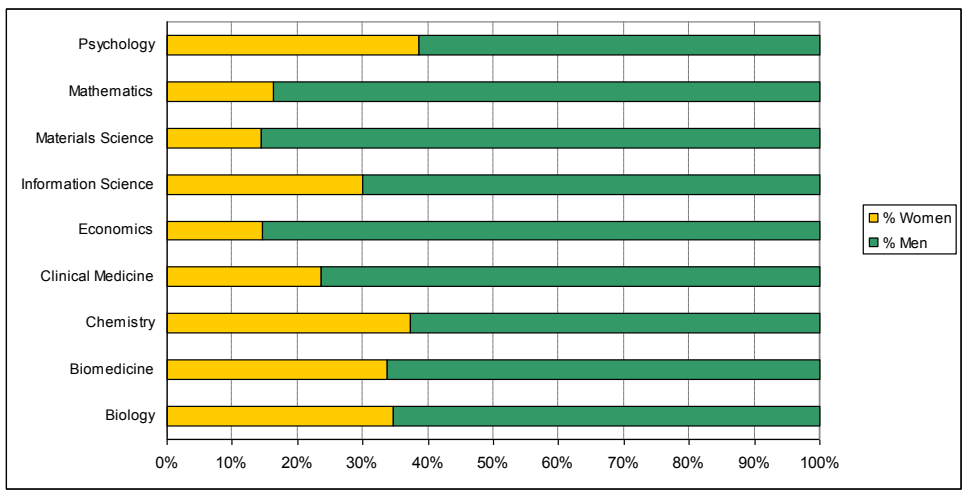


Figure 1. Presence of authors by gender (2008)

c) Collaboration

To analyse inter-gender differences in the trend of authors to collaborate, the presence of men and women according to the number of authors in articles (single-authored vs. multi-authored articles) (table 4) and the type of collaboration (one centre, nationally co-authored articles and internationally co-authored articles) (table 5) are studied. Moreover, the female presence index according to the number of authors and the type of collaboration is calculated.

The percentage of women in single-authored articles is lower than in the multi-authored ones in all fields. Moreover, the share of women in single-authored articles is lower than in the total field in all cases (female presence index<1)(table 4). The case of Materials Science and Chemistry, where there is no woman signing alone, should be mentioned. Very low female activity in one author articles is also observed in Biomedicine, Clinical Medicine and Psychology.

The meaning of single-authorhip may differ by field, since in some fields individual research is still the norm, while in others single-authorship is restricted to special type of articles (i.e.reviews) and it is more frequently used as a sign of scientific recognition. In our study, the share of women in single-authored articles

is higher in the first type of fields, although a relatively high presence of women working alone is found in Biology, where the female presence index is closer to 1.

Table 4. Presence of authors by gender and number of authors (2008)

Field	*Total articles.*		*Single-authored articles*		*Multi-authored articles.*		*Female presence index in*	
	% single-authored articles	*% multi-authored articles*	*Total authors*	*% Women*	*Total authors*	*% Women*	*Single-authored articles*	*Multi-authored articles*
Biology	6.98	93.02	18	27.78	969	34.78	0.80	1.00
Biomedicine	1.48	98.52	11	9.09	3689	33.83	0.27	1.00
Chemistry	2.44	97.56	11	0.00	1871	37.52	0.00	1.01
Clinical Medicine	2.36	97.64	13	7.69	4827	23.58	0.33	1.00
Economics	21.93	78.07	25	12.00	243	14.81	0.82	1.02
Information Science	32.14	67.86	90	25.56	547	30.71	0.85	1.02
Mathematics	17.83	82.17	28	14.29	335	16.42	0.88	1.01
Materials Science	3.76	96.24	13	0.00	1204	14.62	0.00	1.01
Psychology	12.89	87.11	37	13.51	833	39.74	0.35	1.03

Table 5. Presence of authors by gender and type of collaboration (2008)

Field	*Articles by collaboration pattern*			*No collab. articles*		*Nat.collab. articles*		*Int.collab. articles*		*Female presence index*		
	No col.	*Nat. col.*	*Int. col.*	*Total authors*	*% W*	*Total authors*	*% W*	*Total authors*	*% W*	*No col.*	*Nat. col.*	*Int. col.*
Biology	25.30	37.70	37.00	165	46.06	382	35.34	438	29.68	1.33	1.02	0.86
Biomedicine	30.80	38.20	31.00	758	32.85	1430	34.97	1512	33.07	0.97	1.04	0.98
Chemistry	37.90	44.30	17.70	550	35.64	930	40.54	402	32.09	0.96	1.09	0.86
Clinical Medicine	12.50	37.90	49.50	439	21.87	1681	23.02	2673	24.32	0.93	0.98	1.03
Economics	30.40	27.70	42.00	50	8.00	81	20.99	135	13.33	0.55	1.44	0.92
Inf. Science	49.10	27.10	23.80	235	30.21	218	34.40	176	24.43	1.01	1.15	0.81
Materials Science	31.10	41.30	27.60	272	10.66	554	17.51	384	13.02	0.74	1.21	0.90
Mathematics	31.80	39.50	28.70	79	16.46	156	20.51	128	10.94	1.01	1.26	0.67
Psychology	33.10	50.20	16.70	328	36.28	359	43.45	183	33.33	0.94	1.13	0.86

No col.= no inter-centre collaboration (1 centre). Nat.col.= national collaboration (2 or more centres from the same country). Int.col.= international collaboration (2 or more countries). %W = % Women.

With respect to the type of collaboration, in half of the fields a higher presence of women is observed in nationally co-authored articles as compared with those signed by a single centre or by centres from different countries. The presence of women in internationally co-authored articles is lower than in the total field in all

cases except in Clinical Medicine. Women are over-represented in one-centre articles in Biology.

d) Citations

No differences in the share of female authors in non-cited and cited articles are observed. However, the share of female authors in highly cited articles (HCP) (10% most cited articles in each field, 329 documents) was lower than in the whole field in all cases except in Materials Science.

Table 6. Average number of citations per article for female and male authors

Field	*No.Cit./article*		*Female presence index*	
	Women	*Men*	*Non-cited articles*	*HCP*
Biology	7.83	9.03	1.00	0.70
Biomedicine	25.70	26.50	1.00	0.85
Chemistry	6.13	7.16	1.00	0.77
Clinical Medicine	81.57	80.85	1.00	0.91
Economics	13.59	11.37	0.96	0.98
Information Science	8.24	10.37	1.00	0.38
Materials Science	11.96	9.32	1.02	1.42
Mathematics	12.93	12.73	1.00	0.90
Psychology	9.34	9.55	1.01	0.84

Conclusions

Differences by disciplines in the presence of men and women as editorial board members and authors of articles are observed. The presence of women is lower as editorial board members than as authors of articles in all fields, which may suggest a lower presence of women in the scientific elite of each discipline. Women are less likely than men to sign single-authored papers as well as to participate in international collaboration, which is consistent with previous research (Lewison and Markusova, 2011; Mauleón and Bordons, 2013). Although large inter-gender differences in citation rates are not observed, women are under-represented in highly cited articles in most of the fields. Further research concerning the relationship between collaboration, impact and gender of authors is going on. Main limitations of this type of study and implications for science policy purposes will be pointed out. The use of selected international journals as a benchmark for the study of female involvement in other journals in their corresponding fields is proposed.

Acknowledgements

This research was supported by the Spanish Ministry of Science and Innovation (CSO2008-03454-E) and a FECYT Postdoctoral Fellowship contract.

References

Amrein, K.; Langmann, A.; Fahrleitner-Pammer, A.; Pieber, T.R. and Zollner-Schwetz, I. (2011). Women underrepresented on editorial boards of major medical journals. *Gender Medicine*, 8, 6, 378-387.

Crane, D. (1967). The Gatekeepers of Science: Some factors affecting the section of articles for scientific journals. *The American Sociologist*, 2, 4, 195-201.

Lewison, G.; Markusova, V. (2011). Female researchers in Russia: have they become more visible? *Scientometrics* 89, 1, 139-152.

Mauleón, E.; Bordons, M. (2012). Authors and editors in Mathematics journals: a gender perspective. *International Journal of Gender, Science and Technology* 4, 3, 267-293. http://genderandset.open.ac.uk

Mauleón, E.; Hillán, L.; Moreno, L.; Gómez, I. and Bordons, M. (2013). Assessing gender balance among journal authors and editorial board members. *Scientometrics*, 95, 87-114.

Robinson, D.H.; McKay, D.; Katayama, A.D. and Fan, A. (1998). Are women under-represented as authors and editors of educational psychology journals? *Contemporary Educational Psychology*, 23, 3, 331–343.

She Figures (2009). *Statistics and Indicators on Gender Equality in Science*. Brussels: European Commission, Directorate-General for Research. EUR 23856 EN. Retrieved February 2012 from: http://ec.europa.eu/research/science-society/document_library/pdf_06/she_figures_2009_en.pdf

Torres-Salinas, D.; Muñoz, A.M.; Jiménez-Contreras, E. (2011). Análisis bibliométrico de la situación de las mujeres investigadoras de Ciencias Sociales y Jurídicas en España. *Revista Española de Documentación Científica,* 34, 1, 11-28.

ASSESSING INTERNATIONAL COOPERATION IN S&T THROUGH BIBLIOMETRIC METHODS (RIP)

Alexander Degelsegger[1], Dietmar Lampert[1], Katharina Büsel[1], Johannes Simon[1], Juliet Tschank and Isabella Wagner[1]

[1] *degelsegger@zsi.at, lampert@zsi.at*
Centre for Social Innovation (ZSI), Linke Wienzeile 246, A-1150 Vienna (Austria)

Abstract
International co-authorship is frequently used as an indicator for assessing international cooperation in science and technology. On the basis of experiences drawn from bibliometric studies for science and technology policy-makers and funding agencies in Europe, this research-in-progress paper critically reflects on the potential of bibliometric indicators for analysing cooperation patterns. The authors discuss limitations in using quantitative indicators for assessing the nature of cooperation. Several lines of thought for possible future indicators are introduced.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6) and Scientometrics Indicators: Criticism and new developments (Topic 1)

Background and purpose
This paper originates from a series of bibliometric studies on international *science and technology* (S&T) cooperation with emerging research communities. The studies have been carried out by the Centre for Social Innovation (ZSI) for the European Commission (on EU-Southeast Asia co-publications), the Austrian Federal Ministry for Science and Research (Austria-Danube region and Austria-India), and the Research Council of Norway (Norway-India). Their main objective was to analyse international cooperation patterns, with a specific focus on geographic and thematic portfolios. *International co-authorship* has been used as the main indicator for international cooperation.
The purpose of this paper is to use the experience gained in the course of these studies to take stock of the possibilities and limitations of assessing international S&T cooperation[18] with bibliometric methods. We will reflect on the appropriateness of using co-authorship as a proxy for cooperation and discuss a series of dimensions of S&T cooperation that bibliometric indicators can help to scrutinise. The main selection criterion for the dimensions is their relevance for S&T policy-making and programme evaluation. However, for instance regarding

[18] Throughout this paper, the term cooperation is used synonymously with collaboration. The former is preferred in the context of the conducted bibliometric studies, i.e. international S&T policy-making and -evaluation.

the dimension of *cooperation density* and the *nature of cooperation* there are gaps between what would be useful to assess for policy-making and planning and what is possible to conclude with the current data and indicators. In the ongoing research that this paper presents, we address these gaps and introduce lines of thought for possible future indicators and analysis steps.

Methodology underlying the bibliometric studies

The above-mentioned bibliometric studies[19] apply the following methodology: Data is retrieved from both major academic literature and citation databases, i.e. Elsevier's *Scopus* and Thomson Reuters' *Web of Science*. The combination of these sources results in the coverage of an additional 20-25 % of records[20] – compared to using data from just one of the two databases – and therefore a more comprehensive picture of research activities in a given geographic region[21].

Considerable correction of data and improvement of their quality is necessary, especially in those studies where institution-level analysis is required. To this end, the raw data from both sources is first analysed and corrected separately (duplicates, incorrect city-country-pairs, missing data, etc.). Subsequently, a number of algorithms and manual steps enable the unification of the data sets on the journal, record, and affiliation level, further enhancing data quality. To geographically locate records and affiliations as precisely as possible and to allow for institution-level analysis, the organisation strings in the records are disaggregated, its components tagged and analysed via semantic pattern recognition, lexica, and custom-made dictionaries. Despite the algorithm-based analyses and corrections, a significant amount of manual correction work is necessary to increase organisation-level data quality to a level that allows clustering. Clustering unites all name variations that exist in the citation databases for most organisations, under one distinct name. This clustering is largely automated and complemented by a manual correction loop. Through this series of cleaning and correcting, the quality of the data is increased considerably. Flawed bibliometric data, like missing or incorrectly attributed entries, are still an issue.

For the thematic categorisation of the records, the Science-Metrix *Ontology of Science* and the *Ontology of Scientific Journals*[22] serve as a basis. The categories

[19] Available at https://www.zsi.at/en/fe/feprofile/bibliometrische_analysen

[20] A *record* regards an entry in our database that contains the metadata of a distinctly identified publication. In case the very same publication exists in both *Scopus* and *Web of Science*, it appears as only one record.

[21] It goes without saying that despite considering both these databases, large parts of potentially relevant literature cannot be retrieved (especially publications in national language journals; grey literature). The improvements of these two databases by their providers (e.g. towards the coverage of non-English resources), the use of additional databases as well as altmetrics (specifically relevant regarding impact measures, e.g. in the form of download statistics) might alleviate this limitation to a certain extent.

[22] The Science-Metrix Ontology of Science and the Ontology of Scientific Journals are products Éric Archambault and Olivier H. Beauchesne, Science-Metrix, Montreal, Quebec, Canada, first published on 2010-12-01 (v1.00). We thank the authors for their work

of the *Austrian Classification of Science and Technology Fields* (ÖFOS version 2002), the *Web of Science* categories, and the *Scopus All Science Journal Classification* classes are subsequently mapped to the Science-Metrix categorisation system.

Based on the conducted bibliometric studies and their discussion, we were able to collect (1) qualitative information on indicator needs and (2) the interest of policy-makers and -planners in studying international S&T cooperation patterns. This information, complemented with literature research on existing indicators, guides the argumentation introduced below.

Findings and discussion

International S&T cooperation can be characterised by several dimensions. We try to identify at least one indicator per dimension that allows an assessment through bibliometric means. We start with introducing the phenomenon and indicator of international co-authorship in general.

The growth of international co-authorship

It has been shown on several occasions that the share of internationally co-authored indexed publications in overall indexed publications is increasing (e.g. Wagner, 2005; Glänzel & Schubert 2004), which has been read as a clear indication of a higher relevance of international cooperation in the generation of knowledge (Royal Society, 2011). The studies conducted by ZSI confirm this trend (cf. Degelsegger et al. 2012) for growing research communities in South- and Southeast Asia and Europe. Several parallel processes contribute to this growth: The research output of emerging scientific communities (particularly in the BRICS[23] countries) is increasing and these communities' publications are to a higher degree the result of international co-authorship. At the same time, according to our evidence, for instance, on the Danube region countries (cf. Degelsegger et al. 2012), the average number of authors per record is increasing in general in the majority of fields (which increases the statistical probability of international authors participating in a given record), which is at the same time a driver and a result of the higher number of international co-authored papers.

Studies claim that, while the number of internationally co-authored publications is increasing linearly, the number of institution links involved is increasing exponentially (Leydesdorff/Wagner, 2008). Our evidence (Degelsegger et al. 2012) confirms this hypothesis. The exponential growth in institution links over recent years can be explained by the increase in the number of authors – and the number of institutions – per record: n*(n-1)/2 institution links per record. Further scrutiny is needed, however, to examine these growth patterns in detail.

Leydesdorff and Wagner (2008) further state that the growth in international co-authorship takes place in a closely – and ever more closely – knit "core" of

[23] Brazil, Russian Federation, India, China, South Africa

around 14 countries. According to their hypothesis, those 14 can be expected to use the knowledge from global networks very efficiently because they have strong national research and innovation systems. They actually observe that the core group has decreased in size from 1990 to 2005 (from 22 to 14). This would mean that the spread of science as an increasingly global endeavour and the growing share of international co-publications do indeed signal a larger and denser network, but a smaller core group (k-core analysis after normalisation for the size of the countries). Their explanation: "[A]s actors began to experience the phenomenon of globalizing links and distributed research during the 1990s, many of them shifted their choices to incorporate a wider view of the system. But those actors in the scientifically advanced countries made more careful choices to limit their partners to specific countries" (ibid., 322).

Since our studies were not global in scope but instead analysed co-authorship between specific regions with a focus on emerging research areas (Norway-India, Austria-India, Austria-Danube region, EU-Southeast Asia), although in more detail and with improved data coverage and quality, a k-core calculation would not be useful at the national level.

What our data based on a case study on co-publications in food, health, and water-related research between Southeast Asia and the EU do show is that countries with a smaller research community have access to the same networks that Leydesdorff and Wagner describe as core group members. We thus refine their conclusions by observing that network access might not be the problem (with programmes such as the EU's 7th Framework Programme promoting international cooperation with non-EU countries). The challenges for emerging and smaller research communities seem to be the harnessing of the networks they participate in and the question what their role in these networks is. We find ourselves confronted by one of the limitations of bibliometric analyses of S&T cooperation: Co-authorship does not indicate the role of each of the contributing institutions and individuals in a given record. At most, *first author frequencies* can be analysed; apart from data quality issues, the problem is that some disciplines do not necessarily list the most active author first. Examples such as this one raise the question of the value of measuring co-authorship for assessing cooperation.

International co-authorship as a proxy for international cooperation

That international co-authorship is widely used as an important indicator and, partly, an output measure for scientific cooperation does not automatically mean that it is a suitable proxy for S&T cooperation. Even intensive cooperation between scientists can take place without leading to joint publications (Thakur et al. 2011).Moreover, several authors (e.g. Melin & Persson 1996, Katz & Martin 1997, or Jassawalla & Sashittal 1998) called attention to the fact that there is little consensus regarding what constitutes 'collaboration'. The level of formality required to make interacting scientific researchers speak of 'collaboration' varies across disciplines, time, place, etc. Laudel's (2002) inductive qualitative research in the sociology of science sheds light on different forms of collaboration. She

distinguishes collaboration in basic research activities (formulation of research questions, preparation of the research object, development of methods, measurement and interpretation), collaboration in the sense of a division of labour (with one partner carrying out the conceptual and the other the experimental work or with a division of labour within the conceptual or experimental work), service collaboration (routine contributions for sub-contracting collaborators), the provision of access to equipment, the transmission of know-how, mutual intellectual stimulation and trusted assessorship (ibid., 6ff). Interestingly, when considering the rewards for these collaborations, Laudel concludes that, as a general rule, only collaborations involving a division of labour lead to co-authorship: "For other types of collaborations no clear rules exist" (ibid., 13), with practices depending on local (type of collaboration, a research organisation's rules) and global (culture of a specific discipline, etc.) influences.

Thus, at the level of the individual researchers, many factors influence whether co-authorship results from (and can thus indicate) research cooperation. For meta-analyses at a higher level of aggregation, we nevertheless have to rely on co-authorship as an available indicator for cooperation, especially in absence of a bulk of qualitative data, and will refer to "cooperation that results in co-authorship" simply as "cooperation". The following observation serves as justification: Given that journal publications are not only goals and milestones of individual research careers, but are also used as a streamlined performance indicator that scientists are very well aware of and adjust their behaviour to, scientists' propensity of publishing joint work together is usually high. Moreover, strategic considerations common in the practice of co-patenting play a smaller role. In line with these arguments, Gómez et al. (1999), for instance, come to the conclusion that bibliometric indicators are useful for tracking both formal and informal scientific collaboration (ibid., 455).

Katz and Martin (1997) stress the need to distinguish individual, institutional, country, and region-level cooperation. An inter-institutional collaboration, for instance, does not necessarily entail inter-individual collaboration (ibid., 16). This becomes clear, for instance, when considering Memoranda of Understanding signed between universities that are not followed up by actual face-to-face collaborative research. Glänzel and Schubert (2004), who confirm the basic validity of using co-authorship as an indicator for S&T cooperation when analysing it at an aggregated level, also underline that the motivations behind co-authorship of individuals and co-authorship between institutions and countries are different. Moreover, for our meta-analyses at the country- or institution-level, the question of motivations is not as relevant as the question of the extent and nature of cooperation.

Assessing the extent and density of cooperation

At the country-level, with the current data available, the extent of cooperation between two countries can be measured by simple frequency counting of co-

authorship occurrences in a basic dataset (e.g. covering a range of years). Analyses of time series can add information on trends in cooperation.
When deepening the analysis with the aim of assessing the density of the cooperation, several options are available: The density of the cooperation between the countries can be operationalised and assessed by comparing the extent of the co-authorship link with the overall publication output in one or each of the countries. In the latter case, the Jaccard index and Salton's cosine are used as measures (Salton & McGill, 1983 and Hamers et al., 1989; cf. as well Leydesdorff, 2008). Although less frequently done, the co-authored output can also be weighted by relating it to measures on expenditure in research and development (GERD/GDP) or researcher full-time equivalents.

Assessing thematic portfolio of cooperation

Using the thematic categorisation systems of one of the literature databases (*Scopus All Science Journal Classification* or Web of Science thematic categories) or, as in our case, other sources like the Science-Metrix *Ontology of Science*, the co-authored output can be assessed regarding its thematic focus. Here, it is important to take into account that cooperation in thematic area x cannot be considered more relevant or successful simply because the output in absolute numbers is higher than in thematic area y. Normalisations are required to be able to reach this kind of conclusions: One option is to normalise the co-authored output in each thematic area by relating it to the overall output of a country in this particular area. Another option is to compare co-authorship output in one thematic area with one country with co-authored output in the same thematic area, but with another partner country or region.

Assessing the impact and quality of cooperation

The most readily available indicator for the quality of a co-authored record is its *times-cited counts*. It might soon be possible to access *download count* data for records as well, but for the moment data accessibility and coverage is still a problem here.
Times-cited counts are thus still the most common indicator used for assessing the impact of a (citable) journal publication. In assessing co-authored publications, it should be taken into account that they are in general cited more frequently than non-co-authored work, largely because they are fed into broader networks accessible through the co-author group. According to the Royal Society (2011, 59), internationally co-authored papers are cited more frequently than others, and each international co-author up to a turning point of 10 authors adds additional citations (after that, the "marginal gain" from each additional author decreases). On the basis of this diagnosis, times cited counts as an impact measure of cooperation can thus be criticised. However, this critique depends on the definition of impact. If the uptake of results in as many cases as possible is considered impact, which is a reasonable working definition implicitly applied by

many policy-makers, then internationally co-authored publications that are cited more frequently indeed do have a higher impact.
Another important aspect to take into account is varying citation cultures in different scientific disciplines. If field normalised citation scores are available (for the suitable time period and in an appropriate thematic categorisation), Crown indicators (CI) can be calculated to determine whether the *average times cited counts per record* are above or below the average in a specific scientific field (thus combining impact and thematic analyses) and to what extent (this is important in view of the fact that co-authored articles are cited more frequently in general). Gorraiz et al. (2012) compute and compare domestic and collaborative Crown indicators in order to assess the citation gains through co-authorship.
Another way of assessing the quality of a set of co-authored records is to determine the subset which has appeared in high-impact journals (as defined e.g. by using the Scopus SNIP values) or the subset that is cited more often than a threshold of interest (h-index or related indices can be used as thresholds).

Assessing cooperation density in more detail: Networks

The dimension of cooperation density, which was introduced above, can be interpreted and analysed at two distinct more detailed levels, as well, not simply relating co-authored output to general output. One level concerns the relevance of nodes (e.g. countries, but also institutions) in a specific network, which can be assessed with social network analysis (SNA) centrality measures like *betweenness* (cf. also Leydesdorff & Wagner, 2008 for the computing of k-cores in co-authorship networks).
Another one takes into account properties within the record like the number of authors or countries involved. The average number of authors involved in co-authored papers between two countries in a given field can tell us whether the cooperation in this field can be considered as stemming from "big science" collaboration. The evidence from the co-publications between authors affiliated in Austria and some Danube region countries in the field of physics shows that this cooperation mostly takes place within international author networks of more than 100 members. The probability that face-to-face interaction between the authors in Austria and, for instance, Bulgaria is involved is low.
This determination of the nature of co-authored records as either big science or actual face-to-face cooperation partly illuminates the nature of cooperation behind a specific record (or set of records).

Assessing the nature of cooperation

Simply distinguishing big science from non-big science cooperation is not enough. Different forms of cooperation can also be expected to be behind a record co-authored by 3 authors or one where, for instance, 12 authors are listed. Thus it makes sense to further refine the analyses of author numbers in co-authored papers. One way to do this would be to consider the number of authors and number of countries involved in a given record at the same time (then aggregating

the information to the required level – country, institution, etc.). This would allow distinguishing close-knit *multilateral collaborations*. For instance, the multilateral collaborations of four different researchers affiliated with four different countries could be distinguished from (a) dense *bilateral collaborations* of four different researchers from two different countries or (b) from multilateral collaborations made up by multiple affiliations (e.g. two authors affiliated with four different countries). More detailed accounts of multilateral and bilateral cooperation density would be the result.

Another possible way of using co-authorship evidence to assess the nature of S&T cooperation would be to work with distance measures, i.e. the geographical distance between the authors in a co-authored publication could be measured (as long as the data quality allows to *geo-map* their affiliations). While this might as such not seem very useful (with co-authored publications involving many authors obviously spanning larger distances), a normalisation of the distance covered per author-author link could actually be insightful. Considering that language distances might outweigh geographic distances, e.g. it might be easier for a US-based researcher to co-author a paper with an Australian than with a Nicaraguan colleague, "soft" distance measures would need to be introduced that would quantify factors like language or cultural barriers. Spatial econometrics literature has tested hypotheses regarding the impact of social, geographic or territorial distance on scientific collaboration (cf. Autant-Bernard 2011). Our concern, however, lies with a question that spatial econometrics often takes for granted: whether or not co-authorship is a strong indicator for collaboration.

We introduced Laudel's (2002) typology of forms of collaboration, with only some of them leading to co-authorship. Presently, bibliometric indicators do not allow drawing meta-level conclusions on the distribution of conceptual and experimental work between researchers affiliated in two different countries. To analyse the distribution of work, the provision of access to equipment, or the relevance of assessors and intellectual debate for a given record, the acknowledgements or sub-authorship would have to be scrutinised. Although Glänzel and Schubert (2004) observe that proper acknowledgement is less of a problem in international collaboration, it is not reliable enough to serve as an indicator.

With regard to transmission of know-how as a form of cooperation, however, the available co-authorship data might be scrutinised for gaining insight: At the author- or institution-level, co-authorship networks would have to be analysed first with regard to new linkages: Which institutions or authors have not co-authored a publication in a specific field before. As a next step, for each institution or author, earlier publications in the field of relevance would have to be tracked. One of the contributing institutions not having published in the relevant subject area before could indicate (a) the transfer of know-how, (b) the appearance of a new institution or author, or (c) the establishment of a novel thematic focus of a given institution or author. To distinguish the former from the latter two, earlier publications of the institution or author in question could be

tracked. If an earlier publication record exists, the new and recent collaboration could indicate either a transfer of know-how and/or the existence of new know-how because of other sources (e.g. the establishment of a new department in an institution; the acquisition of new knowledge from other sources).

Conclusions

Co-authorship as a bibliometric indicator is useful in assessing various dimensions of international cooperation. With regard to assessing the nature of cooperation, at a meta-level just as well as at the level of individual collaborations, the meta-analyses lack qualitative background information. Indicators taking into account aspects like network size, average distance covered per author-author link, etc. could mitigate these deficiencies and allow drawing some conclusions on the type of cooperation dominant in a given subset of co-authored publications. New data, data integration, and related indicators – such as altmetrics; more consistent data on funding that led to a specific publication; more general tracking of research project granting, project participation or mobility, allowing to relate this to publication output – might open new possibilities. To assess the *type* of actual cooperation underlying a co-authored work, there is still no way around qualitative biographic information from the authors involved.

References

Autant-Bernard, Corinne (2011): Spatial econometrics of innovation: Recent contributions and research perspectives, Working Paper 1120, Ecully: Group d'Analyse et de Théorie Économique Lyon-St Étienne

Degelsegger, Alexander, Dietmar Lampert, Katharina Büsel, Johannes Simon, Juliet Tschank, Isabella Wagner (2012): Analyse der Kopublikationen Österreichs mit Donauraum-Partnern, available (in German) at https://www.zsi.at/en/object/project/2097/

Glänzel, W. & Schubert, A. (2004). Analyzing Scientific Collaboration through Co-Authorship, in: Moed H. F., W. Glänzel & U. Schmoch (Eds), Handbook of Quantitative S&T Research, Dordrecht: Kluwer, 257-276.

Gómez, I., Fernández, M.T. & Sebastián, J. (1999). Analysis of the Structure of International Scientific Cooperation Networks through Bibliometric Indicators. *Scientometrics*, 44(3), 441-457.

Gorraiz, J., Reimann, R. & Gumpenberger, C. (2012). Key factors and considerations in the assessment of international collaboration: a case study for Austria and six countries. *Scientometrics*, 91(2), 417-433.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity Measures In Scientometric Research: The Jaccard Index Versus Salton's Cosine Formula. *Information Processing & Management*, 25, 315-318.

Jassawalla, A. R. & Sashittal, H. C. (1998). An examination of collaboration in high-technology new product development processes. *Journal of Product Innovation Management*, 14(3), 237-254.

Katz, J. S. & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26, 1-18.
Laudel, G. (2002). What do we measure by co-authorships? *Research Evaluation*, 11(1), 3-15.
Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data. Salton's cosine versus the Jaccard Index. *Journal of the American Society for Information Science and Technology*, 59(1), 77-85.
Leydesdorff, L. & Wagner, C. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics*, 2(4), 317-325.
Melin, G. & Persson, O. (1996): Studying research collaboration using co-authorships. *Scientometrics*, 26(3), 363-377.
Royal Society. (2011). Knowledge, networks and nations. Global scientific collaboration in the 21st century, London: The Royal Society.
Salton, G., & McGill, M. J. (1983). Introduction to Modern Information Retrieval. Auckland, et al.: McGraw-Hill.
Thakur, D., Wang, J. & Cozzens, S. (2011). What does International Co-authorship Measure? *Science and Innovation Policy, 2011 Atlanta Conference*, 15-17 Sep 2011, pp.1-7.
Wagner, C. (2005). Six case studies of international collaboration in science. *Scientometrics*, 62(1), 3-26.

ASSESSING OBLITERATION BY INCORPORATION IN A FULL-TEXT DATABASE: JSTOR AND THE CONCEPT OF "BOUNDED RATIONALITY."

Katherine W. McCain
mccainkw@drexel.edu
The iSchool, College of Information Science & Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104

Abstract
To evaluate the usefulness of a full-text database as a source for assessing Obliteration by Incorporation, 3707 article records including the catchphrases "bounded rationality" and/or "boundedly rational" (connected with the work of H.A. Simon) in the article text were retrieved from JSTOR, a full-text database with broad disciplinary coverage. Two subsets were drawn for analysis—a 10% systematic sample of all records (364 articles analysed) and a set of all records with a catchphrase in the Title and/or Abstract field (178 articles analysed). A majority of articles in both subsets came from Economics and Management journals, while Psychology was poorly represented due to database coverage. In the 10% sample, based on the percentage of true implicit citations, a low level of OBI was observed in the 80% of records that had a catchphrase in the body of the article, rather than just in the reference list. Most indirect citations were to sources that themselves cited a relevant work by Simon. Over 90% of the sample articles would not have been retrieved with a database record search because they lacked the catchphrase in the record fields and the percentage of implicit citations was significantly higher in these articles.

Conference Topic
Scientometric Indicators (Topic 1), Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2), and Bibliometrics in Library and Information Science (Topic 14)

Introduction
In the half-century since the first volume of the Science Citation Index was published (Garfield, 1979), citation counts and citation history profiles have demonstrated their usefulness in assessing the visibility and influence of published scholarly works. Some publications are never cited (Burrell, 2012), requiring that any influence or impact be assessed using other sources of data. Publications that receive some measurable number of citations appear to fall into one of a few citation history profile categories (Costas, van Leeuwen & van Raan, 2010). These include "normal" documents (highest citation count in years 3-4, followed by an exponential decline), "flash in the pan" documents (peaking early, sharp decline) or "delayed" documents (a very late peak in citation counts—see

van Raan's "sleeping beauties," van Raan, 2004). A relatively small number of publications can be termed "citation classics" (Garfield, 1977)—their influence, as measured by citation counts, persists at a high level for many years.

The degree to which citation counts for citation classics are "true" indicators of their influence has been debated. On the one hand, citation bias, arising from prior visibility of authors and works may result in overcitation (Barabási & Albert, 1999; MacRoberts & MacRoberts, 1989; Merton, 1968; Price, 1976). On the other, as Zuckerman (1987) notes, citation classics may actually receive fewer citations than would be expected, as their key concepts are incorporated into the body of scholarship without attribution or when more recent works are cited for the contributions first articulated in the original work.

These two sources of undercitation were described by Merton as "Obliteration by Incorporation" and the "Palimpsestic Syndrome."

- **Obliteration by Incorporation**: "the obliteration of the source of ideas, methods, or findings by their incorporation in currently accepted knowledge" (Merton, 1988, p. 622)
- **The Palimpsestic Syndrome:** the covering over of earlier versions of an idea by ascribing it to a comparatively recent author *in whose work the idea was first encountered*" (Merton, 1965, p. xxiii, emphasis mine)

In the forward to Garfield's treatise on citation analysis (Merton, 1979), Merton links this latter notion of (deliberate?) misattribution of intellectual credit with the more general process of Obliteration by Incorporation (OBI). Obliteration by Incorporation is the disappearance of citations to the older work, although the ideas live on; the Palimpsestic Syndrome (citation substitution) is the attribution of the idea to an author who did not originate it but made the work accessible or visible, thus acquiring the citations that should otherwise go to the original author. His discussion appears to assume that the intent of the author or the assumption of the reader was to identify the newer source as the origin of the concept, rather than, say, simply a more useful or approachable discussion than the known original:

> "And since many of us tend to attribute a significant idea or formulation to the author who introduced us to it, *the altogether innocent transmitter sometimes becomes identified as the originator*. In the successive transmission of ideas, repeated use may erase all but the immediately antecedent versions, thus producing an historical palimpsest in which the source of those ideas is obliterated." (Merton, 1979, p. ix, emphasis mine)

In practice, OBI and citation substitution can be detected using Citation-in-Context Analysis (Small, 1978)—cited documents as concept symbols are

represented by the same or very similar turns of phrase in conjunction with the formal reference to the source. Citations to the original work (explicit citations) yield the formal citation counts. OBI would be demonstrated by the use of eponyms or catch phrases in the citation context without accompanying citations (termed "implicit citation" to the original work by Thomas, 1992). Citation substitution ("indirect citation", see Rousseau, 1987 and Thomas, 1992) would be more difficult to observe, since it requires knowledge of what the "original" citation should be and which newer work is being offered in its stead. This can be assumed in specific cases where one is studying the influence of a specific concept and its associated source (one or a small set of works, published or unpublished).

Most prior empirical studies of OBI and citation substitution have been based in large part on retrieval of bibliographic database records to identify sources within which implicit, explicit and indirect citations can be observed, rather than direct access to full-text sources (see discussion in McCain, 2012 and a review of empirical studies in McCain, 2013). OBI (but not citation substitution) can be studied at the database record level – the eponym or catch phrase occurs in searchable fields of the record. The presence or absence of a citation to the source can be tallied and a very large number of records processed. Both OBI and citation substitution can be explored if the full text of articles identified in the database search is examined and the context in which the phrase occurs is noted. The number of articles that can be studied in this way must necessarily be smaller.

Both of these approaches are limited in that the set of potentially citing papers must include the concept phrase in the title, abstract, key words or other searchable database field. In this paper, I explore the usefulness of JSTOR, a full-text-searchable database of scholarly articles, in identifying useful sources for Citation-in-Context Analysis of OBI and citation substitution. The results demonstrate the limitations of reliance on bibliographic database records as well as some interesting phenomena that arise when using a full-text searching approach.

Methods

JSTOR is a full-text database with multidisciplinary coverage of the arts, humanities, natural and social sciences. My university library subscribes to most of the JSTOR journal collections—Arts & Sciences I through VIII and Life Sciences collection. In June, 2012, I searched JSTOR in full-text mode for any occurrence of Herbert A. Simon's phrase "bounded rationality" or the variant "boundedly rational." The search yielded 3707 article records that were downloaded to RefWorks and then to a Filemaker Pro database for coding and analysis.

The results reported here are based on analyses of two subsets of the full retrieval—a 10% systematic sample [24] and a set of all records that included the catch phrases in the Title and/or Abstract field. In both cases article pdfs were retrieved from JSTOR, their text searched, and the catch phrase, along with any associated citation, captured for analysis. Each article was tagged with the broad subject area assigned to its journal (based on *Ulrich's International Periodical Index*) and the nature of the citation context was coded:

- Explicit citation—the mention of "bounded rationality"/"boundedly rational" is connected to a reference to a work by Simon, alone or with works by other authors
- Indirect citation—phrase is connected to a reference only to work by another author
- Implicit citation—phrase occurs in the text without any accompanying citation.

Only articles in English were retained in these two subsets; a small number of articles from *Revue Economique* were not used.

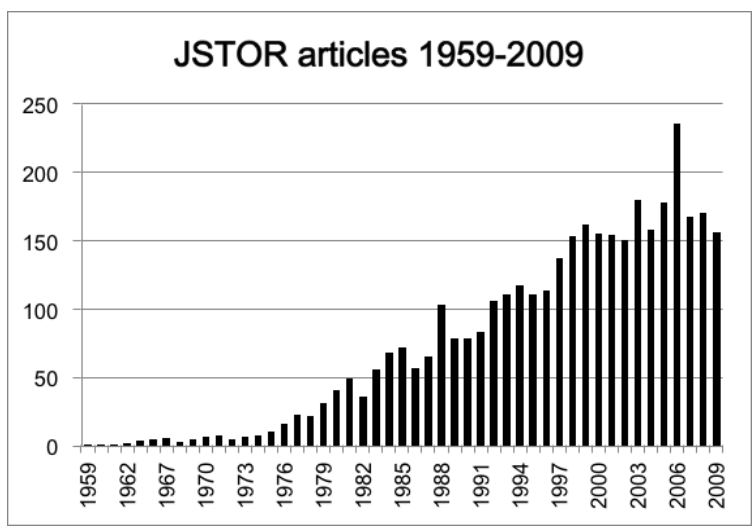


Figure 1. Annual distribution of JSTOR articles with "bounded rationality"/"boundedly rational" in the title, abstract, or text

[24] A systematic sample begins by determining the percent of items to be chosen, in this case 10% or 370 of the 3707. Beginning with a random position number between 1 and 10, every 10th record is chosen for analysis. The results are considered equivalent to a simple random sample for most purposes.

Results

Profile of Full Data Set

The full data set included 3707 articles. Figure 1 shows the annual count of articles in the data set. One limitation of JSTOR is that the window of availability can vary from one journal to another (a "moving" or "fixed wall" imposed by the publisher) resulting in an under-representation of the most recent years. But overall, the number of articles including the catch phrase has increased since the late 1950s (the earliest JSTOR article retrieved).

Table 1. Subject and citation profile of 364 articles retrieved with the phrases "bounded rationality"/"boundedly rational.

Article Subject	*Explicit citation*	*Indirect citation*	*Implicit citation*	*Citation only in reference title*	*Other*	*Total articles*
Economics	16	28	38	19	3	104
Management	23	23	16	24	1	87
Law	10	12	6	5	0	33
Pol Sci	9	6	7	9	0	31
Sociology	4	7	5	6	0	22
Philosophy	7	5	6	1	2	21
Business	4	7	5	3	0	19
Soc Sci	2	6	2	2	0	12
Education	5	0	3	2	0	10
Anthropology	2	1	2	0	0	5
History	4	1	0	0	0	5
Humanities	1	1	0	1	2	5
Gen Sci	1	0	3	0	0	4
Psychology	1	0	1	2	0	4
LIS	0	0	0	1	0	1
Medicine	0	0	0	0	0	1
Total	89	97	94	75	4	364

Ten Percent Sample

Of the 10% systematic sample of 370 articles, 364 were ultimately accessible and readable. The overall Citation-in-Context results are shown in Table 1 and Figures 2 and 3. Across the data set as a whole (Table 1), roughly 1/5 of the sample articles only have the phrase in a cited reference title, not in the article text itself. These cited articles may be in the full JSTOR retrieval as well, but, in any case, can't count toward the assessment of this sample.

Figures 2 and 3 illustrate OBI trends over the years 1992-2009 for articles containing the catchphrase in the record fields and/or text. There was an average

of slightly less than 12 articles a year published (Figure 2) and there appears to be a slight upward trend in Obliteration by Incorporation (Figure 3), based on the annual percentage of these articles that contained implicit citations (catch phrase in text without linked reference).

Eighty-nine articles cited at least one work by Simon (as author or co-author) along with the catchphrase in the text. Two works by Simon were cited most frequently:

- Simon, Herbert A. (various editions) *Administrative Behavior*. New York: MacMillan (12 citations)
- Simon, Herbert A. 1957. *Models of Man: Social and Rational*. New York: Wiley (11 citations)—a canonical citation for the concept

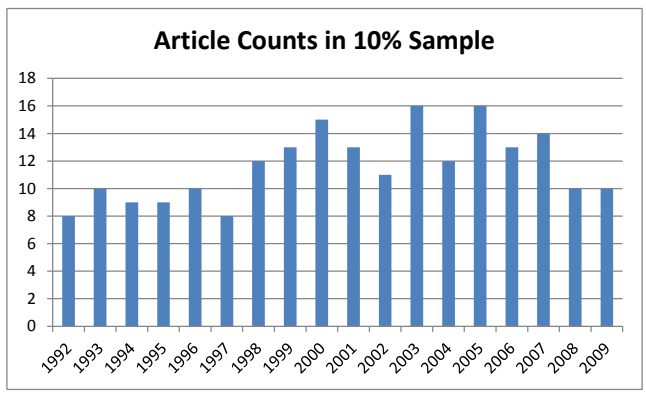


Figure 2: Number of articles retrieved with catchphrase in record or text, 1992-2009

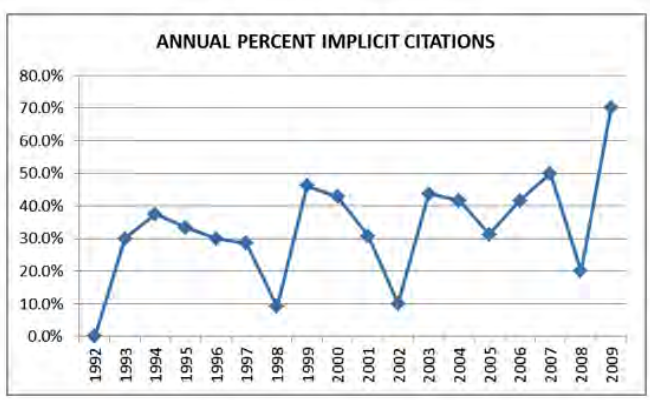


Figure 3: Obliteration by Incorporation trend—annual percentage of implicitly citing articles, 1992-2009

Very few authors and works occurred more than once as indirect citations. The most frequently cited author in this category was Oliver E. Williamson (30 citations) and the most frequently cited work (12 citations) was Williamson, Oliver E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

The articles and books indirectly cited in 77 articles were available for examination, with the following results,

- 59 indirect citations pointed directly to Simon with at least a brief discussion of "bounded rationality"
- 18 had no mention of Simon, the extent of discussion of "bounded rationality" varied
- 20 citing Williamson (who cites Simon) generally had extensive discussion of the concept

Title-Abstract retrieval

One hundred seventy eight articles in the JSTOR retrieval had one of the two catch phrases in the Title and/or Abstract field in the database record. Most of these also included the phrase in the text proper. The subject distribution is shown in Table 2 and the annual publication profile in Figure 4.

Table 2: Subject and citation profile of 178 articles with catch phrase in Title and/or Abstract field

Article Subject	*Explicit*	*Indirect*	*Implicit*	*Implicit Simon*	*TI/AB only**	*Total Articles*
Economics	19	18	14	1	11	63
Management	22	7	5	0	3	37
Political Science	11	8	5	1	1	26
Sociology	9	1	3	0	0	13
Business	4	2	5	0	1	12
Law	3	6	1	0	0	10
General Science	3	1	2	0	0	6
Social Sciences	2	0	1	0	1	4
Life Sciences	1	1	0	0	0	2
Philosophy	0	1	1	0	0	2
Education	1	0	0	0	0	1
Humanities	1	0	0	0	0	1
Math/Stat	0	0	0	0	1	1
Total	76	45	37	2	18	178

* includes three articles that also had the text in a cited reference title, but not in the body of the article

In this set, there are too few articles with implicit citations to look for an annualized trend—18 across the entire set of 178 articles. We can, however, ask a

related question—is there a difference in the relative proportion of implicit citations versus those that link some citation (explicit or indirect) with the catch phrase in the text in the two cases we've examined so far? That is, are articles that include the catchphrase in the Title and/or Abstract fields different from those that only mention the catchphrase in the text proper?

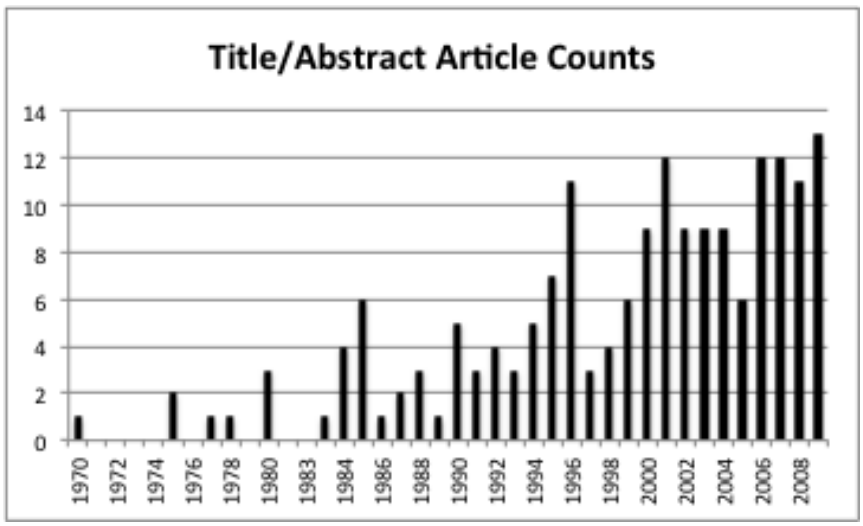


Figure 4: Annual counts of JSTOR articles with catch phrase in the Title and/or Abstract field

Table 3 shows the counts of implicitly citing articles versus articles that explicitly or indirectly cite a source for "bounded rationality" over the period 1992-2009. To more clearly separate the two sets of articles, the 10% sample set excludes articles duplicated in the TI/AB set. A Chi Square test (p=0.0074) suggests a strong association between the occurrence or absence of the catch phrase in the TI/AB field and the nature of the within-text citation.

Table 3: Comparison of implicit vs "citing" article counts, 1992-2009

Article Set	*Explicit & Indirect*	*Implicit*	*Total*
10% no TI/AB *	120	70	190
All TI/AB only*	121	37	158

* excludes articles lacking a catch phrase in the text, and thus having no citation assessment

Simon's most highly cited works are the same as those in the 10% sample, with one addition:

- Simon, Herbert A. (various editions) *Administrative Behavior.* New York: MacMillan (21 citations)

- Simon, Herbert A. 1957. *Models of Man: Social and Rational*. New York: Wiley (13 citations)
- March, James G. & Simon, Herbert A. (various editions). *Organizations.* New York: Wiley (12 citations)

While the list of indirectly cited works is, again, highly diverse, in this set there is no real concentration on clear substitutes for Simon. The most highly cited author is James G. March (5 citations to three different works), while Oliver E. Williamson (the most frequent "Simon substitute" in the 10% sample) was only cited twice. The most highly cited individual works are:

- Jolls, C., Sunstein,C. R.& Thaler, R. (1998). A Behavioral Approach to Law and Economics. *Stanford Law Review*, *50*(5), 1471-1550. (4 citations—3 self-citations by Sunstein)
- McKelvey R.D. & Palfrey, T.R. (1995). Quantal response equilibrium for normal form games. *Games and Economic Behavior*, *10*,6-38 . (4 citations)
- Cyert, RM. & March, J.G. (1963). *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall. (3 citations)
- Rubinstein, A. (1998) *Modeling Bounded Rationality*. Cambridge: MIT Press. (3 citations)

Jolls et al, Cyert & March, and Rubenstein all cite Simon and thus serve as pointers to his work; McKelvey & Palfrey focus on Nash equilibria and do not cite Simon for his notion of bounded rationality.

Discussion

Disciplinary impact of "bounded rationality"

One of the challenges in any study of Obliteration by Incorporation is finding a reasonably satisfactory data source. At the database level, one can search broad-spectrum sources such as Web of Science or Scopus, focus on a particular area such as the life sciences, medicine, engineering, or psychology, or use the disciplinary databases and even individual journal searches to enhance an initial search (see, e.g. McCain, 2012). There are fewer options when a searchable full-text approach is wanted—and challenging trade-offs. On the one hand, JSTOR has some distinct advantages—it is multi-disciplinary, has document text that is searchable (rather than just providing page images) and my University's library subscribes to most of the journal bundles, so that a search can capture most of what JSTOR has to offer. One limitation is, as noted earlier, that publishers may impose constraints on the availability of current issues of the journals provided. Another is that the range of journals provided by JSTOR varies by design as well as by institutional subscription choice. In particular, Psychology, a discipline in which cognitive concepts such as "bounded rationality" would clearly be

important, is not a journal subject "cluster" in JSTOR, and there are only about 10 journals in the JSTOR collection. [25]

So the results reported here must be taken as suggestive, and not representative of Psychology, but likely to be reasonably representative for the remainder of the social sciences, most particularly Economics, Management, Law, Political Science and Sociology. In the 10% sample, Economics and Management contributed more than half of the total article set. To the extent that the sample is representative of the full JSTOR retrieval (a systematic sample with a random start is subject to problems if there are strong repeating patterns in the data), it would appear that Simon's concept has had the most influence in these four fields. In the TI/AB data set, Economics and Management are still ranked #1 and #2, with Law and Political Science still in double digits—the appearance of different disciplines in Table 1 and Table 2 reflects the fact that the first is a 10% sample of the full retrieval, while the second is a full inventory of all articles with the catchphrase in the Title and Abstract field.

Comparing record-level, record+ text, and full-text analyses of OBI

Assessments of OBI at the record level and full text level must necessarily differ because, in the latter case, indirect citations can be identified. They are important because, even though they point to alternative discussions of the concept of interest, still represent the concern of the author that something needs to be cited at that point in the text and it makes sense to consider them as other than "implicit" citations. (They are examples of "citation substitution" and reduce the number of references to Simon's original publications.) Indirect citations cannot be detected when working with citation data at the record level, since it is not possible to structurally connect, say, a reference to OE Williamson (rather than a reference to HA Simon) with a text discussion of "bounded rationality." The absence of an appropriate citation to Simon in the reference list must therefore count as an instance of Obliteration—with the result that OBI is likely to be over-estimated (and a citation to Simon that is NOT connected with a discussion of "bounded rationality" would count, incorrectly as an explicit citation in the context of assessing OBI.).

This over-estimation can be tempered somewhat by using record+text analysis to identify indirect citations which can add to the tally of articles that cite something versus those that cite nothing (=implicit citations =obliteration), but the results are

[25] In contrast, a catch phrase search in Web of Science produced 161 articles in psychology journals (1980 – 2012) with the phrase in a searchable database field; a similar search in PsycINFO yielded 447 database records (1978-2012), while the TI/AB search in JSTOR yielded 4 articles in psychology journals. Unfortunately, my institution lacks a subscription to PsycARTICLES, the full-text article database produced by the American Psychological Association, which would have been the natural choice to expand the JSTOR search and provide good coverage of "bounded rationality" in psychology.

still constrained by the initial requirement that the catchphrase be in a field of the database record. The results reported here suggest that record+text analyses of OBI—searching the database records and then analysing the article text—may be capturing only the tip of the textual iceberg. Twenty-one of the 364 articles (6%) in the 10% sample had the catch phrase in the Title and/or Abstract field and, overall, only 178 articles (7% of the 3707 JSTOR retrievals) met this constraint. More than 90% of the articles in the complete retrieval could be found only with a full-text search. The down-side of a full-text search is that, as the data in Table 1 show, as much as 20% of the retrieval may not represent articles that directly discuss the concept of interest—at least insofar as the catchphrase is not used in the text proper. Seventy-five of the 364 articles are "false positives"—being retrieved only because the catch phrase was in the title of an item in the article's reference list. (No attempt was made to do a deep textual analysis of the context of these citations – the focus here was on a linkage between the specific catchphrases and associated references. Thus the article may have "talked around" the notion of "bounded rationality" without using the phrase.)

Table 3 illustrates another aspect of the assessment of OBI at the text level. It appears that articles retrieved because the catchphrase is part of the searchable database record (as well as being in the text) are more likely to contain a link between catchphrase and reference in the text than are articles with the catchphrase in the text but NOT in the database record. The two article subsets (full-text only from the 10% sample versus all articles with catchphrase in TI/AB field) produced about the same number of "citing" articles—with an explicit or indirect citation associated with the phrase "bounded rationality"/"boundedly rational." But the "in-text only" set had twice as many implicitly citing articles as the "in-the-record" set—resulting in a significantly higher level of Obliteration by Incorporation.

Are indirect citations directional?

The role of "indirect citations" is an interesting one. In the context of OBI and the Palimpsestic Syndrome, Merton characterized this citation substitution essentially as the author deliberately citing a newer work in place of the original *because the citing author learned of the concept from the more recent source* as opposed to the original. To Merton, this pattern of citation contributes to OBI because the citations that should have gone to the original source then go to the newer work. We cannot know the state of mind or level of knowledge of the citing author, or his or her contextual choices. But the results reported here suggest that "indirect citations" are themselves very likely to point directly to the original sources, often with extensive discussion. In the 10% sample 59 of 77 indirect references that could be examined cited one of Simon's canonical references to "bounded rationality. For this reason, it makes sense to combine counts of indirect and explicit citations when assessing OBI.

Is Simon being "obliterated" with respect to "bounded rationality?"

The short answer is: "yes, but how much depends on what you count… ." The most constrained measure of Obliteration by Incorporation (and the one preferred in this study) is the percentage of articles that include the phrase in the text but totally lack any linked reference to a source documenting the concept—true implicit citations. By this measure, the results reported here suggest that OBI is indeed occurring, with a slight upward trend over time, but that the majority of articles discussing the concept still connect the phrase with a source work. There is less obliteration than observed for catchphrases or eponyms such as "sea floor spreading." (Messeri, 1978), "Southern Blot," (Thomas, 1972), and "Nash Equilibrium," (McCain, 2011), but similar to the OBI percentage reported in the text-level analysis of "Evolutionarily Stable Strategies" (McCain, 2012). If we only focus on explicit citations to Simon (combining indirect and implicit citations), the overall percentage of OBI in the 10% sample is almost 70% of articles that include the phrase in the text proper. But I would argue that citing something (particularly if it is a true indirect reference to the original) is very different from omitting any citation to support the discussion of the concept—and that only implicit citations should be used to assess OBI. (Authors such as Merton, who observe that their works are declining in visibility because references are being made to newer works, are likely to have a different opinion.)

Next Steps

The results reported here represent a small sample from one large retrieval from one broad-spectrum, multidisciplinary database, JSTOR. One obvious extension would be to continue to work with the remainder of the 3707 articles in this "bounded rationality" retrieval set. While I doubt that there would be many surprises, if the systematic sample is a reasonable approximation of the full data set, this would provide more data for analysis. It would also allow a focused assessment of OBI in Economics and Management, the two most visible disciplines in which "bounded rationality" has had an impact.

Text-level OBI studies would seem to be desirable, if one wants to get a more accurate picture of the contextual citation patterns relating to the impact of a concept of interest and separate indirect and true implicit citations. In addition to JSTOR, there are other full-text databases that could be used. These include publishers' e-journal offerings, such as the full-text searchable journals of HighWire Press, Wiley-Blackwell, the American Psychological Association (PsycARTICLES), and Elsevier (SciDirect), and disciplinary databases that offer direct full-text searching such as Library Literature & Information Science (HW Wilson), and ABI/Inform Complete (ProQuest). The choice of concept to study would need to be tailored to the subject orientation and source coverage of the database. A third option would be to assemble a relevant list of e-journals to which one has access, and search them directly, building the data set by hand.

Conclusions

From the point of view of assessing the degree of OBI experienced by Simon's notion of "bounded rationality," it appears that more than half of the JSTOR articles examined in the two subsets either cite one of two key works by Simon or a book or article that points to Simon's work—if OBI is increasing, it is doing so very slowly.

From the point of view of the effectiveness of using full-text retrieval to capture a literature for analysis, it appears that relying on database record searches can miss 90% or more of the articles with at least one mention of the chosen catchphrase.

So it appears that OBI researchers are faced with a set of difficult trade-offs—particularly if OBI is going to be determined by the percentage of true implicit citations:

- Estimating OBI solely by relying on the database record+list of references (as in searching the Web of Science) is likely to underestimate the breadth of influence of the concept of interest (missing the majority of papers) and overestimate the degree of Obliteration, due to the inability to identify indirect citations (abstract availability may also be an issue, since there's less searchable text for the catchphrase to appear in). But many thousands of records can be processed and subject areas/journals identified for a more detailed study.
- Estimating OBI by Citation-in-Context analysis of articles retrieved in database record searches (record+text) may underestimate OBI to the extent that inclusion of the catchphrase in the Title or Abstract field is an author's signal that the article deals with the concept at some length (and more likely than not with an explicit or indirect citation to a useful source). Passing mentions and focused discussion with no reference are missed because they're not retrieved. Fewer articles are likely to be processed, if only because analysis is much more labor-intensive. but indirect citations and false positive citations to the original sources can be identified.
- Estimating OBI through examination of articles retrieved by a full-text search can address the overestimation of record-only analysis and underestimation of record+text analysis. Appropriately broad-spectrum full-text databases are less available than similarly diverse bibliographic databases (affecting the ability to generalize), and the labor-intensive issues increased since the full-text retrieval is likely to be much larger than a record-level retrieval.

There no simple answer – just a recommendation to take into account the effects of data source and level of analysis when reporting results and comparing them with prior work.

Acknowledgments

Some results from this study were presented at Metrics 2012: Workshop on Informetric and Scientometric Research (SIG MET) at the 2012 ASIST Annual Meeting in Baltimore, MD. I thank the workshop participants for helpful comments and suggestions.

References

Barabási, A-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509-512

Burrell, Q. L. (2012). Alternative thoughts on uncitedness. *Journal of the American Society for Information Science & Technology, 63*, 1466-1470

Costas, R., van Leeuwen, T. N. & van Raan, A. F. J. (2010). Is scientific literature subject to a "sell-by-date?" A general methodology to analyze the "durability" of scientific documents. *Journal of the American Society for Information Science & Technology, 61*, 329-339

Garfield, E. (1977). Introducing *Citation Classics:* The human side of scientific reports. *Essays of an Information Scientist, 3*, 1-2

Garfield, E. (1979). *Citation Indexing: Its Theory and Application in Science, Technology, and Humanites.* New York: John Wiley & Sons

MacRoberts, M. H. & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, *40*, 342-349

McCain, K.W. (2011). Eponymy and obliteration by incorporation: The case of the "Nash Equilibrium." *Journal of the American Society for Information Science & Technology, 62,* 1412-1424.

McCain, K. W. (2012). Assessing Obliteration by Incorporation: Issues and Caveats. *Journal of the American Society for Information Science & Technology, 63*, 2129-2139.

McCain, K.W. (2013). Obliteration by Incorporation. In B. Cronin & C. Sugimoto (Eds.), *Beyond Bibliometrics: Metrics-based Evaluation of Research,* Cambridge, MA: MIT Press. In review.

Merton, R. K. (1965). *On the Shoulders of Giants: A Shandean Perspective.* Chicago: University of Chicago Press

Merton, R. K. (1968). The Matthew effect in science. *Science* ,*159*, 56-63

Merton, R. K. (1979). Forward. In *Citation Indexing: Its Theory and Application in Science, Technology, and Humanites* by Eugene Garfield (p. ix). New York: John Wiley & Sons.

Merton, R. K. (1988). The Matthew effect in science. 2. Cumulative advantage and the symbolism of intellectual property. *Isis*, *79*, 606-623

Messeri, P. (1978). Obliteration by incorporation: Toward a problematics, theory and metric of the use of scientific literature. Paper presented at the annual meeting for the American Sociological Association, San Francisco, California, September 5.

Price, D. deS. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science & Technology,* 27, 292-306.
Rousseau, R. (1987). The gozinto theorem: Using citations to determine influences on a scientific publication. *Scientometrics, 11*, 217-229.
Simon, H.A. (1957). *Models of Man: Social and Rational*. New York: Wiley
Small, H. G. (1978). Cited documents as concept symbols. Social Studies of Science, 8, 327-340.
Thomas, K. S. (1992). The development of eponymy: A case study of the Southern blot. *Scientometrics, 24*, 405-417.
Van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics, 59*, 467-472
Zuckerman, H. A. (1987). Citation analysis and the complex system of intellectual influence. *Scientometrics,* 12, 329-338.

ASSESSING THE MENDELEY READERSHIP OF SOCIAL SCIENCES AND HUMANITIES RESEARCH

Ehsan Mohammadi[1] and Mike Thelwall[2]

[1] *e.mohammadi@wlv.ac.uk*
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.

[2] *m.thelwall@wlv.ac.uk*
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.

Abstract
There is some evidence that counting the readers of an article in the social reference site, Mendeley, may help to capture the research impact of the article, but the extent to which this is true for different scientific fields is unknown. This study compares Mendeley readership counts with citation counts for different social sciences and humanities disciplines. Mendeley usage data is also used as a novel way to discover patterns of information flow between scientific subjects. The overall correlation between Mendeley readership counts and citations for the social sciences was higher than for the humanities. Low and medium correlations between Mendeley readership and citation counts in all the investigated disciplines suggest that these measures reflect different aspects of research impact. The information flow findings indicate that most users of social sciences and humanities papers are from within the same discipline but some less obvious relationships between scientific disciplines were also discovered. Thus, Mendeley readership can complement citation metrics in many disciplines to help measure broader research impact and to uncover relationships between scholarly disciplines from the reader's perspective.

Keywords: Mendeley, beyond impact, altmetrics

Conference Topic
Scientometrics Indicators (Topic 6), Old and New Data Sources for Scientometric Studies (Topic 2), Webometrics (Topic 2)

Introduction
Research evaluators have often attempted to measure the impact of academic publications. Traditionally, librarians and information professionals have used re-shelving statistics to examine the value of scholarly artefacts (Blecic, 1999) but this is not useful for individual journal articles. The provision of large-scale citation data by the Institute for Scientific Information (ISI), now Thomson Reuters), paved the way for a significant change in the investigation of scholarly commutation and research evaluation. However, citation analysis is restricted to

measuring the impact of publications from an author's perspective but an article could be useful for other contexts such as teaching, commercialisation, and daily working life (Schloegl & Stock, 2004; Haustein & Siebenlist, 2011). In particular, citation metrics are more appropriate for the evaluation of theoretical publications than for applied research. Moreover, there is a worry that a new generation of authors could believe that "citation analysis is a waste of time because authors do not adequately cite those who have influenced their work" (Garfield, 2011).

During the last decade, usage data have been proposed to measure scientific impact to complement citation analysis (Rowlands & Nicholas, 2005; Bollen, Van De Sompel, Smith, & Luce, 2005; Schloegl & Gorraiz, 2011). Usage statistics are able to capture broader research activities (Kurtz & Bollen, 2010) and are obtainable earlier (Brody, Harnad, & Carr, 2006) than citation indicators. As a result, several novel metrics have been suggested based on download data for measuring the impact of scientific publications (Bollen, Van De Sompel, Hagberg, & Chute, 2009). However, most investigations have employed local usage data since global usage statistics are hidden by commercial publishers (Schloegl & Gorraiz, 2010) for privacy and marketing issues. The value of a download also depends on who accessed an article and how it was used (Thelwall, 2012). Moreover, the availability of an article through multiple platforms (Rowlands & Nicholas, 2007) and "data aggregation" are other challenges for accurate usage data (Haustein & Siebenlist, 2011).

The altmetric movement aims to capture new and previously invisible types of impacts of scholarly publications based on crowdsourcing data in social web platforms like blogs, microblogs, social bookmarking tools and online reference managers (Priem, Taraborelli, Groth, & Neylon, 2011). Data collection for altmetrics can often be based on open APIs (Priem, Piwowar, & Hemminger, 2012) which are faster and more accessible than classical usage data and are easy to integrate together (Priem et al., 2011). Amongst web 2.0 platforms, social bookmarking tools, such as CiteULike, Connotea and BibSonomy, may help to overcome the lack of global and "publisher-independent" usage data (Haustein & Siebenlist, 2011). A particularly promising example is Mendeley, a social reference manager that claims to have 2 million users and a database 45 times larger than CiteULike.

Although there has been much discussion about the value of Mendeley as an altmetric source (Priem & Hemminger, 2010; Bar-Ilan et al., 2012; Bar-Ilan, 2012), it has still not been fully evaluated. Previous investigations have found a correlation between Mendeley readership and citation counts for *Nature* and *Science* articles (Li, Thelwall, & Giustini, 2012) and for Genomics and Genetics articles (Li & Thelwall, 2012) but no study so far has examined the relationship between the two measures across different disciplines. The present research addresses this issue by assessing whether the relationship between Mendeley readership and citation counts varies across different social sciences and humanities disciplines. Social sciences and humanities studies are not cumulative and topics are not globally agreed in these disciplines (Becher & Trowler, 2001);

thus citation analysis is less effective for estimating research performance in these areas than in the hard sciences (Nederhof, 2006). As a result, developing appropriate indicators for the research evaluation of the social sciences and humanities has been important for the last three decades (Moed, Linmans, & Nederhof, 2009). Additionally, "usage metrics" are reasonable measures for fields such as social science and humanities with many pure readers (Armbruster, 2008). Moreover, "cross-disciplinary citations" are routinely used to measure the information flow from one discipline to another, but this is not ideal (Rinia, Van Leeuwen, Bruins, Van Vuren, & Van Raan, 2002) due to the inherent limitations of citation analysis. Thus, another objective of this study is to examine whether Mendeley can reflect information flow across different scientific disciplines from the users' perspectives.

Research questions

Although previous studies have found significant moderate correlations between citations and Mendeley readership counts for specific sets of articles, it seems that no previous research has investigated the relationship between Mendeley readership counts and citation measures in a range of specific disciplines. This is important because the citation behaviours of disciplines are known to vary and so Mendeley readership counts may not always correlate with citation counts. The current research partly fills this gap by investigating the correlation between Mendeley readership and citation counts for different social sciences and humanities disciplines. Additionally, measuring knowledge transfer through citation analysis is restricted to author activities while many other scholars, such as students and practitioners, are consumers of research papers. In this study, we also use Mendeley readership data to discover relationships between social sciences and humanities disciplines. The following research questions drive the investigation.

1. Are there significant, substantial and positive correlations between Mendeley readership counts and citation measures in all social sciences and humanities disciplines? If so, are there significant differences between disciplines?
2. Can Mendeley readership reveal patterns of information flow between disciplines?

Related Research

Bookmarking and Mendeley

Social web services connect people (Ding et al., 2009) as well as documents. Scholars can now communicate via web 2.0 products, including social bookmarking tools, Twitter, blogs, and wikis. These tools are potential sources for measuring the impact of scholarly publications at the article and journal levels though many aspects of these social platforms are unknown (Eysenbach, 2011). Altmetrics, a subdivision of scientometrics and webometrics, tries to identify new

metrics based on scholars' activities in online platforms for research evaluation (Priem, Groth, & Taraborelli, 2012). This new approach complements traditional methods and aims to cover broader scientific activities through expanding audiences and using new information sources (Bar-Ilan et al., 2012; Priem, Piwowar, & Hemminger, 2012). In particular, the new generation of personal reference manager tools could provide valuable data for article-level metrics (Neylon & Wu, 2009).

Social bookmarking tools allow users to save and distribute various information resources (Arolas & Ladrón-de-Guevar, 2012). A survey of recent authors found that around 7% of participants used social bookmarking systems (Mark Ware Consulting, 2008). Haustein & Siebenlist (2011) used bookmarking data for 45 physics journals from CiteULike, Connotea and BibSonomy in order to evaluate journals. They defined several indicators based on the bookmarking data. Significant correlations between measures derived from social bookmarking and JIFs (Journal Impact Factors) indicated that social bookmarking data are valuable and could be a useful source for evaluating journals from the reader's perspective. Comparing Mendeley and CiteULike user counts with WoS and Google scholar citation counts for 1613 articles of Nature and Science in 2007, Li, Thelwall and Giustini (2011) found significant correlations between the new measures and citation counts and concluded that Mendeley was more appropriate than CiteULike for research assessment in the studied sample. Bar-Ilan (2012) compared WoS, GS and Scopus citation counts for JASIST between 2001 and 2010 with Mendeley readership counts. Moderate correlations of around 0.5 suggested that "reading and citing are two different scientific activities". Li and Thelwall (2012) examined the relationship between citation measures and two altmetric indicators, Mendeley readership and F1000 article factors (a post-publication peer review score) for a sample of Genomics and Genetics articles published in 2008 that were reviewed by F1000 Faculty Members. They found significant correlations between citation counts and the two altmetric measures. The correlations were stronger for Mendeley readership counts than for FFa scores: evidence for a closer relationship between Mendeley readership and classical citation impact. A comparison between social bookmarking data for PLoS articles with other metrics showed that there was enough data in social media about biomedicine articles for research evaluation purposes (Priem et al., 2012).

Interdisciplinary Knowledge transfer

Science policymakers and funders sometimes promote interdisciplinary research between scholars to overcome sophisticated research problems (Levitt & Thelwall, 2011) and cross-fertilization seems also to be a vital element in modern science (Morillo, Bordons, & Gómez, 2003). Thus, researchers may use publications from outside their disciplines more (Bordons, Morillo, & Gómez, 2005) and it is therefore increasingly important to study the information flow between disciplines. Interdisciplinarity can be conceptualised in two different

ways, *big* and *small* (Rinia, 2007). Small interdisciplinarity deals with interactions between sub-disciplines while big interdisciplinarity refers to relations between different disciplines. It seems that some disciplines are mainly "donors" while others are "receptors" (Pair, 1980).

This review covers studies of different aspects of interdisciplinarity in social sciences and humanities disciplines. Urata (1990) used expert migration and citation flows to identify relationships between social science and humanities disciplines in Japan. The results revealed that sociology and education imported many ideas from other disciplines while psychology, linguistics, philosophy and history exported to other areas. For the social sciences, Gingras and Larivière (2010) found that interdisciplinary decreased from 1965 to 1992, but rose sharply after 1994. Levitt and Thelwall (2011) investigated changes of interdisciplinarity in social sciences disciplines in 1990 and 2000 with similar results: interdisciplinarity diminished between 1980 and 1990 but increased strongly from 1990 to 2000.

Stevens (1990) examined the relationship between planning (Krueckeberg, 1985) and other social sciences disciplines. He found that half of the planning information was from economics whereas geography, environmental studies and economics were the main users of planning publications. An investigation into articles from the four main journals of sociology and political science indicated that the boundaries of these disciplines were not limited (Pierce, 1999). Goldstone and Leydesdorff, (2006) claimed that cognitive science, as an interdisciplinary subject, is like a hub for knowledge exchange between computer science, neuroscience, psychology and education. Cognitive science articles were often used by computer scientists while cognitive science researchers cited psychology publications more. Neeley (1981) applied citation analysis to measure the relationship of management to other social sciences fields, finding that management scholars often cited other disciplines but not vice versa. Another study of management journals revealed that this field was a significant donor for psychology while a large amount of information was imported from economics, psychology, and sociology (Lockett & McWilliams, 2005). Bedeian, (2005) argued that drawing a large amount of information from other disciplines shows a good level of integration with them. Cronin and Pearson (1990) analysed citations to the scholarly artefacts of some senior information scientists and found that few of these publications were used by scholars from outside of the field. Conversely, results of an empirical study in 2005 showed that the pattern of LIS research has changed in terms of interdisciplinarity and LIS articles have been cited by several other disciplines (Tang, 2005). Cronin and Meho (2007) used large-scale data to re-examine the conclusions of Cronin and Pearson (1990), finding that information science transferred ideas to other disciplines more and used publications from computer science, engineering, and business and management more in the last decade. Recently, information science and library science has had the highest increase in interdisciplinarity among the social sciences disciplines (Levitt & Thelwall, 2011).

Data collection

We used two search queries (appendix 1) in the Social Science Citation Index (SSCI) and the Arts and Humanities Citation Index (AHCI) to retrieve all social sciences and humanities publications indexed by Web of Science (WoS) in two separate searches. The results were limited to research articles in English only (reports, editorials, book reviews, etc. removed) from 2008. The year 2008 was selected because the peak time for citations is usually three years after an article is released (Moed, 2005).

In order to classify the results into social sciences and humanities disciplines, we used the ISI subject categories. We used citation counts for each article based on the WoS data at the time of data collection (August 2012).

Table1. Coverage of articles from social sciences and humanities disciplines in Mendeley

Disciplines	*Articles indexed by WoS in 2008*	*Unique articles covered by Mendeley*	*Unique articles with readership statistics*	*Articles without readership statistics*
Psychology	23,811	14,757 (62%)	12,804 (54%)	1,953 (8%)
Interdisciplinary social sciences	6,366	3,763 (59%)	2,416 (38%)	1,347 (21%)
Education and educational research	7,208	3,839 (53%)	2,796 (39%)	1,043 (14%)
Library and information science	2,552	1,617 (63%)	1,343(53%)	274 (10%)
Business and Economics	22,710	12,337 (54%)	8,199 (36%)	4,138 (18%)
Total	62,647	36,313 (58%)	27,558 (44%)	8,755 (14%)
Philosophy	2,833	1,060 (37%)	468 (17%)	592 (21%)
History	2,882	756 (26%)	253 (9%)	503 (17%)
Linguistics	2,245	1,046 (47%)	773 (34%)	273 (12%)
Literature	4,622	643 (14%)	165 (4%)	478 (10%)
Religion	2,058	640 (31%)	255 (12%)	385 (19%)
Total	14,640	4,145 (28%)	1,914 (13%)	2,231 (15%)

We used Webometric Analyst (lexiurl.wlv.ac.uk) to automatically extract Mendeley data for the selected articles via the Mendeley API (Application Programing Interface). As multiple versions of an article sometimes exist in Mendeley, we identified duplicate records based on Mendeley unique IDs, Mendeley URLs, DOIs and probable duplications were checked and removed manually. In the case of duplication, records with the fewest readers were excluded. Out of 41,624 Mendeley records, 1,166 records (3%) were discovered to be duplicates. Some of the articles in the Mendeley catalogue did not have readership statistics and instead of statistical data the phrase “Readership statistics

are being calculated" is displayed. Perhaps Mendeley loaded these articles straight from the publishers' websites or some of the users added own publications to their Mendeley profiles but no one had saved these articles in a personal library. Most of the records removed due to duplication did not have readership statistics. Table 1 shows that 44% of the articles from the chosen social sciences were in the Mendeley catalogue in comparison only 13% of the humanities articles. Library and information science (53%) and linguistics (34%) had the highest coverage in Mendeley among other social sciences and humanities disciplines respectively. Education (39%) and Literature (4%) had the lowest percentage of articles in the Mendeley database. Therefore, 27,558 and 1,914 articles of the social science and humanities disciplines, respectively, which had Mendeley readership statistics were used in this study. Spearman correlation tests were applied to the ISI citations and Mendeley readership counts. Spearman correlation was used rather than Pearson correlation because the frequency distributions of readership and citation counts were skewed.

Findings

Table 2 shows that there is a significant correlation between Mendeley readership and citation counts in all the investigated disciplines. The correlation for social sciences disciplines overall (0.516) is higher than for humanities disciplines (0.428). There were moderate correlations for social sciences disciplines, varying from 0.403 (interdisciplinary social sciences) to 0.573 (business and economics). Amongst humanities disciplines, religion and philosophy have the lowest correlations (0.363 and 0.366) and linguistics has the highest correlation (0.454).

Table 2. Descriptive statistics and correlations between citations and Mendeley readership counts for articles from 2008 with Mendeley readership statistics in different social sciences and humanities disciplines

Disciplines	*WoS citation median*	*Mendeley reader-ship median*	*Correlation (Spearman's rho)*
Psychology	6.00	6.00	.514**
Interdisciplinary social sciences	4.00	4.00	.403**
Education	4.00	6.00	.484**
Library and information science	4.00	8.00	.535**
Business and Economics	5.00	7.00	.573**
All social sciences	5.00	6.00	.516**
Philosophy	1.00	4.00	.366**
History	1.00	2.00	.428**
Linguistics	2.00	4.00	.454**
Literature	0.00	2.00	.403**
Religion	1.00	3.00	.363**
All Humanities	1.00	3.00	.428**

** Significant at the p = 0.01 level

We explored cross-disciplinary readership as an indication of information flow between disciplines based on users' research backgrounds in their Mendeley profiles. Complete statistical data related to readers' background disciplines for each individual article are not accessible through the Mendeley API because only the three most common readers' background disciplines are revealed. The data are provided in percentile format. For each article and each of the three readers' disciplines, we multiplied the percentage of readers from that discipline with the total number of readers of the article and divided by 100 to obtain the estimated number of article readers from that discipline. This process covered 89% and 82% of the readers' background disciplines for social science and humanities articles.

Table 3. Interdisciplinary readership for social sciences disciplines in Mendeley

Read by / Discipline	*Psychology*	*Interdisciplinary social sciences*	*Education*	*LIS**	*Business and Economics*
Psychology	**64.00%**	**15.80%**	**12.40%**	1.80%	**6.50%**
Social Sciences	**6.50%**	**27.80%**	**7.40%**	**20.50%**	**11.60%**
Education	3.80%	5.40%	**54.40%**	4.40%	1.00%
Business& Economics	3.50%	**11.60%**	1.90%	**14.00%**	**55.70%**
Management	0.90%	3.10%	0.50%	3.50%	**11.00%**
Computer and Information Science	3.10%	4.50%	**9.00%**	**45.90%**	4.70%
Medicine	**6.10%**	**7.70%**	4.90%	3.10%	1.00%
Biological Sciences	**6.60%**	4.50%	1.70%	1.40%	1.50%
Philosophy	0.40%	4.50%	0.20%	0.10%	0.10%
Linguistics	1.90%	0.10%	3.00%	0.20%	0.00%
Arts and Literature	0.20%	0.80%	0.40%	0.30%	0.00%
Others	2.90%	14.20%	4.10%	4.70%	6.90%
Total	112898	13436	20817	13000	74080

*LIS=library and information science.

From Table 3 the majority of readers of all investigated social sciences disciplines are from the home disciplines, except for library and information science and interdisciplinary social sciences. However, the percentages vary across different disciplines, from psychology (64%) to interdisciplinary areas of social sciences (28%). This suggests that most Mendeley readers use scientific information mainly from their own disciplines but that this varies substantially between disciplines.

Table 4. Interdisciplinary readership for Humanities disciplines in Mendeley

Read by / Discipline	*Philosophy*	*History**	*Linguistics*	*Literature*	*Religion**
Philosophy	**32.10%**	4.00%	1.20%	0.90%	**6.60%**
Humanities	**7.20%**	**31.70%**	4.70%	**27.80%**	**23.10%**
Linguistics	2.60%	0.70%	**55.00%**	1.20%	2.50%
Arts and Literature	2.60%	3.80%	2.50%	**27.30%**	1.70%
Social Sciences	**12.40%**	**39.60%**	**7.80%**	**20.60%**	**26.90%**
Psychology	**15.60%**	**6.50%**	**8.40%**	1.30%	**21.40%**
Education	3.70%	2.40%	**7.90%**	2.60%	6.40%
Business Administration	1.10%	1.20%	0.10%	1.00%	1.10%
Medicine	2.42%	0.70%	0.50%	1.00%	3.40%
Biological Sciences	5.00%	0.70%	0.90%	0.60%	2.30%
Computer and Information Science	**6.50%**	2.80%	**9.30%**	**10.10%**	1.10%
Others	8.80%	5.90%	1.70%	5.60%	3.50%
Total	1153	911	3760	650	812

*History and religion have been categorized as a humanities sub-discipline in Mendeley.

Also from Table 3, very few psychology articles have an arts and humanities readership while some psychology literature is read by people from biology (7%) and medicine (6%) perhaps reflecting uses of psychology within biomedicine.
The research backgrounds of many readers of articles of library and information science (46%) are computer and information scientists who mainly focus on computer science rather than library science. Moreover, 21% of the library and information science publications were read by individuals from social sciences disciplines.
Table 4 shows that the most readers of philosophy (32%), linguistics (55%) and literature (27%) are from the same discipline but the majority of users of historical (40%) and religious (27%) articles were from the social sciences.

Discussion

This research examined Mendeley usage data for social sciences and humanities publications from 2008. Spearman correlation tests found positive correlations between Mendeley readership counts and citation counts for all the studied disciplines but the values varied across disciplines. The overall correlation for the social sciences (0.516) was higher than for the humanities (0.428). Some social sciences and humanities disciplines are similar to natural and life sciences fields with a high volume of citations while others resemble classical humanities with a lower citation rate (Nederhof, Zwaan, Bruin, & Dekker, 1989). The higher correlations between Mendeley readership and citation counts are in those disciplines that are closer to hard sciences in terms of citation behaviour while the correlations are lower in the disciplines which more resemble traditional humanities.

The median Mendeley readership counts were higher than the median citation counts in all the studied disciplines except psychology. This is consistent with Mendeley readership capturing broader scholarly activities than citations, since different groups from undergraduate students to senior researchers use Mendeley in their academic activities, and corroborates the value of Mendeley readership data.

Cross-disciplinary readership was also used as evidence of knowledge transfer between social sciences, humanities, and other disciplines. Generally, most readers of the studied social science articles were from the home disciplines. Among humanities disciplines, the most readers of historical and religious papers were people with social sciences research backgrounds, however. Part of the results here may be due to the way in which Mendeley classifies people: for example not having a library and information science category but having a computer and information science category instead. The results will also reflect the size of the disciplines involved and the extent to which Mendeley is used within the disciplines. Hence, the results are likely to be skewed towards larger disciplines and biased towards disciplines using Mendeley the most actively (e.g., perhaps library and information science).

A significant amount of psychology information was read by people from biology and medicine, which is not surprising as they have common research borders. Some links were found between interdisciplinary social sciences and biomedicine as previously reported in a citation analysis study (Zhang, Glänzel, & Liang, 2009). Connections were also found between philosophy, computer and information science, and biology. In the case of library and information science, the main importing disciplines were computer and information science, business and economics, management, education and medicine. This agrees with the findings of Cronin and Meho (2007).

Our findings also illustrate that the investigated disciplines are different in terms of the diversity of relationships with other disciplines. For instance, interdisciplinary social science research areas exported ideas to more different disciplines in comparison to others.

One limitation of this research is that readership is limited to the individuals who choose Mendeley for their reference manager while many scholars use EndNote, RefWorks, and ProCite to organize their references. Another limitation is that around 11%-18% of the readers' background disciplines were excluded because they were not accessible via the Mendeley API. Additionally, our studied sample is restricted to journal articles only while books are a fundamental source of research in many humanities and some social sciences disciplines (Huang & Chang, 2008; Nederhof, 2006). However, social sciences and humanities researchers have begun to publish more in ISI ranked journals (Kyvik, 2003; Butler, 2003). Finally, the study excluded all articles that were not found in Mendeley. Whilst it seems likely that these articles will tend to attract few citations and hence the correlations found would not be much affected by adding

these articles to the correlation calculations, this has not been proven in the current paper.

Conclusions

In answer to the first research question, a significant correlation was found between Mendeley readership and citation counts in all social sciences and humanities but the correlations varied from 0.363 (religion) to 0.573 (business and economics). The overall correlation for social sciences is higher than for humanities. In almost all disciplines, the correlation is not strong enough to conclude that Mendeley readership and citation counts measure the same aspect of research impact. As hypothesised by pervious authors, a likely explanation is that Mendeley captures broader scholarly activities from a variety of readers' perspectives in comparison to citation counts. Hence, Mendeley readership data could be a useful supplementary measure to remedy some limitations of citation analysis across the social sciences and humanities. If Mendeley readership data is to be used for important evaluations, however, then steps would need to be taken to ensure that the results cannot be manipulated by those with a vested interest in a particular outcome.

In answer to the second question, our results reveal that patterns of exporting information from social sciences and humanities disciplines to other disciplines can be extracted based on Mendeley readership and agree to some extent with previous citation-based studies. This agreement is some evidence that the results are not random. Nevertheless, other sources of evidence (e.g., questionnaires) would be needed to fully assess the meaning of these results. Mendeley data could thus capture obvious and less obvious relationships between scientific disciplines. The possibility of identifying inter-disciplinary information flows based on Mendeley usage data provides a new way to measure research influences across disciplines. Mendeley and citation sources together may also provide better insights into the relationships between disciplines.

Appendix 1: Search queries for retrieving social science and art and humanities articles from WoS.

(SO=(A* OR B* OR C* OR D* OR E* OR F* OR G* OR H* OR I* OR J* OR K* OR L* OR M* OR N* OR O* OR P* OR Q* OR R* OR S* OR T* OR U* OR V* OR W* OR X* OR Y* OR Z* OR 0* OR 1* OR 2* OR 3* OR 4* OR 5* OR 6* OR 7* OR 8* OR 9*) AND (PY=2008)) AND Language=(English) AND Document Types=(Article) Timespan=All Years. Databases=SSCI.

(SO=(A* OR B* OR C* OR D* OR E* OR F* OR G* OR H* OR I* OR J* OR K* OR L* OR M* OR N* OR O* OR P* OR Q* OR R* OR S* OR T* OR U* OR V* OR W* OR X* OR Y* OR Z* OR 0* OR 1* OR 2* OR 3* OR 4* OR 5* OR 6* OR 7* OR 8* OR 9*) AND AND (PY=2008)) AND Language=(English) AND Document Types=(Article) Timespan=All Years. Databases=A&HCI.

References

Armbruster, C. (2008). Access, Usage and Citation Metrics: What Function for Digital Libraries and Repositories in Research Evaluation? *Social Science Research Network Working Paper Series*, *128*(Pt 6), 1407–17. SSRN. Retrieved December 2, 2012, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1088453.

Arolas, E. E., & Ladrón-de-Guevar, F. G. (2012). Uses of explicit and implicit tags in social bookmarking. *Journal of the American Society for Information Science and Technology*, 63(2), 313–322. doi:10.1002/asi.21663

Bar-Ilan, J, Haustein, S., Peters, I., Priem, J., Shema, H., & Terliesner, J. (2012). Beyond citations: Scholars' visibility on the social Web. *17th International Conference on Science and Technology Indicators*. Retrieved from http://arxiv.org/ftp/arxiv/papers/1205/1205.5611.pdf

Bar-Ilan, J. (2012). JASIST 2001–2010. *Bulletin of the American Society for Information Science and Technology*, 24–28.

Becher, T., & Trowler, P. (2001). *Academic tribes and territories (2ed)*. Milton Keynes, UK: Open University Press.

Bedeian, A. G. (2005). Crossing Disciplinary Boundaries: A Epilegomenon for Lockett and McWilliams. *Journal of Management Inquiry*, *14*(2), 151–155.

Blecic, D. D. (1999). Measurements of journal use: an analysis of the correlations between three methods. *Bulletin of the Medical Library Association*, 87(1), 20–5.

Bollen, J., Van De Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. (T. Mailund, Ed.) *PLoS ONE*, 4(6), e6022. doi:10.1371/journal.pone.0006022.

Bollen, J., Van De Sompel, H., Smith, J. A., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management*, 41(6), 1419–1440. doi:10.1016/j.ipm.2005.03.024

Bordons, M., Morillo, F., & Gómez, and I. (2005). analysis of cross-disciplinary research through bibliometric tools. In Henk F. Moed (Ed.), *Handbook of Quantitative Science and Technology Research* (pp. 437–456). Kluwer Academic Publishers.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Society for Information Science and Technology*, 57(8), 1060–1072.

Butler, L. (2003). Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research Policy*, *32*(1), 143–155.

Cronin, B., & Meho, L. (2007). The shifting balance of intellectual trade in information studies. *Journal of the American Society for Information Science and Technology*, *59*(4), 551–564.

Cronin, B., & Pearson, S. (1990). The export of ideas from information science. *Journal of Information Science*, *16*(6), 381–391.

Bar-Ilan, J. (2012). JASIST 2001–2010. Bulletin of the American Society for Information …, 24–28. Retrieved November 5, 2012, from http://onlinelibrary.wiley.com/doi/10.1002/bult.2012.1720380607/full

Ding, Y., Jacob, E. K., Zhang, Z., Foo, S., Yan, E., George, N. L., & Guo, L. (2009). Perspectives on social tagging. Journal of the American Society for Information Science and Technology, 60(12), 2388–2401.

Eysenbach, G. (2011). Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. Journal of Medical Internet Research, 13(4), e123. doi:10.2196/jmir.2012

Garfield, E. (2011). Full Text downloads and citations: Some reflections. Keynote lecture at the Seminar "Scientific Measurement and Mapping." Santa Fe, New Mexico. Retrieved from http://www.garfield.library.upenn.edu/papers/santafe2011.pdf

Gingras, Y., & Larivière, V. (2010). The historical evolution of interdisciplinarity: 1900-2008. *Eleventh International Conference on Science and Technology Indicators*. Leiden.

Goldstone, R. L., & Leydesdorff, L. (2006). The import and export of cognitive science. *Cognitive science*, *30*(6), 983–93.

Haustein, S., & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. Journal of Informetrics, 5(3), 446–457. doi:10.1016/j.joi.2011.04.002

Huang, M., & Chang, Y. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, *59*(11), 1819–1828.

Krueckeberg, D. A. (1985). The Tuition of American Planning From Dependency toward Self-Reliance. *The Town Planning Review*, *56*(4), 421–441.

Kurtz, M., & Bollen, J. (2010). Usage bibliometrics. *Annual review of information science and Technology*. Retrieved January 4, 2013, from http://onlinelibrary.wiley.com/doi/10.1002/aris.2010.1440440108/full

Kyvik, S. (2003). Changing trends in publishing behaviour among university faculty, 1980-2000. *Scientometrics*, *58*, 35–48.

Levitt, J., & Thelwall, M. (2011). Variations between subjects in the extent to which the social sciences have become more interdisciplinary. *Journal of the American Society for Information Science and Technology*, *62*(6), 1118–1129.

Li, X, Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics, 91(2),* 461-471.

Li, X, & Thelwall, M. (2012). F1000, Mendeley and Traditional Bibliometric Indicators. *17th International Conference on Science and Technology Indicators* (Vol. 3, pp. 1–11).

Lockett, A., & McWilliams, A. (2005). The Balance of Trade Between Disciplines: Do We Effectively Manage Knowledge? *Journal of Management Inquiry*, *14*(2), 139–150.

Mark Ware Consulting. (2008). Peer review in scholarly journals: Perspective of the scholarly community an international study. Information Services and Use - *APE 2008 Academic Publishing in Europe, Quality and Publishing,* 28(2), 109–112.

Moed, H F. (2005). *Citation analysis in research evaluation* (Vol. 9). Kluwer Academic Pub.

Moed, H F., Linmans, J., & Nederhof, A. (2009). Options for a comprehensive database of research outputs in Social Sciences & Humanities. Retrieved January 4, 2013, from http://83.143.5.70/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/annex_2_en.pdf

Morillo, F., Bordons, M., & Gómez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology*, *54*(13), 1237–1249.

Nederhof, A. J., Zwaan, R. A., Bruin, R. E., & Dekker, P. J. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social and behavioural sciences: A comparative study. *Scientometrics*, 15(5-6), 423–435. doi:10.1007/BF02017063

Nederhof. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review. *Scientometrics*, *66*(1), 81–100.

Neeley, J. D. (1981). The management and social science literatures: An interdisciplinary cross-citation analysis. *Journal of the American Society for Information Science*, *32*(3), 217–223.

Neylon, C., & Wu, S. (2009). Article-level metrics and the evolution of scientific impact. *PLoS biology*, 7(11), e1000242. doi:10.1371/journal.pbio.1000242

Pair, C. (1980). Switching between academic disciplines in universities in the Netherlands. *Scientometrics*, *2*(3), 177–191. Retrieved October 28, 2012, from http://www.springerlink.com/index/10.1007/BF02016696

Priem, J., & Hemminger, B. M. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, 15(7). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874

Priem, J., Groth, P., & Taraborelli, D. (2012). The Altmetrics Collection. *PLoS ONE,* 7(11), e48753. doi:10.1371/journal.pone.0048753

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *Arxiv preprint arXiv*:1203.4745. Retrieved from http://arxiv.org/html/1203.4745v1

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). alt-metrics: A manifesto (v 1.01 – September 28, 2011: removed dash in alt-metrics). http://altmetrics.org/manifesto.

Rinia, E. . (2007). *Measurement and evaluation of interdisciplinary research and knowledge transfer*. Universiteit Leiden, The Netherlands.

Rinia, Van Leeuwen, T., Bruins, E., Van Vuren, H., & Van Raan, A. (2002). Measuring knowledge transfer between fields of science. *Scientometrics*,

54(3), 347–362. Akadémiai Kiadó, co-published with Springer Science+Business Media B.V., Formerly Kluwer Academic Publishers B.V.
Rowlands, I., & Nicholas, D. (2005). Scholarly communication in the digital environment: The 2005 survey of journal author behaviour and attitudes. *Aslib Proceedings*, 57(6), 481–497. doi:10.1108/00012530510634226
Rowlands, I., & Nicholas, D. (2007). The missing link: journal usage metrics. *Aslib Proceedings*, 59(3), 222–228. doi:10.1108/00012530710752025
Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: the case of oncology journals. *Scientometrics*, 82(3), 567–580. doi:10.1007/s11192-010-0172-1
Schloegl, C., & Gorraiz, J. (2011). Global usage versus global citation metrics: The case of pharmacology journals. *Journal of the American Society for Information Science and Technology*, *62*(1), 161–170.
Schloegl, C., & Stock, W. G. (2004). Impact and relevance of LIS journals: A scientometric analysis of international and German-language LIS journals - Citation analysis versus reader survey. *Journal of the American Society for Information Science and Technology*, 55(13), 1155–1168. doi:10.1002/asi.20070
Stevens, G. (1990). An Alliance Confirmed Planning Literature and the Social Sciences. *Journal of the American Planning Association*, *56*(3), 341–349.
Tang, R. (2005). Evolution of the interdisciplinary characteristics of information and library science. *Proceedings of the American Society for Information Science and Technology*, *41*(1), 54–63.
Thelwall, M. (2012). Journal impact evaluation: a webometric perspective. *Scientometrics*, 92(2), 429-441. doi:10.1007/s11192-012-0669-x
Urata, H. (1990). Information flows among academic disciplines in Japan. *Scientometrics*, *18*(3-4), 309–319.
Zhang, L., Glänzel, W., & Liang, L. (2009). Tracing the role of individual journals in a cross-citation network based on different indicators. *Scientometrics*, 81(3), 821–838. doi:10.1007/s11192-008-2245-y

ASSOCIATION BETWEEN QUALITY OF CLINICAL PRACTICE GUIDELINES AND CITATIONS GIVEN TO THEIR REFERENCES

Jens Peter Andersen[1,2]

[1] *jepea@rn.dk*
Aalborg University Hospital, Medical Library, DK-9000 Aalborg (Denmark)

[2] *jpa@iva.dk*
Royal School of Library and Information Science, DK-9220 Aalborg E (Denmark)

Abstract
It has been suggested that bibliometric analysis of different document types may reveal new aspects of research performance. In medical research a number of study types play different roles in the research process and it has been shown, that the evidence-level of study types is associated with varying citation rates. This study focuses on clinical practice guidelines, which are supposed to gather the highest evidence on a given topic to give the best possible recommendation for practitioners.
The quality of clinical practice guidelines, measured using the AGREE score, is compared to the citations given to the references used in these guidelines, as it is hypothesised, that better guidelines are based on higher cited references.
AGREE scores are gathered from reviews of clinical practice guidelines on a number of diseases and treatments. Their references are collected from Web of Science and citation counts are normalised using the item-oriented z-score and the $PP_{top\text{-}10\%}$ indicators.
A positive correlation between both citation indicators and the AGREE score of clinical practice guidelines is found. Some potential confounding factors are identified. While confounding cannot be excluded, results indicate low likelihood for the identified confounders. The results provide a new perspective to and application of citation analysis.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3)

Introduction
While most scientometric studies of research publications focus on standard journal articles, it has been suggested several times that other document types may play a role in research assessment as well (Lewison, 2002, 2003; van Leeuwen, Costas, Calero-Medina, & Visser, 2012). Lewison (2002, 2003) in particular has emphasised the possibilities of various document types in the medical fields. One document type particular to that field is the clinical practice guideline. The purpose of these guidelines is to gather the best evidence of the treatment of diseases to ensure the best possible treatment at hospitals, clinics and

general practices (The AGREE Collaboration, 2003). One might therefore expect these guidelines to build on the highest quality research available.

Studies have shown a connection between citation scores and the evidence-level of clinical study types (e.g. Andersen & Schneider, 2011; Kjaergard & Gluud, 2002; Patsopoulos, Analatos, & Ioannidis, 2005). These study types are hierarchically ordered, assigning greater importance to the evidence found in high-level studies, such as meta-analyses and randomised, controlled trials (RCT), than lower level studies, such as case studies (Greenhalgh, 2010). This hierarchy is widely applied in different areas of health research, and the connection between citations and evidence levels indicates that high-evidence studies are indeed, on average, used more than other. We might thus speculate if not references used in clinical practice guidelines are cited more on average than other papers, if the guidelines indeed represent the best available evidence on a topic.

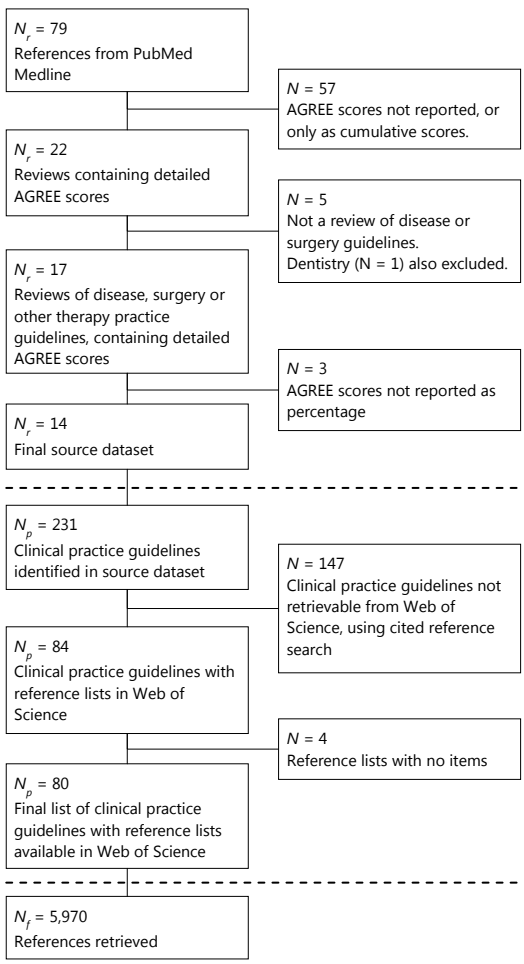


Figure 1 - Flowchart of inclusion and exclusion criteria

Several studies have indicated that clinical practice guidelines are created very differently, with great variation in scope, rigor, clinical recommendation and overall quality (e.g. Burda, Norris, Holmer, Ogden, & Smith, 2011; Ferket et al., 2010, 2011; Freel et al., 2008; Gallardo et al., 2010; Kis et al., 2010). Many of these reviews of clinical practice guidelines use the AGREE[26] instrument (The AGREE Collaboration, 2003) to assess six major aspects of guideline development. Especially one aspect, *rigor of development*, is directly related to the evidence found in the background literature.

[26] Appraisal of Guidelines for Research and Evaluation

The hypothesis presented here is that there is a positive correlation between the rigor of development AGREE score of clinical practice guidelines and the citations given to the literature references in these clinical practice guidelines. If this correlation can be observed it points at an association between clinical evidence, citations and the development of clinical practice guidelines. The association cannot be assumed to be causative, however, but would still provide valuable insights into the inclusion and interpretation of a new document type in research assessment.

Paper outline

The following section outlines the acquisition of data, from reviews of clinical guidelines (top-level) to the actual guidelines and their references. This is followed by a presentation of the citation indicators used to assess the citation impact of the references of the guidelines. These references will also be discussed further in the results section which begins with an analysis of the citations given to the references, to test if they are representative of a standard citation distribution. The section concludes with a correlation analysis of the tested indicators versus the AGREE scores. All results are discussed in the final section and known weaknesses of this study are presented and discussed with respect to the findings.

Materials and Methods

The AGREE instrument consists of 23 key items organised in six domains, with the intention of describing various aspects of guideline development. The six domains are scope and purpose; stakeholder involvement; rigor of development; clarity and presentation; applicability; and editorial independence (The AGREE Collaboration, 2003). These domain scores have been used in a number of reviews of guideline quality as a means of assessment. Each item in the six domains is rated by one or more reviewers on a 4-point scale where 1 = strongly disagree, 2 = disagree, 3 = agree and 4 = strongly agree. In some of the included studies (see below) the overall domain scores were not calculated before all reviewers agreed on one, final item score while others used combined, standardised scores [1], resulting in a more diverse score profile, but also some inter-reviewer inconsistency. In this study the domain scores are included regardless of which procedure had been used, although it could have been preferable if all scores had been standardised. The reasoning behind this decision is elaborated in the discussion.

$$\frac{obtained\ score - minimum\ possible\ score}{maximum\ possible\ score - minimum\ possible\ score} \quad [1]$$

Data collection

To obtain the AGREE scores, reviews of clinical practice guidelines were found in PubMed Medline using the query:

"practice guidelines as topic"[MESH] AND "agree"[TIAB] AND "review"[PTYP] AND "2007":"2011"[PDAT] AND "English"[LANG]*

This resulted in 79 English-language reviews of clinical practice guidelines from 2007 to 2011 with the term agree occurring in the title or abstract. Not all reviews included the detailed AGREE scores, e.g. only reporting a cumulative score although this is not recommended (The AGREE Collaboration, 2003). Also reviews of non-disease topics were excluded, as were those reviews containing pure AGREE scores rather than percentages. An overview of the selection process is presented in figure 1. The final set of reviews (N_r = 14) were used to collect references to clinical practice guidelines (N_p = 231), and to gather AGREE scores for these.

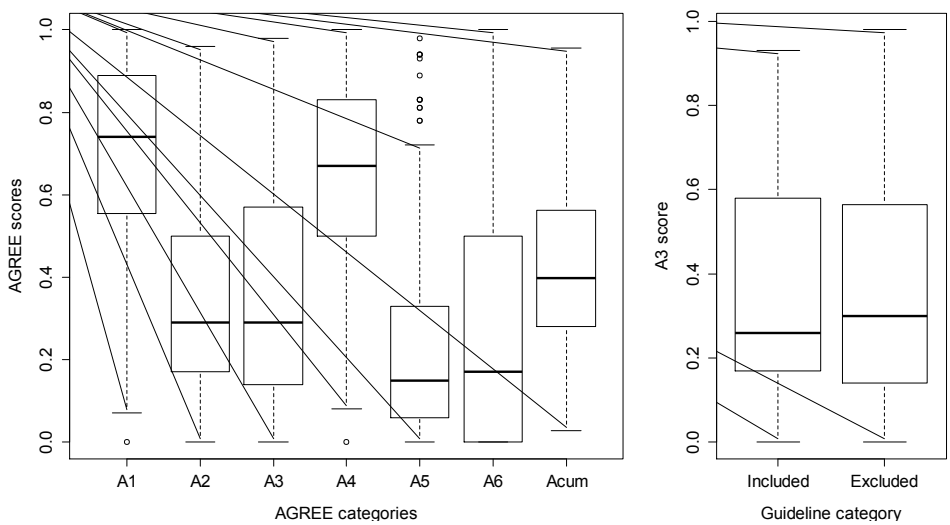


Figure 2 – AGREE scores for all studies (left); A1: scope and purpose, A2: stakeholder involvement, A3: rigor of development, A4: clarity and presentation, A5: applicability, A6: editorial independence, Acum: cumulated, weighted scores. The figure to the right shows the difference in A3 score for guidelines included in the study and those that could not be found in Web of Science (excluded).

As stated in the present hypotheses, it is assumed that there is a connection between the scores of the rigor of development domain (henceforth referred to as A3) and the citations given to references used by the clinical practice guidelines. It is therefore necessary to retrieve the reference lists of the above guidelines in a database where information about citations was also available. The Thomson Reuters Web of Science (WoS) was used for this purpose, and citations from all citation indices were included.

Many of the guidelines are not registered in WoS (N = 147), as they are often published in different channels, e.g. society websites. Those that were retrieved

and contained reference lists (N_p = 80), however, resulted in a total of 5,970 non-unique references in their reference lists. The AGREE scores of all guidelines are shown in Figure 2, illustrating the variation between the six domains, stressing the importance of using A3 rather than a cumulated score for all six domains. The cumulated score, A_{cum}, in Figure 2 was calculated as a weighted average, weighting each domain by the number of items in that domain. Figure 2 also shows the difference between the A3 score for the 80 included guidelines and for the 147 excluded guidelines. As can be seen, the difference in score is very small, and the included sample can be considered representative of the complete 231 guidelines in this respect.

The references from the guidelines were retrieved from WoS, including total number of citations per January 2nd, 2013. To provide a comparison baseline, all papers published in the same journals, the same years as the included references, were also gathered from WoS, resulting in 672,819 items.

Citation analysis

The included clinical practice guidelines were published in a number of different areas of medicine, with varying citation potentials, and their reference lists spanned publications from 1932 to 2010. To enable comparison between these items, citation counts were normalised, using two different approaches. A somewhat direct comparison method was desired, but also an excellence-method could provide different perspectives on the meaning of citation counts as a function of agree scores. For the latter, the $PP_{top\text{-}10\%}$ indicator (Waltman et al., 2012) is considered a sensible choice, as it both focuses on the 10% most highly cited papers (a form of excellence) and remains insensitive to extremely highly cited documents. For the more direct approach, a number of normalisation procedures are available, but the item-oriented *z*-score (Lundberg, 2007) has been chosen here, as it allows normalisation of single items while also incorporating standard scores. The item-oriented *z*-score for a single item, z_i, is denoted as:

$$c' = \ln(c + 1)$$

$$z_i = \frac{c'_i - \bar{c}'}{S(c')}$$

, where c_i is the number of citations given to a document published in a specific journal a specific year, taken from c, the entire distribution of citations to papers in that journal that year. $\bar{c}'$ is the average of the c' distribution and $S(c')$ is the standard deviation of the c' distribution. The cumulative z for a clinical practice guideline is the average z_i for all references in the reference list. Values of z indicate the number of standard deviations the citations diverge from the mean. Positive values indicate higher than average scores.

The $PP_{top\text{-}10\%}$ indicator is simply found by comparing the citations to each paper to those of all other papers published in the same journal and year. If the citation score is placed in the highest decile, it is counted as 1, if not as 0. The $PP_{top\text{-}10\%}$ is the average of this distribution, resulting in a percentage of papers in the reference list that are considered excellent. If this percentage is higher than 10%, the references could be considered to be more excellent than standard publications. One should however be careful to draw the same conclusions in the setting of this study than one would for e.g. the publications of a university, as is the case in the Leiden ranking (Waltman et al., 2012). As all references in a reference list are de facto cited at least once, and there is an increased chance that highly cited documents are used as references, a $PP_{top\text{-}10\%}$ over 0.1 should not be interpreted as better than average. The interpretation of whether the hypothesis of this study has merit thus depends on whether an increase in $PP_{top\text{-}10\%}$ or z can be observed as a function of A3.
Both citation indicators are calculated for individual references and subsequently cumulated for the guidelines. By doing so, both indicators enable comparison between guidelines, without bias from the length of reference lists, which varies greatly (from 4 to 627 with a median of 81).

While there is a long history of and debate about citation normalisation and excellence indicators in the scientometric literature (Beirlant, Glänzel, Carbonez, & Leemans, 2007; Glänzel & Moed, 2012; Leydesdorff & Bornmann, 2011; Leydesdorff & Opthof, 2010; Schubert & Braun, 1986, 1996; van Raan, van Leeuwen, Visser, van Eck, & Waltman, 2010; Vinkler, 1986; Zitt, 2011), the above approaches were selected for their power of interpretation and robustness. Other indicators might have provided equally useful interpretations, but it is not the aim of this paper to discuss these indicator properties.

All data extraction and calculations were performed using R version x64 2.15.2 (R Development Core Team, 2010)

Results

The citations given to references from the guidelines are plotted as a decreasing function of rank in Figure 3. The shape of the curve is as could be expected for a typical citation distribution, indicating that the sample of references is comparable to other citation distributions. The distribution of z-scores reveals that the citations are higher than the background population though, as the median z-score is 0.9, and thus almost one standard deviation higher than the expected average. This is also illustrated in Figure 4, showing the empirical cumulative density function for the z-score of all references.

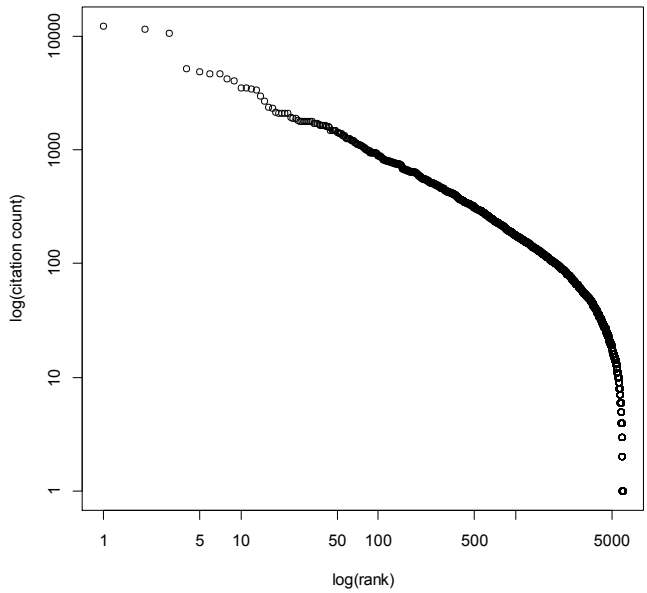


Figure 3 – citation counts as a function of rank, double-log scales

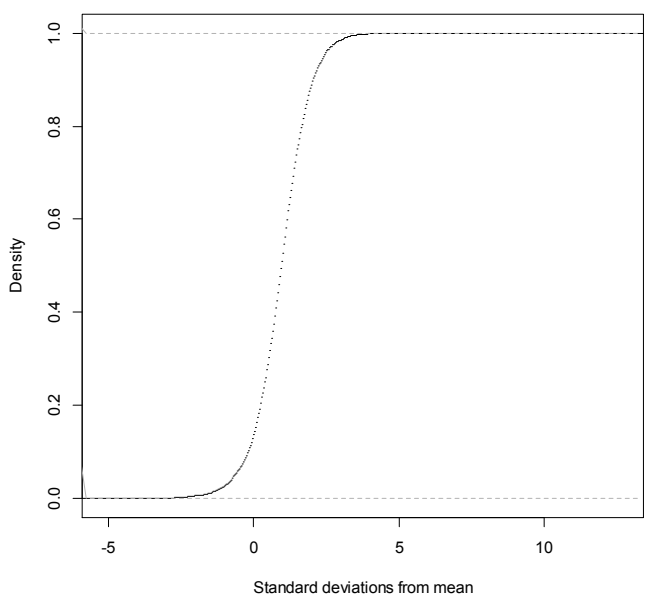


Figure 4 – empirical cumulative density function of z-scores for references

As is stated in the above, one might expect citation distributions for documents retrieved from reference lists to be higher than complete citation distributions of journals, as they represent documents that are actually used as references while a certain proportion of all journal articles remain uncited even after several years. To test the correlation between A3, *z* and PPtop-10% respectively the two citation indicators were plotted as functions of A3 in Figure 5.

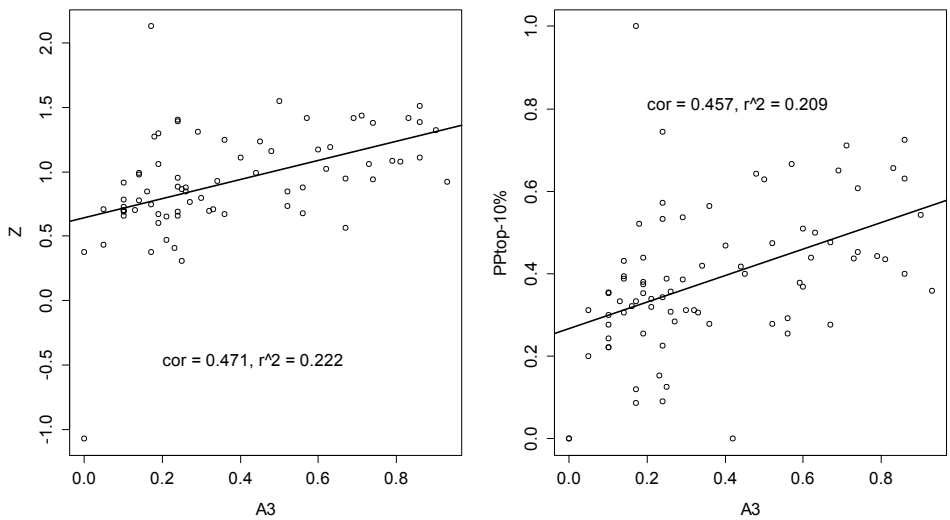


Figure 5 – Cumulative *z*-scores of clinical practice guidelines as a function of A3 (left), and $PP_{top\text{-}10\%}$ of clinical practice guidelines as a function of A3 (right). Solid lines are fitted linear regression lines.

As can be seen from Figure 5, there is a positive correlation for both citation indicators. Linear regression was performed on the data and regression lines were fitted to the plots (solid lines). The residuals from the regressions showed no systematic error, and Q-Q plots of the residuals showed approximately normally distributed residuals (Figure 6), with only few outliers. Thus there is no reason to assume more complex correlations, despite the large variation of data, which results in low r^2-values.

Despite the moderately low goodness of fit for the regression, the increase in the two indicators as a function of A3 is very high. For the *z*-score, the regression line increases from .6 to 1.4 standard deviations. This accounts for a very large increase, and can be interpreted as moving from "mainstream" to "top science". For the $PP_{top\text{-}10\%}$ indicator a similar pattern emerges, with an increase from approximately .27 to .59 (59% of all references belong to the top-10% of the highest cited papers in their respective journals).

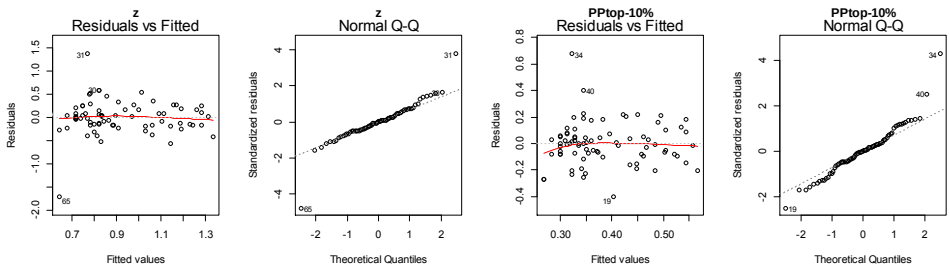


Figure 6 - Diagnostic plots for linear regressions. Residuals and Q-Q for *z* to the left and to the right for $PP_{top\text{-}10\%}$.

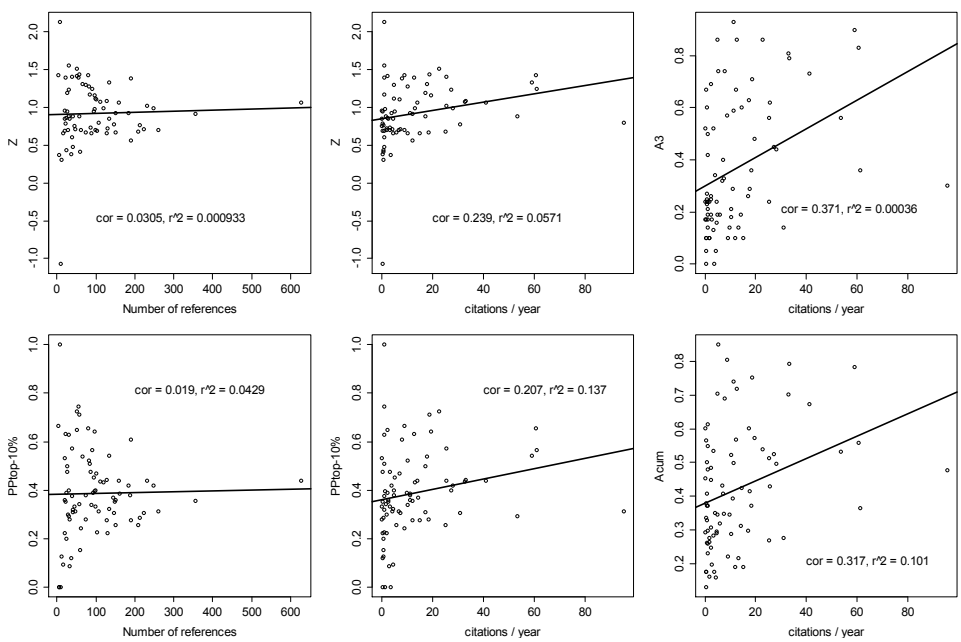


Figure 7 - Citation indicators as functions of number of references in clinical practice guidelines (left), citations per year given to clinical practice guidelines (middle) and correlation between A3 (top-right), A_{cum} (bottom-right) and citations per year.

Several other factors may be influencing the correlations, some of which are not directly related to bibliometric data and some are. There is a risk of confounding error from the number of references of the clinical practice guidelines as well as the citations given to the guidelines themselves. The two citation indicators are thus plotted as functions of number of references and citations per year respectively, displayed in Figure 7. The first column in Figure 7 shows the correlation between the citation indicators and number of references in the clinical

practice guidelines, the middle column the correlation between citation indicators and the annual citations received by the clinical practice guidelines. The latter is potentially related to the mechanism described by this study, therefore the AGREE scores A3 and A_{cum} are plotted in the final column, as a function of annual citations received. All plots have regression lines to show the general tendencies. The statistical strength of the $PP_{top\text{-}10\%}$ regressions is low (not statistically significant, $p>0.05$), while the other regressions are somewhat stronger ($p<0.05$), especially the A3 and A_{cum} regressions ($p<0.01$). This should also be apparent from a visual interpretation of the six plots in Figure 7, and the meaning of these will be discussed further in the following section.

Discussion

The purpose of this study is to investigate the potential association between AGREE assessments of clinical practice guidelines and the citation scores of references used by these guidelines. The hypothesis states a positive correlation, as guidelines scoring highly on the AGREE domain *rigor of development* are assumed to build on better evidence (e.g. in randomised controlled trials) which is here speculated to be associated with higher citation scores based on increased citations to study types associated with higher citation rates (Andersen & Schneider, 2011; Kjaergard & Gluud, 2002; Patsopoulos et al., 2005). Positive correlations are found, and the increased effect from the lowest A3 scores to the highest is considered very large. While this informs us about an association between citations to references and guideline development quality we cannot assume any mechanisms behind the association, as several models might explain the behaviour. From a bibliometric viewpoint, a plausible and interesting cause for the positive relationship is related to the interpretation of citations and the indicated relationship with clinical evidence. This study adds an argument to the hypothesis that clinical studies receive more citations if they provide more (relevant) evidence. An important caveat is the distinction between clinical research and other medical research, e.g. biomedical, as clinical research mostly is concerned with the testing and application of treatments. Therefore clinical practice guidelines will generally rely on clinical research, as it is usually closer to practice as more basic research areas. If biomedical studies are not included in clinical practice guidelines it is thus not a question of missing evidence, but rather of different evidence types, and conclusions about citability or citation scores should thus not be transferred from clinical research to biomedical.

A different, plausible mechanism, which is likely also affecting the results presented here, is the increased focus on the references used in clinical practice guidelines, as well as the reputation of the authors of these references. It is not unlikely that articles will receive more attention, once they have been cited by a clinical practice guideline, ultimately leading to an increase in citations, and it is not unlikely that guidelines will be more likely to cite the work of well-known authors in the field. These mechanisms could be investigated further, but would require more elaborate data than was acquired for this study.

As stated above, there is a danger of confounding errors from other sources when doing these types of analysis. Two potential confounding factors were mentioned previously, namely number of references in the guidelines and the citations given to the guidelines. The number of references might influence the results either way, as e.g. a short reference list might indicate a focus on only the very best evidence, but also a failure to include important evidence. The plots with number of references as the independent variable illustrate much larger variation for short reference lists than for the longer ones, but also that the reference list length is practically unrelated to the citation indicators, with almost horizontal regression lines roughly around the sample mean of the citation indicators. There is however a positive correlation between the annual citation counts of the clinical practice guidelines and the citation impact of their references. This effect can be explained by both of the causative models presented above. If the mention of citations in guidelines leads to higher citation scores for their references, it could be expected that the use of an article as a reference in a highly cited guideline would lead to a greater attention increase than the use in a less cited guideline. The evidence-mechanism, on the other hand, implies that the studies with the highest AGREE-scores would refer to the best evidence and by proxy contain good evidence themselves. Therefore one would also expect the citation scores of the guidelines with the highest AGREE-scores to be higher than those with low AGREE-scores. This is indeed the case, as can be seen in the rightmost plots in Figure 7, showing a clearly increasing function for both A3 and A_{cum} when plotted as a function of citations per year to the clinical practice guidelines. This provides further evidence that the proposed mechanism has merit, but both causative models may yet be active at the same time.

The A3 score used in the present study is far from being a perfect measurement of the degree of foundation on evidence. The measurement is subjective and designed for an entirely different purpose as applied here, which however has the advantage of making the above usage unobtrusive, thereby removing one type of potential bias. More critical weaknesses of the measure, in the current context, are the differences in application and the meaning of the individual items of the A3 domain score. Two factors contribute to this weakness: firstly, there are no clear criteria for the selection of AGREE-assessors, and the personal experience and motivations of (voluntary) assessors might vary from one study to another. Secondly, as was mentioned in the methods section, some guideline reviews implemented an interpretation of the AGREE instrument where all assessors needed to agree on each item score, where other reviews used a standardised approach allowing for inter-assessor inconsistency. One might argue for the benefits of each approach, and the arguments from a clinical standpoint are likely to be different than from a bibliometric one. While a uniform approach would have been preferable, the difference was not considered a major problem in this context, however, as the focus is on the overall increase in effect from the lowest scores to the highest. Given the broadly distributed scores in each AGREE

domain, the potential error from the different approaches is considered negligible, and both approaches are included.
Not all of the seven items in the A3 domain are directly related to the background literature, although most are. It is not possible to deduce how well each guideline scored in each item from the final score, and only very few reviews reported the full background data. Two guidelines with equal scores might thus represent different scores for the items most closely related to the mechanism described in this study. For the extreme cases, i.e. the lowest and highest scores, the problem is not as large as for the mid-range, where the potential for error is much larger. The problem is thus not as relevant when regarding the overall effect, and some of the variance could be explained by this issue.

The final weakness of the study to be discussed here is the use of journal-specific normalisation methods for citation indicators. While the normalisation as such is regarded as a strength, as it allows comparison between papers with different citation potentials, it is debatable whether other normalisation methods might be more appropriate. A common normalisation procedure is the field-normalisation, in which citations are normalised with respect to the entire field rather than merely the journal. It has been argued by e.g. Leydesdorff & Bornmann (2011) that field definitions are not clearly representative of the citation potentials of the individual journals in some field category definitions. It is thus not given that a field-normalisation would necessarily improve the design of this study.
The results presented here indicate an association between clinical evidence, citation rates and the quality of clinical guidelines dependant on the degree to which they are based on evidence. This association may be useful for research assessment, studies on clinical impact and provides insight into one of the many mechanisms behind citations. The use of a study as a reference in a well-developed clinical practice guideline may be seen as a mega-citation in some respects, as it very clearly indicates usefulness in a practice-setting, which is otherwise difficult to capture with traditional citation measures. It has been stressed that impact in a broad perspective has many other aspects than research (citation) impact (Kuruvilla, Mays, Pleasant, & Walt, 2006). While many of the impact types described by Kuruvilla et al. are not related to citation analysis at all, the observations in this study allow us to broaden the application of citation analysis to e.g. health policy and practice impact studies, while in no way claiming to cover all facets of these complex subjects.

References

Andersen, J. P., & Schneider, J. W. (2011). Influence of study design on the citation patterns of Danish, medical research. *Proceedings of the ISSI 2011 Conference* (pp. 46–53).

Beirlant, J., Glänzel, W., Carbonez, A., & Leemans, H. (2007). Scoring research output using statistical quantile plotting. *Journal of Informetrics*, *1*(3), 185–192. doi:10.1016/j.joi.2007.04.002

Burda, B. U., Norris, S. L., Holmer, H. K., Ogden, L. a, & Smith, M. E. B. (2011). Quality varies across clinical practice guidelines for mammography screening in women aged 40-49 years as assessed by AGREE and AMSTAR instruments. *Journal of clinical epidemiology*, *64*(9), 968–76. doi:10.1016/j.jclinepi.2010.12.005

Ferket, B. S., Colkesen, E. B., Visser, J. J., Spronk, S., Kraaijenhagen, R. A., Steyerberg, E. W., & Hunink, M. G. M. (2010). Systematic Review of Guidelines on Cardiovascular Risk Assessment. *Archives of internal medicine*, *170*(1), 27–40.

Ferket, B. S., Genders, T. S. S., Colkesen, E. B., Visser, J. J., Spronk, S., Steyerberg, E. W., & Hunink, M. G. M. (2011). Systematic review of guidelines on imaging of asymptomatic coronary artery disease. *Journal of the American College of Cardiology*, *57*(15), 1591–600. doi:10.1016/j.jacc.2010.10.055

Freel, A. C., Shiloach, M., Weigelt, J. a, Beilman, G. J., Mayberry, J. C., Nirula, R., Stafford, R. E., et al. (2008). American College of Surgeons Guidelines Program: a process for using existing guidelines to generate best practice recommendations for central venous access. *Journal of the American College of Surgeons*, *207*(5), 676–82. doi:10.1016/j.jamcollsurg.2008.06.340

Gallardo, C. R., Rigau, D., Irfan, A., Ferrer, A., Caylà, J. A., & Bonfi, X. (2010). Quality of tuberculosis guidelines : urgent need, *14*(October 2009), 1045–1051.

Glänzel, W., & Moed, H. F. (2012). Opinion paper: thoughts and facts on bibliometric indicators. *Scientometrics*.

Greenhalgh, T. (2010). *How to read a paper: The basics of evidence-based medicine* (4th ed.). Oxford: BMJ Books.

Kis, E., Szegesdi, I., Dobos, E., Nagy, E., Boda, K., Kemény, L., & Horvath, a R. (2010). Quality assessment of clinical practice guidelines for adaptation in burn injury. *Burns : journal of the International Society for Burn Injuries*, *36*(5), 606–15. doi:10.1016/j.burns.2009.08.017

Kjaergard, L. L., & Gluud, C. (2002). Citation bias of hepato-biliary randomized clinical trials. *Journal of clinical epidemiology*, *55*(4), 407–10.

Kuruvilla, S., Mays, N., Pleasant, A., & Walt, G. (2006). Describing the impact of health research: a Research Impact Framework. *BMC health services research*, *6*, 134. doi:10.1186/1472-6963-6-134

Lewison, G. (2002). From biomedical research to health improvement. *Scientometrics*, *54*(2), 179–192.

Lewison, G. (2003). Beyond outputs: new measures of biomedical research impact. *Aslib Proceedings*, *55*(1/2), 32–42. doi:10.1108/00012530310462698

Leydesdorff, L., & Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology*, *62*(2), 217–229. doi:10.1002/asi.21450

Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: fractional counting of citations. *Journal of Informetrics*, *4*(4), 644–646. Digital Libraries; Physics and Society. doi:10.1016/j.joi.2010.05.003

Lundberg, J. (2007). Lifting the crown - citation z-score. *Journal of Informetrics*, *1*(2), 145–154.

Patsopoulos, N. a, Analatos, A. a, & Ioannidis, J. P. A. (2005). Relative citation impact of various study designs in the health sciences. *JAMA : the journal of the American Medical Association*, *293*(19), 2362–6. doi:10.1001/jama.293.19.2362

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, *9*(5-6), 281–291. doi:10.1007/BF02017249

Schubert, A., & Braun, T. (1996). Cross-field normalizations of scientometric indicators. *Scientometrics*, *36*(3), 311–324.

The AGREE Collaboration. (2003). Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. *Quality & safety in health care*, *12*(1), 18–23.

Van Leeuwen, T. N., Costas, R., Calero-Medina, C., & Visser, M. S. (2012). The role of editorial material in bibliometric research performance assessments. In É. Archambault, Y. Gingras, & V. Larivière (Eds.), *Proceedings of STI 2012 Montréal* (pp. 511–522). Montréal, Canada: 17th International Conference on Science and Technology Indicators.

Van Raan, A. F. J., Van Leeuwen, T. N., Visser, M. S., Van Eck, N. J., & Waltman, L. (2010). Rivals for the crown: Reply to Opthof and Leydesdorff, 1–17. Digital Libraries; Physics and Society.

Vinkler, P. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics*, *10*(3-4), 157–177.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., Van Eck, N. J., Van Leeuwen, T. N., et al. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419–2432. doi:10.1002/asi.22708

Zitt, M. (2011). Behind citing-side normalization of citations: some properties of the journal impact factor. *Scientometrics*, (1986). doi:10.1007/s11192-011-0441-7

AUTHOR NAME CO-MENTION ANALYSIS: TESTING A POOR MAN'S AUTHOR CO-CITATION ANALYSIS METHOD (RIP)

Andreas Strotmann[1] and Arnim Bleier[2]

[1] *andreas.strotmann@gesis.org*
[2] *arnim.bleier@gesis.org*
GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, D-50667 Cologne (Germany)

Abstract
As a social science information service for the German language countries, we document research projects, publications, and data in relevant fields. At the same time, we aim to provide well-founded bibliometric studies of these fields. Performing a citation analysis on an area of the German social sciences is, however, a serious challenge given the low and likely significantly biased coverage of these fields in the standard citation databases. Citations, and especially author citations, play a highly significant role in that literature, however.
In this work in progress, we report preliminary methods and results for an author name co-mention analysis of a large fragment of a particularly interesting corpus of German sociology: a quarter century's worth of the full-text proceedings of the Deutsche Gesellschaft für Soziologie (DGS), which celebrated its 100th anniversary meeting in 2012. Results are encouraging for this poor cousin of author co-citation analysis, but considerable refinements, especially of the underlying computational infrastructure for full-text analysis, appear advisable for full-scale deployment of this method.

Conference Topic
Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction
In a way, this paper goes back to one of the very roots of author co-citation analysis, as White (1990) identified Rosengren's (1968) use of "co-mentions" of writers of literary works to visualize an intellectual structure in their reception.
Here, we take Rosengren's (1968) original writer co-mention visualization idea and extend it with the author citation pattern analysis and visualization methodology tradition of White and Griffith (1981), who extended the document co-citation idea generally co-attributed to Marshakova (1973) and Small (1973) to the study of author co-citation patterns, and of Zhao and Strotmann (2008a), who subsequently extended Kessler's (1963) bibliographic coupling idea from the study of document similarities to the study of author bibliographic coupling patterns with similar techniques.

From a technical perspective, the procedures used in this study are rooted in the idea of word co-occurrence analysis (a literature we will not review here), but with a restriction to words which likely denote frequently mentioned author names (or more precisely, their surnames). The analysis and visualization of author surname co-occurrence matrices (i.e., author name co-mention matrices) that we perform here uses methodology taken straight from Strotmann and Zhao (2012) – we refer the reader to that paper as a starting point for tracing its origins.

Author Co-citation Analysis vs. Author Co-mention Analysis

A formal author co-citation analysis of a document set usually proceeds as follows:

1. the reference lists of all documents are collected, usually from Web of Science;
2. for each cited reference found, names of authors of the cited work are identified (usually only the first author);
3. the authors cited most highly in the document set are determined;
4. an author × author co-citation matrix is constructed for the most highly cited authors in the document set – each cell counts the number of papers that are registered as co-citing the corresponding pair of authors;
5. the resulting co-citation matrix is analyzed statistically, and the result visualized and interpreted.

In the quarter-century of DGS proceedings that we selected for this experiment as a potential representation of "German sociology", unfortunately, formal *references* are frequently hard to identify, if they are listed at all. *Names* of sociologists (or other influential thinkers) are mentioned abundantly in these volumes, however: the surname of German sociologist Max Weber, to give a decidedly non-random example, appears at least once in every proceedings text on average. We therefore used *author co-mention analysis* as a poor-man's alternative to author co-citation analysis, as follows:

1. the full text of all documents is collected;
2. for each document full text, a list of candidate author surnames it mentions is compiled;
3. from the collection of all candidate author surnames mentioned, the most frequently mentioned likely surnames are extracted manually;
4. for each document, a weighted co-mention count is determined for each pair of frequently mentioned author surnames that it contains;
5. the per-document author surname co-mention matrices are accumulated into a corpus-level surname × surname co-mention matrix;
6. the co-mention matrix is analysed statistically, and the results are visualized and interpreted.

The DGS 1960-85 corpus

The corpus we use in this experiment is a large fragment (about 25%) of a particularly interesting series of publications in German sociology: a quarter

century's worth of the full-text proceedings of the roughly biannual meetings of the *Deutsche Gesellschaft für Soziologie* (DGS), which celebrated its 100th anniversary meeting in 2012. Our experiment is performed in part to determine if it is possible to perform meaningful citation analysis on this corpus, and to determine areas in which new methods may need to be developed.
The DGS proceedings used for the experiment had previously been scanned, OCRed and catalogued by GESIS Leibniz Institute for the Social Sciences; where possible, this corpus has been made available to the general public on the institution's Social Sciences Open Access Repository (SSOAR.info). The particular fragment of this corpus that we used in this experiment is comprised of 1,212 publications which appeared during the years 1960 to 1985. On average, each document contains 2,471 words, mostly in German.

Identifying frequently mentioned author surnames

In many languages, names are the only words that are capitalized inside a sentence, so that the problem of extracting names from a text is largely reduced to filtering out corporate names. In German orthography, however, which almost all the texts in our corpus adhere to, all nouns are capitalized, not just names, and surnames especially are taken from a wide range of concept nouns (e.g., Vogel = bird; Weber = weaver), place names (e.g., Mannheim), or Christian names (e.g., Walter), each from a range of languages, regions, and cultures. This is further complicated by two forms of attribution suffixes in German, namely, the genitive case marker ("Freuds" for English "Freud's") and its "sch" suffix form ("Freudsche/n/m/r" for English "Freudian"). In our experiment we decided for simplicity's sake to use words as text units rather than compound phrases, which exacerbates the name ambiguity problem.
Identification of author name mentions is thus a major problem in our case, which we addressed only approximately for the sake of this experiment. The approach we opted for was to create a list of words that count as mentions of authors, and to treat everything that does not appear in this list as non-name words. As central criteria for the construction of this list, we would like central and frequently mentioned authors to be included, but at least in the current experiment would prefer to remove from the list any terms that frequently occur as concept words rather than names. We would also like to remove terms that are too likely to name many different individual authors.
We chose to approximate these criteria by creating the list of likely author surnames in the following steps: First, we compiled a list of candidate author surnames from the author metadata of a document collection large enough that it can be assumed to have a similar distribution of surname frequencies as the target corpus. Next, we pruned from the resulting list those names that do not begin with a capital letter, that are very short, or that occur only rarely. We lemmatized this list of names (removing genitive markers and the like as discussed above). Finally, candidate names too likely to refer to multiple authors or to non-author entities were removed.

For the purpose of this experiment the SOWIPORT document collection (see, e.g., Stempfhuber, 2008) presented itself as an attractive source from which to compile the list of candidate author surnames. SOWIPORT covers the German social sciences, something of a superset of our target distribution of author name mentions in the DGS corpus, and it makes authorship metadata readily available. For the subsequent pruning step, we found experimentally that limiting the list of candidate author names to words of at least three characters which occur at least 25 times in the DGS corpus yielded useful results in our setting. We applied lemmatization following the flexion rules of German proper nouns. This resulted in a list of about 1500 different words that frequently appear as author surnames in SOWIPORT and frequently appear in our target corpus. From this list we first hand-picked about 500 to proceed with, lemmatized as above. Finally, we manually removed about another half as being immediately recognizable as concept words, leaving about 220 names in total, each mentioned at least 11 times in the corpus.

Constructing the author name co-mention matrix

Next, for each member in our list the full texts were scanned for lemmatized word matches, resulting in 220 multisets of document author mentions, i.e., one multiset of documents per lemmatized author surname, with each document occurring in the multiset as many times as the corresponding name appears in the document.

Clearly, this list of words, which hopefully denote unique authors cited frequently in German sociology during the years 1960-85, does not qualify as a complete list of authors, nor is it a random sample, since highly cited names are selected preferentially. However, experience teaches us that author co-citation data visualized through factor analysis is quite a stable technique, which means that there is a good chance that intelligible results could be obtained from applying a similar method to our target corpus even with a sub-optimal list of names.

Unlike traditional co-citation analysis based on data from citation databases, where each cited publication appears once in the list of references even if the text refers to it dozens of times, co-mentions between authors allow for a weighting by how many times each name appears in the text. Intuitively, a document that refers frequently to two authors indicates a stronger connection between them than a document that either mentions both authors just once or that mentions one author dozens of times while the other author is mentioned just once.

To calculate the co-mention count of two authors in this set of documents, we therefore first calculate for each author his or her author mention profile as the multiset (rather than the set) of documents that mention the author, weighted by the number of times that author is mentioned in the document. Given these author mention profiles for all the frequently mentioned author surnames identified in the previous step, the co-mention count of two authors is calculated as the multiset size of the multiset intersection of the two authors' author mention profiles.

The typical author co-citation count for a pair of authors would be determined as the size of the intersection of the sets (not multisets) of documents that cite each author. The inspiration for our choice of co-mention counting method is Zhao and Strotmann (2008a), who calculate the author bibliographic coupling of two authors as the multiset size of the multiset intersection of the citing behaviour profiles of the two authors, where an author's citing behaviour profile is the multiset union of all reference lists of the author's oeuvre.
Unlike in the case of all-author co-citation matrices (Zhao, 2006; Zhao & Strotmann, 2008b), it is not possible to distinguish between inclusive and exclusive co-mention counts, i.e., to filter out co-mentions based purely on co-authorship (e.g., Marx & Engels). Indeed, co-mention counts in this experiment even include the co-occurrence between two authors of the mentioning work, and co-occurrences of the names of the mentioning work's authors and the mentioned work's authors. Author co-mention matrices are thus considerably more noisy data sources than author co-citation matrices.

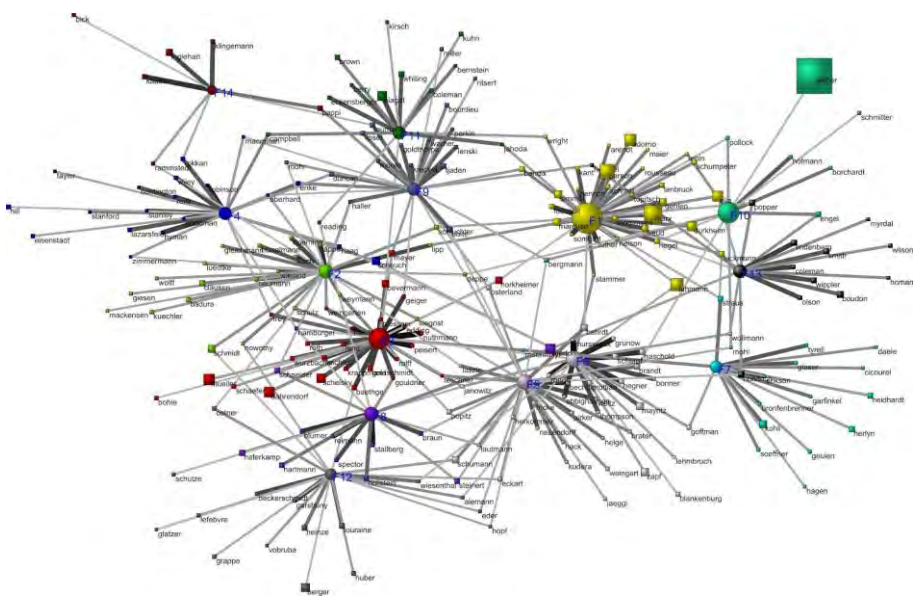


Figure 2: Author name co-mention analysis result visualization, 14 factor solution.

Factor analysis and visualization

Finally, we performed an exploratory factor analysis on the resulting author co-mention matrix, using the Factor Analysis routine of SPSS 19. As parameters, we specified: replacement of missing values (i.e., diagonal values) by the mean; oblique rotation using Oblimin with default parameters; and a maximum of 250 iterations for each of the steps involved in the factor analysis.
Based on visual inspection of the Scree plot, we chose a 14-factor solution rather than the 38-factor one that Kaiser's Rule of eigenvalue > 1 would have suggested.

We visualize the 14 factors × 220 authors pattern matrix result of the factor analysis in Figure 2. In this visualization using the Kamada-Kawai layouting routine of Pajek 3, lemmatized author names and factors are represented as nodes (square or round, resp.), and an author's loading on a factor is represented by a line with gray scale value and width proportional to the absolute loading. Nodes are color coded according to the 14 factors and their memberships (authors who do not load sufficiently on any factor are not displayed). Author node sizes are proportional to the number of mentions, and factor node sizes proportional to the sum of members' mentions weighted by the member's factor loading.

Interpretation

The factors were interpreted and labelled by two social scientists, both colleagues of the authors. Table 1 lists the results of this interpretation, along with some statistical characteristics of the factors: the size of a factor (defined as the number of authors whose maximal loading (of at least 0.3) in the pattern matrix is with this factor) and the highest loading of an author on this factor (which is an indicator of the clarity or distinctness of this factor).

Factor labels ending in a question mark denote factors for which an intellectual interpretation was not easily apparent, and for which an hour's discussion between the social scientists did not lead to a clear agreement. A label for these factors was attempted by the authors using SOWIPORT and Google Scholar in these cases. Generally, though not always, these more "questionable" factors tend to exhibit lower maximal author loadings, as Table 1 shows, and the most questionable one, F10, identified as an artefact of the lack of author name disambiguation in the underlying dataset, has the lowest such characteristic.

Table 1. Factors and their interpretations and characteristics.

Factor	*Label*	*Max Loading*	*# Members*
F1	Theory of Society	.87	30
F2	Biographies?	.79	22
F3	Government Theory	.97	26
F4	Political Science	.85	19
F5	Sociology of Work?	.75	20
F6	Sociology of Organisations	.92	17
F7	Sociology of the Family?	.77	15
F8	Social Problems?	.90	13
F9	Social Inequality?	.70	16
F10	?common names?	.63	8
F11	Psychology?	.72	12
F12	Socioeconomics?	.76	15
F13	Rational Choice Theory	.76	12
F14	Values and elections?	.80	6

Discussion and Outlook

In this paper we present a first experiment using author name co-mention analysis based on the full text of a scientific literature for which no formal citation index and therefore no possibility of a traditional author co-citation analysis is available. Given the quality of the underlying dataset we used here, the results are encouraging. Statistical characteristics of the factor analysis results are reasonable when compared to those from author co-citation analyses of other fields we have performed previously. Interpretation of about half the resulting factors was considered straightforward by the social scientists who looked at the results; again, this suggests reasonable performance in our experience.

The appearance of a factor (F10) that consists mostly of surnames that likely correspond to several distinct authors each reminds us, however, that the author name co-mention methodology we tried here has its limits. Author name disambiguation is a requirement that we attempted to avoid by filtering out words that could, in principle, be either surnames, first names, or dictionary words, or any combination of these. As Strotmann and Zhao (2012) point out in a similar case, this is not always a reasonable approach to take to author name disambiguation, and for author name co-mention analysis to work well, significant effort will need to be invested in improving the identification of individuals from author name mentions.

The fact, on the other hand, that the name "Weber" - almost certainly denoting, in a vast majority of cases, the prominent founding father of sociology in Germany, Max Weber - gets categorized with these multi-individual names is perhaps symptomatic of the extreme degree to which the founders of German sociology are cited across its entire literature. The size of the Weber author node in our visualization - at least an order of magnitude larger than even the factor nodes - illustrates this, too: the name *Weber* is practically synonymous with the term "sociology", being mentioned more than 2000 times in 1200 texts.

We suspect that one reason why this experiment worked quite well despite these short-comings is the fact that author name co-mention counting allowed us to weight higher those who are (co-)mentioned frequently in a text as opposed to those who are mentioned only in passing. This would be expected to significantly increase the relevance of high co-mention counts, and thus improve the signal-noise ratio.

Especially for the purposes of bibliometric analysis of social science literatures, where coverage of standard citation indexes are considered inadequate, this alternative approach may serve as a poor-man's analysis tool as long as the data situation does not improve significantly.

Acknowledgments

The authors wish to thank Howard D. White and Dangzhi Zhao for excellent advice on this paper, and their GESIS colleagues Maria Zens and Matthias Stahl for help interpreting the factor analysis results.

References

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*, 10-25.

Marshakova, I.V.(1973). System of document connections based on references (in Russian). *Nauchno-Tekhnicheskaya Informatsiya, 2*(6), 3-8.

Rosengren, K.E. (1968). *Sociological aspects of the literary system*. Stockholm: Natur och Kultur.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*, 265.

Stempfhuber, M., Schaer, P., & Shen, W. (2008). Enhancing visibility: Integrating grey literature in the SOWIPORT Information Cycle. *The Grey Journal, 4*(3), 121.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science, 24*, 265.

White, H.D. (1990). Author co-citation analysis: overview and defense. In *Bibliometrics and Scholarly Communication*, Christine Borgman, ed. Newbury Park, CA: Sage. 84-106.

White, H.D., & Griffith, B.C. (1981) Author co-citation: a literature measure of intellectual structure, *Journal of the American Society for Information Science*, *32*, 163-172.

Zhao, D. (2006). Towards all-author co-citation analysis. *Information Processing & Management, 42*, 1578.

Zhao, D., & Strotmann, A. (2008a). Evolution of research activities and intellectual influences in Information Science 1996-2005: Introducing author bibliographic coupling analysis. *Journal of the American Society for Information Science and Technology, 59*(13), 2070.

Zhao, D., & Strotmann, A. (2008b). Comparing all-author and first-author co-citation analyses of Information Science. *Journal of Informetrics, 2*(3), 229.

BIBLIOGRAPHIC COUPLING AND HIERARCHICAL CLUSTERING FOR THE VALIDATION AND IMPROVEMENT OF SUBJECT-CLASSIFICATION SCHEMES

Bart Thijs[1], Lin Zhang[2] and Wolfgang Glänzel[3]

[1] *bart.thijs@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)

[2] *zhanglin_1117@126.com*
Department of Management & Economics, North China
University of Water Conservancy and Electric Power, Zhengzhou, (China)

[3] *wolfgang.glanzel@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
An attempt is made to apply bibliographic coupling to journal clustering of the complete Web of Science database. Since the sparseness of the underlying similarity matrix proved inappropriate for this exercise, second-order similarities have been used. Only 0.12% out of 8282 journals had to be removed from the classification as being singletons. The quality at three hierarchical levels with 6, 14 and 24 clusters substantiated the applicability of this method. Cluster labelling was made on the basis of the about 70 subfields of the Leuven-Budapest subject-classification scheme that also allowed the comparison with the existing two-level journal classification system developed in Leuven. The further comparison with the 22 field classification system of the Essential Science Indicators does, however, reveal larger deviations.

Conference Topic
Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction
The issue of subject classification and the creation of coherent journal sets has been a major topic in our field since the seventies (see e.g., Narin et al., 1972; Narin, 1976). The development of computerised methods and the availability of large datasets have shifted the attention from mapping small or single disciplines to the generation of global science maps (Garfield, 1998). Data available from Thomson Reuters' Journal Citation Reports (JCR) has been used by several authors (Bassecoulard and Zitt, 1999; Leydesdorff, 2004). Unlike in Thomson Reuters' Web of Science (WoS) database, where citations are determined for each

paper individually, in the JCR citation data are based on journal information in the papers' reference lists and therefore aggregated to the journal level. However, also WoS data was used at the level of individual publications for the generation of global maps. Jarneving (2005) applied bibliographic coupling to map and to analyse the structure of an annual volume of the Science Citation Index. Janssens et al. (2008; 2009) used a combination of cross-citations and a lexical approach to map journals. Zhang et al. (2010) validated this approach. This paper builds on prior attempts to classify journals relying on computerised techniques. In this study we take a different approach and attempt to build a network among journals based on bibliographic coupling similarities.

The advantage of bibliographic coupling is that there is no delay for the calculation of the link between publications or journals as all data needed are present upon publication or indexing in the database. This also means that link between documents, once established will remain constant over time. Sharing this property with text-based method, new mappings of journals based on bibliographic coupling are able to reflect the current situation as soon as the underlying documents are indexed in the database. However, for this paper and the development and validation of our methodology we use the 2006-2009 publications set to be able to relate our results to those of previous exercises.

In contrast to the above-mentioned advantages of bibliographic coupling, this method has one drawback which is shared with other citation-based approaches such as co-citation analyses.This disadvantage is a result of the very sparse nature of the link matrix (Janssens, 2007; Janssens et al., 2008). The overwhelming number of document pairs does not share any reference at all and thus a large number of zeros occur in the similarity matrix. This deteriorates the quality of the subsequent clustering and may result in an unrealistic large number of singletons (cf. Jarneving, 2005). As cross-citation data suffers from the same problem, Janssens et al. (2008) introduced a hybrid approach, where they combined citation-based with lexical similarities.

Another solution to overcome the sparseness problem is the use of second order similarities (Janssens, 2007; Ahlgren & Colliander, 2009; Thijs et al., 2013). The objective of the present paper is to demonstrate the applicability of bibliographic coupling as link measure in the mapping of journals as well as to compare the results with those of previous cross-citation and hybrid citation-text based studies.

Data sources

A set of journals was compiled from the Web of Science database (SCI-Expanded, SSCI and AHCI). All journals covered in this database between 2006 and 2009 with at least 100 publications in this period are taken into account. This resulted in a set of 8282 journals. For the calculation of the bibliographic coupling between journals we took the following approach. In total more than 134 million references in 4,753,892 publications could be processed on the basis of uniquely coded reference items. All data was uploaded into an *Oracle* database and regular

SQL was used to query for joint references between journals. Analyses are run in *Matlab* and visualizations are made with *Gephi* (Bastian et al., 2009).

Methods

This section describes the choices that have been made for our journal mapping. In order to enhance comparability with the earlier studies (Janssens et al., 2009; Zhang et al., 2010) we adopted the same clustering technique, namely Ward's hierarchical clustering. A short description of this method will follow later. The goal of this paper is, however, to make it possible to create a mapping based on bibliographic coupling and covering all selected journals.

Analogously to document mapping based on bibliographic coupling, all items that appeared in the reference lists of papers published in the journal are taken into account., As references appear only once in the reference list of a paper, a binary approach was chosen assigning the values 0 or 1 according as the reference was shared or not by the two papers. We followed the same principle for journals since weighting according to multiple occurrences of shared references at the journal level resulted in just marginal deviations from the binary approach. Figure 1 presents an example of reference links between two journals. Journal *A* has published 3 articles with six references in total but two papers refer to the same article (*R*4). Journal *A* has thus 5 distinct references. Journal *B* has 4 papers with six references in total, each pointing at a distinct publication. Journal *A* and *B* share 3 distinct references.

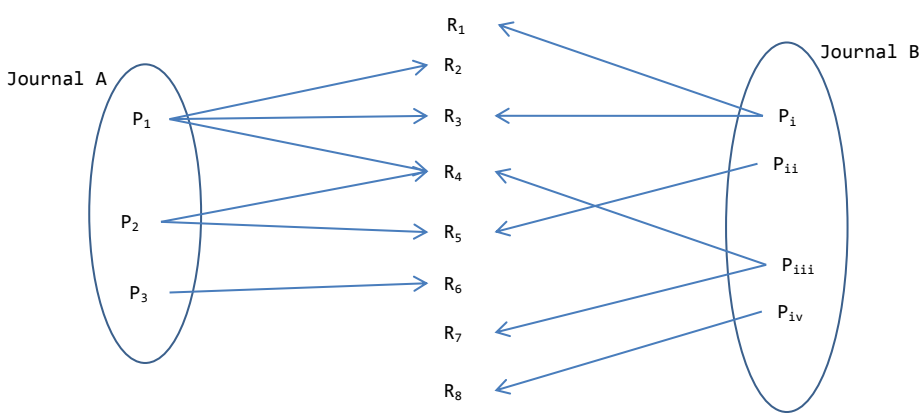


Figure 1. Graphic representation of bibliographic coupling between journals

To express the strength of a link between two journals we calculated a first order similarity based on Salton's cosine measure. The mathematical derivation and interpretation of this similarity measure in the framework of a Boolean vector space model can be found in (Sen & Gan, 1983; Glänzel & Czerwon, 1996). As bibliographic coupling tends to produce very sparse similarity matrices we applied a second order similarity to reduce this effect. While the first-order

similarity is based on the angle between two reference vectors, the second-order similarity is calculated as the cosine of the angle of two vectors holding the first order similarity between two journals. After the calculation of the second-order similarities, ten journals were removed from the set as they appeared to be singletons without any link to the other journals in the set. The network thus included 8272 journals in total.

Hierarchical clustering with Ward's agglomeration method was used to create a hard clustering of all the journals. Given the rather limited set of entities to be clustered, Ward's method already proved its validity in many studies. This method does not provide any automated optimum number of clusters so that the decision was made on the basis of the dendrogram and the silhouette statistics (Rousseeuw, 1987). As Ward assumes distance measures instead of similarities we converted the similarities to distances before clustering.

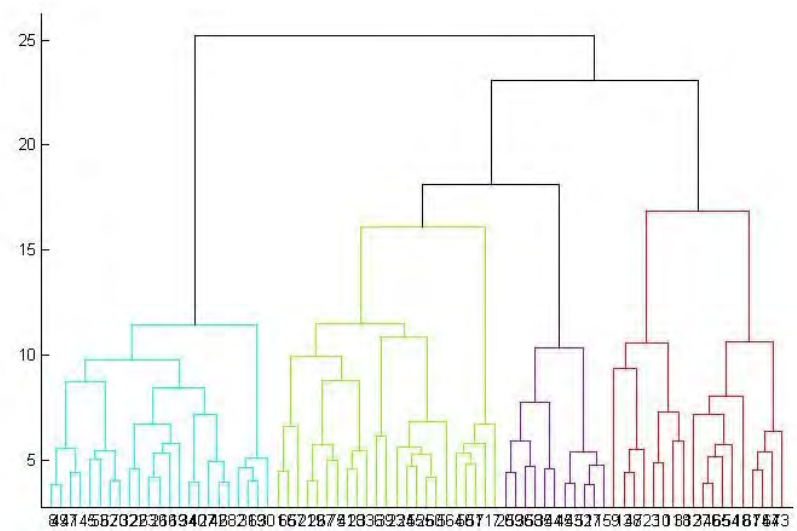


Figure 2. Dendrogram for hierarchical clustering of the 8272 journals based on Ward's method [Data sourced from Thomson Reuters Web of Knowledge]

Results

In this section we present the results of the clustering and discuss the validity of the partitioning of journal set. As pointed out in the previous section, a dendrogram and a silhouette-value plot were used to select an appropriate number of clusters. The two diagrams are presented in Figure 2 and Figure 3. Three different levels were chosen. The dendrogram holds strong arguments for a six cluster partitioning while the silhouette plot shows a first peak at 7 clusters. For the highest hierarchical level in the following analysis we use the six cluster solution. At a lower level, the silhouette plot suggests the solutions with 14 and 24 clusters, respectively. Both will be described in subsequent subsections.

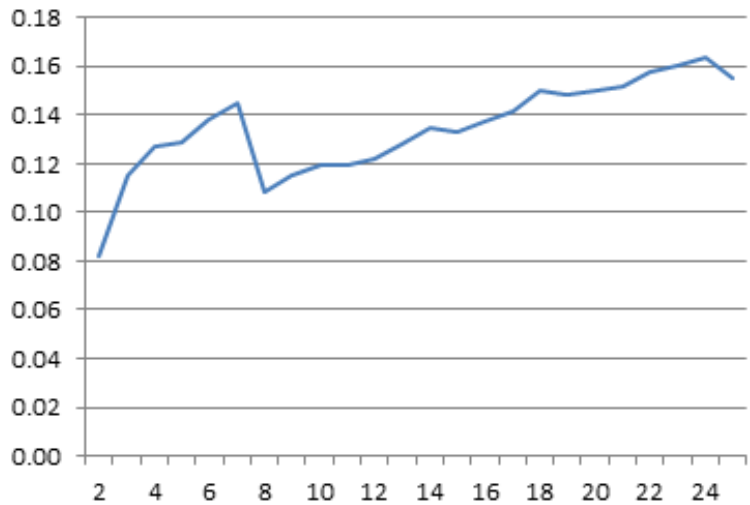


Figure 3. Mean Silhouette values for solutions of 2 up to 25 clusters, with local maxima at 7, 14 and 24 clusters [Data sourced from Thomson Reuters Web of Knowledge]

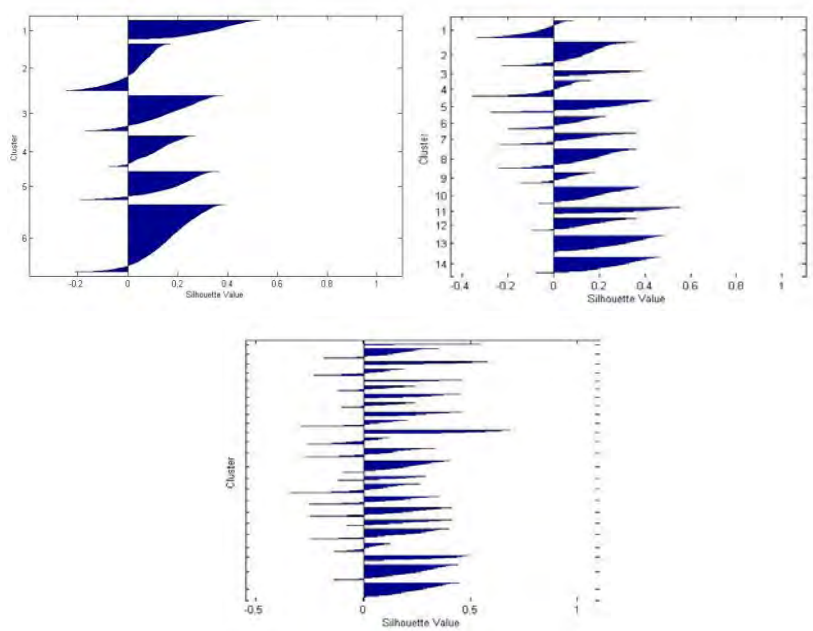


Figure 4 Silhouette values of three distinct clustering solutions with 6, 14 and 24 groups (from left to right) [Data sourced from Thomson Reuters Web of Knowledge]

For the evaluation of the specific cluster solution we can rely on the silhouette graphs presented in Figure 4. Each graph presents the silhouette values of the journals in the respective cluster. For each journal a silhouette value is calculated. These values range between 1 and -1 where positive values indicate an appropriate clustering of the journals. Journals are grouped by cluster and ordered from highest silhouette value to lowest. As a consequence the graph gives a good profile of the quality of each cluster. A larger area at the positive side of the vertical axis thus represents a better partitioning. The most favourable situation is found in the six-cluster solution. Here most journals are assigned to the appropriate cluster and only the second cluster has a larger share of negative values (cf. left-most diagram in Figure 4).

Cluster Description

Unlike in lexical or hybrid citation-textual methods, where clusters can be labelled and described using the textual component, e.g., the best terms or keywords, pure citation-based approaches are put at a severe disadvantage if the content of the clusters have to be described. In order to find an acceptable solution, we decided to use the journal-based subject-classification scheme developed in Leuven (Glänzel & Schubert, 2003). This solution proved most advantageous since both clustering and classification scheme are based on journal assignment. Table 1 presents the hierarchical structure of the three level partitioning. For each cluster the number of journals is mentioned. The labels for the higher levels can be deduced from the lowest level. These labels are taken from the Leuven classification system . The label from the most prominent subject category has been assigned to the corresponding cluster.

Another way to describe the cluster is by using core journals. This notion can be analogously defined as core documents introduced by Glänzel & Czerwon (1996) and extended by Glänzel & Thijs (2011). In this particular application, a core journal can be identified as journal with at least n links with other journals of at least a given strength r on the second order similarity measure. For the identification of core journals in each cluster we set the number of strong links to at least half the set of journals in the cluster. As we are using second order similarities this choice is not unreasonable. The value of the strength is chosen such that 12 journals within each cluster comply with both criteria. This means that for more dense clusters the choice of appropriate r-value is higher than in clusters where the journals are not as strongly linked. Cluster 21 labelled as 'Arts & Humanities' is such a cluster where a lower value of r was required to retain twelve journals. This is a result of the specific citation behaviour in the humanities, where citations play a somewhat different role than in the sciences (cf. Glänzel & Thijs, 2011). A list of selected core journals for each cluster is given in Table 2.

Concerning the results, two striking observations could be made. Above all, chemistry is at each level a separate cluster. One might expect that at the highest level, chemistry is merged with Physics but we found different patterns. The

second noteworthy observation concerns cluster 17 (Public Health & Nursing). This is a cluster within the 'Psychology – Neuroscience' cluster at the highest, six-cluster level. In other partitions or subject classification systems this is attributed to Non-Internal Medicine.

Table 1. Hierarchical structure of the three level partitioning with labels *l(i)* and number of journals *n(i)* according to the level with 6, 14 and 24 clusters [Data sourced from Thomson Reuters Web of Knowledge]

l(6)	*n(6)*	*l(14)*	*n(14)*	*l(24)*	*n(24)*	*Leuven subfield*
I	*n=691*	**n**	*n=691*	**24**	*n=691*	Chemistry; Material Science
II	*n=1704*	**c**	*n=268*	**19**	*n=268*	Geosciences; Geography
		d	*n=632*	**15**	*n=226*	Physics; Astronomy & Astrophysics;
				16	*n=406*	Engineering; Classical Physics
		k	*n=272*	**22**	*n=272*	Pure Mathematics
		l	*n=532*	**1**	*n=80*	Statistics & Probability
				2	*n=452*	Computer Science; Applied Mathematics
III	*n=1285*	**g**	*n=487*	**7**	*n=207*	Neuroscience; Neurology
				8	*n=280*	Psychology; Psychiatry
		h	*n=798*	**17**	*n=381*	Public Health; Nursing
				18	*n=417*	Social Psychology; Therapy; Counseling
IV	*n=1128*	**i**	*n=428*	**21**	*n=428*	Arts & Humanities
		j	*n=700*	**3**	*n=170*	Management; Marketing; Innovation
				4	*n=337*	Sociology; Social & Political Sciences; Law
				11	*n=193*	Economics; Accounting;
V	*n=1032*	**e**	*n=492*	**20**	*n=492*	Biology
		f	*n=540*	**9**	*n=225*	Agriculture; Plant Science
				10	*n=315*	Microbiology; Biotechnology; Food Science
VI	*n=2432*	**a**	*n=712*	**5**	*n=137*	Veterinary Sciences; Animal Sciences
				6	*n=251*	Immunology; Respiratory Medicine
				12	*n=324*	Non-Internal Medicine;
		b	*n=1007*	**13**	*n=432*	Haematology; Oncology; Surgery; Radiology
				14	*n=575*	Internal Medicine; Cardiovascular Medicine
		m	*n=713*	**23**	*n=713*	Biosciences; Biomedical Research

Cluster Structure

To visualise relations between the 24 clusters we created an additional map. Figure 5 shows these relations. The link between the clusters is based on bibliographic coupling. Also for this map we used a binary approach just as we did for the journals. The map was drawn in *Gephi* using the 'Force Atlas 2' layout

method. The thickness of the link represents the similarity. The colours represent the six cluster solution. Here we see the central position of the chemistry cluster between physics, biology and life sciences (especially biosciences and biomedical research). Given the strong links with the three groups the separation of chemistry from physics seems justified.
Cluster 17 (Public Health – Nursing) is linked to several (psychology – neuroscience clusters) medical clusters. This position of the topic is interesting and deserves more attention.

Table 2. Three core journals per cluster (selection does not imply any ranking) [Data sourced from Thomson Reuters Web of Knowledge]

#	*Journal title*	#	*Journal title*
1	biometrika canadian journal of statistics-revue canadienne de statistique computational statistics	13	annals of surgical oncology diseases of the esophagus world journal of gastroenterology
2	elektronika ir elektrotechnika ieee transactions on industrial informatics ieee transactions on systems man and cybernetics part a-systems and humans	14	american journal of the medical sciences annals of medicine clinical and investigative medicine
3	california management review ieee transactions on engineering management journal of business research	15	canadian journal of physics central european journal of physics chinese physics letters
4	china quarterly environment and planning c-government and policy environmental politics	16	acta mechanica sinica advances in engineering software comptes rendus mecanique
5	archivos de medicina veterinaria arquivo brasileiro de medicina veterinaria e zootecnia polish journal of veterinary sciences	17	applied nursing research bmc health services research contemporary clinical trials
6	clinical and vaccine immunology fems immunology and medical microbiology international journal of immunopathology and pharmacology	18	american psychologist canadian journal of behavioural science-revue canadienne des sciences du comportement canadian psychology-psychologie canadienne
7	annals of neurology brain research brain research bulletin	19	canadian journal of earth sciences comptes rendus geoscience earth-science reviews

8	biological psychology developmental neuropsychology international journal of psychophysiology	20	african zoology biological invasions israel journal of zoology
9	annals of applied biology botanical studies journal of horticultural science & biotechnology	21	american historical review new literary history critical inquiry
10	applied biochemistry and biotechnology biotechnology and bioprocess engineering engineering in life sciences	22	archiv der mathematik bulletin des sciences mathematiques chinese annals of mathematics series b
11	canadian journal of economics-revue canadienne d economique economic inquiry australian economic review	23	acta biochimica et biophysica sinica advances in experimental medicine and biology biochemical and biophysical research communications
12	journal of burn care journal of dental research physikalische medizin rehabilitationsmedizin kurortmedizin	24	acta chimica sinica acta physico-chimica sinica chemical journal of chinese universities-chinese

Comparison with the Leuven classification system

The partitioning in 14 clusters is suitable for comparison with the 15 main fields in the Leuven classification system. In this latter system a sixteenth field exists, namely the multidisciplinary sciences but this has been omitted from this analysis for obvious reasons. An important difference between the two systems is that the Leuven classification allows multiple assignments of journals to fields. With the applied Ward methodology this is not possible for the clustering developed in this paper. Despite these multiple assignments we used the Jaccard Index to measure the concordance between the two journal classifications. The results are presented in Table 3. For most fields a good mapping with one of the fourteen clusters can be found. Fields 'Biosciences' and 'Biomedical Research' are jointly mapped on cluster 'm' which explains the reduction by one field. But journals assigned to the field 'Non Internal Medicine Specialties' are spread across four clusters ('a', 'b', 'g', 'h') according to the 14-cluster solution (see column *l*(14) in Table 1). 'Neurosciences & Behaviour' is split into two clusters ('g' and 'h'), both these have also a link to 'Non internal medicine'. Cluster 'h' also has a link to social sciences. In this last cluster we see the common focus in medicine, psychology and social and community issues. Most of the journals assigned to the field

'General, Regional & Community Issues', that have no relevance to medicine or psychology, are assigned to cluster 'j'.

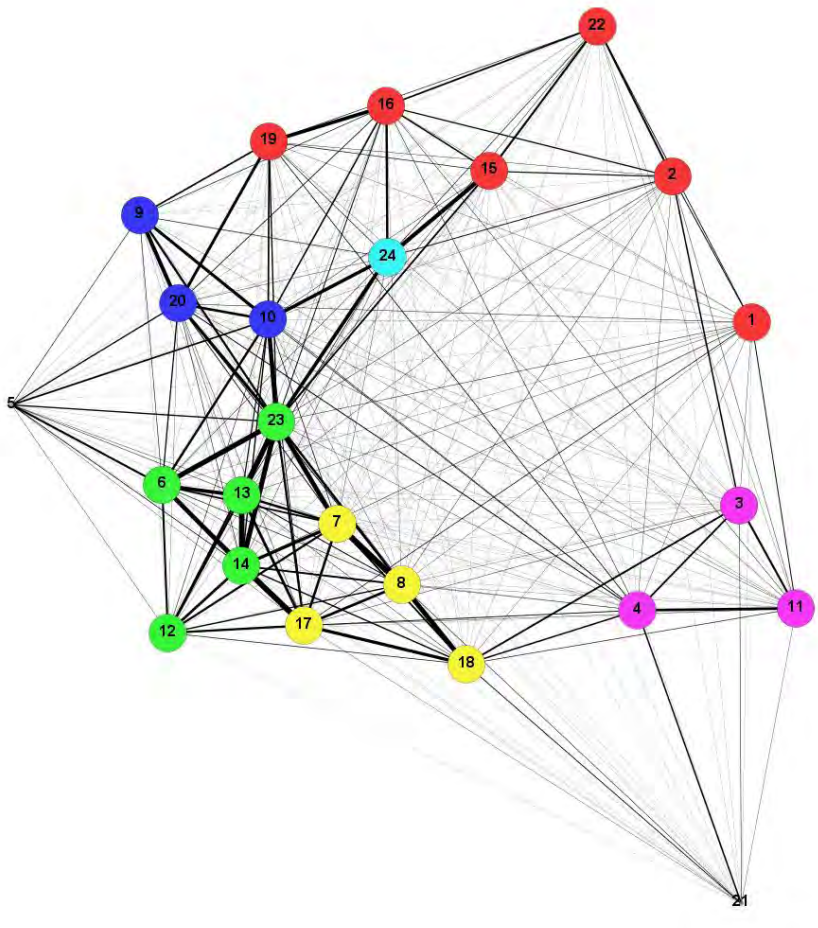


Figure 5. Map with 24 clusters based on bibliographic coupling [Data sourced from Thomson Reuters Web of Knowledge]

Comparison with ESI

A 24 cluster solution can be compared with the 22 categories from the classification of Thomson Reuters' Essential Science Indicators (ESI). Unlike most classification schemes, this classification system provides just like our cluster solutions a structure, where each journal is assigned to only one single category. This means that we can calculate the concordance between the two classification systems. The appendix presents the distribution of journals across both systems. Janssens et al. (2009) showed very low mean silhouette values for the ESI category system in a space with respectively textual distances, cosine

similarities of cross-citation vectors and combined distances. As can be seen from the table in the Appendix the same situation occurs here as well. Also in the present study, not all clusters have a unique counterpart in the ESI classification system and vice versa (cf. Janssens et al., 2009). Notably, the ESI fields *clinical medicin*e and *engineering*, *mathematics* and *social sciences, general* are almost uniformly spread over numerous clusters.

Table 3. Concordance measured with Jaccard Index between 14 clusters and the Leuven subject classification system in 15 disciplines [Data sourced from Thomson Reuters Web of Knowledge]

	a	b	c	d	e	f	g	h	i	j	k	l	m	n
Agriculture & Environment						0.30								
Biosciences													0.32	
Chemistry				0.08		0.06								0.48
Engineering				0.15								0.35		
Geosciences & Space Sciences			0.37	0.12										
Mathematics											0.44	0.18		
General & Internal Medicine	0.08	0.36												
Non-Internal Medicine Specialties	0.19	0.19					0.13	0.16						
Neurosciences & Behaviour							0.28	0.21						
Economical & Political Issues									0.08	0.43				
Physics				0.26										
Biomedical Research													0.20	
General, Regional & Community Issues								0.15	0.08	0.16				
Arts & Humanities									0.55					
Biology	0.13				0.34	0.14								

Conclusions

The application of the second-order similarities proved to be surprisingly stable, and resulted in high-quality cluster solutions. Notably the six-cluster solution provided the best result. The number of singletons, that had to be removed, was marginal: Only ten journals representing 0.12% out of the 8282 journals had to be removed from the classification. The main advantage of this method is that clustering can be made as soon as a new database volume is available. The only issue is the lacking cluster labelling that cannot directly be obtained from the method. As a substitute, intellectual classification schemes can be used as reference system. Cluster labelling was made on the basis of the Leuven-Budapest subject-classification scheme that also allowed the comparison with the existing two-level journal classification system developed in Leuven. In all, the results have been found to provide a well-balanced hierarchical system of 6–14–24 clusters.

The further comparison with the 22 field classification system of the Essential Science Indicators does, however, revealed some striking deviations. These concerned, above all, the fields of clinical medicine, engineering, mathematics and the social sciences. New developments in

computer science, neuroscience and psychology as well as in public health (cf. Glänzel & Thijs, 2011) do certainly contribute to such growing deviation.

The main objective of this study was to analyse whether the proposed methodology is appropriate for multi-level journal clustering and to what extent the solutions fit in the framework of traditional subject classification. Further comparison with other solutions such as cross-citation and hybrid methods will be part of future research.

References

Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: experimental comparison of five approaches. *Journal of Informetrics*, 3(1), 49–63.

Bassecoulard, E. & Zitt, M., (1999). Indicators in a research institute: A multi-level classification of scientific journals. *Scientometrics*, 44(3), 323–345.

Bastian M., Heymann S. & Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Garfield, E. (1998). *Mapping the world of science*. The 150 Anniversary Meeting of the AAAS, Philadelphia, PA, http://www.garfield.library.upenn.edu/papers/mapsciworld.html.

Glänzel, W. & Czerwon, H.J. (1996), A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2), 195–221.

Glänzel, W., Schubert, A. (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56 (3), 357–367.

Glänzel, W. & Thijs, B. (2011), Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88 (1), 297–309.

Janssens, F. (2007). *Clustering of scientific fields by integrating text mining and bibliometrics*. Ph.D. Thesis, Faculty of Engineering, KU Leuven, Belgium. http://www.hdl.handle.net/1979/847

Janssens, F., Glänzel, W. & de Moor, B. (2008), A hybrid mapping of information science. *Scientometrics*, 75(3), 607–631.

Janssens, F., Zhang, L., De Moor, B. & Glänzel, W., (2009) Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management*, 45(6), 683-702.

Jarneving, B. (2005). *The combined application of bibliographic coupling and the complete link cluster method in bibliometric science mapping*. Ph.D. Thesis, University College of Borås/Göteborg University, Sweden.

Leydesdorff, L. (2004), Clusters and maps of science journals based on bi-connected graphs in Journal Citation Reports. *Journal of Documentation*, 60(4), 371–427.

Narin, F., Carpenter, M., & Nancy, C. (1972). Interrelationships of scientific journals. *Journal of the American Society for Information Science*, 23(5), 323–331.

Narin, F. (1976), *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Computer Horizons, Inc., Washington, D.C.

Rousseeuw, P.J. (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65

Sen, S.K. & Gan S.K. (1983), A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library Science and Documentation*, 30, 78–82.

Thijs, B., Schiebel, E., & Glänzel, W. (2013), Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, DOI 10.1007/s11192-012-0896-1, in press.

Zhang, L., Janssens, F., Liang L.M. & Glänzel, W. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research. *Scientometrics*, 82(3), 687–706.

Appendix

Distribution of journals across 24 clusters and 22 ESI fields [Data sourced from Thomson Reuters Web of Knowledge]

ESI field	*24 cluster solution*																								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*
agricultural sciences	2	7	3	4	2	5	1	4	26	33	4	3	4	11	2	2	9	4	4	6	1	2	8	7
biology & biochemistry	4	5	6	8	3	3	6	4	7	23	2	3	10	22	7	8	9	6	4	23	11	6	100	21
chemistry	1	11	8	8	8	4	6	6	7	22	5	8	10	8	10	15	4	16	9	13	23	8	33	148
clinical medicine	8	36	22	39	8	68	50	26	18	17	15	119	178	214	26	37	87	42	23	30	50	14	99	74
computer science	1	73	7	4	2	5		4	2	2		1	8	6	5	19	5	6	2	7	4	11	7	9
economics & business	1	7	22	10		5	2	5	2	9	53	11	6	8	4	5	9	13	8	12	13	7	9	19
engineering	11	127	12	19	3	11	2	15	6	28	9	10	22	29	17	84	14	21	16	26	14	27	28	45
environment/ecology	3	4	2	8	3	3	1		16	29	2	2	7	4	1	9	4	8	14	44	8	4	8	7
geosciences	2	3	1	7	2	3	2	5	2	8	2	7	2	11	7	20	13	4	66	12	8	10	9	18
immunology	1	3		2	1	25		1	1	2	1		3	7		2	4	1	1	1	2	3	5	
materials science	1	9	4	5	1		4	1	5	8	4	5	2	6	5	17	3	8	4	9	12	5	10	71
mathematics	24	27	2	8	1	4	3	4	2	2	4	4	7	5	5	16	11	10	6	8	16	74	12	20
microbiology		1	1		1	12	2	3	1	19		2	6	2	1	4	2	1	2		4	3	10	1
molecular biology & genetics	1		2	4	3	7	3	5	3	3	1	3	8	1	8	5	2	2	3	16			88	4
multidisciplinary				1	1	1		1			1		1		1	3	1			3	3		4	3
neuroscience & behavior		5	1	2		1	70	32	2	3	2	7	7	5		2	4	3	1	8	3	1	7	6
pharmacology & toxicology	2	1	1	1	1	2	1	2	1	3		2	5	8	2	3	6	4	2	4	2	4	55	10
physics		15	2	7	2	3	2	4	9	6	1	14	4	2	49	16	2	8	7	10	14	13	7	29
plant & animal science	1	9	8	19	49	9	6	6	46	13	6	9	8	13	5	20	13	19	23	152	17	9	42	21
psychiatry/psychology	4	9	8	11	2	4	9	76	2	8	9	13	6	14	5	10	17	104	5	9	15	5	13	12
social sciences, general		26	21	119	11	18	2	28	7	16	28	21	19	17	8	28	70	89	18	26	48	17	20	37
space science			1		1		2		1		1			3	20	4	1		2	1	1	1	4	1

BUILDING A MULTI-PERSPECTIVE SCIENTOMETRIC APPROACH ON TENTATIVE GOVERNANCE OF EMERGING TECHNOLOGIES

Daniele Rotolo[1], Ismael Rafols[2], Michael Hopkins[3], and Loet Leydesdorff[4]

[1] *d.rotolo@sussex.ac.uk*
University of Sussex, SPRU - Science and Technology Policy Research, Brighton - BN1 9QE (United Kingdom)

[2] *i.rafols@sussex.ac.uk*
Universitat Politècnica de València, Ingenio (CSIC-UPV), Camí de Vera, s/n, València - 46022 (Spain)
and
University of Sussex, SPRU - Science and Technology Policy Research, Brighton - BN1 9QE (United Kingdom)

[3] *m.m.hopkins@sussex.ac.uk*
University of Sussex, SPRU - Science and Technology Policy Research, Brighton - BN1 9QE (United Kingdom)

[4] *loet@leydesdorff.net*
University of Amsterdam, Amsterdam School of Communication Research (ASCoR), Kloveniersburgwal 48, Amsterdam - 1012 CX (The Netherlands)

Abstract

The present paper proposes a multi-perspective scientometric approach on the phenomenon of tentative governance of emerging technologies. Tentative governance can be conceived as particular forms of governance that aims to flexibly address the uncertainties and dynamics featuring in emerging science and technology, thus stimulating and supporting the emergence process. The development of a multi-perspective scientometric approach is critical to inform researchers and policy makers on this phenomenon in a comprehensive and timely manner. Our approach builds on three distinct but interrelated dimensions of the emergence process, i.e. the (i) cognitive, (ii) social, and (iii) geographical dimensions. Each dimension can be dynamically investigated from different perspectives, which are defined by the combination of units of analysis and sources of data, and by using the wide range of techniques and tools the scientometric community has developed. We discuss and explore the multi-perspective approach across three case studies, namely RNA interference (RNAi), Human Papillomavirus (HPV) and Thiopurine Methyltransferase (TPMT) testing. We selected these case studies since they significantly differ in terms of pace of growth and scale of research thus providing the opportunity for a comprehensive discussion.

Conference Topics

Research Fronts and Emerging Issues (Topic 4), Collaboration Studies and Network Analysis (Topic 6), and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction

To what extent can scientometrics contribute to inform researchers and policy makers on the phenomenon of tentative governance of emerging technologies? Can scientometrics tools and techniques be used to trace this phenomenon? The governance of novel science and technologies is assuming an increasing relevance for policy makers given their potential to generate profound (both positive and negative) social changes such as creating new industries as well as dramatically changing/destroying existing ones (e.g. Day and Schoemaker, 2000; Cozzens et al., 2010). Defining governing arrangements for emerging technologies is however a complex activity given the uncertainties and dynamics that feature in their emergence. Their development may follow certain directions rather than others as a result of a variety of factors. These include the visions, goals and expectations of the actors involved (e.g. Geels, 2002; Wiek et al., 2007; Stirling, 2009). These actors are at the same time being regulated, and actively regulating the emergence process both via intentional government arrangements or by means of non-intentional effects of their activities (Braithwaite and Drahos, 2000). Rip (2010) refers to this situation where un-intential influences matter as *de facto* governance. In such cases, traditional forms of governance are unsuitable because of their highly routinized and structured nature, which in times of more incremental changes gives them legitimacy. Novel governance approaches have begun to appear instead, aiming to address the complexity, interdependencies, and contingencies characterizing the process of emergence. The main characteristic distinguishing these novel approaches to the traditional ones is their 'tentative' nature (e.g. Hagendijk and Irwin, 2006; Stirling, 2006; Wiek et al., 2007; Boon et al., 2011).

Forms of governance that are tentative aim to create a space where the generation of a number of options for the development of emerging technologies is desired and supported. In other words, a space stimulating and sustaining the exploration phase rather than exploitation one, which in turn, by definition, requires narrowing the scope of available options of development (March, 1991). As said, this idea has been associated with the concept of 'tentativeness' according to which the design of governance is such that the governance attempts to flexibly address the uncertainties and dynamics featuring in the emergence process. However, our understanding of this empirical phenomenon is limited. Extensive research needs to be undertaken in this area from both theoretical and methodological point of views. In particular, policy makers and researchers need tools that are capable of capturing, in a timely and informative manner, the dynamics of the emergence process.

From this perspective, scientometrics may represent a valuable source of information. A number of advanced mapping techniques have recently been developed. The value of these techniques resides in the potential to inform policy makers and researchers on *de facto* governance structure and dynamics. Although there are many studies on the dynamics and cognitive or social structures of emerging technologies, studies building the connections between structure and dynamics and the governance are scarce. The aim of the papers is to address this gap. Specifically, we propose a multi-perspective scientometric approach and discuss how this approach can trace the dynamics of emerging technologies and serve as an interpretative tool of *de facto* governance occurring in the process of emergence. A multi-perspective approach has the potential to provide a comprehensive and informative view on science and technology emergence, which is essential given the complexity of the emergence process, broad constellation of involved actors, and rapid dynamics as well as novel technologies crossing multiple domains of which data have been archived in different sources.

The Multi-perspective Approach

To build different perspectives on emerging technologies we combine multiple data sources (e.g. publications, patents, inter-organisational alliances) and units of analysis that portray different analytical lenses (e.g. social, cognitive) at various levels of aggregation (e.g. individual, organisation, discipline). As reported in Figure 1, the combination between the range of data sources and units of analysis defines perspectives, which can be observed across time to capture evolutionary dynamics.[27] Each perspective can be mapped by using a number of techniques scientometricians have developed. These techniques can be specifically used to inform policy makers and researchers on three distinct but interrelated dimensions of the emergence process, i.e. the (i) cognitive, (ii) social, and (iii) geographical dimensions. These dimensions evolve and interact with each other in a nonlinear manner across time leaving signatures on the different perspectives (Leydesdorff et al., forthcoming), i.e. 'data source-unit of analysis' combinations. We now discuss how one can use the multi-perspective scientometric approach to trace the evolutionary dynamics of emerging technologies and then shed light on *de facto* governance across the three aforementioned dimensions.

The Cognitive Dimension

When new sciences and technologies emerges, epistemic developments occur in terms of discoveries, novel theories, or changes in technical developments such as experimental systems, materials, methods and instrumentation (Rheinberger, 1997; Joerges and Shinn, 2002). These developments constitute the cognitive dimension of the emergence process. In this regard, scientometrics has developed robust mapping techniques to dynamically and timely trace the structure of this

[27] Additional data sources and units of analysis can be identified. Yet, for clarity and space limitation we focus on those ones reported in Figure 1.

dimension. These techniques mainly use two types of data sources, i.e. publication and patent data. Until recently, scholars' efforts in using publication data have been focused on the development of maps of science circumscribed only to the publications of the topic which are based on co-citation or bibliographic coupling maps (e.g. Leydesdorff, 2007), and co-words maps (e.g. Cambrosio et al., 2006). However, in the last decade, an important development has been the creation of so-called global maps of science, which represent all science in one map (e.g. Klavans & Boyack, 2009; Rafols, Porter, & Leydesdorff, 2010). The elements of the map can be disciplines (e.g. Leydesdorff and Rafols, 2009), journals (e.g. Leydesdorff, 2006), or research topics (e.g. Waltman & van Eck, forthcoming).

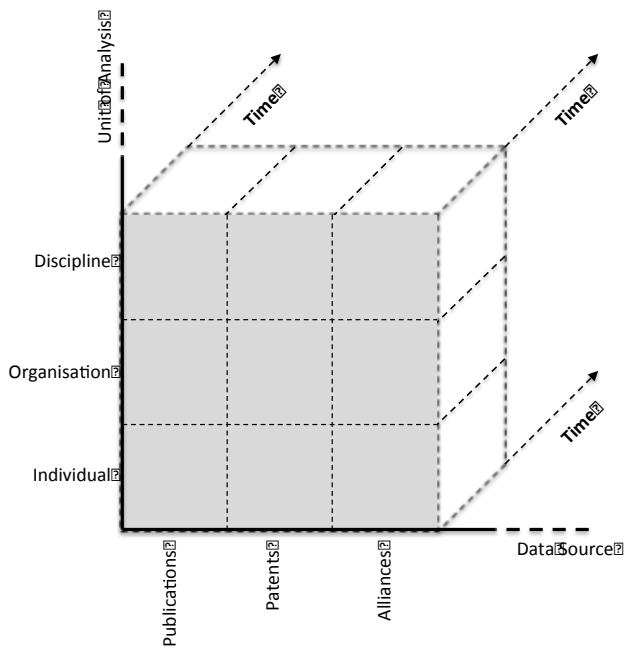


Figure 1. Multi-perspective approach: combining data sources and units of analysis.

Overlay techniques can be applied to this map in order to project an entity's (e.g. individual, organisation, community, research field) publishing activity. By creating overlays across time the evolutionary dynamics of the given entity can be revealed in the overall structure of science. Rafols, Porter, and Leydesdorff (2010), for instance, developed a map of science which elements are represented by scientific disciplines and measured in terms of Web of Science (WoS) subject categories. This map can be animated across time thus showing how scientific research activities spread across the domains of science and social science (Figure 2). Building on the same type of overlay technique, Leydesdorff, Rotolo, and

Rafols (2012) developed a map based on the Medical Subject Headings (MeSH) of PubMed/MEDLINE, which provide the practitioners' view on the use of publications—that is a different type of cognitive perspective. This map however is only suitable to trace the dynamics of emerging technologies in the medical sector. Scientometricians have also developed global maps and overlay techniques to trace patenting activity (Newman et al., 2011; Leydesdorff and Bornmann, 2012; Leydesdorff, Kushnir, & Rafols, in press). The elements of these maps represent technological classes (generally IPC classes) to which patents are assigned.

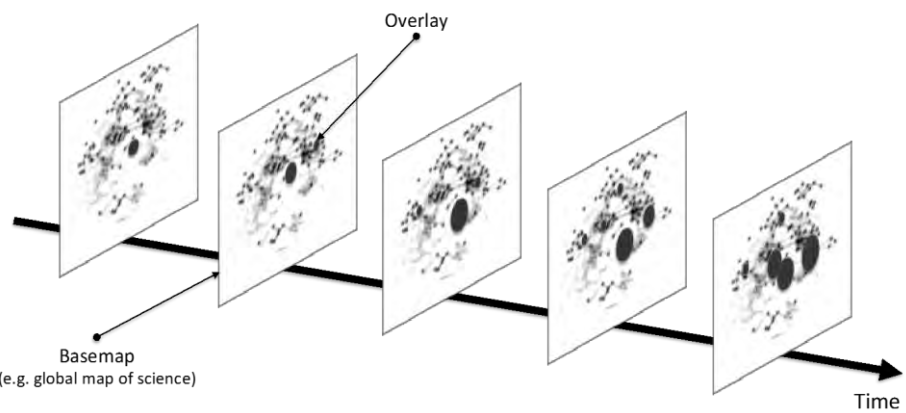


Figure 2. Map of science, overlays, and evolutionary dynamics.

The Social Dimension

The structure of the relationships between actors surrounding the emergence process and its dynamics play a critical role in shaping the development of novel technologies (e.g. Latour, 1993). These connections are channels through which actors gain access to and mobilise knowledge, resources, and power. Networks of agents therefore affect and are affected by emerging technologies (e.g. Klijn and Koppenjan, 2000). The use of co-authorship data in publications to trace network and social dynamics has a long tradition in scientometrics (Crane, 1972). As discussed, these networks can be built at different units of analysis such as individual researchers, organisations as well as disciplines. However, co-authorship data in publications is not the only source to trace the dynamics in the social dimension of the emergence process. Co-invention activities and inter-organisational alliances data represent also valuable sources to build perspectives on this dimension.

The Geographical Dimension

The geographical diffusion of emerging technologies can also be traced. Scientometricians have developed also in this case a number of applications to localise the production of publications and patents (Kwakkel et al., Forthcoming).

For instance, Leydesdorff and Bornmann (2011; 2012) have developed mapping techniques that overlay publications and patents on Google Maps.

Case Studies

We build our discussion of the aforementioned multi-perspective scientometric approach by drawing on three illustrative case studies: (i) RNA interference (RNAi), (ii) Human Papillomavirus (HPV) and (iii) Thiopurine Methyltransferase (TPMT) testing. Uncertainty and rapid dynamics feature in the three cases, which make them suitable examples to discuss the multi-perspective scientometric approach. It is worth noting that the interest in comparing these cases lies in their different position in the innovation chain. RNA interference is a discovery leading to a research technology (Joerges and Shinn, 2002) that can be applied to different purposes in biomedical research—hence it is positioned close to basic research. HPV testing is a diagnostic tool aimed to diagnose a specific disease—hence an emerging technology with a dominant domain of application. TPMT testing is a diagnostic tool adopted for drugs that are used in several diseases treated in oncology, dermatology and gastroenterology. Therefore, analysing these cases provides an opportunity to discuss the multi-perspective scientometric approach across different contexts of the emergence. This diversity will enrich our discussion.

Table 1. Search strings used in WoS and relative number of publications retrieved.

Case Study	Search string	Number of publications (1990-2011)
RNAi	(TS=siRNA OR TS=RNAi OR TS="RNA interference" OR TS="interference RNA")	41,948
HPV testing	(TS="HPV*" OR TS="Human Papilloma Virus*" OR TS="Human Papillomavirus*" OR TS="Human Papilloma*virus*" OR TS="Human*Papilloma*Virus*") AND (TS="Cervical" OR TS="Cervix") AND (TS="diagnos*" OR TS="test*" OR TS="assay" OR TS="detect*" OR TS="screen*" OR TS="predict*")	10,019
TPMT testing	(TS=TPMT OR TS= "Thiopurine Methyltransferase")	1,246

In terms of data, we first retrieved from ISI WoS data on publications up to 2011. We specifically identified for each case study a set of ad hoc keywords by using multiple sources (e.g. interviews with experts, reviews and previous research on the cases). These keywords and their combinations were then searched in scientific articles' titles, abstracts and lists of keywords, i.e. "topic" field of WoS

(see Table 1). Similar results in publishing activity can be found using alternative databases as SCOPUS and PubMed/MEDLINE. Figure 3 shows the rapid emergence of these novel technologies as revealed by the number of published scientific articles. We also reported the top-5 ISI subject categories to which scientific articles have been assigned. While a growing research activity features in all three case studies, the pace of this growth as well as the scale of research is different. For instance, the growth in the number of publications for RNAi is steeper than HPV and TPMT testing, respectively. We now briefly describe the case studies by providing examples on how the multi-perspective scientometric approach can be used to gives us insights on *de facto* governance structure and dynamics.

Case Study 1: RNA interference (RNAi)

The first case study will be focused on RNAi, which is a technique for gene silencing. Genes play a critical role in the progression of cancers, genetic diseases, and infection agents. Theoretically, by silencing specific genes one can stop the progression of a given disease. The RNAi silencing mechanism was discovered in 1998 (Fire et al., 1998) and its discovery reshaped the landscape of research on RNAi creating important expectations on the therapeutic applications (e.g. Sung and Hopkins, 2006; Leydesdorff and Rafols, 2011; Leydesdorff et al., 2012). One of the main characteristics of RNAi is that it can be conceived as a general purpose technology for research in labs. By mapping the publication activity in RNAi area with overlay techniques applied to the global map of science (Rafols et al., 2010; Leydesdorff and Rafols, 2011) the structure of the cognitive dimension of the emergence process is revealed. For instance, Figure 4(a), which depicts the overlay map for the 2007-2011 period, shows how RNAi has diffused across various fields of science.[28]

The social dimension of the emergence process can be also traced by using collaboration networks. We reported as example the inter-organisational alliances networks of companies involved in the development of RNAi in Figure 4(b). The network shows how the two key players in the emergence process of RNAi, i.e. Alnylam Pharmaceuticals and ISIS Pharmaceuticals (grey nodes), are strongly connected and positioned at the centre of the network of relationships in the industry. Centrality in the network reveals capacity to have power and control and therefore capability to affect the emergence process. As discussed, overlays of publishing and patenting activities can be projected on Google Maps to investigate the geographical dimension (Bornmann and Leydesdorff, 2011; Leydesdorff and Bornmann, 2012). For instance, Figure 4(c) shows the collaboration activity (co-authorships data in publications) in RNAi domain across different cities projected onto Google Maps (Leydesdorff and Rafols,

[28] ISI subject categories are grouped in 19 macro-areas. A different colour is assigned to each macro-area (for further details see Rafols et al., 2010).

2011). Locations of and interactions among the constellation of actors involved in *de facto* governance can thus be identified.

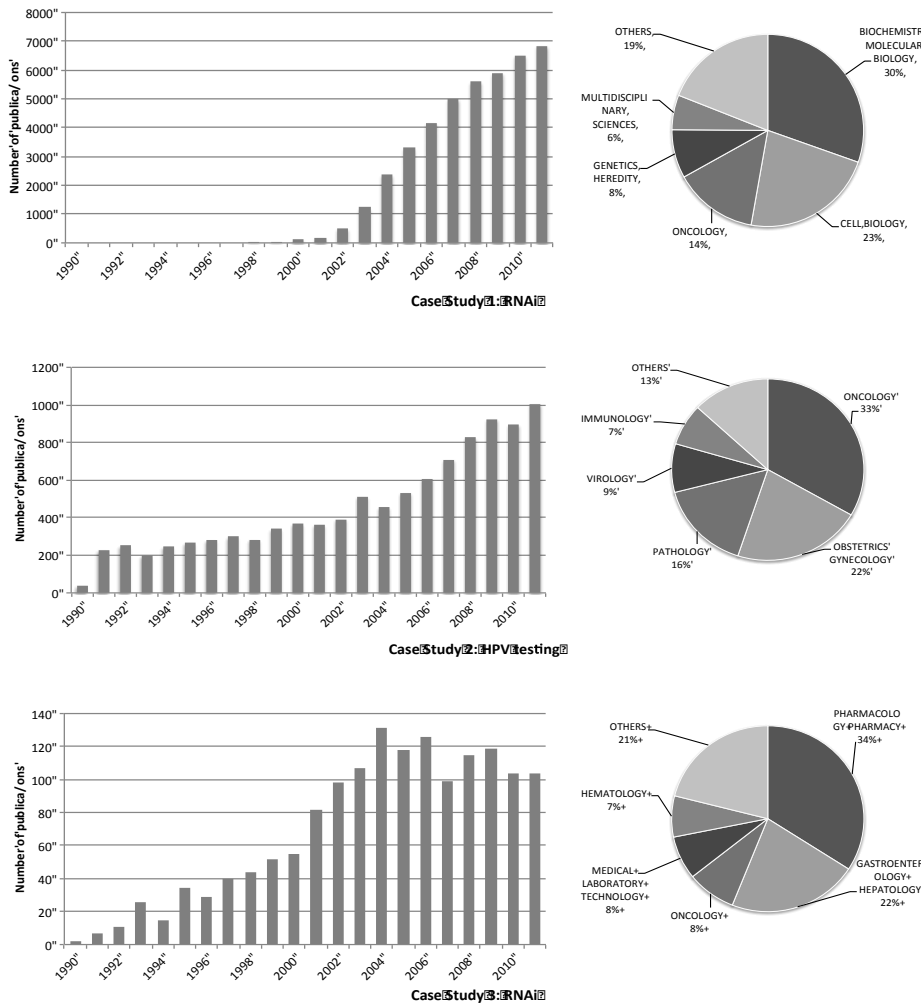


Figure 3. Three case studies: publishing activity.

Case Study 2: Human Papillomavirus (HPV) testing

The second case study is related to the development of a diagnostic technology for the detection of HPV. In the 1980s, HPV viruses were discovered as strongly associated with cervical cancer, which has a significant impact on women population. About 500,000 new cervical cancers occur and cause about 250,000 deaths each year. This has led to the development of a large screening program with 100+ million tests performed annually. While this screening has been mainly performed by using the Pap-test, the discovery on the association between HPV and cervical cancer opened the space for the development of a competing and

more sensitive technology for the detection of the HPV and then of the cervical cancer based on molecular diagnostics technology, namely the HPV testing (Casper and Clarke, 1998; Hogarth et al., 2012). In this process, a private actor, Digene Corp., played a crucial role in establishing the HPV-test as gold standard to use together withthe Pap-test for the cervical cancer screening (Hogarth et al., 2012). We reported in Figure 4(d) the collaboration network (based on co-authorships data in publications from 1997 to 2001) in the HPV testing area at organisational level.[29] Digene and the organisations to which the company was directly connected are represented with yellow and red nodes, respectively. The network reveals the social structure of the *de facto* governance since a detailed analysis shows Digene collaborating with main institutions in the field (e.g. National Cancer Institute, Kaiser Permanente) involved in the regulation of the cervical cancer screening. In other words, while Digene's activity was 'regulated' (e.g. FDA approval), Digene was affecting the developments and dynamics in cervical cancer screening.

Case Study 3: Thiopurine Methyltransferase (TPMT) testing

The third case study is focused on an emerging class of pharmacogenetic tests (which predict adverse events affecting patient's health) (Hopkins et al., 2006), i.e. the TPMT testing. TPMT is an enzyme in the human body responsible for metabolising thiopurine drugs. Cytotoxic Thiopurine drugs such as Azathioprine are used to treat a range of conditions including leukaemia, and autoimmune diseases (such as Lupus, or rheumatoid arthritis). However, where a patient has mutations in the gene encoding TPMT, they may be at increased risk of toxicity from a build up of thiopurines. Therefore, several types of TPMT test started to emerge across a number of clinical fields of use. Some of the tests are based on patented technology (with an IP holder that seeks to aggressively exploit their exclusivity in certain markets such as the USA, but not in others such as the UK). The relatively small market for the (off-patent) thiopurine drugs represents a small 'niche' made up of other niches (several specialist fields—such as transplantation, gastroenterology, rheumatology, paediatric oncology). In these different niches, evidence of clinical utility of the test is highly contested (there is disparity in use of the tests between fields and clinical guidelines). In this case, de facto governance operates through medical guidelines. Interestingly, analyses of medical guidelines reveral significant differences in the use of TPMT testing across disciplines. Investigating the cognitive structure of the emergence process can reveal for instance different translations and interpretations of basic knowledge on TPMT.

[29] We reported only the giant component and the nodes' size is proportional to organisations' degree centrality.

Conclusion

We discussed how scientometrics may represent a valuable source to inform researchers and policy makers on *de facto* governance structure and dynamics of emerging science and technologies. We proposed a multi-perspective approach. Each perspective, resulting from the combination of units of analysis and data sources, can investigate the emergence process across three dimensions - cognitive, social, and geographical. We believe this approach has the potential to timely and comprehensively inform researchers and policy makers on the dynamics featuring in the process of emergence and especially on *de facto* governance.

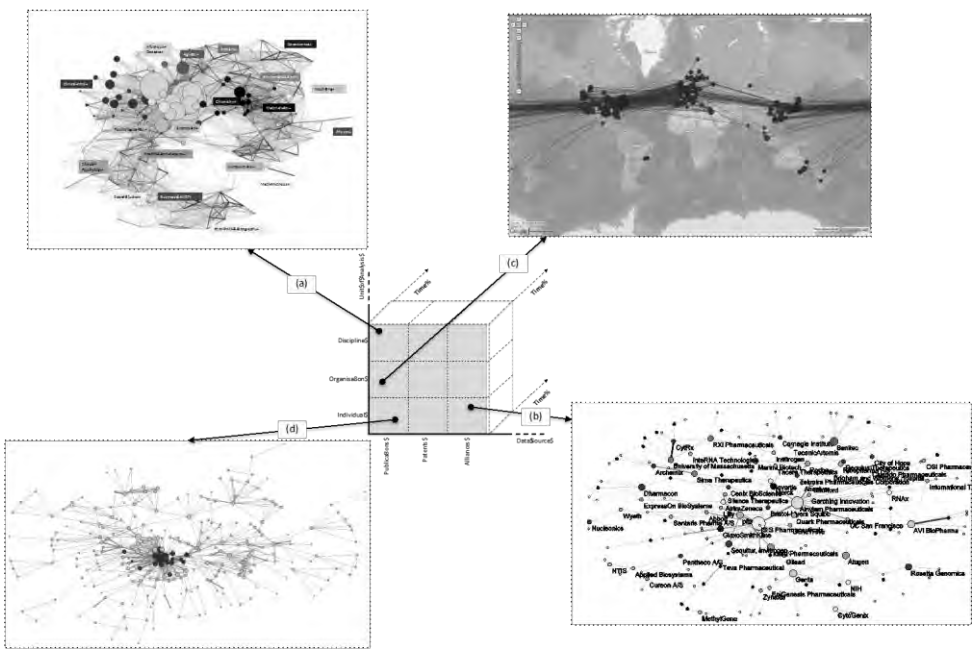


Figure 4. Multi-perspective scientometrics approach, techniques, and examples.

Acknowledgments

The paper is part of the ESRC project 'Mapping the Dynamics of Emergent Technologies' (RES-360-25-0076). We also acknowledge support from the US National Science Foundation (Award #1064146 - "Revealing Innovation Pathways: Hybrid Science Maps for Technology Assessment and Foresight"). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Boon, W. P. C., Moors, E. H. M., Kuhlmann, S., & Smits, R. E. H. M. (2011). Demand articulation in emerging technologies: Intermediary user organisations as co-producers? *Research Policy*, *40*(2), 242–252.

Bornmann, L., & Leydesdorff, L. (2011). Which cities produce more excellent papers than can be expected? A new mapping approach, using Google Maps, based on statistical significance testing. *Journal of the American Society for Information Science and Technology*, *62*(10), 1954–1962.

Braithwaite, J., & Drahos, P. (2000). *Global Business Regulation*. Cambridge University Press.

Cambrosio, A., Keating, P., Mercier, S., Lewison, G., & Mogoutov, A. (2006). Mapping the emergence and development of translational cancer research. *European Journal of Cancer*, *42*(18), 3140–3148.

Casper, M. J., & Clarke, A. E. (1998). Making the Pap smear into the "right tool" for the job: Cervical cancer screening in the USA, circa 1940-95. *Social Studies of Science*, *28*(2), 255–290.

Cozzens, S. E., Gatchair, S., Kang, J., Kim, K.-S., Lee, H. J., Ordóñez, G., & Porter, A. (2010). Emerging technologies: quantitative identification and measurement. *Technology Analysis & Strategic Management*, *22*(3), 361–376.

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. University of Chicago Press.

Day, G. S., & Schoemaker, P. J. H. (2000). Avoiding the pitfalls of emerging technologies. *California Management Review*, *42*(2), 8–33.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, *391*(6669), 806–811.

Geels, F. (2002). Technological transitions as evolutionary reconfiguration processes: a multi-level perspective and a case-study. *Research Policy*, *31*(8–9), 1257–1274.

Hagendijk, R., & Irwin, A. (2006). Public deliberation and governance: Engaging with science and technology in contemporary europe. *Minerva*, *44*(2), 167–184.

Hogarth, S., Hopkins, M. M., & Rodriguez, V. (2012). A molecular monopoly? HPV testing, the Pap smear and the molecularisation of cervical cancer screening in the USA. *Sociology of Health & Illness*, *34*(2), 234–250.

Hopkins, M. M., Ibarreta, D., Gaisser, S., Enzing, C. M., Ryan, J., Martin, P. A., … Forde, T. (2006). Putting pharmacogenetics into practice. *Nature Biotechnology*, *24*(4), 403–410.

Joerges, B., & Shinn, T. (2002). *Instrumentation Between Science, State and Industry*. Springer.

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455–476.

Klijn, E. H., & Koppenjan, J. F. M. (2000). Public management and policy networks: foundations of a network approach to governance. *Public Management*, *2*(2), 135–158.
Kwakkel, J. H., Carley, S., Chase, J., & Cunningham, S. W. (Forthcoming). Visualizing geo-spatial data in science, technology and innovation. *Technological Forecasting and Social Change*.
Latour, B. (1993). *The Pasteurization of France*. Harvard University Press.
Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology*, *57*(5), 601–613.
Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, *58*(9), 1303–1319.
Leydesdorff, L., & Bornmann, L. (2012). Mapping (USPTO) patent data using overlays to Google Maps. *Journal of the American Society for Information Science and Technology*, *63*(7), 1442–1458.
Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362.
Leydesdorff, L., & Rafols, I. (2011). Local emergence and global diffusion of research technologies: An exploration of patterns of network formation. *Journal of the American Society for Information Science and Technology*, *62*(5), 846–860.
Leydesdorff, L., Rotolo, D., & De Nooy, W. (Forthcoming). Innovation as a nonlinear process, the scientometric perspective, and the specification of an "Innovation Opportunities Explorer". *Technology Analysis & Strategic Management*.
Leydesdorff, L., Rotolo, D., & Rafols, I. (2012). Bibliometric perspectives on medical innovation using the medical subject Headings of PubMed. *Journal of the American Society for Information Science and Technology*, *63*(11), 2239–2253.
Leydesdorff, L., Kushnir, D., & Rafols, I. (in press). Interactive Overlay Maps for US Patent (USPTO) Data Based on International Patent Classifications (IPC). *Scientometrics*.
March, J. G. (1991). Exploration and exploitation In organizational learning. *Organization Science*, *2*(1), 71–87.
Newman, N. C., Rafols, I., Porter, A. L., Youtie, J., & Kay, L. (2011). Patent overlay mapping: Visualizing technological distance. *In preparation*.
Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, *61*(9), 1871–1887.
Rheinberger, H.-J. (1997). *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press.

Rip, A. (2010). De facto governance of nanotechnologies. In *Dimensions of Technology Regulation* (Morag Goodwin, Bert-Jaap Koops and Ronald Leenes.). Nijmegen: Wolf Legal Publishers.

Stirling, A. (2006). Precaution, foresight and sustainability: reflection and reflexivity in the governance of science and technology. In *Reflexive Governance for Sustainable Development* (Voß, J-P et al.). Cheltenham: Edward Elgar Publishing.

Stirling, A. (2009). *Direction, distribution and diversity! Pluralising progress in innovation, sustainability and development*. STEPS Working Paper 32. Brighton: University of Sussex.

Sung, J. J., & Hopkins, M. M. (2006). Towards a method for evaluating technological expectations: Revealing uncertainty in gene silencing technology discourse. *Technology Analysis & Strategic Management*, *18*(3-4), 345–359.

Waltman, L., & Van Eck, N. J. (Forthcoming). A new methodology for constructing a publication-level classification system of science. *Journal American Society for Information Science and Technology*.

Wiek, A., Zemp, S., Siegrist, M., & Walter, A. I. (2007). Sustainable governance of emerging technologies—Critical constellations in the agent network of nanotechnology. *Technology in Society*, *29*(4), 388–406.

CAREER AGING AND COHORT SUCCESSION IN THE SCHOLARLY ACTIVITIES OF SOCIOLOGISTS: A PRELIMINARY ANALYSIS (RIP)

Cassidy R. Sugimoto, Staša Milojević, Andrew Tsou, and Ying Ding[1]

[1] *{sugimoto, smilojev, atsou, dingying}@indiana.edu*
School of Library and Information Science, Indiana University Bloomington, Bloomington, IN (USA)

Abstract
The aging of scholars is considered an important factor in creativity, productivity, and collaborative behaviour. However, the literature lacks in both conceptualizations and operationalizations of aging, and empirical studies show wide variation across disciplines. This study focused on two approaches on aging (by career age and cohort) and examined the possible effects of aging on scholarly communication behaviour (i.e., genre, collaboration, and productivity) within sociology. Our research suggests that changes in scholarly communication patterns are related to career aging rather than cohort changes; the data did not reflect significant changes in productivity, genre choice, or collaboration from those who received their degrees in the 1960s to the present. However, there were marked differences in productivity by rank—productivity of sociologists increased rather than decreased with rank.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3) and Sociological and Philosophical Issues and Applications (Topic 13)

Introduction
Merton and Zuckerman's (1972) seminal chapter on age stratification in science bemoaned the paucity of literature on the subject and enumerated several areas of future work. In suggesting potential cohort effects in science, they reflected: "It is an exemplary question for the sociology of science directing us to one form of interaction between the social structure and the cognitive structure of science and inviting the thought that, in some of its aspects, the cognitive structure of a field may appreciably differ for sub-groups of scientists within it" (Merton & Zuckerman, 1972, p. 555). The issue of age has since been included as one social variable which could have an impact upon the intellectual structure of a field and the behavioural activities within it. However, aging has been discussed in many different ways. These are briefly described here.

Age of scientist: this represents the actual age of the scholar in years since birth. This form of aging has been studied largely in relation to the receptivity of young

minds to new ideas (Merton & Zuckerman, 1972, p. 515). Aging is often portrayed as something inherently negative; as Blackburn and Lawrence (1986) summarize: "…some college and university administrators tend to believe that as faculty members become older they will be less productive, less, creative, less innovative, less willing to adapt to a changing environment and less effective as teachers" (p. 265-266). These statements are based on studies showing a negative correlation between chronological age and variables such as productivity, creativity, and impact (e.g., Lehman, 1953; Simonton, 1988). However, there is a great deal of complexity in these results, particularly across disciplines (Simonton, 1988). In addition, there has been some indication that senior authors do not lag behind their junior colleagues when it comes to doing cutting edge research (Milojević, 2012).

Cohort-succession: this model suggests that scholars can be grouped into cohorts (typically by year of doctoral matriculation, graduation, or the receipt of a first academic job) and that scholars within a cohort behave in ways similar to each other and distinct from previous or subsequent cohorts (O'Brien, 2011). Cohort-succession is in line with the concept of codification—a phenomenon in which scholars are encumbered with their views of the discipline at an early age and retain these throughout their career (Merton & Zuckerman, 1972). One might also extend this to scholarly behaviours—the patterns taught during doctoral education may remain embedded in a scholar's work practices.

Career-age of scientist: this model of aging suggests that a scholar's actions change as they meet various milestones in their career (O'Brien, 2011). This differs from cohort-succession in that scholars may meet career milestones at a different rate than other scholars in their cohort. Rank advancement serves as a distinct way to identify career stage for scholars; given this, it is not uncommon for bibliometric analysis to display results by rank (e.g., Shaw & Vaughan, 2008). Studies show modification in the emphasis on different forms of scholarship across rank (Sugimoto, Russell, Meho, & Marchionini, 2008) and changes in productivity and author order through the career, with scholars largely deferring prestigious author positions to junior scholars (Long, McGinnis, & Allison, 1980; Merton & Zuckerman, 1972).

Each of the models of change implies that careers are not stable. However, studies have suggested that scholars choose and retain scholarly publication activities across their lifetime, regardless of aging (e.g., Bayer & Smart, 1991; Sugimoto & Cronin, 2012) or vary independent of aging (Cronin & Meho, 2007). Therefore, a theory of aging must also take into account individuality in the model.

The relationship between age and science is one of vast importance, particularly given the rapid developments in scholarly communication. The system is "less linear, less rigid and less opaque than before; both the process and the end products are being transformed, slowly if inexorably" (Cronin, in press). These changes in collaboration behaviour, communicative genres, and open access to published works represent a "velvet revolution in scholarly communication" (Cronin, 2012). However, inflexibility and discord among cohorts could be

detrimental to the progress of science. Therefore, this work seeks to examine aging differences (by career stage and cohort) in scholarly communication behaviour (i.e., genre, collaboration, and productivity).

Methods

This study will focus on sociologists. Sociology is no stranger to scientometric analyses. Scientometric studies of sociologists have examined productivity by department (e.g., Glenn & Villemez, 1970), ranking of sociology journals (Glenn, 1971), collaborative styles (Leahey & Reikowsky, 2008), applicability of the h-index (Ouimet, Bedard, & Gelineau, 2011), interdisciplinary knowledge exchange (Shafique, 2013), and semantic integrity of the field (Varga, 2011). The validity of scientometric studies has also been brought into question with respect to sociology. The main point of contention is the lack of sources providing comprehensive lists of all publication types used by sociologists (e.g., Najman & Hewitt, 2003; Nederhof, 2006). Therefore, this study uses manual data collection from the sociologists CVs, rather than a standard bibliometric database, in order to account for all publication types.

The lists of faculty members were generated in August 2012. The list of publications was culled from the CVs of active faculty members at "top ten" schools in sociology, as determined by consulting lists published by U.S. News & World Report.[30] Only full-time faculty were counted (i.e., lecturers, emeritus professors, and other such non-tenured individuals were excluded). Official department webpages were consulted in order to generate the lists of faculty members, as well as to ascertain job titles; individual faculty webpages were then used to harvest CV links. If an individual's webpage did not contain a link to a CV (or if the person did not have a webpage), a Google search was employed. The CVs for 21 faculty members could not be located.

CVs were mined for the dates at which individuals attained various ranks (i.e., Assistant Professor, Associate Professor, and Professor), as well as place from and the year in which the individual received his or her Ph.D. Desired publications were coded by genre. Publications that were not subjected to formal peer review were omitted (e.g., editorials), as were book reviews, working papers, and works classified as "in progress," "submitted," "forthcoming" or "under review." Lectures and other oral presentations were generally excluded, unless they were later published in a formal venue. Privately prepared papers (for example, reports to government commissions) were also excluded.

Results and Discussion

In total, data were collected on 273 sociologists in 10 programs in the United States. These individuals produced 1,214 books (including 484 edited books that

[30] Sociology: http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-humanities-schools/sociology-rankings

were not included in the analysis), 3891 chapters, 329 encyclopedia entries, 7969 journal articles, and 3 dictionary entries (not included in the analysis).

Genre and productivity

We examined the productivity across genres for faculty members of different career-ages and cohorts both synchronically and diachronically. We first performed a synchronic analysis of the types of genres published by faculty currently in the assistant, associate, and full professor rank for the last five years (2007-2011). As Figure 1 indicates, contrary to the expected decline in productivity, we find that productivity actually increases with rank. Namely, full professors outperform both associate and assistant professors in publishing their research findings in all genres. However, it must be taken into account that only the professors in the final years of their assistant professor rank will have a full five years of productivity—the data thereby rely on productivity at the doctoral level for some assistant professors.

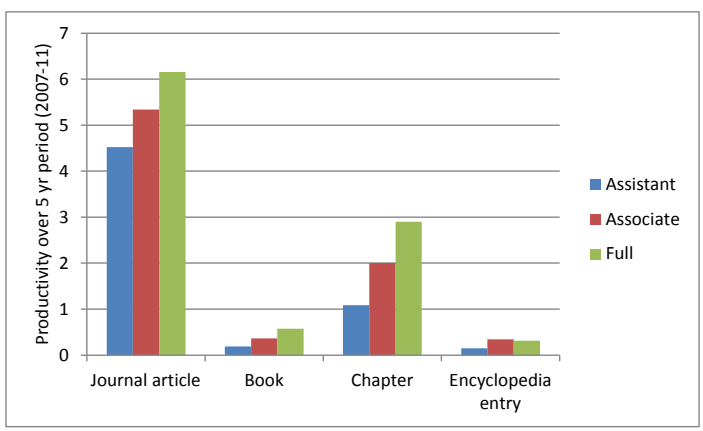


Figure 1. Productivity across genres by rank (2007-2011)

While it would be expected that age and rank (which are usually highly correlated) would have the highest effect on productivity, it is the cohort effect that would be stronger when it comes to proclivity towards particular genres. We found that journal articles comprise the most frequent unit of publication across all ranks, although the share of journal articles decreases across ranks with a corresponding increase in books and book chapters (Figure 2). Thus, it may seem that rank rather than cohort informs the choice of genre. Namely, it is hardly surprising that assistant professors who are working on obtaining their tenure favour journal articles. Journal articles are quicker to produce than books, and may carry more weight with university Promotion and Tenure committees who may be used to thinking in terms of journals and associated metrics. On the other hand, it is also not surprising that full professors produce so many book chapters. Book chapters are often written by invitation, and full professors at leading

institutions would be expected to be invited to make contributions to this literature.

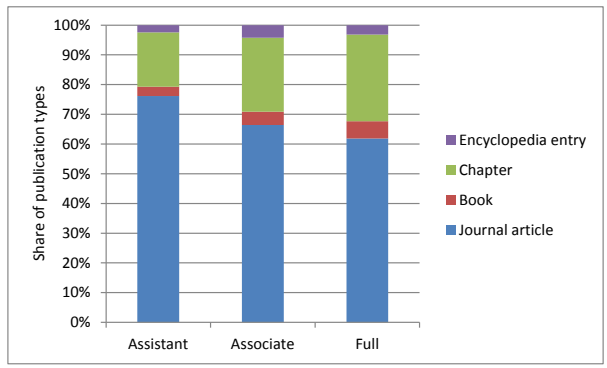


Figure 2. Share of publication types, by rank for the last 5 years (2007-2011)

In the diachronic analysis we focus on examining whether there have been changes in pre-tenure behaviour by academic cohorts since the 1960s. To define the pre-tenure stage we used a fixed time period since receipt of the doctoral degree. We decided to use an eight-year time window because 84% of the faculty in our sample obtained the status of Associate Professor within that time frame. For each faculty member we obtained the data on the number of publications they had up to the eight years after obtaining their PhD, including any publication they might have had prior to obtaining their degree. The results depicted in Figure 3 show averages in five year bins. Note that for the most recent bin (2008-2012) and part of the previous one, the faculty did not have the requisite 8 years, so their output may actually be higher. The analysis indicates that the output across all genres of pre-tenured faculty has been remarkably stable for the faculty who currently work at the elite institutions in the last forty years. This would again indicate that there is no cohort effect when it comes to productivity or the genre preference of researchers through time.

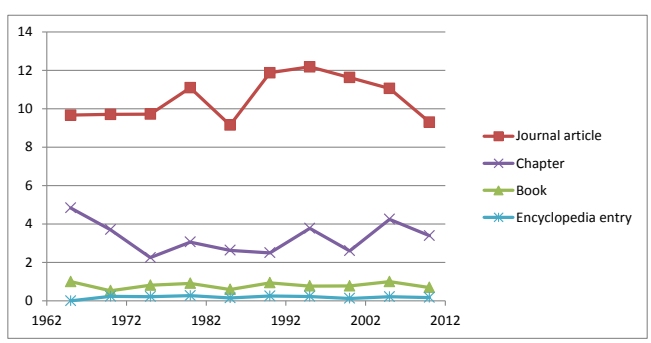


Figure 3. Average number of publications in each genre within 8 years following appointment of associate professor

Collaboration

Studies have shown that science is becoming more collaborative. Thus, one may expect that younger scholars are more open to this mode of scientific production. We examined collaboration both synchronically and diachronically, using coauthorship as an indicator of collaboration. For each rank, we examined the share of publications from the last five years which were collaboratively authored (figure 4). As shown, assistant professors work more collaboratively than associate or full professors. They have fewer single-authored papers than the other ranks. They also have more co-authored papers in which they are not the first author. One needs to keep in mind, though, that for this group the output may predominantly include scholarly works produced before a doctoral degree was obtained. Doctoral students are more likely to be in the role of co-authors (as opposed to first or single authors). This could also explain the higher proportion of journal articles in their output. There are marginal differences between associate and full professors in collaboration behaviour.

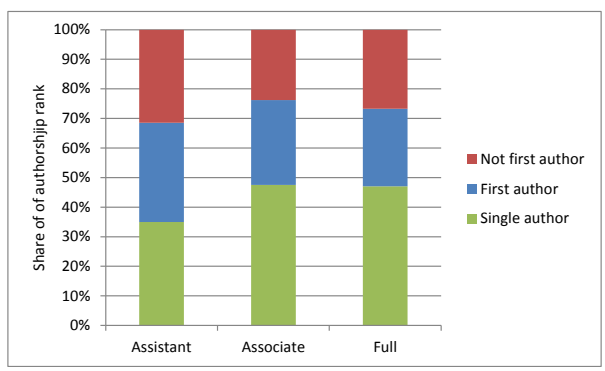


Figure 4. Percentage of collaboratively-authored papers by rank (2007-2011)

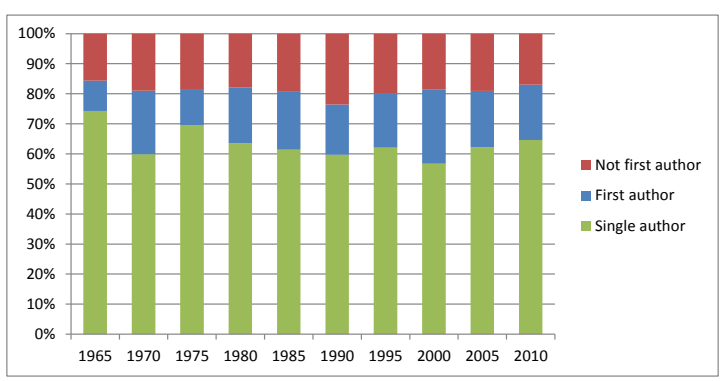


Figure 5. Proportion of collaboratively-authored papers over time

In the diachronic analysis we examine the degree to which these practices have changed over time (by doctoral cohort). In terms of collaborative practices, with the exception of the oldest cohort (who had a higher fraction of single-authored papers in their initial output than any other cohort since), the relative contribution of single-authored, first-authored, and co-authored papers has remained fairly constant over time (Figure 5).

Conclusion

Our research suggests that changes in scholarly communication patterns are more due to career aging than to cohort changes; the data did not reflect significant changes in productivity, genre choice, or collaboration from those who received their degrees in the 1960s to the present. However, there were marked differences in productivity by rank—productivity of sociologists increased rather than decreased with rank, suggesting that sociology may not be the "young man's game" that many scientific disciplines are made out to be (Merton & Zuckerman, 1972). However, this study is limited by looking at faculty members only in elite institutions of one discipline. To increase the robustness of the study, the sample should be enhanced to show variability in institutions and disciplines. Future work should continue to test various models of aging in order to understand more fully the intersection among temporal, social, and individual factors of scientific achievement.

Acknowledgments

This work was funded by the Science of Science Innovation and Policy program of the National Science Foundation (grant no. 1158670), under the grant titled: "Incubators of knowledge: Predicting protégé productivity and impact in the social sciences." The authors would like to thank Jylisa Doney for her assistance in data collection.

References

Bayer, A.E., & Smart, J.C. (1991). Career publication patterns and collaborative "styles" in American academic science. *The Journal of Higher Education*, *62*(6), 613-636.

Cronin, B. (2012, August). *The velvet revolution in scholarly communication*. Keynote address. 2nd International Conference on Integrated Information, Budapest, Hungary.

Cronin, B. (in press). Scholars and scripts, spoors and scores. In B. Cronin & C.R. Sugimoto (Eds.), *Beyond Bibliometrics: Harnessing multi-dimensional indicators of performance*. Boston: The MIT Press.

Cronin, B., & Meho, L.I. (2007). Timelines of creativity: A study of intellectual innovators in information science. *Journal of the American Society for Information Science & Technology, 58*(13), 1948-1959.

Glenn, N.D. (1971). American sociologists' evaluations of sixty-three journals. *The American Sociologist, 6*(4), 298-303.

Glenn, N.D., & Villemez, W. (1970). The productivity of sociologists at 45 American universities. *The American Sociologist, 5*(3), 244-252.
Leahey, E., & Reikowsky, R.C. (2008). Research specialization and collaboration patterns in Sociology. *Social Studies of Science, 38*(3), 425-440.
Lehman, H.C. (1953). *Age and achievement*. Princeton, NJ: Princeton University Press.
Long, J.S., McGinnis, R., & Allison, P.D. (1980). The problem of junior-authored papers in constructing citation counts. *Social Studies of Science, 10*(2), 127-143.
Merton, R.K., & Zuckerman, H. (1972). Age, aging, and age structure in science. In N.W. Storer (Ed), *The sociology of science: Theoretical and empirical investigations*. Chicago, IL: University of Chicago Press.
Milojević, S. (2012). How are academic age, productivity and collaboration related to citing behavior of researchers? *PLoS ONE, 7*(11), e49176.
Najman, J.M., & Hewitt, B. (2003). The validity of publication and citation counts for Sociology and other selected disciplines. *Journal of Sociology, 39*(1), 62-80.
Nederhof, A.J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics, 66*(1), 81-100.
O'Brien, T.L. (2011). Change in academic coauthorship, 1953-2003. *Science, Technology & Human Values*, *37*(3), 210-234.
Ouimet, M., Bedard, P-O., & Gelineau, F. (2011). Are the h-index and some of its alternatives discriminatory of epistemological beliefs and methodological preferences of faculty members? The case of social scientists in Quebec. *Scientometrics, 88*, 91-106.
Shafique, M. (2013). Thinking inside the box? Intellectual structure of the knowledge base of innovation research (1988-2008). *Strategic Management Journal, 34*, 62-93.
Shaw, D., & Vaughan, L. (2008). Publication and citation patterns among LIS faculty: Profiling a "typical professor." *Library & Information Science Research, 30*(1), 47-55.
Simonton, D.K. (1988). Age and outstanding achievement: What do we know after a century of research? *Psychological Bulletin, 104*(2), 251-267.
Sugimoto, C.R., & Cronin, B. (2012). Bio-bibliometric profiling: An examination of multi-faceted approaches to scholarship. *Journal of the American Society for Information Science & Technology, 63*(3), 450-468.
Sugimoto, C.R., Russell, T.G., Meho, L.I., & Marchionini, G. (2008). MPACT and citation impact: Two sides of the same scholarly coin? *Library & Information Science Research, 30*(4), 273-281.
Varga, A.V. (2011). Measuring the semantic integrity of scientific fields: A method and a study of sociology, economics, and biophysics. *Scientometrics, 88*, 163-177.

CITATION IMPACT PREDICTION OF SCIENTIFIC PAPERS BASED ON FEATURES

Tian Yu[1], Guang Yu[2] and Qing-Hua Hu[3]

[1]*yutian.hit@gmail.com*
Harbin Institute of Technology, School of Management, Harbin (People's Republic of China)

[2]*yug@hit.edu.cn*
Harbin Institute of Technology, School of Management, Harbin (People's Republic of China)

[3]*huqinghua@tju.edu.cn*
Tianjin University, School of Computer Science and Technology, Tianjin (People's Republic of China)

Abstract
Researchers pay more attention to the scientific papers published in the last two years, especially the papers which could have a great citation impact in further. But currently citation impact prediction results are still not satisfied. This paper points out that objective features of a scientific paper could make predictions about the citation impact relatively accurately. External features of a paper, features of authors, features of published journal, and features of citations are all considered in constructing papers' feature space. And stepwise multiple regression analysis is used to choose appropriate features from the space and build the regression model for explaining the relationship between the citation impact and the features. The validity of this model is also experimentally verified in the subject of INFORMATION SCIENCE & LIBRARY SCIENCE. The results of this paper show that the regression model is effective in the subject.

Conference Topic
Scientometrics Indicators: Criticism and new developments (Topic 1)

Introduction
Scientific paper is the basic unit of analysis in scientometric research. Papers as knowledge carriers build on existing published papers, which would influence communication and progress in science. Citation Impact that is considered as a count of the number of citations is nowadays a widely used measure of scientific impact of a publication. Individual papers, journals, scientists, institutions, etc. have been evaluated or even ranked based on their citation impacts (*Hargens* and *Schuman*, 1990).

In the era of knowledge explosion, researchers can obtain a large number of papers in a given research subject conveniently. According to counting the number of papers from Web of Science, a researcher in subject of

 reading two papers daily would spend at least 100 years to finish. However, the reading time of individual researcher is scarce, which implies that a researcher does not want to waste time reading a paper of no significance. For those papers which have already been published more than five years, we can easily evaluate which paper has a greater citation impact by their citation count. But for the papers which only have been published one or two years, it is difficult to predict their future citation impacts. While the papers published within a short period of time usually cover the current hotspots and research trends, researchers would pay more attention to them to ensure the novelty of their study. Therefore, it is significant to predict the citation impacts of the papers published in the last two years.

The present studies show that a paper's citation impact could be influenced by the four main factors: the authors, the published journal, the research field and the quality of the paper itself.

Scientific papers are produced by researchers in scientific exploration, so authors' characteristics are indirectly reflected in the papers. There is some evidence that author reputation is the determinants of the allocation of citations (*Stewart*, 1983; *Bornmann* and *Daniel*, 2005; *Danell*, 2011).

It is considered that journals (and their editors) with good reputation can attract high-quality papers. Van Dalen and Henkens (1999; 2001) gave some evidence that papers published in core journals received considerably more citations than papers in second-tier journals, and the majority of papers in the second-tier journals remained uncited in the five years following their publication.

Garfield (1979) underlined that the research field must be taken into account in making comparisons between citation counts generated in different research fields, because the "citation potential" could vary significantly from one field to another. Boyack and Klavans (2011) pointed out the delineations among research fields were defined artificially and fuzzy. Researchers have tried to do field normalization with different methods or consider the non-parametric statistics instead of central tendency statistics to solve the problem (*Radicchi* et al 2008; *Radicchi* et al, 2012; *Moed*, 2010; *Leydesdorff* and *Bornmann*, 2011; *Leydesdorff* et al, 2012).

It is noted that the quality of a scientific paper is one of the most important factors for its citation impact. *Van Dalen* and *Henkens* (2005) stated that the quality of a paper could be approximated by the impact and speed with which knowledge is disseminated in the scientific community. Citations reveal the impact of a paper in the literature. And the speed with which a paper is disseminated in the scientific community is measured by the timing of the first citation.

Some statistical models based on the above features were established to predict future citation behaviors. *Glänzel* and *Schubert* (1995) presented a non-homogeneous birth-process model. *Burrell* (2001; 2003) presented the theory for a stochastic model for the citation process in the presence of obsolescence to predict the future citation pattern of individual papers. Recently *Wang* et al (2011; 2012) established a high-cited papers' prediction model with machine learning

tool. However, classification output is discrete and boundaries among classes are usually relatively fuzzy. Moreover, for the papers published in the last two years, the citation impact prediction results are still not satisfactory. Previous studies have shown that citation features of one paper in its first 5 years after publication is an important manifestation of its quality (*Glänzel* et al, 2003). Therefore we attempt to predict the citation impact in papers' first 5 years after publication by regression analysis which is introduced to perform more detailed classification prediction in our study.

In this paper, we analyze relevant features of scientific papers and seek to examine the relationship between the citation impact and the features by regression analysis, to predict the number of citations in papers' first 5 years after publication.

The feature space of scientific papers

Scientific paper can be described as a vector collection of multi-dimensional information which contains reference, author, research field, etc. In other words, they are multi-dimensional features of papers. The feature space X of scientific papers can be defined above:

$$X=\{x_1,x_2,x_3,\ldots,x_n\}$$

where $x_i(i=1,2,\ldots,n)$ is the feature of papers. And citation impact y of a scientific paper is defined as the total number of citations.

The features describing scientific papers are divided into four types: external features of a paper, features of authors, features of published journal, and features of citations. Some external features, such as the document type, the language, the published time, the number of references are used to describe the paper itself. According to the Matthew effect, the reputation of authors and published journal is likely to influence the total number of citations. There are several scientometric indicators such as total number of publications and citations, citations per journal paper, which may characterize the publications of scientists quantitatively. And a series of journal evaluation indicators from JCR and Eigenfactor™ metrics are used to characterize the quality and impacts of journals and their editorial board. Furthermore, features of citations in the period of the first 2 years after publication are used to describe the capacity of knowledge diffusion.

The features listed in Table 1 are finally extracted to describe the characteristics of scientific papers. They are simple indictors which are widely accepted and easily accessible. To make it more convenient for comparing, we only select the papers whose document type is article and which published in 2007. In addition, it is noted that the reciprocal of the first-cited age takes the place of the first-cited age in this study, because some papers have never been cited in Web of Science. We define the value of the first-cited age of these papers as positive infinity in order to facilitate comparison. So the reciprocal of the first-cited age could be in the range 0-1. The knowledge that one paper with high value of the reciprocal of the first-cited age contains should diffuse more rapidly.

Table 1 The features of scientific papers

Feature type	*Feature description*	*Label*
External features of a paper	the year when published (all were published in 2007)	
	The type (the document type of each selected paper is article)	
	The number of references	x_1
Features of authors	The number of authors	x_2
	The country of author's institution (text type features)	
	The *h* index of the first author before publication of this paper	x_3
	The number of papers published by the first author before this paper	x_4
	The total citations to the papers published by the first author before this paper	x_5
	The average citations to the paper published by the first author before this paper	x_6
	The maximum *h* index of the authors before publication of this paper	x_7
	The maximum number of papers published by the authors before this paper	x_8
	The maximum total citations to the papers published by the authors before this paper	x_9
	The maximum average citations to the paper published by the authors before this paper	x_{10}
Features of citations	The reciprocal of the first-cited age of this paper	x_{11}
	The total citations to this paper in its first 2 years after publication	x_{12}
	The number of countries citing this paper in its first 2 years after publication	x_{13}
	The number of kinds of papers citing this paper in its first 2 years after publication	x_{14}
	The number of journals citing this paper in its first 2 years after publication	x_{15}
	The number of subjects citing this paper in its first 2 years after publication	x_{16}
Features of published journal	The total citations to the journal	x_{17}
	The impact factor of the journal	x_{18}
	The 5-year impact factor of the journal	x_{19}
	The immediacy index of the journal	x_{20}
	The number of papers published in the journal in this year	x_{21}
	The cited half-life of the journal	x_{22}
	The Eigenfactor score of the journal	x_{23}
	The article influence score of the journal	x_{24}

Table 2 The list of 20 journals from INFORMATION SCIENCE & LIBRARY SCIENCE

Num.	*Abbreviated Journal Title*	*ISSN*
1	ASLIB PROC	0001-253X
2	COLL RES LIBR	0010-0870
3	GOV INFORM Q	0740-624X
4	INFORM MANAGE-AMSTER	0378-7206
5	INFORM PROCESS MANAG	0306-4573
6	INFORM RES	1368-1613
7	INFORM SOC	0197-2243
8	INFORM SYST J	1350-1917
9	INFORM SYST RES	1047-7047
10	INT J GEOGR INF SCI	1365-8816
11	INT J INFORM MANAGE	0268-4012
12	J ACAD LIBR	0099-1333
13	J AM MED INFORM ASSN	1067-5027
14	J AM SOC INF SCI TEC	1532-2882
15	J DOC	0022-0418
16	J HEALTH COMMUN	1081-0730
17	J INF SCI	0165-5515
18	J INF TECHNOL	0268-3962
19	J LIBR INF SCI	0961-0006
20	J MANAGE INFORM SYST	0742-1222

Method

Data preparation

The basis for our study is the data provided by Thomson ISI. The 2007 version of the JCR indexed 56 journals in the subject of INFORMATION SCIENCE & LIBRARY SCIENCE. In accordance with the list of JCR 2007 we selected the first 20 journals whose indicators were completed in JCR because of limited time for data collection (listed in Table 2). We used these ISI products with data covering to Jan 2012.

The papers we selected all published in 2007, so the features of published journal are obtained directly in the 2007 version of the JCR.

The web version of Web of Science provides an analysis tool ‘‘Analyze Results’’ for analyzing the characteristics of papers. Basing on this tool, we first extract the citations published in the first 2 years after publication from all citations, and then we could be able to analyze the features of citations in the period of the first 2 years after publication.

Features of authors could be identified in several steps. Firstly, we view “Distinct Author Record Sets” which is a discovery tool in Web of Science showing sets of papers likely written by the same person to get one author’s all publications. Secondly, we exclude the papers written by the other authors with the same name from all publications. It requires intensive labor activities because we have to

separate the papers of different authors with the same name and extract the papers of the desired author in accordance with the author's affiliation, address, email, and so on. Thirdly, we exclude the papers published in 2007-2012. The month when paper published is ignored here, which is convenient for the data statistics. Finally we calculate the features of the author before publication of the paper. In addition, we use the country of the first corresponding author as the country of author's institution, which could be statistically analyzed only because it is a text feature.

Consequently we have identified the features of 1,025 papers published in 20 journals from INFORMATION SCIENCE & LIBRARY SCIENCE by using ISI database. Data collection has been finished in January 2012.

Analysis method

The focus of this paper is on the relationship between the citation impact and the features shown in Table 1, so we adopt multiple regression analysis to learn the scoring function with the feature set.

Multiple regression analysis is an important branch of applied statistics. Gibbons firstly suggested the multivariate regression model methodology to measure the effect of new information on asset prices (*Gibbons*, 1982). It can not only extract the important information hidden in massive data sets, but also take advantage of variables to predict and control a certain variable (*Kleinbaum* et al, 1998). In regression analysis, an output variable is called the dependent variable, and the variables that influence the dependent variable are called independent variables. The dependent variable is changed in response to changes in the independent variables. Therefore, in our research, the number of citations is considered as the dependent variable and 24 features shown in Table 1 as independent variables in the regression analysis. And the SPSS 13.0 for Windows is used to conduct most of our calculations.

Logically it is necessary to prove that the features of scientific papers do influence the number of citations before examining the relationship between the citation impact and the features. Therefore we formulate four hypotheses for this research:

- H_1: The number of references could influence citation impact.
- H_2: Author reputation could influence citation impact.
- H_3: A ranking of published journal could influence citation impact.
- H_4: A paper's quality could influence its citation impact.

In actual data analysis both tests can be conducted in a single model of statistical analysis.

Results and discussion

Statistical analysis of features

We select 1,025 papers published in 20 journals in 2007, and the accumulated total number of citations to these papers is 7,232. Figure 1 shows that citations are skewed in distribution of all papers on the total number of citations *y*, suggesting

that most papers are cited only a few times. It conforms to the overall situation in the subject of INFORMATION SCIENCE & LIBRARY SCIENCE and implies the data we selected are valid.

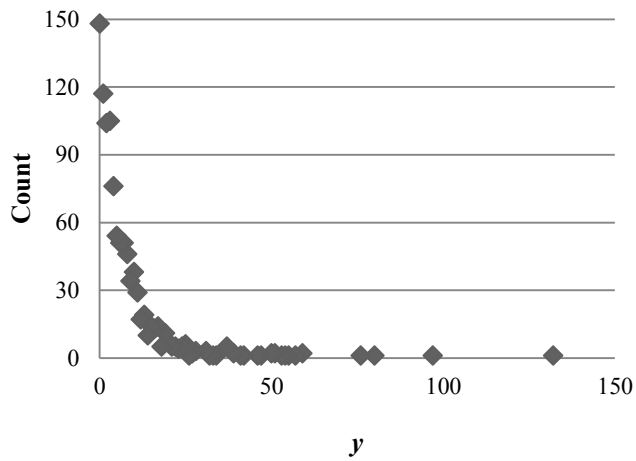


Figure 1 Distribution of 1,025 papers on the total number of citations y

Overall, most of the papers have 10-50 references. More than 80% of all papers have 1-3 authors. The authors of 452 papers, nearly 45% of all papers, come from American institutions. And the authors from England and Canada also published a great number of papers in 2007.

The features x_3, x_4, x_5 and x_6 reflect the reputation of first author. The value of x_3 for about 75% of all papers are not more than 2; the value of x_4 for about 70% are not more than 4; the value of x_5 for about 80% are lower than 60; the value of x_6 for about 70% are lower than 9. It implies that about half of all researchers in the field are new and their prestige is very low. Moreover, the features x_7, x_8, x_9, and x_{10} reflect the best reputation of co-authors. The value of x_7 for about 80% of all papers are not more than 7; the value of x_8 for about 60% are not more than 7; the value of x_9 for over 80% are lower than 1250; the value of x_{10} for about 80% are lower than 30.

In the data more than 50% of all papers were firstly cited in their first 2 years after publication, and about 75% were firstly cited in their first 3 years. Moreover, high-cited papers have strong capability to be cited in their first year after publication, and their impacts and speed of knowledge diffuse are good.

Stepwise multiple linear regression model on citation impact

Based on the features of 1,025 papers, the distributions on the features of scientific papers were obtained. The results indicate that we can use linear regression model, the most common model in regression analysis, to explore the relationship between the citation impact and the features.

Multiple linear regression analysis is used to estimate the parameters of the linear function based on given data. The regression model is trained from a set of known citation impacts of 1,025 papers which provide the determined values of these 24 features. We execute the multiple linear regression analysis with these features in our statistical software (SPSS).

Before executing multiple linear regression analysis, we should analyze the features first. The correlation coefficient matrix of 24 features reveals the high correlations between some features. It may cause the multicollinearity problem that all variables are introduced into a regression model. To solve this problem we apply stepwise regression analysis for choosing good variables from all variables to generate the predictor team. It could not only guarantee the validity and importance of the chose variables, but also reduce additional error caused by the redundant variables.

Table 3 shows that several chose independent variables are significant at the 0.05 level. The R, R-squared and the adjusted R-squared for this model are 0.822, 0.676 and 0.674 respectively, which means that the linear regression model can explain the relationship between the citation impact and the features. Results of a further ANOVA show the model is statistically significant. It is also shown that for the selected features all variance inflation factors are below 1.5 in Table 3. There is virtually no collinearity in this model. Furthermore, the residual plot and the normal PP plot of regression standardized residual indicate that the approach of this paper is computationally feasible.

Table 3 Regression coefficients of the model

Feature	*B*	*Sig.*	*VIF*
x_1	0.061	0.000**	1.081
x_5	0.002	0.012*	1.059
x_{10}	0.017	0.006**	1.121
x_{11}	-2.137	0.004**	1.464
x_{12}	3.470	0.000**	1.446
x_{19}	0.872	0.000**	1.121

*Significant at the 0.05 level. **Significant at the 0.01 level.

Therefore, taking the results of our regression, the regression equation for our analysis can be written as:

$$y = 0.061x_1+0.002x_5+0.017x_{10}–2.137x_{11}+3.470x_{12}+0.872x_{19}+e$$

This can be interpreted as, x_1, x_5, x_{10}, x_{12} and x_{19} have positive impacts on the research performance (citation impact); x_{11} has a negative impact on the research performance (citation impact). In this regression model, four feature types - external feature of paper, feature of authors, feature of published journal, and feature of citations - all contribute greatly to the citation impact prediction.

After establishing the functional relationship between the impact and the features, we use new data to examine the regression model's validity. The papers published

in *Scientometrics* (February 2008, Volume 74, Issue 2) from the subject of INFORMATION SCIENCE & LIBRARY SCIENCE are randomly selected as the test data set. The procedure is as follows: firstly ten papers' features are identified; then the predictive values of the test data could be obtained according to the regression function.

The number of citations in these papers' first 5 years after publication is determined, so we can compare the predictive values with the actual citation impacts. As shown in Figure 2, although there are some errors for the regression results, the accuracy of the predictive values reaches about 65%. It proves that the regression model is relatively effective. Therefore, papers' citation impact in their first 5 years after publication could be predicted by objectively assessed factors.

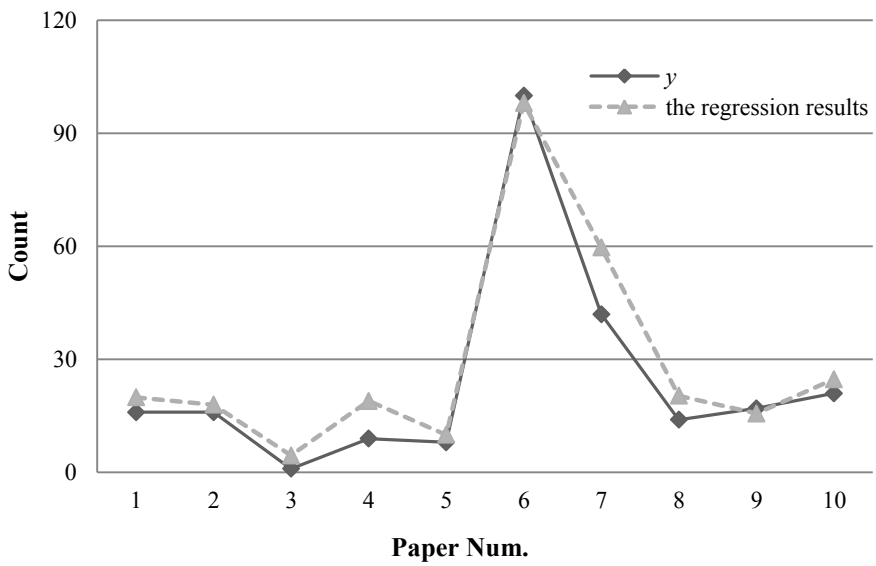


Figure 2 The comparison between the regression results and the actual total citations

In the above section, four hypotheses were proposed for the research. The results prove the validity of the hypotheses: each of these four types of features describing a scientific paper can indeed influence the citation impact. Similar to the findings of some previous studies (*Van Dalen* and *Henkens*, 2001; *Danell*, 2011), our results also indicate that author reputation and published journal's rank can influence a scientific paper's citation impact. It provides the necessary basis for us to carry out regression analyses. Moreover, it exceeds expectations that the number of references has some relationship with the paper's citation impact. This finding is obtained by running a regression model that treated citation impact as a dependent variable, and the result has been proven to be statistically significant.

In addition, the independent variables in our regression model include some citation features in papers' first 2 years after publication. That is, according to our research results, the citation impact of a paper can only be measured after it has

been published several years. So we establish a citation impact model using the features except features of citations by stepwise multiple linear regression analysis to predict the citation impact when a scientific paper has just been published. But the R, R-squared and adjusted R-squared for the obtained regression model are respectively 0.439, 0.193 and 0.177, which means the model cannot explain the relationship between the impact and the features except features of citations. We also tried to predict the citation impact by other regression techniques, but the obtained models fail to explain the relationship between them. Simonton's model of creative productivity considers that scientific creativity is "to some significant degree blind or haphazard" (*Simonton*, 1997). That means there is not a priori way to predict output. And the impact can be only evaluated retrospectively, after recognition has been achieved. Our result proves the validity of Simonton's model.

Explanation for the regression model

The regression model shows that for one scientific paper six features play the significant roles in affecting its citation impact: the number of references, the total citations to the papers published by the first author before this paper, the maximum average citations per paper published by the authors before this paper, the first-cited age of this paper, the total citations to this paper in its first 2 years after publication, and the 5-year impact factor of the journal. It suggests that all four types of features describing scientific papers are significantly correlated with citation impact. However, the strengths of the associations differ: features of citations have the strongest influence, followed by external features of a paper itself, features of published journal, and features of authors.

A paper's quality. The first-cited age and the total citations to a paper in its first 2 years after publication measure the speed with which knowledge is diffused in scientific community and a degree of acceptance by peers and other professionals respectively. Approximately the contents and quality of a paper could be measured by these features. In our regression model these two features are the most important factors associated with citation impact. However we fail to establish an effective model to predict citation impact using the features except features of citation. We confirm that the quality of scientific paper is one of the most significant factors to effect on citation impact.

A paper's external features. The number of references is an external feature to characterize scientific papers. We get this conclusion that the number of references has a significant influence on citation impact. This is probably a consequence of learning a lot of literatures. The more literatures a researcher reads, the deeply he understand the current situation and development trend of his research field. This is an effective method to enhance research capacity.

Journal reputation and *Author reputation*. Author and journal reputation are generally felt to play a role of some significance in gaining attention in science. Our regression model shows that journal reputation has higher influence on citation impact than author reputation. To some extent, this is due to the dominant

role of editors. Editors of core journals tend to have access to a number of high-quality manuscripts, and they perpetuate the status of core journals by publishing high-quality papers. Furthermore, although journal reputation and author reputation are correlated with citation counts, actually reading habits and citation motivations of researchers are significant factors to effect on citation impact. Merton's Matthew effect (1968) is applicable here: researchers are more willing to read and cite the papers written by famous authors or published in core journals.

Overall, our results suggest that characteristics of a scientific paper itself are very important factors to make it influential. Indeed, citation impact is a complex phenomenon involving many explicit and implicit social and scholarly factors. These six variables included in the model are the most apparent ones, yet we need to acknowledge the existence of other factors associated with citation impact.

Conclusion

In summary, our results suggest that a papers' citation impact could be predicted by objective scientometric indicators. External features of a paper, features of authors, features of published journal, and features of citations are all considered in constructing papers' feature space with the mathematical description method. Because the information provided by these features may be redundancy, the method of stepwise regression analysis is applied for choosing good variables from all features and building a model to describe the relationship between citation impact and the features. Because the citation potential can vary significantly between different fields, the papers published in the subject of INFORMATION SCIENCE & LIBRARY SCIENCE are selected only to avoid the error. And we can relatively accurately predict papers' citation impact in their first 5 years after publication in this subject.

Several important caveats should temper these conclusions. Most importantly, our research has obtained the interesting relationship between citation impact and some features. It means that these features are significant factors to indicate citation impact rather than cause it. Although we believe that a scientific paper obtains multidimensional complex features, in this study we only selected the features which are considered available and could be obtained in a relatively simple and fast manner. For instance, we did not consider the characteristics of the *citing* papers as determinants of citations. That may cause the omissions of some features. In terms of data acquisition, we obtained them in the ISI database. Some of the limitations of the ISI database itself such as incompleteness are bound to be brought into the study. However, it is undeniable that the ISI is the largest comprehensive academic information resource database in the world which covers the most subjects. It is the reason for selecting this database. In addition, the sample of scientific papers included in this analysis is quite limited and covers the papers published in one subject category. Therefore, our model may just apply in this subject category.

Even with these caveats, the findings of this study still reveal the interesting relationships between the citation impact and the features of scientific papers. And the feature space constructed by the selected features is effective for the description of scientific papers. We need to further consider the comprehensiveness and effectiveness of the features, involving many aspects of open access status of the paper itself, acceptable level of the audiences, etc. And the data needs to be larger and more comprehensive.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 70973031).

References

Bornmann, L. & Daniel, H.-D. (2005). Selection of research fellowship recipients by committee peer review. Analysis of reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics*, 63, 297-320

Boyack, K. W., & Klavans, R. (2011). Multiple dimensions of journal specificity: Why journals can't be assigned to disciplines. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *The 13th conference of the international society for scientometrics and informetrics* (Vol. I, pp. 123-133). Durban, South Africa: ISSI, Leiden University and the University of Zululand.

Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics*, 52, 3-12.

Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology*, 54(5), 372-378.

Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal of the American Society for Information Science and Technology*, 62(1), 50-60.

Garfield, E. (1979). *Citation indexing. Its theory and application in science, technology and humanities*. New York: Wiley.

Gibbons, M. R. (1982). Multivariate tests of financial models: A new approach. *Journal of Financial Economics*, 10(1), 3-27.

Glänzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing and Management*, 31(1), 69-80.

Glänzel, W., Schlemmer, B. & Thijs, B. (2003). Better later than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, 58(3), 571-586.

Hargens, L.L. & Schuman, H. (1990). Citation counts and social comparisons: Scientists' use and evaluation of citation index data. *Social Science Research*, 19(3), 205-221.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E. & Nizam, A. (1998). *Applied regression analysis and other multivariable methods*. Brooks/Cole Publishing Company, Pacific Grove.

Leydesdorff, L., & Bornmann, L. (2011). Integrated impact indicators (*I3*) compared with impact factors (*IF*s): An alternative design with policy implications. *Journal of the American Society for Information Science and Technology*, 62(7), 1370-1381.
Leydesdorff, L. (2012). Alternatives to the Journal Impact Factor: *I3* and the Top-10% (or Top-25%?) of the Most-Highly Cited Papers. *Scientometrics*, 92(2), 355-365.
Merton, R. K. (1968). The Matthew effect in science. *Science*, 159, 56-63.
Moed, HF (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3), 265-277.
Radicchi, F., Fortunato, S. & Castellano, C. (2008). Universality of citation distributions: toward an objective measure of scientific impact. *PNAS*, 105(45), 17268-17272.
Radicchi F & Castellano C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics*, 6(1), 121-130.
Simonton, D.K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review*, 104(1), 66-89.
Stewart, J.A. (1983). Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces*, 62(1), 166-189.
Van Dalen, H. P., & Henkens, K. (1999), How influential are demography journals? *Population and Development Review*, 25(2), 229-251.
Van Dalen, H. P., & Henkens, K. (2001). What makes a scientific article influential? The case of demographers. *Scientometrics*, 50(3), 455-482.
Van Dalen, H. P., & Henkens, K. (2005). Signals in science-on the importance of signaling in gaining attention in science. *Scientometrics*, 64(2), 209-233.
Wang, M.Y., Yu, G., & Yu, D.R. (2011). Mining typical features for highly cited papers. *Scientometrics*, 87(3) 695-706.
Wang, M.Y., Yu, G., An, S., & Yu, D.R. (2012). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. *Scientometrics*, 93(3), 635-644.

CITATION IMPACTS REVISITED: HOW NOVEL IMPACT MEASURES REFLECT INTERDISCIPLINARITY AND STRUCTURAL CHANGE AT THE LOCAL AND GLOBAL LEVEL

.

Michel Zitt[1] and Jean-Philippe Cointet[2]

1 Michel.Zitt@nantes.inra.fr
INRA-Lereco, Nantes, France

2 Jean-Philippe.Cointet@polytechnique.edu
INRA-SenS, INRA, Marne-la-Vallée, France
3 Complex Systems Institute of Paris Ile-de-France (ISC-PIF), Paris, France

Abstract

Citation networks have fed numerous works in scientific evaluation, science mapping (and more recently large-scale network studies) for decades. The variety of citation behavior across scientific fields is both a research topic in sociology of science, and a problem in scientific evaluation. Normalization, tantamount to a particular weighting of links in the citation network, is necessary for allowing across-field comparisons of citation scores and interdisciplinary studies. In addition to classical normalization which drastically reduces all variability factors altogether, two tracks of research have emerged in the recent years. One is the revival of iterative "influence measures". The second is the "citing-side" normalization, whose only purpose is to control for the main factor of variability, the inequality in citing propensity, letting other aspects play: knowledge export/imports and growth. When all variables are defined at the same field-level, two propositions are established: (a) the gross impact measure identifies with the product of relative growth rate, gross balance of citation exchanges, and relative number of references (b) the normalized impact identifies with the product of relative growth rate and normalized balance. At the science level, the variance of growth rate over domains is a proxy for change in the system, and the variance of balance a measure of inter-disciplinary dependences. This opens a new perspective, where the resulting variance of normalized impact, and a related measure, the sum of these variances proposed as a Change-Exchange Indicator, summarize important aspects of science structure and dynamism. Results based on a decade's data are discussed. The behavior of normalized impact according to scale changes is also briefly discussed. A shift towards a network-based definition of domains, more in the nomenclature-free spirit of citing-side normalization than database classification schemes, appears promising, albeit with technical challenges. An appealing issue is the connection with macro-level life-cycles of domains, and the dynamics of citation network.

Conference topics

Topic 1 - Scientometrics Indicators: - Criticism and new developments

Topic 11 - Modeling the Science System, Science Dynamics and Complex System Science

Introduction

The use of citation measures in science (Garfield, 1955, 2006) is a controversial issue in research evaluation, as shown in the recurrent debates on impact factors[31]. Citations also shape a large-scale network (Price, 1965, Redner, 2005) which, along with collaboration, linguistic and web-communication networks, is a powerful tool for mapping science and understanding knowledge exchanges and self-organization of communities. A lasting issue is the variability of citation practices across fields, which prevents any sensible comparison between gross citation figures or h-indexes, say in mathematics vs. cell biology. A traditional way to deal with this variability is the normalization of citation figures based on fields baseline figures (Murugesan & Moravcsik 1976, Schubert & Braun 1986, Sen, 1992,. Czapski 1997, Vinkler, 2002, see also Raddichi F.et al., 2008). This "*ex post*" or "cited-side" statistical normalization is typically nomenclature-dependent, assuming an explicit delineation of scientific domains, usually from databases classification schemes. Forcing equality of cited domains, it sacrifices the consistency of the network and jeopardises multidisciplinary analysis. An alternative is the citing-side normalization ("*ex ante*", "source-level", "fractional citation"; Zitt & Small 2008, Moed 2010, Glänzel 2011, Leydesdorff & Opthof 2010, Waltman 2012. The citing-side perspective (Zitt et al., 2005) is at the confluence of Garfield's insights on citation density (Garfield, 1979) and fractional weighting to reduce biases in cocitation mapping (Small & Sweeney, 1985). It corrects for the unequal propensity to cite amongst domains: in doing so, it keeps the best part of normalization – by removing undesirable sources of across-fields variability – while keeping the coherence of the citation network.

Especially, the partial normalization brought by the citing-side process, gives interpretable figures of domain-level average impact, which is true neither for usual "cited-side" normalized figures, forced to equality, nor for gross citation figures, blurred in magnitude by the effects of differential propensity to cite amongst fields. Focusing here on the analysis at the aggregate levels, we argue that citing-side approach opens new perspectives on interpretation of citation impacts at the domain level, and on structure and change of science insofar as it can be depicted by citation networks. We shall first establish two basic propositions on the decomposition of gross impacts and citing-side normalized impacts at the domain level. For the latter, we propose to summarize into a "Change-Exchange Index" the variances over domains of its two factors at the domain level, growth rate and dependence. It may seem strange at first to come

[31] in recent literature, see the dedicated issue of *Scientometrics 92(2), (2012)*

across a time-dependent variable such as growth, but the diachronic nature of citations implicitly carries information on change.
The CEI identifies with the variance of normalized impact when the two factors are independent, making the covariance term zero. We shall examine, on a decade's data on citation flows across science and a fixed nomenclature of domains, the empirical value of the variance and covariance terms calling for interpretations in terms of dynamics of the system, and discuss the challenges of shifting the nomenclature-based analysis to a bibliometric classification into topic/domains, more in line with the nomenclature-free spirit of citing-side approach.
In section I, analytical, we shall state two propositions at the domain level: one on gross impacts,one on citing-side normalized impacts. Then considering the science level, we will define the CEI and its relation with the normalized impact. Section II summarizes first results from an on-going empirical analysis on a decade of the Web of Science. The discussion section discusses several aspects: the shift from a database classification scheme (nomenclature) framework to a bibliometric classification of science; the relationship with various aspects of dynamics of science.

I – Analytical bases

If all variables are calculated at the same level of classification (whatever the level: for example the subject category) we get two basic propositions.

proposition 1: domain level, gross impact (not normalized)

The impact $I(A)$ of a domain A is defined as the average number of incoming citations per articles susceptible to get cited in A. If $\phi_{\leftarrow}(A)$ denotes the aggregated number of references citing domain A then : $I(A) = \frac{\phi_{\leftarrow}(A)}{|A|}$.
The growth rate $\rho(A)$ of a domain A is simply defined here by the ratio of publication volumes between the cited and the citing periods, volumes reduced to average volume over each period. We then introduce the balance which compares the total inflow of citation with total outflow emitted by A ($B(A) = \frac{\phi_{\leftarrow}(A)}{\phi_{\rightarrow}(A)}$).
Finally we denote $\kappa(A)$ the average number of references in citing articles in A.
It is then straightforward to deduce the following equation (seen Appendix for further details):

$$I(A) = \rho(A)B(A)\kappa(A)$$

Equation 1

From this equation, it is useful to introduce the notation ^ transforming any domain level index into its relative version normalized with its science level counterpart. Given any domain-level measure $m(A)$ one can compute $\hat{m}(A) = \frac{m(A)}{m(S)}$. Thus the relative impact $\hat{I}(A)$ is obtained by dividing the gross impact by the impact computed at the whole science level ($I(S)$). We will also denote $\hat{\rho}(A)$ the relative growth rate (*i.e.* growth rate normalized by the growth rate at the

global science level) and $\hat{\kappa}(A)$ the relative number of references in citing articles in A.

$$\hat{I}(A) = \hat{\rho}(A)B(A)\hat{\kappa}(A)$$

Equation 2

Proof is given in Appendix.

proposition 2 - domain level: citing-side normalized impact

In order to neutralize the main source of variability, a normalization based on the relative number of active references (the "citing propensity") is introduced. It is implemented by weighting the links of the original directed and unweighted citation network, with options fixing the granularity of the baseline. In a simple device, cited-side normalization weighs links proportionally to average in-links by node within the citable set in a given domain's delineation while citing-side normalization weighs links proportionally to average out-links by node within the citing set in the domain. Those domains can be defined by some neighbourhood of the citing article: journal, cluster, or librarians/database categories. Here, for establishing basic propositions, we shall rely on subject categories as defined by Web of Science (Thomson Reuters). With such a weighting of the citation links it naturally appears that $\hat{\kappa}(A) = 1$. Neutralizing citing propensity variability then defines a new measure of impact which can be decomposed as:

$$\hat{I}_g(A) = \hat{\rho}(A)B_g(A)$$

Equation 3

Proof is given in Appendix. These propositions generalize previous results on the journal impact factor (Zitt, 2011).

proposition [3] - science level: the deviation of citation impacts.

If the domain-level normalized impact is the product of two relative measures linked to interdisciplinary structure (asymmetry of exchange) and local dynamism (relative growth), what can we learn at the science level? All measures being relative, the signs of change are expected in the deviation indexes. We shall limit ourselves to the variances (on the log-transformed variables), in spite of imperfections, but concentration indexes such as the Gini mean difference (Yitzakhi, 2003) with larger scope of application could be envisioned.

For a particular category A at a given level of breakdown $\hat{I}_g(A) = \hat{\rho}(A)B_g(A)$. With logarithmic transformation of variables, suggested by the distribution of impacts at the domain level:

$LI(A) = LG(A) + LB(A)$ where LI, LG, LB designate respective logs of normalized impact, growth rate and normalized balance. Over all domains:

$\mathrm{wVar}(LI) = \mathrm{wVar}(LG) + \mathrm{wVar}(LB) + 2\mathrm{wCov}(LG, LB)$

Equation 4

where wVar stands for variance weighted by the volume of publications of domains, expressed in number of publications. For comparison sake, the unweighted variance has also been used.

In Equation 4 the variance terms have a simple interpretation. $\mathrm{wVar}(LB)$ over domains is a proxy of global interdisciplinary dependences in the system, and $\mathrm{wVar}(LG)$ is a proxy for the intensity of "creative destruction" through differentiation of growth rates over domains. A scientific system where domains do not exchange and are in steady state will associate zero variance and covariance terms, giving a zero variance of impacts. At the opposite end, a scientific system combining a high proportion of growing and declining domains and a strongly asymmetrical balance of flows across fields (exporters and importers) will show a high level of variance terms, but the final value of $\mathrm{wVar}(LI)$ will also depend on the covariance term.

The relationship between growth and balance partly depend on the superposition of domains at various stages of their life-cycle, while the potential value of balance for individual domains, typically reached at maturity stages, can show great dispersion linked to the position of the domain in the cognitive chain. The variance of balance (compared to growth's) may play a dominant role in the shaping of impact dispersion . Domains in emergence both grow rapidly and are quite dependent on imports of knowledge/ information from their parent fields. Hence they are likely to enhance the variance of growth, and to yield negative covariance

In order to summarize asymmetry and growth effects, we propose then to consider only the sum of variance terms, the "structural-change and exchange-asymmetry index", abridged into Change-Exchange Index CEI

$$CEI = \mathrm{wVar}(LG) + \mathrm{wVar}(LB)$$

Equation 5

This index is closely related to the variance of impacts with $CEI = \mathrm{wVar}(LI) - 2\mathrm{wCov}(LG, LB)$

CEI is trivially equal to the variance of impacts if growth rate and balance are independent.

scale issues

If the level of calculation of impact and the level of normalization (at which balances and growth rate are computed) are different, factors of scale come into play. Let us for example compare the normalized impact of sub-disciplines obtained (a) by normalization on the same-level, sub-discipline (b) by normalization at inferior level, the subject category. The growth factor for (a) is the weighted mean of growth factor for the corresponding categories.For the balance factor (b), a correcting coefficient depending on the structure of exchanges is needed, since the global balance of say a discipline is obviously not the average of the balances of the component categories. As far as gross impacts are concerned: the impact, the growth factor and the relative length of

bibliographies are stable in aggregations with appropriate weighting by the volume of publications, whereas gross balances are not. Scale irregularities in standard (cited-side) normalization had also already been stressed by Zitt, Ramanana, Bassecoulard (2005).
In such configurations where the level of definition of impacts and of other variables are not homogeneous (which is the case in many practical uses of normalization), the relations above should be altered by a correcting factor for the balance.

II – A first experiment within a fixed nomenclature

Data are based on OST aggregate figures at the category level, based on primary data and subject categories from the Web of Science (Thomson Reuters).
The citation framework is based on "cited years", on the period 1999-2010, giving an exploitable span 1999-2006 and with caution through 2008 (with reduced but acceptable citation window). In the database (OST-WoS), there are overlaps in assignment of journals and then papers to categories (WoS subject categories) at the lower level, handled by fractional counting. The nomenclature at the sub-discipline and the discipline level is derived from OST scheme, modified for simplicity sake, in order to get an embedded scheme:
specialty (subject category) $\subset$ sub-discipline $\subset$ discipline.
The nomenclature covers all sciences including social sciences and humanities. Specialties with very low citations activity, most of them belonging to humanities where the interpretation of citations is problematic, were discarded[32].
Let us summarize the main results.
Gross impact: As expected the gross impact heavily depends on the propensity to cite. The variance of the impact is essentially shaped by the variance of this factor, which jeopardizes any interpretation of its variance in terms of balance, the variance of which is by an order of magnitude lower, and growth, still behind.
Normalized impact: As soon as citing propensity is corrected, a new avenue is open to interpretation of citation impacts, in terms of dynamism and asymmetrical interdisciplinarity. Fig.***1A,B,C shows the time series of variance (weighted variance) of normalized impact, of its factors growth and balance, of the covariance term, and the series of CEI. A couple of striking points:
- the respective role of the two factors: within this citation window, the influence of growth is small, the CEI is mostly shaped by the asymmetry of exchanges. However, the dominant role of balance increases with the level of aggregation increases. In average over years, the ratio is about 8.2 at the category level, 10.6 at the sub-discipline level, 14.1 at the discipline level. With respect to the reduction of all variances in the aggregation process, it appears that the smoothing effect is stronger for differential growth than for balances. This is not unexpected: the status of exporter of knowledge in generic domains (fundamental biology) tends

[32] The resulting selection is given in Appendix II of the forthcoming Arxiv document corresponding to this submission.

to persist over levels. Conversely, the domains in medical research tend to remain importers of knowledge whatever the level of aggregation, from specialties through "medical research" as a large discipline. This depends on the properties of the network at various scales. We limited ourselves to the three levels mentioned, but a step further if we were to consider "life sciences" as an ensemble, the balance would largely collapse. In terms of trend, both the dispersion of balance and growth slightly decline, except at the discipline level with a quasi-stability of growth between 2001 and 2006.

- the covariance of growth and balance: covariance is almost always negative over the period at the category and sub-discipline level. The negative covariance is still higher in absolute value in the non weighted option: domains' size (volume of publications) does not matter and then small domains gain relative importance, among them emerging ones. There is a clear trend over the period, an increase of covariance which seems to get closer to the zero value, remembering however that the last two years are not strictly comparable to the rest of the series. This trend is watched whatever the level of aggregation.
- the variance of impacts, in terms of trend, is less affected than the CEI by the down-trend, since the increase of covariance (from fairly negative to weakly negative, not to mention the last two years with incomplete information) compensates for the reduction of variances of growth and balance.

A correlation analysis was also conducted, in the same line. To conclude, as far as an analysis based on a fixed nomenclature can be trusted, there is no sign of differentiation of growth rates over the period, nor of increasing asymmetry in the system, whatever the scale.

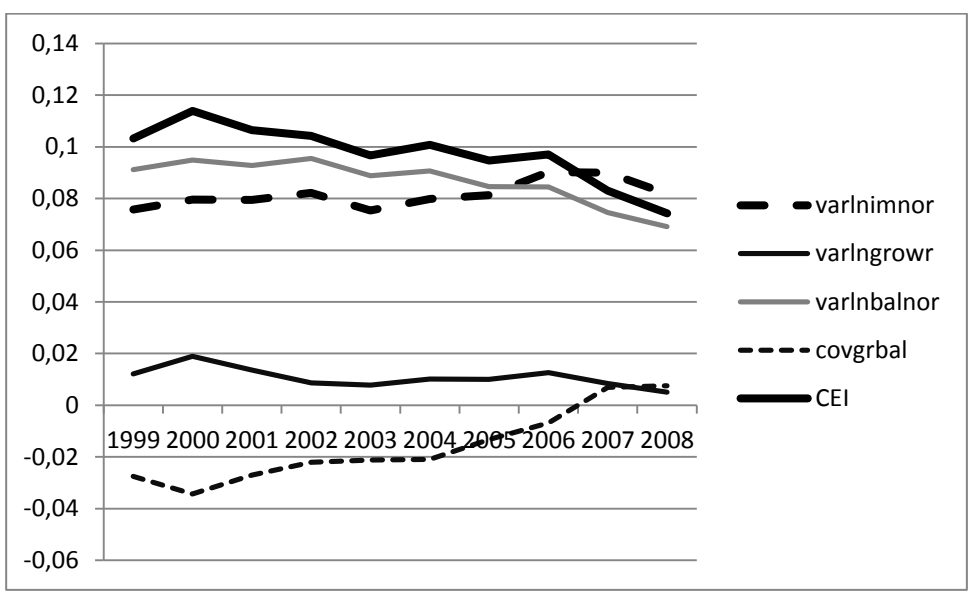


Fig. 1A - category level

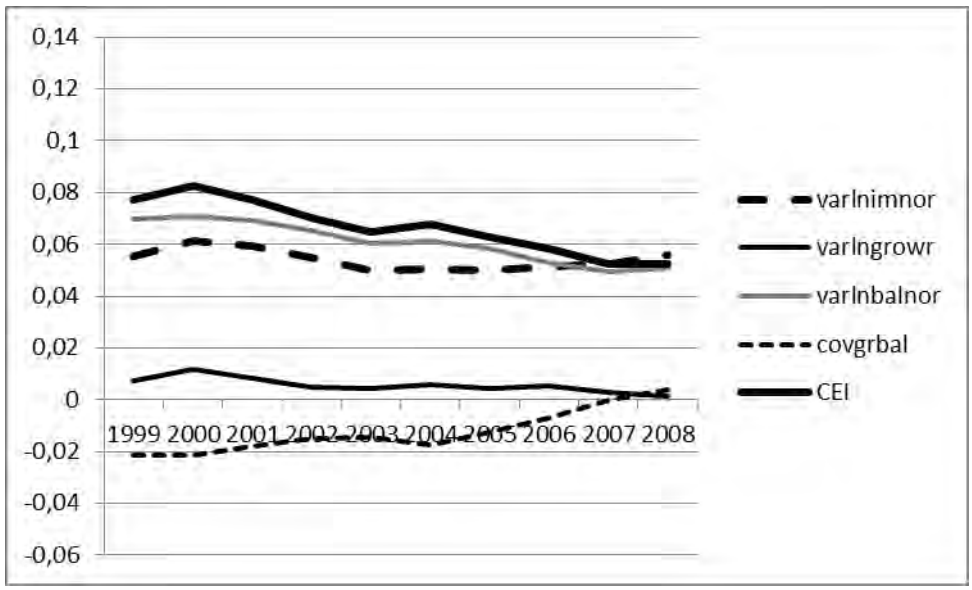


Fig. 1B – sub-discipline level

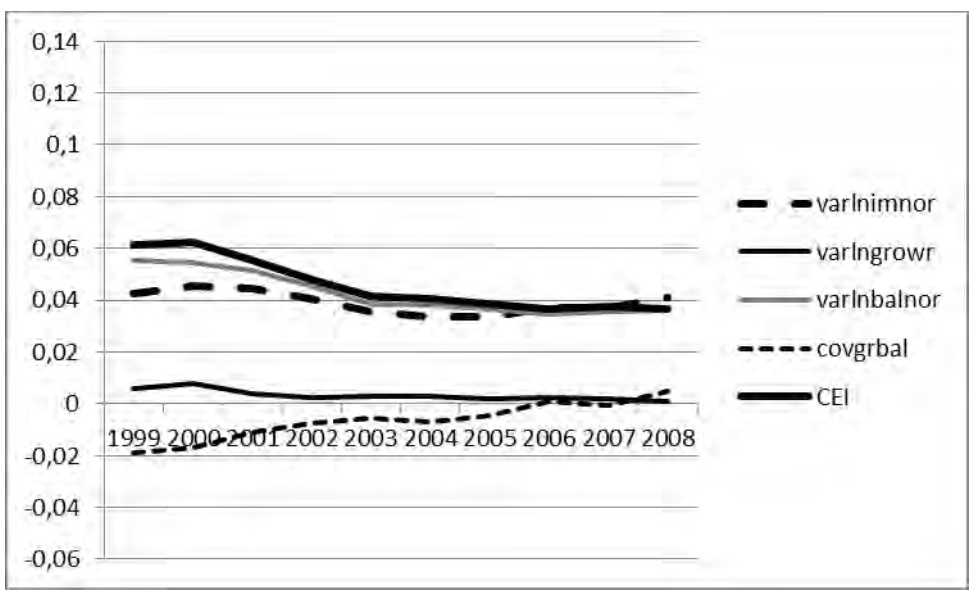


Fig. 1C – discipline level

Fig. 1 – time series of normalized impact and CEI – variances and covariance

varlnimnor: variance of logarithm of normalized impact
varlngrowr: variance of logarithm of relative growth
varlnbalnor: variance of logarithm of citation balance
covgrbal: covariance (logs)growth—balance
CEI : Change-Exchange Index

III discussion

We emphasized that citation impacts corrected for citation propensity identify, at a given domain level, with the product of growth and export-import balance; over all these domains (science), the variance of those factors are markers of differential growth and of interdisciplinarity in a particular form, the citation dependence. These properties suggest that the variance of impacts one the one hand, and a related index, the Change-Exchange Index, on the other hand, help to partly characterise structural dynamism of the scientific system at a particular level of aggregation. Without scale correcting factor, those measures hold iff all variables involved are defined at the same level of nomenclature (identical level for the calculation of impact and the normalization). practical applications of citing-side normalization suppose that a unique level is chosen for averaging the bibliographies lengths (propensity to cite), creating a specific weighted citation network. ,Normalized impacts can then be calculated for various levels of aggregation. Therefore, the constraint for establishing the propositions [1], [2] and [3] (the equality of levels) is not satisfied. Interpretation in terms of growth and balance would need the scale correcting factor mentioned above.

From nomenclature to clusters

The experiment was conducted on a fixed nomenclature. This is clearly a limitation. Nomenclatures such as databases classification schemes suffer shortcomings: artefacts in the delineation of categories, low reactivity in the short term, sensitivity to national context. These schemes are conservative and let tensions accumulate in the system between two revisions: they may for example keep two sub-domains attached, that a data analysis based on bibliometric networks could consider as having parted, and conversely for merges. An alternative is to rely on those networks, especially citations, with various transformations (Kessler 1963, Small 1973, Marshakova 1973, Chen 1999, Boyack, Klavans & Börner, 2005), which proved powerful tools for clustering and mapping science. Clustering reduces tensions by adjusting delineation of domain and trading topics. Substituting clusters/ neighborhoods to categories is therefore expected to yield more realistic representations, minimizing artificial exchange flows. Bibliometric clusters (co-authorship, citations, semantic content) enable scholars to track emergence and life-cycle phenomena (Scharnhorst et al., 2012, Morris, 2005, Chavalarias et al. 2013). If citation approach is preferred, which is logical in our context, a sensible objective is to reach bipartite clusters encompassing both cited and citing items in close relation. The choice of symmetrical metrics (citing<-->cited) seems preferable for building clusters, avoiding mixing areas with asymmetrical positions in the flows of knowledge.

A challenge of network-based clustering is the loss of coverage: in nomenclature schemes or classification based on editorial entities (journals), any citable article is classified, whether cited or not; any citing article is classified, whether its references are "active" (falling into the citation window) or not. For example, the exercise of "audience factor" (Zitt & Small, 2008) based on entire journals on

both sides, escapes this integrity issue, whereas finer granularity exercises (SNIP, Moed, 2010) have to cope with it. There are various ways – easier for citing than for citable articles – to circumvent the problem, by relaxing the citation window, modifying the construction of neighbourhood, introducing a correcting factor (SNIP 2, Waltman & van Eck, 2012).
The application of bibliometric clustering to the questions addressed here is promising. It would be appealing to confirm the slant suggested by the empirical findings,a relative down-trend in differentiation of growth rates and asymmetry of domains' balances.
It should be stressed that unlike the conventional cited-side normalization, the citing-side approach does not aim at a complete normalization. Usualquality tests assessing the performance of the various methods on the ground of the total reduction of variability are inappropriate in the present context. By limiting itself to the correction of propensity to cite, the citing-side approach reveals fruitful. We have focused in this paper especially on these uncontrolled factors.

Normalized Impact, Change-Exchange Index and dynamics of science

Further research is needed to explore the various aspects of these measures. A first is the effect of the citation window's length. A more general issue is the linkage between macro and micro-models. The relationship between growth and balance along the typical life cycles of scientific domains is appealing. We gave empirical evidence that growth rate and dependence are negatively correlated. The equilibrium between values of growth and balance variance on the one hand– with respect to citation windows - their covariance on the other hand are linked to features of local structure and dynamics. The sign of covariance, all things equal, may change over different phases in a domain's life-cycle. Typically negative in emergence phases, it may become positive in the central phases, especially if endogenous growth is echoed by external diffusion in the network overcoming domains' borders. A challenge is to connect model of life-cycle of areas, preferably delineated by citation-based clustering, with various mechanisms of networks dynamics (Powell et al. 2005), among them preferential attachment (Price 1963, Jeong et al., 2007, Eom et al., 2011).
The present approach only addresses aggregate phenomena. Balances at the domain level express a particular aspect of inter-disciplinarity, the asymmetrical linkages: domains equalizing exports and imports of knowledge will tend to reduce the dispersion. Many dimensions of citation networks are not accounted for. Diversity, essential for understanding the science structure and dynamics, is not directly accounted for. It should also be stressed that only relative changes were addressed here, through relative variables. The absolute growth or the average impact over science is corrected for, in contrast with long-range analyses in the wake of Price (1963) which focus on volumes of publications and citations (see for example Larivière et al., 2008 in their study of aging).
To conclude, citing-side approach opens a new perspective for the analysis of knowledge flows, insofar as they can be sketched by citation networks. It is a

promising solution for addressing diversity and interdisciplinary studies (Zitt, 2011, Rafols et al., 2012), with a significant improvement over gross flows analysis (e.g. Rinia et al., 2002). Here, at a macro-level, we have shown that basic relations connect the novel normalized impact and a derived measure, the CEI, connected to important features of dynamism and structure of science. The relation with the parallel and powerful "influence weighting" pioneered by Narin & Pinski (1976) with iterative weighting of citation sources, that has known a revival in the last decade (Palacio-Huerta & Volij, 2004, Bergstrom, 2007) is also appealing.

Acknowledgements

The authors thank OST (S. Ramanana and G. Filliatreau) for help in providing aggregated data and bases of nomenclature. Modifications are the responsibility of the authors. They also thank E. Bassecoulard for her assistance.

Bibliography

Bergstrom, C. (2007). Eigenfactor: measuring the value and prestige of scholarly journals. *College & Research Libraries News, 68*, 5, www.ala.org/ala/acrl/acrlpubs/crlnews/backissues2007/may2007/eigenfactor.cfm.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351-374.

Chavalarias D, Cointet J-P (2013). Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields. PLoS ONE 8(2): e54847. doi:10.1371/journal.pone.0054847

Chen, C. M. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management, 35*(3), 401-420.

Cronin, B. (1984). *The citation process; the role and significance of citations in scientific communication*. London: Taylor Graham.

Czapski, G. (1997). The use of deciles of the citation impact to evaluate different fields of research in Israel. *Scientometrics,* 40, 3, 437-443.

Garfield, E. (1955) Citation Indexes for Science. A new dimension in documentation through association of ideas. *Science*, 122, 108-111.

Garfield, E. (1979). Citation Indexing. Its theory and applications in science, technology and humanities. New York: Wiley.

Garfield E. (2006). The history and meaning of the journal impact factor. *JAMA* 295: 90 – 93,

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011). A priori vs. a posteriori normalization of citation indicators. The case of journal ranking. *Scientometrics,* 87, 2, 415-424

Jeong, H, Néda, Z, & Barabasi, A-L. (2007). Measuring preferential attachment in evolving networks. *EPL (Euro-Physics Letters)*, 61(4), 567.

Kessler M.M. (1963). Bibliographic coupling between scientific papers, *American Documentation,* 14, 1, 10-25.

Larivière V., Archambault E., Gingras Y. (2008) Long-Term Variations in the Aging of Scientific Literature: From Exponential Growth to Steady-State Science (1900–2004), *Journal of the American Society for Information Science,* 59(2): 288-296.
Leydesdorff, L., & Opthof, T. (2010). Scopus's Source Normalized Impact per Paper (SNIP) versus a Journal Impact Factor based on Fractional Counting of Citations. *Journal of the American Society for Information Science & Technology* 61, 11, 2365-2396.
Marshakova, I. V. (1973). Document coupling system based on references taken from ScienceCitation Index (in Russian). *Nauchno-TeknicheskayaInformatsiya, 2* (6.3).
Merton R.K.,(1968). The Matthew effect in science, *Science* 159 (3810) 56-63
Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*, 3, 265-277
Morris S.A.(2005). Manifestation of emerging specialties in journal literature: a growth model of papers, references, exemplars, bibliographic coupling, co-citation, and clustering coefficient distribution, *Journal of the American Society for Information Science* 56, 12, 1250-1273
Murugesan, P., & Moravcsik, M. J. (1978). Variation of the nature of citation measures with journal and scientific specialties. *Journal of the American Society for Information Science,* 29, 3, 141-155.
Palacio-Huerta, I., & Volij, O. (2004). The Measurement of Intellectual Influence. *Econometrica, 72*, 3, 963-977.
Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics. *Information Processing and Management,* 12, 297-312.
Powell W.W., White D.R., Koput K.W., Owen-Smith J. (2005)Network Dynamics and Field Evolution: The growth of interorganizational collaboration in the Life Sciences, *American Journal of Sociology* 110, 4, 1132–1205
Price D.J. de Solla (1963). Little science, big science, Columbia Univ. Press
Price D.J. de Solla (1965). Networks of scientific papers, *Science*, 149, 3683,510-515
Raddichi F., Fortunato S., Castellano C. (2008). Universality of citation distributions: Towards an objective measure of citation impact, *PNAS*, 105, 45, 17268-17272
Rafols I, Leydesdorff L., O'Hare A., Nightingale P., Stirling A. (2012) How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management, Research Policy 41 1262– 1282
Rinia E.D., Van Leewen T.N., Bruins E.W., van Vuren H.G., van Raan A.F.J., (2002) Measuring knowledge transfer between fields of science, *Scientometrics* 54, 3, 347-362

Scharnhorst A., Börner K., van den Besselaar P., (eds., 2012). Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences (Understanding Complex Systems), Springer
Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics, 9*, 5-6, 281-291.
Sen, B. K. (1992). Normalized Impact Factor. Journal of Documentation, 48(3), 318-325.
Small, H., & Sweeney, E. (1985). Clustering the Science Citation Index using co-citations I - a comparison of methods. *Scientometrics, 7*, 3-6, 391-409.
Small, H. (1973). Co-citation in the scientific literature : A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265-269.
Small, H., & Sweeney, E. (1985). Clustering the Science Citation Index using co-citations I – a comparison of methods. Scientometrics, 7 (3-6), 391-409
Vinkler, P., (2002). Subfield problems in applying the Garfield (Impact) Factors in practice, *Scientometrics*, 53, 2, 267-279
Waltman L., Van Eck N.J., (2012). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison, *arxiv.org/pdf/1208.6122*
Yitzakhi S.(2003). Gini's Mean difference: a superior measure of variability for non-normal distribution, *Metron - International Journal of Statistics,* 61, 2, 285-316
Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics, 63*(2), 373-401.
Zitt, M., & Small, H. (2008). Modifying the Journal Impact Factor by Fractional Citation Weighting: the Audience Factor. *Journal of the American Society for Information Science and Technology 59*(11), 1856-1860.
Zitt, M. (2010). Citing-side normalisation of journal impact: A robust variant of the Audience Factor. *Journal of Informetrics, 4*(3), 392-406
Zitt M. (2011). Behind citing-side normalisation of citations: Some properties of the journal impact factor, *Scientometrics* 89, 1, 329-344

APPENDIX

The basic propositions are established by considering that the domains are used as the basis of definition for all variables involved: relative growth; balance of citation exchanges; relative number of references; and the resulting relative impact. "Relative" values are understood as the ratio to the value for all science. The basic equations hold for any sensible level of granularity where it makes sense to align citing and citable literature. For convenience, the empirical illustrations are based three levels (subject category; sub-discipline; discipline) in a fixed nomenclature, albeit the spirit of citing-side normalization is nomenclature-free, in contrast with classical normalization.

For a given domain A of science S ($A \subset S$), we distinguish the set of articles pertaining to A according to their publication period T_0 and T_1: respectively A^{T_0} and A^{T_1}. In the empirical illustration, period T_0 is reduced to one "cited year"; period T_1 a set of "citing years" defined by the citation window (and in our data, containing the cited year T_0). We can define the matrix C which summarizes every citations pertaining to , e.g. every citation links from articles in or pointing to articles in . Obviously C is binary and asymmetric: for two articles i and j, $C(i,j) = 1$ iff i cites j. For the sake of clarity, we assume that only articles written during $T_0$are cited and that these citations are emitted by articles produced during period T_1. It can also be useful to interpret the citation matrix as a directed bipartite graph $G = (S, C)$ featuring a set of articles $i \in S$ tagged according to their domain of science and organized in two sets according to their publication date connected by the set of citations links: C. The total number of citations received by a publication is then simply given by its in-degree in G.

Those citations can be aggregated at the domain level: incoming and out-going citations at the domain level will be respectively denoted: $\phi_{\leftarrow}(A) = \sum_{i \in S, j \in A} C(i,j)$ and : $\phi_{\rightarrow}(A) = \sum_{i \in A, j \in S} C(i,j)$. Those flows can then be detailed according to the origin or the target domain of citations according to the scheme figure 1.

We define the growth rate $\rho(A)$ of the domain A as the publication number growth rate between the two successive periods (periods of same length, or appropriate annual averages on the citing, respectively the cited period) : $\rho(A) = \frac{|A^{T_1}|}{|A^{T_0}|}$. We can also define $\hat{\rho}(A)$, A's relative growth rate with respect with the general growth rate of science by computing the ratio between its growth rate between the two successive time periods with the global growth rate assessed at the whole science scale:

$$\hat{\rho}(A) = \frac{|A^{T_1}|/|A^{T_0}|}{|S^{T_1}|/|S^{T_0}|}$$

It should be kept in mind that the growth rate depends on the citing period with respect to the cited period, just as the balances defined below, the relative length of bibliography when needed, and the final relative impact.

The impact of a domain A is defined by the average number of incoming citations per citable articles in:

$$I(A) = \frac{\phi_{\leftarrow}(A)}{|A^{T_0}|}$$

The relation $I(S) = \frac{\phi_{\leftarrow}(S)}{|S^{T_0}|}$ can be written $I(S) = \frac{\kappa(S^{T_1})|S^{T_1}|}{|S^{T_0}|}$ as the total incoming citation flows can be written as the total number of citing paper times the average number of references in citing articles $\kappa(S^{T_1})$.

$$\mathrm{I(S)} = \kappa(S^{T_1})\rho(S)$$

We can also detail the citation in-flow according to the sources, which yields the impact of the domain:

$$I(A) = \frac{\phi_{\leftarrow}(A)\phi_{\rightarrow}(A)}{\phi_{\rightarrow}(A)|A^{T_0}|}$$

We then define the balance ratio, which compares the total inflow with the total outflow of A, such as :

$$B(A) = \frac{\phi_{\leftarrow}(A)}{\phi_{\rightarrow}(A)}$$

Combining the two previous equations, the impact of a domain A can be described as:

$$I(A) = B(A)\frac{\phi_{\rightarrow}(A)}{|A^{T_0}|}$$

By definition, $\phi_{\rightarrow}(A) = \kappa(A^{T_1})|A^{T_1}|$, the global equation then rewrites:

$$I(A) = B(A)\kappa(A)\rho(A)$$

Defining the relative impact $\hat{I}(A)$ of A as the absolute impact $I(A)$ divided by the absolute impact of all science $I(S)$, it comes:

$$\hat{I}(A) = B(A)\hat{\kappa}(A)\hat{\rho}(A)$$

where $\hat{\kappa}(A) = \frac{\kappa(A)}{\kappa(S)}$ is the average number of out-going citation in A normalized with respect to S, that is the relative length of bibliography.

This gross relative impact then depends on three components linked to dynamic aspects (relative growth $\hat{\rho}$ and exchanges B) and the variations of citation habits $\hat{\kappa}$. Interpretation in terms of dynamics of science could be possible based on a raw citation flows count, differential growth and exchanges being elements of changes in the system but they would be blurred by citation habits.

The purpose of citing side normalization of citation flows is to get rid of those variations which can be quite large (between one and two orders of magnitude at the "subject category" level). The citing-side normalization neutralizes the factor of citation habits $\hat{\kappa} = 1$ and makes comparison possible on the whole system.

From the original citation network, we can derive $G^g = (S, C^g)$ where citations links C^g are normalized with respect to the average propensity to cite of each domain. Each citation coming from a publication in A is assigned a weight. This procedure provides more weight to citations stemming from domains producing fewer citations on average. The edges of C^g coming from publications in A are then weighted according to the formula:

$$C^g(i,j) = C(i,j)\frac{w^g(I)}{w^g(S)}$$

where $w^g(S)$ and $w^g(I)$ are the average number of citations produced by publications published in I or in S: $w^g(I) = \frac{\sum_{i\in I,j} C(i,j)}{|A^{T_1}|}$ and $w^g(S) = \frac{\sum_{i,j} C(i,j)}{|S^{T_1}|}$

The same argument regarding citation flows and related impacts holds with this new normalized definition of the citation matrix, such that the general equation is simplified:

$$\hat{I}(A) = B(A)\hat{\rho}(A)$$

THE *CITER-SUCCESS-INDEX*: AN INDICATOR TO SELECT A SUBSET OF ELITE PAPERS, BASED ON CITERS

Fiorenzo Franceschini[1], Domenico Maisano[2] and Luca Mastrogiacomo[3]

[1] *fiorenzo.franceschini@polito.it* [2] *domenico.maisano@polito.it*
[3] *luca.mastrogiacomo@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract
The goal of this paper is introducing the *citer-success-index* (*cs*-index), i.e., an indicator that uses the number of different citers as a proxy for the impact of a generic set of papers. For each of the articles of interest, it is defined a comparison term – which represents the number of citers that, on average, an article published in a certain period and scientific field is expected to "infect" – to be compared with the actual number of citers of the article. Similarly to the recently proposed *success*-index (Franceschini et al., Scientometrics 92(3):621-6415, 2011), the *cs*-index allows to select a subset of "elite" papers.
The *cs*-index is analyzed from a conceptual and empirical perspective. Special attention is devoted to the study of the link between the number of citers and cited authors relating to articles from different fields, and the possible correlation between the *cs*- and the *success*-index.
Some advantages of the *cs*-index are that (i) it can be applied to multidisciplinary groups of papers, thanks to the field-normalization that it achieves at the level of individual paper and (ii) it is not significantly affected by self citers and recurrent citers. The main drawback is its computational complexity.

Conference Topic
Scientometrics Indicators: Criticism and new developments, Relevance to Science and Technology (Topic 1).

Introduction and Literature Review
In bibliometrics, one of the main analysis dimensions is the impact of scientific publications, which is commonly estimated by counting the number of citations that they accumulate over time (Egghe and Rousseau, 1990). As an alternative to citations, Dieks et al. (1976) and Braun et al. (1985) suggested to use the total number of different citers (or citing authors), i.e., the members of the scientific community who are "infected" by a certain paper. The number of different citers is a proxy which is harder to compute, but more elegant, as only marginally affected by citations from self citers and recurrent citers.

The idea of citers was recently dug up by Ajiferuke and Wolfram (2010), who proposed and implemented an indicator based on citers, without encountering the computational obstacles of the past, thanks to the current evolution of databases and information management tools. The indicator is the *ch*-index, defined for a generic group of papers (e.g., those of a scientist, journal or entire research institution) as *the number (ch) such that, for a general group of papers, ch papers are cited by at least ch different citers while the other papers are cited by no more than ch different citers*. It can be immediately noticed that this definition is similar to that of the *h*-index, with the only exception that, for each publication, the citations obtained are replaced by the number of different citers (Hirsch, 2005).

The *ch*-index was empirically analyzed by Franceschini et al. (2010). This study showed: (i) the general correlation between *ch* and *h*, and (ii) the potential of *ch* in complementing the information given by *h*. E.g., paradoxical situations in which the number of citations obtained by a paper and the number of different citers do not go hand in hand are not so rare, due to the anomalous incidence of recurrent or self citers. A theoretical interpretation of the correlation between *ch* and *h* was recently provided by Egghe (2012).

In this article we focus the attention on the *success*-index (*s*-index), i.e., a recent indicator that, for a generic set of articles, allows to select an "elite" subset, according to a logic different from that of *h* (Franceschini et al., 2012a). The *s*-index is defined as *the number of papers with a number of citations greater than or equal to CT_i, i.e., a generic comparison term associated to the i-th publication.* CT_i is an estimate of the number of citations that articles of the same scientific context and period of time of that of interest (i.e., the *i*-th publication) are likely to achieve.

With the aim of formalizing this definition, a score is associated to each (*i*-th) of the (*P*) publications of interest:

$$\begin{cases} score_i = 1 & \text{when } c_i \geq CT_i \\ score_i = 0 & \text{when } c_i < CT_i \end{cases}, \tag{1}$$

where c_i are the citations obtained by the *i*-th publication. The *s*-index is therefore given by:

$$s\text{-index} = \sum_{i=1}^{P} score_i \ . \tag{2}$$

Apart from *s*, there are other indicators in the literature that allow to select an elite subset, based on the comparison between the number of citations accumulated by each paper and a threshold. E.g., let us consider the selection by $P_{top\ 10\%}$-indicator (Bornmann, 2013), that by π-indicator (Vinkler, 2011), the characteristic scores and scales (CSS) method (Glänzel, 2011) or the ESI's Highly Cited Papers method (ISI Web of Knowledge, 2012). We remark that, differently from *s*, the aforementioned methods require that the set of publications examined are preliminarily categorized into scientific (sub-)disciplines.

As regards the *s*-index, there are several options for constructing the CT_i related to an *i*-th paper of interest. Generally, three issues are crucial (Franceschini et al., 2012b):

1. Defining the procedure for selecting a reference sample of homologous publications. Possible approaches are: (i) the selection of papers of same age, type (e.g. research article, review, letter, etc.) and published by the same journal of the *i*-th paper of interest, (ii) the use of superimposed classifications such as ISI subject categories, (iii) the implementation of "adaptive" techniques in which the sample is determined considering the "neighbourhood" of the paper of interest – typically consisting of the set of papers citing or being cited by it.
2. Deciding whether to consider (i) the distribution of the number of references given or (ii) the citations obtained by the publications of the sample.
3. Identifying a suitable (central tendency) indicator for obtaining CT_i from the distribution of interest, e.g., mean, median, harmonic mean, percentiles, etc..

Regarding point (2), Franceschini et al. (2012a, 2012c) state that indicators based on the distribution of references given – rather than citations obtained – have several advantages:

- The number of references is fixed over time, while the number of citations obtained tends to increase and requires a certain accumulation period to stabilize.
- This stability is also derived by the fact that the number of references is likely to be less variable than the number of citations obtained.
- Bibliographic references are less influenced by journal particularities, such as the average citation impact of articles.

Conceptually, the link between references given (by the papers of the reference sample) and citations obtained (by the papers of interest) originates from a simple consideration: focussing on the totality of the scientific literature in a certain field and according to a simplified model configuration of *isolated* fields – i.e., excluding transfers of citations between different disciplines – the following relationship applies:

$$\sum_{i=1}^{P} c_i = \sum_{i=1}^{P} r_i \,, \tag{3}$$

where

P is the total number of articles (that can cite each other) in the isolated field;
c_i is the number of citations obtained by the *i*-th paper;
r_i is the number of citations given by the *i*-th paper.

The equality of Eq. 3 can also be expressed in terms of average values:

$$\frac{1}{P}\sum_{i=1}^{P} c_i = \frac{1}{P}\sum_{i=1}^{P} r_i \quad \Rightarrow \quad \bar{c} = \bar{r} \,. \tag{4}$$

For more detailed and rigorous information on the relation between the $\bar{c}$ and $\bar{r}$ values concerning a set of documents, we refer the reader to (Egghe & Rousseau, 1990).

Returning to the *s*-index, apart from the simplicity of meaning, a great advantage is that it implements a field-normalization at the level of single paper and can therefore be applied to multidisciplinary groups of articles, for instance the whole production output of a research institution.

Another important quality of the *s*-index is that it is defined on a *ratio* scale. This feature has several practical implications that make this indicator more versatile than others – such as the *h*-index, which is defined on an *ordinal* scale (Franceschini et al., 2012a):

- The *s*-index reflects compositions of the input publication sets (with the corresponding citations). In other terms, the union of two groups of publications with *s*-index of 2 and 5 (with no common publications) will always originate a third group of publications with *s*-index of 2 + 5 = 7. This simple property is very useful for extending the use of the *s*-index to multi-disciplinary institutions, e.g., joining groups of publications from different scientific fields.
- The *s*-index eases normalizations aimed at obtaining the so-called size-independency (Franceschini et al., 2012c). Given a general group of papers and the same capacity of producing successful papers, it is reasonable to assume that thr *s*-index should increase proportionally with the different types of "resources" deployed. In fact, several normalized indicators can be obtained dividing the *s*-index by the resource unit of interest; e.g, the staff number of a research institution, the age of a researcher, the number of articles of a journal, the amount of funding received in a certain period, etc..

The purpose of the paper is introducing the *citer-success*-index (or *cs*-index), i.e., a variant of the *s*-index, which is based on citers instead of citations, according to a logic similar to that of *ch*. Given a set of articles, the *cs*-index identifies a subset for which the number of different citers of an *i*-th article exceeds a specified comparison term cCT_i. Formalizing, a score is associated to each *i*-th of the (*P*) publications of interest:

$$\begin{cases} score_i = 1 & \text{when } \gamma_i \geq cCT_i \\ score_i = 0 & \text{when } \gamma_i < cCT_i \end{cases}, \tag{5}$$

where γ_i are the unique citers related to the *i*-th publication. The word "unique" means that repeated citers are counted only once. The *cs*-index is therefore given by:

$$cs\text{-index} = \sum_{i=1}^{P} score_i \tag{6}$$

Figure 1(a) exemplifies the calculation of the *s*- and *cs*-index for a fictitious set of papers.

In analogy with CT_i, cCT_i is an estimate of the number of unique citers that articles homologous to that of interest are likely to "infect".

(a)

paper no.	c_i	CT_i	γ_i	cCT_i	*s*-elite	*cs*-elite
1	115	20.3	297	60.1	✓	✓
2	86	21.2	187	71.0	✓	✓
3	17	14.5	31	44.8	✓	×
4	15	20.4	68	72.4	×	×
5	12	11.8	30	29.2	✓	✓
6	9	15.7	12	61.9	×	×
					s-index=4	*cs*-index=3

(b)

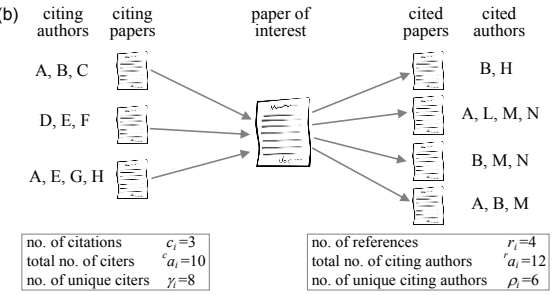


Figure 1. Propaedeutic examples: (a) calculation of the *s*- and *cs*-index for a fictitious set of papers, and (b) introduction of some indicators concerning the authors (represented by letters, e.g., A, B, C, etc.) of papers citing/cited by a fictitious paper of interest.

Similarly to CT_i, there are three basic steps when constructing the cCT_i relating to an *i*-th article of interest:

1. Selecting a sample of articles homologous to that interest.
2. Deciding whether to consider the distribution of (i) unique citers or (ii) unique cited authors, relating to the papers of the sample.
3. Defining cCT_i by an indicator of central tendency, applied to the distribution chosen at point (2).

The choice at point (2) is more delicate than in the case of the *s*-index. Intuitively, it may appear convenient to use the distribution of unique cited authors for the same reasons for which, in the case of the *s*-index, it was convenient to use the distribution of references. However, the link between unique citers and unique cited authors is not necessarily similar to that between r_i and c_i values; even in a model configuration of isolated fields:

$$\sum_{i=1}^{P} \gamma_i \text{ is not necessarily} = \sum_{i=1}^{P} \rho_i , \tag{7}$$

being

P the total number of papers in the isolated field;

γ_i the number of unique citers of the *i*-th paper;

ρ_i the number of unique authors cited by the *i*-th paper.

The reason for this lack of parallelism is twofold and will be examined later in the manuscript.

The rest of the paper is structured in three sections. The section "General link between citers and cited authors" investigates whether it is appropriate to construct the cCT_i by using the distribution of the number of unique authors cited by a sample of papers. The section "Preliminary Empirical analysis of the *cs*-

index" delves into the issue raised in the previous section, examining a large number of papers from different fields. After defining the cCT_i properly, it is studied the correlation between the *s*- and the *cs*-index. Finally, the section "Further remarks" summarizes the original contributions of the paper and the main advantages and disadvantages of the *cs*-index.

General link between citers and cited authors

Before getting into the problem, Figure 1(b) introduces the reader to the indicators and notation that will be used in the remaining of the paper.

Even modelling a scientific field as isolated and considering the totality of the scientific production in it, there are two possible elements of diversity among citing and cited papers: (i) different average number of authors per paper, and (ii) different percentage of unique authors. Let us clarify this point with simple mathematical considerations. The quantity $\sum_{i=1}^{P}\gamma_i$ can be expressed as:

$$\sum_{i=1}^{P}\gamma_i = \left(\sum_{i=1}^{P}\gamma_i \Big/ \sum_{i=1}^{P}{}^{c}a_i\right)\cdot\left(\sum_{i=1}^{P}{}^{c}a_i \Big/ \sum_{i=1}^{P}c_i\right)\cdot\sum_{i=1}^{P}c_i \;=\; {}^{c}p\cdot{}^{c}app\cdot\sum_{i=1}^{P}c_i \quad , \tag{8}$$

in which

γ_i is the number of unique citers of the *i*-th paper in the isolated field;

${}^{c}a_i$ ($\geq \gamma_i$) is the total number of citers (even repeated, in the case that some citing papers are (co-)authored by the same individuals) related to the *i*-th paper;

c_i is the number of citing papers (or the number of citations obtained) relating to the *i*-th paper;

P is the total number of articles in the isolated field.

As shown in Eq. 8, the quantity $\sum_{i=1}^{P}\gamma_i$ can also be seen as the product of three terms:

${}^{c}p$ $= \sum\gamma_i / \sum^{c}a_i$ (≤ 1) i.e., the percentage of unique citers;

${}^{c}app$ $= \sum^{c}a_i / \sum c_i$ (≥ 1) i.e., the average number of authors per citing paper;

$\sum_{i=1}^{P}c_i$ the total number of citations obtained.

A "decomposition" similar to that of Eq. 8 may apply to the quantity $\sum_{i=1}^{P}\rho_i$:

$$\sum_{i=1}^{P}\rho_i = \left(\sum_{i=1}^{P}\rho_i \Big/ \sum_{i=1}^{P}{}^{r}a_i\right)\cdot\left(\sum_{i=1}^{P}{}^{r}a_i \Big/ \sum_{i=1}^{P}r_i\right)\cdot\sum_{i=1}^{P}r_i \;=\; {}^{r}p\cdot{}^{r}app\cdot\sum_{i=1}^{P}r_i \quad , \tag{9}$$

in which

ρ_i is the number of unique authors cited by the *i*-th paper in the isolated field;

${}^{r}a_i$ ($\geq \rho_i$) is the total number of cited authors (even repeated, in the case that some cited papers are (co-)authored by the same individuals) related to the i-th paper;

r_i is the number of papers cited (or the number of bibliographic references) relating to the i-th paper;

P is the total number of articles in the isolated field.

Similarly to $\sum_{i=1}^{P} \gamma_i$, $\sum_{i=1}^{P} \rho_i$ can be seen as the product of three terms:

${}^{r}p = \sum \rho_i / \sum {}^{r}a_i$ (≤ 1) i.e., the percentage of unique cited authors;

${}^{c}app = \sum {}^{r}a_i / \sum r_i$ (≥ 1) i.e., the average number of authors per cited paper.

$\sum_{i=1}^{P} r_i$ the total number of references given.

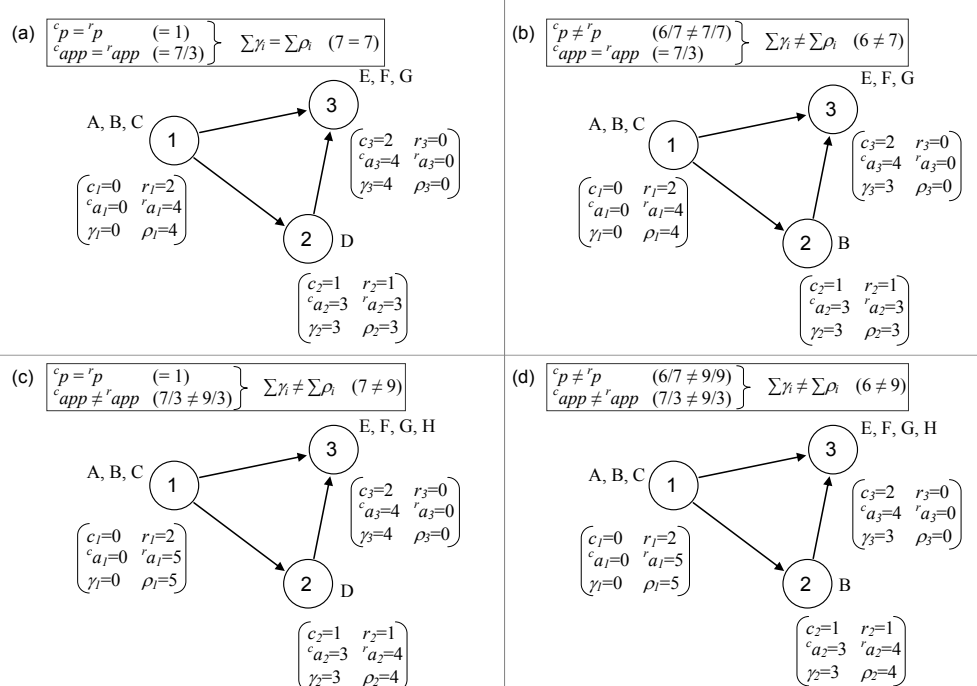


Figure 2. Examples of isolated groups of three papers. Nodes represent the papers (1, 2 and 3), whose authors are A, B, C, D, etc.; arrows represent the citations given by one paper to another. For each paper, it is reported the number of citations obtained (c_i), the number of references given (r_i), the number of total citers (${}^{c}a_i$), the number of total cited authors (${}^{r}a_i$), the number of unique citers (γ_i) and the number of unique cited authors (ρ_i). The equality of Eq. 7 is satisfied in case (a) only, when ${}^{c}p = {}^{r}p$ and ${}^{c}app = {}^{r}app$.

Combining Eqs. 8 and 9 with Eq. 3, it is obtained:

$$\sum_{i=1}^{P}\gamma_i = \left(\frac{{}^{c}p}{{}^{r}p}\cdot\frac{{}^{c}app}{{}^{r}app}\right)\cdot\sum_{i=1}^{P}\rho_i \ . \tag{10}$$

The "balanced" situation $\sum\gamma_i = \sum\rho_i$ can be achieved in the case the following two (sufficient but not necessary) conditions occur (also see the exemplification in Figure 2):

$$\begin{aligned} {}^{c}p &= {}^{r}p \\ {}^{c}app &= {}^{r}app \end{aligned} \tag{11}$$

that is to say, (i) equal average percentage of unique authors and (ii) equal average number of authors for the papers citing and being cited by the total P papers in the isolated field.

Eq. 7 could also be met without necessarily satisfying the two conditions in Eq. 11, that is to say in the case the quantity in brackets in Eq. 10 was unitary. However, there is no practical reason that justify the occurrence of this coincidence, which is purely conjectural. On the other hand, the two conditions of Eq. 11 seem reasonable for (citing and cited) papers within the same field. In any case, they will be tested empirically in the next section.

Table 1. List of journals analyzed within seven ISI subject categories (WoS). For each journal, we considered the research papers issued in the three-year period from 2008 to 2010.

Discipline (ISI Subject Category)	Journal and abbreviation	No. of papers			
		2008	2009	2010	Total
Biology	Bio1 - Bioscience	84	65	66	215
	Bio2 - Biology Direct	46	41	65	152
	Bio3 - Journal of Biosciences	60	65	52	177
Chemistry (analytical)	Che1 - Analytical Sciences	264	238	209	711
	Che2 - Journal of Chemometrics	83	68	76	227
	Che3 - Microchemical Journal	85	114	151	350
Engineering (manufacturing)	Eng1 - International J. of Machine Tools & Manufacture	164	139	118	421
	Eng2 - Robotics and Computer-Integrated Manufacturing	77	96	87	260
	Eng3 - Journal of Intelligent Manufacturing	57	62	71	190
Mathematics	Mat1 - Computational Complexity	20	20	21	61
	Mat2 - Constructive Approximation	31	46	38	115
	Mat3 - Advances in Mathematics	169	146	190	505
Medicine (general & internal)	Med1 - American Journal of Medicine	112	98	119	329
	Med2 - Mayo Clinic Proceedings	86	55	74	215
	Med3 - Medicine	33	40	30	103
Physics (applied)	Phy1 - Applied Physics Express	341	339	345	1025
	Phy2 - Current Applied Physics	177	430	436	1043
	Phy3 - Journal of Magnetic Resonance	230	214	241	685
Psychology	Psy1 - Journal of Experimental Psychology: Learning Memory and Cognition	66	94	52	212
	Psy 2 - Cognitive Psychology	18	26	24	68
	Psy 3 - Health Psychology	125	90	73	288

Preliminary empirical analysis of the *cs*-index

Data collection

A preliminary empirical analysis of the *cs*-index is performed by selecting some papers from a set of journals of seven different ISI subject categories (in brackets the total number of journals indexed by Thomson Scientific in each category): Biology (85), Analytical Chemistry (73), Manufacturing Engineering (37), Mathematics (289), General & Internal Medicine (155), Applied Physics (125), Psychology (75). For each discipline, we selected a random sample of three scientific journals. For each journal, we considered as articles of interest those produced in the three-year period from 2008 to 2010, limiting the selection to research papers only (other document types, such as reviews, conference papers or letters, were excluded). Table 1 contains the journal titles and the number of articles examined for each year. Data are retrieved by querying the Web of Science[1] (WoS) database (Thomson Reuters, 2012).

For each *i*-th article of interest, the following operations are performed.

1. Collection of the citation statistics, consisting of:
   - c_i the number of citing papers published in 2011 and indexed by the database in use;
   - ${}^{c}a_i$ the total number of authors of the (c_i) citing papers (even repeated, if different citing papers are (co-)authored by the same individuals);
   - γ_i the total number of unique citers, obtained by performing the union of the (${}^{c}a_i$) total citers and removing those repeated.

   The choice of a time window for citations accumulation of one year (2011) is to simplify the analysis.
2. Determination of an appropriate cCT_i, which takes into account the propensity to obtain citations from different authors. The construction of cCT_i is based on a sample of *S* articles that are issued in 2011 by the same journal of the (*i*-th) article of interest.

   For each *j*-th of the articles of the sample, we determine:
   - r_j the number of cited papers that were published in the three-year period from 2008 to 2010 and are indexed by the database in use. These constraints were introduced to be consistent with the time window described at point (1) (Moed, 2011);
   - ${}^{r}a_j$ the total number of cited authors (even repeated, if different cited papers are authored by the same individuals);
   - ρ_j the total number of unique cited authors, obtained by the union of the (${}^{r}a_i$) total cited authors, removing those repeated.

   Next, the distribution of the ρ_j values (relating to the papers of the sample) is constructed and the cCT_i is defined by an appropriate central tendency indicator – e.g., the mean ($\overline{\rho}$) or median ($\widetilde{\rho}$). This construction is based on the assumption that, referring to the *i*-th article, the propensity to be cited by

different authors is, on average, reasonably close to the propensity to cite different authors, referring to articles issued by the same journal. According to this construction, articles published in the same journal and in the same year will have the same cCT_i value. Probably, a more rigorous way to estimate the cCT_i – but also computationally more expensive – is to use the distribution of the ρ_j values relating to the articles that cite other articles, issued by the article of interest's journal. For further information about this point, please refer to (Franceschini et al., 2012c).

Table 2. Summary of the analysis results. For each of the journals (in Table 1), we report the indicators illustrated in the "Data collection" sub-section. *Overall* indicators are obtained by aggregating the data relating to the three journals examined in each field.

Field	Journ.	${}^{c}app$	${}^{r}app$	${}^{c}p$	${}^{r}p$	P	C	CPP	h	ch	S	R	cCT_i		cs-index		CT_i		s-index	
													$\bar{\rho}$	$\tilde{\rho}$	$(\bar{\rho})$	$(\tilde{\rho})$	$\bar{r}$	$\tilde{r}$	$(\bar{r})$	$(\tilde{r})$
Bio	Bio1	4.6	5.5	0.95	0.91	215	1131	5.3	14	37	76	792	52.3	35.0	25	38	10.4	9.0	30	35
	Bio2	4.9	6.5	0.94	0.86	152	469	3.1	9	26	59	943	89.4	60.0	3	4	16.0	14.0	2	2
	Bio3	5.3	5.9	0.86	0.93	177	274	1.5	7	19	71	382	29.3	18.0	9	20	5.4	4.0	16	17
	overall	4.8	6.0	0.93	0.89	544	1874	3.4	15	45	206	2117	55.0	35.0	31	57	10.3	8.5	37	52
Che	Che1	4.4	4.5	0.89	0.83	711	905	1.3	7	20	191	1076	21.1	17.0	14	30	5.6	5.0	14	14
	Che2	3.9	3.9	0.92	0.86	227	371	1.6	7	17	65	304	15.8	12.0	22	29	4.7	4.0	15	15
	Che3	4.3	4.3	0.92	0.88	350	948	2.7	9	28	185	1274	25.9	22.0	35	50	6.9	5.0	29	51
	overall	4.3	4.3	0.91	0.86	1288	2224	1.7	10	30	441	2654	22.4	17.0	71	128	6.0	5.0	44	78
Eng	Eng1	3.6	3.3	0.86	0.84	421	1148	2.7	9	23	98	392	11.3	9.0	115	142	4.0	3.0	78	126
	Eng2	3.2	3.1	0.93	0.88	260	374	1.4	6	15	101	229	6.2	5.0	74	86	2.3	2.0	57	57
	Eng3	3.0	2.8	0.90	0.93	190	191	1.0	6	10	78	140	4.6	3.0	41	54	1.8	1.0	43	43
	overall	3.4	3.2	0.88	0.87	871	1713	2.0	10	24	277	761	7.6	5.0	261	341	2.7	2.0	266	266
Mat	Mat1	2.2	2.4	0.92	0.86	61	39	0.6	2	6	19	25	2.7	1.0	11	17	1.3	1.0	11	11
	Mat2	2.5	2.1	0.88	0.80	115	178	1.5	4	8	36	87	4.0	3.0	18	26	2.4	1.0	17	31
	Mat3	1.9	2.0	0.88	0.77	687	912	1.3	7	11	290	819	4.3	3.0	113	157	2.8	2.0	126	126
	overall	2.0	2.0	0.88	0.77	863	1129	1.3	7	13	345	931	4.2	3.0	138	190	2.7	2.0	145	145
Med	Med1	5.3	7.5	0.93	0.91	329	533	1.6	6	25	125	946	51.4	36.0	1	7	7.6	6.0	1	4
	Med2	5.3	6.8	0.92	0.89	215	996	4.6	14	37	75	833	66.8	42.0	12	31	11.1	8.0	18	27
	Med3	5.6	7.7	0.92	0.91	103	489	4.7	10	29	48	424	61.8	45.5	7	12	8.8	7.0	17	20
	overall	5.4	7.3	0.92	0.90	647	2018	3.1	15	44	248	2203	58.1	40.0	26	56	8.9	6.0	45	82
Phy	Phy1	5.8	6.1	0.82	0.81	1025	2919	2.8	17	50	418	2483	29.1	24.0	122	147	5.9	5.0	149	149
	Phy2	4.5	4.8	0.89	0.85	1043	1939	1.9	12	34	526	2573	20.1	14.0	99	160	4.9	4.0	111	111
	Phy3	4.4	4.5	0.87	0.79	685	1579	2.3	11	31	243	1671	24.1	19.0	53	80	6.9	6.0	37	37
	overall	5.1	5.2	0.85	0.82	2753	6437	2.3	17	55	1187	6727	24.1	19.0	270	395	5.7	5.0	287	287
Psy	Psy1	2.9	2.7	0.89	0.79	212	545	2.6	10	18	78	596	16.7	15.0	20	23	7.6	7.0	12	12
	Psy2	2.9	2.5	0.88	0.85	68	298	4.4	7	16	17	172	21.3	19.0	10	11	10.1	9.0	5	5
	Psy3	4.3	4.4	0.93	0.89	288	1245	4.3	12	35	90	738	32.4	26.0	43	58	8.2	7.0	32	41
	overall	3.8	3.5	0.92	0.86	568	2088	3.7	15	37	185	1506	24.7	19.0	87	121	8.1	7.0	50	60

The *cs*-index related to the articles of each journal can be calculated using the cCT_i determined at point (2) (according to Eq. 5). The information at point (2) can also be used to determine the average number of authors (${}^{r}app$) and the percentage of unique authors (${}^{r}p$) of the articles cited by the (S) articles of the sample (see Eq. 9). Similarly, the information at point (1) can be used to determine the average number of authors (${}^{c}app$) and the percentage of unique authors (${}^{c}p$) of the articles that cite the (P) articles of interest (see Eq. 8).

The overall ${}^{c}app$, ${}^{r}app$, ${}^{c}p$ and ${}^{r}p$ values of the seven fields examined can be estimated by aggregating data related to the three journals considered in each discipline.

Information at point (1) can also be used to build other indicators: C (i.e., total number of citations), CPP (i.e., average citations per paper), h, ch and s. As regards the s-index, we will compare the (c_i) citations obtained by each (i-th) paper with a CT_i represented by the mean or median number of references ($\bar{r}_j$ and $\tilde{r}_j$ respectively) that are given by each (j-th) of the articles of the sample.

Conventionally, all indicators are constructed considering the citations obtained in 2011 and the references given to (cited) articles, issued from 2008 to 2010 and indexed by WoS.

Data analysis

Table 2 summarises the results of the empirical analysis. For each journal, the $C = \sum c_i$ total citing papers are those citing each (i-th) of the P papers of interest, and the $R = \sum r_i$ total cited papers are the ones cited by each (j-th) of the S articles of the sample. All statistics were constructed considering the aforementioned time windows and the papers indexed by WoS.

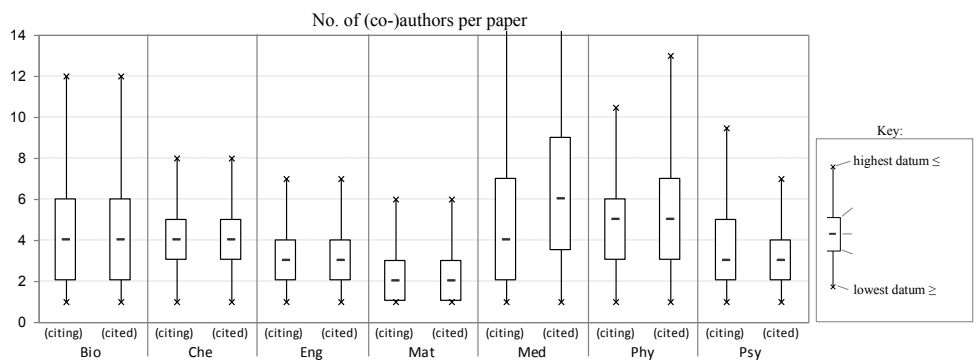


Figure 3. Box-plot of the distribution of the number of (co-)authors relating to the citing and cited papers, concerning the seven fields examined. Citing papers are those that cite the *P* papers of interest while cited papers are those cited by the *S* papers of the macro-sample. $Q^{(1)}$, $Q^{(2)}$ and $Q^{(3)}$ are the first, second and the third quartile of the distributions of interest.

For a specific journal, there are marginal differences between citing and cited authors, as regards (i) the average number of authors per paper (i.e., ${}^{c}app$ and ${}^{r}app$ values) and (ii) the percentage of unique authors (i.e., ${}^{c}p$ and ${}^{r}p$ values).

Besides, there are relatively small variations among the three journals in a specific field. For this reason, it seems appropriate to calculate some aggregated indicators for the whole disciplines (see "overall" indicators in Table 2). The determination of the overall indicators – by joining the data related to the three journals in each discipline – is extended to all the indicators presented in Table 2. In the case of the cs-index and s-index, overall indicators are constructed using cCT_i and CT_i

values determined on the basis of macro-samples obtained by joining the articles issued in 2011 by the three journals selected for each discipline.
Returning to the comparison between ${}^{c}app$ and ${}^{r}app$ values in each field, a simple way to visualize their similarity is through box-plots based on overall statistics. In particular, two distributions are considered; (i) that of the number of authors per paper relating to articles that cite the papers of interest, and (ii) that of the papers cited by the papers of the (macro-)sample (see Figure 3).
It can be seen that, for each discipline, the notches of the two box-plots (respectively for citing and cited papers) almost completely overlap, supporting the view of absence of systematic differences between the two distributions. The same hypothesis can be tested by more rigorous statistical tests, albeit introducing additional assumptions about distributions. On the contrary, when comparing different fields there are systematic differences, confirming what observed in other studies (Glänzel, 2002). For example, let us consider the comparison between the notches relating to Mathematics and Physiscs.
As regards the comparison between ${}^{c}p$ and ${}^{r}p$ values, the question is a bit more complicated: the overall percentages of different authors (respectively citing or cited) can be seen as weighted averages of the same percentages, at the level of individual papers:

$$\begin{aligned} {}^{c}p &= \left(\sum_{i=1}^{P}\gamma_i\right)\Big/\left(\sum_{i=1}^{P}{}^{c}a_i\right) = \left(\sum_{i=1}^{P}{}^{c}p_i\cdot{}^{c}a_i\right)\Big/\left(\sum_{i=1}^{P}{}^{c}a_i\right) \\ {}^{r}p &= \left(\sum_{j=1}^{S}\rho_j\right)\Big/\left(\sum_{j=1}^{S}{}^{c}a_j\right) = \left(\sum_{j=1}^{S}{}^{r}p_j\cdot{}^{r}a_j\right)\Big/\left(\sum_{j=1}^{S}{}^{r}a_j\right) \end{aligned}, \qquad (13)$$

being

- ${}^{c}p_i$ the percentage of unique citers relating to the i-th of the P papers of interest;
- ${}^{c}a_i$ the "weight" of ${}^{c}p_i$, i.e., the number of authors (even repeated) citing the i-th paper;
- ${}^{r}p_j$ the percentage of unique authors cited by the j-th of the S papers of the sample;
- ${}^{r}a_j$ the "weight" of ${}^{r}p_j$, i.e., the number of authors (even repeated) cited by the j-th paper.

Being ${}^{c}p$ and ${}^{r}p$ weighted quantities, one can represent the distributions of ${}^{c}p_i$ and ${}^{r}p_j$ values by special box-plots based on *weighted quartiles*, defined as:

- ${}^{c}Q_w^{(1)}$, ${}^{c}Q_w^{(2)}$ and ${}^{c}Q_w^{(3)}$, i.e., the weighted first, second (or weighted median) and third quartile of the ${}^{c}p_i$ values. These indicators are obtained by ordering in ascending order the ${}^{c}p_i$ values of the articles of interest and considering the values for which the cumulative of weights is equal to respectively the 25%, 50% and 75% of their sum;
- ${}^{r}Q_w^{(1)}$, ${}^{r}Q_w^{(2)}$ and ${}^{r}Q_w^{(3)}$, i.e., the weighted first, second (i.e., the weighted median) and third quartile of the ${}^{r}p_i$ values.

The box-plots relating to weighted quartiles are represented in Figure 4. The differences between the ${}^{c}p_i$ and ${}^{r}p_j$ distributions within the same field seem insignificant. We also note the absence of significant differences between fields.

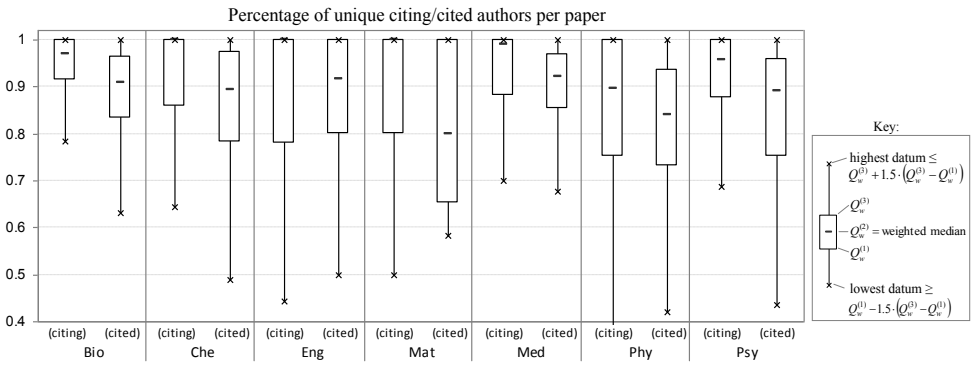


Figure 4. "Weighted" box-plot of the percentage of unique citing (${}^{c}p_i$) and cited authors (${}^{r}p_i$), relating to the papers that cite the papers of interest and are cited by the papers of the macro-sample, in the seven fields examined. $Q_w^{(1)}$, $Q_w^{(2)}$ and $Q_w^{(3)}$ are the first, second and the third weighted quartile of the distributions of interest

Returning to Table 2, there are relatively little differences in terms of cCT_i values (i.e., estimators of the propensity to cite different authors), for journals of the same field. Some exceptions are: Bio2 for Biology and Eng1 for Engineering. This incomplete uniformity is probably due to the fact that some journals are influenced by publications of neighbouring fields, with different citation propensity. For a more rigorous estimate, it would probably be appropriate to define cCT_is using a larger sample of papers/journals.

For each journal, in Table 2 are reported two different cCT_js: i.e., using $\overline{\rho}$ and $\widetilde{\rho}$. In general, the resulting values are higher in the first case. This probably depends on the incidence of papers characterized by hyperauthorship – i.e., literally tens or even hundreds of authors (Cronin, 2001) – which tends to "inflate" $\overline{\rho}$ but not $\widetilde{\rho}$, as the latter indicator is only marginally sensitive to the right tail of the distribution of ρ_j values.

Another interesting aspect is the link between *cs*-index and *s*-index. The diagram in Figure 5 – which is constructed using $cCT_i = \overline{\rho}$ and $CT_i = \overline{r}$ (in Table 2) – shows a strong correlation ($R^2 \approx 89\%$), similar to that between *ch* and *h* (Franceschini et al., 2010; Egghe, 2012). All the points of the graph – although resulting from articles of different scientific fields – tend to be distributed around the same trend line, which is very close to the bisector of the *cs*-*s* plane.

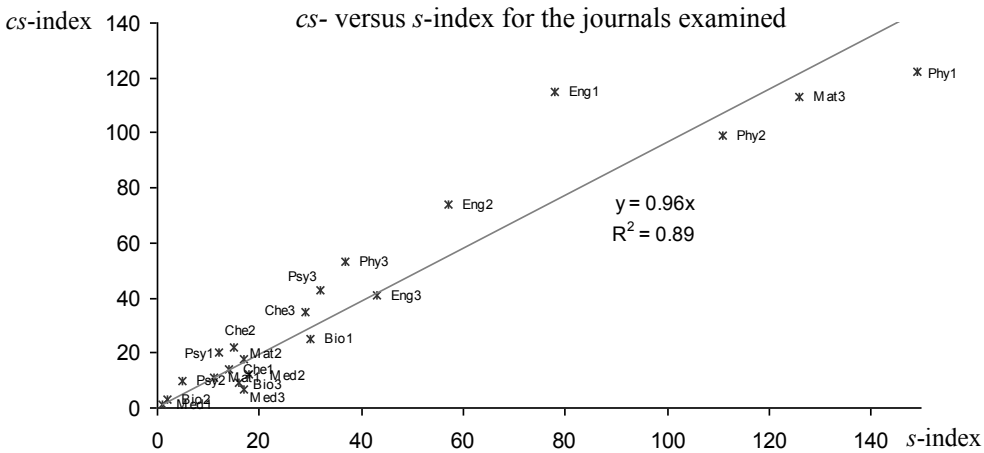


Figure 5. Relationship between the *cs*- and *s*-index for the journals examined. Indicators are calculated considering respectively $cCT_i = \overline{\rho}$ and $CT_i = \overline{r}$ (see Table 2).

In the absence of "anomalies" – e.g., high incidence of self-citations or citations from recurrent citing authors – the *cs*-index and *s*-index should be very close. Therefore, the study of their difference can be useful to highlight abnormal situations. For example, consider the point related to Med3 in Figure 5, which corresponds to a relatively high value of *s*-index, associated to a quite small value of *cs*-index, probably due to a relatively high incidence of self citers and recurrent citing authors. On the contrary, the point related to Eng1 denotes an opposite situation, in which *cs*-index is much larger than *s*-index. probably due to an opposite attitude.

Further remarks

This study revealed some interesting points that it is worth summarizing and developing in the following:

- The analysis suggests that the comparison term (cCT_i) of the *cs*-index can be constructed using the distribution of the ρ_j values related to the papers of a sample. This is justified by the absence of systematic differences between (i) the average number of authors and (ii) the average percentage of unique authors, between citing and cited papers in a certain field. On the other hand, the analysis confirmed some systematic differences between fields, as regards the average number of authors per paper.
- The *cs*-index is an indicator that, although generally correlated with the *s*-index, can complement it, being only marginally affected by self-citations and citations from recurrent citers.
- Similarly to the *s*-index, the *cs*-index has an immediate meaning and is practical for normalizations aimed at obtaining the so-called size-

independency, thanks to the ratio scale property (Franceschini et al., 2012a). For example, scientific journals with a different number (P) of articles could be easily compared by means of the percentage of "successful" papers, i.e., *cs*-index/P.

- Even if it was not shown directly in this paper, another advantage "inherited" by the *s*-index is that cs-index can be calculated for a set of multidisciplinary articles, thanks to the field-normalization that it achieves at the level of individual paper. For example, the *cs*-index can be used as a proxy for synthesizing the productivity and impact of (i) the whole publication output of scientists involved in multiple disciplines (e.g., mathematicians or computer scientists actively involved in bibliometrics), or (ii) that of entire multidisciplinary research institutions.
- A disadvantage of the *cs*-index is the computational complexity of the cCT_i values. E.g., our data collection and analysis – which was performed by an *ad hoc* application software able to query the WoS database automatically – took about twenty consecutive hours.
- Another potential drawback of *cs*-index is represented by hyperauthorship, which could lead to inflate cCT_i values. A partial solution to this problem is (i) to determine cCT_i by indicators that are insensitive to the right-hand tail of the distribution of ρ_j (e.g., $\widetilde{\rho}$), or (ii) to apply some exclusion criteria, so as to curtail the count of the authors of a certain paper, according to a conventional threshold.

[1] The WoS database configuration included the following resources: Citation Index Expanded (SCI-EXPANDED) from 1970 to present, Social Sciences Citation Index (SSCI) from 1970 to present, Arts & Humanities Citation Index (A&HCI) from 1975 to present, Conference Proceedings Citation Index - Science (CPCI-S) from 1990 to present, Conference Proceedings Citation Index - Social Science & Humanities (CPCI-SSH) from 1990 to present.

References

Ajiferuke, I. & Wolfram, D. (2010). Citer analysis as a measure of research impact: Library and information science as a case study. *Scientometrics*, 83(3), 623–638.

Bornmann, L. (2013). A better alternative to the h index. *Journal of Informetrics*, 7(1), 100.

Braun, T., Glänzel, W. & Schubert, A. (1985). Scientometric Indicators: A 32-country comparative evaluation of publishing performance and citation impact. Philadelphia: World Scientific.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, 52(7), 558–569.

Dieks, D. & Chang, K.H. (1976). Differences in impact of scientific publications: Some indices derived from a citation analysis. *Social Studies of Science*, 6(2), 247–267.

Egghe, L. & Rousseau, R. (1990). Introduction to Informetrics: Quantitative Methods in Library, documentation and Information Science, Amsterdam: Elsevier.

Egghe, L. (2012). A rationale for the relation between the citer h-index and the classical h-index of a researcher, to appear in *Scientometrics*, DOI 10.1007/s11192-012-0770-1.

Franceschini, F., Maisano, D., Perotti A. & Proto A. (2010). Analysis of the *ch*-index: an indicator to evaluate the diffusion of scientific research output by citers, *Scientometrics*, 85(1), 203–217.

Franceschini, F., Galetto, M., Maisano, D. & Mastrogiacomo, L. (2012a). The success-index: an alternative approach to the h-index for evaluating an individual's research output, *Scientometrics*, 92(3), 621-641.

Franceschini, F., Galetto, M., Maisano, D. & Mastrogiacomo, L. (2012b). Further clarifications about the success-index, *Journal of Informetrics*, 6(4): 669–673.

Franceschini, F., Maisano, D. & Mastrogiacomo, L. (2012c). Evaluating research institutions: the potential of the success-index, to appear in *Scientometrics*, DOI 10.1007/s11192-012-0887-2.

Glänzel, W. (2002). Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies, *Library Trends*, 50(3), 461–473.

Glänzel, W. (2011). The application of characteristic scores and scales to the evaluation and ranking of scientific journals, *Journal of Information Science*, 37(1), 40–48.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569–16572.

ISI Web of Knowledge (2012). Essential Science Indicators. http://thomsonreuters.com [Nov. 2012]

Moed, H.F. (2011). The Source-Normalized Impact per Paper (SNIP) is a valid and sophisticated indicator of journal citation impact. *Journal of the American Society for Information Science and Technology*, 62(1), 211–213.

Thomson Reuters (2012) 2011 Journal Citation Reports – Science Edition, www.isiknowledge.com.

Vinkler, P. (2011). Application of the distribution of citations among publications in scientometric evaluations, *Journal of the American Society for Information Science and Technology*, 62(10), 1963–1978.

COLLABORATION IN AFRICA: NETWORKS OR CLUSTERS?

Jonathan Adams(1), Karen Gurney(1), Daniel Hook(2) and Loet Leydesdorff(3)

[1] *jonathan.adams@thomsonreuters.com*
Evidence Thomson Reuters, 103 Clarendon Road, LEEDS LS2 9DF, UK
[2] Symplectic, 10 Crinan St, LONDON N1 9XW, UK
[3]Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands

Abstract
Recent discussion about the increase in international research collaboration suggests a comprehensive global network centred around a group of core countries and driven by generic socio-economic factors where the global system influences all national and institutional outcomes. In counterpoint, we demonstrate that the collaboration pattern for countries in Africa is far from universal. Instead, it exhibits layers of internal clusters and external links that are explained not by monotypic global influences but by regional geography and, perhaps even more strongly, by history, culture and language. Analysis of these bottom-up, subjective, human factors is required in order to provide the fuller explanation useful for policy and management purposes.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction
Wagner & Leydesdorff (2005) argued that patterns in international collaboration in science can be considered as network effects and that only the European FPs noted by Georghiou (op. cit.) mediated relationships at that level. Their global network shares features with other complex adaptive systems in which order emerges from interactions between many agents pursuing self-interested strategies. Adams, Gurney & Marshall (2007) pointed to the intense levels of interactions between leading research economies. Leydesdorff & Wagner (2008) suggested that the global network reinforced a core group of (fourteen) cooperative countries with strong national systems. They argued that peripheral countries could be disadvantaged by increased strength at the core.
Wagner (2008) argues, from complex systems theory, that the self-organizing global system influences all lower systems (Wagner et al, in prep). Here, we accept the meta-pattern but contest the network as a sufficient explanatory model and concur with Georghiou, that there are other agents such as major facilities (e.g. CERN – see King, 2012) and cooperative programmes (e.g. WHO, FAO, climate change) which have been important. In addition, we argue that the effects

of history, culture and language continue to have a profound human influence on collaboration patterns, mediated through personal preference rather than strategic logic (Adams, 2012).

In this paper we illustrate these effects through an analysis of collaboration patterns in Africa (see also Adams, King & Hook, 2010). Africa contains more than 50 nations, hundreds of languages, and a welter of ethnic and cultural diversity. OECD's African Economic Outlook (OECD et al., 2011) sets out in stark detail the challenge for the research base in Africa and the extent to which current global economic problems may make this worse and further compromise the value of the commitment made in by developed nations 2005 to double official development assistance to Africa by 2010. More than half the African nations are off-track or regressing on objectives to achieve universal primary education by 2015. Internet penetration is good only in North Africa, constraining communication and access to knowledge. It needs international research partners.

Is the uniform, generic pattern perceived elsewhere also found in Africa, or does the continent exhibit more subtle influences in its patterns of research collaboration? And, picking up the visualisation methods compared by Leydesdorff et al (in press), how can we best represent what we see?

Methods

We focussed on research publications with one or more addresses for a country within Africa as defined by the UN. We sampled data for the period 2000-2012 (data to current indexing, not year-end). We also collated data on GDP for each country for which publication data were available.

Volume and subject analyses used Thomson Reuters *National Science Indicators*. Collaboration analyses were carried out using *Research Performance Profiles* data in *InCites*™, a web-based platform for research evaluation from Thomson Reuters. Database years were used to delineate years, and only article, note and review document types were considered.

We counted all collaborations between countries represented by co-authorship on the publications we collated. The counts are by paper not by number of researchers. For example, a paper co-authored by two researchers from Ghana, three from Nigeria and one from Kenya counts as a single paper in each country's total and as one link between each pair of countries.

Analysis was extended using *Wolfram Mathematica® 7* to create maps and collaboration diagrams. We also had access to the data collected about 2011-publications by Leydesdorff et al. (in press), and extracted the subset of African countries.

Results

Total research output for Africa increased from 13,271 publications indexed on Thomson Reuters Web of Science in 2000 to over 35,000 publications in 2012

(34,528 catalogued at Dec 15, 2012). For reference purposes, the total output of Africa is about the same as that of the Netherlands.

The percentage of Africa's publications that were substantive research papers (that is to say, articles or reviews) declined from 88% to 82.6%, which reflects an increasing number of proceedings papers and other contributions authored within Africa. (Figure 1)

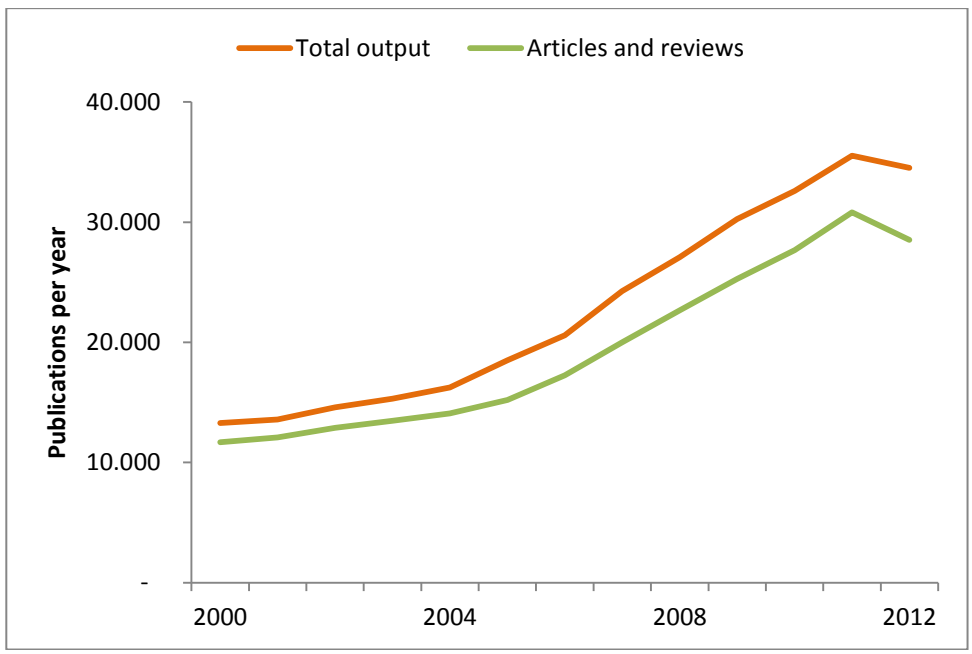


Figure 1 Africa's output of publications indexed on Thomson Reuters *Web of Science* databases between 2000 and 2012

The number of articles and reviews that have been authored wholly within Africa (i.e. that have no collaborative co-author from outside the region) has doubled since 2000 from 6,319 to 12,089. This is a decline as a percentage of total research paper output from 54% to 42%. However, the relative collaborative output of G8 countries rose much faster over the period: collaboration is increasing everywhere. Thus, in fact, the autonomous research output of Africa clearly grew in the last decade and Africa is becoming increasingly self-reliant in this regard.

A breakdown of the figures demonstrates the extent to which each region—and African science as a whole—is dominated by three nations: Egypt in the north, Nigeria in the middle, and South Africa in the south. In this millennium, since 2000, Egypt produced nearly 58,000 publications which was more than twice the total for Tunisia, its next-place and regional neighbour. In west-central Africa, Nigeria's total for the same period was over 20,000, compared to roughly 12,800

for Kenya which is the leading research economy in the east of the continent. South Africa's dominance, as might be expected, is even more pronounced: over 95,000 publications since 2000, compared to the southern region's next-most-prolific nation, Tanzania, which fielded just over 6,300. (Figure 2a, Table 1)

Table 1 Research output and collaboration (all publications on *Web of Science*, 2000-2012) between the USA, UK, France, Saudi Arabia and their most frequent partners in Africa

		USA	*France*	*UK*	*Saudi Arabia*
Africa - Total	*296,351*	*39,292*	*31,421*	*25,753*	*6,285*
South Africa	95,309	14,264	3,801	10,131	
Egypt	57,741	5,900	1,019	2,409	4,939
Kenya	12,769	4,260	460	2,791	18
Uganda	6,317	2,318	231	1,402	11
Nigeria	21,909	1,945	243	1,426	54
Tanzania	6,299	1,693	171	1,625	8
Ghana	4,945	1,159	156	1,003	14
Malawi	2,909	990	124	1,087	
Morocco	17,518	956	5,738	559	186
Ethiopia	5,579	933	218	576	26
Cameroon	5,915	832	1,730	548	
Tunisia	24,724	755	7,400	421	326
Senegal	3,634	573	1,622	275	3
Algeria	14,846	412	5,961	292	367
Gambia	1,294	297	86	857	
Gabon	1,188	241	504	205	

What happens when we break the publication data down by field of research? In our recent Global Research Report on Africa we showed a discernible pattern in Africa's relatively high representation—as a share of world publications—in fields that are relevant to natural resources. The highest percentage of any field, for example, is South Africa's 1.55% share of Plant & Animal Science. Not far behind is the same country's 1.29% share of Environment/Ecology. A review of the more detailed analyses in Thomson Reuters *Essential Science Indicators* shows that many of South Africa's most highly cited papers in this field pertain to climate change and its effects on plant propagation. Following this theme, South Africa's 1.13% share of Geosciences is in keeping with the region's mineral richness.

In short, Africa is a continent abundant in natural resources. How much does Africa itself benefit from those resources? Absolute volume of published papers is one indicator of research activity and—indirectly—of research capacity. It will therefore be obvious that the output of a country reflects how much money is

going in to its research system, and that is likely to be partly dependent on its general economy. Bigger countries with a larger economy should be producing more papers, if they invest at the same level as smaller countries. However, land area, population density and resources vary a great deal. We have compared publications with Gross Domestic Product (GDP) for each country, reasoning that proportionate investment in the knowledge economy is a good index of a government's commitment to maximize the longer term benefit of resource development and exploitation for the general wealth of its people.

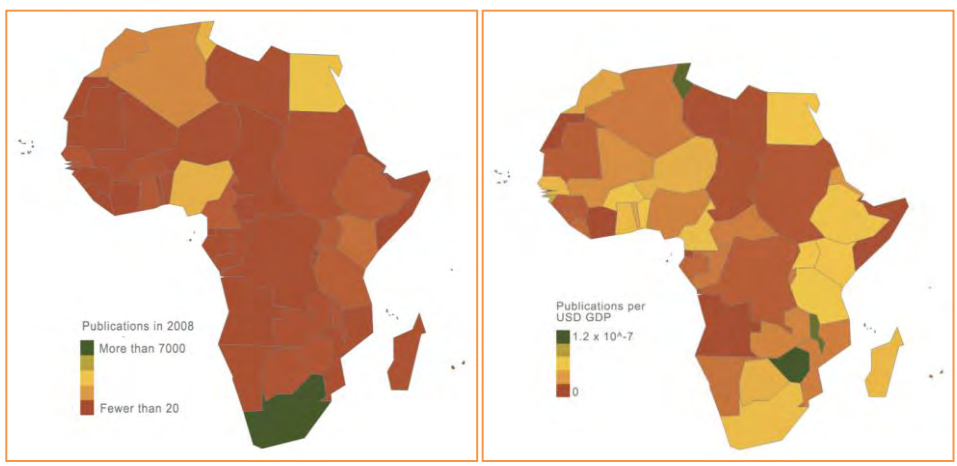


Figure 2. Output per country in 2008 as total volume (Figure 2a) and as volume/GDP (Figure 2b). South Africa is absolutely the most productive country. Zimbabwe appears to be relatively productive but this is an anomaly due to very low recent GDP and a strong historical base. Tunisia is relatively the more productive on current performance.

The leading countries by output are South Africa, Egypt, Nigeria, Tunisia, Algeria and Kenya (Figure 2a). Four of these (South Africa, Egypt, Nigeria and Algeria) are also leading countries in terms of GDP while Kenya and Tunisia fall in a lower GDP tier. Indexing output against GDP (Figure 2b) provides further interpretation. Zimbabwe is highlighted as relatively the most productive country in terms of publications per unit GDP but this is anomalous because it retains its legacy research base despite a collapsing economy and very low current GDP. The real leaders are Tunisia and Malawi with very different economic bases but strong relative productivity in both cases. South Africa, Kenya and Egypt all have significant relative productivity, as do a number of other countries in East Africa (Ethiopia, Uganda, Tanzania) and West Africa. (Cameroon, Ghana).
It is clear, however, that despite Nigeria's high volume output it is not producing as much research as would be expected given the size of its economy. The value of its resources is not yet being felt in its knowledge base. In fact, the same research productivity gap between potential and actual investment applies to

several other countries. This is an area where Africa is not yet benefitting from the best use of its own natural resources.

Africa's research can be boosted by collaborative international partnerships. The countries collaborating most frequently with partners in Africa are the USA (39,292 papers between 2000 and 2012), France (31,421), the UK (25,753), Germany (13,879) and Canada (7,604). This looks like a roll-call of 'the usual suspects' among major research producers. It is therefore worth noting that Saudi Arabia collaborated on 6,285 papers, albeit almost entirely with countries in North Africa of which Egypt (4,939 joint publications) was the pivotal link. Ethiopia's research base is distinctive in being substantial, growing and yet almost entirely domestic. The most substantial links between countries in Africa and the USA, UK, France, Saudi Arabia are summarised in Table 1.

The research axis between Egypt, Saudi Arabia and the USA is an instructive example of new and changing collaboration patterns. The numbers of papers co-authored between Egypt and the USA has grown but has remained around 10% of Egyptian output since 1995. The numbers of papers co-authored between Egypt and Saudi Arabia has been much smaller historically but reached 100 (4% of Egypt's output) in 2002 and exceeded 1000 (15%) in 2011. This is regionally, not globally, driven: only a small fraction of these papers also have the USA as a co-author. (Table 2)

Table 2 Growth of Egypt's research output and its collaboration with the USA and with Saudi Arabia over thirty years from 2000-2012 (part year). Egypt has increased collaboration with Saudi Arabia and little of this is driven by its prior links with the USA.

Year	*Egypt total*	*Egypt + USA*	*USA as % Egypt*	*Triple co-authors*	*Saudi as % Egypt*	*Egypt + Saudi*
2000	2,577	286	11	2	4	95
2001	2,707	227	8	3	3	94
2002	2,894	295	10		4	115
2003	3,238	312	10	7	6	181
2004	3,212	318	10	4	5	169
2005	3,338	326	10	3	5	164
2006	3,847	358	9	6	5	190
2007	4,280	424	10	8	5	199
2008	4,710	439	9	15	6	261
2009	5,725	597	10	20	7	416
2010	6,281	708	11	33	10	614
2011	7,416	823	11	55	15	1,093
2012 (part)	4,386	428	10	47	19	832

France also has a niche relationship with Africa. It is unusual in studies of international collaboration to find it high in any ranking, and here to be 2nd behind the USA, ahead of the UK and with much more than twice the collaboration links

of Germany. Among the 31,421 total co-authorships by partner then we find that Tunisia (Table 1: 7,400 23.6%) leads with Algeria (19%), Morocco (18.3%) and then Cameroon (5.5%). France is focussed on a small set of countries just across the Mediterranean in North Africa.
The USA and the UK, by contrast, collaborate diversely with South Africa, Kenya, Egypt, Nigeria and others (Table 1). The UK has much greater collaboration with specific countries, such as the Gambia and Malawi, than any other partner. Clearly, no collaboration pattern in Africa is general or uniform.

For each of six key research economies in Africa we have analyzed collaborative research links by collating co-authorships with other countries and analyzing collaboration with the USA and the UK as the most frequent partner for most countries, and three other frequent partners. (Figure 3)

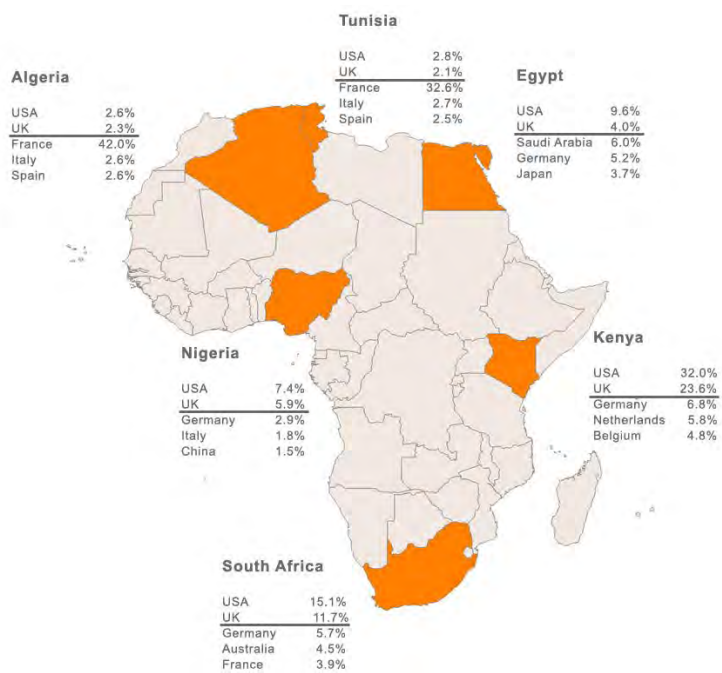


Figure 3 Most frequent intercontinental research collaborations for six key African research economies

How can we best visualise research co-publication within Africa? First, to create a simplified picture useful as an indication of major links for policy purposes, we used a threshold to clarify where relatively strong and persistent collaborations occurred. The threshold was set at a minimum of five papers per year, or 25

papers in total over the recent five-year period. This meant that some countries did not appear at all in the analysis because they had too low a level of recent collaboration. We then used a grouping algorithm to associate the countries around the rim of the wheel until groups with strongest cross-links were placed close together. (Figure 4)
Second, we created a more complete but necessarily more complex picture of the entire Africa network 2011 using VOSViewer. (Figure 5)

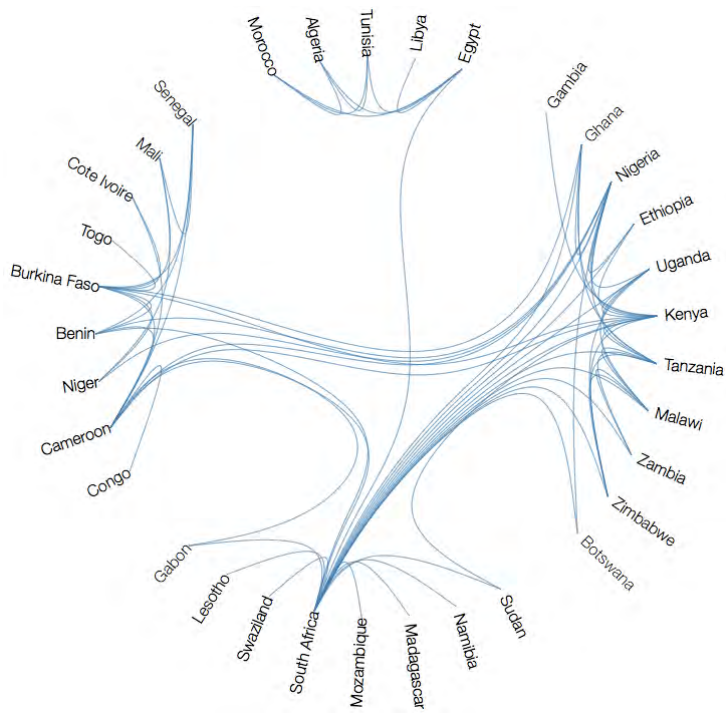


Figure 4: Collaboration between countries within Africa. This dependency graph uses *Wolfram Mathematica® 7 to* provide a new visual interpretation of collaboration, by paper not by number of researchers, and reveals clusters of countries with strong and persistent partnerships. Links displayed between each country meet a threshold of five publications per year for a continuous period of five years.

Discussion

This analysis presents a complex picture of diverse research collaboration links, internationally (Table 1, Figure 3) and within Africa (Figure 4, Figure 5). It is difficult to argue that these outcomes are a response to a common global network phenomenon rather than local, cultural and historical factors that play into research opportunities and create the highly individualistic and specific African outcome. We do not disagree with the concept that international research collaboration is a common phenomenon but we do believe in the need to

determine the bottom-up regional and local factors that properly explain complex outcomes departing from a notional top-down global template. Only by understanding this detail will research performance analysis engage with the theory and practice of research policy and management.

The research output of Africa is growing although remains small compared to established economies (Figure 1). Africa has enormous natural resources but, while there is a broad relationship with investment as GDP (Figure 2), some richer countries have yet to commit to substantial investment in their knowledge economy. It is therefore to be anticipated that further research development will continue to benefit from extensive external support and collaboration.

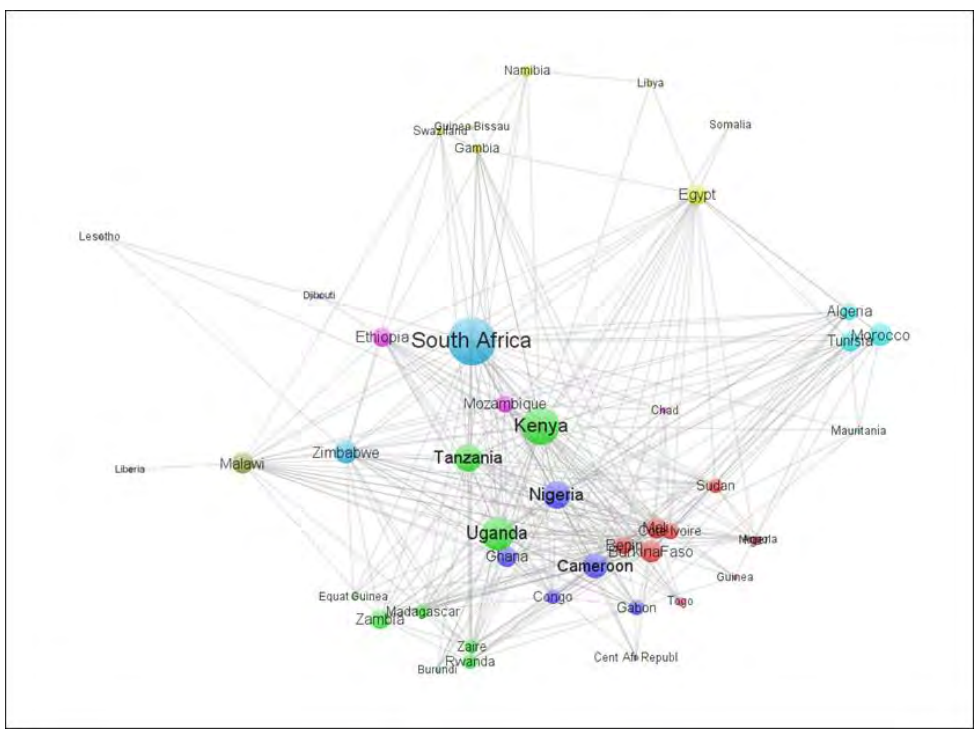


Figure 5: Coauthorship relations between 46 countries within Africa in 2011. VOSViewer was used for the clustering and mapping. This map highlights the pivotal role of South Africa, shows the separated cluster of North Africa with Egypt as an outlier due to its wider attachments, and recognizes not only the East Africa group but also the development of two distinct groups in West Africa. The map can be web-started for further exploration at http://www.vosviewer.com/vosviewer.php?map=http://www.leydesdorff.net/intcoll/afr_map.txt&network=http://www.leydesdorff.net/intcoll/afr_netw.txt&label_size=1.25&n_lines=1000.

External collaborative links vary significantly by country. France is the second most frequent collaborator with Africa, after the USA, with concentrated links to

North-West Africa and to central West Africa. It is interesting to note that, after normalization for size, Leydesdorff & Wagner (2008: 321) found France as highest on betweenness centrality because of its intermediating function with the EU networks. The UK is the most frequent collaborator with other African countries, such as Malawi and Gambia (Table 1). These links are not driven by global phenomena but by local historical and cultural factors and by targeted international cooperative health and food programmes. Many links are mediated through cooperative health and agricultural programs. Gambia is the site for long-term research into tropical diseases for the UK's Medical Research Council (Adams, Gurney & Pendlebury, 2012) which also works in Uganda. The Wellcome Foundation has similar, major research investments in Kenya and Malawi. A significant intellectual benefit is thus secured outside Africa.

Another exceptionally strong link is that between Egypt and Saudi Arabia, which is not mediated by a third party such as the USA (Table 2) but through their axis in supporting regional growth in research capacity in the Arabian Middle East. (Figure 3; Adams et al, 2011)

How can we create a picture of Africa's research network that would be helpful for policy engagement? If we apply a threshold on the strength of interaction we find no single network within Africa. Interface with African countries requires awareness that collaboration is driven partly by geography but also by shared culture and—very strongly—by language. (Figure 4)

- There is a marked interaction between the countries in North Africa which share both language and culture and are also relatively prolific. Thus, this network is probably the strongest group overall since it links countries which are individually research active across multiple fields. The group does little research with the rest of Africa, however, other than through an Egypt-South Africa link.
- A West Africa group (Benin-Togo) pivots around Cameroon, a relatively research productive country. The common factor within this group is almost certainly their common use of French as the cross-national business language.
- Language also gives us the clue to the large group which includes Kenya and geographical neighbours in East Africa but also includes Nigeria, Ghana and Gambia. Those countries appear to have English as a common language or have had a strong Anglophone influence.
- The Southern African Development Community (SADC) does not emerge as a research network since it is split between that group linked to Kenya and Nigeria and a second group most closely linked to South Africa, but which also includes Sudan and Gabon. The overall collaboration network, to the extent that one exists at all, is dependent on a small number of key players linking these regional and cultural groupings.

The simplified collaboration cartwheel of Figure 4 is useful for managers and planners. It is expanded and developed in Figure 5 into a complementary visualisation where completeness adds complexity requiring additional interpretation. It is therefore of greater value for the analyst. The map highlights

the pivotal role of South Africa: the research hub in every sense. The map shows that Egypt is not embedded in the separated cluster of North Africa but is an outlier due to its wider attachments. There is a strong East Africa group, as in Figure 4, but there is also the development of two distinct groups in West Africa.

In Figure 3, there is a striking difference between the countries pulled out in North Africa (Figure 5) and those in other regions. Globally, the most frequent collaborative partner is the USA. Often this is a consequence of researchers who have studied in the USA maintaining links with those research groups when they return home. The UK and Germany are the other common partners to the countries featured here and France has a major role. This is the influence of the global network (pace Wagner *op. cit.*): between them the USA, UK, Germany and France have authors on half the world's research papers every year.

Nigeria sits at a research crossroads between East, West and South Africa. Despite its disappointing level of research investment, it has an important connecting role. Not only is it a part of the Anglophone collaborative network but it also has significant—albeit weaker—connections with its West African neighbours, and it connects strongly to South Africa. South Africa is a similarly strong node with a spread of links into other groups. These two, with Kenya, create strong cross-continent links and are key nodes into global networks.

China and Brazil's rapidly expanding research bases collaborate only weak with Africa. Nigeria's global reach is marked by some collaboration with China. It is theoretically well-positioned to extend its links westwards and partner with the emerging Brazilian research base. It could thus serve as a key doorway into both the West African and the Anglophone African research base for some of the exciting research which is now appearing in Asia and Latin America. But it has yet to realise this opportunity.

References

Adams, J. (2012). Collaborations: the rise of research networks. Nature, 490, 335-336

Adams J, Gurney, K. A. & Pendlebury, D. (2012). Neglected Tropical Diseases, pp 1-16. Global Research Report, Thomson Reuters, Philadelphia. ISBN 1-904431-31-3

Adams, J., King, C., Pendlebury, D., Hook, D., Wilsdon, J. & Zewail, A. (2011). Exploring the changing landscape of Arabian, Persian and Turkish research, pp 1-8. Global Research Report, Thomson Reuters. ISBN 1-904431-27-5

Adams, J., King, C. & Hook, D. (2010). Africa, pp 1-9. Global Research Report, Thomson Reuters. ISBN 1-904431-25-9

Adams, J., Gurney, K. & Marshall, S. (2007). Patterns of international collaboration for the UK and leading partners. Report commissioned by the Office of Science and Innovation, 27 pp. Department of Innovation, Universities and Skills, London. http://www.berr.gov.uk/files/file40396.pdf

Frame, J.D. & Carpenter, M.P. (1979). International research collaboration. Social studies of Science, 9, 481-497.

Georghiou, L. (1998). Global cooperation in research. Research Policy, 27, 611-626.
Greene, M. (2007). The demise of the lone author. Nature, 450, 1165.
King, C. (2012). Multiauthor papers: onward and upward. Science Watch, 23, 1-2
Leydesdorff, L. & Wagner, C.S. (2008). International collaboration in science and the formation of a core group. Journal of Informetrics, 2(4), 317-325.
Leydesdorff, L., Wagner, C.S., Park, H. W., & Adams, J. (in press). International Collaboration in Science: the Global Map and the Network. *El Profesional de la Información*.
OECD, African Development Bank, United Nations Economic Commission for Africa and United Nations Development Programme. (2011). African Economic Outlook 2011: Africa and its emerging partners. OECD Publishing, Paris. ISBN 9789264111752
Persson, O., Glanzel, W. & Danell, R. (2004). Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies. Scientometrics, 60, 421-432.
Wagner, C. S. (2008). The New Invisible College. Washington, DC: Brookings Press.
Wagner, C.S. & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. Research Policy, 34, 1608-1618.
Wagner, C.S., Leydesdorff, L. & Adams, J. (in prep.) Policy implications of the global network of science.

COLLABORATIVE INNOVATIVE NETWORKS: INFLUENCE AND PERFORMANCE

Alireza Abbasi[1] and Kon Shing Kenneth Chung[2]

[1] *alireza.abbasi@unsw.edu.au*
School of Engineering and IT, University of New South Wales Canberra, ACT 2600 (Australia)

[2] *kenneth.chung@sydney.edu.au*
Complex Systems Research Group, Project Management Program, Faculty of Engineering and IT, University of Sydney, NSW 2006 (Australia)

Abstract
With an increase in studies of co-authorship and citations count very few have examined the role of influence of authors in their collaboration network and its effect on performance. In this study, we describe the joint effort of academic publishing as being conducted within a collaborative innovative network, where new knowledge is produced as a result. By assuming that the collaborative process involves social influence embedded within relationships and network structures amongst direct and indirect co-authors, we examine whether such influence is directly associated an author's performance, based on the number of citations. Using a co-authorship and citation data, we use social network analysis to propose a combined degree and centrality metrics to measure social influence among connected scholars in a co-authorship network. We then use Spearman's correlation rank test to examine the association between social influence measures and (citation-based) performance. Results suggest that research performance of authors is positively correlated with their social influence measures. Furthermore, results suggest that our combined degree and centrality measures are statistically very significant and also have a higher positive correlation index with performance than the correlation coefficient between performance and standard centrality alone.

Introduction
As with most large organizations, performance of individuals and teams is measured through a set of metrics that pertain to task and contextual performance. Similarly in academia, scholars and scientists are evaluated based on their academic performance (e.g., research productivity, teaching evaluations, governance capabilities, achieved grants). Such evaluation of academics is not only needed for faculty recruitment, but also for governmental funding allocation and for achieving a high reputation within the research community. The reputation of research organizations indirectly affects the society's welfare, since a high reputation attracts foreign purchases, foreign investments, and highly qualified students from around the world. Most recently, the Australian government's Excellence for Research Excellence scheme has ranked its nation's

universities based on research metrics such as publication output, number of grants and research collaborations in comparison to world standard (Hare, 2011).
The implication of such ranking provides basis and justification for federal funding thus encouraging high research standards and goals. Therefore, on a global level, with respect to governmental funding (i.e., the allocation of funding for a specific project to a scientific research group) and university strategy, it is important to identify key scholars, collaboration areas and research strengths within universities with the aim of maximizing research output, cost optimization, and resource utilization. Furthermore, at a micro level, as universities shift its key focus on research activities and as academics are being asked to 'do more with less', it is useful to understand how research performance is impacted using a holistic view. Thus, in all these cases, the pressing task of how can research productive scientists be identified, clustered, and configured for optimal research synergies needs to be examined carefully (Jiang, 2008). To assess the performance of scholars, many studies suggest quantifying scholars' publication activities as a good measure for the performance of scholars. Further researchers showed the number of citations a publication receives qualifies the quantity of publications (Hirsch, 2005; Lehmann, Jackson, & Lautrup, 2006) and is a good metrics for measuring the combination of quantity and quality of research.
Since individuals have limited capacity to acquire and use knowledge, their interactions with others are necessary for knowledge creation (Demsetz, 1991) through publication productivity. Therefore, many scientific outputs are a result of group work and most research projects are too large for an individual researcher to undertake. Thus, having researchers with different skills, experience and knowledge (in addition to basic shared understanding of each other's knowledge) in a group work is needed (McFadyen, Semadeni, & Cannella, 2009) as diversity of members facilitates the integration of expertise, contribute to the successful projects' implementation and accelerate cycle time for new product development (Cummings, 2004; Eisenhardt & Tabrizi, 1995; Griffin & Hauser, 1992; Pinto, Pinto, & Prescott, 1993). In many ways, Gloor's (Gloor, Paasivaara, Schoder, & Willems, 2006) concept of a collaborative innovative network (CoINs) captures such ideas postulated above.
Recently Abbasi et al. (2010) highlighted the importance of scholars' collaboration activity and proposed a measure (Rc-index) to quantify researcher collaboration activity. In addition, using social network analysis methods and metrics, several studies have shown the applicability of centrality measures for co-authorship networks and how centrality measures are useful to reflect the performance of scholars based on their social position and influence within their collaboration network (A. Abbasi, Chung, & Hossain, 2012; Takeda, Truex III, & Cuellar, 2010; Yan & Ding, 2009; Zhuge & Zhang, 2010). Here also we attempt to assert the importance of co-authors' role and position in their collaboration network. In particular, we study co-authorship network, performance measure of scholars and actors' centrality measures, and test the correlation between an actor performance measure and her social influence metrics considering her co-authors'

centrality measures. In brief, the motivating question for our study is: how social structure of scholars influences their co-authors' performance?
In the following sections, we review existing literature on using social influence theory and social network analysis in analyzing scientific collaboration networks. In Section 3, we explain about our data collection and the measures we proposed to quantify scholar's social influence. Finally, the result of testing association between scholars' social influence measures and their performance is shown in following section. We conclude our paper by discussing about our findings and research limitations.

Social Influence Theory

Social influence is the change in behavior that one person causes in another, intentionally or unintentionally, as a result of the way the changed person perceives themselves in relationship to the influencer, other people and society in general. Thus, the theory of social influence states that behavior is intentionally or unintentionally influenced by others (Strang, 2000). Due to social influence process people's behaviors adapt to those they interact with more (Crandall, Cosley, Huttenlocher, Kleinberg, & Suri, 2008; Friedkin, 1998).
We can see two of three different categories of social influence introduced by (Strang, 2000) can be seen in scholar's social interactions: Social Conformity which is changing how you behave to be more like others. This plays to belonging and esteem needs as we seek the approval and friendship of others. Conformity can run very deep, as we will even change our beliefs and values to be like those of our peers and admired superiors; and Social Compliance which is where a person does something that they are asked to do by another. They may choose to comply or not to comply, although the thoughts of social reward and punishment may lead them to compliance when they really do not want to comply.
Friedkin (1998) developed a formal theoretical approach to influence as a network process and posed the Durkheimian question of how interaction can generate consensus and permit coordination within complex social settings. Friedkin (1998) treated influence as proportional to the strength of direct and short indirect ties linking actors. "His refinement of this balance theoretic process produces perhaps the most thoroughly developed analysis of cohesion within contemporary network analysis" (Strang, 2000).
Scientific collaboration is defined as "interaction taking place within a social context among two or more scientists that facilitates the sharing of meaning and completion of tasks with respect to a mutually shared, super-ordinate goal" (Sonnenwald, 2007). An important result of scientific collaborations is the creation of new scientific knowledge, including new research questions, new research proposals, new theories, and new publications (Stokols, Harvey, Gress, Fuqua, & Phillips, 2005).
Now, in the process of interaction among scholars to create a new knowledge, some scholars are more persuasive than others in terms of influencing others as to

the validity of their ideas (Takeda et al., 2010). This phenomena of being influenced by others (often co-authors as direct contacts in scientific interaction process during development of a new knowledge), we term Social Influence. Since other studies showed the position and role of scholars in their collaboration networks reflects their skills and performance (Abbasi, Altmann, & Hossain, 2011; Abbasi et al., 2012), we quantify a scholars' social influence by considering her co-authors' position in the co-authorship network, which can be shown through their centrality measures in the network.

Social Network Analysis

Social network analysis (SNA) is the mapping and measuring of relationships and flows between nodes of the social network. SNA provides both a visual and a mathematical analysis of human-influenced relationships. The social environment can be expressed as patterns or regularities in relationships among interacting units (Wasserman & Faust, 1994). Each social network can be represented as a graph made of nodes (e.g. individuals, organizations, information) that are tied by one or more specific types of relations, such as financial exchange, friends, trade, and Web links. A link between any two nodes exists, if a relationship between those nodes exists. If the nodes represent people, a link means that those two people know each other in some way.

A method used to understand networks and their participants is to evaluate the location of actors in the network. Measures of SNA, such as network centrality, have the potential to unfold existing informal network patterns and behavior that are not noticed before (Brandes & Fleischer, 2005). Measuring the network location is about determining the centrality of an actor. A point can be central locally or globally. A point is locally central if it has a large neighborhood of direct contacts (actors). It is important to recognize that it doesn't mean there it would be just a unique central point in the network. On the other hand, a point is globally central if it has a position of strategic significance in the overall structure of the network (Scott, 1991).

These measures help determine the importance of a node in the network. Freeman (1979) defined centrality in terms of node degree centrality, betweenness centrality, and closeness, each having important implications on outcomes and processes. While these defined measures are widely used to investigate the role and importance of networks but each one is useful based on especial cases: (i) degree centrality is an indicator of an actor's activity popularity; (ii) closeness centrality indicates the extent to which an actor is close to all others in the network and shows accessibility of an actor and its independence; and, (iii) betweenness is an indicator of an actor's potential control of communication within the network and highlights brokerage (gate keeping) behavior of an actor.

Degree Centrality

The degree is simply the number of other points connected directly to a point. Necessarily, a central point is not physically in the center of the network. As

degree of a point is calculated in terms of the number of its adjacent points, the degree can be regarded as a measure of local centrality (Scott, 1991). Thus, a person (point) in a position with having high degree centrality can influence the group by withholding or distorting information in transmission (Bavelas, 1948; Freeman, 1979). So, degree centrality is an indicator of an actor's popularity and activeness.

Closeness Centrality

Freeman (1979, 1980) proposed closeness as a measure of global centrality in terms of the distance among various points. Sabidussi (1966) had been used the same concept in his work as 'sum distance', the sum of the 'geodesic' distances (the shortest path between any particular pair of points in a network) to all other points in the network. A point is globally central if it lies at the shortest distance from many other points which means it is 'close' to many of the other points in the network. So, simply by calculating the sum of distances of a point to others we will have 'farness', how far the point is from other points and then we need to use the inverse of farness as a measure of closeness. A point in the nearest position on average, to all others, can most efficiently obtain information.

Betweenness Centrality

Freeman (1979) yet proposed another concept of point centrality which measures the number of times a particular point (node) lies 'between' the various other points in the network (graph). Betweenness centrality is defined more precisely as "the number of shortest paths (between all pairs of points) that pass through a given point" (Borgatti, 1995). Betweenness is an indicator of the potential of an actor (or point) which plays the part of a 'broker' or 'gatekeeper' which can most frequently control information flow (communication) in the network.

Data and Methods

Based on the co-authorships of publications of scholars, we construct the research collaboration network of scholars. Nodes of the research collaboration network represent scholars. A link between two nodes represents a publication co-authorship relationship between those scholars.

Data

For this study, we collected data on co-authorship (collaboration) and citation from five north-American based information schools (iSchools): University of Pittsburgh, University of Berkeley, University of Maryland, University of Michigan, and Syracuse University. These schools were chosen primarily because they offer similar programs in the area of information management and systems and, because of the fact that the topic of these schools is new within the university landscape.

The data sources used are school reports, which include the list of publications of researchers, DBLP, Google Scholar and the ACM portal. Citation data has been taken from Google Scholar and the ACM Portal. The relationships (e.g., co-authorships) between researchers were extracted and stored a database.
For our analysis, we followed the Google Scholars approach and did not differentiate between the different types of publications (i.e., proceedings of local conferences, proceedings of international conferences, journals, books, and presentations were weighted equally). Our data covered a period of five years (2001 to 2005), except for the University of Maryland iSchool, which had no data for the year 2002 in their report. To resolve this issue, we substituted the missing data with data of the year 2006. As we neither apply longitudinal analysis nor comparing schools, this will not affect our results. After the cleaning of the publication data of the five iSchools, 2139 publications, 1806 authors, and 5310 co-authorships were finally available for our analysis.

Measures

Scholars' Performance

To assess the performance of scholars, many studies suggest quantifying scholars' publication activities as a good measure for the performance of scholars. The general idea is that a researcher gets a high visibility in the research community, if the researcher publishes and her publications get cited. The number of citations qualifies the quantity of publications (Lehmann et al., 2006). Hirsch introduced the h-index as a simple measure that combines in a simple way the quantity of publications and the quality of publications (i.e., number of citations) (Hirsch, 2005). A scholar with an index of h has published h papers, which have been cited by others at least h times (Hirsch, 2005). Furthermore, the h-index became also the basis for a wide range of new measures (Altmann, Abbasi, & Hwang, 2009; Batista, Campiteli, & Kinouchi, 2006; Egghe, 2006; Jin, 2006; Sidiropoulos, Katsaros, & Manolopoulos, 2007; Tol, 2008).
Although there is considerable debate on the reliability of the h-index and its variants (Haque & Ginsparg, 2009) the h-index is still widely used world-wide amongst academics. While the reliability of the measure is not the subject of this paper per se, it does provide at least an empirical metric so as to gauge a researcher's prolificacy. Thus, we will consider h-index as a citation-based surrogate measure as proxy for performance of research scholars.

Measuring Social Influence

In this study, in order to quantify to what extend a scholar is influenced by their co-authors, we propose new measures which consider the centrality measures (i.e., degree, closeness and betweenness) of the co-authors of the scholar. Thus, we propose three different measures which the basic definition is the same but the difference is on considering each centrality measure separately.

To define social influence measures, we consider a co-authorship network having centrality measures of each actor (scholar) as the weight (or attribute) of the actor and the strength (weight) of the links among each pair of actors (co-authors), which is the frequency of joint publications. Then, we define social influence of an actor as sum of each centrality measures of all direct actors (co-authors) similar to the measures define in (Abbasi & Hossain, 2013). To have generalize measures, considering weighted networks which their links have different strengths, we can extend definitions by considering the weight of the links. Thus, for instance, the social influence measure base on the degree centrality of co-authors of an actor a, SID(a), can be defined as sum of degree centrality of each co-author multiply by the weight of the link between actor (a) and the co-author. The three measure can be shows as below where n is the degree of actor a (number of direct neighbors of actor a) and w(a,i) is the weight of the link between actor a and its neighbors i. $C_D(i)$, $C_C(i)$ and $C_B(i)$ indicate the degree centrality, closeness centrality and betweenness centrality of the co-authors respectively.

$$SI_D(a) = \sum_{i=1}^{n} [\, w(a,i) * C_D(i)]$$

$$SI_C(a) = \sum_{i=1}^{n} [\, w(a,i) * C_C(i)]$$

$$SI_B(a) = \sum_{i=1}^{n} [\, w(a,i) * C_B(i)]$$

Thus, SID measure indicates the actors (scholar) who are connected better to more actors and it reflects the theory that connecting to more powerful actors will give you more power. So, SID centrality indicates actors' power and influence on transmitting and controlling information. It indicates the popularity of an actor based on popularity of its direct neighbors. SIC measure indicates not only an actors' power and influence on transmitting and controlling information but also efficiency for communication with other or efficiency in spearing information within the network. It indicates popularity and accessibility of an actor simultaneously. Also, SIB measure indicates not only an actors' power and influence on transmitting and controlling information but also potential control of communication and information flow within the network. It shows popularity and brokerage attitude of an actor in the network simultaneously.

Analysis and Results

The result of Spearman correlation rank test between new proposed social influence measure and scholars' performance (e.g., citations count, h-index) has been shown in Table 1. In addition the correlation test between standard centrality

measures and scholars' performance measure has been shown for comparison. As it shows almost all measures (new proposed measures of influence and standard centrality measures) are significantly correlated to performance measure except for closeness centrality which has weak or not significant correlations with performance measures.
As shown in Table 1, the new proposed measures show higher correlation coefficients rather than standard centrality measures. It highlights the importance of the role and position of direct contact as an influencer to the actors' performance.

Table 1. Spearman correlation rank test between scholars' influence measure, centrality measures and their performance

Centrality Measures (N=1806)	*Scholars' Performance Measures*	
	citations count	*h-index*
C_D	.332 **	.311 **
C_C	- .012	.052 *
C_B	.388 **	.501 **
SID	.394 **	.426 **
SIC	.385 **	.432 **
SIB	.304 **	.503 **

*. Correlation is significant at the .05 level (2-tailed).

**. Correlation is significant at the .01 level (2-tailed).

Another outcome of this result is that new proposed measure are different from eigenvector centrality and to support this we also applied non-parametric independent t-test (Mann-Whitney U test) to compare the distribution of eigenvector centrality measure between two groups (lower than mean of h-index and above mean) and it was not significant while the t-test was significant for new proposed measures. So, this also supports that new centrality measures are different from eigenvector centrality.
Having more central neighbors will also lead to being more central. Thus, we are going one step more and will find the actors who are central themselves and also connected to direct central actors. These kinds of actors have special and strategically positions which can control the network.

Conclusion

In this paper, we proposed new measures (i.e., SID, SIC and SIB) to quantify social influence among actors in a network. Our analysis showed that they are good indicators of the importance of an actor in a social network by considering standard centrality measures: degree of each node with degree, closeness and betweenness of its direct contacts. Therefore, they are potentially good extensions of standard centrality measures.

SID centrality reflects the extent to which actors have more direct connection to those actors who themselves have high number of connections to others. So, it is an indicator of power and influence of an actor's ability to control communication and information.
SIC is an indicator of the extent to which actors have more direct connection to the actors who are more close to all other actors – this shows how the actor, on an aggregate level, is close to all. So, it reflects at the same time how the actor is popular and active in communication (due to having high degree) and also efficient in spreading information in less time and cost (as it is close to all other actors).
SIB reflects the extent to which actors who have more direct connection to the actors who are more frequently positioned in a path among other actors. So, it is an indicator of power and influence on transmitting and controlling information but also potential control of communication and information flow within the network.
This has implications for academics to focus on diverse array of professional connections when it comes to co-authorship as social influence plays a crucial and conducive role towards performance in collaborative innovation.

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5(4), 594-607.

Abbasi, A., Altmann, J., & Hwang, J. (2010). Evaluating scholars based on their academic collaboration activities: two indices, the RC-index and the CC-index, for quantifying collaboration activities of researchers and scientific communities. *Scientometrics*, 83(1), 1-13.

Abbasi, A., Chung, K. S. K., & Hossain, L. (2012). Egocentric analysis of co-authorship network structure, position and performance. *Information Processing & Management*, 48(4), 671-679.

Abbasi, A., & Hossain, L. (2013). Hybrid centrality measures for binary and weighted networks. In R. Menezes, A. Evsukoff & M. C. González (Eds.), *Complex networks* (Vol. 424, pp. 1-7). Berlin / Heidelberg: Springer.

Altmann, J., Abbasi, A., & Hwang, J. (2009). Evaluating the productivity of researchers and their communities: The RP-index and the CP-index. *International Journal of Computer Science and Applications*, 6(2), 104–118.

Batista, P., Campiteli, M., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1), 179-189.

Bavelas, A. (1948). A mathematical model for group structures. *Human organization*, 7(3), 16-30.

Borgatti, S. (1995). Centrality and AIDS. *Connections*, 18(1), 112-114.

Brandes, U., & Fleischer, D. (2005). *Centrality measures based on current flow*.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Suri, S. (2008). *Feedback effects between similarity and social influence in online communities*.

Cummings, J. N. (2004). Work groups, structural diversity, and knowledge sharing in a global organization. *Management Science*, 50(3), 352-364.

Demsetz, H. (1991). The Theory of the firm revisited. In O. E. Williamson, S. G. Winter & R. H. Coase (Eds.), *The Nature of the Firm: Origins, Evolution, and Development*. (pp. 159–179). New York: Oxford University Press.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131-152.

Eisenhardt, K., & Tabrizi, B. N. (1995). Accelerating Adaptive Processes: Product Innovation in the Global Computer Industry. *Administrative Science Quarterly*, 40(1).

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks*, 1(3), 215-239.

Freeman, L. C. (1980). The gatekeeper, pair-dependency and structural centrality. *Quality and Quantity*, 14(4), 585-592.

Friedkin, N. E. (1998). *A structural theory of social influence*. Cambridge: Cambridge Univ Press.

Gloor, P., Paasivaara, M., Schoder, D., & Willems, P. (2006). Correlating performance with social network structure through teaching social network analysis. *Network-Centric Collaboration and Supporting Frameworks*, 265-272.

Griffin, A., & Hauser, J. R. (1992). Patterns of communication among marketing, engineering and manufacturing-a comparison between two new product teams. *Management Science*, 38(3), 360-373.

Haque, A., & Ginsparg, P. (2009). Positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science and Technology*, 60(11), 2203-2218.

Hare, J. (2011, 1st February 2010). Elite Eight Head University Research Ratings, *The Australian*. Retrieved from http://www.theaustralian.com.au/higher-education/elite-eight-head-university-research-ratings/story-e6frgcjx-1225997293930

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569.

Jiang, Y. (2008). Locating active actors in the scientific collaboration communities based on interaction topology analyses. *Scientometrics*, 74(3), 471-482.

Jin, B. (2006). H-index: an evaluation indicator proposed by scientist. *Science Focus*, 1(1), 8-9.

Lehmann, S., Jackson, A., & Lautrup, B. (2006). Measures for measures. *Nature*, 444(7122), 1003-1004.

McFadyen, M., Semadeni, M., & Cannella, A. A. (2009). Value of strong ties to disconnected others: Examining knowledge creation in biomedicine. *Organization Science*, 20(3), 552-564.
Pinto, M. B., Pinto, J. K., & Prescott, J. E. (1993). Antecedents and consequences of project team cross-functional cooperation. *Management Science*, 39(10), 1281-1297.
Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4), 581-603.
Scott, J. (1991). *Social network analysis: a handbook*.: Sage.
Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2), 253-280.
Sonnenwald, D. (2007). Scientific collaboration: a synthesis of challenges and strategies. *Annual review of information science and technology*, 41, 643-681.
Stokols, D., Harvey, R., Gress, J., Fuqua, J., & Phillips, K. (2005). In vivo studies of transdisciplinary scientific collaboration: Lessons learned and implications for active living research. *American Journal of Preventive Medicine*, 28(2), 202-213.
Strang, D. (2000). Review: A Structural Theory of Social Influence (Vol. 45, pp. 162-164): JSTOR.
Takeda, H., Truex III, D., & Cuellar, M. (2010). Evaluating Scholarly Influence Through Social Network Analysis: the Next Step in Evaluating Scholarly Influence. *AMCIS 2010 Proceedings*, 573.
Tol, R. (2008). A rational, successive g-index applied to economics departments in Ireland. *Journal of Informetrics*, 2(2), 149-155.
Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge; New York: Cambridge Univ Press.
Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10), 2107-2118.
Zhuge, H., & Zhang, J. (2010). Topological centrality and its e-science applications. *Journal of the American Society for Information Science and Technology*, 61(9), 1824–1841.

COMPARATIVE STUDY ON STRUCTURE AND CORRELATION AMONG BIBLIOMETRICS CO-OCCURRENCE NETWORKS AT AUTHOR-LEVEL

Junping Qiu[1] and Ke Dong[2]

[1] *jpqiu@whu.edu.cn,* [2]*dk8047@163.com*
Wuhan University, School of Information Management, Research Centre for Chinese Science Evaluation, NO.299 Ba Yi Road, Wu Han, Hu Bei ,430072 (China)

Abstract
This paper introduces bibliometrics co-occurrence at author-level by discussing its history and contribution to the analysis of scholarly communication and intellectual structure. It proposes five types of bibliometrics co-occurrence networks at author-level: (1) Co-authorship (CA); (2) Author Co-citation (ACC); (3) Author Bibliographic Coupling (ABC); (4) Words-based Author Coupling (WAC); (5) Journals-based Author Coupling (JAC). Networks of 98 highly influential authors from 18 journals indexed by 2011 version of journal Citation Report-SSCI under the Information Science & Library Science（IS&LS）category are constructed for study. Social Network Analysis and Hierarchical Cluster Analysis are applied as methods for identifying sub-networks with results visualized by VOSViewer. QAP test is used to find potential correlation among networks. The results from cluster analysis show that all the five types of networks have the power for revealing intellectual structure of sciences but have differences in describing results. Through the structure analysis of each type, the research groups which have relatively less connections with others are easily identified. ABC identified more sub-structures than other types of network, followed by CA and ACC while the result from WAC is easily affected. Analyzing result from JAC is ambiguous. QAP test result shows that ABC network has the highest proximity with networks of other types while CA network has relatively lower proximity with other networks. A combined use of several methods is suggested to have a better analysis of revealing intellectual structure of sciences.

Conference Topic:
Collaboration Studies and Network Analysis (Topic 6) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction and background
Bibliometrics and scholarly communication seem to be innately related with each other. Borgman (1989) regarded scholarly communication as a research area and bibliometrics as a research method. She maintained that scholarly communication can be examined through producers, artifacts, and concepts. In another article (Borgman & Furner, 2002), she substituted scholars for authors.

Co-occurrence in bibliometrics can be expressed at different levels and in multiple types. Yan and Ding (2012) categorized network based bibliometrics

studies into different levels, including aggregation levels, networks levels and approaches. Aggregation levels involve paper, author, journal, institutions and so on. Network levels include citation, co-citation, bibliographic coupling, co-authorship, etc.. Author-level should be importantly dominated at aggregation levels, because for paper, journal or institution, the networks of each is formed through scholarly relations such as collaboration or citation among authors.

Co-occurrence at author-level has been all the time research foci in bibliometrics field and is a significant approach to analyze scholarly communication and structure of science (Chen & Lien, 2011; White & Griffith, 1982; White & McCain, 1998; Ding & Cronin, 2011). Chasing back to the year 1966 when Price and Beaver (1966) did analysis on the author collaboration among 'invisible collages', they found that productive authors lead the way in their research group but bridged the way between the groups. Later in 21st century, with the development of network analysis and especially the accelerated researches in complex network on scientific collaboration (Newman, 2001a, 2001b, 2001c, 2004), collaboration has been increasingly pervaded in mining intellectual structure (Otte & Rousseau, 2002; Kretshmer, 2004; Adedo et al., 2006; Thijs & Glanzel, 2010).

Researches on knowledge structure of science studies based on author co-citation began in 1981 (White & Griffith, 1981). Author Co-citation Analysis is an analytical method that has been used to trace the intellectual structure. It assumes that two authors are correlated if they are cited together by later works, and if higher is this frequency, more similar are two authors. Author Co-citation has been studied intensively, for example, the operational procedure and diagonal value of matrix (McCain 1990), similarity measure (Ahlgren et al., 2003), visualization (Chen et al., 2001; White, 2003; Zhao, 2008), author co-citation analysis for web environment (Leydesdorff & Vaugh, 2006). Author bibliographic coupling is extended from bibliographic coupling (Zhao & Strotmann, 2008), with the assumption that the more references two authors have in common in their oeuvres, the more similar their research is.

As for ABC, the author coupling is based on cited references. Author coupling is formed not only through references, but also by using the same keywords or topic, so it can be called as words-based author coupling. Through journals on which authors published their research works, author coupling also can be achieved. Since journals respectively have ad hoc subjects and disciplinary interests, journals-based author coupling can also be used to illustrate the relationship of researches among authors and to analyze the proximity of authors' publishing behaviour and preferences in choosing publishing venues.

Here we analyzed five kinds of author-level co-occurrence networks to explore intellectual structure of 98 authors in library and information science: (1) Co-authorship (CA); (2) Author Co-citation (ACC); (3) Author Bibliographic Coupling (ABC); (4) Words-based Author Coupling (WAC); (5) Journals-based Author Coupling (JAC). White, Wellman & Nazer (2004) found that citation and social structure are mutually influenced. Ding (2011) analyzed authors'

endorsement by using co-authorship and citation networks. This article extends the research on correlation among networks to more types of co-occurrence networks by examining the relation among different types of bibliometrics networks at author-level and comparing different network analysis results.

Methodology

Data

First, data were collected on Dec.12, 2012 from 18 journals whose IF>1 indexed by 2011 version of Journal Citation Report-SSCI under the Information Science & Library Science（IS&LS） category. Three kinds of documents: articles, proceedings and review, were all downloaded. The journal whose name was altered has been taken into account during the data collection process. For example, JASIS altered its name as JASIST instead. However, the task of analyzing authors should consider disambiguation of author name (Torvik et al., 2005). So at the first step Thomson Data Analyzer (TDA) is applied to clean the data. Table 1 shows the basic information of the data.

Table 1. Summary of dataset

	value	*cleaned*
Number of papers	25,652	
Number of cited references	499,986	
Number of first authors	14,857	14,151
Number of authors	32,268	30,908
Number of cited authors	214,887	172,295

In table 1, the authors' names signed on documents in 18 types of journals have no obvious problems in labeling. After data-cleaning, the data result has an error rate less than 5% compared with the raw data. But serious problem is identified in authors' name labeling and the error rate is up to 20%.

The second step is to choose intellectual community with the criteria as follows: first author, all authors and cited authors should be ranked forefront according to the number of papers published or times cited. 98 authors are identified in IS&LS domain. Manual processing is required for filtering the data to get the sample data for analyzing.

Data cleaning is a task with circulatory process when situations differently cropped out in dealing with author name and keywords. In Web of Science (WOS) system, if all the indexed papers are cited, hyperlinks would be presented in references; however, it is impossible to get the value for each vertex in matrix from the WOS retrieval since the network structure at author-level is large scaled. In the author-level co-occurrence matrix, the main task of data cleaning is performed on cited authors and words. When cleaning the data of cited authors, it

is found that names are labeled mainly by authors themselves and that each type of journals do not have the same rules in name labeling, so this causes different situations with various types. For example, James J. Cimino is found to be Cimino J J, Cimino James J, or Cimino JJ and in other forms. Three types of situations in name labeling are found as follows: (1) first name and last name are reversed out of order; (2) whether space character is being used; (3) whether names use abbreviations or not and whether a dot is marked at the right side of abbreviations. Name labeling problems are even more complex in Chinese and Korean (Kim & Cho, 2013). Since the authors with high impact are relatively small in number under this research, a detailed discussion about it is omitted. After the range of authors is determined, manual processing is performed on data cleaning for the second time to get the final data as samples for this study.

Method

Social network analysis and traditional bibliometrics analysis are based on different perspectives and methods for structural analysis. For social network analysis, cohesion analysis is used, with methods include component, k-core, p-cliques etc. (Wouter, Mrvar, & Batagel, 2005), and it treats the network as a whole to explore the structure by analyzing sub-networks. For bibliometrics analysis, commonly used method is Hierarchical Clustering which categorizes samples into different types and aggregates them by calculating similarity distances. The difference between them is substantially rested in different perspectives whether it is holistic or atomistic.

In this paper, component analysis from SNA is applied because of its simple and intuitive effect; if it cannot get the sub-network, then Hierarchical Clustering is to be used to do structural analysis. Author proximity is expressed through using cosine similarity. The final results are visualized by using software VOSviewer (Van Eck & Waltman, 2010). A comparison on proximity results was conducted using the Quadratic Assignment Procedure (QAP). Its algorithm is always used in measuring correlations between two networks. QAP statistics are annotated in the documentation of Ucinet (Borgatti, Everett, & Freeman, 2002), a more detailed description can be found in White's paper (2004).

Results

Structure Analysis and Visualization

Figure 1 is the visualization result of co-authorship network. The size of vertices marked by using loops represents the number of papers published. The largest cluster of this network is found to be composed by 38 authors (cluster1), which has three distinct groups. First group is for the authors who are highly productive such as Egghe L, Leydesdroff L, Braun T, Rousseau T, Thelwall M etc. Most of them focus on bibliometrics, informetrics, scientometrics, and webometrics. The second group is centered around Ingwersen P, Croft W B with other members including Borgman CL, Belkin NJ, Rebertson SE and so on. Their research topics

are related with information seeking and retrieval. The third group is centered around Spink A with members Wilson TD, Ford N, Ellis D, Cole C etc., and the main research focuses are related with information behavior. The ties connected three groups indicate the research contents are inter-crossing. The first and second group is linked by Ingwersen P who is the winner of Derek de Solla Price Medal for his research in Scientometrics and Webometrics in 2005, and he is also a professor in information retrieval. The second and third group is linked by Saracevic T who has an extensive research interests mainly resting on digital library, information seeking and retrieval. Moreover, cluster2 is also quite large in scale with medical informatics as its research topic. This cluster is centered around Friedman C and Cimino JJ. Cluster3 is centered around four authors and they are Venkatesh V, Dennis AR, Agarwal R, and Rai A, most of whom are professors of business school with management information system as their research domain. Cluster4 includes 4 authors centered around Grover V whose research topics obviously emphasized on information technology. There are three authors in cluster5 who do the research work about government information. However, cluster6 contains large numbers of authors who do not collaborate with each other or some dyads.

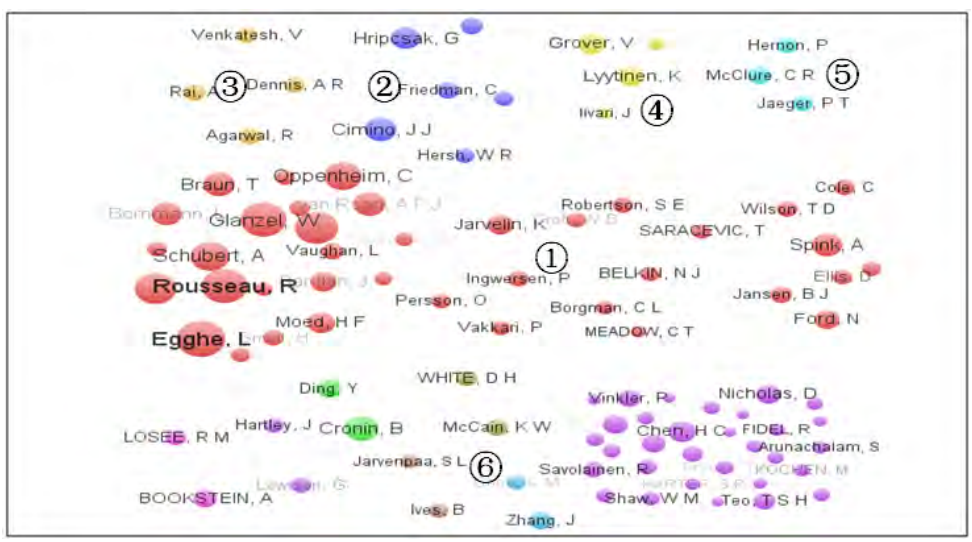


Figure 1. Mapping result of CA network

The clustering result for ACC yields four clusters, displayed in Figure 2. The size of vertices represents author's degree. For example, vertices with high degree include authors Garfield E, Leydesdorff L, Thelwall M, Egghe L, Salton G, Braun T and Kostoff RN etc.. The boundary between different clusters can be clearly displayed. The biggest group named cluster1 which is in the center of the graph is distributed separately. It contains many vertices that cannot be fallen under other

clusters such as Cimino JJ who is an expert in bioinformatics field, and Venkatesh V who is an expert in management information system. This is because the citation count of these authors in local dataset are lower than those in other clusters. Cluster2 on the left of the figure is the second largest in scale and is composed of experts who are connected more closely with each other in bibliometrics, informetrics, scientometrics and citation analysis. The cluster3 which is centered around Salton G, is clustered with members in information retrieval field while cluster4 was composed of authors in text mining and knowledge discovery fields centered around Kostoff RN. With exception to cluster1, the research boundaries of other three clusters can be easily identified. Although degree is used to represent the size of each vertex, the distribution of citation among different authors is not in the same case in the light of cited papers. For example, Garfield E, Small H, Salton G are quite similar in that most of citation come from their classic works in small number (Salton & McGill, 1983; Salton, 1989; Small, 1973; Garfield, 1972, 1979). But cases are different for vertices such as Leydesdorff L, Thelwall M, Egghe L whose citation distribution is clustered more loosely than the previous three authors.

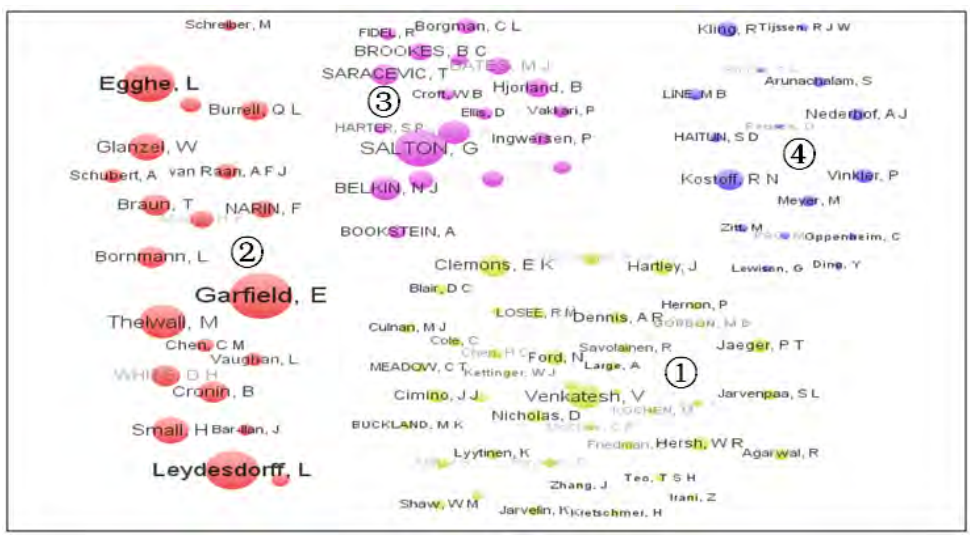


Figure 2. Mapping result of ACC network

Figure 3 shows the clustering result of ABC network. Different from ACC network, the research group on bibliometrics, informetrcs and scientometrics is partitioned into two parts; one is composed of Leydesdorff L, Glanzel W etc. (cluster1), while the other contains six authors including Egghe L, Rousseau R and Bernmann L etc. (cluster2). The six authors in cluster2 largely cited papers on h-index and bibliometrics laws; while Leydesdorff L and Glanzel W and other authors in cluster1 focus more on citation analysis, visualization and the application of bibliometrics methods. Small H, White HD and other authors in

cluster3 are from universities or research institutes in US such as Drexel University, ISI, and Indiana University Bloomington. They are clustered together through large numbers of papers with theme on co-occurrence. The three groups above are all doing relative researches on bibliometrics, informetrics and scientometrics, but research topics are diversified at micro level. The cluster4 formed by Thelwall M and other four authors is quite small in scale, but due to its extensive influence, it has a relatively good visualization result of cohesion analysis. And this group is clustered by papers on aspects of webmetrics, link analysis, and application of informetrics under web environment. Authors on the left in cluster8 are researchers mainly involved in management science. The density of ABC network is due to the size of intersection of references in the authors' published papers. These authors have written articles with paper references in a considerable number, so that the vertices turn out to be large in size in the visualization result. It is showed in the upper part of the figure that a research community on information retrieval is centered around Salton G (cluster5); the Cimino JJ-centered community on medical informatics (cluster6) and the community with member Spink A etc. who focus on users and information behavior (cluster7) are partitioned clearly.

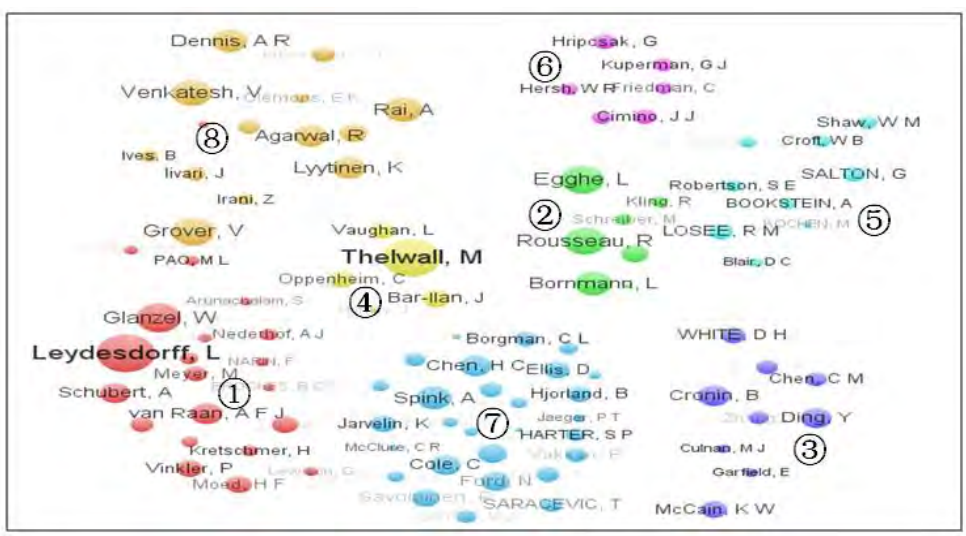


Figure 3. Mapping result of ABC network

Figure 4 shows the visualization analysis result of WAC network. The analysis process is very alike to the process in the previous ABC network in which research proximity is reflected by intersection of cited references while in WAC network research proximity is manifested by intersection of academic terms which are used to express research content. 14841 words used here are chosen from paper titles and are sorted through TDA and processed by manual intervention. The six research groups are partitioned and clearly presented.

Compared to the previous network analysis results, the number of authors in information system (cluster6) in this network is even larger, containing researchers who focus on information system from Information Science and those who focus on management information system from Management Science. Cluster1 and cluster2 in ABC are combined into cluster1 in WAC network with research interests in bibliometrics, informetrics and scientometrics. For cluster2, Thelwall M himself is mainly involved in webometrics, but the boundary of this group is not as clear as the one showed in ABC analysis result because papers published by Thelwall M as a co-author involve research contents such as citation analysis and impact factor etc. Thus, the scale of this group becomes larger. The analysis results for clusters such as medical informatics (cluster4), information seeking and retrieval (cluster3), information behavior (cluster5) are quite similar to the analysis result in ABC without obvious changes.

Figure 5 is the clustering result for JAC. Authors in medical informatics and management information system fields are distinctively partitioned at the upper part of the figure. Author group centered around Cimino JJ have papers published on JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION. Three authors in Cluster7 doing researches in government information and most of the papers are published on GOVERNMENT INFORMATION QUARTERLY. The group which is at the bottom of the figure is mainly composed of authors whose papers published on JOURNAL OF INFORMATION. The boundaries of other groups can be roughly distinguished but the authors fail to be partitioned to specific research communities.

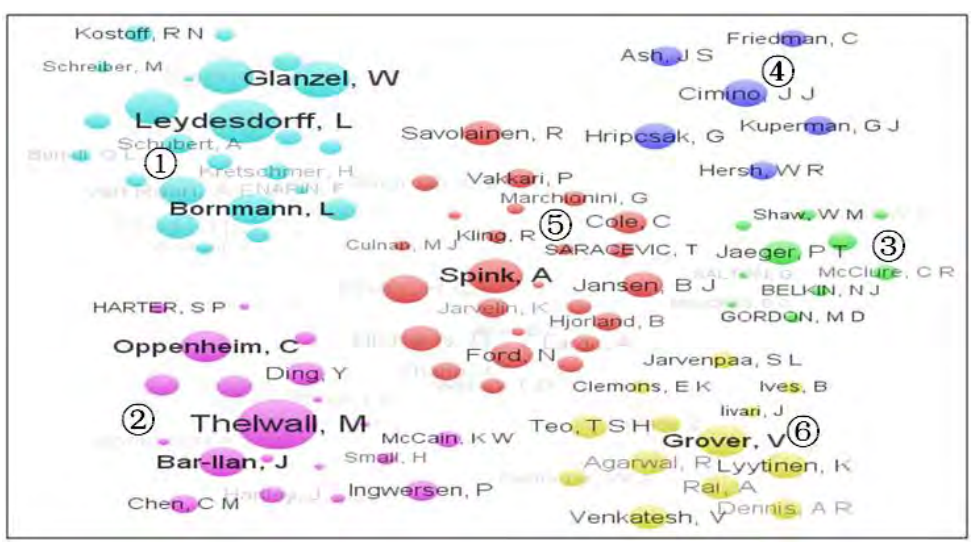


Figure 4. Mapping result of WAC network

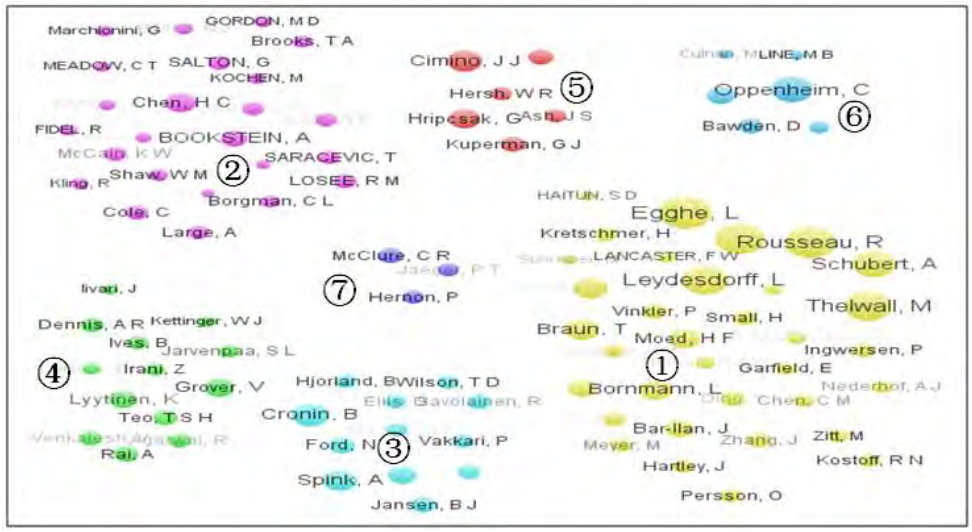


Figure 5. Mapping result of JAC network

Quadratic assignment procedure test

Table2 QAP correlation test of networks

Network	*CA*	*ACC*	*ABC*	*WAC*	*JAC*
CA	1	0.353	0.473	**0.220**	0.306
ACC	0.353	1	0.641	0.364	0.513
ABC	0.473	0.641	1	**0.700**	0.586
WAC	**0.220**	0.364	**0.700**	1	0.584
JAC	0.306	0.513	0.586	0.584	1

Note. number of permutations is 5000 times, with significant level at 0.001.

Table 2 is the result of QAP correlation test calculated among networks, showing that all the relations are significant at 0.001 level. It is noted that co-authorship network is most significantly related with author bibliographic coupling network while it is most irrelevant with words based author coupling network. The second row in table 2 can be read as CA network is most significantly related with ABC network, while ACC network is its second significant related network followed by JAC network and WAC network.

Conclusion and Discussion

Structure analysis results from Figure 1 to Figure 5 show that five types of networks are generally similar in structure. For example, two groups with research topics on medical informatics and management information system are always first identified, whatever network they are in. However, these networks have obvious differences in specific details. ABC can discover the intellectual structure more comprehensively; bibliometrics, informetrics and scientometrics are partitioned into two research domains with research topics more in detail (see

in Figure 3). CA and ACC are ranked second in accuracy, followed by WAC with analyzing result being easily affected while result of JAC is ambiguous.

For the five types of networks, the ACC network is yielded based on the authors cited by others, namely based on the recognition of the authors' research achievements by others. The other four types of networks were formed based on author's choice. CA network is due to author's own choice of collaboration with others; ABC network is formed when authors choose to use same previous research results to support their writings; WAC is based on author's choice of specific words which can express their research results; JAC is based on author's choice of publishing venues. The mechanism for these four types of network is similar when examining their network construction process.

Each type of network reveals academic communication at different levels. Generally, if collaboration among authors has happened, authors must have been acquainted with each other and have communicated and researched on shared research topics, so the CA network among authors are closest to social network in which desire for collaboration is strongest. The CA network is the most loosely constructed, which indicates that if two authors are in the same research domain, it is still not safe to declare the collaboration relationship between them. So it is more real to reveal the structure of academic communication based on CA network, but the power of CA network on revealing disciplinary structure is slightly weakened. With exception to CA network, other four types of network are constructed based on undirected connections. QAP analysis results show that CA network is least related with ACC, ABC, WAC and JAC. This analysis result is coincided with the one by Yan & Ding (2012) who measured institution-level network similarities by using cosine distance.

ACC network is far different with social network. This can be simply explained that ACC network represent authors' degree of recognition by others while the real fact may be that the authors never have had academic communication at all. So ACC network may not an appropriate way to analyze academic communication, rather it can be used to reveal disciplinary structure. With respect to the network proximity between ACC network and other networks, ACC is most related with ABC in that ACC reveals structure of relation in research paper references while author coupling tends to show the structure in authors at research frontier. For author groups of the same community, ACC and ABC are linked and have higher proximity.

As is shown from the forth column of Table 2, ABC network is the one that is most proximate with the other four types. Its proximity value with WAC is the highest up to 0.7. This result is inevitable since paper references which are the existing knowledge base are chosen closely related with author's own research content during scientific research; while as for WAC, authors' description of their own researches through use of words or academic terms can be regarded as summarization when extending further on previous researches.

From figure 4, authors in research areas such as information system in Information Science and management information system in Management are

clustered together by using words related with information system. Literally they are doing researches on information system, but the real fact is that their researches may have more differences than similarities. So the limitation of WAC is obvious that its analysis result and ability of revealing disciplinary structure can be easily affected by words which can cause ambiguity when they have different meanings. QAP test in table 2 explains this at some extent showing that WAC is weakly related with CA and ACC.

According to the researches made by Garfield (1979,1996), a relatively small number of journals publish the majority of significant scholarly results. Papers in these journals reflect the disciplinary affiliation of the journals. Although the dataset in this paper are collected from 18 journals indexed in ISI JCR-SSCI edition and categorized in IS & LS, twelve of them are interdisciplinary journals in six disciplines and have journals on management up to five kinds. With respect to the topics of these journals' interests, journals are generally not confined to single topic. The visualization result of JAC network shown in figure 5 fails to be clearly explained so that only a few groups that is obviously different can be partitioned. In QAP test, JAC network is not closely correlated with other network so it is mentioned here just as a possible analysis method.

Ni, Sugimoto & Cronin (2013) extended Borgman's three facet framework (1989) by adding a fourth gatekeepers. Journals are carriers for scientific research; meanwhile authors are the research conductors. The methodology in this paper seemed similar to theirs; however, this paper was new in its research dimension different from theirs although we have the same research purpose. On top of the proposed author-level couplings, authors can also be coupled through other forms of academic communication. For example, Cabanac (2011) considered author-venue coupling, as a way to measure inter-researcher similarity through the conferences they jointly might have attended. It's also an effective way to measure academic communication.

Although MIS grouping and Medical Informatics grouping are obviously different from those of information science/library science/informetics/scientometrics, they were kept for quick testing whether the 5 types of network can identify the two partitions or not. But modeling author production from the WoS only seems to be an approximation of the diversity in scientific output. For instance, open access journals and conferences are also can be considered to better reflect the true production of authors. It is a limitation of the current approach.

Acknowledgments

This paper is supported by the Major Program of the National Social Science Foundation of China (Grant No. 11&ZD152) and Humanities and Social Sciences project of Wuhan University (Grant No. 2012GSP032).

Reference

Adedo F.J., Barroso C., Casanueva C., & Galan, J.L. (2006). Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies*, 43(5), 957-983.

Ahlgren P., Jarneving B., & Rousseau, R. (2003). Requirements for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. *Journal of the American Society for Information Science and Technology*, 54(6), 550-560.

Borgatti, S.P., Everett, M.G., & Freeman, L.C. (2002). UCINet 6 for Windows: Software for social network analysis. Harvard, MA: Analytic Technologies.

Borgman, C. L. (1989). Bibliometrics and scholarly communication. *Communication Research*, 16 (5), 583.

Borgman, C.L., & Furner, J. (2002). Scholarly Communication and Bibliometrics. In B.Cronin (Ed.), *Annual Review of Information Science and Technology,* 36. Medford, NJ: Information Today, pp. 3-72.

Chen C., Paul R.J., & Keefe B.O. (2001). Fitting the jigsaw of citation: information visualization in domain analysis. *Journal of the American Society for Information Science and Technology*, 52(4), 315-330.

Chen L., & Lien Y. (2011). Using author co-citation analysis to examine the intellectual structure of e-learning: A MIS perspective. *Scientometrics*, 89, 867–886. doi: 10.1007/s11192-011-0458-y.

Ding Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, 5, 187-203.

Ding, Y., & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management*, 47(1), 80–96.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178 (4060), 471-479,.

Garfield, E. (1979). *Citation Indexing: Its theory and Application in Science, Technology and Humanities*. New York: Wiley & Sons.

Garfield, E., (1996). The Significant Scientific Literature Appears in a Small Core of Journals. *The Scientist*, 10(17), 13.

Kim, S., & Cho, S. (2013). Characteristics of Korean personal names. Journal of the American Society for Information Science and Technology, 64(1), 86-95. doi:10.1002/asi.22781.

Kretshmer, H. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics*, 60(3), 409-420.

Leydesdorff, L., Vaughan, L. (2006). Co-occurrence Matrices and Their Applications in Information Science: Extending ACA to the Web Environment. *Journal of the American Society for Information Science and Technology*. 57(12), 1616–1628.

Ma R.M. (2012). Author bibliographic coupling analysis: A test based on a Chinese academic database. *Journal of Informetrics*, 6(4), 532-542. doi:10.1016/j.joi.2012.04.006.

McCain K.W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433-443.

Newman, M.E.J. (2001). Scientific collaboration networks: I. Network construction and fundamental results. *Phys. Rev. E*, 64, 016131.

Newman, M.E.J. (2001). Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64, 016132.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. *PNAS*, 98(2), 404-409.

Newman, M.E.J. (2004). Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101(Supplemet 1), 5200–5205.

Ni, C., & Sugimoto, C. R. (2013). Visualizing and comparing four facets of scholarly communication: producers, artifacts, concepts, and gatekeepers. *Scientometrics*, 94, 1161–1173.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6), 441-453.

Price, D. J., & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist*, 21, 1011-1018.

Salton, G. (1988). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston: Addison-Wesley.

Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265-269.

Thijs, B., & Glanzel, W. (2010). A structural analysis of collaboration between European research institutes. *Research Evaluation*, 19(1), 55–65.

Torvik, V.I., Weeber, M., Swanson, D.R., & Smalheiser, N.R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technolog*, 56(2), 140-158.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.

White H.D. (2003). Pathfinder Networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54(5), 423-434.

White, H.D., & Griffith, B. C. (1981). Author cocitation: a literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32 , 163–171.

White, H.D., & Griffith, B. C. (1982). Authors as markers of intellectual space: Co-citation in studies of science, technology and society. *Journal of Documentation*, 38, 255–272.

White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author cocitation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, 49(4), 327–355.

White, H.D., Wellman, B., & Nazer, N. (2004). Does Citation Reflect Social Structure?Longitudinal Evidence From the "Globenet" Interdisciplinary Research Group. *Journal of the American Society for Information Science and Technolog*, 55(2), 111–126.

Wouter, D.W., Mrvar, A., & Batagel, V. (2005). *Exploratory Social Network Analysis with Pajek*. New York: Cambridge University Press.

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, co-citation networks, topical networks, coauthorship networks, and co-word networks relate to each other. *Journal of the American Society for Information Science and Technology*, 63(7), 1313–1326.

Zhao D. (2008). Information science during the first decade of the web: an enriched author cocitation anlysis. *Journal of the American Society for Information Science and Technology*, 59(6), 916-937.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, 59(13), 2070–2086.

COMPARING BOOK CITATIONS IN HUMANITIES JOURNALS TO LIBRARY HOLDINGS: SCHOLARLY USE VERSUS 'PERCEIVED CULTURAL BENEFIT' (RIP)

Alesia Zuccala[1] and Raf Guns[2]

[1] *a.zuccala@uva.nl*
Institute for Logic, Language and Computation, University of Amsterdam,
P.O. Box 94242, Amsterdam, 1090 GE (The Netherlands)

[2] *raf.guns@ua.ac.be*
University of Antwerp, IBW, City Campus, Venusstraat 35, B-2000 Antwerpen (Belgium)

Abstract
In this paper we examine the statistical relationship between citation counts to books referenced in SCOPUS humanities journals and library holding counts ('libcitations') retrieved from WorldCat®. Our focus is on books (with ISBN numbers) published between 2001-2006, which received citations in *History* and *Literature & Literary Theory* journals during the period of 2007-2011. A Spearman's rank correlation coefficient was used, and our test resulted in significant correlations between the citations and 'libcitations'. We present and discuss the details of our dataset (extracted from a much larger, newly constructed database), and comment on why the 'perceived cultural benefit' of holding a book in a research library can lead to, but may not necessarily lead to use (i.e., a citation) of that book in new humanities research.

Conference Topic
Scientometrics Indicators: Relevance to Humanities (Topic 1), Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2), and Bibliometrics in Library and Information Science (Topic 3).

Introduction
Books or monographs published in the humanities capture the research efforts of scholars concerned with human achievements. These texts are as much a part of our cultural heritage as they are part of scholarship (Garfield, 1979). In books we observe the story of a research discipline, that is, how it has evolved in different regions, over a specific time period, and within a particular "interpretive" community (Fish, 1980). Despite the fact that books are, for many humanities topics, principal modes of output, little is known about their scholarly *impact*. Bibliometricians have been reluctant to approach the subject of impact, because it is normally associated with high citation counts to and from articles published in scientific journals covered by the Web of Science (e.g., the Journal Impact Factor). Some journals published in the humanities are agreeable to impact factors

(see Elsevier, 2010), but for the most part, these measures have been avoided in favour of general citation monitoring (Nederhof, 2006). Since the late 1970s, research has focused primarily on the *characteristics* of cited works in humanities texts or *classifying citations* to or from small monograph collections in disciplines where they are most prevalent (Budd, 1986; Cullars, 1985; 1989; 1998; Frost, 1979; Hammarfelt, 2012; Heinzkill, 1980; Hellqvist, 2010; Jones et al., 1972; Lindholm-Romantschuk & Warner, 1996; Nolan, 2010; Stern, 1983; Thomson, 2002).

For books in general, the absence of source metadata (i.e., internal identification codes) in the main commercial citation indices (i.e., Thomson Reuters' Web of Knowledge and Elsevier SCOPUS) has made it difficult to develop reliable indicators. Books have always been recorded in the Thomson Reuters' Web of Science (i.e., Science-SCI-E, Social Science-SSCI and Arts & Humanities-A&HCI) as 'non-sourced' cited materials, but some 'book chapters' and 'books' started to appear in all three indices as far back as 2005. Growth rates indicate that their appearance has occurred irregularly (Leydesdorff & Felt, 2012). Recently, Leydesdorff & Felt (2012) found that the classification of books in the Web of Science is problematic: many have been misclassified as articles or reviews. Thomson Reuters' new Book Citation Index (BKCI) is expected to be a more accurate resource for bibliometric analyses (Adams & Testa, 2011). With the introduction to this index, we have been promised a 'complete picture' (Thomson Reuters, 2013). Only research based on this new Book Citation Index can tell us how useful it will be for evaluating citation-based impacts over the long term.

While the new indices are still in production, some researchers from the bibliometrics community have been considering alternative ways to study the impact of books. Kousha and Thelwall (2009) confirm that there are substantial numbers of citations to academic books from Google Books and Google Scholar to help evaluate book-oriented disciplines. Torres-Salinas and Moed (2009) as well as White et al. (2009) have focused on the potential of library catalogues for impact-based analyses, where an analogy may be created between journal-based citations and library holdings. Torres-Salinas and Moed (2009) studied the number of catalogue inclusions per book title in WorldCat®, while White et al. (2009) introduced the term 'libcitation' as "an indicator of perceived cultural benefit" (p. 1087). Linmans (2010) later suggested that researchers use a three-level approach for assessing books, focusing on citation counts, library holdings, and productivity.

The present study is motivated by the contributions of Torres-Salinas and Moed (2009) and White et al. (2009). Our objective is to further this earlier work using a special database that we have constructed to include books cited in journal articles covered by SCOPUS (*History* and *Literature & Literary Theory*) and

corresponding library holding counts in both *Association of Research Libraries* (ARL) and non-ARL libraries. These were gleaned from WorldCat®. ARL is a non-profit membership organization of 125 research libraries in North America. Here, we explain how this database was developed for a much larger project (i.e., still a research in-progress) and present some preliminary analyses pertaining to the scholarly use of books (i.e., cited in journals) and their 'perceived cultural benefit' (catalogued in ARL and non-ARL libraries).

Overview of the datasets and database

Data were granted to us from Elsevier through the Elsevier Bibliometrics Research Program. In our application to this program we requested two separate datasets, each limited to citations recorded in journals classified as *History* or *Literature & Literary Theory* (Table1).

Table 1. Journals and journal citation data granted by Elsevier SCOPUS (April 2011).

Journal article publication years: (1996-2000 and 2007-2011)	
Journal Numbers and Classifications	
• History (n=604)	• ASJC 1202 (SCOPUS Classification Code)
• Literature & Literary Theory (n=529)	• ASJC 1208 (SCOPUS Classification Code)

Upon receiving the SCOPUS data, we examined the number of citations recorded in the 1023 journals (two time periods together) to determine the overall frequency to books, research articles (ar), conference proceedings (cp), review papers (re), notes (no) and other non-sourced materials. Cited materials that did not have an internal SCOPUS identification number, or did not meet the criteria that we established for identifying other materials - e.g., a non-sourced journal/proceedings article - were classified as a 'book'. All 'other' documents will be re-examined and classified at a later date.

Table 2. Number of citations according to document type.

CITING - TO-	*CITED*	*NUMBER*
Research Article (ar)	Books	1,647,520
Research Article (ar)	Other documents	1,563,831
Review (re)	Books	1,162,461
Review (re)	Other documents	852,464
Research Article (ar)	Research Article (ar)	133,531
Review (re)	Research Article (ar)	41,855
Conference Proceeding (cp)	Books	41,800
Conference Proceeding (cp)	Other documents	34,854
Research Article (ar)	Review (re)	30,493
Notes (no)	Books	27,341

Table 3, above, presents some descriptive statistics resulting from queries made to our new database. Here we show the total number of documents cited by articles or reviews published in *History* and *Literature & Literary Theory* journals for two citation windows: 1996-2001 and 2007-2011. Some of these cited documents have been categorized as follows: a) sourced in SCOPUS only, b) non-sourced in SCOPUS, but matched in WorldCat®, or c) sourced in SCOPUS and matched in WorldCat®. In this paper, we are concerned with a subset of books that were non-sourced in SCOPUS, but matched in WorldCat®.

Table 3. Cited documents as SCOPUS sourced or non-sourced items.

	All Cited Docs	Sourced in SCOPUS only	Not in SCOPUS, but Matched in WorldCat ®	In SCOPUS & Matched in WorldCat ®	Not in SCOPUS or WorldCat ®	Cited Docs w. Missing Values (?)
HISTORY						
1996-2001	882,155	6,945	303,048	368	564,773	7,021
2007-2011	2,858,005	117,789	**806,985**	2,251	1,915,002	15,978
LITERATURE						
1996-2000	198,606	815	75,840	139	120,445	1,367
2007-2011	1,395,917	36,737	**504,721**	1,546	845,561	7,352

Data analyses and results

The aim of this study is to statistically examine the relationship between citation counts to books in SCOPUS journals and WorldCat® library holding counts ('libcitations') for both ARL and non-ARL libraries worldwide. We expect to find a strong positive relationship, where 'libcitations' or library holdings support or lead to journal citations, but not vice versa. The idea behind this hypothesis is that humanities scholars borrow books from their university/academic library and 'use' these books by reading and citing them in research articles/review papers. Time is required for a book to be published, marketed to and purchased by a library before it is cited; hence we focus on a book publication window of six years (2001-2006), followed by a journal citation window of five years (2007-2011). With respect to library holdings, we assume that a book published in 2001 would not have been added to one of the libraries at least until this date or sometime after (up to and including Nov. 2012).

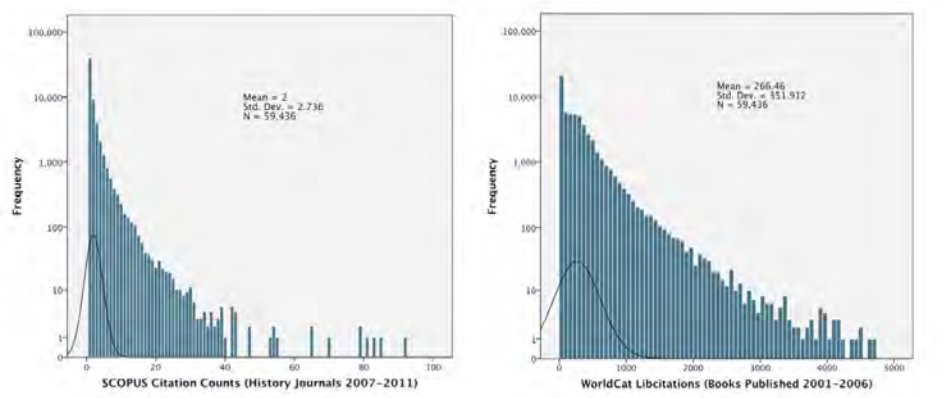


Figure 1. Citation and 'libcitation' frequency distributions for books published in 2001 to 2006 (cited in SCOPUS *History* journals, 2007-2011).

Table 4. Spearman's rank correlation coefficients for citations and libcitations. (*History*, 2007-2011 and *Literature & Literary Theory*, 2007-2011).

		CITATIONS	LIBCITATIONS		NON-PARAMETRIC CORRELATIONS SPEARMAN'S rho		
PUBLISHED BOOKS (ISBN#)	NUMBER	SCOPUS JOURNALS	ASSOCIATION OF RESEARCH LIBRARIES (ARL)	OTHER LIBRARIES (NON_ARL)	CITATIONS + ARL HOLDINGS	CITATIONS + NON_ARL HOLDINGS	CITATIONS + ALL LIBRARY HOLDINGS
HISTORY	59,436	Max. = 92 Min. = 1 Mean = 2.00	Max. = 212 Min. = 0 Mean = 46.36	Max. = 4,603 Min. = 0 Mean = 220.10	.288*	.267*	.275*
LITERATURE	41,853	Max. = 91 Min. = 1 Mean = 1.91	Max. = 215 Min. = 0 Mean = 48.35	Max. = 4,603 Min. = 0 Mean = 247.63	.281*	.244*	.254*
*correlation is significant at the 0.01 level (1-tailed).							

To select the best test for our hypothesis we first observed and compared the frequency distributions for all citation counts and library holding counts in the separate fields. Figure 1, above, presents the book citation and library holding distributions for *History* (*N*=59,436) only. Note that the data are skewed thus do not fit a normal curve. Similar non-normal distributions were found for *Literature & Literary Theory* (*N*=41,853). With this data the most appropriate test to use is the Spearman's rank correlation coefficient. The procedure for performing a Spearman correlation is the same for a Pearson correlation; however, the Spearman rho is less sensitive to strong outliers. Table 4, above, presents some general statistics related to our datasets and indicates that we found significant, though not especially strong correlations between citations and 'libcitations' in *History* as well *Literature & Literary Theory*.

Discussion and Considerations for Further Research

First it is important to comment on the 'cleanliness' of the *History* and *Literature & Literary Theory* datasets. Since we were working with thousands of 'non-sourced' book references, it was necessary to examine repeat iterations of the reference strings to be sure that they were to the same book. For the most part, they were, but without a deep manual cleaning effort, we cannot say that the datasets were 'perfectly' clean. With respect to our correlation results, it is possible that if given access to cited book references from other books as well, the Spearman's rho might even be more significant. Also, there is much to be said about conducting this type of test, especially when correlation measures are not necessarily the best for understanding 'causes' and 'effects'. Many 'in between' variables can influence a correlation, some of which may be the browsing habits of humanities scholars, the concentrated nature of their work, and habit of citing books from their collegial network regardless of whether or not it is present in their institutional research library. Nevertheless, it is the goal of their institution to hold books that are 'perceived' to be beneficial to the culture of their research. From a bibliometric perspective, it is helpful to know if research-oriented libraries are doing for the humanities what they aim to do, which is to make quality books available for scholarly use.

A sizable portion of the books are present in many libraries but infrequently cited in the data set. The reverse (highly cited, but present in few libraries) is less common. When we examined the list of books that were proportionally cited least compared to their holding count, we discovered two main reasons for the divergence. First, reference works such as the *Oxford English Dictionary* have a very high 'perceived cultural benefit' but are not typically cited. The presence of this kind of book indicates that citations and 'libcitations' are not entirely interchangeable – they measure (partially) different dimensions. Second, the list contains many books that stem from other disciplines (e.g., Diagnostic and Statistical Manual of Mental Disorders). It seems likely that most books in this category would be highly cited if journals from their disciplines were part of the data set. More work may be done with larger datasets, for instance, expanding the data to include other humanities subjects, and/or making comparisons with cited books in the social sciences and sciences. There is also a strong opportunity here to further examine the role of book reviews, as 'gateway' documents, i.e., documents that encourage or discourage librarians to purchase books, and the motivation of scholars to read and cite books that were or were not selected based on reviews.

Ackowledgements

The authors are grateful to both the Elsevier Bibliometrics Research Programme (http://ebrp.elsevier.com/) and OCLC WorldCat® for granting access to the data that were used to build the unique database required for this study. We also wish to thank Dr. Roberto Cornacchia for assisting us with the development of our

database, as well as Maurits van Bellen and Robert Iepsma for their data cleaning and standardisation work.

References

Adams, J., & Testa, J. (2011). Thomson Reuters Book Citation Index. In E. Noyons, P. Ngulube & J. Leta (Eds.), The 13th Conference of the International Society for Scientometrics and Informetrics (Vol. I, pp. 13-18). Durban, South Africa: ISSI, Leiden University and the University of Zululand.

Budd, J. (1986). Characteristics of written scholarship in American literature: A citation study. *Library and Information Science Research, 8*, 189–211.

Cullars, J. (1985). Characteristics of the monographic literature of British and American literary studies. *College & Research Libraries, 46,* 511-22.

Cullars, J. (1989). Citation Characteristics of French and German Literary Monographs. *Library Quarterly, 59,* 305-25.

Cullars, John. (1998). Citation characteristics of English language monographs in philosophy. *Library & Information Science Research, 20(1)*, 41–68.

Elsevier (2010). Latest impact factor figures from Elsevier's arts and humanities journals. Retrieved August 1, 2012 from http://about.elsevier.com/impactfactor/author-reports-93964/webpage/author-webpage-93964.html.

Fish, S. (1980). *Is there a text in this class? The authority of interpretive communities*. Harvard, MA: Harvard University Press.

Frost, C. O. (1979). The use of citations in literary research: A preliminary classification of citation functions. *Library Quarterly, 49*, 399-414.

Garfield, E. (1979) Is Information Retrieval in the Arts and Humanities inherently different from that in Science? The effect that ISI®'S Citation Index for the Arts and Humanities is expected to have on future scholarship. *The Library Quarterly, 50(1),* 40-57.

Hammarfelt, (2012). Following the footnotes: A bibliometric analysis of citation patterns in literary studies. Unpublished PhD Thesis, Uppsala University. Retrieved January 7, 2013 from

Heinzkill, R. (1980). Characteristics of references in selected scholarly English literary journals. *Library Quarterly, 50,* 352-365.

Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology, 61(2)*, 310–318.

Jones, C., Chapman, M., & Woods, P. Carr. (1972). Characteristics of the literature used by historians. *Journal of Librarianship, 4,* 137-56.

Kousha, K. & Thelwall, M. (2009) Google book citation for assessing invisible impact? *Journal of the American Society for Information Science and Technology, 60(8)*, 1537-1549.

Ledesdorff, L. & Felt, U. (2012). "Books" and "book chapters" in the book citation index (BKCI) and science citation index (SCI, SoSCI, A&HCI).

Proceedings of the American Society for Information Science and Technology, 49(1), 1-7.
Lindholm-Romantschuk, Y. & Warner, J. (1996). The role of monographs in scholarly communication: An empirical study of philosophy, sociology, and economics. *Journal of Documentation, 52(4),* 389–404.
Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66(1),* 81-100.
Nolan, D. S. (2010). Characteristics of la literatura: A reference study of Spanish and Latin American literature. *College & Research Libraries, 71(1),* 9-19.
Stern, M. (1983). Characteristics of the literature of literary scholarship. *College & Research Libraries, 44(4),* 199-209.
Thompson, J. W. (2002). The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. *Libri, 52(3),* 121–136.
Thomson Reuters. (2013). Putting books back into the library: Completing the research picture. The Book Citation Index[SM]. Retrieved January 8, 2013 from http://wokinfo.com/products_tools/multidisciplinary/bookcitationindex/.
Torres-Salinas, D., & Moed, H.F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. *Journal of Informetrics, 3(1)*, 9–26.
White, H., Boell, S.K, Yu, H., Davis, M., Wilson, C.S., & Cole, F.T.H. 2 (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology, 60(6),* 1083-1096.

A COMPARISON OF TWO HIGHLY DETAILED, DYNAMIC, GLOBAL MODELS AND MAPS OF SCIENCE

Kevin W. Boyack[1] and Richard Klavans[2]

[1] *kboyack@mapofscience.com*
SciTech Strategies, Inc., Albuquerque, NM 87122 (USA)

[2] *rklavans@mapofscience.com*
SciTech Strategies, Inc., Berwyn, PA 19312 (USA)

Abstract
As data availability and computing resources increase, the ability to create more detailed and accurate global models of science is also increasing. This article reports on two advances in methodology aimed at creating more accurate versions of these highly detailed, dynamic, global models and maps of science. 1) A combined co-citation/bibliographic coupling approach for assigning current papers to co-citation clusters is introduced, and is found to significantly increase the accuracy of the resulting clusters. 2) A sequentially hybrid approach to producing useful visual maps from models is introduced. Two maps and models – one based on linked annual co-citation/bibliographic coupling models, and one based on direct citation – are created from a 16-year (1996-2011) set of Scopus data comprising over 20 million documents. The two models are compared and are found to be very complementary to each other.

Conference Topic
Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8)

Introduction
Over the years, our quest has been to increase the accuracy, coverage, and detail of models (or classification schemes) and maps of the structure and dynamics of science. The purpose behind this quest has been one of practicality – detailed, comprehensive, and accurate maps of science can be used to address a host of questions that are currently being asked by decision makers worldwide. Many of these questions fall under the broad headings of 'planning' and 'evaluation'. Much of the metrics (sciento-, biblio-, infor-, alt-) world is focused on evaluation; topics include impact factors, h- and other indexes, rankings, etc. We are far more interested in planning, and in using maps of science and technology to address topics such as portfolio analysis, predicting emergence and organizational structure.

The dance between funding bodies, administrators, and researchers, each asking and answering their own questions, coupled with external stimuli (e.g., social, political, regulatory issues, etc.), is what creates the structure and dynamics of

science. It is an extremely complex dance; individual actors number in the millions. Yet, despite this complexity, the ability to accurately model the structure and dynamics of global science at a highly detailed level is within our reach. It has already been shown that modern science is very robust in its high level structure (Klavans & Boyack, 2009; Leydesdorff & Rafols, 2009). Recent advances in modelling global science at the publication level (Klavans & Boyack, 2011; Waltman & Van Eck, 2012) suggest that clustering of millions of scientific documents into a large number of clusters (tens of thousands) results in partitions that are highly recognizable to subject matter experts (Klavans, Boyack, & Small, 2012). The fact that these small partitions are recognizable gives us confidence that these structures are reasonable representations of the actual topics in science that have resulted from the complex interactions of many actors.

Planning and evaluation are inseparably connected. One cannot answer questions that will impact the future without having a retrospective understanding of how science operates and how those operations are embodied in a model. Thus, researchers develop detailed models (and maps) of past science to gain an understanding of the relationships between structural units (clusters of documents) and how and why they change over time. With this understanding, we hope to learn those features that will allow us to more accurately answer questions related to planning.

To that end, this article reports on multiple advances in the creation of models and maps from millions of scientific documents, and quantifies the effect of each advance on the accuracy of the models. In this article, we use co-citation analysis, bibliographic coupling, and direct citation analysis for modelling, and text analysis in the final visualization step. We use our own methodologies, and we also use the new clustering methodology from CWTS (Waltman & Van Eck, 2012) that can quickly create models from millions of documents.

In the balance of the article, we first review related work to provide context for the advances reported here. We then detail a recent advance in co-citation analysis, and a new sequentially hybrid method for generating a map layout. Each of these was investigated using a different data set. We then create two separate models of science using a dataset comprised of over 20,000,000 documents from Scopus (1996-2011) – one based on linked annual co-citation/bibliographic coupling models, and the other based on direct citation using the CWTS methodology. As an additional step, visual maps were created for each model using textual analysis. The article will close with a comparison of the two models and maps, and with a summary of the advances and findings presented here.

Background

Science mapping, when reduced to its most basic components, is a combination of classification and visualization. We assume there is a structure to science, and then seek to create a representation of that structure by partitioning sets of documents (or journals, authors, grants, etc.) into different groups. This act of partitioning is the classification part of science mapping, and typically takes the

majority of the effort. The resulting classification system is what we call a model of science. The visualization part of science mapping uses the results of the classification as input, and creates a visual representation (or map) of that model as output.

Mapping of scientific structure using data on scientific publications began not long after the introduction of ISI's citation indexes in the 1950s. Since then, science mapping has been done at a variety of scales and with a variety of data types. Many of these studies have been reviewed at intervals in the past (Börner, Chen, & Boyack, 2003; Morris & Martens, 2008; White & McCain, 1997). When it comes to mapping of document sets, most studies have been done using local datasets. The term 'local' is used here to denote a small set of topics or a small subset of the whole of science. While these local studies have successfully been used to improve mapping techniques, and to provide detailed information about the areas they study, we prefer global mapping because of the increased context and accuracy that are enabled by mapping of all of science (Klavans & Boyack, 2011).

The context for this study lies in the efforts undertaken since the 1970s to map all of science at the document level using citation-based techniques. The first map of worldwide science based on documents was created by Griffith, Small, Stonehill & Dey (1974). Their map, based on co-citation analysis, contained 1,310 highly cited references in 115 clusters, showing the most highly cited areas in biomedicine, physics, and chemistry. Henry Small continued generating document level maps using co-citation analysis (Small, 1999; Small, Sweeney, & Greenlee, 1985), using thresholds based on fractional citation counting that ended up keeping roughly the top 1% of highly cited references by discipline. The mapping process and software created by Small at the Institute for Scientific Information (ISI) evolved to generate hierarchically nested maps with four levels. Small (1999) presents a four level map based on nearly 130,000 highly cited references from papers published in 1995, which contained nearly 19,000 clusters at its lowest level. At roughly the same time, the Center for Research Planning (CRP) was creating similar maps for the private sector using similar thresholds and methods (Franklin & Johnston, 1988). One major difference is that CRP's maps only used one level of clustering rather than multiple levels.

The next major step in mapping all of science at the document level took place in the mid-2000's when Klavans & Boyack (2006) created co-citation models of over 700,000 references papers and bibliographic coupling models of over 700,000 current papers from the 2002 fileyear of the combined Science and Social Science Citation Indexes. Later, Boyack (2009) used bibliographic coupling to create a model and map of nearly 1,000,000 documents in 117,000 clusters from the 2003 citation indexes. Through 2004, the citation indexes from ISI were the only comprehensive data source that could be used for such maps. The introduction of the Scopus database in late 2004 provided another data source that could be used for comprehensive models and maps of science. Klavans & Boyack (2010) used Scopus data from 2007 to create a co-citation model of science

comprised of over 2,000,000 reference papers assigned to 84,000 clusters. Over 5,600,000 citing papers from 2003-2007 were assigned to these co-citation clusters based on reference patterns.

We note that all of the models and maps mentioned to this point have been static maps – that is they were all created using data from a single year, and were snapshot pictures of science at a single point in time. It is only recently that researchers have created maps of all of science that are longitudinal in nature. In the first of these, Klavans & Boyack (2011) extended their co-citation mapping approach by linking together a set of nine annual models of science to generate a nine-year global map of science from Scopus data, comprised of 10,360,000 papers from 2000-2008. More recently, Waltman & van Eck (2012) at CWTS clustered nearly 10,000,000 documents from the Web of Science (2001-2010) using direct citation and a modularity-based approach that is similar to their familiar VOS method. This new CWTS approach has advantages over other approaches: it can be used to generate a multi-level hierarchical clustering, and it can handle very large document sets with very modest computational requirements. Although it has been used with direct citation similarities, there is no reason it could not be used with similarities generated from other methods, such as co-citation, bibliographic coupling, or even text-based or hybrid similarity measures.

Combined co-citation/bibliographic coupling approach

The last step in co-citation analysis is to assign current papers to the co-citation clusters using their references. We have long assumed that there should be a much more accurate way of assigning current papers to co-citation clusters. In the past we have done this using simple fractional assignment based on the distribution of references to clusters for each paper. For

example, for a paper with 10 references, if seven of those references appeared in one cluster, and three in a second cluster, the paper would be assigned to those two clusters with fractions of 0.7 and 0.3, respectively.

We recently decided to design and test a new approach. This approach assigns papers fractionally to co-citation clusters by combining cluster solutions from co-citation analysis and bibliographic coupling. The detailed process is as follows:

1) A bibliographic coupling solution for the current papers was calculated using the methodology from Boyack & Klavans (2010). At this point there are two solutions – the original co-citation solution which fractionally assigns papers to clusters (*PID CC wt*), and the bibliographic coupling solution which singly assigns papers to cluster (*PID BC*).
2) The current papers (*PID*) are divided into 3 groups:
    a. Group A – those that are in both solutions
    b. Group B – those that are only in the BC solution
    c. Group C – those that are only in the CC solution

3) For all papers in group A, a figure of merit (*FOM*) based on a combination of the co-citation (*CC*) and bibliographic coupling (*BC*) clusters was calculated:
    a. The *BC* cluster was assigned to each *PID* in the *CC* solution to create a table with entries (*PID CC BC wt*)
    b. Weights were summed up by *CC:BC* pair (*CC BC sumccbc*)
    c. Weights were summed by *BC* cluster (*BC sumbc*) and added to the table in 3b (*CC BC sumccbc sumbc*)
    d. Divide *sumccbc* by *sumbc* to get *FOM* for each *CC:BC* pair (*CC BC FOM*)
    e. This figure of merit replaces the original weights for each *PID* within a *CC:BC* pair (*PID CC BC FOM*)
    f. *FOM* are summed and normalized so that they sum to 1.0 for each *PID* (*PID CC FOMnorm*). These *FOMnorm* become the new weights for *PID* in the co-citation clusters, replacing the old weights.
4) Bibliographic coupling clusters some papers that are not clustered by the co-citation solution. These papers (group B) were assigned to co-citation clusters by:
    a. Creating a *CC:BC* cosine relatedness matrix from the *FOM* in step 3d where the matrix values are *cos = FOM/sqrt(rowsum*columnsum*)
    b. Linking *PID* to *CC* using their *BC* clusters (*PID CC BC cos*)
    c. Singly assigning the *PID* to the *CC* cluster with the highest cosine value (*PID CC 1.0*)
5) Co-citation clusters some papers that are not clustered by the bibliographic coupling solution (group C). The simple fractional assignments from co-citation analysis are used for these papers.

We tested this new current paper assignment approach against our simpler fractional assignment approach using a set of 2.15 million documents (2004-2008) that intersect the Scopus and Medline databases. We used this set because we had previously calculated both co-citation and bibliographic coupling solutions for these data (Boyack & Klavans, 2010), and they were thus available for use without needing additional work. As was done in our previous study of similarity approaches, we compared solutions using two metrics – textual coherence and a concentration (Herfindahl) index based on grant-to-article linkages mined from Medline. The reasoning behind the grant-article linkage metric is that multiple articles that reference the same grant should be concentrated, rather than dispersed, in a cluster solution. In other words, the cluster solution that does the best job of putting articles from the same grant in the same cluster can be assumed to be the more accurate solution. One other positive benefit of using grant-to-article linkages is that they are an extrinsic measure of quality – they are not used in the clustering in any way and thus do not bias the results.

Table 1. Results of a combined co-citation + bibliographic solution compared to simple co-citation and bibliographic solutions.

	BC	*CC*	*CC-BC*
Coh	0.08599 (+5.3)	0.08167	0.08865 (+8.5%)
Herf	0.28486 (+19.8%)	0.23778	0.27516 (+15.7%)

Table 1 shows that the coherence and the grant-to-article concentration (Herf) metrics are both significantly higher for our new combined co-citation/bibliographic coupling (CC-BC) approach than for the simple co-citation (CC) approach. In fact, the coherence for the combined approach is also higher than that for a simple bibliographic coupling (BC) approach. The simple bibliographic coupling approach still has a higher concentration index than the combined approach, but not by much. Given the size and scope of the dataset, we take these results as an indicator that a combined co-citation/bibliographic coupling approach is preferable to using either approach separately, and we have revised our approach to modeling accordingly.

Sequentially hybrid map layout

Once a model (or classification system, or cluster solution) has been created from a set of documents, it is often useful to create a visual map of the model. One straight-forward way of doing this is to create a graph layout of the clusters in the model. A variety of methods have been devised for this type of visualization. However, the most common layout algorithms in use today (e.g., Fruchterman-Reingold, Kamada-Kawai) are typically only used to generate layouts for small datasets (100s of nodes). We use the OpenOrd (formerly DrL) algorithm to generate layouts for sets of hundreds of thousands of nodes (Martin, Brown, Klavans, & Boyack, 2011).

For many years we have generated visual maps of our large-scale co-citation models using co-citation between pairs of clusters. To do this, one takes the original list of citing-cited article pairs, replaces the cited articles with their cluster numbers, and then runs the same algorithm used in the original co-citation similarity calculation, but with cluster numbers as the cited items. The result of this is a set of cluster-cluster similarity values based on co-citation at the cluster level. This set of similarity values can then be used as input to a graph layout algorithm to calculate cluster coordinates which can then be used to create a visual map of the model. As an example, Figure 1 (left) shows the map created from cluster-cluster similarities between 116,163 clusters based on co-citation from the 2010 file year of Scopus data. Although this map shows the relative positions of major areas of science, we have never found this type of map to be as appealing and informative as we would like because everything is so bunched

together. There is very little white space in this visual map. There are few structures that could represent discipline-level structures.

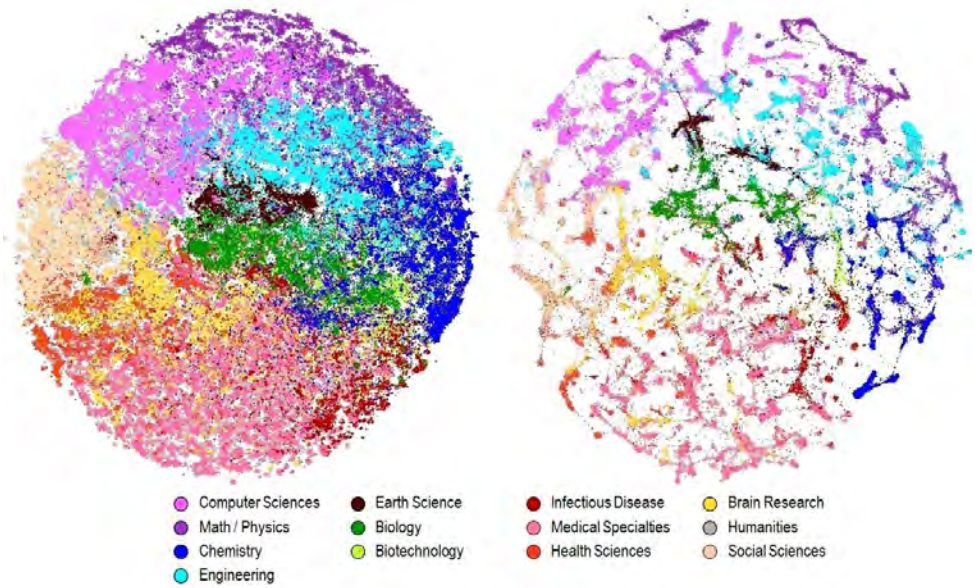


Figure 1. Visual maps of co-citation clusters from our Scopus 2010 model, where the layout was based on co-citation similarities (left), and BM25 text similarities (right) between clusters.

We decided to test an alternative cluster-cluster similarity based on textual analysis. We had resisted this in the past because of the computation requirements of calculating text-based similarities between all pairs of over 100,000 clusters. Although any of several text-based similarities would likely have worked equally well, we decided to use the BM25 measure because it is simple to calculate, is among the least computationally expensive text approaches, and is among the most accurate measures that we had previously tested (Boyack et al., 2011). Each cluster was represented textually as being comprised of the titles and abstracts of its papers. BM25 was then used to calculate cluster-cluster scores. The BM25 similarity between one object q and another object d is calculated as:

$$s(q,d) = \sum_{i=1}^{n} \left(IDF_i \frac{n_i (k_1 + 1)}{n_i + k_1 \left(1 - b + b \left| D \right| / \overline{D} \right)} \right)$$

where n_i is the frequency of term i in object d. Note that $n_i = 0$ for terms that are in q but not in d. Typical values were chosen for the constants k_1 and b (2.0 and 0.75, respectively). In our formulation each cluster was treated as if it were a

single document. Document length $|D|$ was estimated by adding the term frequencies n_i per document. Average document length $\overline{|D|}$ is computed over the entire set of documents. The IDF value for a particular term i is computed as:

$$IDF_i = \log \frac{N - n_i + 0.5}{n_i + 0.5}$$

where N is the total number of documents in the dataset and d_i is the number of documents containing term i. Each individual term in the summation in the first formula is independent of document q. To remove the influence of high frequency terms all IDF scores below 2.0 were discarded.

Figure 1 (right) shows the visual map of co-citation clusters that resulted from using BM25 to calculate cluster-cluster similarities. The similarity file was filtered to the top-N similarities per cluster, and layout was done with OpenOrd using the default edge cutting parameter. These same steps were also used for the map at the left of Figure 1, enabling a comparison of the two maps that can be definitively tied to the differences in similarity measures. A comparison of the maps leads to some interesting observations. First, clusters of a particular color (each color indicates a broad field in science) are in the same relative position in both maps, indicating that co-citation and text-based methods give a similar view at the highest level. This is to be expected given the consensus in high level map structures that has been recently noted by multiple researchers (Klavans & Boyack, 2009; Leydesdorff & Rafols, 2009). Second, clusters in the text-based map are grouped much more densely, leaving a significant amount of white space. The BM25 similarity values are an order of magnitude higher than the co-citation similarity values, which suggests that the lower similarities lead to much more even spacing between nodes in the map, while higher similarity values create a map with much more well defined groupings of clusters. This should also not be surprising. In this map, the clusters were created using co-citation, using up a high fraction of the variance in the system that could be accounted for using co-citation. A second level of similarity using co-citation thus should have far less signal available with which to link clusters than would textual analysis, the majority of whose signal would still be available.

We find the text-based map to be far more visually compelling than the co-citation map because of the localized density of clusters, the greater amount of white space, and the visible strings between localized areas that indicate pathways between discipline-like structures. In addition, following the arguments above, it is likely that the layout of the text-based map is based on more signal indicating similarity between clusters than is the co-citation map. This new mapping technique can be considered as a hybrid technique. Although the first level similarity metric is not a citation+text hybrid, this technique uses a citation-based method to generate clusters, followed by a text-based method to generate a cluster layout, and is thus a *sequentially hybrid* map layout technique.

Two dynamic global maps of science

Two large-scale models of science – one using our new combined CC-BC approach and one based on direct citation using the CWTS clustering approach – were constructed. Maps were created for each model using a sequentially hybrid layout where the citation-based clusters are positioned using cluster-cluster similarities calculated using BM25. The data set for these models is a 16-year (1996-2011) set of Scopus data comprised of over 20 million documents. Although the entire Scopus data from those years contains 25.6 million records, only 20.6 million of those have references. Given that we are using citation-based techniques to model science, these 20.6 million records can be considered as our basis set.

Linked co-citation/bibliographic coupling model

Although our linked co-citation approach is explained in detail elsewhere (Klavans & Boyack, 2011), we give a brief version of our updated linked CC-BC approach here for completeness. Models are calculated for each file year.

- A subset of the cited references is selected using roughly the top 12% of cited references.
- Co-citation counts (C_{ij}) are calculated for each pair of references and then converted to modified frequencies as $a_{ij} = 1/\log(p(C_{ij}+1))$ where $p(C_{ij}+1) = C_{ij}\ (C_{ij}+1)/2$.
- Calculate K50 values from the a_{ij} matrix using the K50 formula above.
- The matrix of K50 values is filtered to the top-N per node, where N varies from 5 to 15, using the method described above.
- References are clustered using OpenOrd using the detailed process explained in Boyack & Klavans (2010). The minimum cluster size was set to five papers.
- Current papers are fractionally assigned to the clusters using the combined co-citation/bibliographic coupling process explained earlier in this article.

Annual models are then linked into a longitudinal model of science by linking clusters of documents from adjacent years together using overlaps in the cited references belonging to each cluster.

We used the above process to create a 16-year (1996-2011) CC-BC model of science from Scopus data. Table 2, which contains numbers of papers and clusters by year, shows that the process is relatively stable in terms of cluster sizes and the fraction of annual articles covered by the model.

The 16 annual models of Table 2 were linked together into a longitudinal model of science using overlaps in the cited references from adjacent years. Linked set of clusters are called *threads*. For each pair of years, linking is done using the superset of references from the two years' models. Typically, ⅓ of the references are present in both models, ⅓ are only present in the first year's model, and the other ⅓ are only present in the second year's model. The majority of the

references that are only in one model are missing from the other only because they did not meet the citation threshold, and can be easily added to the other model using their reference lists. This process generates augmented reference lists for each model. For a given model, the augmented reference list used for linking to the prior year's model is somewhat different than the augmented reference list used for linking to subsequent year's model. Using these augmented reference lists, clusters from adjacent models are linked if a simple cosine index based on the number of overlapping references is above a threshold.

Table 2. Annual details of the CC-BC model of science

Year	*#Clust*	*#Pap*	*Pap/Clust*	*%Pap*	*#Ref*	*Ref/Clust*	*FwdCos*
1996	54,221	752,442	13.88	95.2%	1,072,014	19.77	0.2593
1997	56,225	774,390	13.77	95.3%	1,108,296	19.71	0.2578
1998	57,434	788,643	13.73	95.3%	1,149,310	20.01	0.2572
1999	59,048	808,027	13.68	95.3%	1,215,370	20.58	0.2605
2000	64,072	876,335	13.68	95.4%	1,333,079	20.81	0.2575
2001	70,680	965,106	13.65	95.7%	1,447,172	20.47	0.2540
2002	74,207	1,004,837	13.54	96.0%	1,541,707	20.78	0.2547
2003	79,657	1,080,103	13.56	96.2%	1,665,590	20.91	0.2535
2004	90,074	1,212,349	13.46	96.1%	1,854,537	20.59	0.2500
2005	98,848	1,332,524	13.48	96.2%	2,058,536	20.83	0.2451
2006	107,197	1,451,006	13.54	96.1%	2,243,455	20.93	0.2385
2007	113,426	1,546,811	13.64	95.6%	2,360,593	20.81	0.2357
2008	121,595	1,645,524	13.53	95.6%	2,539,626	20.89	0.2294
2009	130,701	1,754,603	13.42	96.0%	2,759,731	21.11	0.2230
2010	135,836	1,807,757	13.31	96.4%	2,930,351	21.57	0.2227
2011	151,305	2,004,176	13.25	96.6%	3,277,735	21.66	

Although it would be nice to set this cosine threshold based on theory, in practice we find it requires a heuristic approach. If the cosine threshold is set too low, a giant component quickly emerges and the longitudinally-linked sets of clusters become so large as to no longer represent research problems. If the cosine threshold is set too high, there is very little linking between clusters. We also found that, given that science is growing and that the number of clusters increases each year, using a single cosine threshold for all pairs of years created less linking for later years (late 2000s) than for earlier years (late 1990s). This was an undesirable effect. To create consistency in the linking patterns over time, we ran linking calculations at a variety of linkage fractions, where linkage fraction is defined as the fraction of clusters in a given year that link to a cluster in the subsequent year. This required calculation of the cosine threshold for each year that would return the desired linkage fraction. For each set of calculations we examined the average and maximum numbers of forward links per cluster (of those that have forward links), along with the maximum thread size. We chose a linkage fraction of 0.48, meaning that only 48% of the clusters link to a cluster in the subsequent year. This gave us average and maximum numbers of forward

links per cluster of 1.1 and 5, respectively. The corresponding forward linking cosine threshold values for each year are listed in Table 4, and range from 0.260 to 0.223, typically decreasing over time.

When linking clusters this way over many years, one additional problem can arise – two long threads can be linked together in a later year if the cosine value is high enough, creating artificially large threads. For example, using the method and thresholds detailed above, the largest thread had 872 clusters, or 54 clusters per year. Although this type of linkage can reflect the history of how topics link together from a retrospective point of view, it does not necessarily reflect how each thread grows. We thus implemented an additional step in our threading calculation. For cases where a single cluster merges two threads, where one thread is at least 8 years old, and the other is at least 5 years old, the cluster is assigned to the shorter thread and not to the longer one, despite the fact that the cosine threshold is met in both cases. This criterion still allows very long threads to form, but it does not allow retrospectively joining of long existing threads. The effect of this criterion on size was to reduce the largest thread to 66 clusters (or 4 per year). This is a significant improvement in our view; the majority of the threads remain thin – they are not dominated by branching – and thus represent coherent research problems as they move through time.

We have examined the age characteristics of the resulting threads. Roughly 40% of the clusters are what we call *isolates*. These are clusters that link neither forward nor backward within the model. These are research problems that do not have enough momentum to continue into a second year. *Isolates* are typically among the smallest clusters, while the longest threads are comprised of larger clusters on average. 46% of clusters are in threads of 3 years or longer.

Direct citation model

To create our 16-year direct citation model of science, we used the CWTS modularity-based code, which is explained in great detail by Waltman & van Eck (2012). Some details of the calculation we ran are listed here:

- All direct citations within the 16-year data set were selected. Pairwise similarity values between citing-cited pairs of papers were then calculated using the similarity normalization method from Boyack & Klavans (2010). The similarity file was filtered to the top-N similarities per paper, with N ranging from 5 to 15, based on total degree.
- The resulting similarity file was input into the CWTS code, which was run at a single level with a minimum cluster size (n_{min}) of 20, and resolution (r) of 9.0 x 10^{-5}. The code was run 10 times and the solution that maximized the CWTS quality function was used as our completed model.

A total of 19,012,183 papers (92.1% of those with at least one reference) were assigned to 149,613 clusters in the direct citation model. While determining start dates, end dates, and ages of threads from the CC-BC model was based directly on linked clusters, calculating these properties for direct citation clusters requires

some assumptions. Direct citation clusters can have very long ramp-up periods with few papers (e.g., 2 papers per year), and intervening years with zero papers, making it difficult to clearly delineate start and end dates. We decided to calculate start and end dates using the following method:

- Mean year and standard deviation were calculated for each cluster from the publication year of its papers.
- Although the annual distribution of papers in clusters is not normally distributed for most clusters, we nonetheless applied the three-sigma rule, which assumes that 99.7% of observations will lie within three standard deviations of the mean. Upper and lower limits for cluster start and end dates were set at the mean $\pm$ three standard deviations.
- The cluster start date (year) was assigned to be the first year, greater than or equal to the lower limit, in which the cluster had five or more papers. The cluster end date was assigned to be the latest year, less than or equal to the upper limit, in which the cluster had five or more papers.

Although not an exact determination, this methodology produced reasonable results that were better for clusters that started in 2004 or later than for those that started earlier. This protocol could undoubtedly be refined, but is sufficient to provide a first characterization of the model. Using this definition of cluster start and end dates, the number of clusters that are active ($\geq$5 papers) during each year was calculated. It is interesting to note that a large majority of all clusters (90.2%) are active in 2008. This is very different from the CC-BC model, where a much smaller fraction of the threads are active in any one year.

Visual maps of 16-year models

Maps have been created for each of the two models described in the sections above. For each map, BM25 coefficients were calculated between pairs of clusters using their titles and abstracts, as described earlier. The full set of similarities was filtered to the top-N (5 to 15) similarities per cluster, and layout was done using OpenOrd with a default edge cutting setting. For the CC-BC map, isolates were not included; only the 190,151 threads of two years or longer were included in the layout. For the direct citation map, all 149,613 clusters were included in the layout.

Figure 2 shows that the two maps have similar characteristics. The color distributions in each map are similar, as is the visual appearance, with white space and pathways between cluster-like structures in both maps. Although the two models that underlie these maps are dynamic or longitudinal, the maps are static pictures of the threads (CC-BC) or clusters (direct citation). The maps can easily be time-sliced to show growth in the various areas of science over time, but they are not truly dynamic in that they assign each thread or cluster to a single position, and do not allow them to move over time with changing influences from other clusters. Although each map is shown here at high level, smaller sections can be enlarged to show greater detail. These maps can also be used as templates on which additional information, such as cluster ages, or output from a particular

author, journal or institution, can be overlaid (Rafols, Porter, & Leydesdorff, 2010).

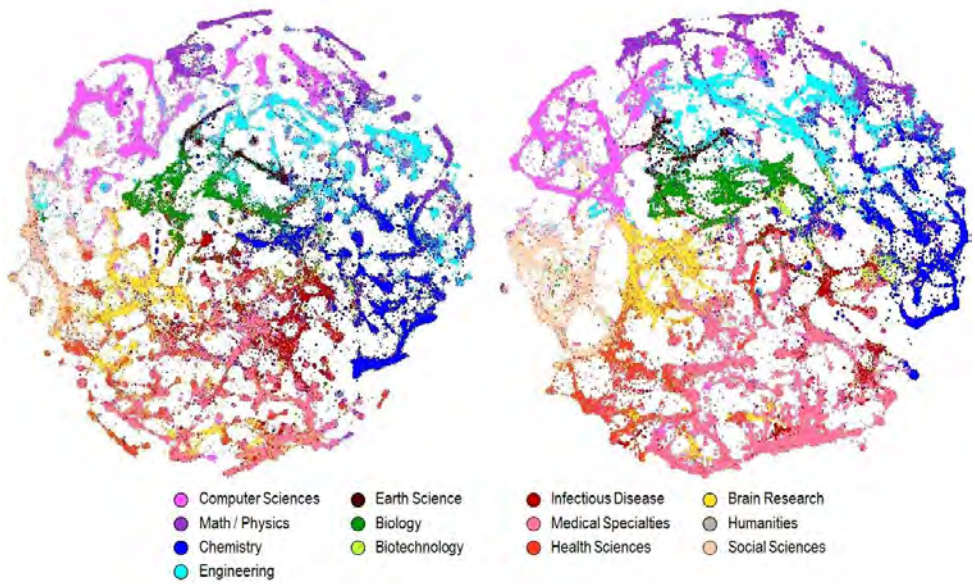


Figure 2. Visual maps of the CC-BC (left) and direct citation (right) 16-year models of science.

These two maps, although created using different similarity measures and different clustering processes, have nearly identical coherence values – the CC-BC map has a coherence of 0.08201, while the direct citation map has a coherence of 0.08152.

Discussion

There are a variety of approaches that can be used to create maps of science. These different approaches are based in different theoretical perspectives of how science operates. We have explored the CC-BC and direct citation global maps, compared their properties, and have also thought a great deal about the theory behind and methodologies used to construct these two types of maps. The direct citation map is based in a theoretical framework that emphasizes *academic lineages*. Direct citation relies inherently on the direct linkages (lineages) between documents. Since direct citation explicitly includes self-citations, this framework tends to preserve the historical (rather than the cognitive) bases of the lines of research conducted by researchers and research groups. The CC-BC map is based in a theoretical perspective that emphasizes *problem frames* – or how researchers cognitively frame the research problems on which they work, and how those frames change over time. Research problems only persist in time if a group of researchers maintain a similar reference frame from year-to-year. The use of co-citation analysis to create clusters of reference papers aligns primarily with this

theoretical perspective. Clusters of cited references are the intellectual bases that are used to frame current research problems.

The direct citation approach lends itself to clusters of long duration; it is highly retrospective and takes into account citation linkages of multiple ages, including self-citations, and it is not segmented into annual models. By contrast, co-citation models can change much more rapidly because they are based on a second-order process and self-citations are not explicitly accounted for. (We note that they are accounted for to a lesser degree through the second-order processes.) It should not be surprising that a set of linked annual co-citation or CC-BC models will have thread durations that are much shorter on average than the durations of direct citation clusters.

Both approaches do a very good job of clustering the database content. The high level map views are similar, but the details are different. Given that the two models and maps are based on different theoretical perspectives, we feel no need to choose one over the other; these two models are extremely complementary. For example, preliminary analyses suggest that analysis based on a combination of both maps can be extremely useful for the identification of emerging topics (Small, Boyack, & Klavans, 2013).

Summary

This article has detailed advances in methodology aimed at creating more accurate versions of highly detailed, dynamic, global models and maps of science.

- A combined co-citation/bibliographic coupling approach for assigning current papers to co-citation clusters was introduced, and was found to significantly increase the accuracy of the resulting clusters.
- A sequentially hybrid approach to producing useful visual maps from models was introduced. We advocate the use of citation-based approaches to create a model (or classification system) from data, followed by use of cluster-cluster similarities generated using a text-based approach for map layout.

In addition to these advances, we constructed two maps and models – one based on linked annual co-citation/bibliographic coupling models, and one based on direct citation – were created from a 16-year (1996-2011) set of Scopus data comprising over 20 million documents. The two models were compared and found to be very complementary to each other. We consider these two dynamic models of science to be the current standards to which any who wish to create document level models of all of science must compare their work.

Acknowledgments

We thank Michael Patek for extracting the necessary data and creating and running the scripts that create our models of science, and Ludo Waltman at CWTS for making both code and data available to us for this study. The sequentially hybrid layout experiment described here was supported by the Center for Scientific Review (CSR) at the U.S. National Institutes of Health.

References

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*, 179-255.

Boyack, K. W. (2009). Using detailed maps of science to identify potential collaborations. *Scientometrics, 79*(1), 27-44.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., et al. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One, 6*(3), e18029.

Franklin, J. J., & Johnston, R. (1988). Co-citation bibliometric modeling as a tool for S&T policy and R&D management: Issues, applications, and developments. In A. F. J. van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 325-389). North-Holland: Elsevier Science Publishers, B.V.

Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). Structure of scientific literatures. 2. Toward a macrostructure and microstructure for science. *Science Studies, 4*(4), 339-365.

Klavans, R., & Boyack, K. W. (2006). Quantitative evaluation of large maps of science. *Scientometrics, 68*(3), 475-499.

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology, 60*(3), 455-476.

Klavans, R., & Boyack, K. W. (2010). Toward an objective, reliable and accurate method for measuring research leadership. *Scientometrics, 82*(3), 539-553.

Klavans, R., & Boyack, K. W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology, 62*(1), 1-18.

Klavans, R., Boyack, K. W., & Small, H. (2012). *Indicators and precursors of 'hot science'*. Paper presented at the 17th International Conference on Science and Technology Indicators.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348-362.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: An open-source toolbox for large graph layout. *Proceedings of SPIE - The International Society for Optical Engineering, 7868*, 786806.

Morris, S. A., & Martens, B. V. (2008). Mapping research specialties. *Annual Review of Information Science and Technology, 42*, 213-295.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871-1887.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799-813.

Small, H., Boyack, K. W., & Klavans, R. (2013). *Identifying emerging topics by combining direct citation and co-citation*. Paper presented at the 14th International Conference of the International Society for Scientometrics and Informetrics.

Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the Science Citation Index using co-citations. II. Mapping science. *Scientometrics, 8*(5-6), 321-340.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378-2392.

White, H. D., & McCain, K. W. (1997). Visualization of literatures. *Annual Review of Information Science and Technology, 32*, 99-168.

A COMPREHENSIVE INDEX TO ASSESS A SINGLE ACADEMIC PAPER IN THE CONTEXT OF CITATION NETWORK[33] (RIP)

Yi Han, Hui Xia & Ying Tong

Hanyi72@swu.edu.cn, 826414600 @qq.com, 564571440 @qq.com
College of Computer and Information Science, Southwest University,
1 Tiansheng Road, Beibei, Chongqing ,400715 (P. R.China)

Abstract
The influence of a single academic paper should be assessed under the context of citation network, in which a literature's references reveal the source of knowledge and its inheritance relationship and a literature's citing papers reflect the flow and diffusion of knowledge. ID index and its derived RID index are designed on the basis of such inheritance and diffusion to evaluate the influence of a single academic paper, which not only takes into account direct references and direct citing papers but also the contribution of indirect references and indirect citing papers. With the academic papers of Library and Information Science in Web of Science as a sample, the present research selects six sample vertexes in main path of its citation network to calculate their ID and RID. The correlation coefficient between the ID and traversal value verifies that the ID index has higher validity and effectiveness to measure the influence of a single academic paper.

Keywords
single academic paper ; academic influence assessment ; ID index ; citation network ; main path

Conference Topic
Scientometrics Indicators: Criticism and new developments (Topic 1) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches(Topic 3)

Introduction
A scientific, objective and fair measuring of the influence of academic literature is of great significance for its objective assessment of scientific creation and academic works of researchers as well as its guidance to management work, such as selection of scientific research project, establishment of the assessing system of researcher performance, application of research funds, formulation of scientific plan and policy, and so on. However, it is always a puzzling problem to get a scientific, objective and fair evaluation method to measure the influence of academic literature, especially that of a single academic paper.

[33] Research for this article has been funded by scientific and technological foundation of Southwest University (No. SWU112001)

Measuring the influence of academic literature in the context of citation network is of great necessity since every literature is firmly grounded in corresponding citation network system. In citation network citing and cited relationship among literatures can be included, the characteristics of direct and indirect citation can be reflected, and the particular position of a specific literature in the structure of citation network can be located. An influential evaluation index designed on the basis of citation network structure would be undoubtedly an objective mapping for the holographic information of citation network. Literature in citation network would inherit the influence of cited literature by citing and simultaneously diffuse its own academic influence if cited. Inheritance, therefore, is the accumulation of influence; diffusion, meanwhile, the penetration of influence. There would be an increase in an academic paper's influence, academic strength and the power of back penetration if with more authorized knowledge source and more creative theories and knowledge. Thus, the influence of an academic paper is the integration of knowledge accumulation and penetration ability. Citation network is a faithful record of the process of influence accumulation and penetration of an academic paper. So how to quantify the influence of an academic paper based on this process is the major problem to be solved in this paper.

Relative Researches on the Measurement of the Influence of a Single Academic Paper

Various methods are adopted by scholars to assess the influence of a single academic paper scientifically, objectively and fairly. Generally speaking, two kinds of methods are in current use: qualitative evaluation and quantitative evaluation.

Qualitative Evaluation

Commonly used method of qualitative evaluation is peer review which refers to the process of evaluating a scientific activity and its result by an evaluation committee composed of experts in a given scientific field (Geisler, 2000, pp.217-242). Peer review can provide quality control for the distribution of scientific and technological resources in any level, from individual to institutions and to nation; and it is one of the methods to do after-evaluations for the scientific and technological activities and their performers. With definite goals and standards, objective and fair result can be achieved through peer review. But peer review usually costs much time and is liable to the influence of subjective and emotional factors of reviewers, which causes difficulty in ensuring an objective, fair and impartial evaluation.

In the biomedical field, f1000 (faculty of 1000) system was developed based on peer review (Bornmann & Leydsdorff, 2013). The system is the quantitative manifestation of peer review with aim of an objective evaluation of important papers collected in the biomedical database, such as SCI and PUBMED, on the basis of academic achievement rather than its source journals. The evaluation results have been catalogued into three grades: outstanding literature (score more

than 9), the required ones (score between 6 to 9) and recommended ones (score between 3 to 6). An asserted involvement of more than ten thousand experts in the evaluation processes cannot cover the drawback of peer review itself. Moreover, evaluated papers are limited to the biomedical field and the number of the evaluated paper is relatively small (up till the present altogether about 130000 articles). The application of f1000 is restricted as the result of its charging service system. Despite these drawbacks, the evaluation system is of great value to be popularized, especially so far as the periodical papers open to all are concerned.

Quantitative Evaluation

With regard to the quantitative evaluation, scholars study the evaluation method mainly on two levels: using a single index or using integrated indices.

Easily obtainable, the journal impact factor becomes one of the common indexes to evaluate the academic quality and influence of an academic paper, which makes it the simplest method to evaluate a single academic paper by using journal impact factor directly (Hoeffel, 1998; Garfield, 2006). But the inter-causal relationship between impact factors and paper quality renders this method a target of criticism by scholars. Then why is this method still used in practice? As Hoeffel (1998, p.1225) and Garfield (2006) argued, the journal impact factor is not the perfect tool to evaluate paper quality, but no better ones are available at present. Owing to the great differences in impact factor among different disciplines, Impact Factor Point Average (IFPA) is proposed to solve the problem of the comparison of impact factors among different disciplines (Sombatsompop, et al, 2005).

Except the journal impact factor, there are some other single index to measure the academic influence of a single paper, such as Paper Quality Index (Qiu, et al, 2007), academic papers quality index based on citation strength (Wu, 2007), academic paper assessment in the same field based on principal component analysis (Long, 2007), single paper h Index (Schubert, 2009; Thor & Bornmann ,2011).

Since the paper quality is determined by multiple factors, a comprehensive evaluation index should be designed to integrate the advantages of various methods and avoid their disadvantages. Many suggestions have been proposed, for example, the integrated evaluation method based on indices such as periodical literature type, periodical influence, international clearer defect display, and fund assistance (Zhang & Pan, 2004); the integrated academic quality indices based on the periodical influence factor, the paper cited frequency and non self citation frequency (Guo,2005); the comprehensive evaluation system based on non self citation amount and journal impact factor (Jin, et al,2009).

In spite of their special characteristics and respective disadvantages, a common problem occurs: emphasis is merely put on direct citation relationship among papers without considering various indirect citations and the value of references, and little attention is paid to citation context structure, namely a paper's position in the citation network. Taking direct and indirect references, direct and indirect

citation, and citation network into consideration, the present research tries to establish a new evaluation index.

Methodology and Data

Research Method

Academic papers form a self-organized network by citing each other. Figure 1 is a simple citation network. In a citation network, some nodes or elements occupy central status because of their positional relationship, in which they play a key part in inheriting and diffusing the field knowledge. For a paper, except for its citation number, citation structure has provided important background information to present its influence. In figure 1, if we simply use the citation number as a measuring standard, the citation number of node 2,8,10 all are 3, and their influence should be the same; but if we take the structure information into account, then the influence of node 8 and 10 may be greater than that of node 2. Particular attention should be paid to node 8 which is the bridge of the entire network and whose influence should be more greater. Therefore, the academic value of a paper is both associated with the backtracking depth of the references and the extending breadth of the cited papers. Backtracking depth and extending breadth is a comprehensive reflection of the quality effect of a paper in citation network.

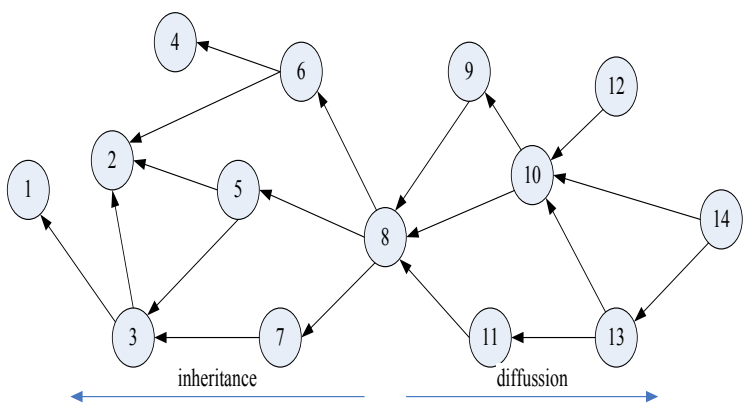


Fig 1 a simple citation network

If treating the nodes in each level equally (True a refiner processing requires that different weights should be put on the different elements structure and the different number relationship, but for the convenience of data processing, they are temporarily regarded as having the same weight), we can construct a simplified evaluation index to measure a single academic paper influence as follows:

$$ID = \sum_i \frac{r_i}{i^2} + \sum_j \frac{c_j}{j^2}$$

In the index, i indicates the backtracking depth of the cited papers of the assessed literatures, j indicates the extending breadth of the citing papers, r_i indicates the cited number of the *i*th backtracking level, and c_j indicates the citing number of the *j*th extending level. The calculation result is named ID, which is the comprehensive interacting effect of knowledge inheritance and diffusion of the assessed paper.

Generally, closer to the measured paper, greater the contribution of its influence is, so different weights should be given to different cited or citing levels. When i=1, r_1 is the direct references number of the assessed paper ; when $i \geq 2$,r_i is the indirect references; we use $1/i^2$ as the adjustment factor to every backtracking level, which exemplifies out greater attention to short-distance effects. $1/i^2$ is chosen as the adjustment factor mainly because, on the one hand it is the manifestation of the principle that closer the distance is, greater the importance is; on the other hand it can reduce the negative impact caused by dramatic increase in indirect references as the distance increases. It is the same to j, that is, the direct cited papers are given greater emphasis, and $1/j^2$ is used as adjustment factor. Take Node 8 in Figure 1 as an example. Its backtracking level is 3, and its extending breadth is 2. In terms of backtracking, there are 3, 3 and 1 cited papers on the first, second, and third level respectively. So far as extending breadth is concerned, there are both 3 citing papers in the first and second level. So to node 8, $ID_8 = 7.61$.And similarly, $ID_2 = 4.01$ and $ID_{10} = 7.30$.Comparing their ID value, we can know that the influence of node 8 is greater than that of node 2 and 10, while the influence of node 10 is greater than node 2. The ID value is totally different from the result of measuring simply on the basis of citation number. This indicates that: even the citation number is completely the same, the paper influence is not the same because of the different location in citation network. Hence, using the citation number as the mere evaluation criterion may not fully reflect the real influence of academic literature, and the citation structure should be considered.

According to the ID index of a single paper influence measurement, we can also calculate the relative contribution rate of the references and the citing papers:

$$RID = \frac{\sum_i \frac{r_i}{i^2}}{\sum_j \frac{c_j}{j^2}}$$

If $RID \geq 1$, it indicates that the contribution of knowledge inheritance factors is greater than that of knowledge diffusion factors, and vice versa. In fig.1, we can get the results: $RID_2 = 0$, $RID_8 = 1.03$, $RID_{10} = 1.43$. As can be seen, node 2 is

without knowledge inheritance, which means this node is the source of the citation network and its main effect is knowledge diffusion. Knowledge inheritance contribution of node 8 is weaker than that of node 10. Of course, for those papers without citations, namely the sink of citation network, the relative contribution rate will be infinite with the indication that their full contribution is knowledge inheritance.

From the above, a problem will be raised naturally: whether the ID index can be used in the actual citation network? Whether the ID values reflect the importance of the node itself? In the following part, it will be verified by actual citation network of some sample field.

Data Collection

Sixteen kinds of core journals of Library and Information Science, collected by SCI and tabulated in Table 1(Yan & Ding, 2009), are chosen in this research as a sample. The sample data was retrieved on the Web of Science On November 30, 2011. Document type of these data includes *Article*、*Book*、*Book Chapter*、*Book Review*、*Discussion*、*Hardware Review*、*Letter*、*Note*、*Proceedings Paper*、*Review*、*Software Review*. They come from *SCI-EXPANDED*、*SSCI*、*CPCI-S* databases and their time range is "all year "(1900-2011).

First of all, main path analysis is used to sort all the nodes by traversal value. Secondly, some sample nodes in the main path are selected and their ID values are calculated. Finally, the correlation analysis between the traversal values of sample nodes and their ID values are done.

Tab.1 Journal name of Library and Information Science sample

NO.	Journal Name	NO.	Journal Name
1	Annual review of information science and technology	9	Journal of the american society for information science and technology
2	Information processing & management	10	Information society
3	Scientometrics	11	Online information review
4	College and research libraries	12	Library quarterly
5	Journal of documentation	13	Library resources and technical services
6	Journal of information science	14	Journal of academic librarianship
7	Information research	15	Library trends
8	Library & information science research	16	Reference and user services quarterly

Research Results

The main path of Citation network of Library and Information Science

The main path is a path from source to sink in acyclic network, whose arcs have the highest traversal values (De Nooy, et al, 2005).The main path analysis focuses

on the connectivity of citation network (Hummon & Doreian, 1989). Main-path techniques examine connectivity in acyclic networks, and are especially interesting when nodes are time dependent, as it selects the most representative nodes at different moments of time.

Three models to identify the most important part of a citation network can be distinguished: the node-pair projection count, which accounts for the number of times each link is involved in connecting all node pairs; the search-path link count, which accounts for the number of all possible search paths through the network emanating from an origin; and the search-path node pair, which accounts for all connected vertex pairs along the paths (Hummon & Doreian, 1989). Of these three methods, the latter two algorithms are included in Pajek and a new algorithm, search path count, is designed(Batagelj, 2003). The search-path link count is the preferred algorithm for this analysis because all citation relations are taken into account. In this study, we use the search path count algorithm to get the main path of the sample (Fig 2).

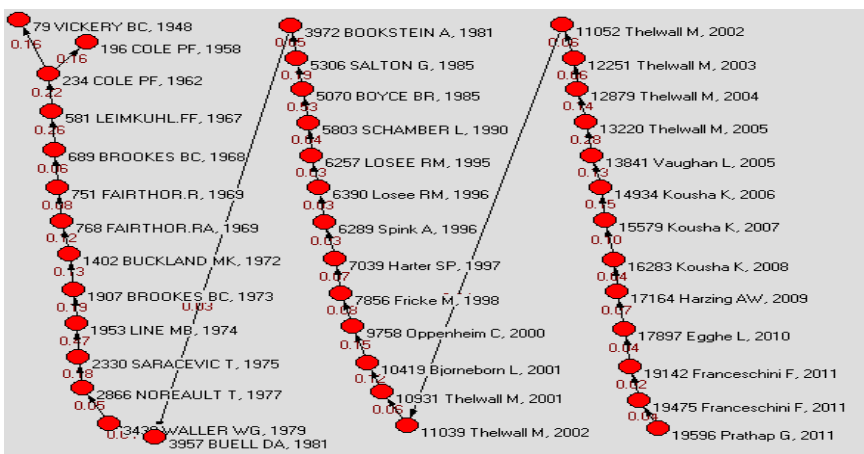


Fig 2 the main path and traversal values of Library and Information science sample

ID value and RID value of sample nodes

Traversal value reflects the degree of the importance of a node and the paper chooses those nodes with different traversal value to calculate ID value. ID value is also able to reflect the citation structure information of nodes. A higher linear relationship between traversal value and ID value means that the ID index is valid in theory.

Six representative nodes were chosen from the main path(Tab.2). 2 nodes have a higher traversal values, 2 nodes have a middle traversal values, and 2 nodes have a lower traversal values. These nodes were analyzed with Pajek to achieve k-out-neighbors (k-level backtracking references) and K-in-neighbors (k-level diffusion citation) of these sample nodes.

Tab.2 Label and traversal value of sample nodes

lable	Traversal value
581 LEIMKUHL.FF,1967	0.26
1953 Line MB,1974	0.47
5306 SALTON G,1985	0.19
9758 Oppenheim C, 2000	0.15
16283 Kousha K 2008	0.04
19142 France Schini F,2011	0.02

In order to display the composition of ID value in details, a bigger neighbor level (k=10) is chosen in the present research. To the sample nodes, the level numbers of neighbor, namely out-degree (backtracking level) and in-degree (diffusion level), are from 1 to 10, and the r_i and r_j can be obtained. So the ID value and RID value can be calculated. Taking the node 581 as an example, we can see that the number of each level backtracking document is 2、1、0、0、0、0、0、0、0、0 and the number of each level diffusion document is 24、66、156、661、1550、907、291、75、26、3. Thus the influence of node 581 can be calculated, ID602=196.05, and RID602=0.01161.The detailed data of other nodes are in Tab.3

As seen from Tab 3, ID value differs greatly from the direct references number and citation number. Let's take the node 16283 as an example. Its ID value is equal to 131.32, but the number of direct reference of (the paper published in 2008) is 14 (r_i=14) and direct citation is 15 (r_j=15), and the latter is dramatically smaller to the former.

Generally citation number is taken as the only criteria to measure academic influence. Thus, the academic influence of node 16283 is more important than that of node 9758 whose direct citation is 9 (r_j=9); and node 16823 has a better quality than node 5306 whose direct citation is 2 (r_j=2). On the contrary, in terms of ID values, node 5306(ID=185.60) is more important than node 9758(ID=149.75) and node 9758 is better than node 16283(ID=131.32).

According to RID value, the values of the first four sample nodes are less than 1,which indicates knowledge diffusion factors have a larger contribution than knowledge inheritance factors in these documents' influential elements. RID value is smaller, and the influence of knowledge diffusion is more profound. RID value of the fifth and sixth nodes is more than 1, which shows the greater contribution of knowledge inheritance factors than knowledge diffusion factors. The larger RID value is, the longer the history of knowledge inheritance is. In a further sense, the node in the front of the network has more knowledge diffusion elements, while the nodes at the back of the network have more knowledge inheritance elements.

Correlation analysis of ID value and traversal value

A positive correlation has been found between the competence of knowledge diffusion in citation network and traversal value. Hence, the document has great

power of influence and is the core document in the course of discipline evolution; accordingly, the author has great influence in this discipline. As seen from Tab. 2, node 1953 has the greatest influence (traversal value is 0.47) and node 19142 has the least influence in the network (traversal value is 0.02).

Tab 3 ID value and HID value of each sample node (i=j=10)

NO.	i(out-degree)	1	2	3	4	5	6	7	8	9	10
581	r_i	2	1	0	0	0	0	0	0	0	0
	j(in-degree)	1	2	3	4	5	6	7	8	9	10
	r_j	24	66	156	661	1550	907	291	75	26	3
	ID value	**196.05**									
	RID value	**0.01161**									
1953	i(out-degree)	1	2	3	4	5	6	7	8	9	10
	r_i	27	12	5	4	0	0	0	0	0	0
	j(in-degree)	1	2	3	4	5	6	7	8	9	10
	r_j	21	132	810	1500	710	210	54	12	3	2
	ID value	**304.14**									
	RID value	**0.11270**									
5306	i(out-degree)	1	2	3	4	5	6	7	8	9	10
	r_i	26	59	72	50	39	7	0	0	0	0
	j(in-degree)	1	2	3	4	5	6	7	8	9	10
	r_j	2	9	137	701	1090	598	279	136	47	8
	ID value	**185.60**									
	RID value	**0.40636**									
9758	i(out-degree)	1	2	3	4	5	6	7	8	9	10
	r_i	3	68	170	302	176	53	25	15	4	0
	j(in-degree)	1	2	3	4	5	6	7	8	9	10
	r_j	9	72	235	311	216	50	5	0	0	0
	ID value	**149.75**									
	RID value	**0.81121**									
16283	i(out-degree)	1	2	3	4	5	6	7	8	9	10
	r_i	14	73	175	462	430	198	65	19	13	4
	j(in-degree)	1	2	3	4	5	6	7	8	9	10
	r_j	15	34	20	8	0	0	0	0	0	0
	ID value	**131.32**									
	RID value	**4.0078**									
19142	i(out-degree)	1	2	3	4	5	6	7	8	9	10
	r_i	7	61	133	317	548	485	189	64	15	2
	j(in-degree)	1	2	3	4	5	6	7	8	9	10
	r_j	2	1	0	0	0	0	0	0	0	0
	ID value	**99.545**									
	RID value	**43.242**									

A contrast between traversal value of sample nodes and ID value reveals that: if traversal value is larger, ID value is greater, which indicates a high relevance between the two measured values. From another perspective, the constructed ID index can be proved available if relevance reaches a fairly high significance level.

The research calculates and inspects the Pearson correlation coefficient of ID value and traversal value by using SPSS 13.0.

Tab.4 The Pearson correlation coefficient of ID value and traversal value

		IDVALUE	TRANVALUE
IDVALUE	Pearson Correlation	1	.985(**)
	Sig. (2-tailed)		.000
	N	6	6
TRANVALUE	Pearson Correlation	.985(**)	1
	Sig. (2-tailed)	.000	
	N	6	6

(** Correlation is significant at the 0.01 level (2-tailed).)

It can be found from the analysis (Tab.4): there is a fairly high positive linear correlation (correlation coefficient is 0.985) between ID value and traversal value. The two-tailed testing result at 0.01 confidence level is 0.000, far less than the critical value of 0.01.The significant linear correlation is established in between. Thus, the constructed ID index scientifically reflects the citation network's structural features and the importance of the measured object. Values reflect the importance and influence of the measured document and they can be used to measure the influence of a single academic paper in the context of citation network.

Conclusion and Discussion

The influence of a single academic paper is rooted in the citation network structure. The number of direct citation and direct references cannot fully reflect its influence and indirect citation and reference in the citation network also contribute to different degrees. Among them, all of the citation, direct and indirect, show the ability to diffuse knowledge, which manifests seeping effect of the influence. All the references, direct and indirect, reflect the competence of knowledge inheritance, which manifests the accumulative effect of the influence.

In this research, ID index based the seeping effect and accumulative effect and RID reflecting constituent ratios of the two effects are constructed. The index not only changes the simple way to evaluate the academic influence of literature by means of the mere use of direct reference and direct citation, but also reveals the comprehensive and in-depth structural information of a single academic paper in the whole citation network. It is a more comprehensive and influential evaluation index.

The research chose 6 nodes, which came from the 16 Library and Information Science journals collected by Web of Science, to calculate ID index and RID index based on the traversal value of the main path nodes with the result of an indication that ID value grows with traversal value. Pearson correlation

coefficient test shows that there is a fairly high positive linear correlation in between at a strict inspection level (α= 0.01).All the results has shown ID index has a relatively high effectiveness and reliability.

Though it could fully reflect inheriting and diffusing characteristics of the measured object, the calculation of the ID index is relatively complex and must be with the aid of certain software. Thus, the practical application of the ID index would be somewhat limited. In addition, how to determine the adjustment factors of different levels (i.e. weight) is an issue for further study.

References

Batagelj, V(2003). Efficient algorithms for citation network analysis. Retrieved on 2012-11-15 http://www.imfm.si/preprinti/PDF/00897.pdf

Bornmann L, Leydesdorff L(2013). The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from Incites and F1000. Journal of Informetrics,7(2),286-291

De Nooy W, Mrvar A, Batagelj V(2005). Exploratory social network analysis with Pajek. New York: Cambridge University Press

Garfield E(2006).The History and Meaning of the Journal Impact Factor. The Journal of American Medical Associtation, 295(1), 90-93

Geisler R(2000). The metrics of science and technology. Greenwood Publishing Group Inc.

Guo Li-Fang(2005). Research on bibliometric indicator to assess the quality of academic papers. Modern Intelligence,(3),11-12(in Chinese)

Hoeffel C(1998). Journal impact factor. Allergy, 53(12), 1225

Hummon, N P, Doreian, P(1989). Connectivity in a citation network: The development of DNA theory. Social Networks, 11(1), 39–63

Jin Jing, He Miao,Wang Xiao-Ning, et al(2010). Feasibility research of evaluation and comparison of natural science papers in different fields[J].Science and Technology Management Research, 30(14), 279- 284(in Chinese).

Long Sha, Ge Xin-Quan(2007). Evaluation of academic level of scientific papers. Sci-Technology and Management,(1), 133-135,138(in Chinese)

Qiu Jun-Ping, Ma Rui-Min, Cheng Ni(2007). New approaches for evaluation of scientific researches with SCI. Journal of Library Science in China, (4),11-16(in Chinese)

Schubert, A(2009). Using the h-index for assessing single publications. Scientometrics, 78(3),559-565

Sombatsompop N, Markpin T, Yochai W, et al(2005). An evaluation of research performance for different subject categories using impact factor point average (IFPA) Index: Thailand Case Study. Scientometrics, 65(3), 293-305

Thor A, Bornmann L(2011). The calculation of the single publication h index and related performance measures: A web application based on Google Scholar data.Online Information Review, 35(2),291-300

Wu Qin(2007). Research on quality evaluation in the academic articles based on the intensity of citation. Journal of China Society for Scientific and Technical Information,26(4),522-526(in Chinese)
Yan E,Ding Y(2009). Applying centrality measures to impact analysis: a co-authorship network analysis. Journal Of The American Society For Information Science and Technology, 60(10):2107-2118
Zhang Yu-Hua, Pan Yun-Tao, Ma Zheng(2004).Evaluation methods for scientific papers. Acta Editologica,(8), 243-244(in Chinese)

THE CONSTRUCTION OF THE ACADEMIC WORLD-SYSTEM: REGRESSION AND SOCIAL NETWORK APPROACHES TO ANALYSIS OF INTERNATIONAL ACADEMIC TIES

Maria Safonova[1] and Mikhail Sokolov[2]

[1] *msafonova@eu.spb.ru*
Higher School of Economics, Soyuza Pechatnikov Str. 16, 190008, St.Petersburg (Russia)

[2] *msokolov@eu.spb.ru*
European university at St.Petersburg, Gagarinskaya Str. 3ª, 191187, St.Petersburg (Russia)

Abstract
This paper explores factors responsible for strength of various forms of academic ties between countries. It begins with examining several theoretical models of international academic collaboration: "the republic of letters", "academic (neo)colonialism", "the classical world-system", and "the world-society". Propositions about factors affecting intensity of ties between countries and configuration of their overall network are then derived from each of the models. These propositions are then tested against empirical data on two kinds of academic ties: volumes of international student flows between pairs of countries (UNESCO statistics) and number of co-authored papers (Web of Science database). Negative binomial regression is used to estimate influence of various independent variables (funding of science, distance, historical experience of dependency) about the significance of which the models make different predictions. We discover that expectations associated with "the classical world-system" fit the data best, with "academic neo-colonialist" factors also important in the case of international student flows. To account for possible differences between disciplines and to capture the directions of evolution of the system, we then explore changes in international collaboration network in two fields: geoscience and economics during a 30-year interval (1980-2010).

Conference Topic
Collaboration Studies and Network Analysis (Topic 6)

The theoretical models
Our thinking about the global system of academic collaborations is torn apart between two conflicting images. One of them, essentially optimistic, is the vision of the *international republic of letters* "as a prototype of truly open and democratic society" (Polanyi) governed by egalitarian and meritocratic norms (Merton). While modern sciences originally emerged in the West, and were exported to the rest of the world in the course of colonization, the classical modernization theory held that formerly peripheral countries would eventually

pass the stage of colonial science and develop national academies attaining full-fledged membership in the global system of division of intellectual labour (Basalla, 1967).

Another vision, essentially pessimistic, is that of hierarchical and exploitive "academic world system". This argument has been developed since the 1960s in a variety of forms, all of which vigorously opposed the earlier idealistic vision. Considering each other as allies, the adherents of these views used to downplay the disagreements in their own camp. The further classification of three types of scepticism concerning global science is thus not present in the literature itself, but can be derived through its careful review.

According to different *neo-colonial theories*, the former colonies never attain full-fledged membership in the global academic system as they remain bound to their metropolitan countries by various institutional and symbolic ties, traditional considerations of prestige, etc. (see a collection of such arguments in Sardar, 1989). The colonial infrastructure (especially educational system which was built following the imperial centre model) reproduces imperial language use and certain type of dispositions and identities (Altbach, 2004; Foner, 1979; Murphy-Lejeune, 2001; Tremblay, 2002). Moreover, since the contours of old colonial relationships are re-created at the level of contemporary international agreements in the educational sphere, application procedures and conditions of educations are simplified for young people from former colonies. That results in so-called "brain circulation": a phenomenon created by return of former skilled-labour migrants into their home countries (Cheng and Yang, 1998). Their employment in the home-country universities and research centres generates, firstly, international collaboration teams on the basis of personal networks of former migrants and, secondly, new incentives for student mobility between a former centre and a colony. That reproduces dominance of the metropolitan countries over its former colony even in absence of direct political dependency.

According to the *classical world-system theory*, in conditions of initial economic and technological inequality between the centre and the periphery most forms of interaction work to further detriment of the latter. Scholars of the "core" countries specialize on the most advanced forms of research, while the peripheral academies produce raw data, perform technical tasks, and send away their best students. The unequal division of labour arises from the fact that scholars from wealthier countries can contribute more in terms of funding, costly equipment, and infrastructure. That makes them sought-after partners and gives them an advantage in negotiating conditions of collaboration. Moreover, due to resources at hand, they have more opportunities to develop ideas produced elsewhere. The regime of academic openness gives an advantage to scholars from economically more advanced countries which also benefit from their greater potential for technological implementation of ideas. Classical world-system and neo-colonial approaches are indiscriminately united under the heading "dependency theories" (e.g. Arnove, 1980), although the former stress economic, while the latter –

institutional and cultural aspects of dependency[34], and in some respects they propose contradicting implications for how the network of global academic ties would look like. Firstly, the neo-colonial theories suggest that the networks will be clustered along the lines of former colonial allegiances, while the classical world-system implicates existence of a single relatively unified core. Secondly, the world-system approach suggests that there will be strong interaction between wealth of a national academic system and distance of its ties, with the richer having more long-distance partnerships, and the poorer less (we are not aware of this hypothesis being discussed in the literature on academic collaboration, but it parallels one well familiar from studies of international trade). The neo-colonial theories believe that transportation costs are secondary to different types of transaction costs arising from institutional and cultural closeness (North, 1991).

In its original formulation, world-system theorizing suggested that their will be lack of scholarly activity in the peripheral countries altogether, with most kinds of intellectual production concentrated in the core countries (the possible exceptions were types of research directly meeting demands from backward peripheral economies). Empirical studies demonstrated, however, that the spread of higher education and research sectors in the XX century was surprisingly uniform in all countries irrespective of their level of economic development. The *world-society theories* developed by John Meyer and his many associates sought to explain this fact by pointing to emergence of single rationalized global culture which is imitated even in absence of any direct economic pressure to do so (Schoffer, 2003; Meyer and Schoffer, 2005). A few studies in the "world society" tradition demonstrated that, counter to what economic determinism of the classical world-system analysis assumes, scale of national investments in research and research education neither responds, nor immediately contributes to economic growth, or may be even detrimental to it (Shenhav, 1993; Schofer, Meyer and Ramirez, 2000). The world-society perspective differs from the neo-colonial approach in focusing on singular world-society, rather than dispersed academic empires. The patterns of collaboration in this world-society, however, emerge under the pressure of cultural, rather than economic, necessities. The classical world-system implies that centrality of a given national academy in the network of international academic ties is directly related to its economic prosperity. This pattern is likely to be most salient in the case of capital-intensive disciplines, involving high-cost experimental or field research. The world-society assumes that prosperity will be secondary to traditional intellectual prestige of a given country (not necessarily related to its prosperity), and that there will be no differences between disciplines.

Parallel to that theoretically-driven efforts, a bulk of more empirical research on international collaboration emerged in the scientometric tradition which demonstrated, among other things, a strong tendency for geographic localization

[34] The distinction between the two arguments becomes somewhat blurred if we consider institutional-economic factors, which are typically omitted from world-system theorizing, but figure prominently in sociological approaches to migration (namely, migration systems theory (Kritz, Lim and Zlotnik, 1992) and migration network theory (Gold, 2005; Massey at all., 1993).

of academic collaboration (e.g. Luukonen *et. al* 1992; Zitt *et al*. 2000). Regretfully, to our knowledge, there were no attempts so far to control for influence of other variables (e.g historical or economic, which are likely to be intertwined with purely geographic). The only partial exception seems to be (Nagpaul, 2006), although his paper does not account for cultural or historical factors. There were little attempts to bring closer the quantitative bibliometric and more historical and sociological literatures (but see Schott, 1998 for an early exception). The most recent theoretical formulations emerging in scientometric literature tended to downplay the role of external factors in formation of academic ties altogether, pointing to self-organization system properties of networks (Wagner and Leydesdorff, 2005a; Wagner and Leydesdorff, 2005b). This contradicts, however, many early findings which demonstrate prominence of extrinsic factors in tie-formation on micro-level. The network science, nevertheless, offers a valuable null-hypothesis which states that none of the characters of the pairs of countries will influence the intensity of ties between them, except their sheer size.

Table 1. Theoretical models and their empirical implications: effects of various variables on intensity of academic ties

Table	*Republic of letters*	*Neo-colonial*	*Classical world-system*	*World-society*
Wealth and research funding	Secondary importance or none	Secondary importance or none	Primary importance, especially for capital-intensive disciplines	Secondary importance or none
Physical distance	Secondary importance or none	Secondary importance or none	Primary importance for poorer countries	Secondary importance or none
Institutional ties	Secondary importance or none	Primary importance	Secondary importance or none	None
Overall network pattern	Cohesive. Clustering, if any, based on national research priorities (in applied fields).	Strongly clustered, homophily between institutionally coupled countries	Classical core-periphery, with position solely dependent on wealth; clustering at periphery based on proximity	May be cohesive or core-periphery; centrality (if any) based on established prestige, no clustering.

The aim of this paper is to evaluate the theoretical models listed above by exploring systematically factors responsible for formation and strength of academic ties. We take two types of ties, corresponding to different stages in academic careers and different, though overlapping, sectors of academic institutions: (1) international student migration flows and (2) scholarly paper co-

authorships. We then try to evaluate the models in two ways: directly, by using regression analysis to predict intensity of migrations and collaborations between pairs of countries, and indirectly, by applying social network analysis measures to evaluate overall network pattern (see Moody, 2004 for an exemplary studies). The rationale behind these steps is that different models have different implications for which factors will strengthen the ties, and how the whole network will be organized. These implications are summarized in Table 1.

The expectation based on the republic of letters model is the existence of relatively cohesive network, the centrality of position of a specific country in which is primarily determined by the number of students and high-school teachers (in the case of student migrations) and total academic personnel (in the case of co-authorships) available in it. Homophily, if any, occurs between countries which similar research priorities (which might arise from common economic necessities, especially in the case of applied sciences). Other factors are deemed secondary.
In the case of neo-colonial model, the network becomes highly clustered with clusters corresponding to former colonial empires. Here instead of a single core-periphery structure, a series of such structures emerges, with each core country having its own periphery, most likely, the one to which it has exported language and educational institutions as a colonial centre. The student flows from former colonies go to the former metropolitan countries, with metropolitan students mostly studying at home. Co-authorship also occurs mostly inside the boundaries of former empires. Prestige of a traditional centre might be reproduced even in presence of economically strong rivals, thus making wealth secondary. We could expect that, due to low transaction costs (common language and institutional similarity), intensive academic ties will emerge between pairs of former colonies of a single imperial centre as well.
In the case of the classical world-system, our expectation would be that the whole system is patterned as a prototypical core-periphery structure with few, if any, contacts between peripheral agents situated in close proximity to each other. Its exact shape may vary with the character of the discipline. In the case of capital-intensive disciplines, scholars from one prosperous academy would prefer partners from academies which are also prosperous, especially when production and consumption of knowledge in a given discipline is global. Similarly, students from wealthier countries have more chances to study abroad as they are more likely to get scholarships at home or to invest family resources; that makes them much more attractive entrants the point of view of universities, especially the private ones. Thus, reasoning by analogy with what sociologists of science observed at the intra-national level, a system aptly called “academic castes” emerges (Burris, 2004). According to it, the academic world is a stratified system, in which exchange is limited to the members of the same strata. Projecting it on the global system of academic collaboration, one might expect that the academics from the core-countries are likely to be overrepresented among the co-authors of academics from other core countries, while semi-peripheral academics would

have to look for partners in their own league; the academics from the periphery probably would find themselves isolated. The picture may be different in disciplines which are labour-intensive, or in which knowledge is locally produced (e.g. involving excavations or cross-cultural comparisons). Here the caste barriers disappear, although direct collaboration between peripheral countries still rare. Corresponding pattern in organization of student flows would look like system of asymmetric exchanges with upper-caste countries sending incoming flows to each other, while lower castes send flows to them without receiving any students in response.

Finally, in the world-society model scholars and academic institutions from all countries are under equally strong pressure to collaborate internationally, as that increases legitimacy of their work. They might be quite indifferent between particular partners (producing cohesive network), or to prefer partners belonging to the academic systems which are considered to be in highest compliance with the requirements of the "world society" (producing a core-periphery structure). Being a paragon of "world society", however, does not necessarily depend on wealth or funding. Here we do not expect to find principal differences between disciplines as all of them have to demonstrate compliance to a single legitimate pattern.

Data, measures and methods

The major sources of data were, firstly, UNESCO Institute of Statistics, and, secondly, Thomson Reuter's "Web of Science" databases. For all regression calculations, 2007 year was used as the data on it were the most complete of those available. We included only those countries for which at least population and the GDP data were available, which gave us a sample of 181 cases. In addition to that, we gathered co-authorship data on two disciplines (geoscience and economics) for 1980, 1990, 2000 and 2010 years. The rationale for choosing these particular disciplines was that we wanted to have a natural and social science to compare. Of the social sciences, only economics seemed suitable as other disciplines simply do not produce enough cases of international co-authorship in a year. To match it on the part of the natural sciences, we wanted to find one which would be the closest in the sense of its results being at least partly locally produced and locally consumed. Geoscience seemed the best fit from this point of view.

The dependent variables were two kinds of links between pairs of countries – (1) volume of international student flows, (2) number of papers scholars from them co-authored in the Web of Science database. The independent variables were either attributes of the countries (GDP, tertiary student population) or characters of relations between them (proximity, experience of colonial dependency or co-dependency).

Dependent variables

Student flows

The data on student flows between pairs of countries were taken from UNESCO datasets which accumulates reports from recipient countries on the numbers of foreign students coming to study in them. The data for 2007 were available from 73 countries (of 209 UNESCO recognizes). Western countries were heavily over-represented in this sub-sample (as nearly all EU countries have produced required statistics). That posed a problem for further analysis. Including all existing data on international flows (12977 valid cases) would probably result in over-estimation of whatever factors influenced volume of flows from non-Western to Western countries as the cases of flows between non-Western countries would be disproportionally under-represented; at the same time, limiting the sample to the 73 countries which have published statistics (5285 valid cases) would exclude most non-Western countries altogether, and thus under-estimate influence of variables pertaining to core-periphery differences. As a solution, the analysis had been performed on both extended and reduced samples. Predictable changes in coefficients occurred, but no significant differences were observed. Below calculations performed on extended sample are reported.

Co-authorships

The data on co-authorships between pairs of all 209 countries included into UNESCO dataset were extracted from Web of Knowledge Science, Social Sciences, and Arts & Humanities databases for 2007; papers in all languages were included, but conference proceedings omitted. A difficulty arouse from the fact that the UNESCO and WS lists of countries differed. For example, WS does not provide users with separate data on Macao or Hong Cong (which are treated by UNESCO as state-type entities); at the same time, it recognized England, Scotland, Wales, and Northern Ireland as separate states, but we had to merge them as UNESCO provides only aggregated statistics. That reduced the list of cases suitable for analysis to 177 countries; we thus had 15576 ($(177^2 - 177) / 2$) valid pairs.

Independent variables

Populations

All four models recognize the importance of the size of academic populations which thus functions as a control variable. UNESCO gathers data on (a) *numbers of tertiary students* studying in the country (estimate of potential student flow from a country); (b) *numbers of higher education teachers* (estimate of the accommodating capacity of a given national higher education system)[35]; (c)

[35] This variable is not truly independent as, in the long run, it is endogenously determined by the size of the flow. Thus, it was not included in the analysis.

numbers of researchers (estimate of numbers of potential co-authors). The problem which plagued these data were missing values; as of 2007, 149 countries (of our 177) provided data on tertiary enrolment, 127 – on numbers of high school teachers, and only 97 on researchers in head count; 4 more did that in full time equivalent.[36] Again, the Western cases were heavily over-represented.

Proximity

We used the UNESCO classification of countries into 21 regions and converted these data into binary variables, assigning "1" if both countries belonged to the same or adjacent regions, and "0" otherwise.

Wealth and gross academic expenditures

(a) Country wealth was estimated by GDP per capita (available for all countries in all three of the samples, source – UN statistics); *(b) National academy's wealth* was estimated by Gross Expenditure on Research and Development (GERD) per researcher. GERD data was available for 105 countries (in 15 cases data were extrapolated from adjacent years in the interval from 2004 to 2008). In 94 of these 120 cases, the data on numbers of researchers were also provided. Availability of GERD data was the single most important limiting factor on selecting the valid cases for analysis of co-authorships; what is more, selection of cases on the basis of availability of statistics on research again favors Western cases against non-Western which is likely to somewhat downplay the importance of the next group of factors. We used GDP per capita as a proxy for wealth of the national academy for most calculations, as it allowed avoiding loss of cases and the correlation between this measure and GERD per researcher reaches .5 size.

Political dependency and co-dependency

We used historical experience of dependency both to directly test the neo-colonial model and as a most general proxy for probability of massive institutional import. We created a binary variable, assigning "1" if one of the countries at certain moment since 1648 were governed by the central government situated at the territory of the other, and "0" otherwise.[37] A former colony can be economically and political successful, and create strong national academy, or even establish its own quasi-colonial system (as the US did, see Mann (2008). In that case it would benefit from primary language and institutional export of its former metropolitan country, and compete with it for oversees resources (students and collaborators). The US and the UK, or Germany and Austria could serve as examples. To account for this fact, we created an additional *"political co-dependency"* variable,

[36] We converted FTE in HC by dividing it by 0.62 (average, S.D. = 0.06) to receive 101 valid attributes.

[37] That not necessarily means colonial dependency. In some cases we dealt with dissolved political unions of a more egalitarian character, e.g. Czechoslovakia. Colonial empires, however were by far a modal case.

"1" for countries which were in certain moment under rule of a central government situated in a territory of a third country and "0" otherwise.
A square matrix was created, where relations of a colony or a dependent territory and a metropolitan country coded as a link between the two countries (binary).[38] Data on absence or presence of a tie were used as an independent variable ("historical experience of political dependency of B from A"). Matrix with geodesic distances was created from the first matrix, and data on geodesic distance of two were extracted to include relationship of belonging to one "colonial neighborhood" as an independent variable into the model ("historical experience of political co-dependency of B and C from some A").

The Regression Model

Both dependent variables were distributed obviously non-normally, with zero being the modal value. Moreover, their standard deviations were much greater, than mean, signaling overdispersion. The distribution closest to the observed would be the negative binomial one. To deal with overdispersion, the scale parameter has been set equal to deviation. Interaction terms for countries' wealth, and research expenditures were included. Regression with robust error variance was used to help remedy non-independence of cases.

Results

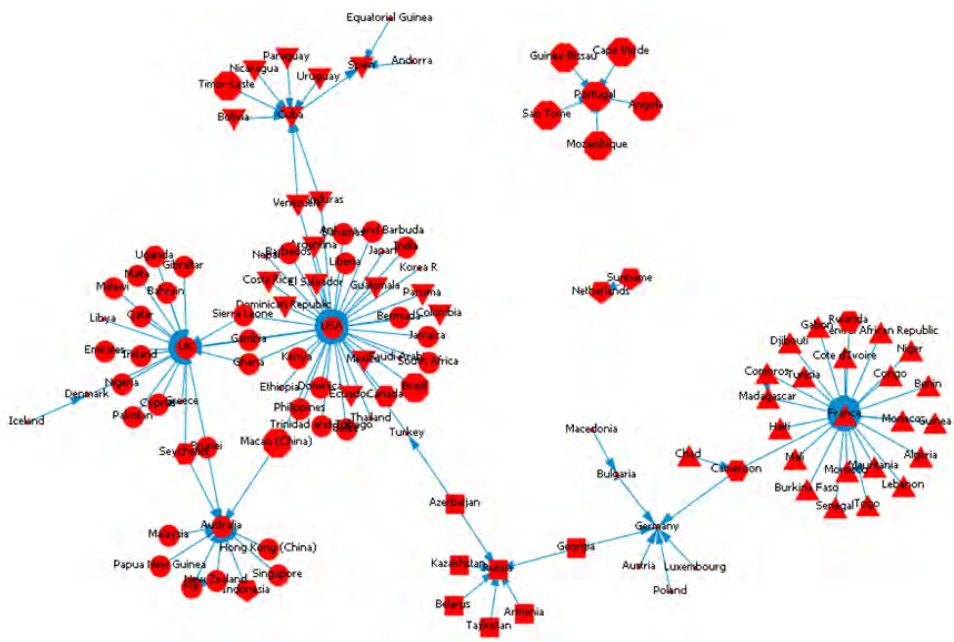
Table 2 shows results of regression of volumes of student flows between countries on independent variables (extended sample, 73*181).

Table 2. Regression model predicting volume of a student flow between pairs of countries

Table	*Wald Chi-Square*	*Sig*	*Exp (B)*
Intercept	221,826	,000	2,866
Tertiary student in the sending country (MLN)	277,694	,000	1,313
Proximity	613,969	,000	11,746
Sending country GDP per capita (PPP $000)	28, 235	,000	,981
Receiving country GDP per capita (PPP $000)	1016,419	,000	1,084
Dependency	215,605	,000	28,827
Co-dependency	144,314	,000	3,308
Interaction: Proximity & Sending Country GDP	47,434	,000	,968
Interaction: Dependency & Sending Country GDP	22,707	,000	,953
Interaction: Co-dependency & Sending country GDP	,432	,511	1,004
Interaction: Countries 1&2 GDP	103,988	,000	1,001
Likelihood Ratio Chi-Square 5431,613		,000	

[38] The date was chosen rather arbitrarily, as a traditional landmark in the history of international relations. If part of a presently existing country were colonized by another, while others were not (e.g. parts of China under British and French rule), we assigned "1" if the respective part exceeded 10% of the present country's territory.

The coefficients demonstrate, first of all, predictable importance of size of the emitting academic world. The volume of student flows between countries is positively correlated with wealth of the destination country, but negatively – with wealth of the country of origin. Richer academic systems posses a sort of social gravitation which attracts student flows from outside, at the same time keeping students from inside from leaving. Other things being equal, students from wealthier countries are less likely to study abroad, and if they become internationally mobile, they chose other prosperous countries. At the same time, the directions of mobility are heavily pre-determined by historical and institutional factors. That finding could be easily supported by inspection of a map of flows, bringing with them more, than 30% of international students from a given country (ORA visualizer used). These deep migration channels link former imperial centres with their once-colonies.[39]



Picture 2. "Deep channels" in international student migration

Co-dependency is significant as well, but less so. Proximity also plays role, albeit two and a half times less massive, than former dependence, judging from Exp(B) coefficients. Finally, proximity and wealth, and dependence and wealth interact, showing that (1) there is a difference in the range of educational migration by students from poorer and wealthier countries with the latter travelling further; (2)

[39] The form of the sign corresponding to a node shows the language which is used in the country as official (English, Spanish, French, Russian, Dutch and Portuguese respectively). I-E Index for language attribute is -0.349, $p < 0.0001$.

students from poorer post-colonial countries are more likely to travel to their metropolitan country, benefiting from its paternalistic policies and relying on existing migration channels to save on transaction costs.

These findings obviously do not fit with the vision of "flat world", equally open to all, which "republic of letters" implies. In addition to evidence of effects of all kinds of economic and political factors, one observes that the countries gravitating towards each other tend to be dissimilar in terms of academic development, and, thus, are unlikely to have similar political priorities. The results even less fit with the world society theory which is probably overestimating spread of rationalizing culture around the globe, especially as far as higher education sphere is concerned. The "core" of the academic world system is fractured between older and newer colonial powers. Overall, both versions of the dependency theory receive some support: we do see academic castes, and we do find a heavily clustered network, especially at the periphery.

The picture changes as we turn to international network of co-authorships. Table 3 summarizes what we observe there.

Table 3. Regression model for co-authorship

Table	*Wald Chi-Square*	*Sig*	*Exp (B)*
Intercept	270,576	,000	,319
Researchers in both countries (UNESCO head count, 000)	1738,267	,000	1,004
Proximity	367,200	,000	5,283
GDP per capita in country 1 (PPP $000)	459,746	,000	1,059
GDP per capita in country 2 (PPP $000)	491,871	,000	1,058
Dependency	22,072	,000	2,836
Co-dependency	88,928	,000	2,651
Interaction: Countries 1&2 GDP	20,695	,000	1,001
Interaction: Proximity & GDP per capita	48,518	,000	,974
Interaction: Dependency & GDP per capita	,008	,930	1,001
Interaction: Co-Dependency & GDP per capita	22,158	,000	0,976
Likelihood Ratio Chi-Square 13353,875		,000	

As co-authorship relationships are essentially symmetric (or, at least, there are little opportunities to decipher any asymmetries authors lists conceal), we summarized data on populations of researchers in both countries to obtain combined variable. Not surprisingly, it is highly significant. GDP per capita in both countries and their interaction are significant as well, signaling the tendency of academics from more prosperous academic worlds to look for other resourceful partners. Producers of scientific papers are divided into economic strata.[40] Finally,

[40] We are not discussing here the possibility that participation of scholars from less resourceful academic systems is not recognized by authorship. More detailed case research is necessary to prove or falsify this disquieting suspicion

colonial variables retained their significance, albeit at a diminished scales. Proximity matters more, and we encounter again interaction between long-distance collaborations and wealth. A significant detail is that while in migration equations exp (B) coefficients for colonial dependency were twice as large as they were for colonial co-dependency, here they draw much closer. An interpretation of this might be that policies of former metropolitan countries which advantage students from former colonies are usually not spread to adult academics. Finally, we find interaction between wealth and co-dependency, but not wealth and dependency. It means that scholars from poorer academies tend to co-author papers with scholars from other countries formerly dependent from the same colonial centre, but not from the centre itself, probably pointing to the fact that former colonies and former metropolitan countries tend to belong to different "academic castes".

Overall, we see that the pattern of international co-authorship even more clearly follows the expectations based on classical world-system, than that of student migration flows. Formation of research partnerships are obviously not completely a stochastic results of network growth, as proponents of the network science would like us to think (Wagner and Leydesdorff, 2005a); the geographic, economic, and institutional factors together produce McFadden R^2 of 0.215. The academic castes are quite salient with scholars from wealthier academic worlds preferring their likes. Cultural and historical legacies remain significant, although at a lesser scale.

The cases of geoscience and economics

At the final stage of our analysis, we looked at evolution of two specific fields to find out, if there are differences between disciplines, and if the development occurring in them is in one and the same direction. Table 4 presents data on economics, Table 5 – on geoscience.

Table 4. Parameters of economics network

Table	*1980*	*1990*	*2000*	*2010*
Nodes	56	63	107	140
Edges	102	164	752	1790
Density	0.033	0.042	0.066	0.092
Transitivity	8.83%	6.46%	16.14%	21.73%
Clustering coefficient	0.225	0.172	0.366	0.454
Centralization (Degree)	8.54%	5.82%	5.00%	4.32%
GK Gamma correlation with Dependency	0.421	0.685*	0.778***	0.810***

In accordance with already reported findings (Wagner and Leydesdorff, 2005b), a rapid growth occurs in both networks, which are also becoming denser and less

centralized.[41] In contradiction to what Wagner and Leydesdorff propose, however, transitivity and clustering coefficients in both networks grow significantly as well, meaning that international collaboration in both disciplines becomes at the same time more, rather than less, fragmented. We calculated Goodman-Kruskal Gamma correlation between dependency and intensity of tie. Astonishingly, the correlation rose from insignificant to very strong. The growth of international collaboration makes the contours of academic empires more, rather than less, visible. Equally surprisingly, there were no marked differences between disciplines, hinting that the processes of academic globalization do not depend on usually assumed epistemological differences between social and natural sciences as such.

Table 5. Parameters of geoscience network

Table	*1980*	*1990*	*2000*	*2010*
Nodes	59	86	140	158
Edges	190	462	1522	3322
Density	0.055	0.063	0.078	0.134
Transitivity	14.27%	15.71%	19.62%	23.05%
Clustering coefficient	0.333	0.359	0.423	0.473
Centralization (Degree)	9.02%	7.89%	5.67%	4.95%
GK Gamma correlation with Dependency	0.594*	0.852*	0.852***	0.898***

Concluding remarks

Obviously, the attractive vision of "republic of letters" is far from harsh realities of international Academy in which economic inequality is central to setting patterns of collaboration and mobility, and inherited cultural and institutional divisions remain all-pervasive. There is a tendency for the scholars from the core countries to form closed clubs by choosing co-authors from wealthier countries as partners. Overall, it seems that academic co-authorships tend to form academic caste structures (as the world-system theory predicts), while student mobility flows are more segmented by colonial legacies (as neo-colonial theory predicts). Finally, the uniformity of the world-societal pressures are probably strongly overestimated as far as academic world is concerned.

No doubt, taking into account the limitations of data processed, these conclusions are to be treated as tentative at best. Including more formal measures of similarity of research profiles of countries (e.g. based on distribution of their publications among different categories in Web of Science) is necessary to do more justice to the "republic of letters" model. An obvious omission of this study is

[41] These considerations do not take into account distortion which may arise from logic of growth of the Web of Science database. Increasing density might be an outcome of wider inclusion of peripheral periodicals, rather than actual growth of collaboration (Passi, 2005). To our knowledge, however, no remedy for potential bias emerging from this has been offered so far.

indiscriminate usage of one measure for "dependency". A variety of measures should be computed to take into account (a) the longevity of belonging to a common political system; (b) the particular historical period of belonging; (c) the part of territory covered by it; and (d) other historical particulars of colonization. Imperial centres differed in their approach to exporting educational institutions to the colonized territories, and some of them attempted to meticulously reproduce metropolitan Academia on the new soil, while others did not care much about institutional export at all, or even imported institutions from territories they happened to acquire (as Muscovy, and later the Russian Empire, from Ukraine and the Baltic region). More historical analysis is necessary to account for such differences. Finally, larger sample of cases of academic specialties is necessary to reach any reliable conclusion about differences between disciplines. This list is to include capital- and labour-intensive specialties (intuitively, geoscience seems much more capital-intensive, than economics, but some more formal measures are desirable here). Varieties along the dimensions of local-global production and consumption of knowledge are to be appreciated as well. All these suggests some avenues for further work.

References

Altbach P.G. (2004) Globalization and the university: myths and realities in an unequal world. *Tertiary Education and Management*, 10: 3–25.

Arnove R. A.(1980) Comparative education and world-system analysis. *Comparative Education Analysis,* 24(1): 48-62

Basalla, George. (1967) 'The spread of western science.' *Science*, 156(3775): 611-622

Burris V. (2004) The academic caste system: prestige hierarchies in PhD exchange networks. *The American Sociological Review*, 69(2):234-264

Cheng L. and Yang P.Q. (1998) Global interaction, global inequality, and migration of the highly trained to the United States. *International Migration Review*, 32(3): 626-653.

Fawcett J. T. (1989) Networks, linkages, and migration systems. *International Migration Review*, 23: 671-680.

Foner N. (1979) West Indians in New York City and London: a comparative analysis. *International Migration Review*, 13(2): 284-297.

Gold S.J. (2005) Migrant network: a summary and critique of relational approaches to international migration. In: Romero M. and E. Margolis (eds.) *The Blackwell Companion to Social Inequality*. (pp. 257-285). Oxford: Blackwell Publishing,.

Kritz M., Lim L.L. and Zlotnik H (eds.) (1992) *International Migration Systems: A Global Approach*. Oxford: Clarendon Press

Luukkonen, Tertu et al. (1992) Understanding patterns of international scientific collaboration. *Science, Technology, and Human Values*, 17(1): 106-126

Mann M. (2008) American empires: past and present. *Canadian Review of Sociology*, 45 (1): 7–50.

Moody, James (2004) The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999.' *American Sociological Review*, 69(2): 213-238

Meyer, John & Evan Schofer. (2005) The worldwide expansion of higher education in the twentieth century.' *American Sociological Review*, 70(6):898-920

Massey D.S. at all. (1993) Theories of international migration: a review and appraisal. *Population and Development Review*, 19 (3): 431-466.

Murphy-Lejeune E. (2001) *Student Mobility and Narrative in Europe: The New Strangers*. London: Routledge

Nagpaul, P.S. (2006) Exploring a pseudo-regression model of transnational cooperation in science. *Scientometrics,* 56(3): 403-416

Passi, Anssi. (2005) Globalization, academic capitalism, and the uneven geographies of international journal publishing spaces. *Environment and Planning*, 37: 769-789

Portes A. (1997) 'Immigration theory for a new century: some problems and opportunities. *International Migration Review*, 31(4):799-825.

Zardar, Z. (Ed.) (1989) *The Revenge of Athena. Science, Exploitation and the Third World*. Mansell

Schofer, E. (2003) 'The global institutionalization of geological science, 1800 to 1990.'

American Sociological Review, 68(5): 730-759

Schott. T. (1998) Ties between centre and periphery in the scientific world-system: accumulation of rewards, dominance and self-reliance in the centre. *Journal of World-Systems Research*, 4: 112 - 144.

Shenhav, Y. (1991). 'The 'costs' of institutional isomorphism: science in non-western countries.' *Social Studies of Science*, 21(3): 527-545

Schofer, E, Ramires, F., & Meyer, J. (2000) 'The effects of science on national economic development, 1970-1990.' *American Sociological Review*, 65(6): 866-887

Tremblay K. (2002) Student mobility between and towards OECD countries: a comparative analysis, in: *International mobility of the highly skilled*. (pp. 39-67) Paris: OECD

Wagner, C. Leydesdorff, L. (2005a) 'Network structure, self-organization and the growth of international collaboration in science' *Research Policy*, 34:1608-1618

Wagner, C. and Leydersdorf, L. (2005b) 'Mapping global science using international co-authorships: a comparison of 1990 and 2000.' *International Journal of Technology and Globalization*, 1(2): 185-208

Zitt, M., Bassecoulard, E., Okubo, Y. (2000) Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics*, 67: 627-657

CONSTRUCTION OF TYPOLOGY OF SUB-DISCIPLINES BASED ON KNOWLEDGE INTEGRATION

Qiuju Zhou Fuhai Leng

zhouqj@mail.las.ac.cn, lengfh@mail.las.ac.cn
National Science Library, Chinese Academy of Sciences, Beijing, 100190, People's Republic of China

Abstract
This investigation recalled the frameworks proposed by Stirling (2007), Rafols & Meyer (2010), Liu et al., (2012) and Zhou et al., (2012). The bibliometric methodology presented here provides an overview of scientific sub-disciplines, with special attention to their interrelation. This work aims to establish a tentative typology of disciplines and research areas according to their degree of knowledge integration. Knowledge integration is measured through diversity based on the Subject Categories mapping from the references of the articles set in sub-disciplines. The similarity-weighted cosine was used to measure the interrelations between sub-disciplines.

Conference Topic
Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8)

Introduction
In a recent article, Rafols and Meyer (2010) presented an analytic framework for the study of interdisciplinarity. And the framework was enriched by Rafols et al. (2012), Wagner et al. (2011) and Leydesdorff and Rafols (2011). The two main factors of this framework are diversity and coherence. Zhou et al. (2012) – inspired by Rafols and Meyer (2010), Stirling (2007) and related ecological research (Nei & Li 1979; Shriver et al. 1995) on the relationship of diversity within populations and the similarity between populations – proposed a generalize framework to study systems' diversity and the similarity (homogeneity) of systems. Zhou et al. (2012) then applied it to the research profile of countries to present the unbalanced and concentrated disciplinary structure of 32 countries.
Liu et al. (2012) synthesized the main points of the Rafols-Meyer approach, and showed how these ideas can be applied to knowledge diffusion and knowledge integration.
In this article, we want to apply the framework proposed by Zhou et al. (2012) to knowledge integration, following Rafols and Meyer (2010), Leydesdorff and Rafols (2011) and Liu et al. (2012). We further aim to analyze the sub-disciplinary structures based on knowledge integration. For case studies we use articles from the various sub-disciplines of ecology and analyze their knowledge integration, as revealed through their references.

The objectives of this paper are threefold.
1) Give an overview of the three analytical frameworks.
2) Find the theoretical foundations for using knowledge integration to explain the framework proposed by Zhou et al. (2012) in the construction of a disciplines structure.
3) Apply the framework proposed by Zhou et al. (2012) to analyze the knowledge integration of selected sub-disciplines in ecology and use this to evaluate the homogeneity of sub-disciplinary structure.
The article is organized as follows. In Section 2 we recall the main points from the three frameworks proposed by Rafols and Meyer (2010), Liu et al. (2012) and Zhou et al. (2012). Section 3 gives the details for a study of knowledge integration within the framework proposed by Zhou et al. (2012). Section 4 provides a case study related to sub-disciplinary structures based on knowledge integration and discusses the results. Section 5 concludes by summarizing the results and discussing their implications and limitations within current research and also discusses issues for further research.

Overview of the three analytical frameworks

In this section we mainly describe the work by Stirling (2007), Rafols and Meyer (2010), Zhou et al. (2012), and Liu et al. (2012).

Overview of the framework for the study of diversity and coherence proposed by Rafols and Meyer (2010) and Liu et al. (2012)

Stirling (2007) proposed a framework of diversity for understanding any system of science and technology. Rafols and Meyer (2010) further developed this by proposing a framework for understanding interdisciplinary through diversity and coherence. Their understanding of diversity was as a measure of the variety of categories used, while coherence explains the interrelatedness of categories and topics. Rafols and Meyer (2010) first applied this framework to a single article. Then, Rafols et al. (2012) applied the concepts to whole groups of related articles, using as case studies an entire university's and a department of a university's output (Rafols et al., 2012). In Rafols and Meyer (2010) diversity measures are based on the JCR categories of the references-of-references, while coherence is understood through the strength of bibliographic coupling in the network of references.
Figure 1 show the Conceptualisation of interdisciplinarity in terms on knowledge integration.

Liu et al. (2012) introduce a general framework for the analysis of knowledge integration and diffusion using bibliometric data. They considered a framework that consists of three entities: (1) the source; (2) the intermediary set (IM) derived from the source; and (3) a target set.
The specific operationalization of diversity and coherence may differ in these empirical studies (due to their different goals, focus, sample size, etc.), but having

a conceptually well-defined framework is important for the sake of clarity and in order to be able to compare cases.

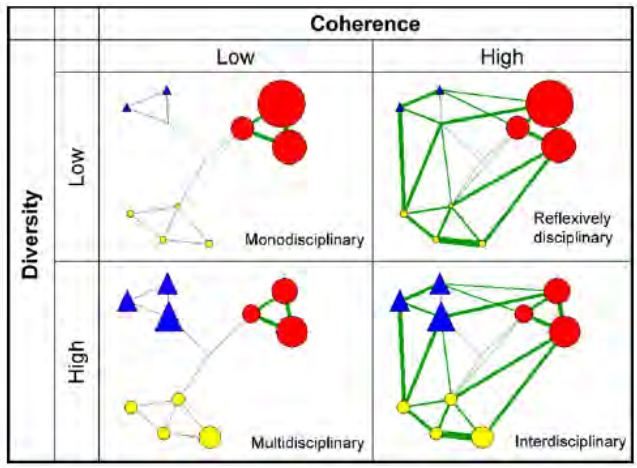


Figure 1. Overview of the proposed framework for the study of diversity and coherence

Overview of the framework for the study of diversity and similarity proposed by Zhou et al. (2012)

Recalling the framework proposed by Zhou et al. (2012), the key innovation was the addition of a measure of similarity between systems. They were inspired by Rafols and Meyer (2010), Stirling (2007) and related ecological research (Nei & Li, 1979; Shriver et al., 1995) on the relationship of diversity within populations and the similarity between populations. They proposed a general framework to study systems' diversity and the similarity of systems.

Diversity within systems can be described, using the number of categories (number of species in ecology; number of scientific fields in scientometric studies), as variety and evenness; or one can go one step further and include a measure for the *disparity* between categories. So, diversity can be said to contain variety, eveness, and disparity. Many early cases used a classical measure, such as the Gini evenness index or the Simpson diversity measure. If there exists any disparity between categories, it is more appropriate to use the Rao-Stirling diversity measure proposed by Rao (1982) and Stirling (1998, 2007).

Similarity between systems can be studied by using a classical similarity measure such as Salton's cosine measure or a weighted form taking category similarity into account. For this purpose Zhou et al. (2012) proposed the similarity-weighted cosine measure. All this is presented in Table 6. This is an expansion on the ecological work of Nei & Lei (1979) and Shriver et al. (1995). It shows that their framework can be applied to any system.

Table 1. Overview of the framework for the study of diversity and similarity proposed by Zhou et al. (2012)

	Diversity within systems	**Similarity between systems**
With a given number of **independent categories** in the system	Simpson diversity measure, $L_X = 1 - \sum_{i=1}^{N} p_{Xi}^2$	Salton's cosine measure, $\lambda(X,Y) = \frac{\lambda_{X,Y}}{\sqrt{\lambda_X . \lambda_Y}} = \frac{\sum_{i=1}^{N} p_{Xi} p_{Yi}}{\sqrt{(\sum_{i=1}^{N} p_{Xi}^2)(\sum_{i=1}^{N} p_{Yi}^2)}}$
Taking **disparity (or similarity) between categories** into account	Stirling-Rao measure, $SR_X = \sum_{i,j=1}^{N} p_{Xi} p_{Xj} D_{ij} = 1 - \sum_{i,j=1}^{N} p_{Xi} p_{Xj} S_{ij}$	similarity-weighted cosine measure, $\varphi(X,Y) = \frac{\varphi_{X,Y}}{\sqrt{\varphi_X \varphi_Y}} = \frac{\sum_{i,j=1}^{N} p_{Xi} p_{Yj} S_{ij}}{\sqrt{(\sum_{i,j=1}^{N} p_{Xi} p_{Xj} S_{ij}) . (\sum_{i,j=1}^{N} p_{Yi} p_{Yj} S_{ij})}}$

Our understanding of knowledge integration and disciplinary structure

Knowledge integration

Knowledge integration can be described as a property of an article (Porter et al., 2007) or a set of articles (Liu et al., 2012). In this research, we want to focus on the analysis of specific areas (a set of articles). The case study is based on the publications in SCI database of several sub-disciplines of ecology.

What are the knowledge integration's breadth and intensity? From how many SCs is the knowledge of a certain sub-discipline derived? Which SCs contribute the most to the sub-discipline? What is the relationship between sub-disciplines? Can we construct the typology of sub-disciplines based on knowledge integration? We are trying to answer these questions with the frameworks proposed by Rafols and Meyer (2010), Liu et al. (2012) and Zhou et al. (2012).

Diversity as one attribute of knowledge integration

Diversity is the property of how the elements of a system are apportioned into categories (Stirling, 2007). As an aspect of knowledge integration, diversity is now based on the image of the categories-mapping (Liu et al., 2012). As explained by Rafols and Meyer (2010) the best approach is to take the three aspects of diversity – i.e. variety, balance and disparity – into account. If a distance or dissimilarity measure exists in the categories, this suggests the need to use the Rao-Stirling diversity measure.

We can consider systems to be the sub-disciplines, the elements to be the references of the sub-disciplines, and the categories to be the SCs to which the references belong. This way the three aspects of diversity can measure the different aspects of the level of knowledge integration in contributing sets:

- Knowledge Integration Breadth (Variety)

Variety is the number of categories of elements, in this case, the SCs into which publications or references can be partitioned. We followed Liu and Rousseau (2010) and Liu et al. (2012) calling variety "SCs knowledge integration breadth", the number of SCs (or ESI fields) in which a set of articles is cited.

- Knowledge Integration Intensity (Balance) – The distribution of the 172 Subject Categories (172SCs). The 172SCs are defined by the JCR system.

We define SCs knowledge integration intensity as the distribution of the 172SCs.

- Knowledge Integration Intensity (Balance + Disparity) – The distribution of the 14 disciplines (or 22 broad fields).

Within the Essential Science Indicators (ESI) the science system is divided into 22 broad fields in such a way that every journal belongs to exactly one field. It is implied that the 22 broad fields are independent. Unlike the 22 broad fields, the ISI subject categories are not disjointed or hierarchically organized, but interconnected, because more than one category is often attributed to a journal (Leydesdorff and Rafols, 2009). Furthermore, these are more specific and therefore contain more information. As we use a method that is designed especially to take similarity between categories into account, we prefer the more detailed approach with more than 170 subdivisions.

Leydesdorff and Rafols (2009) applied factor analysis based on the citation matrix of 172 Subject Categories, suggesting a 14-factor solution with a minimum loss of information. Factor analysis is used to identify clusters of inter-correlated variables (here mean Subject Categories). Factors consist of relatively homogeneous variables. That is to say factor analysis taking category similarity into account. Considering this we confirm that Leydesdorff's 14 disciplines can interpret the results of the field similarity-weighted cosine similarity.

Another definition is SCs knowledge integration intensity as the distribution of 14 disciplines. The 14 disciplines can also help us to distinguish which discipline is contributed more in the process of knowledge integration.

Coherence and similarity are the attributes of knowledge integration at different levels

Knowledge integration is not only about how diverse the knowledge is, but also about making connections between the various bodies of knowledge drawn upon.

Coherence is another attribute of interdisciplinary knowledge integration (Liu et al., 2012). It is the property describing how the elements of a system are related to each other. Rafols et al. (2012) ensured that this measure of coherence is orthogonal to diversity.

Our interest lies in using the framework proposed by Zhou et al. (2012) to track knowledge integration. In the process of knowledge integration, if two sets of articles derive the ideas from the same SCs, we can say they are close in relationship. If two sets of articles seldom share ideas from the same SCs, we can say they are relatively distant in relation. So we can use the cosine measure and

the similarity-weighted cosine measure to explore the relation (homogeneity) between different systems (sets of articles or sub-disciplines).
It is clear the coherence is critical in the relationship between elements of a system. But similarity between systems is also very important in the process of knowledge integration.

Disciplinary structure

Mapping of documents has been a discussion topic in scientometric research for a number of years (Boerner et al., 2003). In general, the procedure follows a three-step process (Sternitzke & Bergmann, 2009).

- First, bibliographic coupling, co-citation analysis, co-word analysis, co-authorship, and semantic structures of texts are the common methodologies to select the (bibliographic) data from documents (Kessler, 1963, Small, 1973, Marshakova, 1973, Rip & Courtial, 1984, Callon et al., 1991 and Tsourikov et al., 2000).
- In the second step, similarities are computed based on the above-mentioned data. Measures such as the Pearson correlation coefficient, Salton's Cosine formula, the Jaccard Index, or the Inclusion Index are possible methods of normalizing this data (Hamers et al., 1989; Peters et al., 1995; Qin, 2000; Ahlgren et al., 2003).
- Finally, in the third step, the previously computed data is visualized by means of multivariate analyses such as cluster analysis or multidimensional scaling (MDS) (see e.g. Leydesdorff, 1987) or social network analysis (see e.g. Leydesdorff and Rafols, 2009).

Bibliometric methods have been used in the study of interdisciplinarity, especially those based on the "maps of science," built upon co-word, co-authorship, or co-citation analysis, which aim to identify structural relations between various subfields and to show them in graphical representations (Tijssen, 1992; Kessler, 1963, Small, 1973, Marshakova, 1973, Rip & Courtial, 1984; Callon et al., 1991] and Tsourikov et al., 2000).
We sort out the frameworks of Zhou et al. (2012) into co-classification methods, which can also identify structural relations between various subfields.

Similarity as a measure to construct discipline structure based on knowledge integration

Two closely related individuals have a lot of genetic information in common (biology, ecological point of view). Kessler (1963) suggested the use of the references contained in papers, given that documents with the same references are regarded as very similar in nature. This approach is known as bibliographic coupling. The same for two set of articles (such as two sub-disciplines), if they have references from the same SCs, they are also considered related. This is Reference Co-Classification (RCC).

So we can say bibliographic coupling and Reference Co-Classification (RCC) are similar in approach in capturing similarity in the process of knowledge integration.
The Salton's cosine index is a commonly used similarity measure, the greater the relationship between two given sets of articles, the higher the similarity between them. Its value ranges from 0 (no relation at all) to 1 (maximum relation).
The Similarity-Weighted cosine measure, proposed by Zhou et al. (2012) considers the similarity between the SCs, which can also identify structural relations between various sub-disciplines.
In the following sections we show the results of an empirical study. To refine similarity results based on the cosine index we will use the similarity-weighted cosine index. Diversity within sub-disciplines is measured using the Rao-Stirling diversity. An investigation of the similarities between sub-disciplines leads to a general view on the homogeneity of the group of sub-disciplines under study.

The empirical study

Data

Data are extracted from the Thomson Reuters database. 7 sub-disciplines of ecology were chosen: GLOBAL CHANGE BIOLOGY, LANDSCAPE ECOLOGY, MICROBIAL ECOLOGY, WILDLIFE BIOLOGY, MOLECULAR ECOLOGY, RESTORATION ECOLOGY, and SOIL BIOLOGY. For each sub-discipline, sample bibliometric records come from the following journals: *GLOBAL CHANGE BIOLOGY, LANDSCAPE ECOLOGY, MICROBIAL ECOLOGY, WILDLIFE BIOLOGY, MOLECULAR ECOLOGY, RESTORATION ECOLOGY, and EUROPEAN JOURNAL OF SOIL BIOLOGY*.
The references of the publications of each sub-discipline were counted and grouped into 172 SCs.

Method

We use a case study of sub-disciplines in ecology to analyze interdisciplinarity as revealed through the set of references. The analysis yields quantitative measures of: (1) the level of knowledge integration in contributing sub-disciplines; (2) the strength of knowledge integration relations between these sub-disciplines. A topological structure based on disciplinary similarity of sub-disciplines is constructed.
Since categories are scientific fields we use a field-similarity weighted cosine measure. In order to obtain the Rao-Stirling diversity index, a field-similarity matrix S_{ij} in the cited dimension provided by Leydesdorff and Rafols (2009) is chosen (see Table 3). The S_{ij} describes the similarity in the citation patterns for each pair of SCs in 2006.

Table 2. Basic sample data information

No.	Short Name of System	System *Sub-discipline*	Sample *Journal*	Element *Record*	Intermediary Set *Reference*
1	GLOB	GLOBAL CHANGE BIOLOGY	*GLOBAL CHANGE BIOLOGY*	2174	70295
2	LANDE	LANDSCAPE ECOLOGY	*LANDSCAPE ECOLOGY*	1157	36821
3	MICE	MICROBIAL ECOLOGY	*MICROBIAL ECOLOGY*	2359	62268
4	WILDB	WILDLIFE BIOLOGY	*WILDLIFE BIOLOGY*	506	15188
5	MOLE	MOLECULAR ECOLOGY	*MOLECULAR ECOLOGY*	4882	122398
6	TROE	TROPICAL ECOLOGY	*JOURNAL OF TROPICAL ECOLOGY*	1416	28549
7	SOIB	SOIL BIOLOGY	*EUROPEAN JOURNAL OF SOIL BIOLOGY*	721	20443

The similarity between sub-disciplines is shown with the Reference Co-Classification (RCC) which indicates the breadth of the basis of knowledge in common between sub-disciplines.

From the ISI Web of Science we downloaded full bibliometric records for the publications. These records were processed using the bibliometric program TDA, the statistical packet SPSS (2007), the network analysis software Ucinet, and Excel.

The Ucinet was used for multidimensional scaling techniques (MDS). And the sub-disciplines were grouped according to their normalized disciplinary similarity through hierarchical clustering analysis (SPSS, Ward Method).

Results and Discussion

(1) The level of knowledge integration in contributing sub-disciplines:

- Knowledge Integration Breadth (Variety)

MOLE (Molecular Ecology), MICE (Microbial Ecology) and GLOB (Global Change Biology) have the highest Knowledge Integration Breadth; the knowledge of the three sub-disciplines comes from 128, 125, and 119 SCs.

- Knowledge Integration Intensity (Balance + Disparity) – The distribution of the 14 disciplines and 3 sup-disciplines.

The 14 disciplines can also help us to distinguish which discipline contributes more to the process of knowledge integration (see Table 5). We give some example of this index to explain the diversity measure and the relationship between sub-disciplines.

- The profile of knowledge integration--Simpson diversity and Rao-Stirling diversity.

Diversity is a combined index. It can give us the profile of the knowledge integration. From Table 6, we can see that GLOB has the highest Knowledge

Integration Intensity. It shows the highest values (0.9066 in Simpson diversity and 0.6408 in Rao-Stirling diversity, respectively).
Three sub-disciplines WILDB, TROE (Tropical Ecology) and MOLE show the lowest level of interdisciplinarity, since the disciplinary diversities within these sub-disciplines are the lowest (Table 6).

Table 3. Similarity matrix Sij of 172 SCI subject categories – partim

Number	172 SCs	1	2	3	4	5	6	…
1	Biochemistry & Molecular Biology	1.0000	0.9760	0.9489	0.0226	0.3372	0.1406	…
2	Biophysics	0.9760	1.0000	0.9041	0.0375	0.3444	0.1947	…
3	Cell Biology	0.9489	0.9041	1.0000	0.0126	0.2492	0.0928	…
4	Thermodynamics	0.0226	0.0375	0.0126	1.0000	0.1738	0.2883	…
5	Chemistry, Applied	0.3372	0.3444	0.2492	0.1738	1.0000	0.5053	…
6	Chemistry, Physical	0.1406	0.1947	0.0928	0.2883	0.5053	1.0000	…
…	……	……	……	……	……	……	……	…

Table 4. Distribution of Top SCs for 7 sub-disciplines based on the Knowledge Integration Intensity

14 disciplines	172SCs	MOLE	MICE	GLOB	SOIB	TROE	LANDE	WILDB
Ecology	Ecology	4.73%	3.67%	***17.00%***	***19.66%***	***47.34%***	***31.46%***	***19.18%***
Ecology	Forestry	0.68%	0.40%	7.90%	1.72%	6.17%	6.10%	1.65%
Ecology	Marine & Freshwater Biology	3.76%	***10.96%***	2.36%	0.96%	1.08%	1.74%	0.54%
Ecology	Oceanography	0.97%	9.78%	2.24%	0.27%	0.36%	0.57%	0.29%
Ecology	Evolutionary Biology	***34.81%***	1.54%	1.58%	1.30%	6.61%	4.79%	5.21%
Ecology	Zoology	5.62%	0.46%	1.04%	3.45%	8.15%	5.59%	***49.01%***
Ecology	Ornithology	0.88%	0.07%	0.44%	0.01%	1.44%	1.91%	6.50%
Biomedical Sciences	Multidisciplinary Sciences	9.44%	5.25%	***10.83%***	2.33%	5.06%	3.85%	2.26%
Biomedical Sciences	Biochemistry & Molecular Biology	3.35%	3.97%	0.59%	2.26%	0.23%	0.08%	0.35%
Biomedical Sciences	Biotechnology & Applied Microbiology	1.36%	***21.82%***	0.47%	5.41%	0.07%	0.02%	0.03%
Biomedical Sciences	Genetics & Heredity	***13.81%***	0.54%	0.07%	0.33%	0.20%	0.35%	0.74%
Agriculture	Plant Sciences	3.92%	3.71%	***10.77%***	4.96%	6.50%	1.93%	0.19%
Agriculture	Soil Science	0.13%	4.57%	5.99%	**25.83%**	1.31%	0.75%	0.01%
Environmental Sciences	Environmental Sciences	2.18%	1.89%	***13.49%***	6.61%	5.93%	***17.22%***	6.87%
Geosciences	Geosciences, Multidisciplinary	1.04%	1.73%	8.83%	1.20%	1.09%	***11.94%***	0.50%
Infectious Diseases	Microbiology	0.65%	***13.61%***	0.42%	4.24%	0.05%	0.01%	0.00%
…	…	…	…	…	…	…	…	…

(2) The strength of knowledge integration relations between these sub-disciplines:
A topological structure based on disciplinary similarity of sub-disciplines is constructed. It is a comparison between the results obtained with the cosine and with the field-similarity weighted cosine index (Shown in Table 1).
Figure 2 shows the final dendrograms using Ward's Method based on the cosine index and field-similarity weighted cosine index. Figure 3 shows the results of the MDS analysis. The distribution of the 14 disciplines and 3 sup-disciplines of 7 sub-disciplines are used to explain Figure 2.
Since Ward's method is a bottom-up agglomerative clustering method (each observation starts in its own cluster, and pairs of clusters are merged as one

moves up the hierarchy) we compare the results of the two methods in a "bottom-up" fashion.

Table5. The distribution of the 14 disciplines and 3 sup-disciplines of 7 sub-disciplines

		MOLE	MICE	GLOB	SOIB	TROE	LANDE	WILDB
14 disciplines	Biomedical Sciences	***31.20%***	***35.29%***	***15.81%***	***14.81%***	7.94%	7.11%	5.82%
	Clinical Sciences	0.04%	0.27%	0.02%	0.05%	0.04%	0.04%	0.11%
	Neuro-Sciences	0.22%	0.02%	0.04%	0.01%	0.07%	0.06%	0.16%
	Infectious Diseases	2.19%	***17.57%***	0.60%	6.37%	0.55%	0.12%	1.22%
	Gen. Medicine; Health	0.43%	0.46%	0.14%	0.18%	0.25%	0.53%	0.74%
	Materials Sciences	0.01%	0.18%	0.13%	0.41%	0.01%	0.04%	0.01%
	Chemistry	0.20%	1.15%	0.40%	1.21%	0.23%	0.08%	0.05%
	Computer Sciences	1.14%	0.26%	0.09%	0.18%	0.09%	0.66%	0.05%
	Engineering	0.04%	0.04%	0.05%	0.03%	0.04%	0.22%	0.08%
	Physics	0.01%	0.08%	0.07%	0.01%	0.03%	0.18%	0.02%
	Ecology	***54.74%***	***27.55%***	***33.57%***	***29.68%***	***73.58%***	***53.62%***	***82.84%***
	Environmental Sciences	2.24%	4.17%	***14.68%***	8.43%	6.36%	***18.29%***	7.10%
	Geosciences	1.38%	2.53%	***14.16%***	1.54%	1.70%	***15.64%***	0.62%
	Agriculture	6.16%	***10.40%***	***20.24%***	***37.04%***	9.11%	3.43%	1.16%
3 sup-disciplines	**Life Sciences**	***34.08%***	***53.61%***	***16.60%***	***21.42%***	8.85%	7.85%	8.05%
	Physical Sciences	1.39%	1.71%	0.74%	1.85%	0.39%	1.17%	0.21%
	Environmental Sciences	***64.52%***	***44.64%***	***82.65%***	***76.70%***	***90.75%***	***90.97%***	***91.72%***

Table 6. Simpson and Rao-Stirling diversity values for seven sub-disciplines

Sub-disciplines	Variety	Simpson diversity	Rao-Stirling diversity
GLOB	119	0.9066	0.6408
SOIB	89	0.8751	0.6284
MICE	125	0.9000	0.6218
LANDE	109	0.8434	0.5632
MOLE	128	0.8389	0.4642
TROE	86	0.7488	0.4349
WILDB	79	0.7095	0.3707

Corresponding to the divisions of the dendrogram is the cluster-enhanced MDS map below:

In the MDS analysis based on the cosine index (Figure 3), WILDB belongs to the same cluster as MOLE and MICE (Cluster 1). Taking field similarity into account, based on field-similarity weighted cosine index, WILDB joins TROE, LANDE (Landscape Ecology), and the others in Profile 2.

From Table 5, we can see WILDB has 82.84% knowledge derived from Ecology. The percentage is at the same level as TROE (73.58% knowledge derived from Ecology).

The data shows that the 54.74% and 27.55% of the knowledge of MOLE and MICE comes from Ecology, while the 31.20% and 35.29% of the knowledge of MOLE and MICE is integrated from Biomedical Sciences. WILDB has the same knowledge basis as TROE, and LANDE, not MOLE and MICE.

At the level of 3 sup-disciplines, “Life Sciences” contributes 34.08% and 53.61% knowledge to MOLE and MICE. While the three sub-disciplines (TROE, LANDE

and WILDB) inherit 90% of their knowledge from “Environmental Sciences”. It can be seen that the MDS map based on the field-similarity weighted cosine index gives us the correct description of the relationships between 7 sub-disciplines.

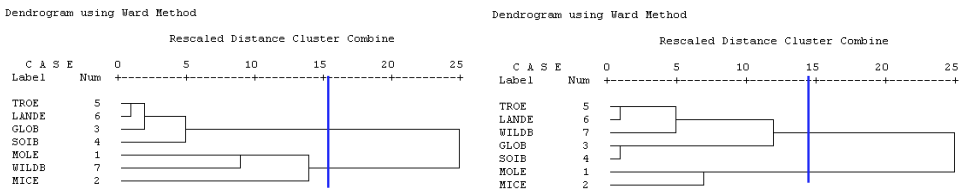


Figure 2. Dendrograms using Ward’s clustering analysis of sub-disciplinary structure. Left: based on the cosine index; Right: based on the similarity-weighted cosine index

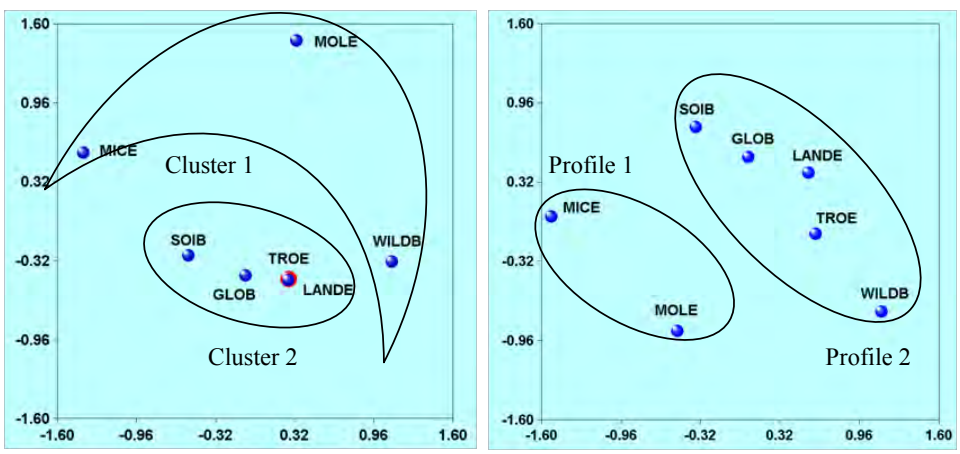


Figure 3. Maps resulting from MDS analysis of sub-disciplinary structure. Left: based on the cosine index; Right: based on the similarity-weighted cosine index

Conclusion and further research

A key point in the study is determining to what extent our indicators are based on categories’ distribution of references into measured knowledge integration. From a conceptual standpoint, we consider the indicators proposed to be valid.

As to the most relevant results of our study, we would like to stress the following:

● Disciplinary diversity at the level of knowledge integration is observed, with GLOB at the upper range of the scale. The Rao-Stirling diversity can give the whole profile of the level of the knowledge integration.

The three aspects of diversity can measure the different aspects of the level of knowledge integration in contributing sets:

Variety – the SCs into which publications can be partitioned can measure the Knowledge Integration Breadth of a set of articles.

Balance_+ Disparity – The distribution of the 14 disciplines can measure the Knowledge Integration Intensity of a set of articles.

●The similarity of sets of articles (sub-disciplines) based on the Reference Co-Classification (RCC) has proven useful in providing a deeper understanding of the relations between sub-disciplines. Since the field-similarity weighted cosine is derived from Rao-Stirling diversity, so the aspects of Balance + Disparity (the distribution of the 14 disciplines) can be used to explain to what extent those two sub-disciplines are related based on knowledge integration.

In summary, we propose that the bibliometric methodology presented here provides a compelling overview of science's structure with a special focus on the inter-relationship between sub-disciplines that makes knowledge integration possible.

However, the results should be analyzed with caution since they are highly dependent on the ISI classification scheme, which is not perfect and has a metaphorically coarse granularity. Future research will include the dynamic structure of the literature over time.

References

Ahlgren, P., Jarneving, B. & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient, *Journal of the American Society for Information Science*, 54, 550–560.

Boerner, K., Chen, C. & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37, 179–255.

Callon, M., Courtial, J. P. & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22, 155–205.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R. & Vanhoutte, A. (1989), Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. *Information Processing and Management*, 25, 315–318.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers, *American Documentation*, 14, 10–25.

Leydesdorff, L. (1987). Various methods for the mapping of science, *Scientometrics*, 11, 295–324.

Leydesdorff, L. & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, 60 (2), 348–362.

Leydesdorff , L. & Rafols, I.(2011). Local Emergence and Global Diffusion of Research Technologies: An Exploration of Patterns of Network Formation. *Journal of the American Society for Information Science and Technology,* 62(5):846–860.

Liu, Y.X., Rafols I., & Rousseau, R. (2012). A framework for knowledge integration and diffusion, *Journal of Documentation*, 68 ,(1),31 – 44.

Liu, Y.X. & Rousseau, R. (2010). Knowledge diffusion through publications and citations: a case study using ESI-fields as unit of diffusion". *Journal of the American Society for Information Science and Technology*, 61(2), 340-51.
Marshakova, I. V. (1973). System of document connections based on references. *Scientific and Technical Information Serial of VINITI*, 6, 3–8.
Morillo, F., Bordons, M. & Gomez, I. (2003). Interdisciplinarity in science: a tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology*, 54(13), 1237-1249.
Peters, H., Braam, R. & Raan, A. (1995). Cognitive resemblance and citation relations in chemical engineering publications. *Journal of the American Society for Information Science*, 46, 9–21.
Qin, J. (2000). Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. *Journal of the American Society for Information Science*, 51, 166–180.
Nei,M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583–590
Nei, M. & Li W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Science USA* 76, 10, 5269-5273
Rafols, I. & Meyer, M. (2007). How cross-disciplinary is bionanotechnology? Explorations in the specialty of molecular motors. *Scientometrics,* 70(3), 633-650.
Rafols, I. & Meyer, M. (2010). Diversity and Network Coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics* 82, (2), 263-287.
Rafols,I., Leydesdorff,L. ,Hare,A., Nightingalea, P. & Stirling, A. (2012) .How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management. *Research Policy*, 41, 1262-1282.
Rao, C.R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology,* 21(1), 24-43.
Rip, A. & Courtial, J. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6, 381–400.
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265–269.
Spasserm, A. (1997). Mapping the terrain of Pharmacy: co-classification analysis of the International Pharmaceutical Abstract Database. *Scientometrics*, 39(1), 77-97.
Sternitzke, C. & Bergmann, I. (2009) . Similarity measures for document mapping: A comparative study on the level of an individual scientist. Scientometrics, 78, (1), 113–130.
Stirling, A. (1998). On the economics and analysis of diversity. SPRU Electronic Working

Paper.http://www.sussex.ac.uk/Units/spru/publications/imprint/sewps/sewp28/sewp28.pdf Accessed 01-04-2006.
Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interfa*ce, 4 (15), 707-719.
Solow, A., Polasky, S. & Broadus, J. (1993). On the measurement of biological diversity. – *J. Environ. Econ. Manage*, 24, 60–68.
Tijssen, R.J.W. (1992). A quantitative assessment of interdisciplinary structures in science and technology: Co-classification analysis of energy research. *Research Policy*, 22, 27–44.
Tsourikov, V. M., Batchilo, L. S. & Sovpel, I. V. (2000). Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures, United States Patent No. 6167370.
Zhou,Q.J., Rousseau, R., Yang, L.Y.,Yue,T. & Yang, G.L. (2012). A general framework for describing diversity within systems and similarity between systems with applications in informetrics. *Scientometrics*, 93(3), 787-812.

CONTRIBUTION AND INFLUENCE OF PROCEEDINGS PAPERS TO CITATION IMPACT IN SEVEN CONFERENCE AND JOURNAL-DRIVEN SUB-FIELDS OF ENERGY RESEARCH 2005-11 (RIP)

Peter Ingwersen[1; 3], Birger Larsen[1], J. Carlos Garcia-Zorita[2], Antonio Eleazar Serrano-López[2] and Elias Sanz-Casado[2]

[1] *{pi; blar}@iva.dk*
Royal School of Information and Library Science, University of Copenhagen, Birketinget 6, DK 2300 Copenhagen S (Denmark)

[2] *{ czorita; aeserran; elias }@uc3m.es*
[2]University Carlos III of Madrid, Laboratory of Information Metric Studies (LEMI), Associated Unit IEDCYT-LEMI. C/ Madrid 126, Getafe 28903 Madrid (Spain)

[3] Oslo University College, St. Olavs Plass, 0130 Oslo (Norway)

Abstract

This paper analyses the following seven sub-fields of Sustainable Energy Research with respect to the influence of conference paper dominance on citation patterns across citing and cited document types, overall sub-field and document type impacts and citedness: Wind Power, Renewable Energy, Solar and Wave Energy, Geo-thermal, Bio-fuel and Bio-mass energy sub-fields. The analyses cover research and review articles as well as conference proceeding papers excluding meeting abstracts published 2005-09 and cited 2005-11 through Web of Science.

Central findings are: The *distribution* across document types and cited vs. citing documents is *highly asymmetric*. Predominantly proceeding papers cite research articles. With decreasing conference dominance the segment of proceeding papers *citing* proceeding papers decreases (from 22 % to 14 %). Simultaneously, the share of all publication types that actually are proceeding papers themselves *citing* proceeding papers decreases (from 35 % to 11 %).The proceeding paper citation impact increases in line with the probabilities that the sub-field's overall as well as proceeding paper citedness increase; and progressively more citations to proceeding papers derive from journal sources. Distribution of citations from review articles shows that *novel knowledge* predominantly derives from research articles – much less from proceeding publications.

Conference Topics

Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability; Visualisation and Science Mapping: Tools, Methods and Applications

Introduction

Commonly journal articles in the form of peer reviewed research articles and review articles are regarded the main vehicles for scientific communication in the natural science, bio-medical and some social science fields (Waltman et al., 2012). However, in several engineering fields as well as for computer science and other social science fields peer reviewed conference proceeding papers form the main scientific communication channel. With the inclusion of conference proceeding publications in the Thomson-Reuters Web of Science citation index (WoS) it is possible to observe how conference papers actually perform compared to journal articles in selected research fields in a controlled manner.

The present analysis investigates seven sub-fields of Sustainable Energy research published 2005-09 with a citation window of max. seven years (2005-11): the Wind Power and Renewable Energy subfields representing strong conference dependence (40-60 % of publications); Solar and Wave Energy subfields signifying medium conference dependence (26-39 %); and Geo-thermal, Bio-fuel and Bio-mass energy fields demonstrating low conference dependence (< 25 %). The analysis distinguishes between conference proceeding papers[42], research articles, review articles and 'other types', containing editorials, book reviews, errata, etc. as defined in WoS. As for journals WoS does not cover all conferences in the analysed energy sub-fields. Monographic materials are not included in the analyses.

Earlier studies of conference paper citation impact have demonstrated their feasibility, e.g. Butler & Visser (2006) who investigated the degree to which WoS contributes adequate data with respect to a variety of document source types, including conference proceeding and meeting publications. Martins et al. (2011) tested comprehensive conference paper indicators in the Electrical Engineering and Computer Science fields, comparing to journal-based indicators. How proceeding paper citations are distributed across a range of document types in computer science was investigated by Wainer, de Oliveira & Anido (2011). They studied the references from all (predominantly proceeding) papers published in the ACM digital library 2006. They found that around 40% of the references were to earlier conference proceedings papers, around 30% were to journal papers, and around 8% were references to books.

Based on these findings founded on a *reference analysis* one might form the hypothesis that in strong conference-dependent fields the conference papers themselves are the main contributor to the impact of the field or, at least, are the major supplier of citations to conference papers. This is measured by means of contingency tables and compared to citation impact and citedness across the three document types involved. One might also speculate that review articles in such conference-dominant areas would tend to cite conference papers rather than journal articles. However, a recent study of the conference-dominated engineering

[42] In the remaining of the paper the notion 'proceeding papers' excludes the WoS document category 'Meeting abstracts'; books are omitted and do not form part of the source set of documents to be cited..

field Wind Power research 1995-2011 (Sanz-Casado et al., 2013) demonstrates that these hypotheses and ideas might not hold true for all conference-dependent fields. Hence the motivation for the present *citation-based* analyses, which aim at observing the characteristics of citations given to defined source documents of various types. If conference papers do play a crucial role in the knowledge distribution and crediting process they ought to be taken more into account, for instance in research evaluation studies.
The paper is organized as follows. The data collection and analysis methods are described, followed by the findings of the investigation. Initially we show the distribution of document types across the seven selected Sustainable Energy research sub-fields. This is followed by findings related to the distribution of citations by document type and associated with field characteristics of conference dependency analysed across the seven sub-fields, including the distribution of citedness. A discussion section and conclusions complete the paper.

Methodology

The study made use of the already existing retrieval strategies and profiles developed and tested in the context of the SAPIENS project for the use in Web of Science (WoS). The SAPIENS Project (Scientometric Analyses of the Productivity and Impact of Eco-economy of Spain) has as main goal the analysis of scientific and technological capacities of Eco-economy in Spain 1995–2009, cited 1995–2011, seen in a global context through quantitative and qualitative R&D indicators and is reported in Sanz-Casado et al. (2013).
The seven Energy research sub-fields were extracted online in December 2012 through WoS. Elaborated search profiles were executed[43]. The following WoS citation databases were applied: SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH. For each sub-field the online set of publications 2005-09 was divided into the three document types examined in this study and analyzed by means of the WoS tools for citations published 2005-11. In addition, each set was sorted according to citation scores and the exact citedness ratio observed. Intermediate analyses and calculations were necessary for each set of a document type to 1) exclude the 2012-citations, 2) limit the citing set of publications to the required time period, and 3) define the distribution across document types of the citing set of publications. In case of sub-field sets too large for WoS to handle when generating online citation reports, i.e. sets above 10,000 items, the set was logically divided into subsets for which the analyses were aggregated later. The sub-field on Solar Energy constitutes such a large set (26,691 documents). In total the analyses deal with almost 60,000 source documents (Table 1) and more than 686,000 citations.

[43] Example of search profile for Wind Power: TS=("wind power" OR "wind turbine*" OR "wind energy*" OR "wind farm*" OR "wind generation" OR "wind systems") AND PY=(2005-2009). Refined by: Document Types=(PROCEEDINGS PAPER OR ARTICLE OR REVIEW) AND [excluding] Web of Science Categories=(ASTRONOMY ASTROPHYSICS).

Findings

Table 1 displays the distribution of document types across the seven selected sub-fields. Aside from an overlap the sub-fields in between a share of documents is indexed both as proceeding paper and article within each field. They are proceeding papers published in thematic serial issues. Hence the larger sums displayed in the 'Total' row above the Online Set figures.

Table 1. Document type distribution 2005–09 in seven sub-fields of Energy research (WoS 2012); the darker the shade the more conference-dominant the sub-field.

	Wind Power		Renewable		Wave Energy		Solar Energy		Geo-Thermal		Bio Mass		Bio Fuel		Total	
Doc. Type	Publ.	%	Publ.	%	Publ.	%	Publ.	%	Publ.	%	Publ.	%	Publ.	%	Publ.	%
Article	2754	37.1	3775	49.3	959	62.4	19794	66.5	2068	73.0	4100	77.0	7502	76.0	40952	63.6
Proc. Paper	4485	60.4	3335	43.6	554	36.0	9068	30.5	605	21.4	878	16.5	1609	16.3	20534	31.9
Review Art.	189	2.5	532	7.0	23	1.5	891	3.0	156	5.5	339	6.4	731	7.4	2861	4.4
Other	1	0.0	9	0.1	2	0.1	23	0.1	2	0.1	7	0.1	27	0.3	71	0.1
Total:	7429	100	7651	100	1538	100	29776	100	2831	100	5324	100	9869	100	64418	100
Online set:	7123		7149		1441		26691		2630		4973		9222		59229	

Conference dependency and citation distribution patterns

Tables 2 through 4 demonstrate the distribution of citations from the pool of citing publications to each of the three different types of source (cited) documents across the seven sub-fields, grouped according to conference ratios as in Table 1. "Articles" refer to research articles published in journals. The Bio Mass and Bio Fuel sub-fields are merged as one field for presentation purposes, named "Bio Energy"; the two sub-fields possess similar conference and citedness ratios, although they are dissimilar in impact[44].

Table 2. High conference-dominant Energy sub-fields. Distribution of citing publications 2005-11 to documents published 2005-09; (a): absolute numbers; (b): ratios. Analysis at document level and including overlap between types (WoS, 2012)

Table 2a.	*Wind Power Research*						*Renewable Energy Research*					
Cited:	Publ.	Citations	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*	Publ.	Citations	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*
Articles	2754	27626	8898	3952	742	13223	3775	41045	20524	4020	2585	26428
Proceed.	4485	4667	2295	1228	181	3511	3335	6323	3766	1179	541	5165
Review	189	3319	1648	425	399	2415	532	11786	7192	1077	1343	9302
Total	7428	35612	12841	5605	1322	19149	7642	59154	31482	6276	4469	40895
Online set	7123	31675				14715	7149	53311				32631

Table 2b.	*Wind Power Research (60.4 % conferences)*						*Renewable Energy Research (43.6 % conferences)*					
Cited:	*Impact*	*Citedness*	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*	*Impact*	*Citedness*	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*
Articles	10.0	0.88	0.69	0.71	0.56	0.69	10.9	0.85	0.65	0.64	0.58	0.65
Proceed.	1.0	0.16	0.18	0.22	0.14	0.18	1.9	0.21	0.12	0.19	0.12	0.13
Review	17.6	0.98	0.13	0.08	0.30	0.13	22.2	0.98	0.23	0.17	0.30	0.23
Total	4.8		1	1	1	1	7.7		1	1	1	1
Online set	4.4	0.45					7.5	0.58				

Citation impact scores and conference dependency

Diagram 1 displays the *ratio* of research article impact *versus* proceeding paper impact 2005-09, cited 2005-11 for the seven Energy research sub-fields sorted

[44] Overall citation impact for Bio Mass is 10.2 and 16.2 for Bio Fuel.

according to conference-dependency. In addition the diagram compares the overall citation impact per sub-field with the impact from proceeding papers, research articles and review articles separately, and that calculated for research and review articles combined. This "Journal impact" signifies the sub-field impact score usually applied in research assessments, which commonly also includes citations to journal and review articles *from* proceeding papers (as indexed by WoS). Further, the diagram displays the large differences between the conference impact scores for the four most conference-dependent sub-fields: Wind Power, Renewable Energy, Wave Energy and Solar Energy – and their research article impact scores: the Research Article impact is from 10:1 to 4.8:1 times that of the corresponding proceeding paper impact per sub-field – the Res.Art./Conf ratio. Diagram 2 shows the citedness across the seven sorted sub-fields.

Table 3. Medium conference-dominant Energy sub-fields. Distribution of citing publications 2005-11 to documents published 2005-09; (a): absolute numbers; (b): ratios. Analysis at document level and including overlap between types (WoS, 2012)

Table 3a.	*Wave Energy Research*						*Solar Energy Research*					
Cited:	Publ.	Citations	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*	Publ.	Citations	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*
Articles	959	6663	4357	730	253	5143	19794	324235	150670	20744	9717	173988
Proceed.	554	738	499	165	42	656	9068	31092	17637	5045	1195	21668
Review	23	252	160	34	47	233	891	48859	28583	2399	3138	33356
Total	1536	7653	5016	929	342	6032	29753	404186	196890	28188	14050	229012
Online set	1441	7071				5381	26691	375006				199071

Table 3b.	*Wave Energy Research (36.1 % conferences)*						*Solar Energy Research (30.5 % conferences)*					
Cited:	*Impact*	*Citedness*	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*	*Impact*	*Citedness*	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*
Articles	6.9	0.85	0.87	0.79	0.74	0.85	16.4	0.93	0.77	0.74	0.69	0.76
Proceed.	1.3	0.27	0.10	0.18	0.12	0.11	3.4	0.42	0.09	0.18	0.09	0.09
Review	11.0	1.0	0.03	0.04	0.14	0.04	54.8	0.98	0.15	0.09	0.22	0.15
Total	5.0		1	1	1	1	13.6		1	1	1	1
Online set	4.9	0.64					14.0	0.77				

Table 4. Low conference-dominant Energy sub-fields. Distribution of citing publications 2005-11 to documents published 2005-09; (a): absolute numbers; (b): ratios. Analysis at document level and including overlap between types (WoS, 2012)

Table 4a.	*Geo-Thermal Energy*						*Bio Energy*					
Cited:	Publ.	Citations	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*	Publ.	Citations	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*
Articles	2068	15638	9241	1010	1051	10952	11602	157618	61300	5243	6836	71742
Proceed.	605	2068	1439	282	180	1762	2487	12969	8623	1166	1308	10514
Review	156	2651	1757	236	401	2324	1070	42962	22465	1805	3396	27720
Total	2829	20357	12437	1528	1632	15038	15159	213549	92388	8214	12140	109976
Online set	2630	18511				12767	14195	200973				

Table 4b.	*Geo-Thermal Energy (21.4 % conferences)*						*Bio Energy Research (16.4 % conferences)*					
Cited:	*Impact*	*Citedness*	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*	*Impact*	*Citedness*	*Citing Art*	*Citing Proc*	*Citing Rev*	*Citing Docs.*
Articles	7.6	0.89	0.74	0.66	0.64	0.73	13.6	0.92	0.66	0.64	0.56	0.65
Proceed.	3.4	0.42	0.12	0.18	0.11	0.12	5.2	0.44	0.09	0.14	0.11	0.10
Review	17.0	0.96	0.14	0.15	0.25	0.15	40.2	0.97	0.24	0.22	0.28	0.25
Total	7.2		1	1	1	1	14.1		1	1	1	1
Online set	7.0	0.79					14.2	0.84				

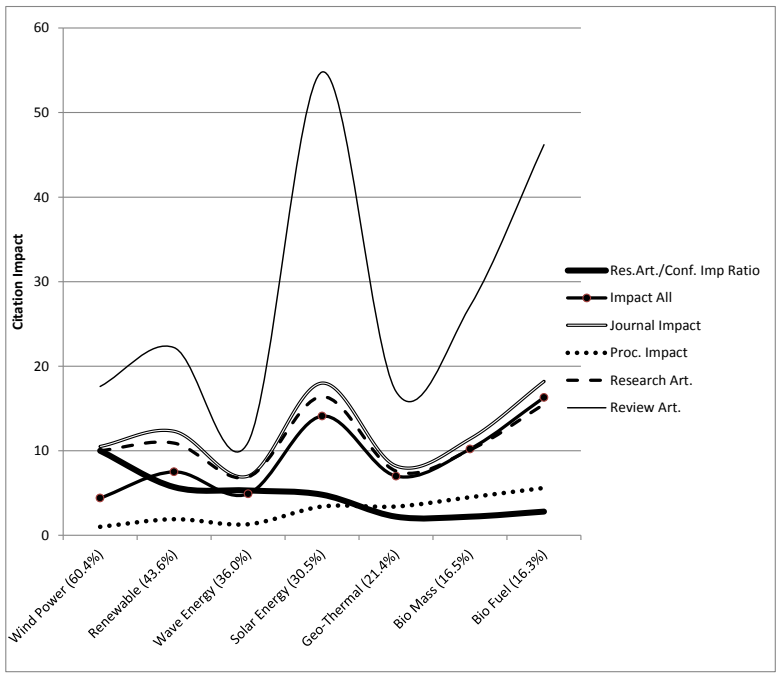


Diagram 1. Document type impact scores per Energy sub-field 2005-09(11) (WoS, 2012)

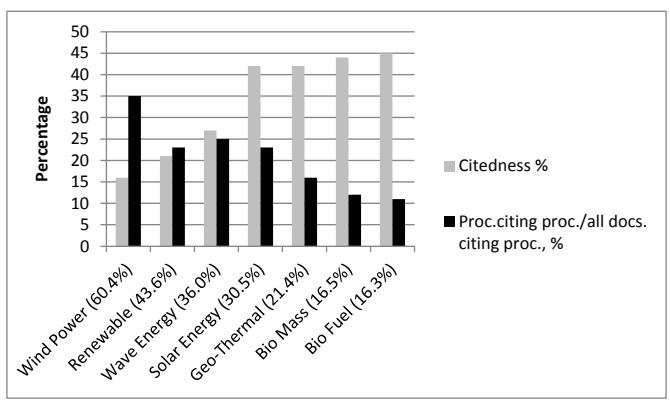


Diagram 2. Proceeding paper citedness (%) and percentage of all documents citing proceeding papers that are also proceeding papers (WoS, 2012).

Discussion

The presented findings concern the Web of Science citation index[45]. In other citation index configurations results might thus differ slightly. The initial hypothesis that in strong conference-dependent fields the proceeding papers themselves are the main contributor to the impact of the field or, at least, are the

[45] Excluding the recent addition of book citation indexing.

major supplier of citations to conference papers, does not hold. The distribution is *highly asymmetric*: All the document types investigated, including the proceeding papers, predominantly provide citations to *research articles* – far less to proceeding papers – even from proceeding papers themselves. Proceeding papers may consequently be regarded a significant (negative) player in the scientific communication process and a crucial factor in research evaluation.
Some distinct trends are observed with *decreasing conference dominance* in the Energy sub-fields:

a) The segment of proceeding papers *citing* proceeding papers decreases (from 22 % to 14 %, Tables 2-4);
b) The share of all publication types that actually are proceeding papers themselves, *citing* proceeding papers decreases (from 35 % in Wind Power to 11 %, Diagram 2). This maximum share is close to the 40 % found by Wainer, de Oliveira & Anido (2011) in their reference analysis on the ACM Computer Science digital library.
c) The ratio drops between research article and proceeding paper impacts, Diagram 1;
d) The conference citation impact increases, Diagram 1; and
e) The gap diminishes between the overall sub-field impact and the isolated impact of research articles as well as the 'journal' impact, Diagram 1.
f) The probability increases that also the sub-field's overall citedness increases (and thus its overall citation impact), Tables 2-4;
g) The probability increases that the sub-field's proceeding paper citedness increases (and thus its proceeding paper citation impact), Diagrams 1-2; and
h) Increasingly citations to proceeding papers derive from journal sources; Diagram 2.

This latter trend is not a paradox but nevertheless an interesting observation and contrasts heavily the initial hypotheses and speculations. It is noticeable that in this citedness game the country profiles may be influential. For instance, the Chinese focus on scarcely cited proceeding papers in Wind Power (Casado et al., 2013) may indeed influence the overall impact of that field – a similar case is observed by He & Guan (2008) for proceeding papers in Chinese Computer Science.

Conclusions

Based on the findings it is recommendable not simply to rely on journal article analyses, but to take all the research and innovation-producing types of documents into account in research evaluation studies – including proceeding papers – because this document type does have significant (negative) influence on the overall citation impact of an Energy research field, in particular in proceeding-dominant fields. This recommendation may probably extend even to all

engineering-like fields, but should be further investigated. However, proceeding papers and their impact pattern alone is *not* a good predictor of a conference-dependent field's overall impact.

For the Energy research fields, which encompass scientific as well as technological and innovative engineering subject areas, the findings demonstrate that with decreasing conference dominance a sub-field's proceeding paper citedness and citation impact increase and increasingly citations to proceeding papers derive from journal sources.

Acknowledgments

This research was funded by the Spanish Ministry of Economy and Competitiveness under the project CSO2010-21759-C02-01 titled ''Análisis de las capacidades científicas y tecnológicas de la eco-economìa en España a partir de indicadores cuantitativos y cualitativos de I+D+i'' (Analysis of scientific and technological capacities of Eco-economy in Spain, throughout I+D+i quantitative and qualitative indicators), and by Carlos III University of Madrid-Banco de Santander Chairs of Excellence Program for 2011/2012 academic year.

References

Butler, L. & Visser, M.S. (2006). Extending citation analysis to non-source items. *Scientometrics*, 66(10), 327-343.

He, Y. & Guan, J.C. (2008). Contribution of Chinese publications in computer science: A case study on LNCS. *Scientometrics*, 75(28), 519-534.

Martins, W.S., Goncalves, M.A., Laender, A.H.F. & Ziviani, N. (2010). Assessing the quality of scientific conferences based on bibliographic citations. *Scientometrics*, 83(19), 133-155.

Sanz-Casado, E., Garcia-Zorita, J.C., Serrano-Lopez, A.E., Larsen, B. & Ingwersen, P. (2012). Renewable energy research 1995–2009: a case study of wind power research in EU, Spain, Germany and Denmark. *Scientometrics*, 95(1), 197-224.

Wainer, J., de Oliveira, H.P. & Anido, R. (2011). Patterns of bibliographic references in the ACM published papers. *Information Processing & Management*, 47(13), 135-142.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., van Eck, N.J., van Leeuwen, T.N., van Raan, A.F.J., Visser, M.S., Wouters, P. (2012). The Leiden ranking 011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419-2432. DOI: 10.1002/asi.22708.

CORE-PERIPHERY STRUCTURES IN NATIONAL HIGHER EDUCATION SYSTEMS. A CROSS-COUNTRY ANALYSIS USING INTERLINKING DATA

Benedetto Lepori[1], Vitaliano Barberio[2],Marco Seeber[3]

[1]*benedetto.lepori@usi.ch*
Centre for Organizational Research, University of Lugano, Via Lambertenghi, 6900, Lugano (Switzerland)

[2]*vitaliano.barberio@gmail.com*
Department of Public Management, WU University, Vienna (Austria)

[3]*marco.seeber@usi.ch*
Centre for Organizational Research, University of Lugano (Switzerland)

Abstract
This paper presents a comparative analysis of the structure of higher education national networks in six European countries using interlinking data. We show that national HE systems display a common core-periphery structure, which we explain by the lasting reputational differences in sciences, as well as by the process of expansion and integration of HE systems. Further, we demonstrate that centrality in national networks (*coreness*) is associated with organizational characteristics, reflecting that interlinking is motivated by access to resources and status of the concerned organizations; national policies impact on network structure by influencing the level of inequality in the distribution of resources and status. Finally, we show that, as an outcome of the core-periphery structure, the strength of ties between two HEIs is largely determined by their individual coreness, while the impact of distance is too small-scale to alter the network structure generated by organizational attributes.

Conference Topic
Webometrics (Topic 7).

Introduction
In recent years, a growing body of literature investigated relational patterns between higher education institutions (HEI), using data from co-publications (Glänzel and Schubert 2005, Jones, Wuchty and Uzzi 2008), collaborations in European projects (Heller-Schuh, Barber, Henriques, et al 2011) and weblinks (Bar-Ilan 2009; Thelwall and Zuccala 2008). Following social network theory, we contrast two types of studies: those focusing on *connectedness* (Laumann, Galaskiewicz and Marsden 1978), i.e. understanding determinants of the relationship between two units (Rivera, Soderstrom and Uzzi 2010), including

geographical distance (Hoekman, Frenken and Tijssen 2010; Thelwall 2002b), size (Thelwall 2002a) and international reputation (Seeber, Lepori, Lomi, Aguillo and Barberio 2012) on the one side; studies focusing on the network structure, dealing with concepts like structural equivalence (White, Boorman and Breiger 1976), network centrality (Freeman 1978/79.; Abbasi, Hossain and Leydesdorff 2012) and core-periphery structures (Borgatti and Everett 1999) on the other side (see for example Thelwall, Tang and Price 2003, Ortega, Aguillo, Cothey and Scharnhorst 2008, Thelwall and Zuccala 2008).

In this context, this paper aims at providing empirical evidence on the structure of higher education (HE) *national* networks, highlighting cross-country patterns, as well as differences related to national policies. More specifically, we focus on three main questions: first, we show that national HE systems display a common core-periphery structure (Borgatti and Everett 1999), which we explain by the lasting reputational differences in sciences, as well as by the process of expansion and integration of HE systems (Kyvik 2004). Second, we demonstrate that centrality in national networks (*coreness*) is associated with organizational characteristics, reflecting that interlinking is motivated by access to resources and status of the concerned organizations (Gonzalez-Bailon 2009); national policies impact on network structure by influencing the level of inequality in the distribution of resources and status. Third, we show that, as an outcome of the core-periphery structure, the strength of ties between two HEIs is largely determined by their individual coreness, while the impact of distance is too small-scale to alter the network structure generated by organizational attributes (Frenken, Hardeman and Hoekman 2009; Holmberg and Thelwall 2009, Hoekman, Frenken and Tijssen 2010).

We explore our propositions through a cross-comparative analysis of six European national systems (Germany, Italy, Netherlands, Norway, Switzerland and UK). We measure relationships between HEIs through counts of web-links among their websites (Bar-Ilan 2009, Thelwall and Sud 2011), while organizational data are extracted from the EUMIDA census of European HEIs (Lepori and Bonaccorsi 2013). Finally, we draw on the literature on higher education policies and funding systems to characterize national systems in terms of competition for funding (Nieminen and Auranen 2010) and functional differentiation between HEI types (Lepori and Kyvik 2010).

Background and theoretical framework

Core periphery models

In a broader meaning, core/periphery models designate a relational pattern in which a group of central organizations can be identified, strongly interacting among themselves, as well as a group of peripheral organizations interacting mainly with the core and to a minor extent among themselves (Borgatti and Everett 1999). This notion comes up with an understanding that the network is

organized around a single center and that the strength of the relationships between two nodes is determined by their closeness to the center.

Core-periphery structures have been important since early social network studies (Snyder and Kick 1979), while more recent empirical tests range from organization theory (Cattani, Ferriani, Negro and Perretti 2008) to physics (Holme 2005). They tend to underline the assumption that a status hierarchy is in place between the two roles, with the core clustering actors with higher status (Owen-Smith and Powell 2008). In science studies, even if formal models of core/periphery have been rarely investigated (see however Kronegger, Ferligoj and Doreian 2011, Chinchilla-Rodriguéz, Ferligoi, Miguel, Kronegger and de Moya-Anegón 2012), there is an understanding that scientific collaboration networks display such a structure as an outcome of reputational effects (Wagner and Leydesdorff 2005; Burris 2004).

The hypothesis that *national* higher education (HE) fields display a *common* core/periphery structure is justified by the importance of reputational differences among HEIs on the one side, by the process of integration and structuring that these fields underwent in the last decades, under the pressure of increasing the demand for tertiary education on the other side (Schofer, E., Meyer, J. 2005). While stratified HE systems, like US and the UK after the 1992 reform, are historically characterized by a well-defined status hierarchy, other European HE system were based on functional differentiation of types of educational organizations, which constituted largely distinct social spaces (Kyvik 2004). In the last decades these systems moved towards a stronger integration in a single HE field characterized by common regulations (albeit with distinction by types), implying a clearer and more formalized hierarchy of status ordering (Bleiklie 2003). An important driver of this process was the introduction of quasi-market governance arrangements (Ferlie, Musselin and Andresani 2008), leading to increased freedom for customers to choose the HE provider and thus requiring clearer signals in terms of quality of offerings, as expressed by various types of rankings (Buela-Casal, Gutierrez-Martinez O., Bermudez-Sanchez M. P. and Vadillo-Munoz O. 2007). Accordingly, it is expected that a core/periphery structure has become a general feature of European HE system independently of their governance arrangements.

We expect national variations in this structure to be related to differences in the regulatory arrangements through which integration was managed (Paradeise, Reale, Bleiklie and Ferlie 2009). The adoption of a binary policy is expected to sharpen the distinction between core and periphery as the national system includes two types of HEIs with different mission and legal status (Kyvik 2004). The introduction of quasi-market arrangements and competition for funding (Nieminen and Auranen 2010) should increase the level of contrast between core and periphery, as competitive logics will tend to reinforce status hierarchies via a selective distribution of resources (Lepori 2011). In non-competitive systems, boundaries between core and periphery are expected to be blurred and the core to include a greater share of HEIs.

Coreness and organizational characteristics

Core/periphery models allow computing a continuous *coreness* measure, which can be broadly interpreted as a measure of the closeness of the HEI to the network center and is expected to be correlated to indegree and outdegree centrality measures, as core/periphery is associated with loglinear independence (Borgatti and Everett 1999).

Social network studies display that the formation of ties between organization can be explained by a number of mechanisms, including identity (belonging to the same social space; Rivera, Soderstrom and Uzzi 2010), legitimacy seeking (linking preferentially to high-status organizations (Cattani, Ferriani, Negro and Perretti 2008) and resource mobilization (linking to organizations controlling a large share of resources). Previous studies support the assumption that interlinking patterns on the web are motivated by strategic behavior of organizations reflecting unequal distribution of resources (as measured by organizational size) and status in the real world (Seeber, Lepori, Lomi, Aguillo and Barberio 2012). Accordingly, interlinking networks tends to display a less skewed distribution than in reputation-based networks, like citation counts of scientific publications (Gonzalez-Bailon 2009).

Since in core/periphery networks coreness is closely related to degree, we expect coreness of an *individual* HEI to be associated to a similar set of attributes that determines the likelihood of one HEIs linking another one. More specifically, we test associations between coreness and following characteristics:

- Size, as larger organizations have a larger volume of activities and hence of relationships, but at the same time control a larger share of resources (and are more desirable partners for establishing ties).
- Status, as organizations will link preferentially to high-status organizations both because of their higher value and of legitimacy-seeking behavior.
- Age, since older organizations are likely to be more established and to attract a larger number of ties.
- Research intensity, as research represents in the academic world the most valuable activity and thus research-oriented HEIs are expected to be more attractive partners.
- Disciplinary specialization, as generalist HEIs are expected to be more central and to develop a higher number of ties because of their broader coverage of scientific domains.

Connectivity and geography

A core/periphery model implies that the number of links between two nodes is associated to the proximity of individual organizations to the network center (as measured by coreness; Borgatti and Everett 1999) and, thus, depends only on their individual attributes (rather than their relative position, like similarity and

proximity). In other words, a direct relationship between *structural position* and *connectedness* is expected.
Yet, micro-level studies of connectivity demonstrate that the spatial distance influences the likelihood of linking and the number of ties (Rivera, Soderstrom and Uzzi 2010), as confirmed by empirical studies on scientific collaborations (Hoekman, Frenken and Tijssen 2010) and on interlinking between HEI websites (Holmberg and Thelwall 2009, Seeber, Lepori, Lomi, Aguillo and Barberio 2012).
The interaction between network structural characteristics (as modeled by the core/periphery model) on the one side, spatial distance on the other side, represents a central issue in social network studies (Adams, Faust and Lovasi 2012); specific questions concern the independence of structural and geographical effects (Daraganova, Pattison, Koskinen, et al 2012) on the one side, the extent to which heterogeneity in the distribution of human activities might generate specific network structures, like the emergence of spatially bounded clusters, on the other side (Butts, Acton, Hipp and Nagle 2012).
To this aim, we model the number of ties between two HEIs as a function of the organizational characteristics explaining network centrality and of the distance between nodes; this allows investigating the relative contribution of distance and position in the core/periphery structure on interlinking patterns, identifying the geographical scale where distance is more important and explaining why in the geography of the considered countries has a limited influence on network structure. Further, this allows speculating under which conditions geography might have a significant impact on network structure (and not only on connectivity).

Methods

Sources and data

HEI sample. Organizational data have been derived from the EUMIDA (European Micro Data) dataset, which includes information on HEIs in 28 European HE systems (Seeber, Lepori, Lomi, Aguillo and Barberio 2012). We consider only 'research active' institutions for reasons of more complete data availability and as these constitute the largest part of the system; these comprise almost all doctorate-awarding institutions, as well as most non-university type institutions in binary countries. A few HEIs have been excluded because of lack of data or because they are focused on research and graduate education, with very few students at the undergraduate level. Our sample is composed by 643 HEIs comprising 96,4% of the students in the full HEI perimeter in the six countries considered. The data mostly refer to the year 2008.
A key advantage of these data for the purpose of studying national HE networks is that they allow extending the analysis of relational structure well beyond the core of internationally-reputed HEIs.

Relational data. To characterize the relational structure of HEI systems, we make use of interlinking data between the webdomains of HEIs. The data were obtained from the Cybermetrics lab (Ortega et al., 2008) by using commercial public search engines following the methodology described in Aguillo, Granadino, Ortega, and Prieto (Aguillo, Granadino, Ortega and Prieto 2006). Two mirrors of the "Yahoo Search!" database were used, the Spanish and the British ones, to avoid collection problems derived from restrictions in the limited bandwidth available or from errors in the automatic scripts used for extracting the data. If the results for the same request were not identical, then the maximum value of the two was used. The collection took place in January 2011. From the original dataset, national matrixes were created considering interlinks between HEIs in the same country.

An extensive literature in webometrics shows that weblinks between HEIs are related to all kinds of academic activities (research, education, institutional cooperation; Bar-Ilan 2004; Wilkinson, Harries, Thelwall and Price 2003), while aggregated numbers display statistical regularities – like depending on distance, reputation, country -, supporting their interpretation as indicators of relationships between HEIs (Seeber, Lepori, Lomi, Aguillo and Barberio 2012). As a matter of fact, it can be argued that weblinks are a better measure of aggregate social relationships than indicators referring only to research collaboration (like co-authorships), and thus are better suited for the purposes of our analysis.

Organizational characteristics. A set of variables are introduced to explore the extent to which they are associated with the position of an HEI in the network. These are: (a) the *type* of organization, as a dummy variable taking the value 1 for universities and 0 for non-universities; (b) the *research intensity,* which estimates the orientation of HEIs towards research, as the ratio between the number of PhD students and the number of undergraduate students; (c) the *organizational size,* measured as the number of academic staff; (d) the *discipline concentration* calculated as the Herfindahl index of concentration of the undergraduate students across the nine fields comprised in the EUMIDA of educational statistics, ranging from 1 (all students in one field) to 1/9 (students evenly distributed across the nine fields), (e) *age* is a dummy variable set to 1 for organizations founded after the year 1970.

We introduce two measures of geography: first, we measure *urban centrality* of individual nodes using Globalization and World Cities Network (GARC) classification of cities 2010 (Taylor 2004; http://www.lboro.ac.uk/gawc/world2010.html); the index takes the value 1 for London, 0.33 for Frankfurt, Madrid and Milan and then decreases towards 0. Second, we measure *geographical distance* in kilometres between two HEIs. Each web domain corresponds to an IP, which has been related to the latitude and longitude coordinates used to compute the distances. Manual data cleaning identified the cases when IP did not correctly locate the HEI.

A measure of *international reputation* is constructed as the product between normalized impact factor and total number of publications of the concerned HEIs

("brute force" indicator; van Raan 2007), normalized by the number of academic staff; this indicator builds on the insight that the international visibility of a HEI is related both to the quality and the volume of the output. Data are derived from the SCIMAGO institutions rankings for the year 2011 (http://www.scimagoir.com/). We hold data for 240 HEIs in our sample – the other HEIs had less than 100 publications in Scopus in the reference year 2009. Despite normalization by size, this index remains correlated with output (as a result of scaling properties of research output; van Raan 2007); accordingly, when the level of output approaches the threshold, the index approaches 0 as well. For the remaining HEIs, we compute expected values of output based on the correlation between output and academic staff (0.866**) with a threshold of 100 publications; we then calculate reputation by setting the impact to the world average. As an outcome, 262 HEIs with less than 200 academic staff receive an international reputation of 0, while 141 HEIs receive a low reputation score below the HEIs included in the ranking.

Characterization of national systems. National HE systems are distinguished between unitary and binary. In unitary systems, all HEIs have the same legal status and are entitled to perform research and award PhD degrees; in binary systems, there is a legal distinction between two institutional types, with non-university HEIs being oriented towards professional education and in most cases not having the right to award the PhD. In strong binary systems, the distinction is clear-cut, while in soft binary non university HEIs can get the right of awarding the PhD and a university status through an accreditation procedure (Kyvik and Lepori 2010). Finally, we characterize the level of competition in HE funding through i) the level of output vs. input orientation in institutional funding and ii) the share of third party funding (Nieminen and Auranen 2010).

Analysis

Core-periphery structure. We test the fit of web-links data to a core/periphery model using two models as specified in Borgatti and Everett (1999). The first model entails a clustering of nodes into two discrete classes (the core and the periphery), while the second a ranking of nodes according to their continuously distributed property of being core (coreness). The procedure takes as an input the observed web-links as an asymmetric weighted matrix and fits both a continuous and a discrete core/periphery model to it. We applied a logarithmic transformation to the weblinks ($y=\log(x+1)$), as it can be assumed that the strength of a relationships is better measured by proportions, rather than by their absolute number, while the transformation strongly reduces skewedness of data.

To find the partitions that maximize the correlation between observed and ideal structures UCINET uses a combinatorial optimization technique – genetic algorithm – then the result will be statistically significant by design (Borgatti, Everett and Freeman 2002). The continuous model differs from the discrete one to the extent that a measure of "coreness" is assigned to the nodes. As a measure of

the sharpness of the core-periphery model, we computed the Gini coefficient which measures the inequality in the distribution of coreness scores.

Determinants of coreness. We compute descriptive statistics for HE organizations' *coreness* and organizational characteristics. To test associations with organizational variables, we run a linear regression by using as dependent variable national coreness normalized on a scale from 0 to 100 (to take into account differences between national models). We apply a square root transformation which strongly reduces skewedness (from .791 to 0.091), while slightly increasing kurtosis (from -657 to -1.162) of the dependent variable (Kolmogorov-Smirnov statistics for normality decreasing from 3.818 to 1.705). The regression allows computing the best combination of organizational variables explaining the observed network coreness, which we call *relational mass*.

Connectedness. Finally after providing some link-level descriptive statistics on distribution of links by distance of the nodes, we perform regressions between links count (dependent variable) on the one side, relational mass and distance on the other side. Since we deal with count data, we use a *negative binomial regression* which includes a parameter to model overdispersion. Further, since the number of null dyads (dyads with no links) is rather high (74% of the sample), we use a *hurdle* negative binomial, which specifies a separate model for predicting zeros – the underlying assumptions being that factors explaining zeros might be different from those explaining counts (Mullahy 1986). This type of models is robust against non-normality of distributions and the presence of outliers (Seeber, Lepori, Lomi, Aguillo and Barberio 2012).

The interpretation of the regression results differs from ordinary regressions, as the model provides a probability distribution for different values of counts and, especially when there is overdispersion, the distribution of probabilities is not normal around the expected mean; accordingly, the expected count value is not necessarily a good predictor of observed counts, but has to be complemented with measures based on probabilities (for example the probability that a dyad has no links). Further, binomial regression coefficients are exponential and multiplicative - changes in different antecedents have a multiplicative impact on expected number of weblinks.

Results

Descriptive statistics

Descriptive statistics on the six HE education systems considered displays a number of relevant differences (Table 1).

UK and Italy are unitary systems, where all HEIs are granted the same status, while the other countries are binary. The Norwegian system can be characterized as a soft binary, as UAS can be accredited to award PhD degrees, while colleges can request accreditation to become universities – as a matter of fact three colleges became universities in 2005 and 2007.

Table 1. Descriptive statistics on national HE systems. Source: EUMIDA. Reference year 2008.

		CH	DE	IT	NL	NO	UK
HEI characteristics	HEI total	35	291	75	55	42	143
	N. universities	12	117	75	15	7	143
	Size (average)	857	518	1298	721	410	939
	Reputation (average)	1.84	1.21	3.32	3.25	.67	3.64
	Research intensity (average)	.07	.03	.02	.01	.02	.05
	Disc-conc (average)	.64	.51	.40	.66	.42	.34
Policy	Level of competition	medium	medium	weak	medium	Medium-strong	strong
	Functional differentiation	Strong b.	Strong b.	Unitary	Strong b.	Weak b.	Unitary
Weblinks statistics	Mean count of links	100	13	36	21	59	18
	% of dyads with 0 links	55%	79%	32%	75%	35%	45%
	Maximum	37'700	39'100	27'100	8'080	11400	16'600
	Average distance of dyads (km)	132	388	456	152	656	301
	Average distance of links (km)	59	326	357	110	526	259
	Average distance of active dyads	125	351	450	123	635	286

Concerning the level of competition in resources allocation, UK represents the extreme case of high competition (output-oriented, high share of external funding), while Italy represents the extreme case of low competition (input-oriented, small share of external funding). All the other countries lay in intermediate positions, with Norway being probably more competitive than the other countries (Lepori, Benninghoff, Jongbloed, Salerno and Slipersaeter 2007, Nieminen and Auranen 2010).

Further, in the whole sample, size and reputation are strongly correlated (.631**) despite the fact that the latter has been normalized by size; both display also moderate correlations with research intensity (.437** and .462** respectively). In binary systems, organizational characteristics of the two types of HEIs are quite different: non-university HEIs are more numerous, but smaller and their research intensity and reputation is low, consistently with the fact that they don't have the right to award the doctorate and have a mission oriented towards education and transfer.

Statistics on weblinks display the well-known skewed distribution, with a large number of non-active dyads, as a well as a few very high counts (Seeber, Lepori, Lomi, Aguillo and Barberio 2012); expectedly, the average distance of links is smaller than the one of dyads, but the difference is not very large showing that strong connections are not short-range (in Switzerland the highest count is between two HEIs in the same city, namely EPFL and UNIL). The same applies for the average distance of active connections, showing that connectivity is by large national.

Table 2. Test of the core-periphery hypothesis and descriptive statistics of organizational variables per country. Test of differences of Medians, Mann-Whitney, two-tailed; *<0.001, **<0.01, *<0.05.**

		CH	DE	IT	NL	NO	UK
Test of core-periphery	Correlation	0.845	0.864	0.859	0.873	0.798	0.852
	Gini Coeff. coreness	0.48	0.59	0.32	0.59	0.25	0.39
Dimension of the core	% of HEIs	37%	23%	55%	23%	16%	42%
	% of academic staff	77%	76%	83%	56%	61%	71%
	% of undg. students	69%	64%	83%	36%	46%	58%
	% of phd students	99%	95%	87%	100%	85%	84%
Connectivity (% of ties active)	Core-core	.99	1.00	.99	1.00	1.00	.99
	Periphery -core	.54	.35	.76	.35	.97	.59
	Core - periphery	.45	.36	.57	.38	.93	.60
	Periphery - periphery	.22	.06	.27	.11	.48	.23
Medians of attributes within the core	N Org.	13	66	41	13	7	61
	N. Universities	11	65	41	13	4	61
	research intensity	.17***	.06***	.02*	.03***	.09	.07***
	size	1675***	1553***	1654***	1694***	1135***	1157***
	disc_conc	.33***	.24***	.24***	.45	.24	.22***
	Reputation	5.36***	4.97***	4.32***	11.17***	4.19***	7.09***
Medians of attributes within the periphery	N Org	22	225	34	42	35	82
	N. Universities	1	52	34	2	3	82
	research intensity	.00***	.00***	.02*	.00***	.00	.01***
	size	84***	111***	436***	186***	149***	439***
	disc_conc	.00***	.49***	.42***	.90	.28	.27***
	Reputation	.00***	.00***	.19***	.02***	.00***	.23***

Testing the core-periphery structure

As shown in table 2, the fit between the core/periphery model and our data, expressed as the correlation between ideal models and observed data, is very high for all countries reaching the maximum level for Netherland (.873) and the minimum level for Norway (.798). The Gini coefficients show that countries adopting a binary policy display a sharper core-periphery hierarchy with the exception of Norway. Measures of connectivity display expected differences: in all countries, relationships within the core being active (at least one weblink), whereas most relationships in the periphery are not.

As foreseen in a core/periphery structure, coreness is closely associated to the total number of links sent and received, the correlation coefficient between coreness normalized on the one size, indegree and outdegree on the matrix of loglinks being 0.910** and 0.918** respectively.

In all countries, there is a clear distinction between core and periphery and characteristics of organizations in the two groups are significantly different. Core organizations are larger, have higher research intensity and reputation, and cover a wider spectrum of disciplines; differences are statistically non-significant only in Netherlands (disciplinary concentration), Norway (research intensity and disciplinary concentration) and Italy (research intensity). These associations are

confirmed by the fact that a binomial regression with reputation as independent variable correctly classifies 91.8 % of the cases between core and periphery (size provides a slightly less fit). In Germany and Norway, reputation allows to classify HEIs much better than the binary type: in Germany, it distinguishes correctly between core and periphery universities (106 out of 117 are classified correctly), whereas in Norway it discriminates between core and periphery universities, but does not identify the three colleges belonging to the core.

There is some evidence of impact of national policies on the core/periphery structure: in binary systems the core includes a lower share of HEIs and the distinction is clearer (with the exception of Norway). Even if comprising less than half of the institutions, the core includes most of the resources (as measured by academic staff) and research activities (as measured by PhD students); concentration is lower for undergraduate students, with the exception of Italy. The inequality of the distribution of coreness is larger in the binary countries (CH, DE, NL) than in UK, IT and in Norway.

In Switzerland and Netherlands, there is a close correspondence between the core/periphery distinction and HEI types. In Germany, the core is composed by universities, but a significant number of university-type HEIs is in periphery. These comprise specialized universities (for example medical universities), very small and specialized universities, as well as teacher training and theological HEIs (which have a university status). In Norway, the three colleges accredited to universities after 2000 (Agder, life sciences and Stavanger) belong to the periphery as well. Thus, if the university type is extended beyond research universities, it does not imply that those HEIs display a high level of network centrality.

Despite its current unitary system, UK display the traces of the integration process: the core is composed by the oldest (pre-1992) universities, as well as by some of the former Polytechnics which were integrated in the university system in 1992. The periphery is composed by the remaining post-1992 universities, as well as by a number of specialized HEI, arts and educational colleges. Competitive allocation of resources largely maintained the pre-existing hierarchy (Stiles 2000), which was however softened by some mobility of the former Polytechnics.

Italy displays a large core, comprising more than half of the HEIs, as well as 83% of staff; the distinction between core and periphery matches almost exactly the one of students – setting a threshold of 15’000 students would correctly classify 69 over 77 HEIs. This can be interpreted as an outcome of the lack of distinction between types of HEIs, as well as of a system of allocation of resources to a large extent related, directly or indirectly, to the number of students, lacking the concentration effect associated to research activities. The massification of higher education was tackled through the foundation of new universities, especially in the south of Italy; once these reached the students’ threshold, they moved into the core which expanded to comprise most of the HE system (with a few exceptions due to geographic position).

In Norway, the four historical universities with high international reputation are the most central and display a very high level of coreness. Large colleges located in the largest cities (Oslo and Bergen) develop strong relationships with universities and move towards the network center, reaching a level of coreness larger than the three "new" universities (colleges upgraded to universities in 2005). We consider that the flat distribution of coreness and the less good fit to a core/periphery model is explained by three factors: (1) the blurring of the distinction between universities and colleges, (2) the specific geographical structure of the country, where most HEIs are clustered in the large cities which are very far apart and (3) the very small number of historical universities with high international reputation.

Interestingly, Norway is the only country where even the smallest and least reputed HEIs are connected to the core of the system – the minimum of coreness is 0 in all countries except Norway (49), while the minimum total degree (sum of inlinks and outlinks) is below 20 in all countries except in Norway where it is 120.

Table 3. Determinants of coreness. Ordinary Least square model. Dependent variable: square root of coreness normalized. N=643

	Staff only		Staff only		Staff and reputation		All variables			
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Beta	VIF
Intercept	1.241	.100***	-.246	.138'	.192	.141***	.780	.255**		
SQRT academic staff	.156	.004***	.306	.011**	.248	.013***	.230	.015***	1.264	25.355
Academic staff			-.003	.000***	-.002	.000***	-.002	.000***	-.698	16.749
SQRT Int. reputation					.609	.073***	.531	.075***	.214	3.509
Res Intensity							5.758	1.031***	.108	1.430
Disc. Concentration							-.834	.203***	-.090	1.845
Found. year (dummy)							.024	.102	.004	1.232
		Df		Df		Df		Df		
Adjusted Rsquare	..735	643	.796	643	.816	643	.833	641		
Residual mean sum of squares	2.081	641	1.598	640	1.444	639	1.312	634		
F statistics	1781.168	1	1257.011	2	950.376	3	533	6		
Durbin-Watson	1.245		1.620		1.644		1.681			
Rsquare original data	.692		.794		.832		.845			

Determinants of coreness

Table 3 presents the results of the regression on the square root of coreness (normalized on a scale from 0 to 100 for each country).

The level of *national coreness* (normalized) is predicted with a high level of precision from the organizational attributes and all coefficients have the expected sign and a high level of significance. Unstardardized residuals fulfill normality test (Kolmogorov-Smirnov statistics .940, p=0.340) thus showing that the transformation of the dependent was effective in addressing normality problems. The introduction of international reputation affects the coefficient of staff, but both are significant and the model is statistically superior to the one with staff only (while the Variance Inflation Factor for international reputation remains moderate).

The standardized coefficients display that size is the most important factor influencing coreness. The negative sign of the quadratic term implies that its impact decreases with size. For small HEIs size has by large the most important effects, whereas quality has only a minor influence on coreness. The only HEI with high reputation and low size in the sample (the London Business School) reaches a level of coreness (normalized) of only 10. In the middle range region (500-1500 academic staff), size remains the main factor determining coreness, but reputation becomes increasingly important and thus middle size international universities tend to be more central in national network than non-university HEIs of similar size. Finally, for the largest HEIs coreness depends only on international reputation.

On the contrary, foundation year and national type for binary countries are statistically not significant. We also tested the urban centrality variable which turns not to be significant both for the general regression and for the specific case of UK (where there is spatial concentration around London). This can be explained by the fact that large cities not only host some of the largest and most reputed HEIs, but as well as a number of smaller and more specialized ones.

At the country level, the model explains between 73% (Switzerland) and 89% of the variance (Netherlands) in the original data, while this drops to only 49% in Norway. Accordingly, there is substantial evidence that the relationship between organizational variables and coreness is largely independent of the specific national characteristics concerning the level of competition, as well as the presence of a binary divide. National policies do not influence directly the network structure, but might do it indirectly through the inequality in the distribution of resources and status (which tends to be larger in binary countries than in the UK and even more in Italy).

Geography, coreness and connectivity

To analyze the interplay between organizational characteristics and geographic distance, we characterize organizations by their *relational mass*, defined as the value of coreness predicted by the regression; this is the best possible combination of organizational characteristics explaining network centrality. We first classify dyads by their total mass and by distance and then we analyze the percentage of counts in each category.

Table 5 shows that, first, the effect of mass prevalent: the share of non-active dyads (0 links) is consistently larger for low mass independently of distance. Second, the effect of distance is stronger for peripheral HEIs and, when mass increases, it moves towards higher counts: if the sum of masses is below 60, distance strongly influences the likelihood of having at least 1 link, whereas between 60 and 120 it influences mostly the likelihood of counts above 100 links; finally, when total mass is very high, the effects is not significant for all levels of counts considered.

Table 5. Dyads by number of links, coreness and distance. Distribution of weblinks by class divided by distance and sum of coreness of sender and receiver

	Sum of mass < 60			Sum of mass between 60 and 119			Sum of mass > 119		
	0 links	1-99 links	>100 links	0 links	1-99 links	>100 links	0 links	1-99 links	>100 links
<10 km	70%	28%	1%	22%	68%	10%	6%	57%	38%
10-100 km	80%	20%	0%	32%	64%	4%	3%	51%	47%
100 - 500 km	90%	10%	0%	46%	53%	1%	4%	65%	31%
>500 km	92%	8%	0%	51%	48%	1%	2%	70%	29%

A binomial hurdle regression provides similar results (Table 6).

Table 6. Results of the binomial regression

	Null model		Mass only		Coreness and distance	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	-8.318	8.151	-10.10	6.638	-10.12	12.80
Mass-sender			.02695	.00053***	.02794	.00054***
Mass-receiver			.02815	.00055***	.02878	.00058***
Log_distance					-.5424	.02154***
Log(theta)	-14.106	8.151’	-12.84	6.638’	-14.00	12.80
Zero hurdle model coefficients (binomial with logit link)						
(Intercept)	-.8415		-3.664	0.020***	-2.141	0.0439***
Mass-sender			.0479	.00034***	.0484	0.00034***
Mass-receiver			.0481	.00034***	.0487	0.00035***
Log_distance					-.4419	0.0170***
Number of iterations	13		29		35	
Log-likelihood	-2.063e+05 on 3 df		-1.776e+05 on 7 df		-1.764e+05 on 9 df	

Signif. Codes 0*** 0.001** 0.01* 0.05’

The model with mass only performs quite well in terms of predictive ability of counts of weblinks. Namely, it identifies 64% of the non-zero dyads and, when it predicts a count higher than 0, it is correct in 78% of the cases. Further, the predictive ability of the model is rather similar for the considered countries, except for Norway where the model identifies only 20% of the non-zero counts and thus the model does not behave well. As expected, the coefficients of sender and receiver mass are almost identical. The model including the log of distance is

statistically superior, but increases only slightly the predictive ability. As a matter of fact, the mass only model provides a largely equivalent result to a full model including separately all organizational and geographical variables, showing that the measure of mass captures almost all organizational effects on interlinking.
Estimates of the predicted probability of interlinking, as well as of the expected counts of links, help disentangling the interaction between mass and distance (Figure 3).

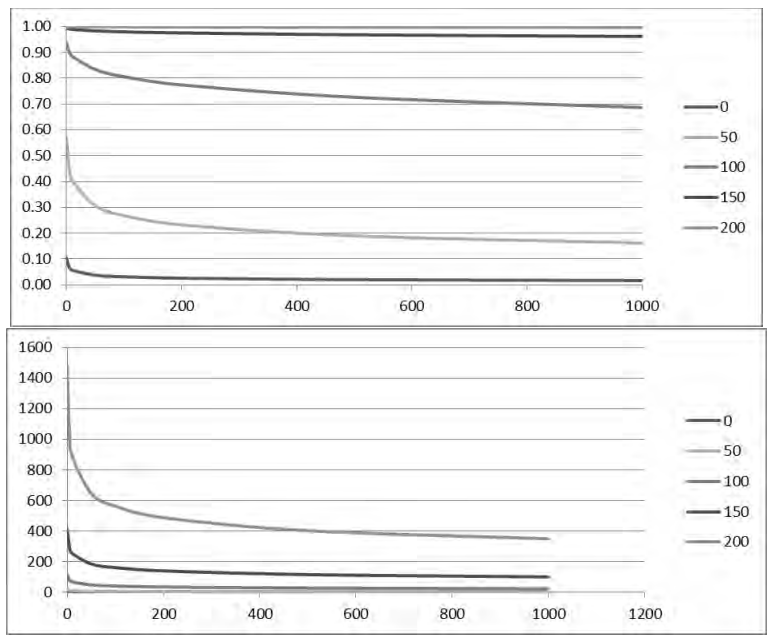


Figure 3. Predicted values of weblinks. Top: probability of interlinking; bottom: expected counts as generated by the model. X: distance in km. Series: sum of the coreness of sender and receiver.

Result are consistent with descriptive analysis. The probability of interlinking remains consistently high when the sum of masses is sufficiently large: core HEIs will be connected independently of the distance, whereas the most peripheral HEIs will be connected only if they are very near (below a scale of about 50 km). In the core-periphery connections (mass near to 100), the likelihood of linking decreases with distance, but remains relatively high at the largest distance found in the countries considered. Thus, core HEIs function as national attractors independently of distance – for an HEI with mass near to 0 the likelihood of linking to a very central HEI (mass=100) at any distance is larger than to a HEI with mass 50 in the same city.
Second, the impact on counts of links is large only at very small distance (below 50 km) and is generally less strong than the one of mass: two HEIs with total mass 200 and 500 km apart are expected to have the same of number of links that

two HEIs of total mass 150 located in the same city. Dyads with high mass and low distance are rather few since large HEIs tend to be distributed across a country in order to respond to the demand for students –only 20% of the dyads in our dataset with total mass above 150 have distance below 100 km. This implies that dyads with large counts will tend to be distributed at a national level and thus distance will not have a strong impact on the overall core/periphery structure (while influencing individual counts when two HEIs become very near).
This analysis helps as well identifying when geography is likely to have a stronger impact on network structure, namely if there is a small number of regional clusters comprising at least one of the largest HEIs (in terms of relational mass) and most of the smaller HEIs, while the geographical size of the clusters is much smaller than the distance between them. Under that condition, connections between large HEIs will remain distributed to the whole country, whereas peripheral HEIs are expected to display larger levels of connectivity thanks to geographical proximity (both to the core and within the periphery).
In Norway there are only four large attractors (the historical universities in Oslo, Bergen, NIST Trondheim and Tromsø) whose average distance approaches 1000 km and clustering most of remaining HEIs (15 out of the 38 remaining HEIs are located in one of these cities). Our model provides evidence that this geographical structure accounts for the characteristics of the Norwegian network, with the lower fit to the core/periphery and a flat distribution of coreness despite inequality in the repartition of resources.
We speculate that similar findings might apply to systems where regional clusters are very far apartand the spatial density of large attractors is very small, , like in the US between West and East coast,; in that case, we would expect regional core/periphery systems to emerge, interconnected by a higher-level network between the largest universities.

Discussion and conclusion

Findings can be interpreted at two levels, a more technical one on the structure of HE interlinking networks, and a more organizational one concerning the structuring of social relationships in HE fields.
First, these results go beyond existing studies, which mainly tried to analyze the determinants of interlinking between two HEIs, to analyze the structural characteristics of the network emerging from connectivity and to which extent they generate regularities in the network structure; this kind of structural investigations has been very common for publication and citation networks, but at our knowledge not frequently adopted for weblinks analysis in science.
We thus demonstrated that national HE interlinking networks display a simple core/periphery structure with a unique center and that the level of centrality in this network is a predictor of the strength of the connection between two HEIs; further, we demonstrated that centrality is closely associated to organizational characteristics and that, for small HEIs, it depends essentially on size, whereas for the largest ones on international reputation. This implies a well-defined repartition

of HEIs in the network, with the center occupied by the largest research universities, the middle range by smaller universities, as well as large non-university HEI and the periphery by the smaller HEIs in the system.
Further we demonstrated that these relationships are basically the same for the considered countries, despite large differences in national policies and in the composition of systems. In our opinion, this hints to the fact that there are deep mechanisms generating weblinks, which are related to organizational activities and characteristics of HEIs. Finally, the models we developed explain why in the considered countries geography, despite having a clear impact on interlinking between HEIs, does not affect the general network structure and why, for example, we do not observe regional clustering. The case of Norway, where departures are observed, suggests that large heterogeneities in the distribution of HEIs, with clearly-defined regional clusters, are likely to impact strongly on network structure. Testing this relationships on countries display a very different geographical scale and organization (like the US) would then advance our understanding of the impact of geography on HE network structure.
Second, these characteristics are consistent with the assumption that weblinks are not just connections between documents published on the web, but rather markers of underlying social relationships between the concerned HEIs, as related to their activities. Weblinks are closely and systematically related to organizational attributes which refer to HEI resources and status and display different distributional properties than citation networks – degree in the considered countries displays a loglinear distribution rather than a power law distribution.
If we accept this assumption, our results can be interpreted in terms of their implication for the structure of HE fields. First, they imply that, despite different policy narratives on HEIs having different status or being similar (like in binary systems), all the considered HE systems have developed a very similar status hierarchy and that binary systems display an even steeper hierarchy than the unitary ones. This conforms to widespread expectations that integration into a unique system leads to a stronger process of hierarchization, as different types of HEIs end providing similar offerings (like bachelor and master studies) and thus hierarchy is required to allow audiences to make choices (Bleiklie 2003). Second, economic sociology considers that the position in social network is closely associated to access to resources (White 1981, Burt 1988); accordingly, the more central HEIs in the network benefit of better opportunities to access to resources, collaborations, people and thus strengthen further their position. Accordingly, in the structuring process of the HE organizational field status hierarchy and network centrality coevolve and reinforce each other to produce the observed core/periphery structure (Owen-Smith and Powell 2008). This implies that relational structures become a central element in ensuring the stability of the status layering of national HE systems – hence the broader interest of developing methods and techniques for observing them.

Acknowledgements

The authors would like to than Isidro Aguillo, Webometrics lab, Madrid for providing the weblink data, as well as Alessandro Lomi (Lugano) for useful advice on the paper.

References

Abbasi, A., Hossain, L. & Leydesdorff, L. (2012). Betwenenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics, 6(3)*, 403-412.

Adams, J., Faust, K. & Lovasi, G. (2012). Capturing context: Integrating spatial and social network analyses. *Social Networks, 34(1)(1-5)*.

Aguillo, I., Granadino, B., Ortega, J. L. & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetric indicators. *Journal of the American Society for Information Science and Technology, 57(10)*, 1296-1302.

Bar-Ilan, J. (2009). Infometrics at the beginning of the 21st century - A review. *Journal of Infometrics, 2(1)*, 1-52.

Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country - the case of Israel. *Scientometrics, 59(3)*, 391-403.

Bleiklie, I. (2003). Hierarchy and Specialisation: on the institutional integration of higher education systems. *European Journal of Education, 38(4)*, 341-355.

Borgatti, S. P. & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks, 21*, 375–395.

Borgatti, S. P., Everett & Freeman (2002). Ucinet 6 for Windows.

Buela-Casal, G., Gutierrez-Martinez O., Bermudez-Sanchez M. P. & Vadillo-Munoz O. (2007). Comparative Study of international academic rankings of universities. *Scientometrics, 71(3)*, 587-596.

Burris, V. (2004). The Academic Caste System: Prestige Hierarchies in PhD Exchange Networks. *American Sociological Review, 69(2)*, 239-264.

Burt, R. S. (1988). The Stability of American Markets. *American Journal of Sociology, 93*, 356–95.

Butts, C. T., Acton, R. M., Hipp, J. R. & Nagle, N. N. (2012). Geographical variability and network structure. *Social Networks, 34(1)*, 82-100.

Cattani, G., Ferriani, S., Negro, G. & Perretti, F. (2008). The Structure of Consensus: Network Ties, Legitimation, and Exit Rates of U.S. Feature Film Producer Organizations. *Administrative Science Quarterly, 53(1)*, 145-182.

Chinchilla-Rodriguéz, Z., Ferligoi, A., Miguel, S., Kronegger, L. & de Moya-Anegón, F. (2012). Blockmodeling of co-authorship networks in library and information science in Argentina: a case study. *Scientometrics, 93(3)*, 699-717.

Daraganova, G., Pattison, P., Koskinen, J., Mitchell, B., Bill, A., Watts, M. & Baum, S. (2012). Networks and geography: Modelling comunity network structures as the outcome of both spatial and network processes. *Social Networks, 34(1)*, 6-17.

Ferlie, E., Musselin & Andresani (2008). The steering of higher education systems: a public management perspective. *Higher Education, 56(3)*, 325-348.
Freeman, L. C. (1978/79.). Centrality in Social Networks. Conceptual Clarifications. *Social Networks, 1*, 215-239.
Frenken, K., Hardeman, S. & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics, 3(3)*, 222-232.
Glänzel, W. & Schubert, A. (2005). Analysing scientific networks through co-authorship. In H. F. Moed, W. Glänzel & U. Schmoch(Eds.) *Handbook of Quantitative Science and Technology Research* (pp. 257-276). Dordrecht: Kluwer Academic Publications.
Gonzalez-Bailon, S. (2009). Opening the black box of link formation: Social factors underlying the structure of the Web. *Social Networks, 31(4)*, 271-280.
Heller-Schuh, B., Barber, M., Henriques, L., Paier, M., Pontikakis, D., Scherngell, T., Veltri, G. A. & Weber, M. (2011). *Analysis of Networks in European Framework Programmes (1984-2006)* Luxembourg: Publications Office of the European Union.
Hoekman, J., Frenken, K. & Tijssen, R. (2010). Research collaboration at distance: Changing spatial patterns of scientific collaboration in Europe. *Research Policy, 39(5)*, 662-673.
Holmberg, K. & Thelwall, M. (2009). Local government websites in Finland: A geographic and webometric analysis. *Scientometrics, 1*, 157-169.
Holme, P. (2005). Core-periphery organization of complex networks. *Physical Review E, 72(4)*.
Jones, B. F., Wuchty, S. & Uzzi, B. (2008). Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science, 322(5905)*, 1259-1262.
Kronegger, L., Ferligoj, A. & Doreian, P. (2011). On the dynamics of national scientific systems. *Quality and Quantity, 45*, 989-1015.
Kyvik, S. (2004). Structural Changes in Higher Education Systems in Western Europe. *Higher Education in Europe, 29 (3)*, 393-409.
Kyvik, S. & Lepori, B. (2010). *Research in the non-university higher education sector in Europe* Dordrecht: Springer.
Laumann, E. O., Galaskiewicz, J. & Marsden, V. P. (1978). Community Structure as Interorganizational Linkages. *Annual Review of Sociology, 4*, 455-484.
Lepori, B. & Bonaccorsi, A. (2013). Towards an European census of higher education institutions. Design, methodological and comparability issues. *Minerva, .*
Lepori, B. & Kyvik, S. (2010). The research mission of Universities of Applied Science and the future configuration of Higher Education systems in Europe. *Higher Education Policy, 23*, 295-316.
Lepori, B. (2011). Coordination modes in public funding systems. *Research Policy, 40(3)*, 355-367.

Lepori, B., Benninghoff, M., Jongbloed, B., Salerno, C. & Slipersaeter, S. (2007). Changing models and patterns of higher education funding: Some empirical evidence. In A. Bonaccorsi & C. Daraio(Eds.) *Universities and Strategic Knowledge Creation. Specialization and Performance in Europe* (pp. 85-111). Bodmin, Cornwall: MPG Books Limited.

Mullahy, J. (1986). Speci cation and Testing of Some Modified Count Data Models. *Journal of Econometrics, 33*, 341-365.

Nieminen, M. & Auranen, O. (2010). University research funding and publication performance - an international comparison. *39*, 822-834.

Ortega, J. L., Aguillo, I., Cothey, V. & Scharnhorst, A. (2008). Maps of the academic web in the European Higher Education Area - an exploration of visual web indicators. *Scientometrics, 74(2)*, 295-308.

Owen-Smith, J. & Powell, W. W. (2008). Networks and institutions. In R. Greenwood, C. Oliver, K. Shalin & R. Suddaby(Eds.) *The SAGE Handbook of Organizational Institutionalism* (pp. 594-621). London.

Paradeise, C., Reale, E., Bleiklie, I. & Ferlie, E. (2009). *University Governance. Western European Comparative Perspectives.*Dordrecht: Springer.

Rivera, M. T., Soderstrom, S. B. & Uzzi, B. (2010). Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms. *Annual Review of Sociology, 36(91-115)*.

Schofer, E., Meyer, J. (2005). The Worldwide Expansion of Higher Education in the Twentieth Century. *American Sociological Review, 70(6)*, 898-920.

Seeber, M., Lepori, B., Lomi, A., Aguillo, I. & Barberio, V. (2012). Factors affecting weblink connections between European higher education institutions. *Journal of Infometrics, 6(3)*, 435-447.

Snyder, D. & Kick, E. L. (1979). Structural Position in the World System and Economic Growth, 1955-1970: A Multiple-Network Analysis of Transnational Interactions. *American Journal of Sociology, 84(5)*, pp. 1096-1126.

Stiles, D. R. (2000). Higher Education Funding Patterns Since 1990: A New Perspective. *Public Money & Management, 20(4)*, 51-57.

Taylor, P. J. (2004). *World City Network: a Global Urban Analysis* London: Routledge.

Thelwall, M. (2002a). A research and institutional size based model for national university web site interlinking. *Journal of Documentation, 58(6)*, 683-694.

Thelwall, M. (2002b). Evidence for the existence of geographic trends in university web site interlinking. *Journal of Documentation, 58(5)*, 563-574.

Thelwall, M. & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organizations. *Journal of the American Society for Information Science and Technology, 62(8)*, 1488-1497.

Thelwall, M., Tang, R. & Price, E. (2003). Linguistic patterns of academic web use in Western Europe. *Scientometrics, 56(3)*, 417-432.

Thelwall, M. & Zuccala, A. (2008). A university-centred European Union link analysis. *Scientometrics, 75(3)*, 407-420.

van Raan, A. F. J. (2007). Bibliometric statistical properties of the 100 largest European universities: prevalent scaling rules in the science system. *Eprint arXiv:0704.0889,* .

Wagner, C. S. & Leydesdorff, L. (2005). Network Structure, Self Organization, and the Growth of International Collaboration in Science. *Research Policy, 34 (10)*, 1608-1618.

White, H., Boorman, S. & Breiger, R. (1976). Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. *The American Journal of Sociology, 81(4)*, 730.

White, H. C. (1981). Where Do Markets Come From? *American Journal of Sociology, 87*, 517–47.

Wilkinson, D., Harries, G., Thelwall, M. & Price, L. (2003). Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science, 29(1)*, 49-56.

CORRELATION AMONG THE SCIENTIFIC PRODUCTION, SUPERVISIONS AND PARTICIPATION IN DEFENSE EXAMINATION COMMITTEES IN THE BRAZILIAN PHYSICISTS COMMUNITY (RIP)

Rogério Mugnaini[1], Luciano A. Digiampietri[1], and Jesús P. Mena-Chalco[2]

[1] *mugnaini@usp.br*
Universidade de São Paulo, Escola de Artes, Ciências e Humanidades, Av. Arlindo Bettio, 1000, CEP 03828-000, São Paulo, SP, Brazil

[1] *digiampietri@usp.br*
Universidade de São Paulo, Escola de Artes, Ciências e Humanidades, Av. Arlindo Bettio, 1000, CEP 03828-000, São Paulo, SP, Brazil

[2] *jesus.mena@ufabc.edu.br*
Universidade Federal do ABC,Centro de Matemática, Computação e Cognição, Rua Santa Adélia, 166, CEP 09210-170, Santo André, SP, Brazil

Abstract
This paper analyses the correlation among the scientific production, supervisions and participation in defense committees in the Brazilian physicists' community. 4,649 curricula of PhD in Physics were evaluated and 16 performance indicators were considered: 6 types of scientific bibliographic production; 5 types of supervisions; 4 types of participation in defense committees; and the number of years after the doctoral defense. Some of the most relevant correlations among these indicators are presented in this paper, including a discussion of the characteristics and behaviours that are inherent to this academic community. Over the past sixty years, 1951-2010, a substantial correlation between the publications and the number of doctorate supervisions was identified.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3)

Introduction
The analysis and assessment of researchers' activities are important tasks in order to understand the scientific/academic communities and promote successful initiatives. In the last decades, the Brazilian production has greatly increased. This growth can be measured, for example, by the number of bibliographical production that the Brazilian academic communities has published (Coutinho, et

al. 2012; de Meis, et al. 2003; Glanzel, et al., 2006; Velho & Krige,1984; Zorzetto, et al., 2006).
Considering that the production of knowledge is influenced by the interaction, exchange and collaboration among actors of the Science & Technology system (Sanz-Menendez, 2001), would be interesting to measure the influences of other academic activities – such as supervision of students and participation in defense examination committees – in the researchers' productivity. Some studies were devoted to explain the doctoral researchers' productivity (Salmi, Gana & Mouillet, 2001; Mallette, 2006; Larivière, 2012; Tuesta et al., 2012). In this context, López-Cózar et al. (2006) analyzed supervision and participation in examination committees aiming to understand the social structure of the research in a specific subject among Spanish universities.
However, little is known about how the scientific production, supervisions and participation in defense examination committees are related, as well the degree in which they influence the behaviour of academic researchers. This paper aims to analyse the correlations among indicators of scientific production, supervisions and participation in defense examination committees.
The results presented in this manuscript are part of an ongoing project focused on characterizing the Brazilian Scientific Community according to different performance metrics, including the scientific production, visibility and academic network analyses. The academic data was organized according to the number of years after the PhD defense in order to identify the behavior of the Brazilian physicists along the years.

Materials and Methods

The research presented in this paper was developed following three main steps: data gathering; organization; and data analyses. All data used in this paper was automatically gathered from the Lattes Platform.
Lattes Platform[46] is an online academic system maintained by the Brazilian National Council for Scientific and Technological Development (CNPq) to congregate curricula and other information about the main professionals and researchers in Brazil. The database contains more than 2 million curricula, and each curriculum is composed of information about academics degrees; bibliographic production; artistic production; supervisions; professional experience; among others (Amorin, 2003).
The curricula are available in HTML format and can be accessed by a unique identifier of 16 digits assigned automatically by the platform to each researcher. Another way to access academic curricula is using a web search tool provide by the Lattes Platform where one can search curricula by the researchers' name or main area of interesting.
This study is part of a project that aims to analyze all the Brazilian Scientific Community and, in order do to this, the curricula from all major areas were

[46] http://lattes.cnpq.br

download in HTML format (*data gathering step*) (Mena-Chalco & Cesar-Jr, 2011).

The *organization step* was divided into two activities. The first one was the parsing of the HTML curricula files in order to identify and separate each of the fields from the curricula and save it in a relational database. In order to do this automatically, some computational procedures were developed. The second activity was related with the selection among the curricula those belonging to PhD in Physics and associated to the knowledge area of Physics. Thus, Physicists who work in other areas are not being considered in this analysis. The selection was made querying the database constructed in the first activity, and it was able to identify 4,649 curricula (of Physics' PhD which doctoral defense occurred before 2011). From each Lattes curriculum were considered 16 indicators - all of them extracted from the database using database queries:

- ***Bibliographic production*** (6 indicators): Article in scientific journals, Complete work published in proceedings of conferences, Expanded abstract published in proceedings of conferences, Abstract published in proceedings of conferences, Book published/organized and Book chapter published;
- ***Supervisions*** (5 indicators): Post-doctoral, Ph.D. thesis, Master thesis, Undergraduate Research and Works of completion for graduation (monograph);
- ***Participation in defense examination committees*** (4 indicators): Ph.D. thesis, Master thesis, Graduation monograph and Public concourse;
- ***Number of years after the doctoral defense***: the curricula were grouped according to eight groups (0 to 4 years since the defense, 5 to 9, and so on, until to 30 to 34, and 35 or more years).

In the *data analyses step* all the correlations between each pair of indicators were calculated and the most relevant results are presented and discussed in the next section.

Findings and discussion

The cross correlations between the different sets of indicators can be observed in Figure 1.

Figure 1(A) shows that the total number of supervisions are more correlated with publications than the total number of participations in defense committees; and, among the different types of scientific production, total number of abstracts is the most correlated. Since advisees tend to publish any piece of research in conferences, it may explain such correlation.

In Figure 1(B), it is possible to observe that the number of supervision of doctoral and post-doctoral researchers are more correlated with total production while undergraduate research and graduation monograph are correlated with total participations in defense committees. Considering master supervision, the total number of productions and participations in defense committees correlates

equally. This situation suggests that: the higher the degree of the supervision, it is more correlated with total production; and contrarily, the lower the degree of the supervision, smaller is the correlation with total production and a bigger the correlation with participations in defense committees.

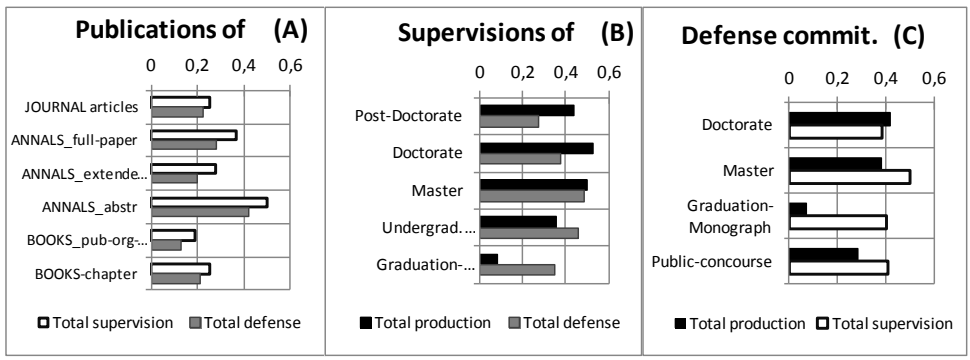


Figure 1. Correlations between: (A) types of publications and the total number of supervisions and participations in defense committees; (B) supervision degree and total number of productions and participations in defense committees; (C) defense committee degree and total number of productions and supervisions.

Figure 1(C) shows that the participations in defense committees of all degrees are correlated to the total number of supervisions, however the total number of production is mostly correlated with supervisions of doctorate and master, indicating that the most prolific are more frequently invited to defense committees of these degrees.

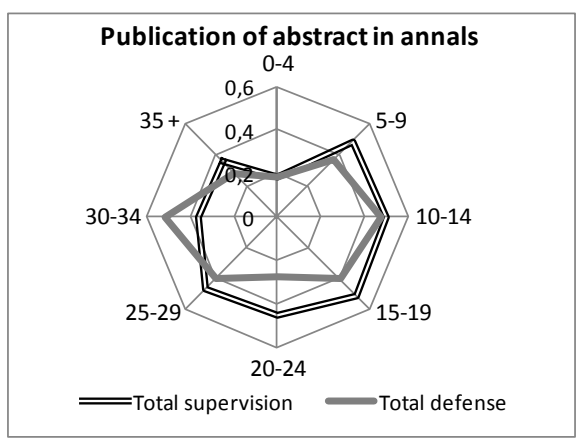


Figure 2. Correlations between publication of abstracts in annals and total of supervisions and total participations in defense committees, according to the elapsed time (years) since the end of the doctorate.

Considering specifically the correlation of abstracts published in annals, in the different elapsed time since the end of the doctorate (Figure 2), the correlation with supervisions is getting lower over time while participations in defense committees does not present an evidence of correlation.

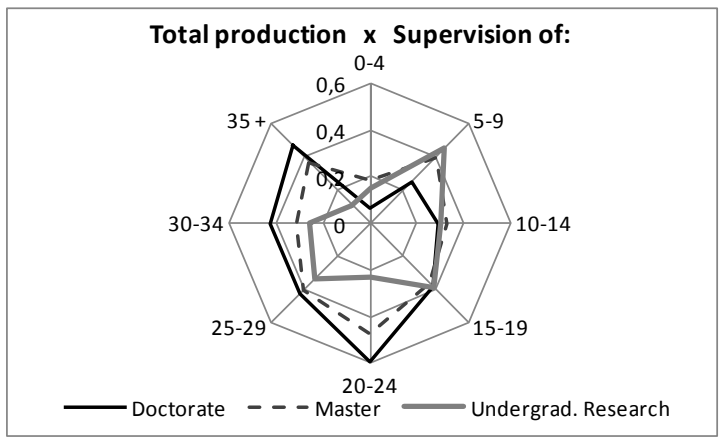


Figure 3. Correlations between supervision of different degrees and total production, according to the elapsed time since the end of the doctorate.

Figure 3 shows that the correlation between total production and supervision of undergraduate research gets lower over time, while it decreases more slowly considering masters' supervision. Doctorate supervision has opposite behavior, getting higher over time.

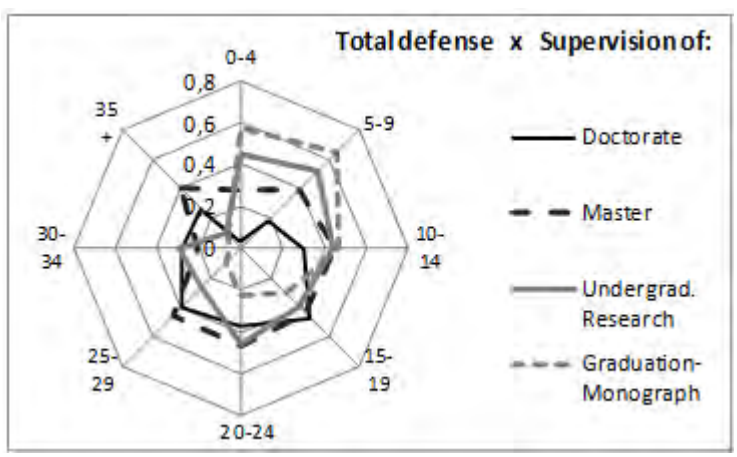


Figure 4. Correlations between supervision of different degrees and total participations in defense committees, according to the elapsed time since the end of the doctorate.

Even though when correlating supervision indicators with total participations in defense committees (Figure 4), one can observe that the lower the degree of the supervision, the faster is the decreasing of correlation over time. The correlation of masters' supervision keeps almost the same over time, while doctorate supervision increases, but both dropped after 30 years from doctorate.
Figure 5 (A) reveals that the correlation between total of supervisions and defense committee of graduation monograph decreases quickly over time, showing that one's participation in defense committee of master and doctorate becomes more correlated, over time, to the total number of supervisions.
The correlations between total production and participation in defense committee of master and doctorate are practically the same until 29 years since the end of doctorate while the correlations between total production and defense committee of public concourse starts very low, aligning with a doctorate after 19 years (Figure 5 (B)). The constant decreasing of the correlation of participation in masters' defense committee corroborates the idea that experienced and prolific advisors stop participating in those committees.

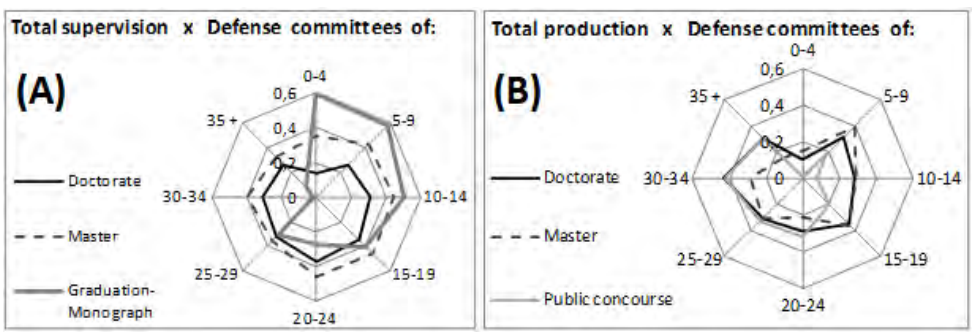


Figure 5. Correlations according to the elapsed time since the end of the doctorate: (A) between participations in defense committees of different degrees and total of supervisions,; and (B) participations in defense committees of different degrees and total production,.

Final remarks

This paper analyzed the correlation among the scientific production, supervisions and participation in defense committees of 4,649 PhD in Physics considering 16 indicators.
A substantial correlation between the total number of publications of a researcher and the number of doctorate supervisions was identified. The importance of doctorate supervisions in the academic production of researchers was evidenced in other studies including in different areas (Larivière, 2012).
It is also worth to notice the high correlation between the number of abstracts published in annals and the total number of supervisions. It suggests that the advisors try to publish any piece of research developed with their advisees (at least as an abstract).

Another interesting correlation was found between the participation in master defense committees and the total number of master supervisions. It may indicate that by inviting other researchers to participate in the defense committee of your advisees there will be a good chance to be invited to participate in the defense committee of the advisees of these other researcher (as a type of favor exchange).
We believed that the information presented in this work could be of great value to Brazilian policy-makers in government, academia, and industry in order to explore, quantify and understand the academic correlation among the scientific production, supervisions and participation in defense examination committees, as well as, the degree in which each indicator influence the behaviour of academic researchers.
This paper is part of an ongoing project which aims to evaluate the Brazilian Scientific Community using different metrics (including bibliographic production, visibility, participation in committees, and metrics from the social network analysis). The research interaction, in the form of bibliographic co-authorship, has potential that will be explored.

References

Amorin, C.V. (2003). Curriculum vitae organization: the Lattes software platform. *Pesquisa Odontológica Brasileira,* 17(1):18-22.

Coutinho, R., Davila, E., dos Santos, W., Rocha, J., Souza, D., Folmer, V., & Puntel, R. (2012). Brazilian scientific production in science education. *Scientometrics*, 92, 697-710.

de Meis, L., Velloso, A., Lannes, D., Carmo, M. S., & de Meis, C. (2003). The growing competition in Brazilian science: Rites of passage, stress and burnout. *Brazilian Journal of Medical and Biological Research*, 36, 1135-1141.

Glanzel, W., Leta, J., & Thijs, B. (2006). Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics*, 67(1), 67-86.

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics*, 90, 463-481.

López-Cózar, E.D. et al. (2006). Análisis bibliométrico y de redes sociales aplicado a las tesis bibliométricas defendidas en España (1976-2002): temas, escuelas científicas y redes académicas. *Revista Española de Documentación Científica*, 29 (4), 493-524.

Mallette, L.A. (2006). *Publishing rates of graduates education Ph.D. and Ed.D. students:* A longitudinal study of University of California schools (Doctoral dissertation, Pepperdine University). Retrieved from http://gradworks.umi.com/32/39/3239922.html

Mena-Chalco, J.P. & Cesar-Jr, R. M. (2011). Towards automatic discovery of co-authorship networks in the brazilian academic areas. In *IEEE Seventh International Conference on e-Science Workshops 2011*, pp 53-60.

Salmi, L.R., Gana, S. & Mouillet, E. (2001). Publication pattern of medical theses, France 1993-98. *Medical Education*, 35(1), 18-21.

Sanz-Menéndez, L. *Indicadoresrelacionales y redessociales en el estudio de los efectos de las políticas de ciencia y tecnología*. Madrid: Consejo Superior de Investigaciones Científicas, 2001. Retrieved from http://digital.csic.es/bitstream/10261/1476/1/dt-0109.pdf

Tuesta, E. F. et al. Análise temporal da relação orientador-orientado: um estudo de caso sobre a produtividade dos pesquisadores doutores da área de Ciência da Computação. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2012*. Retrieved from http:// http://www.lbd.dcc.ufmg.br/colecoes/brasnam/2012/0011.pdf

Velho, L., & Krige, J. (1984). Publication and citation practices of Brazilian agricultural scientists. *Social Studies of Science*, 14(1), 45-62.

Zorzetto, R., Razzouk, D., Dubugras, M. T. B., Gerolin, J., Schor, N., Guimaraes, J. A., & Mari, J. J. (2006). The scientific production in health and biological sciences of the top 20 Brazilian universities. *Brazilian Journal of Medical and Biological Research*, 39, 1513–1520.

COUNTING PUBLICATIONS AND CITATIONS: IS MORE ALWAYS BETTER?

Ludo Waltman, Nees Jan van Eck, and Paul Wouters

{waltmanlr, ecknjpvan, p.f.wouters}@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

Abstract
Is more always better? We address this question in the context of bibliometric indices that aim to assess the scientific impact of individual researchers by counting their number of highly cited publications. We propose a simple model in which the number of citations of a publication depends not only on the scientific impact of the publication but also on other 'random' factors. Our model indicates that more need not always be better. It turns out that the most influential researchers may have a systematically lower performance, in terms of highly cited publications, than some of their less influential colleagues. The model also suggests an improved way of counting highly cited publications.

Conference Topic
Scientometrics Indicators (Topic 1) and Modeling the Science System, Science Dynamics and Complex System Science (Topic 11).

Introduction
When bibliometrics is used for research assessment purposes, a general presumption seems to be that more is always better: The more publications, the better; the more citations, the better. At the same time, there is an increasing awareness that 'more is always better' should not be taken too literally. For instance, interpreting the number of citations of a publication as an *approximate* measure of the scientific impact of the publication, having more citations does not *always* coincide with having more impact. Publications with more citations may *on average* have more impact, but *individual* publications may deviate from this pattern. One could hypothesize, for instance, that authors of a publication tend to copy a substantial part of their reference list from the reference lists of earlier publications, often without paying serious attention to the contents of the referenced works (Simkin & Roychowdhury, 2003, 2005). If there is indeed some truth in this idea, it does not seem unlikely that publications sometimes become highly cited without actually having a lot of impact on subsequent scientific research. This illustrates that there does not exist a perfect relationship between scientific impact and citations. In addition to scientific impact, there are many other factors that may influence a publication's number of citations (Bornmann & Daniel, 2008; Martin & Irvine, 1983; Moed, 2005; Nicolaisen, 2007). Some of these factors are of a systematic nature, while others can be considered to have a

more random character. In this paper, we are especially interested in these random factors.

Also when assessing the scientific impact of an oeuvre of publications rather than a single individual work, the more-is-better idea should be treated with care. It is not obvious, for instance, whether comparing the oeuvres of two researchers based on each researcher's total number of citations is a good approach. One researcher may have more citations than another researcher, but it could be that the latter researcher has authored a number of highly cited publications while the former researcher has earned his citations by producing an extensive oeuvre consisting exclusively of lowly and moderately cited works. In this situation, the researcher with the highly cited publications may actually have been more influential, despite his smaller overall number of citations. When assessing a researcher's scientific impact based on the total number of citations of his publications, the implicit assumption is that the number of citations of a publication is proportional to the scientific impact of the publication. This is a rather strong assumption. As argued by Ravallion and Wagstaff (2011), the true relationship between scientific impact and citations may well be non-linear.

In recent years, a large number of bibliometric indices were introduced that may serve as an alternative to counting a researcher's total number of citations. The best-known example is the *h*-index (Hirsch, 2005). This index is robust both to publications with only a small number of citations and to publications with a very large number of citations. This robustness is often considered a strong property of the *h*-index. Unfortunately, however, the *h*-index has other properties that are difficult to justify and that may cause inconsistencies in the results produced by the index (Waltman & Van Eck, 2012). An attractive alternative to the *h*-index is the highly cited publications (HCP) index (Bornmann, 2013; Waltman & Van Eck, 2012). This index counts the number of publications of a researcher that have received at least a certain minimum number of citations (e.g., Plomp, 1990, 1994). The HCP index has a similar robustness property as the *h*-index, but it does not suffer from the inconsistencies of this index.

In this paper, our focus is on the HCP index. The research question that we consider is whether more is always better when counting highly cited publications. To address this question, we introduce a simple model of the relationship between scientific impact and citations. The model shows that, as a consequence of random factors that influence the number of citations of a publication, the answer to our research question is negative. In itself, this may not be considered surprising. When working with small numbers of publications, it is to be expected that random factors may cause deviations from the more-is-better principle. For instance, a researcher with one highly cited publication need not always be more influential than a researcher who does not have any highly cited publications. However, our model reveals that random factors may result in deviations from the more-is-better principle that are of a systematic nature. These deviations occur even when dealing with large numbers of publications. In concrete terms, the model demonstrates how random effects may lead to

paradoxical situations in which the most influential researchers have a systematically lower performance, in terms of highly cited publications, than some of their less influential colleagues. The model also suggests how the HCP index can be modified to avoid these paradoxical situations.

Before proceeding with our analysis, it is important to emphasize that the problem studied in this paper does not relate specifically to the HCP index. We focus on the HCP index because it is an important bibliometric index that, due to its simplicity, can be analyzed in a convenient way. However, findings similar to ours can be made for other bibliometric indices as well. Examples of such indices include the *h*-index (Hirsch, 2005) and its many variants, but also the generalizations of the HCP index recently proposed by Leydesdorff, Bornmann, Mutz, and Opthof (2011).

Scientific impact vs. citations

A crucial distinction in our analysis is between the scientific impact of a publication and the number of citations the publication has received. The scientific impact of a publication is the influence a publication has on subsequent scientific research. The number of citations of a publication partly reflects the scientific impact of the publication, but it also depends on a multitude of other factors (Bornmann & Daniel, 2008; Martin & Irvine, 1983; Moed, 2005; Nicolaisen, 2007). For instance, the number of citations of a publication may depend on the reputation of the authors, of the institutions with which the authors are affiliated, or even of the countries in which the authors are located. The citation behavior of researchers may play a role as well. If a researcher has a strong tendency to cite his own work, this obviously increases the number of citations of his publications. Scientific impact, reputation, and citation behavior are examples of factors that can be expected to have a systematic effect on the number of citations of a researcher's publications. If a researcher produces influential work, has a good reputation, or has a strong self citation tendency, this is likely to increase the number of citations of his publications in a systematic way.

The number of citations of a publication also depends on factors that can be considered to be more of a random nature (e.g., Dieks & Chang, 1976). Unlike the factors mentioned above, these random factors do not create a systematic advantage for the publications of one researcher compared with the publications of another research. It has been argued, for instance, that a substantial proportion of the references in a publication tend to be of a perfunctory nature (e.g., Moravcsik & Murugesan, 1975). These references are not essential for the citing publication but just serve to indicate that more work has been done on the same topic. The choice of perfunctory references tends to be quite arbitrary, since in many cases just a few publications are cited from a much larger set of publications that could all be cited equally well. Because of this arbitrariness, perfunctory references can be seen as a random factor influencing the number of times a publication is cited. Each researcher now and then benefits from

perfunctory referencing, and there is no reason to expect the publications of one researcher to be advantaged in a systematic way over the publications of another researcher.

Although the choice of perfunctory references involves a significant degree of arbitrariness, one may expect that perfunctory references are more likely to refer to publications that already have a substantial number of citations than to publications with only a few citations. The former publications are more visible in the scientific literature and may therefore be more likely to receive additional citations. This would for instance be the case if researchers choose perfunctory references by more or less randomly selecting references from the reference lists of earlier publications (Simkin & Roychowdhury, 2003, 2005) or if researchers simply choose to refer to publications that are highly ranked by a search engine such as Google Scholar (i.e., a search engine that gives a substantial weight to citations to determine the ranking of publications). So random factors influencing the number of citations of a publication may create a self-reinforcing effect (often referred to as 'cumulative advantage', 'Matthew effect', or 'preferential attachment'; e.g., Price, 1976). The more citations a publication has, the more likely the publication is to receive additional citations.

More need not always be better

To address the question whether more is always better when counting highly cited publications, we introduce a simple model of the relationship between scientific impact and citations. Our model does not intend to provide an accurate representation of the many different factors influencing the number of citations of a publication. Instead, by introducing a number of simplifications, we aim to create an easy-to-understand model that still gives relevant insights into the more-is-better question.

In our model, we assume that scientific impact is the only systematic factor influencing the number of citations of a publication. Other systematic factors, such as reputation and citation behavior, are disregarded. Very importantly, however, we do incorporate in our model the idea that the number of citations of a publication may be influenced by random factors. To keep the model as simple as possible, we treat the scientific impact of a publication as a binary variable. A publication either does or does not have scientific impact. This is of course a highly unrealistic assumption. We will come back to this at the end of the paper.

We are interested in measuring researchers' overall scientific impact. We assume that the overall scientific impact of a researcher is determined by the number of high-impact publications the researcher has produced. We also assume that 10% of the publications in a scientific field have a high impact. The other 90% of the publications have a low impact. The scientific impact of low-impact publications is considered to be negligible.

The scientific impact of a publication cannot be directly observed, and we therefore look at the number of citations of a publication. We distinguish between two classes of publications: Publications that belong to the top 10% of their field

in terms of citations and publications that, based on their number of citations, do not belong to the top 10% of their field. We refer to publications belonging to the top 10% most frequently cited of their field as highly cited publications.[47] Publications that do not belong to the top 10% most frequently cited of their field are referred to as lowly cited publications. Counting the number of highly cited publications of a researcher yields the above-mentioned HCP index.

In an ideal world in which there is a perfect relationship between the scientific impact of a publication and a publication's number of citations, being highly cited coincides with having a high impact. In other words, each highly cited publication is also a high-impact publication, and the other way around. In such an ideal world, the HCP index perfectly indicates the number of high-impact publications of a researcher, and the index therefore always provides a correct assessment of a researcher's overall scientific impact.

However, as we have discussed, the idea of a perfect relationship between scientific impact and citations is difficult to justify. In our model, random factors cause some publications to be highly cited even though they have only a limited scientific impact. Conversely, some publications do not belong to the top 10% most highly cited publications of their field even though they do belong to the 10% high-impact publications. A possible scenario is illustrated in Table 1. In this scenario, 3% of the publications in a field have a high impact and are also highly cited, while 7% of the publications have a high impact but are not highly cited and another 7% of the publications are highly cited but do not have a high impact. The remaining 83% of the publications have a low impact and are also lowly cited. In the scenario illustrated in Table 1, if a publication has a high impact, there is a probability of 3% / 10% = 0.30 that the publication is highly cited. If a publication has a low impact, this probability is just 7% / 90% ≈ 0.08. Hence, high-impact publications are (3% / 10%) / (7% / 90%) ≈ 3.86 times as likely to be highly cited as low-impact publications.

Table 1. Illustration of a scenario in which there is no perfect relationship between the scientific impact of a publication and a publication's number of citations.

	Lowly cited pub.	Highly cited pub.	Total
Low-impact pub.	83%	7%	90%
High-impact pub.	7%	3%	10%
Total	90%	10%	100%

In the scenario illustrated in Table 1, we may have the following interesting situation. Suppose we have two researchers, researcher A and researcher B (see Table 2). Researcher A has produced 100 publications, all of them of high impact. Researcher B has produced 500 publications, so five times as many as researcher

[47] For the purpose of our analysis, practical difficulties in determining whether a publication belongs to the top 10% most frequently cited (Waltman & Schreiber, 2013) can be ignored.

A, but none of these publications is of a high impact.[48] Given our assumption that a researcher's overall scientific impact is determined by the number of high-impact publications the researcher has produced, we must conclude that researcher A has been highly influential while the scientific impact of researcher B has been negligible, despite the large publication output of this researcher.

Table 2. Four hypothetical researchers that are used to illustrate the consequences of different approaches to counting highly cited publications.

	Number of publications		Number of publications	
	low-impact	high-impact	lowly cited	highly cited
Researcher A	0	100	70	30
Researcher B	500	0	461	39
Researcher C	50	200	186	64
Researcher D	270	70	298	42

The interesting question is whether the HCP index confirms this conclusion. Given the percentages reported in Table 1, we can expect researcher A to have (3% / 10%) × 100 = 30 highly cited publications. For researcher B, the expected number of highly cited publications is (7% / 90%) × 500 ≈ 39. If researchers A and B indeed each have their statistically expected number of highly cited publications, we end up in the paradoxical situation in which the HCP index indicates that researcher B, with an HCP value of 39, appears to be more influential than researcher A, with an HCP value of 30. Hence, the HCP index provides an incorrect assessment of the overall scientific impact of the two researchers. Moreover, this incorrect assessment is not caused by an incidental statistical fluctuation. Since researchers A and B each have their statistically expected number of highly cited publications, the HCP index is systematically wrong in situations like ours.

Why does the HCP index in certain situations provide systematically incorrect assessments of researchers' overall scientific impact? This is because, as long as there is no perfect relationship between scientific impact and citations, a researcher with a given number of high-impact publications can always be outperformed, in terms of highly cited publications, by another researcher with a sufficiently large number of low-impact publications. Low-impact publications are less likely to become highly cited than high-impact publications, but by producing lots of low-impact publications it is still possible to obtain a large number of highly cited publications.

The above scenario demonstrates that more need not always be better when counting highly cited publications. There can be systematic deviations from the more-is-better principle. In particular, the HCP index may overestimate the

[48] In the theoretical examples presented in this paper, we know each publication's impact. This is helpful to illustrate our ideas. In practice, however, the impact of a publication cannot be directly observed.

scientific impact of researchers who focus on producing lots of publications without paying much attention to the impact of their work.

Table 3 shows a generalization of the scenario illustrated in Table 1. The parameter α determines the degree to which scientific impact and citations are correlated. A perfect correlation is obtained by setting α equal to zero. The other extreme is to set α equal to 0.09, in which case scientific impact and citations are completely uncorrelated and the number of citations of a publication provides no indication at all of the scientific impact of the publication. The absence of any correlation between scientific impact and citations for $\alpha = 0.09$ follows from the fact that setting α equal to 0.09 causes each cell in Table 3 to be equal to the product of the corresponding row and column totals, making scientific impact and citations statistically independent from each other. The possibility of setting α equal to a value above 0.09 can be ignored. This would lead to the implausible situation of a negative correlation between scientific impact and citations. Setting α equal to 0.07 yields the scenario illustrated in Table 1. In the end, the value of α that one considers most realistic depends on how much trust one has in the ability of citations to indicate the scientific impact of a publication. It also depends on the exact interpretation that one gives to the notion of scientific impact. Moreover, since citation cultures differ across scientific fields, it may well be that different fields require different values of α.

Table 3. Scientific impact vs. citations. The parameter α determines the degree of correlation ($0 \leq \alpha \leq 0.09$).

	Lowly cited pub.	Highly cited pub.	Total
Low-impact pub.	$0.9 - \alpha$	α	0.9
High-impact pub.	α	$0.1 - \alpha$	0.1
Total	0.9	0.1	1

Based on Table 3, it can be seen that producing n_{HI} high-impact publications on average yields $[(0.1 - \alpha) / 0.1] \times n_{HI}$ highly cited publications. Similarly, producing n_{LI} low-impact publications on average yields $[\alpha / 0.9] \times n_{LI}$ highly cited publications. It follows that obtaining a single highly cited publication on average requires $1 / [(0.1 - \alpha) / 0.1]$ high-impact publications or $1 / [\alpha / 0.9]$ low-impact publications. Clearly, the lower the value of α, the more the HCP index rewards the production of high-impact publications. Nevertheless, for any non-zero value of α, a researcher with a given number of high-impact publications can be systematically outperformed, in terms of highly cited publications, by a researcher with lots of low-impact publications. More precisely, a researcher who produces more than $[(0.1 - \alpha) / 0.1] / [\alpha / 0.9] \times n_{HI} = (0.9 - 9\alpha) / \alpha \times n_{HI}$ low-impact publications on average outperforms a colleague producing n_{HI} high-impact publications. Of course, if the value of α is close to zero, the number of

low-impact publications required to outperform a researcher with n_{HI} high-impact publications becomes very large, and in practice it may not be possible to have such a large publication output.

An improved counting approach

An obvious question is whether the HCP index can be modified in such a way that it no longer suffers from systematic errors in the assessment of researchers' overall scientific impact. In other words, is it possible to develop an improved way of counting highly cited publications?

One possibility might be to move from a size-dependent HCP index to a size-independent one. In that case, instead of calculating the *number* of highly cited publications of a researcher, one would calculate a researcher's *proportion* of highly cited publications. In some situations, this would indeed lead to improved results. For instance, consider the scenario illustrated in Table 1, and take the situation of researchers A and B, as discussed in the previous section (see Table 2). Researcher A has produced 100 high-impact publications, of which 30 are highly cited. Researcher B has produced 500 low-impact publications, of which 39 are highly cited. As we have seen, when looking at a researcher's number of highly cited publications, researcher B outperforms researcher A, even though researcher B's scientific impact is negligible compared with researcher A's. Now suppose we look at the proportion of highly cited publications of a researcher, that is, a researcher's number of highly cited publications divided by his total number of publications. Researcher A has 30 / 100 = 30% highly cited publications, while researcher B has only 39 / 500 = 7.8% highly cited publications. Hence, when looking at a researcher's proportion of highly cited publications, researchers A and B are ranked correctly with respect to each other.

Unfortunately, a size-independent HCP index also has problems. To demonstrate this, we introduce a third researcher, researcher C. Suppose researcher C has produced 200 high-impact publications and 50 low-impact ones (see Table 2). In line with the percentages reported in Table 1, this has resulted in (3% / 10%) × 200 + (7% / 90%) × 50 ≈ 64 highly cited publications. Since researcher C has produced twice as many high-impact publications as researcher A, researcher C's scientific impact is also twice as large as researcher A's. However, researcher A has 30% highly cited publications, while researcher C has only 64 / (200 + 50) = 25.6% highly cited publications. Hence, according to a size-independent HCP index, researcher A outperforms researcher C. It is clear that this is an incorrect assessment of the scientific impact of the two researchers.

From the point of view of assessing researchers' overall scientific impact, the fundamental problem of a size-independent HCP index is that productivity is not rewarded. If two researchers have the same proportion of highly cited publications, their scientific impact is assessed to be the same as well. This makes no sense if one researcher for instance has a publication output twice as large as another researcher. Other things being equal, the overall scientific impact of a

researcher should be assessed proportionally to his publication output.[49] If one researcher has both twice as many highly cited and twice as many lowly cited publications as another researcher, then the scientific impact of the former researcher should be assessed to be twice as large as the scientific impact of the latter researcher. A size-independent HCP index fails to take such productivity considerations into account.

There turns out to be a better way in which the HCP index can be modified to make sure that it provides proper assessments of researchers' scientific impact. The HCP index can be seen as a weighted sum of the publications of a researcher, where a highly cited publication has a weight of one while a lowly cited publication has a weight of zero. We now show that the weights used in the HCP index can be modified in such a way that on average the HCP value of a researcher is exactly equal to the number of high-impact publications the researcher has produced.

Our starting point is the general scenario shown in Table 3, with the parameter α ($0 \leq \alpha \leq 0.09$) determining the degree to which scientific impact and citations are correlated. We propose to weight highly cited publications by

$$w_{\mathrm{HC}} = \frac{0.1\alpha - 0.09}{\alpha - 0.09} \tag{1}$$

and lowly cited publications by

$$w_{\mathrm{LC}} = \frac{0.1\alpha}{\alpha - 0.09}. \tag{2}$$

Hence, the HCP value of a researcher is given by

$$\mathrm{HCP} = n_{\mathrm{LC}} w_{\mathrm{LC}} + n_{\mathrm{HC}} w_{\mathrm{HC}}, \tag{3}$$

where n_{LC} and n_{HC} denote the number of lowly and highly cited publications of the researcher. Notice that setting α equal to zero yields $w_{\mathrm{HC}} = 1$ and $w_{\mathrm{LC}} = 0$, which means that (3) reduces to the standard HCP index discussed in the previous section. Notice also that w_{HC} and w_{LC} are not defined if α is set equal to 0.09. As we have seen in the previous section, if α is set equal to 0.09, the number of citations of a publication does not provide any indication of the scientific impact of the publication.

Suppose a researcher has produced n_{HI} high-impact publications and n_{LI} low-impact publications. The expected HCP value of the researcher calculated using

[49] In practice, other things need not always be equal. For instance, one researcher may have more research time than another. For the purpose of our analysis, however, we assume researchers to find themselves in comparable situations.

(1), (2), and (3) then equals n_{HI}. This can be seen as follows. Based on Table 3, we obtain

$$\mathrm{E}(n_{HC}) = \frac{0.1-\alpha}{0.1} n_{HI} + \frac{\alpha}{0.9} n_{LI} \tag{4}$$

and

$$\mathrm{E}(n_{LC}) = \frac{\alpha}{0.1} n_{HI} + \frac{0.9-\alpha}{0.9} n_{LI}, \tag{5}$$

where E(•) denotes the expected value operator. It follows from (3) that

$$\mathrm{E}(\mathrm{HCP}) = \mathrm{E}(n_{LC}) w_{LC} + \mathrm{E}(n_{HC}) w_{HC}. \tag{6}$$

Substitution of (1), (2), (4), and (5) into (6) results in

$$\mathrm{E}(\mathrm{HCP}) = n_{HI}. \tag{7}$$

This proves that on average the HCP value of a researcher calculated using (1), (2), and (3) is exactly equal to the number of high-impact publications the researcher has produced. Unlike the standard HCP index, our modified HCP index therefore does not suffer from systematic errors in the assessment of researchers' scientific impact.

To understand the mechanism of our modified HCP index, it is important to see that w_{LC} in (2) is always negative (except if α is set equal to zero). Hence, lowly cited publications are given a negative weight in our modified HCP index. Other things equal, the more lowly cited publications one has, the lower one's HCP value. Why do we give a negative weight to lowly cited publications? Given our assumption that the scientific impact of low-impact publications is negligible, we want the contribution of a low-impact publication to a researcher's HCP value to be zero on average. However, due to random factors influencing the number of citations of a publication, some low-impact publications end up being highly cited, and these publications make a positive contribution to a researcher's HCP value. To compensate for this, we give a negative weight to lowly cited publications. This negative weight is chosen in such a way that on average the contribution of a low-impact publication to a researcher's HCP value is zero. For a high-impact publication, we want the contribution to a researcher's HCP value to be one on average. Using the weights in (1) and (2), we accomplish both of our objectives: Low-impact publications make an average contribution of zero, and high-impact publications on average contribute one.

Finally, there is an interesting property of our modified HCP index that we want to demonstrate. We again consider the scenario illustrated in Table 1. Let us introduce a new researcher, researcher D. Suppose this researcher has produced 70 high-impact publications and 270 low-impact ones (see Table 2). In this way, he has obtained the expected number of (3% / 10%) × 70 + (7% / 90%) × 270 = 42 highly cited publications. His remaining 70 + 270 – 42 = 298 publications are lowly cited. Setting α equal to 0.07 in (1) and (2), we obtain $w_{HC} = 4.15$ and $w_{LC} = -0.35$. Using (3), we then find that the HCP value of researcher D equals 298 × (–0.35) + 42 × 4.15 = 70. Hence, as expected, researcher D's HCP value equals his number of high-impact publications. A similar calculation can be made for researcher A introduced earlier (see Table 2). Recall that this researcher has produced 100 high-impact publications, which has resulted in 30 highly cited publications and 70 lowly cited ones. Based on his number of highly and lowly cited publications, we obtain a HCP value of 100 for researcher A, which is exactly the number of high-impact publications this researcher has produced. Comparing researchers A and D, our modified HCP index correctly identifies researcher A as the one with the larger scientific impact.

What is interesting in the comparison of researchers A and D is that researcher A is outperformed by researcher D in terms of both highly cited publications (30 vs. 42) and lowly cited publications (70 vs. 298). Intuitively, this may seem sufficient evidence to conclude that researcher D must have a larger scientific impact than researcher A. However, as we have seen, researcher A is the one with the larger scientific impact. Hence, based on simple more-is-better logic, one would easily draw an incorrect conclusion in the comparison of researchers A and D. By deviating from the more-is-better logic, our modified HCP index reaches the correct conclusion.

Discussion and conclusion

The more-is-better principle plays a central role in evaluative bibliometrics. In this paper, we have given examples of situations in which more need not always be better. When the overall scientific impact of researchers is determined by their number of high-impact publications, having more highly cited publications need not always coincide with having a larger scientific impact. This is caused by random factors that may influence the number of citations of a publication. The stronger these random factors, the more difficult it becomes to maintain the more-is-better principle. Importantly, the deviations from the more-is-better principle that we have studied are of a systematic nature. They do not simply result from incidental statistical fluctuations. This shows that, contrary to what sometimes seems to be claimed (e.g., Van Raan, 1998), random effects on citations need not cancel out. Instead, random effects may have systematic consequences, at least when using certain types of bibliometric indices.

The model that we have analyzed in this paper is extremely stylized. On the one hand this makes the model easy to study, but on the other hand it also means that the model has significant weaknesses. The most important weakness may be that

the scientific impact of a publication is assumed to be a binary variable: A publication either does or does not have scientific impact. Although this is of course a highly unrealistic assumption, it does match well with the idea of counting highly cited publications, which also relies on a binary distinction, albeit based on citations rather than impact.[50] Future work could focus on constructing more detailed models of the relationship between scientific impact and citations to find out under what types of conditions our findings do or do not remain valid.

We emphasize that we consider the modified HCP index introduced in this paper to be mainly of theoretical interest. To obtain appropriate weights for lowly and highly cited publications, one would need to have a realistic value for the parameter α. It is not evident how such a value could be determined empirically. Moreover, our modified HCP index is completely based on our very simple model of the relationship between scientific impact and citations. This makes the index vulnerable to the weaknesses of this model.

Nevertheless, we do believe that our modified HCP index provides interesting insights. The index illustrates how random effects on the number of citations of a publication can be corrected for while staying within the framework of simple additive indices with their many attractive properties (Marchant, 2009; Ravallion & Wagstaff, 2011). In addition, our modified HCP index introduces the idea of giving a negative weight to certain publications, not because these publications have a 'negative impact', but simply as a kind of correction factor to ensure that the index on average produces correct results. We emphasize that the insights we have obtained for HCP indices may be applicable to other bibliometric indices as well.

We hope that this paper will stimulate more research into the development of bibliometric indices within a model-based framework, in particular within a framework in which the relationship between citations on the one hand and concepts such as scientific impact and scientific quality on the other hand is made explicit (see also Ravallion & Wagstaff, 2011).

References

Bornmann, L. (2013). A better alternative to the *h* index. *Journal of Informetrics*, *7*(1), 100.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.

Dieks, D., & Chang, H. (1976). Differences in impact of scientific publications: Some indices derived from a citation analysis. *Social Studies of Science*, *6*(2), 247–267.

[50] By assuming a binary concept of scientific impact, our model serves as a kind of ideal world for the HCP index. In a model with a continuous concept of impact, it would be fundamentally impossible for the HCP index to provide perfect measurements of impact. In a model with a binary concept of impact, it is theoretically possible for the HCP index to provide perfect measurements of impact, as we have shown in this paper.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.
Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, *62*(7), 1370–1381.
Marchant, T. (2009). Score-based bibliometric rankings of authors. *Journal of the American Society for Information Science and Technology*, *60*(6), 1132–1137.
Martin, B.R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*(2), 61–90.
Moed, H.F. (2005). *Citation analysis in research evaluation*. Springer.
Moravcsik, M.J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*(1), 86–92.
Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, *41*, 609–641.
Plomp, R. (1990). The significance of the number of highly cited papers as an indicator of scientific prolificacy. *Scientometrics*, *19*(3–4), 185–197.
Plomp, R. (1994). The highly cited papers of professors as an indicator of a research group's scientific performance. *Scientometrics*, *29*(3), 377–393.
Price, D. de S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306.
Ravallion, M., & Wagstaff, A. (2011). On measuring scholarly influence by citations. *Scientometrics*, *88*(1), 321–337.
Simkin, M.V., & Roychowdhury, V.P. (2003). Read before you cite! *Complex Systems*, *14*(3), 269–274.
Simkin, M.V., & Roychowdhury, V.P. (2005). Stochastic modeling of citation slips. *Scientometrics*, *62*(3), 367–384.
Van Raan, A.F.J. (1998). In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics*, *43*(1), 129–139.
Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, *64*(2), 372–379.
Waltman, L., & Van Eck, N.J. (2012). The inconsistency of the *h*-index. *Journal of the American Society for Information Science and Technology*, *63*(2), 406–415.

COVERAGE AND ADOPTION OF ALTMETRICS SOURCES IN THE BIBLIOMETRIC COMMUNITY

Stefanie Haustein[1], Isabella Peters[2], Judit Bar-Ilan[3], Jason Priem[4], Hadas Shema[3],and Jens Terliesner[2]

[1] *stefanie.haustein@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal (Canada) and Science-Metrix, 1335 Avenue Mont-Royal Est, Montréal, H2J 1Y6 (Canada)

[2] *isabella.peters@hhu.de; jens.terliesner@hhu.de*
Department of Information Science, Heinrich-Heine-University, Universitätsstr. 1, Düsseldorf, 40225 (Germany)

[3] *Judit.Bar-Ilan@.biu.ac.il; dassysh@gmail.com*
Department of Information Science, Bar-Ilan University, Ramat-Gan, 52900 (Israel)

[4] *priem@email.unc.edu*
School of Information & Library Science, University of North Carolina at Chapel Hill, 216 Lenoir Drive, CB #3360100 Manning Hall, Chapel Hill (USA)

Abstract
Altmetrics, indices based on social media platforms and tools, have recently emerged as alternative means of measuring scholarly impact. Such indices assume that scholars in fact populate online social environments, and interact with scholarly products there. We tested this assumption by examining the use and coverage of social media environments amongst a sample of bibliometricians. As expected, coverage varied: 82% of articles published by sampled bibliometricians were included in Mendeley libraries, while only 28% were included in CiteULike. Mendeley bookmarking was moderately correlated (.45) with Scopus citation. Over half of respondents asserted that social media tools were affecting their professional lives, although uptake of online tools varied widely. 68% of those surveyed had LinkedIn accounts, while Academia.edu, Mendeley, and ResearchGate each claimed a fifth of respondents. Nearly half of those responding had Twitter accounts, which they used both personally and professionally. Surveyed bibliometricians had mixed opinions on altmetrics' potential; 72% valued download counts, while a third saw potential in tracking articles' influence in blogs, Wikipedia, reference managers, and social media. Altogether, these findings suggest that some online tools are seeing substantial use by bibliometricians, and that they present a potentially valuable source of impact data.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2); Webometrics (Topic 7)

Introduction

Altmetrics, indices based on activity in social media environments, have recently emerged as alternative means of measuring scholarly impact (Priem, 2010; Priem et al., 2010). The idea of impact measuring which moves beyond citation analysis, however, emerged long before the advent of social media (Martin & Irvine, 1983; Cronin & Overfelt, 1994). One of the underlying problems with citation analysis as basis for evaluating scientific impact is that citations paint a limited picture of impact (Haustein, in press). On the one hand, researchers often fail to cite all influences (MacRoberts & MacRoberts, 1989). On the other hand, the total readership population includes not only authors but also "pure," i.e. non-publishing, readers, who are estimated to constitute one third of the scientific community (Price & Gürsey, 1976; Tenopir & King, 2000). Publications are used in the development of new technologies, applied in daily work of professionals, support teaching, and have other societal effects (Schlögl & Stock, 2004; Rowlands & Nicholas, 2007; Research Councils UK, 2011; Thelwall, 2012).
Thus, a better way of approaching scholarly impact is to consider citation as just one in a broader spectrum of possible uses. Webometrics and electronic readership studies gathered impact and usage data in a broader sense, but have been restricted by scalability problems and access to data (Thelwall, Vaughan, & Björneborn, 2005; Thelwall, 2010). As altmetrics are based on clearly defined social media platforms, that often provide free access to usage data through Web APIs, data collection is less problematic, although accuracy is still a problem (Priem, in press). With these new sources comes the possibility of analyzing online usage of scholarly resources independently of publishers. Tracking the use of scholarly content in social media means that researchers are able to analyze impact more broadly (Li, Thelwall, & Guistini, 2012; Piwowar, 2013). Moreover, many online tools and environments surface evidence of impact relatively early in the research cycle, exposing essential but traditionally invisible precursors like reading, bookmarking, saving, annotating, discussing, and recommending articles.

In order to explore the potential of altmetrics, this work studies the applicability and use of altmetrics sources and indicators in the bibliometric community. As it is still unclear how broadly these platforms are used, by whom and for what purposes, this study aims to evaluate the representativeness and validity of altmetrics indicators using the bibliometric community and literature as an initial reference set. We focus on measuring the impact of conventional peer-reviewed publications, such as journal articles and proceedings papers, on the social web as well as how bibliometricians perceive and use social media tools in their daily work routine. New forms of output, such as research results published in blogs, comments and tweets, are not addressed in this paper.

We apply a two-sided approach, aiming to answer the following sets of research questions:

- RQ 1: To what extent are bibliometrics papers present on social media platforms? How comprehensive is the coverage of the literature on platforms like Mendeley and CiteULike? How many users do they have and how many times are they used?
- RQ 2: To what extent is the bibliometric community present on social media platforms? Who uses these platforms and for what purposes?

We answered the first set of questions by evaluating the coverage and intensity of use of bibliometrics literature in social reference managers. Publications by presenters of the 2010 STI conference served as a reference set, as they represent a group of both established and new bibliometricians. The second set of research questions was approached by surveying the attendees of the 2012 STI conference in Montréal regarding their use of social media.

Altmetrics Literature Review

Altmetrics research to date has focused on exploring potential data sources, correlating alternative impact data with citations and analyzing it from a content perspective; for overviews of this research see Bar-Ilan, Shema, and Thelwall (in press), Haustein (in press), and Priem (in press). When it comes to monitoring the impact of scholarly publications, Mendeley (mendeley.com) and CiteULike (citeulike.org) have proven particularly useful. They combine social bookmarking and reference management functionalities and allow users to save literature, share them with other users, and add keywords and comments (Henning & Reichelt, 2008; Reher & Haustein, 2010). Both social bookmarking systems use a bag model for resources, meaning that a particular resource can be simultaneously saved or bookmarked by several users. This functionality allows for counting resource-specific bookmarking actions like how many users saved a particular resource. According to self-reported numbers, Mendeley is considerably larger than CiteULike (CuL). During data collection in March 2012, CuL claimed to have 5.9 million unique papers in CuL vs. more than 34 million in Mendeley (Bar-Ilan, Haustein, Peters, Priem, Shema, & Terliesner, 2012). As of August 2012, Mendeley claims to be the largest research catalog with 280 million bookmarks to 68 million unique documents uploaded by 1.8 million users (Ganegan, 2012). In November 2012 Mendeley reached 2 million users (Mendeley, 2012).

Case studies focusing on the coverage of social reference managers support Mendeley's position as a leader in the field. Li, Thelwall, and Giuistini (2012) investigated how bookmarks in Mendeley and CuL reflect papers' scholarly impact and found that 92% of sampled Nature and Science articles had been bookmarked by at least one Mendeley user, and 60% by one or more CuL users. Bar-Ilan (2012a; 2012b) found 97% coverage of recent JASIST articles in Mendeley. Priem, Piwowar, and Hemminger (2012) showed that the coverage of articles published in the PLoS journals was 80% in Mendeley and 31% in CuL. Li

and Thelwall (2012) sampled 1,397 F1000 Genomics and Genetics papers and found that 1,389 of those had Mendeley users.

Studies have found moderate correlation between bookmarks and Web of Science (WoS) citations. Li, Thelwall, and Giustini (2012) reported r=.55 of Mendeley and r=.34 of CuL readers with WoS citations, respectively. Weller and Peters (2012) arrived at slightly higher correlation values for a different article set between Mendeley, CuL, BibSonomy, and Scopus. Bar-Ilan (2012a; 2012b) found a correlation of .46 between Mendeley readership counts and WoS citations for the JASIST articles. Li and Thelwall (2012) found high correlation (.69) between Mendeley and WoS for the articles recommended on F1000. User-citation correlations for the Nature and Science publications were .56 (Li, Thelwall, & Guistini, 2012) and Priem, Piwowar, and Hemminger (2012) found a correlation of .5 between WoS citations and Mendeley users for the PLoS publications.

While bookmarks in reference managers reflect readership of scholarly articles, Twitter activity reflects discussion around these articles. Several studies have analyzed tweets "citing" scholarly publications. Priem and Costello (2010) and Priem, Costello, and Dzuba (2011) found that scholars use Twitter as a professional medium for sharing and discussing articles, while Eysenbach (2011) showed that highly-tweeted articles were 11 times more likely become highly-cited later. Weller and Puschmann (2011), and Letierce, Passant, Decker, and Breslin (2010) analyzed the use of Twitter during scientific conferences and revealed that there was discipline-specific tweeting behavior regarding topic and number of tweets as well as references to different document types (i.e., blogs, journal articles, presentation slides). Along with Twitter, other studies have examined citation from Wikipedia articles (Nielsen, 2007) and blogs (Groth & Gurney, 2010; Shema, Bar-Ilan, & Thelwall, 2012) as potential sources reflecting alternative impact of scholarly documents.

Apart from aforementioned studies, which focused on quantitative analysis of social media impact, there is a more content-oriented research approach which particularly examines tags attached to products of scholarly practice. Bar-Ilan (2011) studied the items tagged with "bibliometrics" on Mendeley and CuL, whereas Haustein and Peters (2012) and Haustein et al. (2010) showed that tags represent a reader-specific view on articles' content which could be used to analyze journal content from a readers perspective (as opposed to the author and indexer perspectives).

Although altmetric indicators and data sources are increasingly applied in evaluation studies, little is yet known about the users of such social media platforms or how researchers integrate them into their research environment (Mahrt, Weller, & Peters, in press). Understanding who is using social media tools for which purpose is, however, crucial to the application of altmetrics for

evaluation purposes. Given that a representative share of documents are covered by social media tools and the user community can be identified, social media platforms can be valuable sources for measuring research impact from the readers' point of view, functioning as supplements to citation analysis. In contrast to citations, altmetrics potentially cover the whole readership and are available in real time.

RQ 1: Coverage of Bibliometrics Papers on Altmetrics Platforms

Before analyzing the alternative impact of bibliometrics literature and authors from the bibliometric community, it is necessary to explore which sources are suitable and provide the best coverage. Comparing them to traditional sources of impact evaluation provides information about the differences between use in citation and use in other contexts.

Method

In order to create a list of bibliometrics publications, all documents authored by presenters of the 2010 STI conference in Leiden were collected on WoS and Scopus. We chose this author-based, bottom-up approach to facilitate linking altmetrics data to authors as well as just documents. The group of presenters at the STI conference was considered to represent a core group of both established and new members of the current bibliometric community. The presenters' names were retrieved from the conference program. The final list contained 57 researchers, who together had authored 1,136 papers[51] covered in Scopus. Mendeley publication and readership information was retrieved manually via the Mendeley Web search interface from mendeley.com. At the time of data collection in March 2012 the manual approach proved more comprehensive, as the API, searched via the ImpactStory tool[52], only returned one of multiple entries matching the search criteria. More recent searches seem to indicate this problem has since been resolved. In CuL, publications can be searched by DOI. However, it should be noted that bibliographic data in CuL or Mendeley is incomplete (Haustein & Siebenlist, 2011). The number of articles bookmarked in CuL might thus be higher than the number retrieved via DOI. The manual search in Mendeley showed that 33% of the documents retrieved did not contain a DOI.

Results

As shown in Table 1, the coverage of the 1,136 bibliometrics documents in Mendeley was good: 928 (82%) of the documents had at least one Mendeley bookmark, while only 319 (28%) of articles were in CuL. Although coverage in

[51] Some presenters were omitted either because they had not published in sources covered by Scopus or WoS or due to ambiguous names, for which relevant papers could not be identified. Documents without a DOI were not considered as it was needed to identify papers on the altmetrics platforms. For a more detailed description of data collection see Bar-Ilan et al. (2012).
[52] http://impactstory.org

CuL may be underestimated because bookmarks without a correct DOI were not retrieved, this confirms the results found by other studies (e.g., Li, Thelwall, & Guistini, 2012; Priem, Piwowar, & Hemminger, 2012). Unsurprisingly given Mendeley's very recent founding, older articles are less bookmarked. Of the 85 sample articles published before 1990, only 44% have readers in Mendeley, while 88% of those published since 2000 have Mendeley bookmarks (see Figure 1). Mendeley's popularity is not only reflected in the coverage of documents but also by the average activity on bookmarked documents: in Mendeley each document was bookmarked by a mean of 9.5 users, compared to a usage rate of 2.4 in CuL. Correlations between Scopus citations and users counts were .45 for Mendeley and .23 for CuL. These moderate correlations confirm previous findings for other samples and suggest that altmetrics may indeed reflect impact not reflected in citation counts.

Table 1. Coverage and citation or usage rates of a sample of 1,136 bibliometrics documents. "Events" are either bookmarks or citations, depending on the database.

	Scopus	*Web of Science*	*Mendeley*	*CiteULike*
Number of indexed documents	1,136	957	928	319
Total event counts	18,755	17,858	8,847	777
Percent sampled with nonzero event counts (total)	85% (961)	74% (845)	82% (928)	28% (319)
Mean events per article with nonzero count	19.5	21.1	13.4	2.4

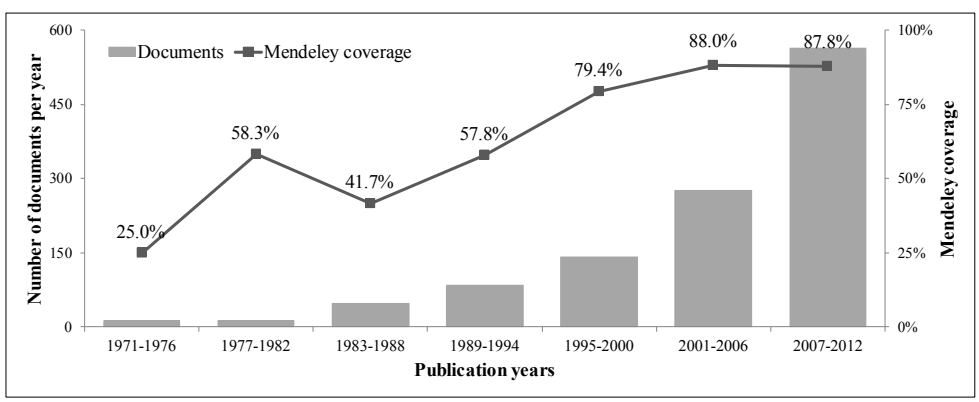


Figure 1. Coverage of sampled documents in Mendeley per publication year. Overall coverage is 82% (*n=1,136*).

RQ 2: Use of Altmetrics Platforms by the Bibliometric Community

Since the results of RQ 1 confirmed that reference managers (Mendeley in particular) were a rich source for usage data and impact measurements of bibliometrics publications, we wanted to study who generates this usage data. To do this, we surveyed a sample of the bibliometrics community to learn how,

when, and why they use various online environments; our goal was to better understand the significance of altmetrics indicators drawn from these environments.

Method

The paper and pencil survey was conducted among participants of the 17th International Conference on Science and Technology Indicators (STI) in Montréal. Participants filled out the survey during the conference from September 5th to 8th 2012. The survey contained open and closed questions; these mainly asked if and how members of the bibliometric community used social media with regards to organizing their literature and promoting their work, as well as how such tools influenced their professional lives. SPSS and Open Code were used for the analysis of the survey. All openly designed questions were coded using the Grounded Theory approach (Glaser & Strauss, 1967): codes were assigned to participants' statements, and these were then used to generate broader categories reflecting patterns of answering behavior.

Results

Of the 166 participants of the STI 2012 as indicated on the attendee list, 71 returned the questionnaire, resulting in a response rate of about 42.8%. Of the survey participants 63.4% were male and 33.8% were female, while 2.8% did not indicate their gender. Compared to the conference, females were somewhat overrepresented in our sample. While the youngest participant was 26 and the oldest 64, most respondents were between 31 and 40 years old. The mean age was 41.5 years. The respondents came from a mixed professional background, as 14.1% were research scientists and 14.1% worked in the R&D industry. 15.5% indicated that they had another background, 12.7% were doctoral candidates, 11.3% research managers, 8.5% government employees and 7.0% librarians. 4.2 % were associate professors/readers, 2.8% students, 2.8% postdocs, 2.8% assistant professors/lecturers and 2.8% full professors. One participant (1.4%) did not indicate his professional background.

Sixty people answered the question about reference management, 35 (58.3%) of whom use reference management software to organize scientific literature. The category "reference management software" includes desktop based software and web reference management services. A "personal solution" of literature management was described by 38.3% of respondents, which summarizes storing documents on personal drives on the desktop or on the Web as well as organizing literature on book shelves or in Word documents. Alerts from journals, bibliographic databases, or libraries fall in the category "information suppliers", which was described by 12 people (20.0%) as their way to find literature. Four people stated explicitly that they do not manage literature, because there is no need since they are not researchers.

When asked in a multiple choice question about whether they had heard of and used any of the social bookmarking services BibSonomy, CuL, Connotea, Delicious, or Mendeley, the latter was the most popular among respondents. Table 2 shows the percentage of the 70 respondents who knew and used the different bookmarking services and reference managers. Note that 77.1% of the respondents had heard about Mendeley, but only 25.7% actually used it. A similar percentage of the respondents had heard about CuL (72.9%), but only 12.9% of the respondents were actual users. The category "perceived usefulness" represents the percentage of a given platform's actual users compared to the number who have heard about it. By this measure, BibSonomy and CuL, were perceived to be relatively less useful; only 4.0% and 8.0% of those who knew the tools, respectively, actually use them. Mendeley was not only the most known tool, but also the one with the highest number of users. A third of all who had heard of the tool, used it, even though usage was rather occasional.

Table 2. Knowledge and usage of social bookmarking services and reference managers.

	BibSonomy	*Connotea*	*CiteULike*	*Delicious*	*Mendeley*
heard about the service *(n=70)*	35.7%	35.7%	72.9%	64.3%	77.1%
used the service *(n=70)*	1.4%	2.9%	12.9%	11.4%	25.7%
perceived usefulness	4.0% *(n=25)*	8.0% *(n=25)*	17.6% *(n=51)*	17.8% *(n=45)*	33.3% *(n=54)*

While there were more male than female users, the age structure of the Mendeley users corresponds to that of all participants. Both the youngest and the oldest respondent were Mendeley users. Although the numbers are too low to be representative, there seems a tendency towards a professional background in research of Mendeley users: the share of full professors, postdocs, doctoral candidates, and research scientists is higher among Mendeley users compared to the overall percentage of participants, while the percentage of research managers and members of R&D industry is lower. Thirteen of the 18 people who used Mendeley indicated for which purposes they used the tool. Managing references and connecting with people were equally important reasons to use Mendeley. This emphasizes that Mendeley connects literature management with the social aspect of connecting people who are interested in the same contents whereas CuL is mostly used for literature search.

The survey showed that Facebook, LinkedIn, Twitter, and Google+ were the most popular social networks. Figure 2 summarizes how many survey participants used the different social media tools. 52 people (73.2%) had a profile on Facebook, 48 (67.6%) on LinkedIn, 31 (43.7%) on Twitter, and 28 (39.4%) on Google+. Xing was used from 9.9% of users and 7.0% used MySpace. Among the tools focusing

on the research community, Mendeley (23.9%), Academia.edu (21.1%), and ResearchGate (21.1%) have almost the same number of users in our sample, i.e. about one fifth of the participants had a profile on each of these platforms.

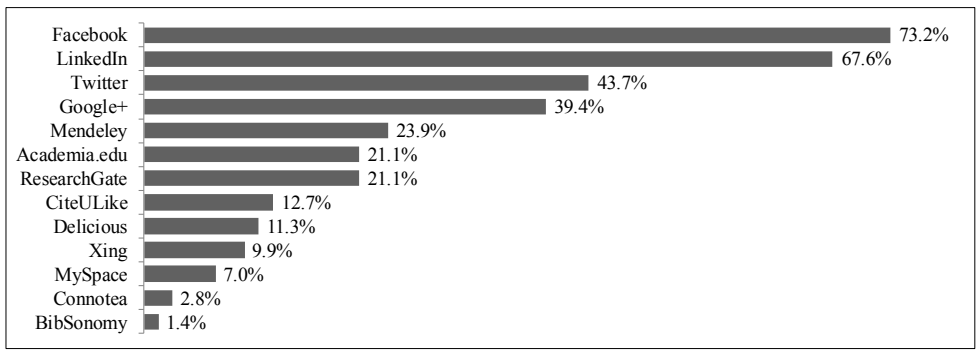


Figure 2. Percentage of participants having a profile on or using social media tools mentioned in the survey (*n=71*).

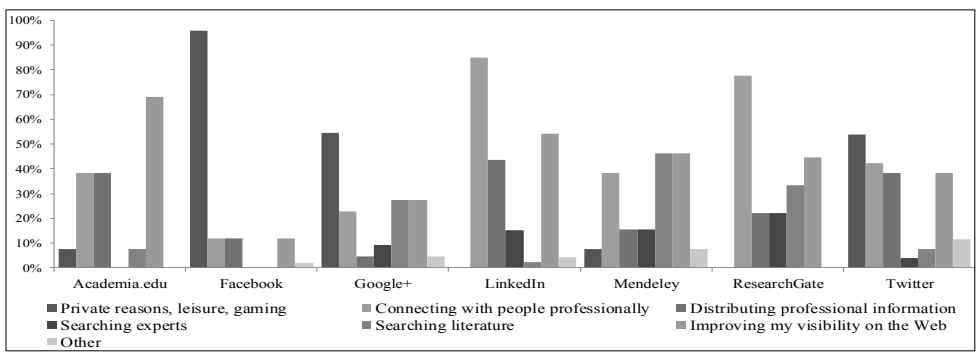


Figure 3. What are participants using particular social networks for? Question allowed for multiple answers (Academia.edu: *n=13*; Facebook: *n=50*; Google+: *n=22*; LinkedIn: *n=46*; Mendeley: *n=13*; ResearchGate: *n=9*; Twitter: *n=26*. *MySpace (n=4) and Xing (n=5) are not shown)*.

Asking participants for the purpose of using these nine social networking platforms shows that Facebook, Google+, and MySpace are above all used for private purposes, while LinkedIn, ResearchGate, and Xing fulfill the main purpose of connecting with the professional community. LinkedIn is by far the most popular tool to connect with professional contacts; 84.8% indicated that this was the reason to use that platform. They also used LinkedIn to improve their own visibility (54.3%) and distribute professional information (43.5%). Twitter and Facebook were mostly used for private reasons, but Twitter was also important to connect with people professionally, distributing professional information and improving one's visibility. Although the overall use of

Academia.edu was rather low (21.1% had a profile, but only 18.3% used it), 69.2% of the 13 Academia.edu users applied it to improve their visibility. Figure 3 shows the reasons for which respondents use social networks for each of the platforms.

Asked for personal publication profiles on Academia.edu, Google Scholar Citations, Mendeley, Microsoft Academic Search, ResearcherID (WoS), or ResearchGate, 32 participants listed their publications at least at one of these platforms. The most popular tool was Google Scholar Citations (22 respondents with profile; 68.8% of those with publication profiles), followed by ResearcherID (14: 43.8%), which can probably be attributed of the popularity and significance of Google and WoS. Google Scholar Citations (see Figure 4) was mostly used to check citations, WoS was used to check citations and add publications to the ResearcherID, while Academia.edu, Mendeley, and ResearchGate profiles were mostly used to add missing publications. In Microsoft Academic Search, people delete "wrong" publications from their profiles.

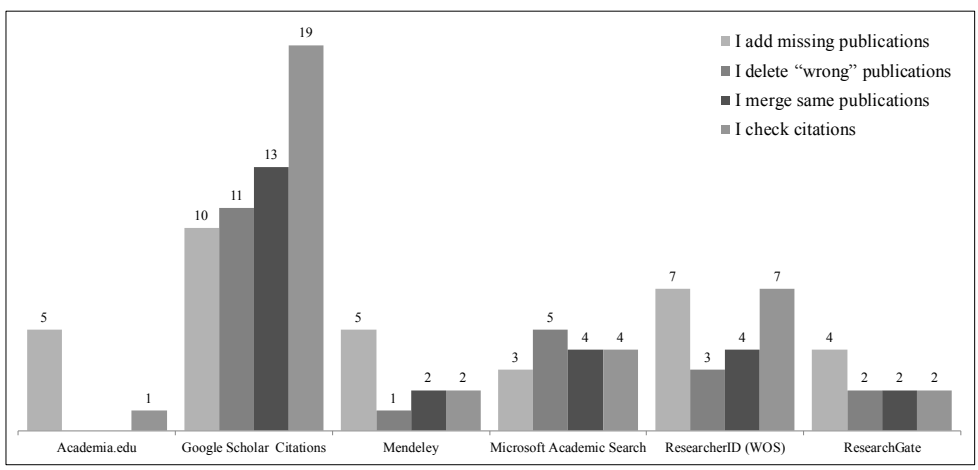


Figure 4. What are participants doing with their publication profile? Question allowed for multiple answers (Academia.edu: *n=5*; Google Scholar Citations: *n=22*; Mendeley: *n=8*; Microsoft Academic Search: *n=7*; Researcher ID (WoS): *n=14*; ResearchGate: *n=9*).

49.3% of the participants used some kind of repository to deposit their work. To 7 respondents the question did not apply, as they do not or no longer actively publish. Among those who used a repository, the most common was the institutional repository (57.1%), the second most popular was arXiv (21.4%). 47.9% of the respondents provided access to fulltexts on their homepages.

Although use of altmetrics platforms was quite low among survey participants, 85.9% thought that altmetrics had some potential in author or article evaluation.

The majority, (71.8%) believed that the number of article downloads or views could be of use in author or article evaluation (see Figure 5 and Kurtz & Bollen, 2010 for a review of usage bibliometrics). Other sources such as citations in blogs (38.0%), Wikipedia links or mentions (33.8%), bookmarks on reference managers (33.8%), and discussions on Web 2.0 platforms (31.0%) were believed to have potential as altmetrics indicators as well.

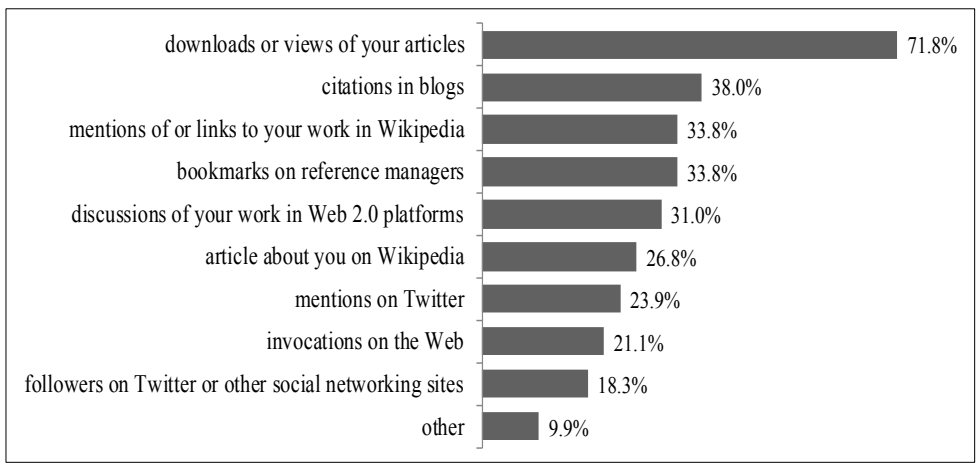


Figure 5. Which alternative metrics are believed to have potential for article or author evaluation? Question allowed for multiple answers (*n=71*).

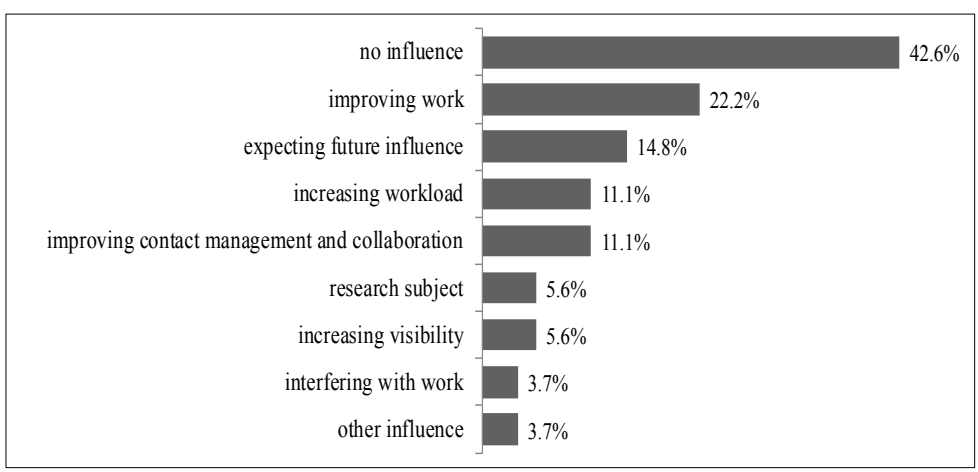


Figure 6. In what ways do social network and bookmarking systems affect your professional life and/or work flow? Openly designed question (*n=54*).

An openly designed question asked about in what ways social network and bookmarking systems affected their professional life and work flow (see Figure 6). Twenty-three (42.6%) of the 54 respondents said they were not at all

influenced by these tools and 8 (14.8%) were not yet influenced but expected some impact in the future. 22.2% of respondents answered that the tools improved their work in terms of finding new information, fast distribution of information, and organization of research material. Two of these stated that social networks and social bookmarking systems "made my life much easier". For 11.1% the tools improved contact management and collaboration and 5.6% felt like they improved their visibility. On the other hand, 11.1% stated that social media tools increased their workload and 3.7% said that it interfered with their daily work, i.e. causing procrastination and getting lost in discussions on social media sites while delaying work.

Conclusions and Outlook

This study has followed a two-sided approach to explore the representativeness and validity of social media platforms to be used as data sources for altmetrics indicators evaluating impact of scholarly documents. It has shown that bibliometrics literature is well represented on social media platforms (i.e., Mendeley), making them a valuable source for evaluating the influence of scholarly documents in a broader way than citation analysis. The coverage of the sampled documents was as high as 82% overall with an even higher coverage of recent documents. Although this age bias was expected, as Mendeley was only launched in 2009, this bias needs to be considered when evaluating older documents. Mendeley did not only dominate in terms of coverage, but had also a much greater number of readers per document than CuL.

Having analyzed how bibliometrics documents are used on social reference managers, the second part of the study aimed to find out who was generating this use. A survey distributed among the core of the bibliometric community present at the 2012 STI conference in Montréal asked for social media use and its influence on the working environment of participants. Over half of those surveyed asserted that social media tools were affecting their professional lives, or that they were expecting future influence. Actual uptake of the platforms varied. Two-thirds of survey participants had LinkedIn accounts, which they used to connect professionally, while social networks with a scholarly focus such as Academia.edu, Mendeley, and ResearchGate were each used by only a fifth of respondents. Nearly half of those responding had Twitter accounts, which is extremely high compared to findings by Priem, Costello, and Dzuba (2011) and Ponte and Simon (2011), who found a Twitter usage rate of 2.5% and 18% among scholars, respectively; this may be due to growth in Twitter use, disproportionate use by bibliometricians, or the different methodologies employed.

Although Mendeley was the most popular social reference manager among the 71 participants, only one third surveyed use the tool, and and their use was rather sporadic. This is surprising given the high coverage of bibliometrics articles in Mendeley; it is unclear who is generating the high reader counts observed. A

survey targeted directly at Mendeley users could clarify whether groups not at the conference (for example, people from other disciplines, or students, or practitioners) are using Mendeley heavily. The surveyed conference participants may also not properly represent the typical social media users and therefore reflect a biased picture of actual usage, although this assumption has to be proven in detailed studies. When altmetrics is broadly defined to include download data, 85% of bibliometricians surveyed expect at least one altmetrics indicator to become influential in future research evaluation. Around a third of respondents expected such influence from altmetrics based on blogs, Wikipedia, reference managers, and social media. Thus, although their use of social media tools remains modest as yet, survey participants are increasingly aware of the potential of altmetric indicators to supplement traditional evaluation indicators.

This study is limited by the specificity of its sample, and by potential non-response bias (enthusiastic users of social media may have been more likely to complete the survey). Results are thus not generalizable. Hence, further research should include the systematic analysis of all scholarly disciplines using this two-sided approach. Thus it would be possible to define the extent to which social media platforms cover a discipline's publication output as well as determine who is generating the use and for what purpose. This will help to validate altmetrics indicators as supplements to traditional metrics in research evaluation.

References

Bar-Ilan, J. (2011). Articles tagged by 'bibliometrics' on Mendeley and CiteULike. Paper presented at the *Metrics 2011 Symposium on Informetric and Scientometric Research*.

Bar-Ilan, J. (2012a). JASIST@mendeley. Presented at the *ACM Web Science Conference Workshop on Altmetrics. Evanston, IL*. Retrieved January 21, 2013 from http://altmetrics.org/altmetrics12/bar-ilan

Bar-Ilan, J. (2012b). JASIST 2001-2010. *Bulletin of the American Society for Information Science and Technology*, *38*(6), 24-28.

Bar-Ilan, J., Shema, H., & Thelwall (in press). Bibliographic References in Web 2.0. In B. Cronin, & C. Sugimoto (eds.), *Bibliometrics and Beyond: Metrics-Based Evaluation of Scholarly Research*, Cambridge: MIT Press.

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H., & Terliesner, J. (2012). Beyond citations: Scholars' visibility on the social Web. In *Proceedings of the 17th International Conference on Science and Technology Indicators, Montréal, Canada* (Vol. 1, pp. 98–109).

Cronin, B., & Overfelt, K. (1994). The scholar's courtesy: A survey of acknowledgement behaviour. *Journal of Documentation*, *50*, 165-196.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, *13*(4). Retrieved January 21, 2013 from http://www.jmir.org/2011/4/e123

Ganegan, F. (2012, August). *Filtering the research record and farming big data.* Retrieved January 21, 2013 from http://www.swets.com/blog/filtering-the-research-record-and-farming-big-data#

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory. Strategies for Qualitative Research*. New Brunswick, London: Aldine Transactions.

Groth, P., & Gurney, T. (2010). Studying scientific discourse on the Web using bibliometrics: A chemistry blogging case study. Presented at the *WebSci10: Extending the Frontiers of Society On-Line, Raleigh, NC, USA*.

Haustein, S. (in press). Readership Metrics. In B. Cronin, & C. Sugimoto (eds.), *Bibliometrics and Beyond: Metrics-Based Evaluation of Scholarly Research*, Cambridge: MIT Press.

Haustein, S., & Peters, I. (2012). Using Social Bookmarks and Tags as Alternative Indicators of Journal Content Description. *First Monday*, *17*(11). Retrieved January 21, 2013 from www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/4110/3357

Haustein, S., & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, *5*(3), 446–457.

Haustein, S., Golov, E., Luckanus, K., Reher, S., & Terliesner, J. (2010). Journal evaluation and science 2.0. Using social bookmarks to analyze reader perception. In *Book of Abstracts of the 11th International Conference on Science and Technology Indicators, Leiden, The Netherlands* (pp. 117-119).

Henning, V., & Reichelt, J. (2008). Mendeley: A Last.fm for research? In *Proceedings of 4th IEEE International Conference on Escience, Indianapolis, IN, USA* (pp. 327-328).

Kurtz, M. J. & Bollen, J. (2010). Usage bibliometrics. *Annual Review of Information Science and Technology*, 44, 1-64.

Letierce, J., Passant, A., Decker, S., & Breslin, J.G. (2010). Understanding how Twitter is used to spread scientific messages. In *Proceedings of the Web Science Conference, Raleigh, NC, USA*.

Li, X., & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. In *Proceedings of the 17th International Conference on Science and Technology Indicators, Montréal, Canada* (Vol. 2, pp. 451-551).

Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics*, *91*(2), 461-471.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis – A critical review. *Journal of the American Society for Information Science*, *40*(5), 342-349.

Mahrt, M., Weller, K., & Peters, I. (in press). Twitter in Scholarly Communication. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (eds.), *Twitter and Society*. New York, NY: Peter Lang.

Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*(2), 61-90.

Mendeley (2012). *Mendeley Global Research Report*. Retrieved January 21, 2013 from http://www.mendeley.com/global-research-report/#.UPxyUqPi58E

Nielsen, F. (2007). Scientific citations in Wikipedia. *First Monday*, *12*(8). Retrieved January 21, 2013 from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1997/1872

Piwowar, H. (2013). Value all research products. *Nature*, *493*, 159.

Ponte, D., & Simon, J. (2011). Scholarly communication 2.0: Exploring researchers' opinions on Web 2.0 for scientific knowledge creation, evaluation and dissemination. *Serials Review*, 37(3), 149–156.

Price, D. J. de Solla, & Gürsey, S. (1976). Studies in Scientometrics I. Transience and continuance in scientific authorship. *International Forum on Information and Documentation*, *1*(2), 17-24.

Priem, J. (in press). Altmetrics. In B. Cronin, & C. Sugimoto (eds.), *Bibliometrics and Beyond: Metrics-Based Evaluation of Scholarly Research*, Cambridge: MIT Press.

Priem, J. (2010). Tweet by Jason Priem on September 28, 2010. Retrieved January 21, 2013 from https://twitter.com/#!/jasonpriem/status/25844968813

Priem, J., & Costello, K. (2010). How and why scholars cite on Twitter. In *Proceedings of the 73rd Annual Meeting of the American Society for Information Science and Technology, Pittsburgh, PA, USA*. doi: 10.1002/meet.14504701201/full

Priem, J., Costello, K., & Dzuba, T. (2011). First-year graduate students just wasting time? Prevalence and use of Twitter among scholars. Presented at the *Metrics 2011 Symposium on Informetric and Scientometric Research, New Orleans, LA, USA*. Retrieved January 21, 2013 from http://jasonpriem.org/self-archived/5uni-poster.png

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). *Altmetrics in the wild: Using social media to explore scholarly impact*. Retrieved January 21, 2013 from http://arxiv.org/abs/1203.4745

Priem, J., Taraborelli, D., Groth, P., & Nylon, C. (2010). *alt-metrics: a manifesto*. Retrieved January 21, 2013 from http://altmetrics.org/manifesto

Reher, S., & Haustein, S. (2010). Social bookmarking in STM: Putting services to the acid test. *Online - Leading Magazine for Information Professionals*, *34*(6), 34-42.

Research Councils UK. (2011, March). *Types of impact*. Retrieved January 21, 2013 from http://www.rcuk.ac.uk/documents/impacts/TypologyofResearchImpacts.pdf

Rowlands, I., & Nicholas, D. (2007). The missing link: Journal usage metrics. *Aslib Proceedings*, *59*(3), 222-228.

Schlögl, C., & Stock, W. G. (2004). Impact and relevance of LIS journals: A scientometric analysis of international and German-language LIS journals – Citation analysis versus reader survey. *Journal of the American Society for Information Science and Technology*, *55*(13), 1155-1168.

Shema. H., Bar-Ilan. J., & Thelwall, M. (2012). Research Blogs and the Discussion of Scholarly Information. *PLoS ONE*, 7(5): e35869. doi:10.1371/journal.pone.0035869

Tenopir, C., & King, D. W. (2000). *Towards electronic journals: Realities for scientists, librarians, and publishers*. Washington, D.C.: Special Libraries Association.

Thelwall, M. (2010). Webometrics: emergent or doomed? *Information Research: An International Electronic Journal*, *15*(4). Retrieved January 21, 2013 from http://informationr.net/ir/15-4/colis713.html

Thelwall, M. (2012). Journal impact evaluation: A webometric perspective. *Scientometrics*, *92*, 429-441.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, *39*(1), 81-135.

Weller, K., & Peters, I. (2012). Citations in Web 2.0. In A. Tokar, M. Beurskens, S. Keuneke, M. Mahrt, I. Peters, C. Puschmann, et al. (eds.), *Science and the Internet* (pp. 211-224). Düsseldorf: Düsseldorf University Press.

Weller, K, & Puschmann, C. (2011). Twitter for scientific communication: How can citations/references be identified and measured? In *Proceedings of the 3rd ACM International Conference on Web Science, Koblenz, Germany*. Retrieved January 21, 2013 from http://journal.webscience.org/500/1/153_paper.pdf

CROWDSOURCING THE NAMES-GAME: A PROTOTYPE FOR NAME DISAMBIGUATION OF AUTHOR-INVENTORS (RIP)

Matthijs den Besten[53]

m.den-besten@supco-montpellier.fr
Groupe Sup de Co Montpellier Business School/Montpellier Research in Management, Montpellier

Catalina Martinez

catalina.martinez@csic.es
Institute of Public Goods and Policies (CSIC-IPP), Madrid

Nicolas Maissonneuve

n.maisonneuve@gmail.com
Independent researcher, Paris

Stéphane Maraut

stephane.maraut@gmail.com
IT expert, Madrid

Abstract
Crowdsourcing is a process for outsourcing micro-tasks to a distributed group of anonymous people, as in Amazon Mechanical Turk. The purpose of this paper is to present an exploration of the extant literature on crowdsourcing to identify best practices and describe the results of the implementation of a prototype that uses crowdsourcing to help with the name disambiguation of Spanish author-inventors. Our aim is to investigate whether and how the use of crowdsourcing for the Names-Game, as this activity is called, can help increase the efficiency and accuracy of human raters.

Conference topic:
Research Fronts and Emerging Issues (topic 4).
Also related to Technology and Innovation Including Patent Analysis (topic 5).

[53] Corresponding author: m.den-besten@supco-montpellier.fr

Introduction

Crowdsourcing is a relatively recent phenomenon, which holds a lot of promise for the scientific community. In particular, it might help address seemingly unconnected issues like the need to engage with citizens, the need to reduce costs, and the need to increase the reproducibility of research.

> "Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers."[54]

The purpose of this paper is to explore the potential of crowdsourcing to help with name disambiguation exercises. The Names-Game is the term coined by Trajtenberg *et al* (2006) to the different solutions used to address the 'Who is Who' problem in patent data, in particular to uniquely identify patent inventors based on their names, location and any other useful information. Based on a review recent research on name disambiguation in the context of patent and publications data, as well a review of experiences with Amazon Mechanical Turk (AMT), a platform for crowd-sourcing, we propose a design for an experiment with the Name Game on Mechanical Turk. Assuming that the outcome of the experiment will confirm the feasibility and appropriateness of the deployment of AMT for the Name Game, others will be able to build upon our work and extent crowdsourcing to other activities concerning data imputation.

Background: The Names-Game

Matching and disambiguating database records related to single individuals is a long standing problem in computer science for which different names have been given over the years: record linkage, entity resolution, entity disambiguation, record matching, object identification, data integration (Winkler 2006; Elmagarmid *et al.* 2007). It is an active area of research. Uniquely assigning documents to individuals is a challenging endeavor, because of the existence of synonyms, homonyms, abbreviations, spelling mistakes and poor quality of reported personal information. The final aim of this area of research in computer science is to design fully automated techniques that can be used efficiently for large volumes of data. Human raters, however, are still needed, either to build learning sets or to control the quality of the automated results. The increasing availability of large datasets makes the cost of human intervention unaffordable in some cases, but without manual validation, the quality of the final result is not always as good as needed.

[54] Howe, Jeff (June 2, 2006). "Crowdsourcing: A Definition". Crowdsourcing Blog. Retrieved January 2, 2013.

Social scientists are increasingly confronted with the need to apply matching and disambiguation techniques to disambiguate data, but with some exceptions, most work was until recently done manually and not much information was given in the research articles about the data and disambiguation techniques used to get the final data used for the analysis. One of these exemptions are the studies pioneered by Trajtenberg *et al.* (2006) that use disambiguation techniques to reclassify patent data at the inventor level in innovation studies, solving what it is often referred to as the patent 'Names-Game' (Raffo and Lhuillery 2009).[55] Considerable efforts have been devoted by different research groups over the past years to disambiguate inventors listed in patents and identify academic researchers amongst them. This has been mainly done in three different ways: i) matching inventors to research staff lists (e.g. Lissoni *et al*. 2008); ii) searching for the "professor" title in the inventors' name fields (e.g. Czarnitzki *et al*. 2007); and ii) matching inventors to authors of scientific publications (Dornbusch *et al*. 2012).

The latter is the basis for the experiment presented in the current paper, for which we rely on the database of Spanish author-inventors created by Maraut and Martinez (2013). The mix of the specific features of Spanish names (e.g. multiple surnames), the lack of structure of person and institution name fields in large bibliographic databases (for patents and publications) and the frequent existence of input errors due to poor understanding of the Spanish name patterns makes this data particularly useful for as it reduces the efficiency of off-the-shelf matching algorithms and exact matching techniques and increases the importance of including quality control through manual validation by human raters.

Best practice for engaging with Mechanical Turk

AMT has been adopted by scientists for a wide variety of activities ranging from data collection (e.g. Snow *et al.,* 2008), image analysis (Maisonneuve and Chopard, 2012), to interview transcription (Marge *et al.,* 2010), and copy-editing (Bernstein *et al.,* 2010). It has also been deployed for activities that are very similar to the Names-Game such as Entity Resolution (Wang *et al*., 2011; Demartini *et al*., 2012). It has been previously observed that the quality of task formulation strongly influences the quality of the results obtained (Kittur *et al*., 2008). Our framing of the Names-Game adopts the template provided by Wang *et al*. (2012) as starting point. That is, we present the AMT workers with a list of items to be compared. The items are preselected by a clustering algorithm to ensure that the comparisons are sufficiently challenging. AMT workers select the tasks they want to carry out among the ones that are available. Typically, a limited number of workers will end up doing the brunt of the work (Bernstein *et al.,* 2010). It is possible for the requester to require that workers pass a qualification first. Alonso and Mizzaro (2012) find that workers who have passed a test are

[55] For information on most recent developments see the European Science Foundation Research Networking Programme – Academic Patenting in Europe (APE-INV) at http://www.esf-ape-inv.eu/ . The current project has been developed in the framework of that programme.

more likely to complete the tasks. Furthermore, Wang *et al*. (2012) find that the workers who have passed the qualification tests deliver work of slightly higher quality.

In order to attract the attention of workers, it helps if the tasks are relatively easy to grasp. It also helps if there are not too many other tasks competing for attention. Ipeirotis (2009) observed that most tasks are launched during weekdays and that most workers are active during weekend. If this still holds, it would be better to launch the task during the weekend. It also helps to offer higher pay than other requesters. According to Horton and Chilton (2010) a higher effort level can be expected in return for a higher pay. They also discovered that a number of workers clearly prefer earning total amounts that are evenly divisible by 5 and speculate that this might be because these workers pursue earning targets. The quality of the work does not seem to be affected by the level of payment, however (Mason and Watts, 2009; Mason and Suri, 2012).

In order to improve quality, Shaw *et al.* (2011) find that it helps to indicate that payment will be linked to the extent in which responses conform to responses given by peers. For this to work, the lists have to be given to a large enough number of different workers. Overall, cheating seems to have become more prevalent at AMT over time (Eickhoff and de Vries 2012). Sun *et al.* (2011) observe that workers are more likely to continue or complete a task if they enjoy doing it, yet according to one worker interviewed by Kittur *et al.* (2012) tasks are often monotonous. In order to make the Names-Game more interesting we consider adding an additional question asking workers to identify the gender of the people in the list. Hopefully, this will alleviate the complaint of a worker interviewed by Kittur *et al.* (2012) that many task assignments are monotonous. The gender assignments thus obtained can provide a further indication of the seriousness of the workers and can be used to correct for misbehavior ex post (Shaw *et al.* 2011). Among the other measures to improve quality, Ipeirotis (2010b) advises that one should announce the rules of the game clearly in the task description an announce sanctions if deficiencies are observed. Finally, Kittur *et al.* (2008) found a significant increase in the quality of the data obtained after the inclusion of additional questions with verifiable answers. The inclusion of feedback once every so often might also be useful in case of the Names-Game.

Task Protocol

So, with regards to the design of the Names-Game task for AMT, it appears that clarity and attractiveness of the formulation of the task is crucial. In addition, proper selection of workers will improve quality and proper timing is important to attract attention. Hence, we announce the main tasks on a Saturday and try different rewards per task. In order to ensure the participation of many different workers, we limit the number of tasks a worker can carry out to a maximum of five. Each task is presented as a list of records with which persons are to be associated. In order to make the task more interesting, we include a checkbox for gender (male/female) next to each patent application record. The tasks are

composed with help of a clustering algorithm, which helps ensure they are sufficiently challenging. With respect to worker selection, we test four types of filters: accept everyone (1); accept qualified workers only (2); accept workers with skills in Spanish only (3).

Data

We use data on Spanish author-inventors from Maraut and Martinez (2013). In particular, we focus on the set of validated matches included in that database corresponding to EPO patent applications with Spanish applicants filed in 2007-2008 (2,727) with all scientific publications of 2008 indexed in SCOPUS (55,980). After discarding the most obvious non-matches, it includes 14,869 author-publication/inventor-patent pairs broken down into 1,722 distinct clusters (a cluster includes articles and patent applications likely to belong to the same author-inventor). We then limit the sample for our experiment by considering only 'journal articles', for which additional information would be easier to find online by AMT workers if necessary,[56] and select a set of 100 clusters. In particular, we split the data set into five subsets based on the proportion of author-applicant pairs, which are considered to be the same by our expert reviewer. From each subset we randomly draw 5 clusters (i.e. 5 with 0-20% pair-agreement, 5 with 20-40% agreement and so on). This to make sure that there is sufficient variety among our tasks.

Table 1. Example of a task in the AMT Names-Game prototype

Person	*Gender*	*Document Type*	*First Name*	*Last Name*	*Address*	*Affiliation*	*Patent Applicant*	*Document Title*
1	F	PATENT	María Dolores	Toro García	Madrid	Universidad A	Firm A	Title 5
2	M	PATENT	Manuel	Toro López	Pontevedra		Toro González, José	Title 6
3	M	PATENT	Maximino	Toro González	Mallorca		Firm B	Title 7
0	M	ARTICLE	José María	del Toro	Madrid	Hospital A		Title 1
1	F	ARTICLE	María	Toro	Madrid	Universidad A		Title 2
0	M	ARTICLE	Mario	del Toro	Madrid	Hospital B		Title 3
1	U	ARTICLE	M. D.	del Toro	Ciudad Real	Hospital C		Title 4

Note: Gender M stands for male, F for female and U for 'unidentifiable' due to lack of information (i.e. initials only).

[56] About 70% of all SCOPUS publications are journal articles (original research or opinion published in peer reviewed journals).

We propose 10 different tasks (clusters) to AMT workers in each experiment, which are randomly drawn from the 100 clusters we pre-selected. For each task we provide information as set out in the columns 'document type'; 'first name; 'last name; 'address'; 'affiliation'; 'patent applicant' and 'document title' of Table 1. The document title is linked to a version of the document available online, so that the worker can get additional information if needed (e.g. abstract, coauthors). The first two columns in Table 1 ('person' and 'gender') show the true responses that corresponds to this fictitious example, against which we compare the responses of the workers. They are empty in the version shown to workers.

Research in progress

We launched five experiments in April 2013, after trying some beta versions in 2012. The increase in the price radically increased the number of workers submitting tasks. The fact that the hits with the high reward were launched on a Saturday might also have a positive influence on its accomplishment, but the price seems to be determinant. Since we only allowed a maximum of 5 and 3 hits per assignment in the Saturday experiments, we received 50 responses for the first and 30 responses for the second, the maximum allowed for 10 workers participating in each. While the batch is on progress, AMT provides information on average time per hit spent by workers and effectively hour reward they get from working on each hit. Our first two tests, at a low price, were useful to estimate the average time spent per hit, which was about 6-7 minutes per hit, which at 0.10$ per hit represented an hourly rate of about 1$. Previous analysis and online blogs suggest that a correct reward is between 5 and 10$ per hour, so we realized we were paying too little rewards relatively to the complexity of the hits and attractiveness of the task. We then shortened the text and increased the reward.

Table 2. AMT Names-Game experiments, April 2013

Launched	Time life	Maximum time per hit	Reward	Maximum number of assignments per hit	Qualifi-cations	Tasks visible to	Introductory text	Number of workers	Average time per hit
Wednesday	24 hours	30 minutes	0.05 $	5	Hit approval rate >=95%	Only qualified	Long, with examples	1	6 minutes
Thursday	24 hours	30 minutes	0.10 $	5	Nothing	Everyone	Short, no examples	3	7 minutes
Saturday	24 hours	30 minutes	0.50$	5	Nothing	Everyone	Short, no examples	10	4 minutes
Saturday	24 hours	30 minutes	0.50$	3	Hit approval rate >=60%	Only qualified	Short, no examples	10	4 minutes

Figure 1 below provides a preliminary glimpse at the results from our experiments. It shows the amount of time each worker spent on the tasks. The workers are identified by the color of the dots and the tasks can be identified based on the proportion of valid pairs that our expert reviewer associated with them (except for the three tasks with a proportion of valid pairs equal to zero). Note that the number of tasks per worker varies greatly. Also note that some

workers are consistently faster than others. Nonetheless, the figure suggests there might be an U-shaped relationship between the time spent on a task and the difficulty of the task as expressed by the proportion of valid pairs in the cluster.

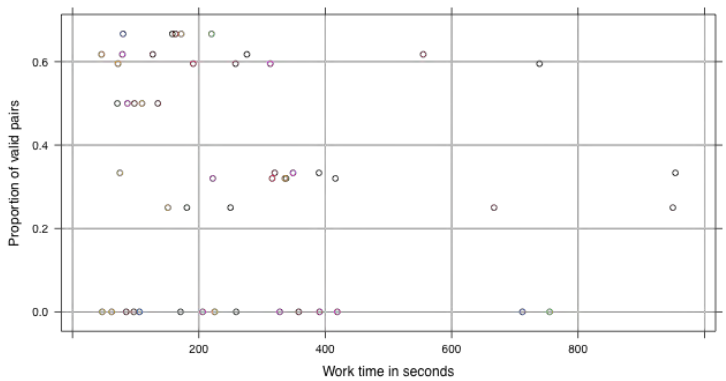


Figure 1. Average time per hit v proportion of valid pairs in experiment with 50 responses

References

Alonso, O., Mizzaro, S., 2012. Using crowdsourcing for TREC relevance assessment. Information Processing & Management 48, 1053–1066.

Bernstein, Michael S., Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. "Soylent: a Word Processor with a Crowd Inside." In Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, 313–322. UIST '10. New York, NY, USA: ACM, 2010. doi:10.1145/1866029.1866078.

Czarnitzki, D., Hussinger, K, and C. Schneider (2007), "Commercialising academic research: the quality of faculty patenting", Industrial and Corporate Change, 20, 1403-1437.

Dornbusch, F., Schmoch, U., Schulze, N. and N. Bethke (2012), "Identification of university-based patents: a new large-scale approach". Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis No. 32. Karlsruhe, July 2012.

Elmagarmid, A., Ipeirotis, P. and V. Verykios (2007), "Duplicate record detection: a survey", IEEE Transactions on Knowledge and Data Engineering, 19, 1, 1-16.

Ipeirotis, P. (2009). When to Post Tasks on Mechanical Turk? | A Computer Scientist in a Business School [WWW Document], URL http://www.behind-the-enemy-lines.com/2009/08/when-to-post-tasks-on-mechanical-turk.html (accessed 1.28.13).

Ipeirotis, P. (2010). Be a Top Mechanical Turk Worker: You Need $5 and 5 Minutes | A Computer Scientist in a Business School [WWW Document].

URL http://www.behind-the-enemy-lines.com/2010/10/be-top-mechanical-turk-worker-you-need.html (accessed 1.28.13).
Kittur, Aniket, Ed H. Chi, and Bongwon Suh. "Crowdsourcing User Studies with Mechanical Turk." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 453–456. CHI '08. New York, NY, USA: ACM, 2008. doi:10.1145/1357054.1357127.
Lissoni, F., Llerena, P., McKelvey, M. and B. Sanditov (2008), "Academic patenting in Europe: new evidence from the KEINS database", Research Evaluation, 17, 2, 87-102.
Maisonneuve, N., Chopard, B., 2012. Crowdsourcing Satellite Imagery Analysis: Study of Parallel and Iterative Models, in: Xiao, N., Kwan, M.-P., Goodchild, M.F., Shekhar, S. (Eds.), Geographic Information Science, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 116–131.
Maraut, S. and C. Martinez (2013), "Identifying Spanish author-inventors: methods and first insight into results", Working Paper, CSIC Institute of Public Goods and Policies.
Marge, Matthew, Satanjeev Banerjee, and Alexander I. Rudnicky. "Using the Amazon Mechanical Turk to Transcribe and Annotate Meeting Speech for Extractive Summarization." In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 99–107. CSLDAMT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. http://dl.acm.org/citation.cfm?id=1866696.1866712.
Mason, W., Suri, S., 2012. Conducting behavioral research on Amazon's Mechanical Turk. Behav Res 44, 1–23.
Mason, W., Watts, D.J., 2009. Financial incentives and the "performance of crowds", in: Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '09. ACM, New York, NY, USA, pp. 77–85.
Raffo, J. and S. Lhuillery (2009), "How to play the "Names Game": Patent retrieval comparing different heuristics" Research Policy, 38 (2009) 1617–1627
Schmoch, U., Dornbusch, F., Mallig, N., Michels, C., Schulze, N. and N. Bethke (2012), Vollstandige Erfassung von Patentanmeldungen aus Universitaten. Bericht an das Bundesministerium fur Bildung und Forschung (BMBF). Revidierte Fassung, Karlsruhe: Fraunhofer ISI.
Shaw, A.D., Horton, J.J., Chen, D.L., 2011. Designing incentives for inexpert human raters, in: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW '11. ACM, New York, NY, USA, pp. 275–284.
Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. "Cheap and Fast—but Is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 254–263. EMNLP '08. Stroudsburg, PA, USA:

Association for Computational Linguistics, 2008. http://dl.acm.org/citation.cfm?id=1613715.1613751.
Sun, Y., Wang, N., Peng, Z., 2011. Working for one penny: Understanding why people would like to participate in online tasks with low payment. Computers in Human Behavior 27, 1033–1041.
Trajtenberg M., Shiff G. and R. Melamed (2006), "The 'Names Game': Harnessing inventors' patent data for economic research", NBER working paper 12479.
Wang, J., Kraska, T., Franklin, M.J., Feng, J., 2012. CrowdER: crowdsourcing entity resolution. Proc. VLDB Endow. 5, 1483–1494.
Winkler, W.E. (2006), "Overview of record linkage and current research directions", Statistical Research Division U.S. Census Bureau, Research Report Series, Statistics #2006-2.

DETECTING THE HISTORICAL ROOTS OF RESEARCH FIELDS BY REFERENCE PUBLICATION YEAR SPECTROSCOPY (RPYS)

Werner Marx[1], Lutz Bornmann[2], Andreas Barth[3]

[1] *w.marx@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstraße 1, D-70569 Stuttgart (Germany)

[2] *bornmann@gv.mpg.de*
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich (Germany)

[3] *andreas.barth@fiz-karlsruhe.de*
FIZ Karlsruhe, Hermann-von-Helmholtz-Platz 1, D-76344 Eggenstein-Leopoldshafen (Germany)

Abstract
Scientific progress is almost always based on historical predecessors who have already provided important contributions to the new research field. However, it is rather difficult to identify historical roots in a systematic manner and to the best knowledge of the authors there is no method available which can be used independently of the research field. In this paper we introduce a new quantitative method to determine the historical roots of research fields and to quantify their impact on current research. Our method is based on the frequency of citations within a specific research field as a function of the publication year. Major historical contributions appear as more or less pronounced peaks, depending on the total number of citations within this research field. In most cases, these peaks are caused by high citation rates of individual historical publications. In analogy to spectroscopy which shows physical phenomena as peaks in a spectrum we have named our new method reference publication year spectroscopy (RPYS). In this study, we use research on graphene (a recently prepared new material) to illustrate how RPYS functions and what results it can deliver.

Conference Topic
Scientometrics in the History of Science (Topic 12) and Sociological and Philosophical Issues and Applications (Topic 13).

Introduction
Research activity usually evolves on the basis of previous investigations and discussions between the experts in a scientific community: “Original ideas seldom come entirely ‘out of the blue’. They are typically novel combinations of existing ideas” (Ziman, 2000, p. 212). Earlier findings are re-combined and developed

further on, resulting in the accumulation of knowledge and thus scientific progress.

According to Popper (1961) scientists formulate empirically falsifiable hypotheses, develop empirical tests for these hypotheses and apply them. Some hypotheses remain intact as this process is repeated or applied in different contexts and some are rejected. Thus, knowledge is acquired when hypotheses are formed on the basis of earlier findings and of the empirical testing they undergo. In Kuhn's alternative view (Kuhn, 1962) knowledge is acquired when scientists work on certain problems or puzzles. According to Kuhn (1962) scientists working under normal circumstances are guided by certain paradigms or exemplars which provide a framework for the work (puzzle-solving). Paradigms are "a set of guiding concepts, theories and methods on which most members of the relevant community agree" (Kaiser, 2012, p. 166). When scientists question what represents good evidence and reason in a research field, and a different framework offers a better alternative, one paradigm replaces the other. Kuhn therefore believes that knowledge is acquired through changes in paradigms in a non-cumulative process. Popper (1961), on the other hand, sees a cumulative process. While "Popper is more concerned with the normative and prescriptive question of how science *should* be carried out, and Kuhn is more concerned with the descriptive question of how science *is* carried out" (Feist, 2006, p. 30).

Although there are many differences between these two theories of scientific development, the relationship of current research to past literature plays a significant role in both: knowledge cannot be acquired without this relationship. The relationship to earlier publications is expressed in the form of references to or citations of them in later publications. The content of an earlier publication and that of the later publication which refers to it are usually related and the former is usually of significance to the existence of the latter. The premiss of citation analysis and its application to the evaluation of research is that, in terms of statistics, the more frequently scientific publications are cited, the more important they are for the advancement of knowledge (Merton, 1965; Bornmann et al., 2010). Therefore, citation data also provides interesting insight into the historical science context, in terms of the significance of the previous historical publications on which the later publications in a field of research are based. In this study we introduce the quantitative method named as reference publication year spectroscopy (RPYS) and show examples of how it is possible to determine and further analyse the historical roots referred to in the publications cited within a single research field. This method is based on the citation-assisted background (CAB) method proposed by Kostoff and Shlesinger (2005) which is a "systematic approach for identifying seminal references" (p. 199) in a specific field (Kostoff et al., 2006).

Methods

Citation analyses are usually based on a publication set comprising the publications of a researcher, of a research institution or in a journal. The number

of times these publications are cited is analysed to evaluate research performance. As a rule, citations from every research field and not only those of the citing publications within a certain research field are taken into account.

In a previous publication (Bornmann & Marx, 2013) it has been proposed for certain issues to reverse the perspective of citation analysis from a forward view on the overall citation impact of the publications to a backward view, where the impact of publications, authors, institutions or journals within a specific research field can be determined (Kostoff & Shlesinger, 2005).[57] We have shown that it is possible to limit citation analyses to single research fields by first selecting their publications and then analysing the references cited (fully) in them. A cited reference analysis of this kind can also be used to determine the historical roots of a research field and to quantify the significance of historical publications.

Empirically, it appears that most references refer to more recent specialist literature in the discipline in which the citing publication has appeared – only a relatively small proportion of the cited publications is older and derives from different disciplines, respectively. The distribution of the cited publications over their publication years (that is, reference publication years, RPY) is typically at a maximum a few years before the publication year of the citing publications and then tails off significantly into the past. The (steep) decline over time is not only associated with the fact that specialist literature as a rule becomes less interesting and important as time passes (ageing). It is also the result of an abrupt increase in specialist literature in every discipline which started around 1960 ("Sputnik shock") and continues to this day. For example, just 2% of the literature on physics in the 20th century was published before 1950 (Marx, 2011).

Quantitative analysis of the publication years of all the publications cited in the publications in one research field shows that RPYs lying further back in the past are not represented equally, but that some RPYs appear particularly frequently in the references. These frequently occurring RPYs become more differentiated towards the past and mostly show up as distinct peaks in the RPY distribution curves. If one analyses the publications underlying these peaks, it is possible to see that during the 19th and the first half of the 20th century they are predominantly formed by single relatively highly cited publications. These few, particularly frequently cited publications as a rule contain the historical roots to the research field in question. The publications can be found with cited reference analysis (Bornmann & Marx, 2013) and it is possible to determine how the relationship to earlier publications developed over time; that is, at which stage in the development of the research field these publications were (re-)discovered and then cited more frequently. Towards the present, the peaks of individual publications lie over a broad continuum of newer publications and are less pronounced. Due to the many publications cited in the more recent RPYs, the proportion of individual, much-cited publications in the RPYs falls steadily.

[57] It has already been proposed in another publication to analyse the typical use of bibliographical references by individual scientists (Costas et al., 2012).

The focus on the important historical publications in one research field is a special application of the method known as cited reference analysis (Bornmann & Marx, 2013). In an analogy to the spectra in the natural sciences, which are characterised by pronounced peaks in the quantification of certain properties (such as the absorption or reflection of light as a function of its colour), we call this special application RPYS. To illustrate RPYS we present here an example of research on graphene.

The results of the RPYS on graphene presented here are based on the Science Citation Index (SCI) which is accessible via the SCISEARCH database offered by the database provider STN International (http://www.stn-international.com/). This database combined with the STN search system enables sophisticated citation analyses. Among many other options, the SCISEARCH database searched via STN International makes it possible to ask which historical publications in the various fields of the natural sciences have been cited most frequently by the publications since 1974, the period covered by the SCISEARCH database. The Web of Science (WoS) provided by Thomson Reuters, the most common search platform of the Thomson Reuters citation indexes, stretches back to 1900. However, the WoS search functions have not been optimized for the bibliometric analysis presented in this study. The selection of numerous references from large sets of citing publications and their further analysis is not possible under WoS.

STN's retrieval language, Messenger, allows the publications from a specific research field to be selected and all the references they cite to be extracted. Instead of the complete references it is also possible to select and analyse just the authors of the publications in the cited references, the journals or the RPYs. In this publication we are concerned mainly with the analysis of the RPYs and especially the early publications cited particularly frequently as the historical roots of a research field. The first step in RPYS is to select the publications for a certain research field and extract all the cited publications (the references) from them. The second is to establish the distribution of the frequencies of the cited references over the RPYs and from this determine the early RPYs cited relatively frequently. The third is to analyse these RPYs for frequently cited historical publications.

Results

Single planar layers of graphite one atom thick are named graphene, the newest member of the carbon structural family. Graphene has been called a rising star among new materials (Geim & Novoselov, 2007; Barth & Marx, 2008). Although it has been discussed since 1947, it was not believed to exist in a free state. In 2004, however, graphene was found unexpectedly when it was isolated from graphite crystals (Novoselov et al., 2004; Novoselov et al., 2005). This defined a new allotrope of carbon in addition to diamond and graphite, nanotubes and fullerenes. Graphene exhibits some remarkable properties which feature in particular highly efficient electrical conductivity combined with extremely fast charge transport and extraordinary strength. These properties make the material

potentially useful in a wide range of applications such as in electronics (high speed transistors, and single-electron transistors) and in materials science (composite materials) (Geim & Novoselov, 2007; Geim & Kim, 2008).
The experimental discovery of free-standing graphene sheets as a new member of the carbon structural family caused a "gold rush" to surround this interesting and promising research field, leading to a substantial rise in the number of publications. Since research on graphene has become a "hot topic" for scientists, it is not surprising that the publication (and citation) pattern of such a new research field is also of great interest for scientometric studies (see e.g. Winnink, 2012).

Table 1. Search query for the RPYS of the literature on graphene.

```
=> dis hist

     (FILE 'SCISEARCH' ENTERED AT 09:35:22 ON 14 AUG 2012)
                DEL HIST Y
L1         19356 S GRAPHENE
                SET TERM L#
L2            SEL L1 1- RPY :     185 TERMS

=> dis l2 1- alpha delim
L2            SEL L1 1- RPY :     185 TERMS

…
34;2;1;0.01;1850
35;2;2;0.01;1852
36;4;4;0.02;1853
37;3;2;0.01;1854
38;14;14;0.07;1855
39;1;1;0.01;1856
40;2;2;0.01;1857
41;1;1;0.01;1858
42;120;120;0.62;1859
43;89;89;0.46;1860
44;2;2;0.01;1865
45;5;4;0.02;1866
46;5;5;0.03;1867
47;1;1;0.01;1870
…
```

Notes. L1: Selection of the graphene publications. L2: Extraction of the RPYs from all of the cited references (both list number entries marked in light grey). The number of references with RPYs from 1850 to 1870 (cut-outs of the full STN specific display list including the earliest pronounced peak with n=120/89 cited references in 1859/1860, again marked in light grey) are displayed here for demonstration.
Source: SCISEARCH under STN International.

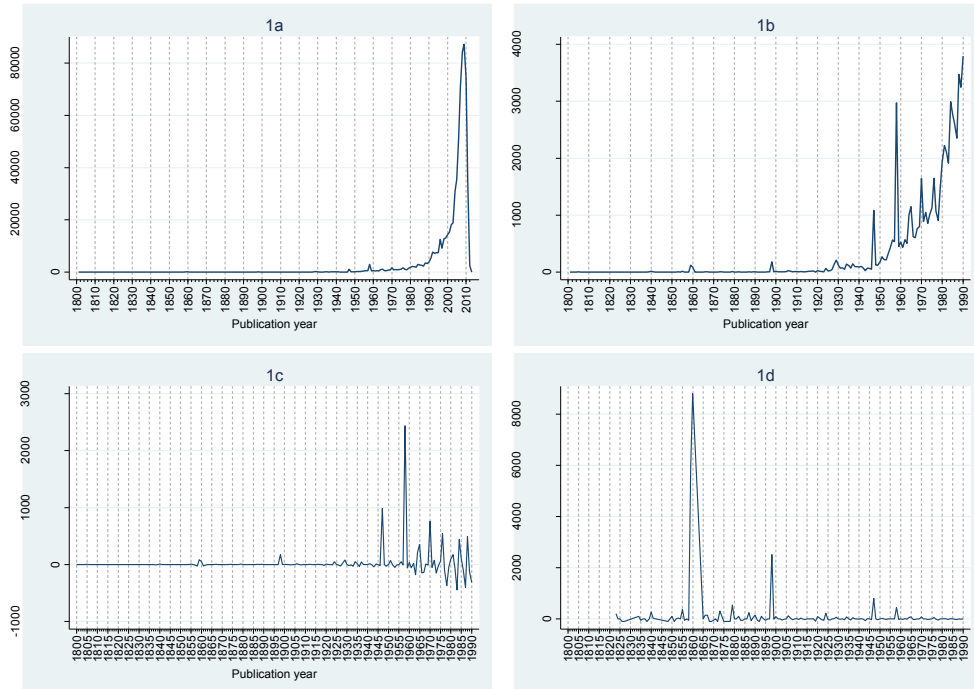


Figure 1 a-d. Annual distributions of cited references in publications of research on graphene.

In this study, the publications dealing with graphene were selected by searching for the term "graphene" in the title and abstract search fields of the SCISEARCH. There is no need for a search in a field-specific database such as the Chemical Abstracts Service (CAS) literature database because the core literature covered by the SCISEARCH is sufficient to reveal the most frequently cited historical publications. The STN search query for the RPYS of the graphene literature is given in Table 1. Of the complete set of 19356 publications on graphene research published since 1974 (the time period covered by the SCISEARCH database accessible under STN International) in the journals covered by SCISEARCH (SCISEARCH source journals), all the cited references (n=679023) have been extracted (date of the literature search: 14-08-2012).

The distribution of the number of references cited in graphene literature across the publication years is presented in Figures 1a-d. Figure 1a shows the distribution of the number of all the references cited in graphene publications across their publication years. The most frequently cited RPY is 2009, showing the strong contemporary relevance of this newly emerging research field. The RPYs are presented here back to the year 1800.

Table 2. Search query for the cited references in 1859/60.

```
=> dis hist

     (FILE 'SCISEARCH' ENTERED AT 09:35:22 ON 14 AUG 2012)
                DEL HIST Y
L1          19356 S GRAPHENE
                SET TERM L#
L2            SEL L1 1- RPY :     185 TERMS
L3            205 S L1 AND (1859 OR 1860)/RPY
L4            183 S L1 AND 1898/RPY
L5           1073 S L1 AND 1947/RPY
L6           2862 S L1 AND 1958/RPY
L7            SEL L3 1- RE HIT :      14 TERMS
L8            SEL L4 1- RE HIT :      10 TERMS
L9            SEL L5 1- RE HIT :      72 TERMS
L10           SEL L6 1- RE HIT :     251 TERMS

=> dis l7 1- occ delim
L7            SEL L3 1- RE HIT :      14 TERMS

1;112;112;54.63;BRODIE B C, 1859, V149, P249, PHILOS T ROY
SOC LON
2;77;77;37.56;BRODIE B C, 1860, V59, P466, ANN CHIM PHYS
3;5;5;2.44;BRODIE M B C, 1860, V59, P466, ANN CHIM PHYS
4;3;3;1.46;BRODIE B, 1859, V149, P249, PHILOS T R SOC LONDO
5;2;2;0.98;BRODIE B C, 1859, V10, P249, P ROY SOC LONDON
6;2;2;0.98;BRODIE B C, 1860, V12, P261, Q J CHEM SOC
7;1;1;0.49;BRODIE B C, 1859, V10, P11, P R SOC LONDON
8;1;1;0.49;BRODIE B C, 1859, V149, P10, PHILOS T R SOC
9;1;1;0.49;BRODIE B C, 1860, V114, P6, LIEBIGS ANN CHEM
10;1;1;0.49;BRODIE B, 1860, P59, ANN CHIM PHYS
11;1;1;0.49;BRODIE B, 1860, V59, P17, NN CHIM PHYS
12;1;1;0.49;BRODIE B, 1860, V59, P7, ANN CHIM PHYS
13;1;1;0.49;BRODIE E C, 1860, V59, P466, ANN CHIM PHYS
14;1;1;0.49;BRODIE F R S, 1859, V149, P249, PHILOS T R SOC
LONDO
...
```

Notes. L3-L10: List numbers comprising the search steps of the analysis of the RPYs 1958/60 with peaks and demonstrating the analysis method by displaying the reference variants of the publication by Brodie (1859/1860) as an example (with the relevant search steps and displayed results marked in light grey).
Source: SCISEARCH under STN International.

Figure 1b shows a cut-out limiting the RPYs to 1800-1990 with the distinct peaks of the most frequently cited historical publications more clearly visible. The citing graphene publications were published between 1974 and the present (mainly since 2004), whereas the time window of the cited publications (the references cited within the citing graphene publications and analysed here) extends from 1800 to

1990 in order to focus on historical publications and to provide suitable scaling to reveal the peaks. Figures 1c and 1d show the deviation of the number of cited references in one year from the median for the number of cited references in the two previous, the current and the two following years. While Figure 1c shows the absolute deviation from the median, Figure 1d illustrates the deviation in percent. It is particularly easy to see the peaks created by the frequently cited historical publications in the deviations expressed by percentage.

The search query for the citation analysis of the peak in the RPYs 1859/60 via the SCISEARCH database under STN International is given in Table 2. As the list of references shows many references have turned out to be erroneous. Misspelled citations (e.g. incorrect with regard to the numerical data: volume, starting page, and publication year) are a general problem in citation analysis. The references in earlier publications, however, are particularly susceptible to 'mutations' (Marx, 2011).

The four most clearly pronounced peaks in Figure 1d can be attributed to early publications on graphite oxide which are most important for graphene research. Table 3 specifies the four most frequently cited historical publications, including their bibliographic data and comments on the publications taken from a review on graphene research (Dreyer et al., 2010). The relevance of the publications as the historical roots of this newly emerging research field was highlighted in this review. The review cites the four publications in Table 3 and also two further publications with less pronounced peaks (but no other publications published before 1960 which were not identified in our study). The two publications of Schafhaeutl (1840a; 1840b) cited additionally in the review can be seen as precursors to Brodie's publications (Brodie, 1859; 1860) (see Table 3). One publication by Schafhaeutl (1840a) appeared in a German journal where fewer citations can be expected.

Table 3. The four most frequently cited early (pre-1990) references in graphene literature. In each case, the relevant RPY, the number of references in the graphene literature attributed to the specific publication, the total number of references in the graphene literature with regard to the given RPY, the overall number of citations of the specific publication until October 2012 (TC=Times Cited), and the relevant comment from Dreyer et al. (2010) are listed.

RPY	Reference / Comment	TC
1859/1860	204 of 205 references refer to: Brodie, B.C. (1859). On the atomic weight of graphite. Philosophical Transaction of the Royal Society of London, 149, 249-259. Brodie, B.C. (1860). Sur le poids atomique du graphite [On the atomic weight of graphite]. Annales de Chimie et de Physique, 59, 466-472. *"In 1859, the British chemist Brodie used what may be recognized as*	324

	modifications of the methods described by Schafhaeutl in an effort to characterize the molecular weight of graphite by using strong acids (sulfuric and nitric), as well as oxidants, such as KClO3" (p. 9337).	
1898	177 of 183 publications refer to: Staudenmaier, L. (1898). Verfahren zur Darstellung der Graphitsäure [Method for the preparation of graphitic acid]. Berichte der Deutschen Chemischen Gesellschaft, 31, 1481-1487. *"Nearly 40 years later, Staudenmaier reported a slightly different version of the oxidation method used by Brodie for the preparation of GO by adding the chlorate salt in multiple aliquots over the course of the reaction instead of in a single portion" (p. 9338).*	270
1947	962 of 1073 publications refer to: Wallace, P.R. (1947). The band theory of graphite. Physical Review 71, 622-634. *"As early as the 1940s, a series of theoretical analyses suggested that these layers—if isolated—might exhibit extraordinary electronic characteristics (e.g., 100 times greater conductivity within a plane than between planes) " (p. 9336).*	1467
1958	2095 out of 2862 publications refer to: Hummers , W.S. (Jr.) & Offeman, R.E. (1958). Preparation of graphite oxide. Journal of the American Chemical Society, 80(6), 1339-1339. *"Graphite oxide: A berthollide layered material prepared by treating graphite with strong oxidants, whereby the graphite surface and edges undergo covalent chemical oxidation. The degree of oxidation may vary, though strongly oxidized graphite oxide typically exhibits a C/O ratio of approximately 2:1" (p. 9342).*	2511

The question arises at which point in time the historical publications were cited most frequently. Are such publications already taken account of at the start point of a new research field (in the case of graphene research this is 2004) since the research is directly based on them? Or are they detected, for example, as forerunners not before literature reviews are published (which discuss the historical background)? Figure 3 shows the evolution over time (citation history) of the four most frequently cited historical publications mentioned above against the backdrop of the time curve for the literature on graphene in total. The citation numbers of the four publications are limited to citing publications dealing with graphene.

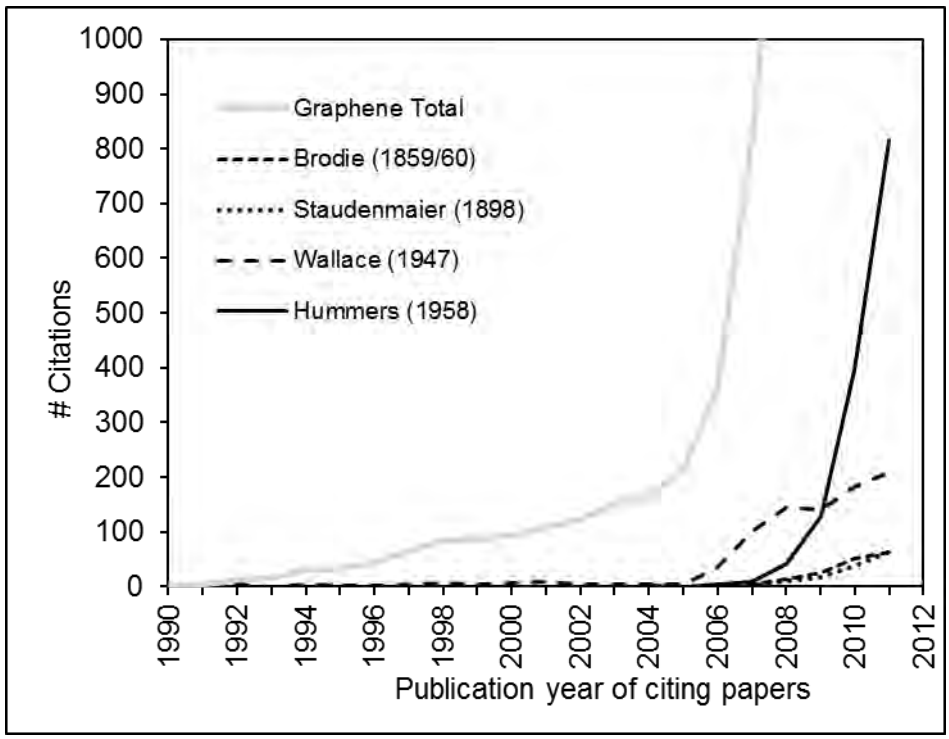


Figure 3. Citation history of the four most frequently cited historical publications in graphene literature. The overall number of graphene publications published per year is shown for comparison.

Only the publication of Wallace (1947) was cited more frequently immediately after the discovery of graphene in the year 2004, whereas the other three historical publications (Brodie, 1859/60; Staudenmaier, 1898; Hummers & Offeman, 1958) did not receive a boost until two years later. This can be explained by the fact that the boom in graphene research was triggered by a physical preparation method and the focus initially was on the physical properties predicted in theory. Accordingly, as a theoretical physics publication, the publication of Wallace (1947) was immediately cited more frequently. Over 85% of the citing publications are classified as physics research. Researchers into chemistry only subsequently started looking at the question of how graphene could be synthesized chemically, which made the other historical publications (Brodie, 1859/60; Staudenmaier, 1898; Hummers & Offeman, 1958) on graphite oxide relevant. Around 70% of the publications cited here are from research into chemistry. A comparison of all the literature on graphene in these two research areas shows that generally speaking, the reaction of the chemistry community to the discovery of graphene came two years after that of the physics community.

As described above, the discovery of free-standing graphene goes back to the publications by Novoselov et al. (2004). The earliest references in these

publications are from the 1980s (1981). One possible reason for the absence of historical publication data could be that the publications are relatively short and focus on the discovery of free-standing graphene. Furthermore, the physical and chemical proofs for the new discovery were given priority. The authors did not discuss the history of the discovery until three years later (Geim & Novoselov, 2007).

Discussion

In this study we proposed RPYS, a bibliometric method with which it is possible to determine the historical roots of research fields and quantify their impact on current research. "If you want to know how science is carried out, then in one way or another, you are going to have to look at the history of science" (Lehoux & Foster, 2012, p. 885). The RPYS method is based on an analysis of the frequency references are cited in the publications in a single research field by publication year. The origins show up in the form of more or less pronounced peaks mostly caused by individual historical publications which are cited particularly frequently. As the RPYS can only indicate the possible origins, a second step is required in which specialist experts verify which publications genuinely played a significant part in a research field. When those publications which resulted in a peak are identified, each of them should be reviewed for their significance in the particular research field and what contribution they made. RPYS is a very simple method which can be applied in different disciplines.

One method which approaches the quantification of historical events in a way similar to RPYS and which can be used to examine historical events on the basis of very different sources of data and mathematical models was proposed by Turchin (2003) and called cliodynamics (Spinney, 2012). Alternative methods for analysing historical papers are (1) the concept of co-citations and research fronts (Garfield & Sher, 1993) and (2) the so called "algorithmic historiography" (Garfield et al., 2003; Leydesdorff, 2010). HistCite developed by Eugene Garfield (http://garfield.library.upenn.edu/algorithmichistoriographyhistcite.html) enables a citation graph (called historiogram or historiograph) visualizing the citation network among historical publication sets. Whereas the RPYS method proposed here reveals quantitatively which historical papers are of particular interest for the specific research field or research topic, HistCite visualizes the citation network of the historical papers.

We used research on graphene to illustrate how RPYS functions and what results it can deliver. Many research fields refer in their literature to historical publications which are cited comparatively frequently and can be investigated. However, sometimes, the methods and topics of a research field are so new (e.g., of molecular biology or genetics) that the roots do not extend very far into the past. These should be looked at individually. According to Smith (2012) RPYS can be included in the "newly emerging field of 'historical bibliometrics'" (Holmes, 2012). Smith says that it is a "relatively under-researched area" in which new studies would be very welcome. For example, it would be possible to use

RPYS to examine Stigler's Law of Eponomy (Stigler, 1980), which says that "no scientific law is named after its discoverer". In a recent study, Grünbaum (2012) for example, looks at the question – without the help of bibliometrics – of whether Napoleon's theorem really is Napoleon's. Another phenomenon in the history of science that would be interesting for RPYS are multiple independent discoveries (Merton, 1973), whereby it would be possible to use bibliometrics to examine the form in which the relevant historical publications on multiple independent discoveries are cited.

References

Barth, A. & Marx, W. (2008). Graphene: a rising star in view of scientometrics, http://arxiv.org/abs/0808.3320.

Bornmann, L., de Moya-Anegón, F., & Leydesdorff, L. (2010). Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. PLoS ONE, 5(10), e11344. DOI: 10.1371/journal.pone.0013327.

Bornmann, L. & Marx, W. (2013). The proposal of a broadening of perspective in evaluative bibliometrics by complementing the times cited with a cited reference analysis. Journal of Informetrics, 7(1), 84–88. DOI: 10.1016/j.joi.2012.09.003.

Brodie, B.C. (1859). On the atomic weight of graphite. Philosophical Transaction of the Royal Society of London, 149, 249-259.

Brodie, B.C. (1860). Sur le poids atomique du graphite [On the atomic weight of graphite]. Annales de Chimie et de Physique, 59, 466-472.

Costas, R., van Leeuwen, T.N., & Bordons, M. (2012). Referencing patterns of individual researchers: Do top scientists rely on more extensive information sources? Journal of the American Society for Information Science and Technology, 63(12), 2433-2450. DOI: 10.1002/asi.22662.

Dreyer, D.R., Ruoff, R.S., & Bielawsky, C.W. (2010). From conception to realization: an historial account of graphene and some perspectives for its future. Angewandte Chemie -International Edition, 49, 9336-9345. DOI: 10.1002/anie.201003024.

Feist, G.J. (2006). The psychology of science and the origins of the scientific mind. New Haven, CT, USA: Yale University Press.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? Journal of the American Society for Information Science and Technology, 54(5), 400-412. DOI: 10.1002/asi.10226

Garfield, E., & Sher, I. H. (1993). Key Words Plus [TM]-Algorithmic Derivative Indexing. Journal of the American Society for Information Science, 44(5), 298-298.

Geim, A.K., & Kim, P. (2008). Carbon wonderland. Scientific American, 298, 90-97.

Geim, A.K. & Novoselov, K.S. (2007). The rise of graphene. Nature Materials, 6, 183-191. DOI: 10.1038/nmat1849.

Grünbaum, B. (2012). Is Napoleon's theorem really Napoleon's theorem? The American Mathematical Monthly, 119(6), 495-501. DOI: 10.4169/amer.math.monthly.119.06.495.
Holmes, R. (2012). Biography: the scientist within. Nature, 489(7417), 498-499. DOI: 10.1038/489498a.
Hummers , W.S. (Jr.) & Offeman, R.E. (1958). Preparation of graphite oxide. Journal of the American Chemical Society, 80(6), 1339-1339. DOI: 10.1021/ja01539a017.
Kaiser, D. (2012). The structure of scientific revolutions: 50th Anniversary Edition. Nature, 484(7393), 164-166.
Kuhn, T.S. (1962). The structure of scientific revolutions. Chicago, IL, USA: University of Chicago Press.
Kostoff, R. N. & Shlesinger, M. F. (2005). CAB: citation-assisted background. Scientometrics, 62(2), 199-212. DOI: 10.1007/s11192-005-0014-8.
Kostoff, R. N. et al. (2006). The seminal literature of nanotechnology research. Journal of Nanoparticle Research, 8(2), 193-213. DOI: 10.1007/s11051-005-9034-9.
Lehoux, D. & Foster, J. (2012). A revolution of its own. Science, 338(6109), 885-886. DOI: 10.1126/science.1230708.
Leydesdorff, L. (2010). Eugene Garfield and Algorithmic Historiography: Co-Words, Co-Authors, and Journal Names. Annals of Library and Information Studies, 57(3), 248-260.
Marx, W. (2011). Special features of historical papers from the viewpoint of bibliometrics. Journal of the American Society for Information Science and Technology, 62(3), 433-439. DOI: 10.1002/asi.21479.
Merton, R.K. (1965). On the shoulders of giants. New York, NY, USA: Free Press.
Merton, R. K. (1973). The sociology of science: theoretical and empirical investigations. Chicago, IL, USA: University of Chicago Press.
Novoselov, K.S., et al. (2004). Electric field effect in atomically thin carbon films. Science, 306, 666-669. DOI: 10.1126/science.1102896.
Novoselov, K.S., et al. (2005). Two-dimensional atomic crystals. Proceedings of the National Academy of Sciences of the USA, 102, 10451-10453. DOI: 10.1073/pnas.0502848102.
Popper, K.R. (1961). The logic of scientific discovery (2nd Edition). New York, NY, USA: Basic Books.
Schafhaeutl, C. (1840a). Über die Verbindungen des Kohlenstoffes mit Silicium, Eisen und anderen Metallen, welche die verschiedenen Gallungen von Roheisen, Stahl und Schmiedeeisen bilden [On the combinations of carbon with silicon and iron, and other metals, forming the different species of cast iron, steel, and malleable iron]. Journal der Praktischen Chemie, 21(1), 129-157.

Schafhaeutl, C. (1840b). On the combinations of carbon with silicon and iron, and other metals, forming the different species of cast iron, steel, and malleable iron. Philosophical Magazine, 16(106), 570-590.
Smith, D.R. (2012). Impact factors, scientometrics and the history of citation-based research. Scientometrics, 92, 419-427. DOI: 10.1007/s11192-012-0685-x.
Spinney, L. (2012). History as science. Nature, 488, 24-26.
Staudenmaier, L. (1898). Verfahren zur Darstellung der Graphitsäure [Method for the preparation of graphitic acid]. Berichte der Deutschen Chemischen Gesellschaft, 31, 1481-1487.
Stigler, S. M. (1980). Stigler's law of eponomy. Transactions of the New York Academy of Sciences, 39(1 Series II), 147-157. DOI: 10.1111/j.2164-0947.1980.tb02775.x.
Turchin, P. (2003). Historical dynamics: why states rise and fall. Princeton, NJ, USA: Princeton University Press.
Wallace, P.R. (1947). The band theory of graphite. Physical Review, 71, 622-634. DOI: 10.1103/PhysRev.71.622.
Winnink, J.J. (2012). Searching for structural shifts in science: Graphene R&D before and after Novoselov et al. (2004). In E. Archambault, Y. Gingras, & V. Lariviere (eds.), The 17th International Conference on Science and Technology Indicators (pp. 835-846). Montreal, Canada: Repro-UQAM.
Ziman, J. (2000). Real science. What it is, and what it means. Cambridge, UK: Cambridge University Press.

DETECTION OF NEXT RESEARCHES USING TIME TRANSITION IN FLUORESCENT PROTEINS

Shino Iwami[1], Junichiro Mori[2], Yuya Kajikawa[3] and Ichiro Sakata[4]

[1] *iwami@ipr-ctr.t.u-tokyo.ac.jp*
The University of Tokyo, Graduate School of Engineering, Dept of Technology Management for Innovation, Hongo 7-3-1, Bunkyo-ku Tokyo (Japan)

[2] *jmori@platinum.u-tokyo.ac.jp*
The University of Tokyo, Presidential Endowed Chair for "Platinum Society", Hongo 7-3-1, Bunkyo-ku Tokyo (Japan)

[3] *kajikawa@mot.titech.ac.jp*
Tokyo Institute of Technology, Graduate School of Innovation Management, Shibaura3-3-6, Minato-ku Tokyo (Japan)

[4] *isakata@ipr-ctr.t.u-tokyo.ac.jp*
The University of Tokyo, Policy Alternatives Research Institute, Hongo 7-3-1, Bunkyo-ku Tokyo (Japan)

Abstract
To survive worldwide competitions of research and development in the current rapid increase of information, decision-makers and researchers need to be supported to find promising research fields and papers. To find an available data in too much heavy flood of information become difficult. We aim to find leading papers of the next generation with bibliometric approach.
The analyses in this work consist of two parts: the citation network analysis and the time transition analysis. The bibliographic information of papers about fluorescent proteins is collected from Thomson Reuters' Web of Science. In the citation network analysis, each citation network is made from citation relations and divided into clusters. In the time transition analysis, the features of the leading papers are extracted, and we proposed the ways to detect the leading papers.
This work will contribute to finding the leading paper, and it is useful for decision-makers and researchers to decide the worthy research topic to invest their resources.

Conference Topic
Research Fronts and Emerging Issues (Topic 4) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction
In 1962, Shimomura, who is one of the 2008 Nobel Prize winners, discovered the green fluorescent protein (GFP) from Aequorea victoria, which is a luminous jellyfish in the sea. In addition, he found the photoprotein “aequorin”, which emits light by itself, and the phenomena that pH values and calcium ion (Ca^{2+})

concentration are capable to control the luminance of GFP. GFPs emit green light after they receive blue light from photoproteins within jellyfishes (Shimomura, Johnson & Saiga, 1962). Then, photoproteins are the mainstream research topics. However, 30 years later, Chalfie implemented GFP into other creatures in 1993 (Chalfie, Yuan & Prasher, 1993). Tsien succeeded to make various colored fluorescent proteins in 1998 (Tsien, 1998). Chalfie and Tsien are also the 2008 Nobel Prize winners.

The fluorescent proteins contribute to the elaboration of life science. The fluorescent proteins enable to trace individual proteins in live cells without autopsy agglomerated dead tissues. Different colors of fluorescent proteins are useful to examine relations between several proteins. For example, the fluorescent proteins are used for revealing the spread of cancer and the structure of neurons.

Social Issues

Developed countries need to invest for fostering industries against the recent economic recession, but some of them are forced to reduce the budget for research and development. Thus, it is important for decision-makers to determine the field for their investment, aiming to strengthen industries efficiently. In addition, some industries in Japan often select research and development strategies to develop cutting-edge areas of the world and pursue first-mover advantages, for example, in the field of supercomputers. Thus, it is needed to find the cutting-edge areas at the earlier stage. Meanwhile, information has increased year after year since the information revolution, and too much information makes it difficult for people to find their suitable information.

Methodological Issues

There are methodological issues besides social problems. Indicators using times cited, which means how many papers a paper is cited by, are considered historically as effective indicators to know leading papers, such as the impact factor by Garfield (1955). However, times-cited-based indicators give older papers an advantage and don't deal with change of importance over the years. Then, all fields are treated as the same regardless of its importance, density and time scale of fields (Vanclay, 2012). Therefore, the pervading indicator is inadequate for finding the next generation of researches.

Analyses using temporal changes were done in some research. Topological measures in citation networks of scientific publications (Shibata, Kajikawa, Takeda & Matsushima, 2008) proposed a methodology for detecting emerging researches using temporal change and relations between papers. Citation lag analysis (Nakamura, Suzuki, Tomobe, Kajikawa & Sakata, 2011) revealed that intra-cluster and inter-cluster have a time lag to contribute to develop interdisciplinary research. What these researches want to find are fields, which needs some time to build. However, any fields come from a paper at the embryo stage. We can't get away from the delay by publication as long as papers are used,

but the delay by citation will be shortened by finding a first paper instead of an emerging field.

In this work, scientific and technological detailed structures in the field of fluorescent proteins are identified, using the bibliographic information of academic papers. Then, the features of leading papers are extracted from the time transition analysis to foresee the promising papers, and we proposed the ways to detect the leading papers.

Methodology

In this work, we perform the citation network analysis of papers to reveal the detailed structure, and for the time transition analysis, we use clusters extracted from the citation network from the first year to each year.

Selected Knowledge Domain and Papers as Leading Papers

To select the query for gathering information of papers, leading papers on fluorescent proteins are decided. In 2008, Osamu Shimomura, Martin Chalfie and Roger Y. Tsien won the Nobel Prize in Chemistry for the discovery and development of the green fluorescent protein (GFP). Papers written by these winners, especially listed in advanced information "Scientific Background on the Nobel Prize in Chemistry 2008" of the official nobelprize.org (Ehrenberg, 2008), are defined as the leading papers. However, papers published from 1974 are used on account of a change of the largest graph component of the citation networks as described later in the section: Time Transition in Results and Discussion.
In addition to the papers of the Nobel Prize, papers cited by many papers are also used as the leading papers for the purpose of getting many verified results.

Citation Network Analysis and Identifying Clusters

The citation network analysis was begun by Garfield (1955), and it became an efficient tool to extract popular topics and important papers (Borner, Chen & Boyack, 2003). In the citation networks, a node is defined as a paper, and an edge is defined as a citation relation between papers. The citation networks show authors' thought about the contents of other papers related to the authors' paper.
In this paper, we do the network analysis with the following five steps. Step (1) involves collecting the bibliographic information of papers in the field of fluorescent proteins and step (2) constructs citation networks of direct citations. Direct citation is adopted for the purpose of our work, because Shibata (2009) says that direct citation is suitable for detecting emerging fields. In step (3), only the largest graph component of the citation networks is used, because this paper focuses on the relationship among papers, and we should therefore eliminate papers that have no citation from or to any others. The using relationship is the direct citation between the citing paper and the cited paper. The direct citation has less amount of calculation and clearly describes relationships, though the direct citation has a flaw that the relationships published simultaneously are not

available. After extracting the largest connected component, in step (4), the network is divided into clusters where papers are densely connected by citations from papers belonging to the same cluster by a topological clustering method. A clustered network is visualized in a manner that links, that is citations in the same cluster are visualized in the same color. A fast clustering algorithm developed by Newman (2004) was used for clustering. In the step (5) after clustering, topics of each cluster are detected by the way that an expert checked cluster contents from keywords and abstract of papers. Charts are also made from the result of clustering. A sequence of the procedure between step 2 and 4 is performed at the academic landscape system (Innovation Policy Research Center, 2013) after you throw the bibliographic information of papers in the academic landscape system.
On the step (1), papers are selected with the query '"fluorescent protein*" OR "bioluminescent protein*" OR "luminescent protein*" OR "photoprotein*"' in topic from all the papers published between 1900 and 2011 using Science Citation Index Expanded (SCI-EXPANDED), Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI) updated 2012-11-16, which are provided as a service of academic papers' database by Thomson Reuters. 32,439 papers are retrieved, and their bibliographic records are used for cluster analysis. For the largest graph component, 28,926 papers are divided into 244 clusters.

Time Transition to Find the Feature of Leading Papers

In this work, the time transitions of the centralities are used. On the basis of the above citation networks, centralities are calculated in each dataset by each year. The in-degree centrality of a paper is defined by the total number of cited papers. Each citation has a direction, but only the in-bound direction is used. The in-degree centrality values are normalized by dividing by the maximum possible in-degree in a simple graph n-1 where n is the number of nodes in the dataset. If a paper has higher in-degree centrality, it plays a role as a hub within its network. Closeness centrality at a node is the inverse number of average distance to all other nodes. The closeness centrality is also normalized. Betweenness centrality of a node V is the sum of the fraction of all-pairs shortest paths that pass through V. To calculate these centralities, the NetworkX module (NetworkX developer Team, 2013) of python is used.

Results and Discussion

Identifying Clusters

Figure 1 shows the perspective map of fluorescent proteins clustered from citation relation between papers. Shimomura's three papers in 1962, 1979 and 2005 belong to the third cluster #3. Chalfie's one paper in 1994 belongs to the top cluster #1. Tsien's paper in 1998 and his co-authors 10 papers (included in Table 2) belong to the third cluster #3. The cluster #1 includes papers how GFPs are increased for various applications and used in the living cells of many heterologous organisms as a marker. The cluster #3 includes papers about the

discovery of GFP and the feature in the proteins themselves. The gaps among the average published year (2004.0 – 2006.2) of the top 7 cluster in Figure 1 are small.

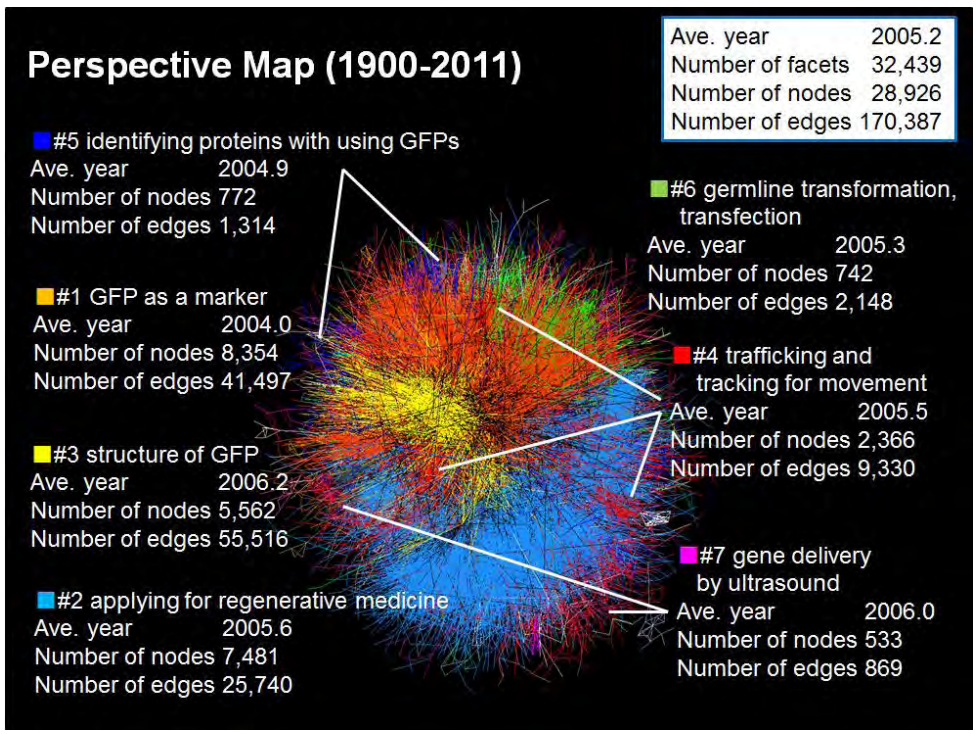


Figure 1 : Perspective Map of Fluorescent Proteins (1900-2011)

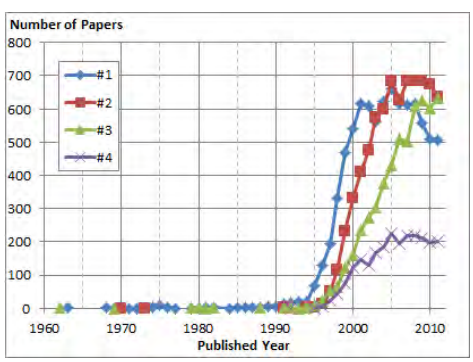


Figure 2 : Number of Papers about #1, #2, #3 and #4

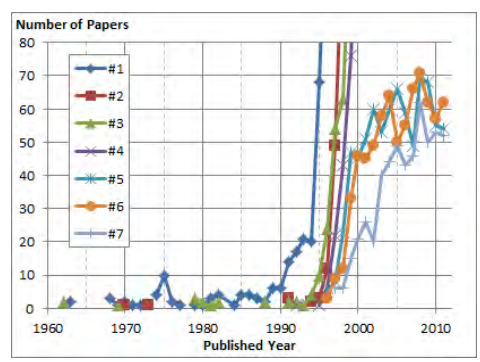


Figure 3 : Number of Papers about mainly #5, #6, and #7

Seeing number of papers about the clusters of #1, #2, #3 and #4, a rapid increase started between 1994 and 2007. The each clusters of #5, #6 and #7 got its first

papers from 1996 or 1997. In these years, Chalfie and Tsien achieved their researches, as in Table 2.
In the last decade, number of papers on #1 and #4 decreased. This is considered that the research topics are changing from #1 and #4 of basic application into specific application such as the use for medicine. Number of papers on #3, that is the basic of basic researches, is still increasing.

Time Transition

In-degree centralities can identify the leading papers. The leading papers tend to have the longer ages from the published year to the year of the highest point of in-degree centrality and the higher value at the highest point.
Figure 4 shows the time transitions of the in-degree centralities, including 1,486 papers published between 1900 and 1998, when the three Nobel Prize winners had already appeared. Judging by Figure 4, the time transition of in-degree centrality brings the leading papers to the surface. One line shows one paper. Blue lines have the top 20% value of times cited between the maximum value and the minimum value, and times cited are derived from the Web of Science at 2012-11-16. Values of times cited decrease in an order of: blue, aqua, green, orange and red, and 20% of the value range are assigned to each colors. In Figure 4, a blue line (b) is one of the defined leading papers, published by Chalfie et al. in 1994, belonging to the cluster #1, cited by 3,627 papers. An aqua line (c) is also one of the defined leading papers, published by Tsien in 1998, belonging to the cluster #3, cited by 2,761 papers. An orange line (a) is the oldest defined leading papers, published by Shimomura in 1962, belonging to the cluster #3, cited by 861 papers, but these in-degree centralities have had valid values since 1974, seeing Figure 4. The reason is that a change of the largest graph component of the citation networks occurs between 1973 and 1974, and that paper of Shimomura is included in the second largest graph component, which is not analyzed in this work. The papers included the largest graph component to 1973 return into the largest graph component from 2006. The keyword and journal trends of these two terms are different as in Table 1, such as “bioluminescent” is used as the keyword of “fluorescent” after 1974. Some papers by Shimomura, who used the keyword “bioluminescent” instead of “fluorescent” in his initial papers, cross over the change of the largest graph component, including the orange line (a).
Figure 5 shows the time transitions of closeness centralities by 1,486 papers published between 1900 and 1998. Figure 6 shows the time transitions of betweenness centralities by 1,460 papers published between 1974 and 1998. The lines of (a) - (c) in Figure 5 and Figure 6 are the same papers as those in Figure 4.
From Figure 4, we can get the hypothesis that the leading papers have gradual summit of in-degree centrality over long years. The cause is that the leading paper could gain newly cited papers every year. Adams (2005) strengthens our hypothesis. Meanwhile, unimportant papers could hardly gain cited papers, so the in-degree centralities become less and less over years. The years of the highest point are: (b) 1997, 1994, 1996, and (c) 2004, 1998, 2000, ordered by in-degree

centrality, closeness centrality and betweenness centrality. Regarding (b) and (c), the betweenness centralities tend to have the highest points before those of the corresponding in-degree centralities, as in Table 2. However, the betweenness centralities don't bring the leading papers to the surface, compared with the in-degree centralities. The most of highest points of closeness centralities occur at the same year of publication. Thus, the following discussions focus on the in-degree centralities. The published year, the years when values become over 3σ and 2σ, and the year of the highest point are: (b) 1994, 1995, 1995, 1997, and (c) 1998, 2000, 2000, 2004. Here, over 3σ or 2σ of values means three or two times more of standard deviations than the arithmetic average. The extraordinarily leading papers, which are definitely separated from others, are found earlier with standard deviation than detecting the highest point. However, the moderately leading papers could not exceed 2σ and 3σ, so our hypothesis about the highest points of in-degree centralities has a role towards the moderately leading papers.

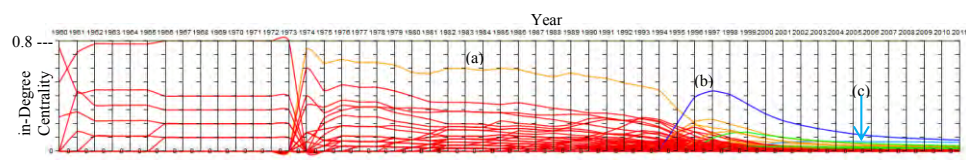


Figure 4: Time Transition of in-Degree Centrality (1,486 Papers Published in 1900-1998; View of 1960 - 2011)

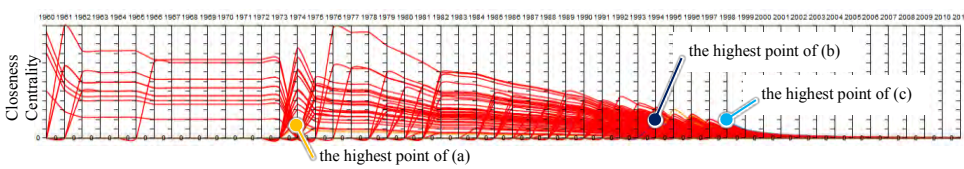


Figure 5: Time Transition of Closeness Centrality (1,486 Papers Published in 1900-1998; View of 1960 - 2011)

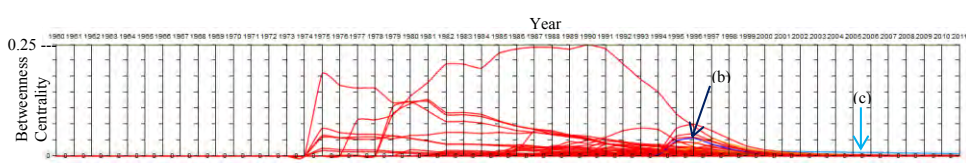


Figure 6: Time Transition of Betweenness Centrality (1,460 Papers Published in 1974-1998; View of 1960 - 2011)

Table 1 shows the top 5 keywords and journals in the two datasets of 1900 - 1973 and 1900 - 1974. The trends of journals are different by a change of the largest graph component between 1973 and 1974. The trends of keywords are also different. The keyword of "fluorescent" appears in the dataset of 1900 - 1973,

whereas the keywords of "luminescent", "bioluminescent" and "photoprotein" are used in the dataset of 1900 - 1974. It means that there is no link between papers about fluorescent proteins and papers about luminescent proteins.

Table 1 : The Top 5 Keywords and Journals before and after a Change of the Largest Graph Component

dataset		*papers published in 1900 - 1973*		*papaers published in 1900 - 1974*	
		keywords	*journals*	*keywords*	*journals*
rank	*1*	fluorescent	IMMUNOLOGY	protein	BIOCHEMISTRY-US
	2	fluorescent protein	CHEM REV	calcium	J CELL COMPAR PHYSL
	3	protein	J PATHOL BACTERIOL	activated	BIOCHEM J
	4	protein tracer	ENDEAVOUR	bioluminescent protein	FED PROC
	5	fluorescent protein tracer	NATURE	aequorin	SCIENCE

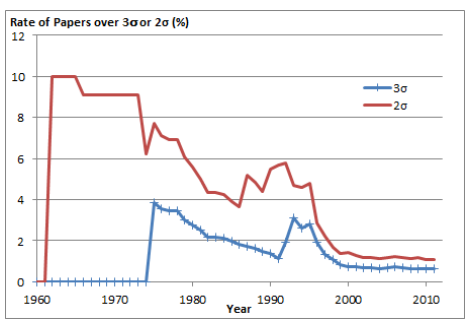


Figure 7 : Rate of Papers having over 3σ or 2σ of in-Degree Centralities

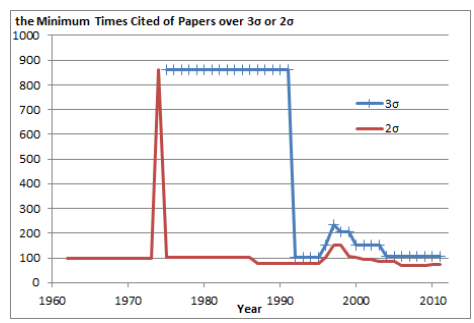


Figure 8 : The Minimum Times Cited of Papers having over 3σ or 2σ of in-Degree Centralities

The Detection of the Leading Papers by Outliers

Some of the top papers have over 3σ or 2σ of the in-degree centralities. Figure 7 shows the rate of papers having over 3σ or 2σ of the in-degree centralities. Figure 8 shows the minimum times cited among papers having over 3σ or 2σ of the in-degree centralities. If papers having over 3σ of the in-degree centralities are defined as the top papers, 0.64% - 0.73% of all papers have been found since 2000, as the blue line in Figure 7. These papers are cited by more than 100 papers as of 2012-11-16, as the blue line in Figure 8. If papers having over 2σ of the in-degree centralities are defined as the top papers, 1.08% - 1.41% of all papers have been found since 2000, as the blue line in Figure 7. These papers are cited by more than 60 papers as of 2012-11-16, as the blue line in Figure 8. On the

contrary, especially for over 3σ of the in-degree centralities in 1980s, more than 1% of papers cited by more than 800 papers are identified. That means that the leading papers by outliers come closer to the average papers.

Table 2 : The List of Papers related to the three Nobel Prize winners (Ehrenberg, 2008; Only Papers having Bibliographic Information available for Analyses)

	Leading Paper	*Times Cited*	*The Highest Point Year*	
			In-Degree	*Betweenness*
(a)	Shimomura, O., Johnson, F.H. and Saiga, Y. (1962) J. Cell. Comp. Physiol. 59 223-240.	861	-	-
-	Shimomura, O. (1979) FEBS Letters 104 220-222.	155	1988	1996
(b)	Chalfie, M. et al. (1994) Science 263 802-805.	3,627	1997	1996
-	Heim, R. et al., (1994) Proc. Natl. Acad. Sci. USA 91 12501-12504.	974	1996	1996
-	Cubitt, A.B. et al. (1995) Trends Biochem. Sci. 20 448-455.	901	1998	1997
-	Ormo, M. et al. (1996) Science 273 1392-1395.	1,171	1999	1997
-	Heim, R. and Tsien, R. (1996) Curr. Biol. 6 178-182.	11	1997	1997
-	Brejc, K. et al. (1997) Proc. Natl. Acad. Sci. USA 94 2306-2311.	378	2000	1998
(c)	Tsien, R. (1998) Annu. Rev. Biochem. 67 509-544.	2,761	2004	2001
-	Miyawaki, A. et al. (1999) Proc. Natl. Acad. Sci. USA 96 2135.	454	2002	2001
-	Baird, G.S. et al. (2000) Proc. Natl. Acad. Sci. 97 11984-11989.	497	2005	2004
-	Gross, L.A. et al. (2000) Proc. Natl. Acad. Sci. USA 97 11990-11995.	312	2010	2005
-	Shaner, N.C. et al. (2004) Nature Biotechnology 22 1562-1572.	1,551	2011	2011
-	Shimomura, O. (2005) Journal of Microscopy 217 3-15.	24	2010	2011
-	Shaner, N.C. et al. (2008) Nature Methods 5 545-551.	191	2011	2011

* Bold year means the earliest year among years of the highest point at each paper.

The Detection of the Leading Papers by the Age to the Highest Point

To confirm our hypothesis about Figure 4, Figure 9 shows average ages from the published year to the year of the highest point for each class of times cited, and Figure 10 shows average heights of the highest point for each class of times cited. In Figure 9 and Figure 10, averages are calculated for each decade to compare

temporal changes. Table 3 shows the coefficient of determination of each decade in Figure 9 and Figure 10. The coefficients of determination for the ages are between 0.3928 and 0.5323, and the coefficients of determination for the heights are between 0.3471 and 0.426. However, the tendency is: the higher times cited the papers have, the longer ages and the higher heights they have. Especially, the tendency appears more clearly on the lower classes of times cited, because those classes have so many papers that outliers influence less, as in Table 4. For example, regarding to papers published in 2001 - 2010, the coefficient of determination for the age is 0.8724 when only the lower classes of 1 - 80 times cited are adopted.

Table 3 : The Coefficient of Determination (For All Classes of Times Cited)

Published Year	*1971 - 1980*	*1981 - 1990*	*1991 - 2000*	*2001 - 2010*
for the Ages (Figure)	0.4323	0.4638	0.3928	0.5232
for the Heights (Figure)	0.3631	0.3917	0.426	0.3471

Table 4 : The Fluctuations of the Coefficient of Determination (Only Papers Published in 2001-2010)

Classes of Times Cited	*1 - 100*	*1 - 200*	*1 - 300*	*1- 400*	*1 - 500*	*1 - 600*
for the Ages (Figure)	0.4812	0.5679	0.6388	0.3976	0.4319	0.5776
for the Heights (Figure)	0.3497	0.5381	0.1804	0.1608	0.1508	0.2591

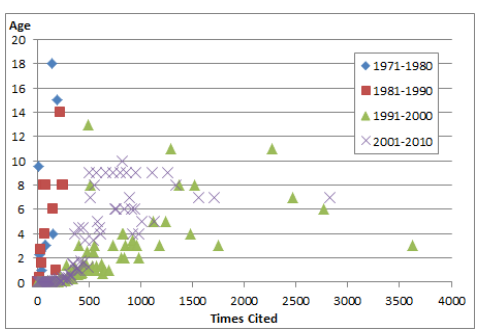


Figure 9 : Average Ages to the Highest Points (Averages are calculated for each 10 times cited.)

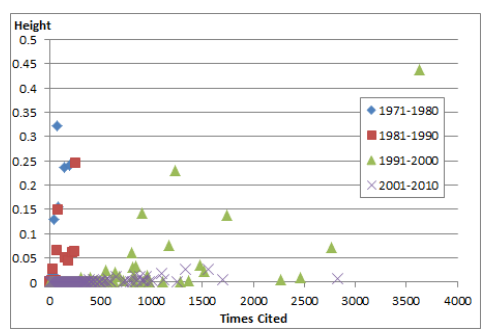


Figure 10 : Average Heights to the Highest Points (Averages are calculated for each 10 times cited.)

Proposition of Indicators to Finding the Papers of Next Generation

The detection by outliers enables to find the leading papers earlier than the detection by the age to the highest point. However, the detection by outliers can't find the moderately leading papers which never exceed 3σ or 2σ. Thus, we proposed the both use of outliers and the age to the highest point to find both the extraordinarily leading papers and the moderately leading papers. According to the above two section, the early detection by outliers catch papers cited by more

than 60 - 100 papers, and the age to the highest point have a strong correlation with times cited within around 80, in the case of 2001 - 2011.

Conclusion

In this paper, we identify the field of fluorescent proteins and analyze the time transitions to find leading papers of the next generation. The bibliographic information of papers about fluorescent proteins is collected from Thomson Reuters' Web of Science.

In the citation network analysis, each citation network is made from citation relations and divided into clusters. In the time transition analysis, the features of the leading papers are extracted, and we find that the in-degree centralities of the leading papers are obviously increasing over several years, because the leading papers have collected citations over years. Meanwhile most of newly published papers can't collect citations, and then their in-degree centralities never have increased. About closeness centrality and betweenness centrality, the feature of the leading papers could not be extracted or is difficult to use the detection of the leading papers.

To quantify the feature of the leading papers, we propose the combined usage of two ways to detect the leading papers. One way is that the extraordinarily leading papers are identified by outliers. Another is that the moderately leading papers are selected by the ages from their publication to the highest point and the height of the highest point of in-degree centrality. This work will contribute to find the candidates of the leading papers.

In the future, we should confirm if these ways are universal and applicable to other academic fields. Then, we had better search a lot of unknown clues to find the leading papers or the promising academic fields.

References

Adams, J (2005). Early citation counts correlate with accumulated impact. *Scientometrics*, 63(3), pp.567–581.

Borner, K., Chen, C. & Boyack, K.W. (2003). Visualizing Knowledge Domains. *Annual Review of Information Science and Technology*, pp.179–255.

Chalfie, M., Yuan, T. & Prasher, D. C. (1993). Glow Worms - A New Method of Looking at C. elegans Gene Expression. Worm Breeder's Gazette 13.

Ehrenberg, M. (2008). *The green fluorescent protein : discovery , expression and development*, Available at: http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2008/advanced-chemistryprize2008.pdf [Accessed November 16 2012].

Garfield, E. (2006). Citation indexes for science. A new dimension in documentation through association of ideas. 1955. *International journal of epidemiology*, 35(5), pp.1123–1127.

Innovation Policy Research Center, the University of Tokyo (2013). Academic Landscape System. Available at: http:// academic-landscape.com / [Accessed April 30, 2013].

Nakamura, H., Suzuki, S., Tomobe, H., Kajikawa, Y. & Sakata, I. (2011). Citation lag analysis in supply chain research. *Scientometrics*, 87(2), pp.221–232.
NetworkX developer Team (2013). NetworkX. Available at: http://networkx.github.com/ [Accessed January 1 2013].
Newman, M.E.J. (2004). Fast Algorithm for detecting community structure in networks. *Physical Review E*, 69(066133).
Shibata, N., Kajikawa, Y., Takeda, Y. & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, 28(11), pp.758–775.
Shibata, N. et al., 2009. Comparative study on methods of detecting research fronts using different types of citation. *J. Am. Soc. Inf. Sci. Technol.*, 60(3), pp.571–580.
Shimomura, O., Johnson, F.H. & Saiga, Y. (1962). Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *J. Cell. Comp. Physiol.* 59: 223–239.
Tsien, R.Y. (1998). The green fluorescent protein. *Annual Review of Biochemistry*, 67, pp.509-544.
Vanclay, J.K. (2011). Impact factor: outdated artefact or stepping-stone to journal certification? *Scientometrics*, 92(2), pp.211–238.

DIFFERENCES AND SIMILARITIES IN USAGE VERSUS CITATION BEHAVIOURS OBSERVED FOR FIVE SUBJECT AREAS

Juan Gorraiz[1], Christian Gumpenberger[1] and Christian Schlögl[2]

[1]*juan.gorraiz@univie.ac.at; christian.gumpenberger@univie.ac.at*
University of Vienna, Vienna University Library, Bibliometrics Department,
Boltzmanngasse 5, A-1090 Vienna (Austria)

[2]*christian.schloegl@uni-graz.at*
University of Graz, Institute of Information Science and Information Systems,
Universitätsstr. 15, A-8010 Graz (Austria)

Abstract
This study puts an emphasis on the disciplinary differences observed for the behaviour of citations and downloads. This was exemplified by means of 5 selected fields, namely "Arts and Humanities", "Computer Science", "Economics, Econometrics and Finance", "Oncology" and "Psychology", for the last 10 years. Differences in obsolescence characteristics were studied using synchronic as well as diachronic counts. Furthermore, differences between document types were taken into consideration and correlations between journal impact and journal usage measures were calculated.
The results show that the diachronic timelines for downloads are very similar for all subject categories, namely a steady and steep curve progression, and corroborate the rapid acceptance of electronic journals, which have speeded up the process of scholarly communication in the last decade. Synchronic trend lines are very similar as well. Here the first two years post publication account for the highest downloads and need to be taken into account for the calculation of a solid journal usage factor. On the contrary to downloads, diachronic and synchronic citation timelines differ considerably from one field to the other.
Usage metrics should consider the special nature of downloads and ought to reflect their intrinsic differences to citations. Moreover, they should also incorporate the characteristics of document types evolved from the digital era like "Articles in Press".

Keywords
Downloads, citations, usage metric, citation metric, obsolescence, synchronic, diachronic

Conference Topic
Scientometrics Indicators (Topic 1)

1. Introduction

In the course of the steadily increasing popularity of electronic journals the tracking and collection of usage data has become much easier compared to the print-only era. Thanks to the global availability of e-journals it is now possible to observe scholarly communication also from the reader's perspective (Rowlands and Nicholas, 2007). In comparison to citation data, usage data have apparent advantages like easier and cheaper data collection, earlier availability, and the reflection of a broader usage scope (Bollen et al., 2005; Brody, Harnad and Carr, 2006; Duy and Vaughan, 2006; Haustein, 2011). Several usage indicators have been suggested in recent years. Most of them are based on the classical citation indicators from the Journal Citation Reports (JCR), using download data (usually full-text article requests) instead of citations. The corresponding usage metrics are "usage impact factor" (Rowlands and Nicholas, 2007; Bollen and Van de Sompel, 2008), "usage immediacy index" (Rowlands and Nicholas, 2007) or "download immediacy index" (Wan et al., 2008), and "usage half-life" (Rowlands and Nicholas, 2007).

The authors of this study have already performed a few analyses focusing on usage data for oncology and pharmacology journals provided by ScienceDirect (Schloegl and Gorraiz, 2010; Schloegl and Gorraiz, 2011). Major outcomes were as follows:

- strong increase in the usage of e-journals for ScienceDirect journals from the fields of oncology and pharmacology between 2001 and 2006
- significant correlation between article downloads and citation frequencies at journal level, which were slightly lower at article level
- medium to high correlation between relative indicators (usage impact factor and Garfield's impact factor)
- unequally observed obsolescence characteristics: the download half-lives amounted to approximately 2 years, whereas the cited half-lives were three times higher on average.

In this study particularly the following issues have been addressed:

- comparison of download and citation frequencies at category level: disciplinary differences exemplified by means of 5 selected fields, namely "Arts and Humanities", "Computer Science", "Economics, Econometrics and Finance", "Oncology" and "Psychology"
- disciplinary differences in obsolescence characteristics between citations and downloads using synchronic and diachronic counts
- differences between document types
- comparison and correlations between different journal impact and journal usage measures.

2. Methodology and data

2.1. Data

All data were provided within the scope of the Elsevier Bibliometric Research Program (EBRP) 2012. The analysed data pool includes usage data for the 5 ScienceDirect categories "Arts and Humanities" (37 journals), "Computer Science" (150 journals), "Economics, Econometrics and Finance" (133 journals), "Oncology" (42 journals) and "Psychology" (9 journals).

The following data from ScienceDirect have been used at journal level (all for the period 2002-2011):

- total number of downloadable items for each year
- number of downloadable items disaggregated by document types for each year
- download counts disaggregated by document types for each download year as well as for each publication year available within the given time period
- corresponding citation counts from Scopus for each citation year and disaggregated by the various publication years (from citation year back to 2002).

2.2. Analyses at category level

All journals within a subject category were aggregated and considered as "one big journal". That way the number of all downloads within the category and the number of citations to all journals in the category were taken into account. Resulting values are averages per document.
Used metrics were applied at synchronic (= reference point for the calculation is the download or citation year) as well as at diachronic level (= reference point for the calculation is the publication year addressing subsequent citation or download years).
Timelines for downloads per item as well as for citations per item have been provided in order to study the occurring obsolescence patterns. The common document types in ScienceDirect - articles, reviews, conference papers, editorial materials, letters, notes, and short communications – were differentiated accordingly. Notes and Research Notes could not be distinguished. In addition the evolution of AIPs (Articles in Press) was analysed. Correlations between downloads and citations were calculated at synchronic as well as at diachronic level for each of the 5 ScienceDirect categories using Spearman's correlation coefficient.

2.3. Correlation between journal usage and journal impact indicators

Due to the fact that the majority of downloads are effectuated in the current and subsequent years of publication (Schlögl and Gorraiz, 2010), the use of a usage impact factor relying on the same time window as the impact factor is flawed. It is rather suggested to deploy a "journal usage factor" (JUF), which not only reflects the two retrospective years but also includes the current reference year. The JUF is therefore defined as the number of downloads in the reference year from journal items published in this year as well as in the previous two years divided by the number of items published in these three years. In contrast to the so far usual two year time window, this three year time interval allows for a significant amount of downloads in most of the cases (Gorraiz and Gumpenberger, 2010). Correspondingly an adapted version of "Garfield's Impact Factor" (GIF) is used in this study considering also the year of reference along with the previous two years. This indicator is labelled as "total impact factor" (TIF), as it also includes the "immediacy index".

In order to test the stability of the above defined journal usage factor (JUF(2)), we calculated also versions of this indicator with longer time windows:

- JUF(5) = number of downloads in 2010 to documents published in the years 2010-2005 divided by the number of documents published in 2010-2005 (reference year plus 5 years window)
- JUF(8) = number of downloads in 2010 to documents published in the years 2010-2002 divided by the number of documents published in 2010-2002 (reference year plus 8 years window).

Equivalently and using citations instead of downloads, we calculated TIF(2), TIF(5) and TIF(8). GIF(2) and GIF(5) correspond to Garfield's Journal Impact Factor for 2 and 5 years, respectively without consideration of the first year (= reference year) but including all document types. GIF (8) is an extension of Garfield's Impact factor to all the data available (till 2002).
Correlations were then performed for all journals comprised in each category.

3. Results and discussion

3.1. Synchronic counts: timelines of downloads and citations per document (item)

The timelines of downloads and citations are comparatively shown for all 5 categories in Figures 1-5 below. The x-axis always represents the publication years of the downloaded/cited documents, whereas the multi-coloured lines represent the different download/citation years.

3.1.1. Arts and Humanities

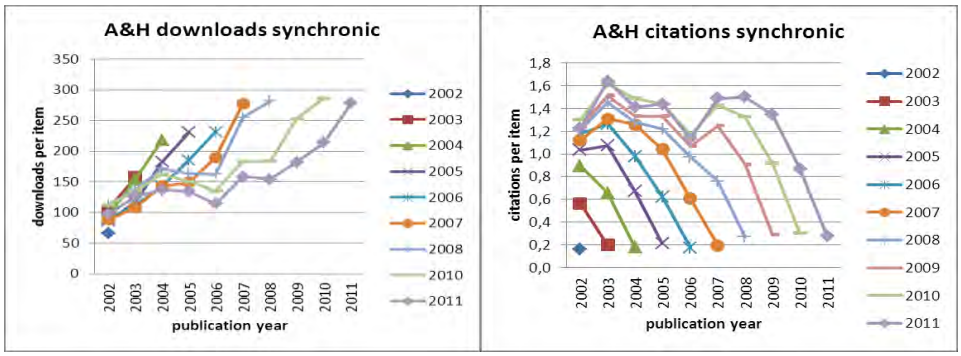


Figure 1. Timelines of downloads vs. citations (synchronic counts) in Arts & Humanities (n=37 journals)

3.1.2. Computer Science

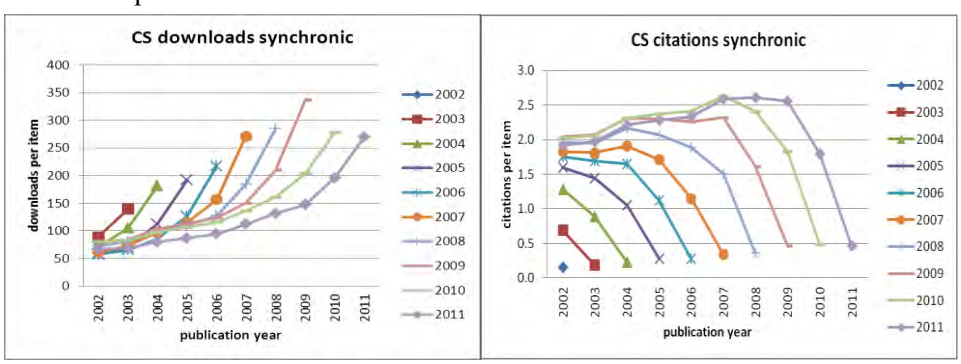


Figure 2: Timelines of downloads vs. citations (synchronic counts) in Computer Science (n=150 journals)

3.1.3. Economics, Econometrics and Finance

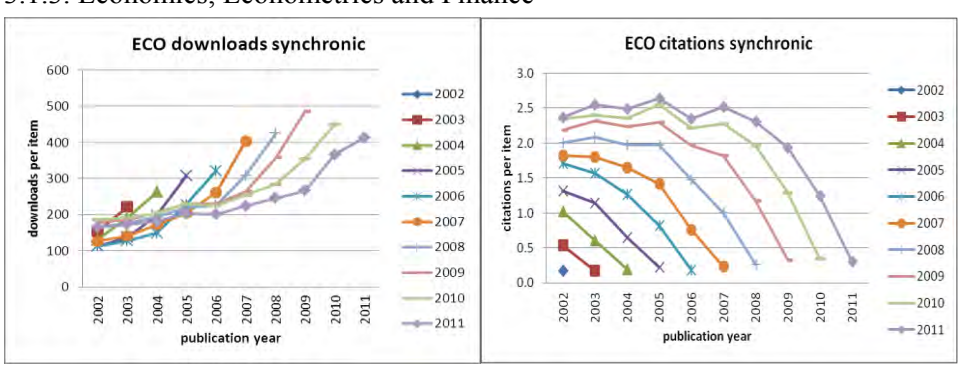


Figure 3: Timelines of downloads vs. citations (synchronic counts) in Economics, Econometrics and Finance (n=133 journals)

3.1.4. Oncology

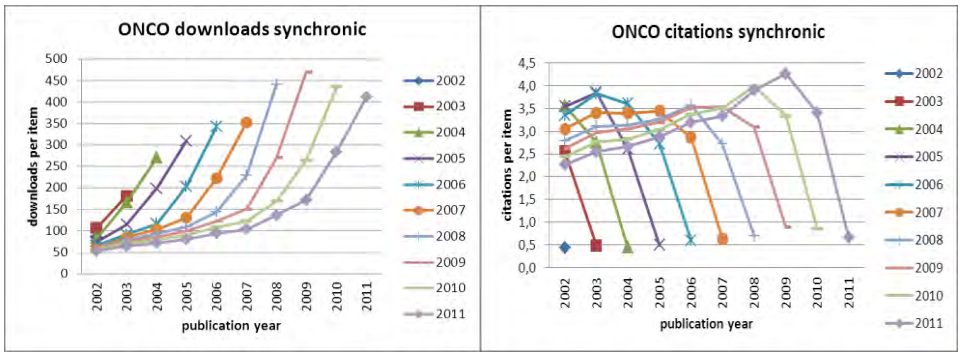


Figure 4: Timelines of downloads vs. citations (synchronic counts) in Oncology (n=42 journals)

3.1.5. Psychology

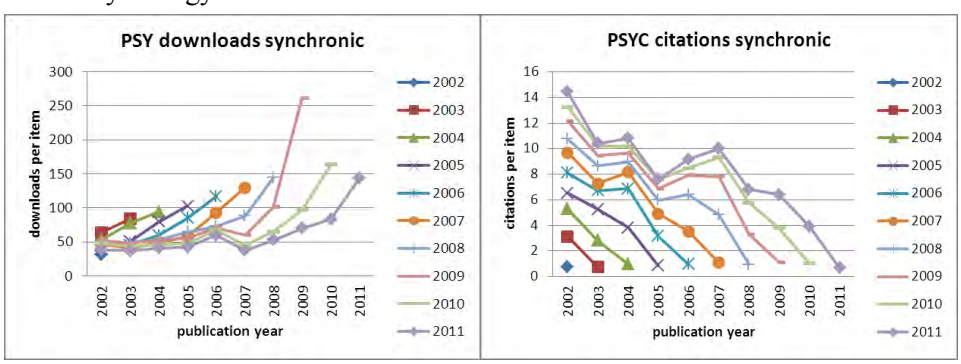


Figure 5: Timelines of downloads vs. citations (synchronic counts) in Psychology (n=9 journals)

Considering downloads, similar trend lines can be observed for all 5 categories. They have also in common that the first two years post publication account for the highest downloads. Disciplinary differences only occur regarding the absolute download values, as illustrated by the different values of the y-axis in Figures 1 to 5.

Synchronic citation counts differ also in their development from discipline to discipline. For Oncology, the citation maximum is reached two years after publication, followed by a decrease afterwards. For Computer Science this interval increases to 3-4 years, and for Economics, Econometrics and Finance even to 5-6 years. After these intervals, stagnation rather than a decrease can be observed. For Arts & Humanities this interval is overall longer, for Psychology it is probably more than 10 years.

3.2. Diachronic counts: timelines of downloads and citations per document

The timelines of downloads and citations are comparatively shown for all 5 categories in Figures 6-10 below. The x-axis always represents the download/citation years of the downloaded/cited documents, whereas the multi-coloured lines represent the different publication years.

3.2.1. Arts and Humanities

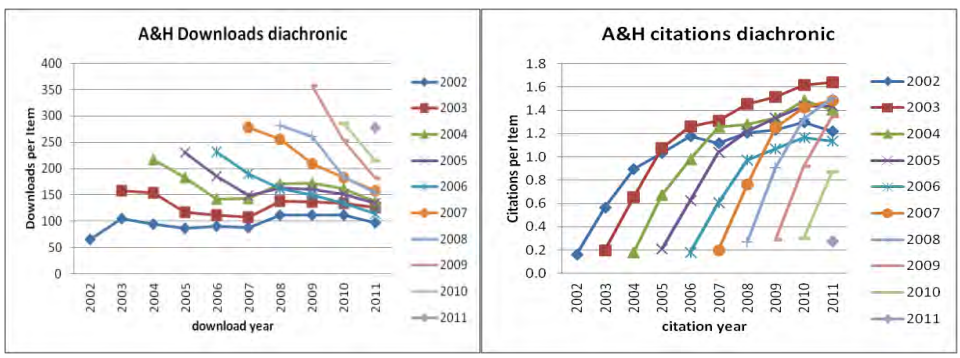


Figure 6. Timelines of downloads vs. citations (diachronic counts) in Arts & Humanities (n=37 journals)

3.2.2. Computer Science

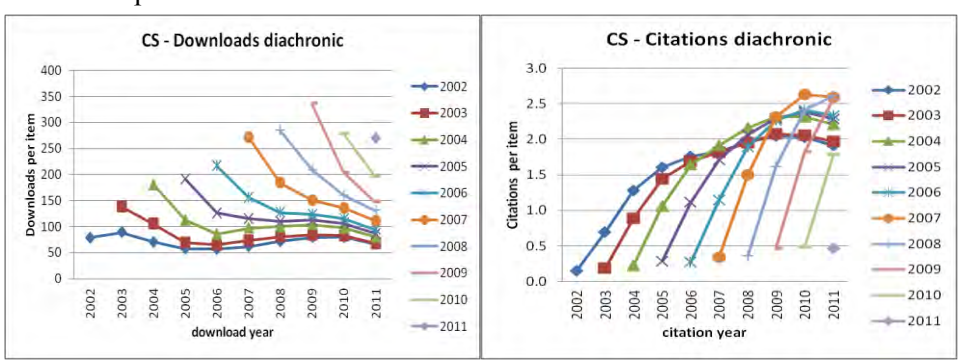


Figure 7: Timelines of downloads vs. citations (diachronic counts) in Computer Science (n=150 journals)

3.2.3. Economics, Econometrics and Finance

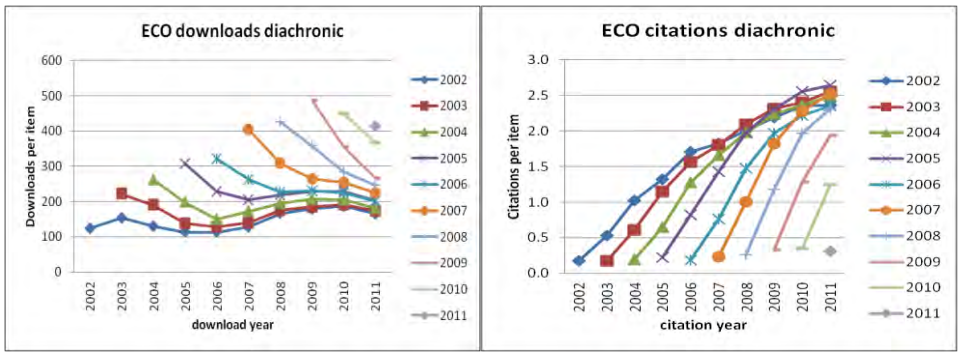


Figure 8: Timelines of downloads vs. citations (diachronic counts) in Economics, Econometrics and Finance (n=133 journals)

3.2.4. Oncology

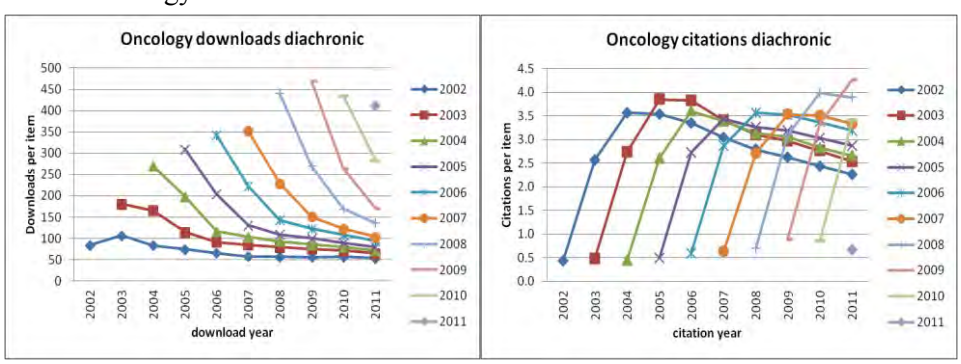


Figure 9: Timelines of downloads vs. citations (diachronic counts) in Oncology (n=42 journals)

3.2.5. Psychology

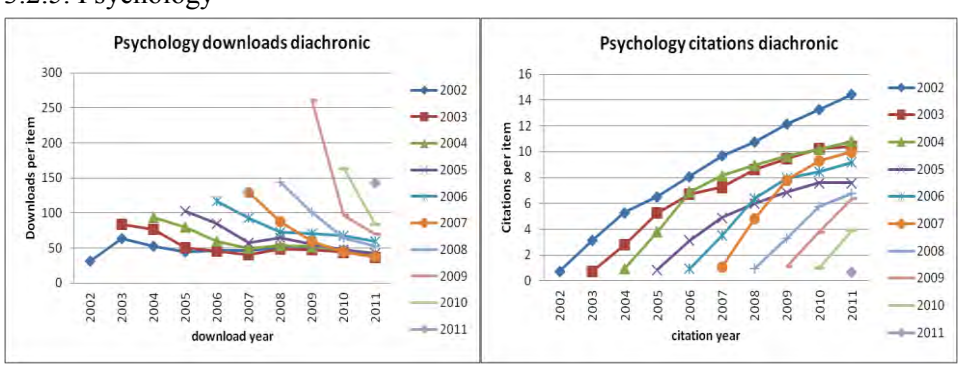


Figure 10: Timelines of downloads vs. citations (diachronic counts) in Psychology (n=9 journals)

Considering downloads, the results do not differ for the diachronic counts. The trend lines show a very similar run for all 5 analyzed subject categories, namely a steady and steep curve progression.
Higher download averages have been identified for Oncology and Economics, Econometrics and Finance (see Fig. 8 and 9), with maximum values between 450 and 500 in 2009 for publications of the same year, followed by Computer Science and Arts &Humanities (see Fig. 6 and 7) with maximum values between 300-350 in 2009 for publications of the same year), and finally by Psychology (see Fig. 10) with an outlier reaching 250 in 2009 for publications of the same year.

For citations, the results from diachronic counts show different obsolescence patterns depending on the research field. There is a steady increase in citations within the first 10 years for Arts & Humanities (Fig. 6), Economics, Econometrics and Finance (Fig. 8) as well as for Psychology (Fig. 10). Whereas in Computer Science (Fig. 7) stagnation occurs after the first 6 to 7 years for the older articles (2002-2004). For the other years, data availability is too sparse for a solid evidence. Oncology (Fig. 9) is the only exception where a decrease can be observed after the second year.

Average citation frequency is also different for the various categories. Average counts are below 2 for Arts & Humanities, below 3 for Computer Science and Economics, Econometrics and Finance and below 4.5 for Oncology. Rather surprising are the higher averages for Psychology, even reaching 14 citations in the citation year 2011 for publications of the year 2002.

3.3. Diachronic counts for different document types: timelines of downloads and citations per document

The diachronic count mode with the fixed publication years gives a good picture of the citation and download trends for each document type over the last 10 years. Their timelines (aggregated for all 5 subject categories) can be seen in Figures 11-16 below. The x-axis always represents the download/citations years of the downloaded/cited documents, whereas the multi-coloured lines represent the different publication years.

3.3.1. Articles

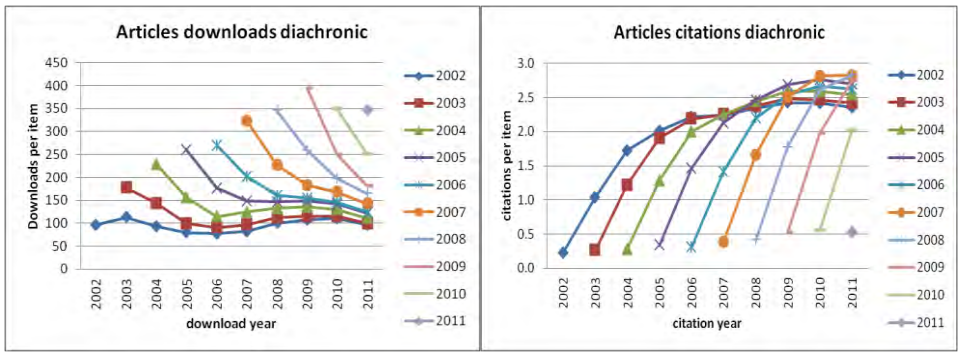


Figure 11: Timelines of downloads vs. citations (diachronic counts) for articles

3.3.2. Reviews

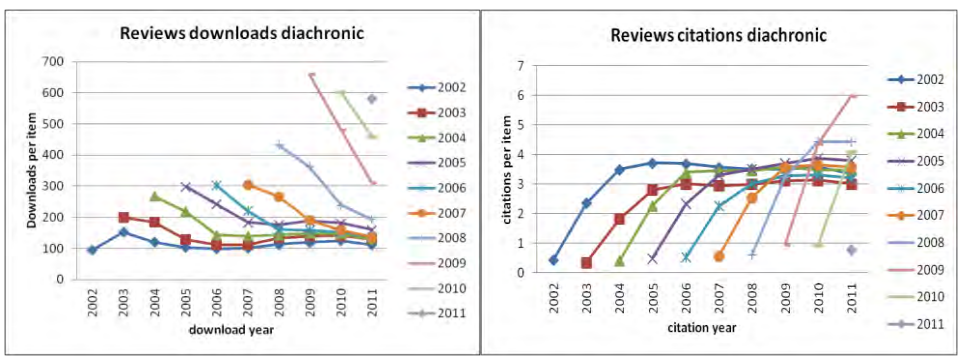


Figure 12: Timelines of downloads vs. citations (diachronic counts) for reviews

3.3.3. Conference Proceedings

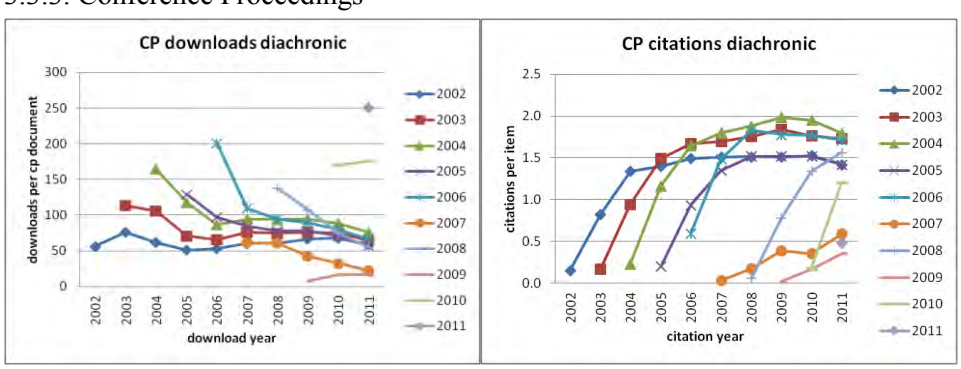


Figure 13: Timelines of downloads vs. citations (diachronic counts) for conference proceedings

3.3.4. Editorials

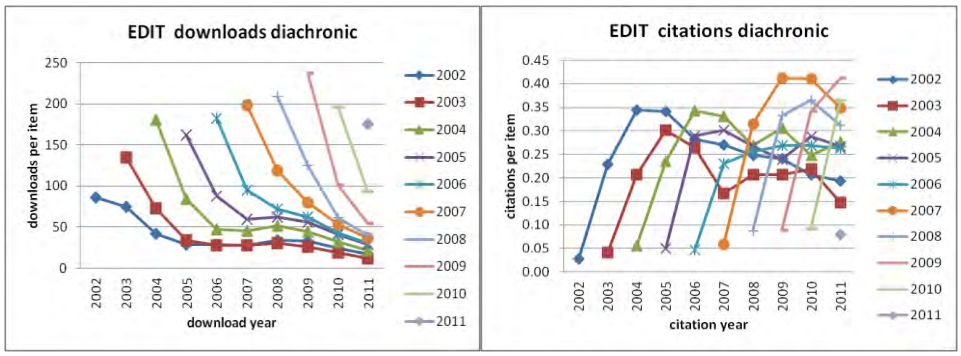


Figure 14: Timelines of downloads vs. citations (diachronic counts) for editorials

3.3.5. Letters

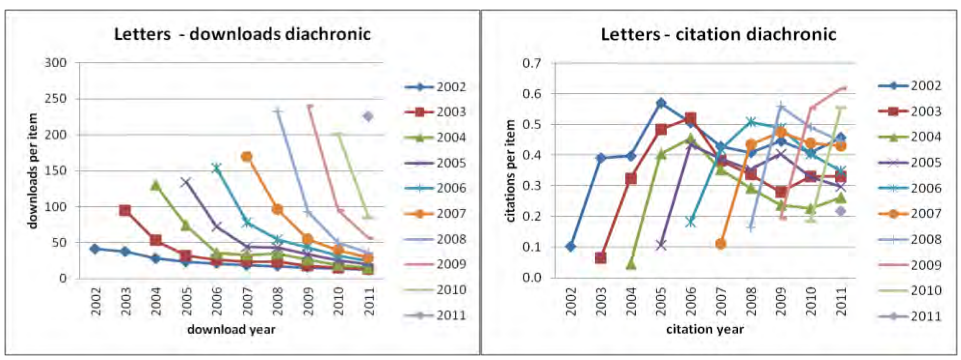


Figure 15: Timelines of downloads vs. citations (diachronic counts) for letters

3.3.6. Short communications and notes

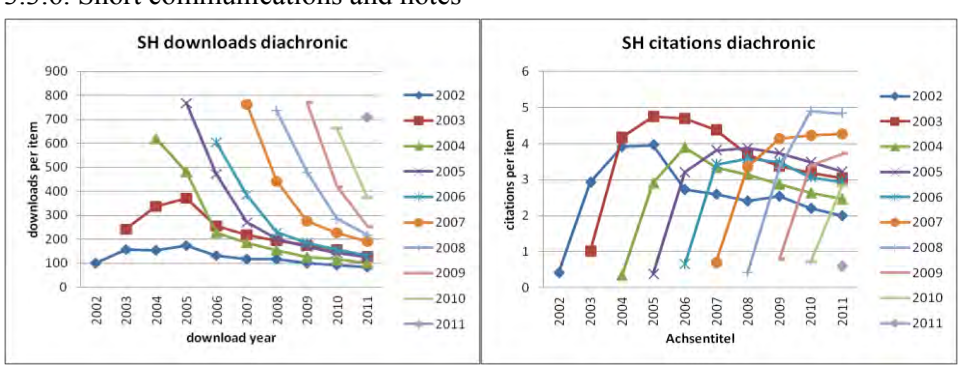


Figure 16: Timelines of downloads vs. citations (diachronic counts) for short communications

Figures 11 to 16 show very similar download timelines for all document types. The number of downloads of Review Articles is about twice as high as of Articles for the last 3 years (2009-2011). Articles in turn are downloaded almost twice as often as Letters. The timeline results for Short Communications are similar to the ones observed for Letters, with the difference that the latter document type is approximately three times less often downloaded. The availability of Notes was restricted and therefore the obtained results were too sparse to be presented here.
Citation timelines are all similar for Articles, Review Articles and Conference Proceedings, showing a steady increase at the beginning and reaching stagnation after a while. On the one hand, Review Articles accrue clearly more citations than Articles, on the other hand they reach the stagnation phase earlier. Conference Proceedings remain less cited than Articles. Editorials and Letters are mostly cited within the first 3 years after publication, although at a very low level.

Table 1. Evolution of Articles in press (AIPs) and their download frequencies for each category (2007-2011).

Subject category	*PY*	*# AIPs*	*# downloads*	*downloads/AIP*
A&H	2007	4	36	9.0
	2008	**20**	**4888**	**244.4**
	2009	**2**	**0**	**0.0**
	2010	49	6021	122.9
	2011	88	12195	138.6
	TOTAL	**163**	**23140**	**142.0**
CS	2007	1	41	41.0
	2008	**160**	**34369**	**214.8**
	2009	**33**	**1878**	**56.9**
	2010	578	101614	175.8
	2011	1800	194306	108.0
	TOTAL	**2572**	**332208**	**129.2**
ECON	2007	3	147	49.0
	2008	**68**	**31665**	**465.7**
	2009	**16**	**367**	**22.9**
	2010	127	26543	209.0
	2011	650	125440	193.0
	TOTAL	**864**	**184162**	**213.2**
ONCO	2008	**54**	**23523**	**435.6**
	2009	**1**	**403**	**403.0**
	2010	169	35369	209.3
	2011	652	109670	168.2
	TOTAL	**876**	**168965**	**192.9**
PSYCH	2008	**2**	**836**	**418.0**
	2011	9	1245	138.3
	TOTAL	**11**	**2081**	**189.2**

3.3.7. Articles in press

Data about "Articles in press" (AIPs) were only available from 2007 onwards. Their growth and the evolution of their download rates are represented in Table 1. Although values in red suggest inconsistencies in the data provided by Elsevier for the years 2008 and 2009, this analysis proves an overall growth in number and download frequencies of AIPs.

3.4. Correlations between synchronic and diachronic downloads and citations (absolute values) at journal level for each category

Spearman correlations between the total number of downloads and the total number of citations were calculated for each publication year (diachronic mode) as well as for each download/citation year (synchronic mode) for all journals with almost complete data for the interval 2002-2011 (see Table 2).

The diachronic count mode, considering the total number of citations and downloads for each publication year, should be the most appropriate way to determine the strength of the correlation between downloads and citations at journal level.

Table 2. Correlations (Spearman) between total number of downloads and total number of citations.

Diachronic:		***Publication years***									
Subject category	***Journals***	***2002***	***2003***	***2004***	***2005***	***2006***	***2007***	***2008***	***2009***	***2010***	***2011***
A&H	30	0.6	0.85	0.83	0.8	0.84	0.78	0.82	0.82	0.86	0.64
CS	127	0.77	0.79	0.82	0.82	0.87	0.86	0.88	0.9	0.86	0.83
ECON	83	0.83	0.88	0.88	0.89	0.9	0.92	0.91	0.87	0.88	0.84
Onco	31	0.77	0.8	0.85	0.92	0.93	0.95	0.95	0.95	0.95	0.94
PSYCH	8	0.29	0.19	0.17	0.45	0.33	0.36	0.45	0.43	0.33	0.14
Synchronic:		***Download/citation years***									
Subject category	***Journals***	***2002***	***2003***	***2004***	***2005***	***2006***	***2007***	***2008***	***2009***	***2010***	***2011***
A&H	30	0.42	0.59	0.72	0.76	0.77	0.81	0.78	0.8	0.86	0.86
CS	126	0.57	0.65	0.69	0.75	0.8	0.8	0.83	0.85	0.85	0.86
ECON	84	0.49	0.77	0.82	0.85	0.9	0.87	0.89	0.89	0.89	0.88
Onco	31	0.51	0.56	0.77	0.79	0.82	0.86	0.85	0.92	0.95	0.96
PSYCH	7	-0.23	-0.21	0	0.04	0.04	0	-0.04	0	0	0

However, the download and citation windows need to be long enough. Thus, significant correlations were expected for the former publication years (2002 and 2003), where the citation windows are large enough. Nevertheless, the results presented in Table 2 are not in agreement with this assumption. The reason might be the strong increase in e-journal usage between 2003 and 2009, consolidating

afterwards, causing a certain distortion of download counts in the transition years. The same assumption holds true for the synchronic correlations reported for the latter years (for instance, 10 years window in 2011), which are considerably higher for all subject areas besides Psychology than the diachronic ones for 2002, the corresponding year with the largest citation/download window (10 years).
In spite of all these observations, high "diachronic" correlations were observed for Economics, Econometrics and Finance as well as for Oncology. Correlations were also high for Computer Science and for Arts & Humanities, but only very low for Psychology. Also for the synchronic correlations the highest values can be observed for Oncology and Economics, Econometrics and Finance, followed by Computer Science and Arts & Humanities, whereas they were inexistent for Psychology.

3.5. Correlations between JUF, TIF and GIF at journal level for each category

Spearman correlations among JUF, TIF and GIF were compiled for the year 2010 for each journal with almost complete data availability for the interval 2002-2011 in each category (see Table 3). The same correlations were furthermore calculated for the year 2011 with no appreciable differences.

Table 3 shows that the application of different time windows (2, 5 or 8 years) has nearly no influence, since the corresponding correlations are all very high. Furthermore, it makes nearly no difference whether the reference year is considered (TIF) or not (GIF) when calculating the impact factor. Significant correlations between JUF and TIF were observed in all subject categories except Psychology.

Table 3. Correlations (Spearman) between JUF, TIF and GIF for the year 2010.

Correlations between		***A&H***	***CS***	***Econ***	***Onco***	***Psych***
JUF(2)	JUF(8)	0.98	0.96	0.99	0.98	1
TIF(2)	TIF(8)	0.97	0.94	0.93	0.98	0.93
GIF(2)	GIF(8)	0.97	0.94	0.93	0.97	0.95
JUF(2)	JUF(5)	0.98	0.98	0.99	0.99	1
TIF(2)	TIF(5)	0.97	0.97	0.96	0.98	0.98
GIF(2)	GIF(5)	0.97	0.97	0.96	0.97	1
JUF(5)	JUF(8)	1	0.99	1	0.99	1
TIF(5)	TIF(8)	0.99	0.99	0.99	0.99	0.97
GIF(5)	GIF(8)	0.99	0.99	0.99	0.99	0.95
JUF(8)	**TIF(8)**	**0.74**	**0.67**	**0.79**	**0.75**	**0.52**
JUF(5)	**TIF(5)**	**0.72**	**0.66**	**0.79**	**0.77**	**0.52**
JUF(2)	**TIF(2)**	**0.65**	**0.6**	**0.73**	**0.77**	**0.58**
GIF(8)	TIF(8)	1	0.99	1	0.99	0.98
GIF(5)	TIF(5)	1	1	1	0.99	0.98
GIF(2)	TIF(2)	0.98	0.98	0.99	1	1

4. Conclusions

For all 5 subject categories, the results of this study corroborate in most instances the findings of previous analyses by Schloegl and Gorraiz (2010, 2011), who already observed a significant increase in the usage of ScienceDirect e-journals in oncology and also pharmacology in the period 2002-2006.

The diachronic count mode with fixed publication years is more suitable to analyze the overall increase in e-journals usage over time. The trend lines for downloads are very similar for all 5 analyzed subject categories. The steady and steep curve progressions illustrate the rapid adoption of electronic journals by the research community, which has definitely speeded up the process of scholarly communication in the last decade.

Results from synchronic download counts have proven that the first two years post publication account for the highest downloads. The exclusion of the reference year is therefore no longer arguable for the sound construction of the journal usage factor.
Usage metrics should consider the special nature of downloads and ought to reflect their intrinsic differences to citations. In citation metrics, the common non-consideration of the "immediacy year" (like for Garfield's Impact Factor and almost all journal impact measures like SJR or SNIP) is well-grounded in the existing citation delay. This is also confirmed by our study since the "inclusion" of the "immediacy index" in the impact factor (GIF) (which we named TIF) did not result in considerable changes in any of the disciplines.

Contrary to the download analyses, the results for diachronic and synchronic citation counts reveal not only rather different obsolescence patterns depending on the research field, but also different citation frequencies.
Regarding document types, the time lines of downloads are very similar in general. Differences only occur in the download rates per document type. For citations, similarities only exist between Articles, Review Articles and Conference Papers. Average citation frequencies differ from document type to document type. Review Articles are overall more cited than Articles but they reach the stagnation phase earlier.
The correlations between impact and usage factors were lower than those between the absolute values. Furthermore, the obtained results of this study suggest that different time windows for the calculation of JUF, TIF or GIF seem to be indiscriminative. The high correlations observed for GIF(2) and GIF(5) are in agreement with the ones comparatively calculated in the 2010 editions of the Journal Citation Reports (JCR) for Oncology (0.99; 145 journals), Computer Sciences (0.92; 395 journals), Psychology (0.96; 370 journals) and Economics & Business- Finance (0.95; 237 journals). The correlation for all the journals of the JCR-SCI edition (6717 journals) was 0.97, and for the overall JCR-SSCI edition (1995 journals) 0.94.

A new document type evolved from the digital era. “Articles in Press” have become more and more common in recent years and are particularly interesting regarding usage metrics. They could even play an important role to project future downloads or even citations. However, further analyses at publication level are required to gain more insight to underpin this argument.

Overall the results obtained for Psychology need to be taken with a pinch of salt due to the small sample size of only 13 journals. Outliers may possibly skew the calculated averages in this group.

Acknowledgments

This paper is partly based on anonymous ScienceDirect usage data and/or Scopus citation data kindly provided by Elsevier within the framework of the Elsevier Bibliometric Research Program (EBRP).

References

Bollen, J.; Van de Sompel, H.; Smith, J.A. & Luce, R. (2005). Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management,* 41, 1419-1440. online available at URL: http://public.lanl.gov/herbertv/papers/ipm05jb-final.pdf [26 November 2008].

Bollen, J. & Van de Sompel, H. (2008). Usage impact factor: the effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1), 136-149.

Brody, T.; Harnad, S.; Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8), 1060-1072.

Duy, J. & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. *The Journal of Academic Librarianship*, 32(5), 512-517.

Gorraiz, J. & Gumpenberger, C. (2010). Going beyond citations: SERUM – a new tool provided by a network of libraries. *Liber Quarterly* 20, 80-93.

Haustein, S. (2011). Taking a multidimensional approach toward journal evaluation. *Proceedings of the ISSI Conference*, Durban, South Africa, 04-07 July, Vol. 1, 280-291.

Moed, H.F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56(10), 1088-1097.

Rowlands, I. & Nicholas, D. (2007). The missing link: journal usage metrics. *Aslib Proceedings*, 59(3), 222-228.

Schloegl, C. & Gorraiz, J. (2010): Comparison of citation and usage indicators: the case of oncology journals. *Scientometrics*, 82(3), 567-580.

Schloegl, C. & Gorraiz, J. (2011): Global Usage versus Global Citation Metrics: The Case of Pharmacology Journals. *Journal of the American Society for Information Science and Technology*, 61(1), 161-170.

Wan, J.-K., Hua, P.-H., Rousseau, R. & Sun, X.-K. (2010), The download immediacy index (DII): experiences using the CNKI full-text database, *Scientometrics*, 82(3), 555-566.

DIFFERENCES IN CITATION IMPACT ACROSS COUNTRIES

Pedro Albarrán[1], Antonio Perianes-Rodriguez[2] and Javier Ruiz-Castillo[3]

[1] *albarran@merlin.fae.ua*
Universidad de Alicante, Departamento de Fundamentos de Análisis Económico, Campus de San Vicente, 03080 Alicante (Spain)

[2] *antonio.perianes@uc3m.es*
Universidad Carlos III de Madrid, Department of Library and Information Science, SCImago Research Group, C/Madrid, 128, 28903 Getafe, Madrid (Spain)

[3] *jrc@eco.uc3m.es*
Universidad Carlos III de Madrid, Departamento de Economía, C/Madrid, 126, 28903 Getafe, Madrid (Spain)

Abstract
Recent results indicate that the ranking of research units that focus on the upper tail of citation distributions is quite similar to the ranking one obtains with average-based indicators. This paper explores the conjecture that in international comparisons this can be explained because differences in country citation distributions have a strong scale factor component. If this is the case, it is argued that the effect on overall citation inequality that these differences cause should be drastically reduced when raw citation counts are normalized with the countries' mean citations. This is what we find for Physics and the all-sciences case, and a partition of the world into 36 countries and two residual geographical areas. We use a large Thomson Reuters dataset of articles published in 1998-2003 with a five-year citation window. We conclude that international comparisons in terms of countries' mean citations appear to capture most of the differences over the entire support of country citation distributions.

Conference Topic
Scientometrics Indicators: Criticism and new developments (Topic 1)

Introduction
It is well known that citation distributions are highly skewed in the sense that a large proportion of articles get none or few citations while a small percentage of them account for a disproportionate amount of all citations (see *inter alia* Seglen, 1992, Shubert *et al.*, 1987, Glänzel, 2007, Albarrán and Ruiz-Castillo, 2011, and Albarrán *et al.*, 2011a). In this situation, the mean –or any central-tendency statistic– may not provide a good representation of the citation distribution. Consequently, "*colleagues have begun a search to find other indicators that do not depend on averages*" (Rousseau, 2012). For example, Tijssen *et al.* (2002) and Tijssen and van Leeuwen (2006) argue that the top-10% of papers with the

highest citation counts in a publication set can be considered highly cited. Consequently, well established institutions –such as the CWTS of Leiden University, and SCImago– have recently started to rank research units in terms of scientific excellence using the $PP_{top\ 10\%}$ indicator, defined as the percentage of an institution's scientific output included into the set formed by the 10% of the most cited papers in their respective scientific fields.

In this paper we confront the following surprising results obtained in two of the first contributions that have applied this indicator to a large body of data, namely, Waltman *et al.* (2012), and Albarrán and Ruiz-Castillo (2012). Waltman *et al.* (2012) discuss the 2011/2012 edition of the Leiden Ranking of 500 universities world-wide. The citation impact indicators of universities in this edition include the $PP_{top\ 10\%}$ indicator, and the Mean Normalized Citation Score (MNCS hereafter). Albarrán and Ruiz-Castillo (2012) evaluate the citation impact of a partition of the world into 39 countries and eight geographical areas using a Thomson Reuters dataset consisting of the 4.4 million articles published in 1998-2003, and the citations they receive during a five-year citation window for each year in that period. The research impact of the 47 countries and geographical areas in each of the 20 natural sciences and the two social sciences distinguished by Thomson Reuters are compared in terms of the $PP_{top\ 10\%}$ indicator and the mean citation (MC hereafter).

We expected large differences between the results obtained using these two indicators. However, Waltman *et al.* (2012) find that there is a strong, more or less linear relationship between the $PP_{top\ 10\%}$ and the MNCS (Figure 2 in that paper). However, as Waltman *et al.* (2012) emphasize, the two indicators are very different in other important respects: in particular, the $PP_{top\ 10\%}$ indicator, but not the MNCS, is robust to extreme observations with a dramatically large number of citations. Similarly, Albarrán and Ruiz-Castillo (2012) find that the correlation coefficient between MC and $PP_{top\ 10\%}$ taking together the results for the 22 fields is 0.933. Thus, to a first approximation, the two indicators lead to rather similar results.

In principle, differences in resources, intellectual traditions, organization, incentives and many other factors determine the characteristics of citation distributions in any science in every country. How is it possible that, in spite of these differences across countries, an average-based indicator, and an indicator that focus on the upper tail of citation distributions give similar results? A possibility is that citation distributions across universities or countries are fairly similar. For illustrative purposes, this paper investigates this conjecture for the partition of the world studied in Albarrán and Ruiz-Castillo (2012) in only two instances: in Physics, and in the all-sciences case. The approach we follow can be summarized as follows.

In the first place, the Characteristic Scores and Scales (CSS hereafter) technique, first introduced in scientometrics by Schubert *et al.* (1987), is applied to country citation distributions. We find that (i) they are highly skewed, in the sense that a large proportion of articles gets none or few citations while a small percentage of them account for a disproportionate amount of all citations, and (ii) appear to differ mostly by a scale factor. In the second place, Crespo *et al.* (2012a) introduce a method for quantifying the effect on overall citation inequality of differences in publication and citation practices across the 22 scientific fields already mentioned. In this paper, we apply this method to quantify the effect on overall citation inequality of differences in citation impact across countries. Using an additively decomposable citation inequality index, this effect is seen to be well captured by a between-group term, denoted by *IDCC* (citation *I*nequality due to *D*ifferences in *C*itation impact across *C*ountries), in a certain partition by countries and quantiles. Then we test the hypothesis that differences in countries' citation distributions have a strong scale factor component by normalizing the raw citation data using the country mean citations as normalization factors. If the conjecture is correct, then the *IDCC* term for the normalized distributions should be considerably smaller than in the original distributions. If this is the case, then we can assess the fundamental differences in citation impact over the entire support of country citation distributions by their mean citation –a very convenient conclusion.

Articles are assigned to countries according to the institutional affiliation of their authors on the basis of what had been indicated in the by-line of the publications. We must confront the technical difficulty posed by international cooperation, namely, the existence of articles written by authors belonging to two or more countries. The problem, of course, is that international articles as opposed to, say domestic articles, tend to be highly cited. Although this old question admits different solutions (see *inter alia* Anderson *et al.*, 1988, for a discussion), in this paper we focus on a *multiplicative* strategy according to which in every internationally co-authored article a whole count is credited to each contributing country. However, we report on the robustness of our results to a *fractional* strategy where each international article is fractioned into as many pieces as countries appear among its authors.

The rest of the paper is organized into four Sections. Firstly, we describe the method for quantifying the effect on overall citation inequality of differences in citation impact across countries under a multiplicative strategy. Secondly, we present the data and document some basic characteristics common to all country citation distributions. Thirdly, we present the empirical results for Physics and the all-sciences case. Finally, we offer some concluding comments and suggestions for possible extensions.

The effect on citation inequality of differences in citation impact across countries

Notation

Consider a certain scientific field, say Physics, consisting of N distinct articles, *indexed by* $l = 1,\ldots, N$. Let $\boldsymbol{Q} = (c_1,\ldots, c_l, \ldots, c_N)$ be the initial citation distribution, where c_l is the number of citations received by article l. Assume that there are P countries, indexed by $p = 1,\ldots, P$. For any l, let X_l be the non-empty set of countries to which the author(s) of article l belongs to, and let x_l be the cardinal of this set, i. e. $x_l = |X_l|$. Since, at most, an article can be written by authors in P countries, we have that $x_l \in [1, P]$.

Let N_p be the total number of distinct articles in p, indexed by $i = 1,\ldots, N_p$. In the multiplicative strategy, country p's ordered citation distribution can be described by $\boldsymbol{C_p} = (c_{p1}, \ldots, c_{pi}, \ldots, c_{pNp})$, where $c_{pi} = c_l$ for some article l in the initial distribution Q, and $c_{p1} \leq c_{p2} \leq \ldots \leq c_{pNp}$. What we call the *geographical extended count* is simply the union of these distributions, $\boldsymbol{C} = \cup_{\boldsymbol{p}} \boldsymbol{C_p}$, whose total number of articles is $M = \Sigma_p N_p = \Sigma_l x_l$. Only domestic articles, or articles exclusively authored by one or more scientists affiliated to research centers in a single country are counted once, in which case $x_l = 1$. Otherwise, $x_l \in [2, P]$. As long as $x_l > 1$ for some l, we have that $M > N$.

For any p, let us partition the citation distribution $\boldsymbol{C_p}$ into Π quantiles of size N_p/Π. That is, let $\boldsymbol{C_p} = (\boldsymbol{C_p^1},\ldots, \boldsymbol{C_p^{\pi}},\ldots, \boldsymbol{C_p^{\Pi}})$, where $\boldsymbol{C_p^{\pi}} = \{c_{pj}^{\pi}\}$ is the vector of the number of citations received by the N_p/Π articles in the π-th quantile of distribution $\boldsymbol{C_p}$, with $j = 1,\ldots, N_p/\Pi$, and $c_{pj} = c_k$ for some article k in distribution $\boldsymbol{Q}$. Assume for a moment that we disregard the citation inequality within every vector $\boldsymbol{C_p^{\pi}}$ by assigning to every article in that vector the mean citation of the vector itself, m_p^{π}, defined by

$$m_p^{\pi} = (\Sigma_j c_{pj}^{\pi})/(N_p/\Pi)$$

The interpretation of the fact that, for example, $m_p^{\pi} = 2\, m_q^{\pi}$ is that, on average, country p receives twice the number of citations as country q to represent the same underlying phenomenon, namely, the same degree of citation impact in both countries. In other words, for any p, the difference between m_p^{π} and m_q^{π} is entirely attributable to the difference in the citation performance that prevails in the two countries for articles that represent the same degree of citation impact within each of them.

For any π, consider the distribution ($\boldsymbol{m_1^{\pi}}$,…, $\boldsymbol{m_P^{\pi}}$,…, $\boldsymbol{m_P^{\pi}}$) where, for each p, each article in vector $\boldsymbol{C_p^{p}}$ receives the mean citation of the vector itself, m_p^{π}. The citation inequality of this distribution according to any relative inequality index I, $I(m_1^{\pi},\ldots, m_P^{\pi})$, abbreviated $I(\pi)$, is entirely attributable to differences in citation impact across the P countries at quantile π. Hence, any weighted average of these quantities, denoted by *IDCC* (citation *I*nequality due to *D*ifferences in *C*itation impact across *C*ountries), such as

$$IDCC = \Sigma_{\boldsymbol{\pi}} \beta^{\boldsymbol{\pi}}\, I(\pi) \tag{1}$$

with $\beta^{\boldsymbol{\pi}} \geq 0$, and $\Sigma_{\boldsymbol{\pi}}\beta^{\boldsymbol{\pi}} = 1$, provides a good measure of the citation inequality due to such differences. In the next Sub-section we introduce an appropriate citation inequality index, and a convenient weighting system.

The Measurement of the Effect of Differences in Citation Impact

For each π, define the vector $\boldsymbol{C^{\pi}} = (\boldsymbol{C_1^{\pi}},\ldots, \boldsymbol{C_p^{\pi}},\ldots, \boldsymbol{C_P^{\pi}})$ of size $(\Sigma_p N_p)/\Pi = N/\Pi$. Clearly, $\boldsymbol{C} = (\boldsymbol{C^1},\ldots, \boldsymbol{C^{\pi}},\ldots, \boldsymbol{C^{\Pi}})$, and the set of vectors $\boldsymbol{C^{\pi}}$, $\pi = 1,\ldots, \Pi$, form a partition of $\boldsymbol{C}$. As in Crespo *et al.* (2012), it is useful to develop the following measurement framework in terms of an additively decomposable inequality index, denoted by I_1. For any distribution $\boldsymbol{Z}$ with K elements, indexed by $k = 1,\ldots, K$, $\boldsymbol{Z} = (z_1,\ldots, z_K)$, I_1 is defined as:

$$I_1(\boldsymbol{Z}) = (1/K)\, \Sigma_k\, (z_k/m)\, log\, (z_k/m),$$

where m is the mean of distribution $\boldsymbol{Z}$. Apply the decomposability property of citation inequality index I_1 to the partition $\boldsymbol{C} = (\boldsymbol{C^1},\ldots, \boldsymbol{C^{\pi}},\ldots, \boldsymbol{C^{\Pi}})$:

$$\mathrm{I1(C)} = \Sigma\pi\ \mathrm{V}\pi\ \mathrm{I1(C}\pi) + \mathrm{I1(m1,..., m}\Pi) \tag{2}$$

where V^{π} is the share of total citations in $\boldsymbol{C}$ received by articles in $\boldsymbol{C^{\pi}}$, and ($\boldsymbol{m^1}$,..., $\boldsymbol{m^{\Pi}}$) is the distribution where each article in sub-group $\boldsymbol{C^{\pi}}$ is assigned the citation mean of the sub-group, $m^{\pi} = \Sigma_{\boldsymbol{p}}\ (N_{\boldsymbol{p}}/\Pi)\ m_p^{\pi}$. Next, apply the decomposability property of I_1 to the partition $\boldsymbol{C^{\pi}} = (\boldsymbol{C_1^{\pi}},\ldots, \boldsymbol{C_p^{\pi}},\ldots, \boldsymbol{C_P^{\pi}})$:

$$\mathrm{I1(C}\pi) = \Sigma\mathrm{p}\ \mathrm{V}\pi,\mathrm{p}\ \mathrm{I1}(Cp\pi) + \mathrm{I1(m1}\pi,..., \mathrm{mP}\pi) \tag{3}$$

where $V^{\pi,p}$ is the share of total citations in $\boldsymbol{C^{\pi}}$ received by articles in $\boldsymbol{C_p^{\pi}}$, and ($\boldsymbol{m_1^{\pi}}$,..., $\boldsymbol{m_P^{\pi}}$) is the distribution where each article in quantile $\boldsymbol{C_p^{\pi}}$ is assigned the citation mean of the quantile, $m_{\boldsymbol{p}}^{\pi}$. Substituting (3) into (2), we obtain that the overall citation inequality in the double partition of distribution $\boldsymbol{C}$ into P

countries and Π quantiles within each country can be decompose into the following three terms:

$$I_1(\boldsymbol{C}) = W + S + IDCC, \quad (4)$$

where: $W = \Sigma_\pi \Sigma_p V^{\pi,p} I_1(\boldsymbol{C_p^\pi})$,

$$S = I_1(\boldsymbol{m^1},..., \boldsymbol{m^p}),$$

$$IDCC = \Sigma_\pi V^\pi I_1(\boldsymbol{m_1^\pi},\ldots, \boldsymbol{m_P^\pi}) = \Sigma_\pi V^\pi I(\pi).$$

The term W in Eq. 4 is a within-group term that captures the weighted citation inequality within each quantile in every country. Clearly, for large Π, $I_1(\boldsymbol{c_p^\pi})$, and hence W is expected to be small. The S term is the citation inequality of the distribution $(\boldsymbol{m^1},..., \boldsymbol{m^\Pi})$ in which each article in the vector $\boldsymbol{C^\pi}$ is assigned the vector's citation mean, m^π. Thus, S is a measure of citation inequality at different degrees of citation impact that captures well the skewness of science characterizing citation distributions in different contexts. Consequently, S is expected to be large. Finally, the expression $I_1(\pi)$ is the citation inequality according to I_1 attributable to differences across countries at the degree p of citation impact. Thus, the weighted average of $I_1(\pi)$ for all π in Eq. 4 –which is a version of the *IDCC* term introduced in Eq. 1– provides a convenient measure of the citation inequality due to such differences.

Data

The Distribution of Articles By Field and Country Under the Multiplicative Approach

Since we wish to address a homogeneous population, in this paper only research articles or, simply, articles are studied. As indicated in the Introduction, we begin with a large sample, consisting of more than 4.4 million articles published in 1998-2003, as well as the citations these articles receive using a five-year citation window for each one. In our dataset, the number of distinct articles in the original dataset is N = 4,472,332, while the number of articles in the extended count is M = 5,450,309, a total which is 21.9% larger than N. In turn, the number of distinct articles in Physics is N_P = 456,144, while the number of articles in the corresponding geographically extended count is 626,304, a total which is 37.3% larger than N_P.

We consider 36 countries and two residual geographical areas that have published about 10,000 articles in all sciences in 1998-2003. In the all-sciences case, the U.S. publishes about 27% of the total, while the EU is responsible for

approximately one third in the extended count. The remaining 23 countries and the two geographical areas taken together publish almost 39% of the total. In Physics, the U.S., and the EU publish 18%, and 34% of the total in the extended count. For reasons of space, further descriptive statistics are only available on request.

The All-sciences Case

Given the wide differences in publication and citation practices, in scientometrics is customary to proceed to some normalization before aggregating all fields into what we call the all-sciences case. Recent results indicate that the standard practice of using field MCs as normalization factors generates good results (Radicchi and Castellano, 2012a, b, Leydesdorff *et al.*, 2012, and Crespo *et al.*, 2012, a, b). Therefore, in the sequel all references to the all-sciences case take place after the standard field normalization.

Characteristics of Country Citation Distributions

It is important to know whether or not country citation distributions present the fundamental features that have been appreciated for entire sub-fields, broad fields, and other aggregates (Albarrán and Ruiz-Castillo, 2010, Albarrán *et al.*, 2011, and Herranz and Ruiz-Castillo, 2011). For this purpose, as indicated in the Introduction we use a scale and size invariant statistical method that allows us to focus on the shape of citation distributions: the CSS technique. The following *characteristic scores* are determined: m_1 = mean citation for the entire distribution, and m_2 = mean citation for articles with citations above m_1. Consider the partition of the distribution into three broad classes: articles with none or few citations below m_1; fairly cited articles, with citations above m_1 and below m_2; and articles with a remarkable or outstanding number of citations above m_2. Both for Physics and the all-sciences case, Table 1 includes the average and standard deviation over all countries and geographical areas for the percentage of articles in the three classes, as well as the corresponding statistics for the percentages of the total number of citations accounted by each class.

Table 1. The Skewness of Country Citation Distributions in Physics and the All-fields Case. Averages (and Standard Deviations) over 38 Countries of the Percentages of Articles and the Percentages of Total Citations by Category

	Percentage of Articles In Category			Percentage of Total Citations Accounted For By Category		
	1	**2**	**3**	**1**	**2**	**3**
Physics	71.1 (2.4)	20.7 (1.6)	8.1 (1.2)	22.5 (2.3)	33.5 (1.1)	44.0 (2.2)
All Sciences	75.7 (2.6)	17.2 (1.9)	7.1 (0.8)	29.7 (0.7)	32.0 (0.5)	38.4 (0.8)

The results are truly remarkable. A complex set of economic, sociological, political, and intellectual factors are influencing the research performance of each country in every field and, consequently, the shape of their citation distributions. Nevertheless, the small standard deviations in Table 1 indicate that country citation distributions in Physics and the all-sciences case tend to share some fundamental characteristics. Specifically, between 71% and 76% of all articles receive citations below the mean and account for, approximately, between 22% and 30% of all citations, while articles with a remarkable or outstanding number of citations represent about 7% or 8% of the total, and account for, approximately, between 38% and 44% of all citations. Thus, we can conclude that, both within Physics and the all-sciences case, country citation distributions are very similar and highly skewed.

These results closely resemble those concerning the shapes of citation distributions across a wide array of 219 sub-fields identified with the Web of Science subject categories distinguished by Thomson Reuters (Albarrán *et al.*, 2011, p. 391), where approximately 69% of all articles receive citations below the mean and account for 21% of all citations, while articles with a remarkable or outstanding number of citations represent about 9% or 10% of the total, and account for approximately 44% of all citations. This similarity between citation distributions paves the way for meaningful comparisons of citation counts across heterogeneous scientific disciplines (see *inter alia* Radicchi and Castellano, 2012 a, b). In particular, Crespo *et al.* (2012a, b) establish that standard procedures that use field and sub-field mean citations as normalization factors dramatically reduces the citation inequality attributed to differences in publication and citation practices across them. Analogously, the similarity between country citation distributions observed in Table 1 suggests the possibility of searching for normalization procedures that considerably reduce the citation inequality attributable to differences in citation impact across countries in Physics and the all-sciences case –a task pursued in the next Section.

Results

Physics

The results concerning the decomposition of Eq. 4 for Physics when $\Pi = 100$ appear in Panel A in Table 2. As expected, the *W* and *S* terms are small and large, respectively, representing 4.6% and 90.7% of overall citation inequality. Consequently, the *IDCC* term only represents 4.7% of overall citation inequality. Thus, once we control for the skewness of science –namely, differences in citation counts from poorly, fairly, and highly cited articles in all countries–, the effect of differences in performance across countries in the field of Physics is relatively small (however, see the discussion below). In any case, once we normalize the raw citation data with the countries' MCs, the *IDCC* in absolute terms goes down

by 81%. Since overall inequality also decreases, after normalization the *IDCC* term only represents approximately 0.9% of overall citation inequality.

Table 2. Total Citation Inequality Decomposition Before and After Country Mean Normalization. Physics

	Within-group Term, *W*	Skew. of Science Term, *S*	*ICDP* Term	Total Citation Ineq., *I1*(*C*)	In % (1)/(4)	In % (2)/(4)	In % (3)/(4)
A. All Physics	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Raw Data	0.0427	0.844	0.0435	0.9305	4.6	90.7	4.7
Mean Normalization	0.0432	0.86	0.0087	0.9123	4.6	92.5	0.9
B. Domestic Articles							
Raw Data	0.0165	0.8179	0.0942	0.9286	1.8	88.1	10.15
Mean Normalization	0.0164	0.8291	0.0111	0.8566	1.9	96.8	1.3
C. International Articles							
Raw Data	0.0525	0.8172	0.0185	0.8882	5.9	92	2.09
Mean Normalization	0.0507	0.818	0.0054	0.8742	5.8	93.6	0.62

The similarity between country citation distributions in Table 1 can be interpreted as indicating that a large part of differences in citation impact between any pair of countries appears to be due to a scale factor. The country MCs seem to capture well such scale factors in the sense that, once they are used as normalization factors, the citation inequality attributable to differences in citation impact across countries is vastly reduced: the relative importance of the *IDCC* term goes down by a greater than five factor –a large order of magnitude.

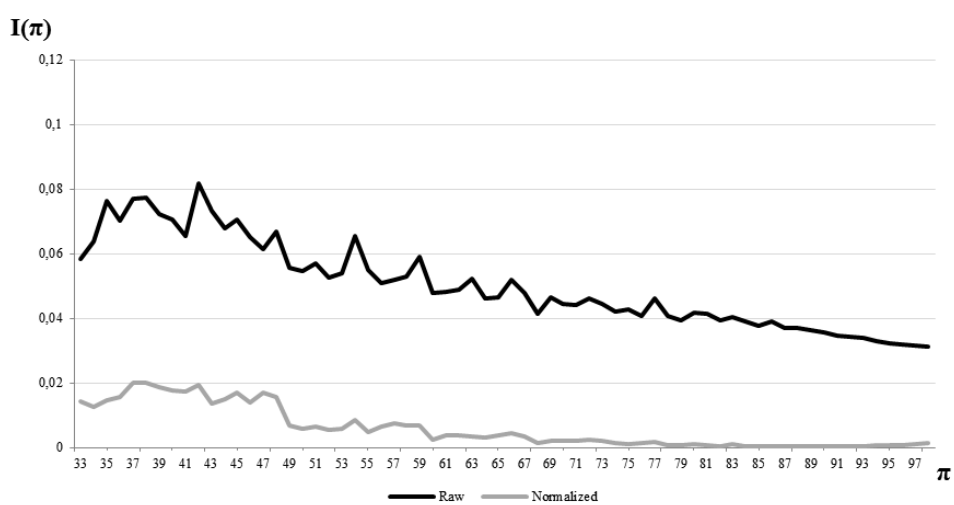


Figure 1. Citation Inequality Due to Differences in Citation Impact Across Countries, I(π), as a function of π. Raw and Country Normalized Data for Physics

As in Crespo *et al.* (2012a, b), it is illuminating to graphically observe the evolution of $I(\pi)$ as a function of π in Figure 1. A key fact is that the effect of differences in citation impact across countries in the raw data tends to decline as we advance towards higher quantiles. This partially offsets the rapidly increasing pattern of the weights V^{π}, leading to a relatively low value for the *IDCC* term. In any case, the effect of mean country normalization is clearly apparent in Figure 1 (Since the values of $I(\pi)$ in the interval (10, 32) for the raw data are very high, for clarity all values for $\pi < 33$ are omitted in Figure 1).

A possible explanation of the declining pattern of $I(\pi)$ as a function of π is that, as illustrated in Figure 2, articles written under international co-authorship are typically highly cited. As a matter of fact, in both cases there are more international than domestic articles in the upper tail: from the 70th to the last percentile in Physics, and from the 98th percentile onwards in the all-sciences case.

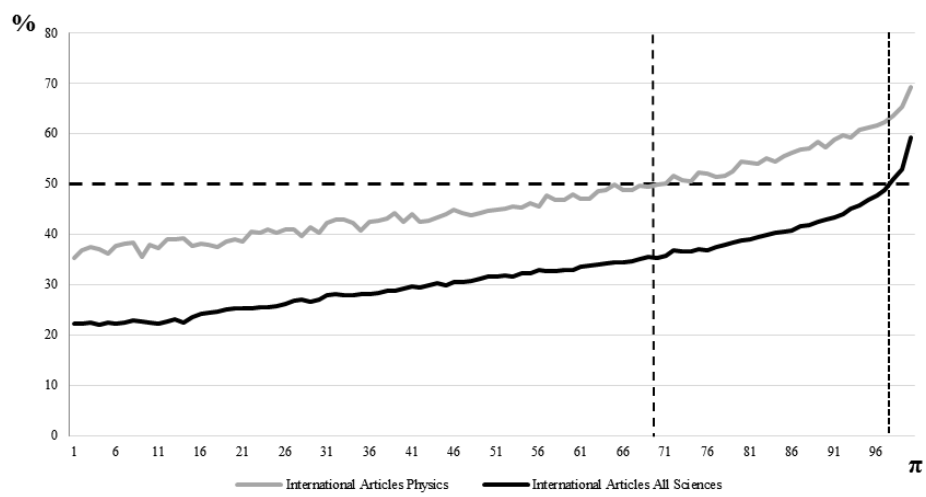


Figure 2. Percentage of international articles by percentile. Physics and All-sciences case

Thus, as we approach higher values of π, differences across countries who cooperate in highly cited articles should decrease –whoever are the cooperating countries involved. This is exactly what we observe in Figure 3 where the expression $I(\pi)$ is represented as a function of π for two types of articles: domestic articles (bold curve), and international articles (grey curve). The expression $I(\pi)$ for domestic articles shows a relatively small variation as a function of π, and a steady increase for the last few values of π. Instead, as expected, for international articles $I(\pi)$ is a decreasing function of π.

The pattern of $I(\pi)$ for domestic articles, together with the rapidly increasing weighting system, causes *IDCC* to be rather large, representing 10.1% of overall citation inequality (Panel B in Table 2, raw data). Given its own $I(\pi)$ pattern, the opposite is the case for international articles (Panel C in Table 2, raw data). Naturally, in both cases country mean normalization has very large effects, leading to IDCC terms representing 1,3% and 0.6% of the corresponding overall citation inequality.

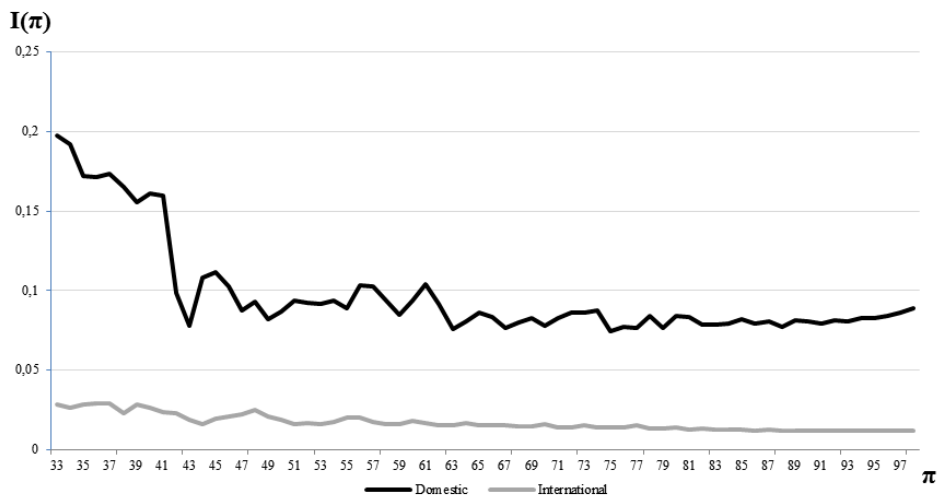


Figure 3. Citation Inequality Due to Differences in Citation Impact Across Countries, I(π), as a function of π. Domestic and International Articles in Physics

The All-sciences Case

Results for the decomposition of overall inequality in the all-sciences case after field normalization are included in Table 3. The variation of the expression $I(\pi)$ as a function of π is illustrated in Figure 4 (for values of $\pi > 46$). Firstly, for the raw data the *IDCC* term represents, approximately, 4% of overall citation inequality. After using the countries' MCs as normalization factors, this is reduced to less than 1%. In absolute terms, the *IDCC* term is reduced by 80% (Panel A in Table 3 and Figure 3). Secondly, the numerical pattern of domestic and international articles resembles the one in Physics (Panels B and C in Table 3). However, as illustrated in Figure 3, the expression $I(\pi)$ is a relatively constant function of π over a large quantile interval. This opens up the possibility of computing what Crespo *et al.* (2012a, b) call *exchange values*, namely relative average-based indicators over this interval that may serve to compare countries' performances in citation impact. Thirdly, a key finding is that the order of magnitude of the above results is very similar to what we found in Physics. A possible explanation, that deserves further research, is that differences in citation

impact across countries are of a similar order of magnitude in all fields –a truly surprising result.

Table 3. Total Citation Inequality Decomposition Before and After Country Mean Normalization. All-sciences case

	Within-group Term, W	Skew. of Science Term, S	*ICDP* Term	Total Citation Ineq., $I1(C)$	In % (1)/(4)	In % (2)/(4)	In % (3)/(4)
A. All Articles	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Raw Data	0.0257	0.7434	0.0341	0.9305	3.2	92.5	4.2
Mean Normalization	0.0258	0.7586	0.0070	0.9123	3.3	95.8	0.9
B. Domestic Articles							
Raw Data	0.0162	0.7405	0.0582	0.9286	2.0	90.9	7.1
Mean Normalization	0.0156	0.7579	0.0105	0.8566	2.0	96.7	1.3
C. International Articles							
Raw Data	0.0305	0.7773	0.0240	0.8149	3.7	93.4	2.09
Mean Normalization	0.0312	0.7807	0.0026	0.7840	3.8	95.9	0.3

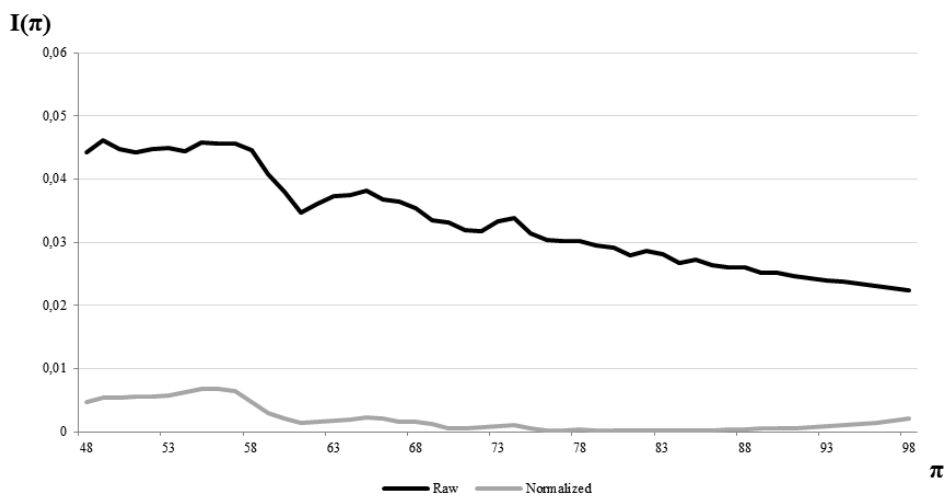


Figure 4. Citation Inequality Due to Differences in Citation Impact Across Countries, I(π), as a function of π. Raw and Country Normalized Articles In the All-sciences Case

Conclusions

This paper, which has adopted a multiplicative strategy in the treatment of internationally co-authored articles, has achieved the following two aims. Firstly, using the CSS technique, we have presented convincing evidence concerning the similarity of the shape of citation distributions in a large number of countries in two instances: articles in Physics and the all-sciences case published in 1998-2003 with a five-year citation window. In Physics, for example, on average the partition into three classes of poorly cited, fairly cited, and highly cited articles is,

approximately, 71/21/8. These three classes account for 22/34/44 of all citations. These percentages are extremely similar to those found in previous research at the level of 219 scientific sub-fields and other aggregate categories.

Secondly, we have introduced a decomposition of overall citation inequality that includes a *IDCC* term capturing the effect on citation inequality of differences in citation impact across countries. In this scenario, we have explored the following idea: if differences in country citation distributions are essentially due to a scale factor, then after normalizing the raw citation counts by the countries' MCs, the *IDCC* term should be very much reduced. We have found that both in Physics and in the all-sciences case the *IDCC* term represents, approximately, 4.2%-4.7% and 0.9% of overall citation inequality before and after country normalization, respectively. In brief, approximately 80% of all differences in citation impact across countries seem to be due to a scale factor well captured by the countries' MCs.

This largely explains the results in Albarrán and Ruiz-Castillo (2012) concerning the similarity of the ranking of countries and geographical areas when using an average-based or a percentile rank citation indicator focusing on the world top-10% of articles with the highest citation counts. We conclude that, once we verify the similarity of citation distributions –using, for example, the CSS approach– international comparison carried in terms of countries' MCs seem to capture most of the country differences in citation impact over the entire support of country citation distributions.

This conclusion, however, needs to be qualified. First of all, we should say that our results are essentially maintained when we follow a fractional approach for the treatment of internationally co-authored articles (results are available on request). However, our methods must be applied to other scientific fields different from Physics, and the robustness of the above results must be investigated with other datasets: other publication years; other citation windows, and other sources different from Thomson Reuters. On the other hand, it seems interesting to apply this methodology to universities for which, as indicated in the Introduction, Waltman *et al.* (2012) have provided evidence concerning the similarity of the 2011/2012 Leiden ranking using average-based and the $PP_{top\ 10\%}$ indicator.

As far as extensions is concerned, there are several interesting alternatives to the countries' MCs that can serve as normalization factors. Firstly, as in Crespo *et al.* (2012a, b) one could compute exchange rates for each country. As indicated in the previous Section, there is evidence in the all-sciences case to warrant this computation. In Physics, and perhaps in other fields, this can be at least attempted for domestic articles. Secondly, one could use as normalization factors the relative country's $PP_{top\ 10\%}$ indicator. Thirdly, one could use the mean citation over articles in the top 10% of every country citation distribution, which corresponds

to the second member of the family of high-impact indicators introduced in Albarrán *et al.* (2011b, c). Naturally, one could assess the adequacy of any of these alternatives in the measurement framework developed in this paper by studying the extent of the reduction they generate in the *IDCC* term.

In a completely different direction, future research should confront the striking similarity of country, field, and sub-field citation distributions, as well as individuals' productivity distributions. In the first place, one should systematically explore whether these similarities are also present not only in university departments, but also in field journals, research institutes, and other types of research units. But more importantly, we should start asking: what type of behavioral model could explain the existing similarities in scenarios so different as scientific fields and international comparisons within specific fields or within the all-sciences case? The evidence indicates that, against all expectations, non-random samples of individual scientists in a field, or in a country within a field give rise to strikingly similar citation distributions. Quite apart from the modeling of the restrictions under which they work, it would appear that one may assume that the distribution of talent of individual scientists in different contexts resembles the distribution one would obtain with random sampling from a highly skewed pool of individual talents.

References

Albarrán, P. and J. Ruiz-Castillo (2011), "References Made and Citations Received By Scientific Articles", *Journal of the American Society for Information Science and Technology*, 62: 40-49.

Albarrán, P. and J. Ruiz-Castillo (2012), "The Measurement of Scientific Excellence Around the World", Working Paper, Economic Series 12-08, Universidad Carlos III (http://hdl.handle.net/10016/13896).

Albarrán, P., J. Crespo, I. Ortuño, and J. Ruiz-Castillo (2011a), "The Skewness of Science In 219 Sub-fields and A Number of Aggregates", *Scientometrics*, 88: 385-397.

Albarrán, P., I. Ortuño, and J. Ruiz-Castillo (2011b). "The Measurement of Low- and High-impact In Citation Distributions: Technical Results", *Journal of Informetrics*, 5: 48-63.

Albarrán, P., I. Ortuño and J. Ruiz-Castillo (2011c), "High- and Low-impact Citation Measures: Empirical Applications", *Journal of Informetrics*, 5: 122-145.

Anderson, J., P. Collins, J. Irvine, P. Isard, B. Martin, F. Narin, and K. Stevens (1988), "On-line Approaches to Measuring National Scientific Output: A Cautionary tale", *Science and Public Policy*, 15: 153-161.

Crespo, J. A., Li, Y., and Ruiz-Castillo, J. (2012a), "Differences in Citation Practices Across Scientific Fields", Working Paper, Economic Series 12-06, Universidad Carlos III (http://hdl.handle.net/10016/14771).

Crespo, J. A., Li, Yunrong, Herranz, N., and Ruiz-Castillo, J. (2012b), "Field Normalization at Different Aggregation Levels", Working Paper 12-022, Universidad Carlos III (htpp:/hdl.handle.net/10016/15344).

Herranz, N., and Ruiz-Castillo, J. (2012), "Multiplicative and Fractional Strategies When Journals Are Assigned to Several Sub-fields", in press, *Journal of the American Society for Information Science and Technology* (DOI:10.1002/asi.22629).

Glänzel, W. (2007), "Characteristic Scores and Scales: A Bibliometric Analysis of Subject Characteristics Based On Long-term Citation Observation", *Journal of Informetrics*, 1: 92-102.

Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., and de Nooye, W. (2012), "Field-normalized Impact Factors: A Comparison of Rescaling versus Fractionally Counted IFs", in press, *Journal of the American Society for Information Science and Technology*.

Radicchi, F., and Castellano, C. (2012a), "A Reverse Engineering Approach to the Suppression of Citation Biases Reveals Universal Properties of Citation Distributions", *Plos One*, 7, e33833, 1-7.

Radicchi, F., and Castellano, C. (2012b), "Testing the fairness of citation indicators for comparisons across scientific domains: The case of fractional citation counts", *Journal of Informetrics*, 6: 121-130.

Rousseau (2012), "Basic Properties of Both Percentile Rank Scores and the I3 Indicator", *Journal of the American Society for Information Science*, 63: 416-420.

Schubert, A., W. Glänzel and T. Braun (1987), "A New Methodology for Ranking Scientific Institutions", *Scientometrics*, 12: 267-292.

Seglen, P. (1992), "The Skewness of Science", *Journal of the American Society for Information Science*, 43: 628-638.

Tijssen, R. M., Visser, M., & van Leeuwen, T. (2002), "Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference?", *Scientometrics*, 54: 381–397.

Tijssen, R. M., van Leeuwen, T. (2006) "Centers of Research Excellence and Science Indicators. Can 'excellence' Be Captured In Numbers?", in W. Glänzel (ed), *Ninth International conference on Science and Technological Indicators*, Leuven, Belgium: Katholieke Universiteit Leuven.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., van Leeuwen, T. H., van Raan, A. F. J., Visser, M. S.. and Wouters, P. (2012), The Leiden Ranking 2011/2012: Data Collection, Indicators, and Interpretation, *Journal of the American Society for Information Science*, 63: 2419-2432.

DIRECTIONAL RETURNS TO SCALE OF BIOLOGICAL INSTITUTES IN CHINESE ACADEMY OF SCIENCES

Guo-liang Yang[1], Li-ying Yang[2], Wen-bin Liu[3], Xiao-xuan Li[4] , and Chun-liang Fan[5]

[1]*glyang@casipm.ac.cn;* [4]*Xiaoxuan@casipm.ac.cn;* [5]*fcl@casipm.ac.cn*
Institute of Policy and Management, Chinese Academy of Sciences, 100190 Beijing (China)

[2]*yangly@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, 100190 Beijing (China)

[3]*w.b.liu@kent.ac.uk*
Kent Business School, University of Kent, CT2 7PE Canterbury (United Kingdom)

Abstract
This paper aims to investigate the directional returns to scale of 15 biological institutes in Chinese Academy of Sciences. Firstly, the input-output indicators are proposed, including Staff, Research funding, SCI papers, High-quality papers and Graduates training, etc. Secondly, this paper uses the methods proposed by Yang (2012) to analyze the directional returns to scale and the effect of directional congestion of biological institutes in Chinese Academy of Sciences. Based on the analytical results, we have the following findings: (1) we can detect the region of increasing (constant, decreasing) directional returns to scale for each biological institute. This information can be used as one of the basis of decision-making on organizational adjustment; (2) we find that the effect of congestion and directional congestion occurs in several biological institutes. On this occasion, the outputs of these institutes will decrease with the inputs increase. So, these institutes should analyze the deep reason for the occurrence of congestion effect so that science and technology (S&T) resources can be used more effectively.
Keywords: Research institute, returns to scale, directional returns to scale, congestion, directional congestion

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3)

Background
The efficiency of scientific and technological resources utilization is one of the important issues of concern in the S&T management department in China. From 2007 to 2011, the national financial allocation on science and technology (S&T) reaches 1,340.8 billion yuan (RMB), 2.7 times of that in the period of 2001-2005

and the average annual growth rate reaches 23%. Meanwhile, China's R&D staff reaches nearly 260 million people in 2010 and ranks first in the world. However, S&T resources are still scarce so how to use these resources efficiently and effectively becomes a topic of public concern and an important issue facing the S&T management department. All these issues relate to the rational allocation of the limited S&T resources to maximize the efficiencies of resource utilization, and are the important and urgent issues needed to understand and master in national macro S&T management levels. To address these issues, the returns to scale (RTS) and efficiencies of S&T resources utilization of research institutions should be investigated first.

Chinese Academy of Sciences (CAS) is a leading academic institution and comprehensive research and development center in natural science, technological science and high-tech innovation in China. Today's CAS has 12 branch offices, 117 institutes with legal entity, more than 100 national key laboratories and national engineering research centers, and about 1,000 field stations throughout the country. Its staff even surpassed 50,000. In 1998, with the approval of the Chinese Government, the CAS launched the Pilot Project of the Knowledge Innovation Program (PPKIP) in an effort to build China's national innovation system.

In the period of PPKIP, the inputs and outputs of CAS grow significantly. We take several indicators for example. The amount of total income of CAS increases from 493.598 million yuan(RMB) in 1998 to 1,703.971 million yuan (RMB) in 2007. Meanwhile, the full time equivalence (FTE) increases from 30,611 in 1998 to 44,307 in 2007. The number of SCI papers increases from 5,474 in 1998 to 23,674 in 2010, an increase of 3.3 times. The ability to gain external funding of CAS as a whole improves continually in the period of PPKIP. We take the projects from Natural Science Foundation of China (NSFC) and "973" and "863" projects from Ministry of Science and Technology (MOST) undertaken by researchers in CAS in 1998-2009 as example. The magnitude of funding grows from less than 600 million yuan (RMB) to about 30 billion yuan (RMB). In the period of PPKIP, the input-output indicators of the biological institutes in CAS grow significantly also. The quantitative monitoring results from 2004 to 2009 in CAS indicate that the research funding and high-quality papers increase exponentially over this 6-year period. In this context, how to analyze the development status of these institutes is an important issue.

The institute evaluation began in 1993. Over the past decade, the models and methodologies of institute evaluation in CAS were adjusted constantly according to characteristics of different stages of institute development and experienced different stages, including the "Blue Book" evaluation, "PPKIP" evaluation, "Innovation Capability Index" evaluation, "Comprehensive Quality" evaluation and "Major R&D Outcome-oriented System" evaluation. The quantitative monitoring system is an important constituent part of the current evaluation system in CAS and is formed of the indicators (such as the amount of research

funding, the number of high-quality paper, etc.) for multiple inputs and outputs of each institute in CAS.

RTS is an important issue in the analysis of organizational performance, which can help decision-makers (DMs) decide if the size of the organization should be expanded or reduced. RTS is a classic economic concept that is tied to the relationship between production factors and variation of outputs. If the scale of production changes due to the proportional increase (or decrease) of all production factors, the RTS measures the change rate of output(s) with that of all input(s) proportionally (Pindyck and Rubinfeld, 2000). There are three types of RTS in production processes, which is classified as constant returns to scale (CRS), or decreasing returns to scale (DRS), or increasing returns to scale (IRS) if proportional change of outputs is the same, or less, or more than that of inputs respectively. The traditional definition of RTS in economics is based on the idea to measure radial changes in outputs caused by those of all inputs. In some real applications, however, the increase in scale is often caused by the inputs changing in unequal proportions. Based on the above thinking, Yang (2012) introduces directional RTS from a global and local (directional scale elasticity) perspective under the Pareto preference, and gives specific formulations of directional RTS.

This paper aims to estimate the directional returns to scale of 15 biological institutes in CAS. For these institutes, whether the massive financial investment on S&T by the government is used efficiently? What are the efficiencies of S&T resources usage? Are these institutes running on optimal scale? All these issues relate to the rational allocation of the limited S&T resources and are basic information for related S&T policies making by national S&T macro-management departments.

The paper is organized as follows: Section 2 presents the methodology used in this paper to analyze the directional returns to scale (RTS) of these institutes using the methods proposed by Yang (2012). In Section 3, we will show the results of analysis, including the directional RTS, optimal inputs direction and the congestion effect, and the conclusion is given in Section 4.

Methodology

Input-output Indicators

Performance management has become much more common in government managed organizations in the past few years as a consequence of two principal factors: a). increased demand for accountability by governing bodies, press, and the general public, and b). a growing commitment by organizations to focus on results to improve performance (Poister, 2003). The development of performance management during the past decade indicates a change from the "output (result)" model to the "objective - process - result" model (Zhang et al., 2011). Geisler (2000) defined metrics for evaluating scientific work as "a system of measurement that includes: (1) the objective being measured; (2) the units to be measured; (3) the value of the units." Keeney and Gregory (2005) studied how to

select measures effectively to determine whether bodies that operate in such an environment are meeting their targets (i.e. assessment indicators). Roper et al. (2004) discussed the indicators for the pre-assessment of public R&D funding, based on the beneficial outcomes that increased knowledge provides to society. Moreover, Soft Systems Methodology (SSM) can be used to analyze systematically the operation of research institutions to build a relatively more complete and reasonable set of evaluation indicators based on the "3E" theory (Efficacy, Efficiency, Effectiveness) (Meng et al., 2007; Mingers, 2009). Zhang et al. (2011) proposed strategy maps for National Research Institutes (NRIs) based on the discussions on the general rules of research activities so that the managers can describe the strategies of their organizations more clearly, accurately and logically.

In the real practice of evaluation of national research institutes (NRIs), due to the characteristics that NRIs tend to be more concerned about the long-term accumulation of scientific issues and the needs of national economic and social development, as well as national security than universities, how to improve the efficiencies of S&T resources utilization so that they can play more important role in the fields of economic development, social progress and national defense is an important issue for the management of NRIs (Li, 2005). In the evaluation practice in CAS, dozens of quantitative indicators (e.g., publications, awards, patents, staff, talents, funds, graduates training) are used to monitor the annual development status of affiliated institutes.

Based on above analysis, we can see that there exist dozens of quantitative indicators for the inputs and outputs of NRIs. In this paper, we follow the succinct indicators used in Liu et al. (2011) to analyze the directional RTS for these biological institutes in CAS. These indicators include: (1) input indicators: Staff and Research Expenditure (Res.Expen.); (2) output indicators: SCI publications (SCI Pub.), Publications in high-impact journals (High Pub.) and Graduate Enrollment (Grad.Enroll). As input indicators, Staff and Res.Expen. denote the number (FTE) of regular staff total in each biological institute and the total income obtained by each institute respectively. As output indicators, SCI Pub. and High Pub. denote the number of published papers indexed by Science Citation Index (SCI) and the number of published papers on the journals with top 15% impact factor in each JCR field respectively, and Grad.Enroll. denotes the number of studying master and Ph.D students currently. Based on these indicators, we investigate the directional RTS of 15 biological institutes in CAS in 2010.

Data

The input-output data of 15 biological institutes in CAS in 2010 is shown in Table 1 as follows.

Table 1. The input-output data of 15 biological institutes in CAS in 2010.

DMU	*Inputs*		*Outputs*		
	Staff (FTE)	*Res.Expen. (RMB million)*	*SCI Pub. (Number)*	*High Pub. (Number)*	*Grad.Enroll. (Number)*
DMU_1	640	253.70	325	105	604
DMU_2	367	251.15	368	109	477
DMU_3	172	91.74	207	66	241
DMU_4	435	189.63	256	62	388
DMU_5	472	395.86	259	96	500
DMU_6	543	497.32	216	93	553
DMU_7	236	89.45	112	39	190
DMU_8	1910	930.44	785	323	1488
DMU_9	608	537.2	385	125	417
DMU_{10}	198	111.28	118	36	235
DMU_{11}	289	182.04	216	63	481
DMU_{12}	335	101.85	125	37	267
DMU_{13}	356	113.97	189	66	232
DMU_{14}	413	214.94	313	64	302
DMU_{15}	180	56.91	83	17	126

Data Source: (1) Monitoring data of institutes in CAS, 2011; (2) Statistical yearbook of CAS, 2011.
Note: These data were derived from these institutes in the period of Jan.01,2010~Dec.31,2010

Analysis methods

Yang (2012) proposed the definition of directional scale elasticity in economics in the case of multiple inputs and multiple outputs as follows.

$$e(Y_0,X_0)=\left(-\sum_{i=1}^{m}\frac{\partial F}{\partial x_i}x_i\omega_i \Big/ \sum_{r=1}^{s}\frac{\partial F}{\partial y_r}y_r\delta_r\right)\Big|(Y_0,X_0)$$

where $F(Y,X)=0$ denotes the continuously directional differentiable production function and $\sum_{r=1}^{s}\delta_r=s;\sum_{i=1}^{m}\omega_i=m$. Basically the traditional scale elasticity only considers the elasticity for the input change along the diagonal directional, while ours consider for all possible directions.

Remark 1. *In many applications, the above formula may not hold. Then the differential in the above formula has to be replaced by the left-hand directional derivative* $(t\to 0^-)$ *and the right-hand directional derivative* $(t\to 0^+)$ *so we will have the left-hand and right-hand scale elasticities (see Banker(1984), Podinovski and Førsund (2010), and Atici and Podinovski (2012)).*

In public sector, the production function cannot often be formulated as $F(Y,X)=0$.Therefore, in practice, DEA method is one of the most commonly used approaches for the analysis of RTS on public sector (e.g., research

institutions) (Fox, 2002; Meng et al., 2006; Zhou and Li, 2009a, 2009b). The estimation of RTS of DMUs using DEA method is investigated first by Banker (1984) and Banker et al. (1984). Banker (1984) introduced the definition of the RTS from classical economics into the framework of the DEA method, and he used CCR-DEA model with radial measure to estimate the RTS of evaluated DMUs. Soon after that, Banker et al. (1984) proposed BCC-DEA model under the assumption of variable RTS, and investigated how to apply the BCC-DEA model to estimate the RTS of DMUs. The existed RTS measurements in DEA models are all based on the definition of RTS in the DEA framework made by Banker (1984). Banker (1984) introduced the RTS in economics into the DEA framework and proposed the method to determine the RTS of DMUs in DEA models, which extended the application area of DEA from relative efficiency evaluation to RTS measurement. The RTS is a classic economic concept describing the relationship between changes in the scale of production and output. The traditional definition of RTS in economics is based on the idea to measure radial changes in outputs caused by those of all inputs. Yang (2012) argued that due to the complexity of research activities in research institutions, it often can be observed that production factors are not necessarily tied together proportionally, and inputs change non-proportionally. Based on this thinking, he introduced directional RTS from a global and local (directional scale elasticity) perspective under the Pareto preference, and gives specific formulations of directional RTS and corresponding models. In addition, he demonstrated that traditional RTS is a special case of directional RTS in the radial direction, so that directional RTS can provide a basis for decision-making concerning the further development of such production processes. He gives the definition of directional RTS in DEA framework based on the production possibility set (PPS) as follows.

Definition 1(directional RTS):

Assuming $DMU\left(Y_0, X_0\right) \in PPS$ and $X_0 \in R_m^+, Y_0 \in R_s^+$, we let

$$\beta(t) = \max\left\{\beta \middle| \left(\Omega_t X_0, \Phi_\beta Y_0\right) \in PPS, t \neq 0\right\}$$

where $\Omega_t = diag\left\{1+\omega_1 t, ..., 1+\omega_m t\right\}$ and $\Phi_\beta = diag\left\{1+\delta_1\beta, ..., 1+\delta_s\beta\right\}$. $\left(\omega_1, ..., \omega_m\right)^T$ ($\omega_i \geq 0, i = 1, ..., m$) and $\left(\delta_1, ..., \delta_s\right)^T$ ($\delta_r \geq 0, r = 1, ..., s$) represent inputs and outputs directions respectively and satisfy $\sum_{i=1}^{m} \omega_i = m; \sum_{r=1}^{s} \delta_r = s$ where t, β are input and output scaling factors respectively. We let

$$\rho^- = \lim_{t \to 0^-} \frac{\beta(t)}{t} \qquad (1) \qquad\qquad \rho^+ = \lim_{t \to 0^+} \frac{\beta(t)}{t} \qquad (2)$$

Then we have

(a) if $\rho^- > 1$ (or $\rho^+ > 1$) holds, then increasing directional RTS prevails on the left-hand (or right-hand) side of this point (Y_0, X_0) in the direction of $(\omega_1, \omega_2, ..., \omega_m)$ and $(\delta_1, \delta_2, ..., \delta_s)$;

(b) if $\rho^- = 1$ (or $\rho^+ = 1$) holds, then constant directional RTS prevails on the left-hand (or right-hand) side of this point (Y_0, X_0) in the direction of $(\omega_1, \omega_2, ..., \omega_m)$ and $(\delta_1, \delta_2, ..., \delta_s)$;

(c) if $\rho^- < 1$ (or $\rho^+ < 1$) holds, then decreasing directional RTS prevails on the left-hand (or right-hand) side of this point (Y_0, X_0) in the direction of $(\omega_1, \omega_2, ..., \omega_m)$ and $(\delta_1, \delta_2, ..., \delta_s)$.

The methods he proposed are as follows. For the strong efficient DMU (X_0, Y_0) on the strongly efficient frontier in BCC-DEA model, its directional scale elasticity can be determined through the following Model (3):

$$\bar{\rho}(\underline{\rho}) = \max(\min) \frac{V^T W X_0}{U^T \Delta Y_0}$$

$$s.t. \begin{cases} U^T Y_j - V^T X_j + \mu_0 \le 0, j = 1, ..., n \\ U^T Y_0 - V^T X_0 + \mu_0 = 0 \\ V^T X_0 = 1 \\ U \ge \mathbf{0}, V \ge \mathbf{0}, \mu_0 \text{ free} \end{cases} \tag{3}$$

where $U = (u_1, u_2, ..., u_s)^T$ and $V = (v_1, v_2, ..., v_m)^T$ are vectors of multipliers, and $\Delta = diag\{\delta_1, \delta_2, ..., \delta_s\}$ and $W = diag\{\omega_1, \omega_2, ..., \omega_m\}$ are matrixes of inputs and outputs directions respectively.

Based on the optimal solutions of Model (3), we have the following procedure for determining the directional RTS of DMU (X_0, Y_0) in the direction of $(\omega_1, ..., \omega_m)^T$ and $(\delta_1, ..., \delta_s)^T$.

(1) The directional RTS to the "right" of DMU (X_0, Y_0): (a) $\underline{\rho}(X_0, Y_0) > 1$, increasing directional RTS prevails; (b) $\underline{\rho}(X_0, Y_0) = 1$, constant directional RTS prevails; (c) $\underline{\rho}(X_0, Y_0) < 1$, decreasing directional RTS prevails;

(2) The directional RTS to the "left" of DMU (X_0, Y_0): (a) $\bar{\rho}(X_0, Y_0) > 1$, increasing directional RTS prevails; (b) $\bar{\rho}(X_0, Y_0) = 1$, constant directional RTS prevails; (c) $\bar{\rho}(X_0, Y_0) < 1$, decreasing directional RTS prevails; (d) if

Model (3) has no optimal solution, there is no data to determine the directional RTS to the "left" of DMU (X_0, Y_0).

For inefficient or weakly efficient DMUs, we can project them onto the strongly efficient frontier using DEA models so that we can estimate the directional RTS to the "right" and "left" of them according to the directional RTS of these projections.

Model (3) is fractional programming and difficult to solve, so we transform them into equivalent mathematical programming (Model (4)) through Charnes-Cooper transformation (Charnes et al., 1962).

$$\begin{aligned} &\bar{\rho}\left(\underline{\rho}\right) = \max\left(\min\right)\Gamma^T X_0 \\ &s.t.\begin{cases} \Lambda^T\Delta^{-1}Y_j - \Gamma^T W^{-1}X_j + \mu_0^{'} \le 0, j = 1,\ldots,n \\ \Lambda^T\Delta^{-1}Y_0 - \Gamma^T W^{-1}X_0 + \mu_0^{'} = 0 \\ \Gamma^T W^{-1}X_0 = t \\ \Lambda^T Y_0 = 1 \\ \Gamma \ge \mathbf{0}, \Lambda \ge \mathbf{0}, t \ge 0, \mu_0^{'} \text{ free} \end{cases} \end{aligned} \tag{4}$$

Solving Model (4), we can obtain the directional scale elasticity and directional RTS of DMU (X_0, Y_0) and its optimal inputs direction.

In the process of analyzing RTS, the concept of congestion effect is often involved. Congestion effect means the reduction of one (or some) input(s) will result in the increase of maximum possible of one (or some) output(s) under the premise that other inputs or outputs do not become deteriorated (Cooper et al, 2004). Essentially, congestion effect describes the issue of excessive inputs (Wei and Yan, 2004). Färe and Grosskopf (1983, 1985) investigated congestion effect using quantitative methods and proposed corresponding DEA models to deal with this issue. Soon after that, Cooper et al. (1996) proposed another model to study congestion effect. Cooper et al. (2001) compared the similarities and differences of the above two models. Wei and Yan (2004) and Tone and Sahoo (2004) built a new DEA model based on the new production possibility set under the assumption of weak disposal to detect the congestion effect of DMUs. The above methods are all based on the idea of radial changes in all inputs. Yang (2012) argued that due to the complexity of research activities in research institutions, it often can be observed that production factors are not necessarily tied together proportionally, and inputs change non-proportionally. Based on this thinking, he introduced the concept of directional congestion under the Pareto preference, and give specific formulations and models. The methods he proposed are as follows.

For the strong efficient DMU$\left(X_0,Y_0\right)$ on the strongly efficient frontier of the production possibility set determined in Model (5), its directional scale elasticity can be determined through the following Model (6):

$$\max \theta = \theta_0$$

$$s.t.\begin{cases} \sum_j \lambda_j x_{ij} = x_{i0}, i = 1,...,m \\ \sum_j \lambda_j y_{rj} - s_r^+ = \theta_0 y_{r0}, r = 1,...,s \\ \sum_j \lambda_j = 1 \\ \lambda_j, s_r^+ \geq 0, r = 1,...,s; i = 1,...,m; j = 1,...,n \end{cases} \tag{5}$$

$$\bar{\rho}\left(\underline{\rho}\right) = \max\left(\min\right)\frac{V^T W X_0}{U^T \Delta Y_0}$$

$$s.t.\begin{cases} U^T Y_j - V^T X_j + \mu_0 \leq 0, j = 1,...,n \\ U^T Y_0 - V^T X_0 + \mu_0 = 0 \\ V^T X_0 = 1 \\ U \geq \mathbf{0}, V, \mu_0 \text{ free} \end{cases} \tag{6}$$

We transform Model (6) into equivalent mathematical programming (Model (7)) through Charnes-Cooper transformation (Charnes et al., 1962) as follows.

$$\bar{\rho}\left(\underline{\rho}\right) = \max\left(\min\right)\Gamma^T X_0$$

$$s.t.\begin{cases} \Lambda^T \Delta^{-1} Y_j - \Gamma^T W^{-1} X_j + \mu_0^{'} \leq 0, j = 1,...,n \\ \Lambda^T \Delta^{-1} Y_0 - \Gamma^T W^{-1} X_0 + \mu_0^{'} = 0 \\ \Gamma^T W^{-1} X_0 = t \\ \Lambda^T Y_0 = 1 \\ \Lambda \geq \mathbf{0}, t \geq 0, \Gamma, \mu_0^{'} \text{ free} \end{cases} \tag{7}$$

Based on the results of Model (7), we have the following procedure for determining the congestion effect of strongly efficient DMU$\left(X_0,Y_0\right)$ on strongly efficient frontier of $P_{convex}\left(X,Y\right)$ in the direction of $\left(\omega_1,...,\omega_m\right)^T$ and $\left(\delta_1,...,\delta_s\right)^T$.

(1) If there exists optimal solution in Model (7) and the optimal value of objective function $\underline{\rho}\left(X_0,Y_0\right)<0$, directional congestion effect occurs to the "right" of the DMU$\left(X_0,Y_0\right)$. If there

does not exist optimal solution in Model (7), there is no data to determine the directional congestion effect to the "right" of DMU$\left(X_0, Y_0\right)$.

(2) If there exists optimal solution in Model (7) and the optimal value of objective function $\bar{\rho}\left(X_0, Y_0\right) < 0$, directional congestion effect occurs to the "left" of the DMU$\left(X_0, Y_0\right)$. If there does not exist optimal solution in Model (7), there is no data to determine the directional congestion effect to the "left" of DMU$\left(X_0, Y_0\right)$.

For inefficient or weakly efficient DMUs, we can project them onto the strongly efficient frontier using DEA models so that we can detect the directional congestion effect to the "right" and "left" of them according to those of these projections.

Analysis results of directional RTS and direction congestion effect

Directional RTS

Firstly, we determine the strongly efficient frontier using input-based BCC-DEA model (Model (8)) with radial measurement.

$$\min\left\{\theta = \theta_0 - \varepsilon\left(\sum_r s_r^+ + \sum_i s_i^-\right)\middle| \begin{array}{c} \sum_j \lambda_j x_{ij} + s_i^- = \theta_0 x_{i0}; \sum_j \lambda_j y_{rj} - s_r^+ = y_{r0}, \\ \sum_j \lambda_j = 1, \lambda_j, s_i^-, s_r^+ \geq 0, i = 1,...,m, r = 1,...,s, j = 1,...,n \end{array}\right\} \tag{8}$$

According to Model (8), we can get the projections of 15 biological institutes on the strongly efficient frontier (See Table 2) through the following formula (9).

$$\tilde{x}_{i0} \leftarrow \theta_0^* x_{i0} - s_i^{-*}, i = 1,...,m; \tilde{y}_{r0} \leftarrow y_{r0} + s_r^{+*}, r = 1,...,s \tag{9}$$

Secondly, we can determine the directional RTS of 15 biological institutes in CAS in 2010 using the methods mentioned in Section 2. We take DMU_1 and DMU_2 as examples. Without loss of generality, we set the outputs direction as $\delta_1 = \delta_2 = \delta_3 = 1$. See Figure 1 ~ Figure 4.

Through the above analysis, we have the following findings.

(a) The directional RTS to the "right" of DMU_1 and DMU_2

(a-1) For DMU_1, on the basis of existing inputs, if Staff and Res. Expen. increase in any proportion (under Pareto preference), decreasing directional RTS prevails on DMU_1, i.e., DMU_1 locates on the region with decreasing directional RTS in any direction of inputs increase. See Figure 1.

(a-2) For DMU_2, on the basis of existing inputs, if Staff and Res. Expen. increase in any proportion (under Pareto preference), decreasing directional RTS prevails on DMU_2, i.e., DMU_2 locates on the region with decreasing directional RTS in any direction of inputs increase. See Figure 3.

Table 2: The projections of 15 biological institutes on the strong efficient frontier.

DMU	Inputs		Outputs		
	Staff (FTE)	*Res.Expen. (RMB million)*	*SCI Pub. (Number)*	*High Pub. (Number)*	*Grad.Enroll. (Number)*
DMU_1	640	253.7000	325	105	604
DMU_2	367	251.1500	368	109	477
DMU_3	172	91.7400	207	66	241
DMU_4	359.9457	158.1071	213.4442	71.7412	323.5013
DMU_5	370.0233	239.2184	262.9233	75.2590	391.9738
DMU_6	425.3855	253.0375	247.3208	72.8561	433.2195
DMU_7	175.5478	76.2936	135.5357	38.5334	162.0547
DMU_8	1910	930.4400	785	323	1488
DMU_9	482.3645	301.9380	319.6222	99.1703	466.4210
DMU_{10}	172.0000	91.7400	191.5051	61.2727	210.1414
DMU_{11}	289	182.0400	216	63	481
DMU_{12}	184.6750	101.5225	207.5731	65.5560	266.1415
DMU_{13}	172.0000	91.7400	170.1353	53.1266	195.7481
DMU_{14}	300.3851	196.6932	286.4286	88.8774	370.7413
DMU_{15}	180	56.9100	83	17	126

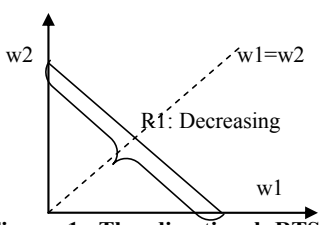


Figure 1. The directional RTS to the right of DMU_1

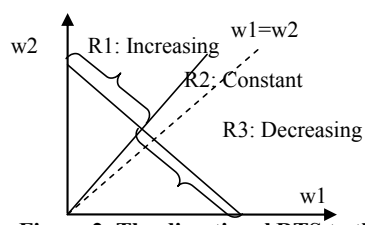


Figure 2. The directional RTS to the left of DMU_1

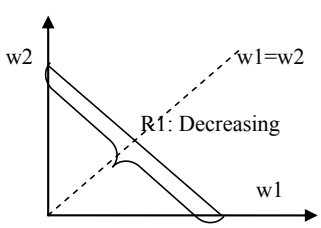


Figure 3. The directional RTS to the right of DMU_2

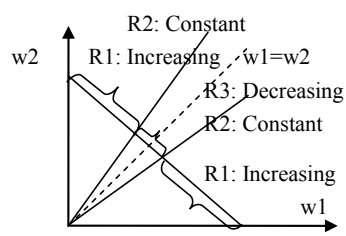


Figure 4. The directional RTS to the left of DMU_2

(b) The directional RTS to the "left" of DMU_1 and DMU_2

(b-1) For DMU_1, on the basis of existing inputs, if Staff and Res. Expen. decrease in radial proportion, decreasing directional RTS prevails. If the proportion of Staff and Res. Expen. locates in the area R1 in Figure 2, increasing directional RTS prevails. If the proportion of inputs decrease locates in the area R3, decreasing

directional RTS prevails. If the proportion of inputs decrease locates in the area R2, constant directional RTS prevails. See Figure 2.
(b-2) For DMU_2, on the basis of existing inputs, if Staff and Res. Expen. decrease in radial proportion, decreasing directional RTS prevails. If the proportion of Staff and Res. Expen. locates in the area R1 in Figure 4, increasing directional RTS prevails. If the proportion of inputs decrease locates in the area R3, decreasing directional RTS prevails. If the proportion of inputs decrease locates in the area R2, constant directional RTS prevails. See Figure 4.
Similarly, we can have the directional RTS to the "right" and "left" of other DMUs.

Directional congestion effect

Firstly, we detect the congestion effect of 15 biological institutes using WY-TS model (Wei & Yan, 2004; Tone & Sahoo, 2004) based on the input-output data of these institutes. We can see that congestion effect occurs on DMU_4, DMU_5, DMU_6, DMU_7, DMU_9, DMU_{12}, DMU_{13} and DMU_{14}. See Table 3 for details.

Table 3. The congestion effect of 15 biological institutes using WY-TS model.

DMUs	*Outputs*			*Inputs*		*Congestion effect* $\varphi = \eta^* / \theta^*$ *(WY-TS model)*
	SCI Pub. (Number)	*High Pub. (Number)*	*Grad.Enroll. (Number)*	*Staff (FTE)*	*Res.Expen. (RMB million)*	
DMU_1	325	105	604	640	253.7	1
DMU_2	368	109	477	367	251.15	1
DMU_3	207	66	241	172	91.74	1
DMU_4	256	62	388	435	189.63	0.9799
DMU_5	259	96	500	472	395.86	0.9177
DMU_6	216	93	553	543	497.32	0.8719
DMU_7	112	39	190	236	89.45	0.9770
DMU_8	785	323	1488	1910	930.44	1
DMU_9	385	125	417	608	537.2	0.8889
DMU_{10}	118	36	235	198	111.28	1
DMU_{11}	216	63	481	289	182.04	1
DMU_{12}	125	37	267	335	101.85	0.9968
DMU_{13}	189	66	232	356	113.97	0.9064
DMU_{14}	313	64	302	413	214.94	0.9540
DMU_{15}	83	17	126	180	56.91	1

Data Source: (1) Monitoring data of institutes in CAS, 2011; (2) Statistical yearbook in CAS, 2011.

Secondly, we can analyze the directional congestion effect of the above DMUs using the methods mentioned in Subsection 2 (Directional congestion effect). We take DMU_1 and DMU_9 as examples. Without loss of generality, we set the outputs direction as $\delta_1 = \delta_2 = \delta_3 = 1$ and we can have the directional congestion effect of these two DMUs in different inputs directions. See Table 4 for details.

Table 4. The directional congestion effect of DMU1 and DMU9 in different inputs directions.

ω_1	ω_2	DMU_1				DMU_9			
		$\underline{\rho}$	$\overline{\rho}$	*Directional congestion effect (right)*	*Directional congestion effect (left)*	$\underline{\rho}$	$\overline{\rho}$	*Directional congestion effect (right)*	*Directional congestion effect (left)*
0.3	1.7	0.68	6.81	No	No	-18.81	-0.44	Yes	Yes
0.5	1.5	0.66	4.86	No	No	-13.44	-0.18	Yes	Yes
0.7	1.3	0.64	2.92	No	No	-8.06	0.08	Yes	No
0.9	1.1	0.41	1.07	No	No	-2.69	0.49	Yes	No
1	1	0	0.92	No	No	0	1.82	No	No
1.1	0.9	-0.97	0.8	Yes	No	0.15	3.5	No	No
1.3	0.7	-2.92	0.68	Yes	No	0.44	8.06	No	No
1.5	0.5	-4.86	0.63	Yes	No	0.73	13.44	No	No
1.7	0.3	-6.81	0.6	Yes	No	0.93	18.81	No	No

Based on the above analysis, we can find that congestion effect occurs on DMU_9 when using WY-TS model. However, for DMU_9, directional congestion effect occurs in certain directions (e.g., $\omega_1 = 1.7, \omega_2 = 0.3; \delta_1 = \delta_2 = \delta_3 = 1$) and does not occur in other directions (e.g., $\omega_1 = 0.3, \omega_2 = 1.7; \delta_1 = \delta_2 = \delta_3 = 1$) . The deep reason for this phenomenon should be investigate thoroughly. The congestion effect does not occur on DMU_1 in WY-TS model and the directional congestion effect does not occur to the left of DMU_1 also. Similarly, we can analyze the directional congestion effect for other DMUs.

From the above analysis and findings, we can see that (1) the regions of increasing (constant, decreasing) directional RTS and optimal inputs direction can be detected through the methods in Section 2 so that DMs can refer to this information when making related decisions or S&T policies. We take DMU_1 and DMU_2 in Section 3 as examples, decreasing RTS prevails on these two DMUs when inputs increase so the inputs of these two DMUs should be reduced to improve their scale efficiencies. In the directions of inputs decrease, if the proportion of Staff and Res. Expen. locates in the area R1 in Figure 4 and Figure 6 for DMU_1 and DMU_2 respectively, increasing directional RTS prevails so that their scale efficiencies can be improved; (2) congestion effect occurs on DMU_4, DMU_5, DMU_6, DMU_7, DMU_9, DMU_{12}, DMU_{13} and DMU_{14} using traditional WY-TS model. However in certain directions, the directional congestion effect does not occur on the same DMUs. Therefore, these institutes should analyze their own strengths carefully and identify their own strength carefully and identify their own development path for resources so that their scale efficiencies could be improved further.

Conclusions and discussions

This paper investigates the directional returns to scale of 15 biological institutes in CAS. Firstly, the input-output indicators are proposed, including Staff, Research funding, SCI papers, High-quality papers and Graduates training. Secondly, this paper uses the methods proposed by Yang (2012) to analyze the directional returns to scale, optimal input direction and the effect of directional congestion of biological institutes in CAS. Based on the analytical results, we have the following findings: (1) we can detect the regions of increasing (constant, decreasing) directional returns to scale for each biological institute. This information can be used as one of the basis of decision-making on organizational adjustment; (2) we find that the effect of congestion and directional congestion occurs in several biological institutes. On this occasion, the outputs of these institutes will decrease with the inputs increase. So, these institutes should analyze the deep reason for the occurrence of congestion effect so that science and technology (S&T) resources can be used more effectively.

Acknowledgements

We acknowledge the support of the National Science Foundation of China (No. 71201158).

References

Atici, K.B.& Podinovski, V.V. (2012). Mixed partial elasticities in constant returns-to-scale production technologies. *European Journal of Operational Research*, 220, 262-269.

Banker, R.D. (1984). Estimating the most productive scale size using data envelopment analysis. *European Journal of Operational Research*, 17, 35-44.

Banker, R.D., Charnes, A. & Cooper, W.W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science,* 30(9), 1078-1092.

Charnes, A. & Cooper, W.W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9, 181-196.

Cooper, W.W., Seiford, L.M. & Zhu, J. (2004). Handbook on data envelopment analysis. Kluwer Academic Publishers, Massachusetts, USA.

Cooper, W.W., Huang, Z. & Li, S. (1996). Satisficing DEA models under chance constraints. *Annals of Operations Research*, 66, 279-295.

Cooper, W.W., Gu, B.S. & Li, S.L. (2001). Comparisons and evaluation of alternative approaches to the treatment of congestion in DEA. *European Journal of Operational Research*, 132, 62-67.

Färe, R. & Grosskopf, S. (1983). Measuring congestion in production. *Zeitschrilft für Nationalökonomie*, 257-271.

Färe, R. & Grosskopf, S. (1985). A nonparametric cost approach to scale efficiency. *Scandinavian Journal of Economics,* 87, 594-604.

Geisler, E. (2000). The metrics of science and technology. Quorum Books.

Fox, K. J. (2002). Efficiency in the Public Sector. Springer.
Keeney, R.L. & Raiffa, H. (1976). Decisions with Multiple Objectives. New York: Wiley.
Keeney, R.L. & Gregory, R.S. (2005). Selecting Attributes to Measure the Achievement of Objectives. *Operations Research*, 53(1), 1-11.
Li, X.X. (2005). Practice and thoughts on performance evaluation in China's state-run institutes of scientific research. *Bulletin of the Chinese Academy of Sciences*, 20(5), 395-398.(in Chinese).
Liu, W.B., Zhang, D.Q., Meng, W., Li, X.X. & Xu, F. (2011). A study of DEA models without explicit inputs. *Omega-The International Journal of Management Science,* 39, 472-480.
Meng, W., Mingers, J. & Liu, W.B. (2007). Studies on framework for science-technology evaluation using soft system methodology. *Science research management*, 28(2), 1-8. (in Chinese).
Meng, W., Huang, M. & Liu, W.B. (2006). Scale efficiency analysis of institutions using DEA models. *Science research management*, 27(4), 19, 20-25. (in Chinese).
Mingers, J., Liu, W.B. & Meng, W. (2009). Using SSM to Structure the Identification of Inputs and Outputs in DEA. *Journal of the Operational Research Society*, 60(2), 168-179.
Pindyck, R.S.& Rubinfeld, D.L.(2000). Microeconomics. Prentice Hall, 4th Edition.
Podinovski, V.V.&Førsund, F.R. (2010). Differential Characteristics of Efficient Frontiers in Data Envelopment Analysis. *Operations Research,* 58(6), 1743-1754.
Poister, T. H. (2003). Measuring performance in public and nonprofit organizations. San Francisco, Calif.: Jossey-Bass.
Roper, S., Hewitt-Dundas, N. & Love, J. H. (2004). An ex ante evaluation framework for the regional benefits of publicly supported R&D projects. *Research Policy,* 33, 487–509.
Tone, K. & Sahoo, B.K. (2004). Degree of scale economies and congestion: a unified DEA approach. *European Journal of Operational Research,* 158, 755-772.
Wei, Q.L. & Yan, H. (2004). Congestion and returns to scale in data envelopment analysis. *European Journal of Operational Research,* 153, 641-660.
Yang, G.L. (2012). On relative efficiencies and directional returns to scale for research institutions. Ph.D thesis, University of Chinese Academy of Sciences, Beijing. (in Chinese).
Zhang, D.Q., Yang, G.L. & Li, X.X.(2011). Strategy maps and performance measures for national research institutes. *Studies in Science of Science*, 29(12), 1835-1844. (in Chinese).
Zhou, W. & Li, J.S. (2009a). Higher education scale efficiency in central China: An empirical study based on Data Envelopment Analysis. *Fudan Education Forum*, 7(4), 53-57. (in Chinese).

Zhou, W. & Li, J.S. (2009b). Empirical Analysis on Scale Efficiency of Key Universities in Western Region. *Journal of Xidian university (Social sciences edition),* 19(5), 115-120. (in Chinese).

DISCIPLINARY DIFFERENCES IN TWITTER SCHOLARLY COMMUNICATION

Kim Holmberg[1] and Mike Thelwall[2]

[1] k.holmberg@wlv.ac.uk | [2] m.thelwall@wlv.ac.uk
School of Technology, University of Wolverhampton
Wulfruna Street, Wolverhampton WV1 1LY, UK

Abstract
This paper investigates disciplinary differences in how researchers use the microblogging site Twitter. Tweets from researchers in five disciplines (astrophysics, biochemistry, digital humanities, economics, and history of science) were collected and analyzed both statistically and qualitatively. The results suggest that researchers tend to share more links and retweet more than the average Twitter users in earlier research. The results also suggest that there are clear disciplinary differences in how researchers use Twitter. Biochemists retweet substantially more than researchers in the other disciplines. Researchers in digital humanities use Twitter more for conversations, while researchers in economics share more links than other researchers. The results also suggest that researchers in biochemistry, astrophysics and digital humanities are using Twitter for scholarly communication, while scientific use of Twitter in economics and history of science is marginal.

Conference Topic
Webometrics (Topic 7) and Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2)

Introduction
Scholarly communication is changing as researchers increasingly use social media to discover new research opportunities, discuss research with colleagues and disseminate research information. Traditionally, scholarly communication may be seen as a process that starts with a research idea and ends with a formal peer reviewed scientific publication. During this process, ideas may be informally discussed with colleagues or presented at seminars and conferences and, after publication, the results may be read and formally cited by other researchers. With the advent of the web both formal and informal scholarly communication has changed. Because of the web, ideas can be more easily and quickly discussed with colleagues over email or video conferencing tools and articles can be published on the web in institutional repositories, online full text databases or online open access journals. Now it seems that social media are triggering another evolution of scholarly communication.
Citations are important in scholarly communication. They are the link that connects earlier research to new research. They indicate use of earlier research,

and hence it can be argued that they indicate something about the value of the cited research. Citations are also part of the academic reward system (Merton, 1968), with highly cited authors tending to be recognized as having made a significant contribution to science. Counting citations is at the core of scientometric methods; they have been used to measure scholarly work and intellectual influence and to map collaboration networks between scholars (Moed et al., 1995; Cole, 2000; Borgman, 2000). However, citations can be created for many different reasons (Borgman & Furner, 2002) and because both publishing and citation traditions vary between disciplines, new ways are needed to measure the visibility and impact of research. In this context, social media may generate new ways to measure scientific output (Priem & Hemminger, 2010). Social bookmarking sites such as Connotea and CiteULike, or recommendation systems like Reddit and Digg, may prove to be fruitful sources for new scientific visibility metrics (Priem & Hemminger, 2010). One of the new social media services that researchers can use in scholarly communication and that has some potential in providing new ways to measure research impact is Twitter.

Twitter is a real-time microblog network; users can publish their opinions, ideas, stories, and news in messages that are up to 140 characters long. Twitter had over 500 million users worldwide in 2012 (Semiocast, 2012) and has gained a lot of media coverage as an efficient and rapid tool for sharing emergency information (Ash, 2011). The service has also been researched from a wide range of disciplines and research goals from political elections (Hong and Nadler, 2012), electronic word of mouth (Jansen et al. 2009), and natural disasters (Earle et al., 2011), to protest movements (Harlow and Johnson, 2011) and health information sharing (Scanfeld et al., 2010). Some earlier research has investigated how researchers are using Twitter at conferences (e.g., Ross et al., 2010; Letierce et al., 2010; Weller & Puschmann, 2011; Weller, Kröge, & Puschmann, 2011) but, to the best of our knowledge, scholarly communication in general, rather than for specific purposes, on Twitter has not been researched before, with the partial exception of a small-scale study of tweets with links from 28 scholars (Priem & Costello, 2010). To fill this gap, the current study investigates how researchers in five diverse disciplines use Twitter. The results can both help researchers to understand how others are using Twitter, and hence how they may use it, and also help scientometricians to decide if and how Twitter can be used as a scientometric data source.

Literature review

Since Twitter is relatively new, this review covers general aspects of its use as well as its scholarly context.

General use of Twitter

Twitter has three special features that aid communication. Forwarded tweets are called retweets and are usually marked by RT or MT for modified tweet. A second feature is the use of @ followed by a username. This can be used to send a

message to another Twitter user or users. Including *@username* in a tweet can also let that person know that he or she has been mentioned. The third feature is the use of hashtags. By adding #-character followed by a freely chosen word the user can tag the tweet and hence group it together with other tweets about the same topic. Hashtags are frequently used at scientific conferences as a convenient way to collect all tweets about the conference together because users can set up real-time monitoring of hashtags through Twitter to ensure that they are able to quickly access relevant tweets. Because of the unique features of these types of tweets (RT, @username, #hashtag) they can be extracted automatically from a corpus of tweets.

In a large scale study on Twitter Ediger et al. (2010) discovered that retweeting on Twitter has power law-like characteristics: a few tweets are extensively retweeted whereas most tweets are not retweeted or are only retweeted a few times. Ediger et al. (2010) found that retweets tend to refer to a relatively small group of original tweets, which is a behavior more common in one-to-many broadcasting rather than many-to-many communication patterns. Many-to-many broadcasting patterns were also identified in their study but in significantly smaller subsets of the complete graph they had built from the collected tweets. This supports the belief that we are moving away from broadcasting and broadcasted media towards networked media and information dissemination in networks (e.g. Boyd, 2010). Twitter supports information sharing in networks because of the social networks created by users following other users.

Roughly 30% of all tweets have been found to be conversational in nature (Honeycutt & Herring, 2009), in the sense of using the @ convention. Huberman et al. (2008) arrived at a similar number (25.4%) in an earlier study. Honeycutt and Herring (2009) investigated tweets containing the @-sign and concluded that a clear majority (90%) of tweets containing the sign were conversational. The study therefore showed that some, but perhaps not all, conversational tweets can fairly easily be collected from Twitter, as they are usually identifiable by the @-sign.

In their sample of 720,000 random tweets Boyd et al. (2010) found that about one third of tweets were addressing someone (using *@username* in the tweet), about one fifth contained a URL, 5% contained a hashtag and only 3% were retweets. In a random sample of retweets they discovered that over half of the retweets contained a URL and that about one fifth contained a hashtag. The use of hashtags and URLs was therefore significantly higher in retweets than in tweets. Suh et al. (2010) found that only about 20% of tweets contain a URL or URLs and that almost 30% of retweets contain a URL or URLs. They also concluded that hashtags and the type of hashtags have an impact on "retweetability". The number of followers also has an impact, which is quite expected. The more followers a user has the more likely his or her tweets are to be retweeted.

People retweet for a variety of different reasons. Earlier research (Boyd et al., 2010) has shown that people retweet because they want to spread information to new audiences or a specific audience of followers, they may retweet because they

want to comment on someone's tweet or make the original writer aware that they are reading their tweets. People also retweet to publicly agree with or to validate someone's thoughts, to be friendly, and to refer to less popular content in order to give it some visibility, but also for egoistic reasons such as to gain more followers or to gain reciprocity. People also retweet to save tweets for later access.

Social media and scholarly communication

The change in scholarly communication has not been rapid because many researchers are cautious in changing traditional scholarly communication patterns. Weller (2011, p. 55) writes that "*... research is at the core of what it means to be a scholar, and issues around quality and reliability are essential in maintaining the status and reputation of universities. A cautious approach is therefore not surprising as researchers seek to understand where the potential of these new tools can enhance their practice, while simultaneously maintaining the key characteristics of quality research*". But as more and more scholars start to use social media it is possible that it may have an impact on tenure and promotion processes at academic institutions (Gruzd et al., 2011).

Social media has become important for discovering and sharing research. Scholars use tools such as wikis for collaborative authoring, tools for conferencing and instant messaging for conversations with colleagues, scheduling tools to schedule meetings and various tools to share images and videos (Rowlands et al., 2011). Microblogging had not yet gained significant popularity among scholars, as only 9.2% stated that they used microblogging in their research. Rowlands et al. (2011) showed that there are some disciplinary differences in how researchers are using social media in general, as natural scientists in their study were the biggest users. However, they suggest that it may not take long before social scientists and humanities researchers catch up. While there were some differences between disciplines, differences between how different age groups use social media were not discovered.

Scholarly communication and information sharing is changing as academics increasingly use Social Networking Sites (SNSs) such as Facebook and Twitter for professional purposes. SNSs may promote information sharing (Forkosh-Baruch & Hershkovitz, 2011) in both formal and informal ways. It has been shown that scholars use Twitter to cite to scientific articles and hence Twitter could potentially be used to measure scholarly impact (Priem & Costello, 2010). Weller and Puschmann (2011) and Weller, Kröge and Puschmann (2011) considered all tweets containing one or more URLs as a form of citation, while Priem and Costello (2010) considered a tweet as a citation only if it included a URL directly to a scientific article or to an intermediary web page that has a link to a scientific article. In a dataset collected from 28 researchers' tweets Priem and Costello (2010) found that 6% of the tweets including a URL were links to peer-reviewed articles or to web pages that link to peer-reviewed articles. However, sharing links and citations are not the only scholarly activity on Twitter. At scientific conferences for instance, Twitter is often used as a backchannel to share

notes and resources, and for discussions about topics at the conference (e.g. Ross et al., 2010; Letierce et al., 2010; Weller & Puschmann, 2011; Weller, Kröge, & Puschmann, 2011).

Research Questions

The goal of the research is exploratory and descriptive, driven by the following basic research questions.

1. What do researchers typically tweet about?
2. Are there disciplinary differences in the types of tweet sent by researchers?

The approach used to answer these questions was to gather a large corpus of tweets sent by selected researchers in five different disciplines and then to apply a content analysis to a random sample of tweets to identify the types of content posted.

Methods

The main purpose of this research is to investigate disciplinary differences in the use of Twitter in scholarly communication and sharing of scientific information. The five disciplines chosen for this are astrophysics, biochemistry, digital humanities, economics, and history of science. These were chosen to represent variations in the traditional publishing and scholarly communication patterns and to represent disciplines of varying size and focus.

The differences were investigated by collecting tweets sent by researchers from each of the disciplines. First, the most productive researchers based on the number of publications from each discipline were queried from the ISI Web of Knowledge (WoK) database. We chose to search for most productive rather than most cited researchers in order to find seasoned, established researchers that already have had a long career, not just the most influential or prestigious (assuming that citations can indicate this). This was achieved through a topical search for each discipline. Then a list of the most productive authors based on a count of WoS records was extracted. Next we checked which of the top authors were active on Twitter. We visited the homepages of the authors and searched for them on Twitter. However, this was a very time consuming method and in the end it was not possible to find many top researchers using Twitter in this way; hence Twitter's search function and discipline relevant keywords (e.g. astrophysics, biochemistry, etc.) were used to find other relevant researchers from the selected disciplines. The selection criterion was that the person should be active on Twitter and clearly be an established researcher in one of the chosen fields. This meant that only tenure tracked researchers were chosen and for instance PhD students were excluded from the sample. This information was obtained from the persons' profiles on Twitter and in cases where this was not mentioned in the profile the

user was not included to the sample. Then a snowball sampling method was used, which proved to be a good method to collect tweeting researchers as many researchers on Twitter follow other researchers in their own field. In the end we found 45 researchers in astrophysics, 45 in biochemistry, 51 in digital humanities, 45 in economics, and 42 in history of science. The 20 most productive researchers from WoK included only 1 Twitter user in astrophysics, 2 in biochemistry, 6 in digital humanities, none in economics, and 1 in history of science. Hence the results do not reflect top researchers in the disciplines but established Twitter using researchers instead.

The tweets were collected between 4 March 2012 and 16 October 2012, although some earlier tweets will be included from the first queries. Twitter was queried at least daily for updates by the selected users by a program accessing the main Twitter API. A few days were dropped due to system malfunctions but since the queries could retrieve tweets from the missing period it seems unlikely that any tweets were lost and so the collection should be quite comprehensive. However, Twitter restricts the collection of tweets sent by certain users to approximately 3,200 tweets. This means that for users that are not very active on Twitter we can collect all their tweets, while from active users we only get about the 3,200 latest tweets.

Within the time period of data collection a total of 59,742 astrophysics tweets, 40,128 biochemistry tweets, 89,106 digital humanities tweets, 57,673 economics tweets, and 58,414 history of science tweets were sent by the researchers. There were disciplinary differences in the amount of tweeting: in astrophysics the researchers posted on average 1328 tweets each, in biochemistry 892 tweets per researcher, in digital humanities 1747 tweets per researcher, in economics 1282 tweets per researcher, and in history of science 1391 tweets per researcher. This shows that biochemists were least active Twitter users, while digital humanities researchers were the most active.

From each discipline 200 tweets were randomly selected using a random number generator for a faceted content analysis. The 200 tweets from each of the disciplines were grouped into four categories for facet 1: *Retweets*, *Conversations*, *Links*, and *Other*. The category *Retweets* included tweets that were identified by RT or MT (modified tweets), or tweets that were otherwise marked as having been sent via someone else. The *Conversations* category contained tweets that were not retweets and that were identified by *@username*, indicating that the tweet was sent to someone. The categories do not therefore include any conversations that have been held without using the @username convention, but as earlier research suggests (Honeycutt & Herring, 2009), it should be possible to collect most of the conversational tweets with this method. The *Links* category contained tweets that were not retweets or conversations and contained a URL (usually shortened). The *Other* category contained all the remaining tweets.

For facet 2, the tweets were categorized according to scientific and disciplinary content. These categories were: *Scholarly communication*, *Discipline-relevant*, *Not clear*, and *Not about science* (Table 1). The first category contained tweets

that clearly were about science and clearly on topic about the chosen discipline. Tweets in the second category were clearly about the discipline but not clearly about science in the sense of conducting or discussing scientific research. In the third category it was not clear if the tweets were about science or if they even were about the discipline. Tweets in the final category were clearly not about science nor were they about the discipline in question. A conservative approach was used when classifying the tweets. This means that when in doubt a less scientific category was chosen in order to prevent overestimation of the scientific content in the analyzed tweets. Also, every tweet was classified into only one category. The whole sample was coded by the first author and a random set of 25% (50 tweets) of the tweets from each discipline were coded by another researcher to check for inter-coder reliability. After the first round of coding the researchers talked through the cases where they did not agree and refined the coding scheme based on the discussion. Then a second round of coding was conducted with a new random set of 25% of the tweets and the standard Cohen's Kappa statistic was used to assess the reliability of the classification in this second round.

Table 1. Categorizing tweets according to scientific and disciplinary content

Category	Description	Example of tweet
Scholarly communication	Tweets that are clearly scientific and on topic of the discipline. This includes tweets with links to scientific papers or journals, sharing research results, comments, questions and answers of a scientific nature. Tweets in this category clearly have some scientific value for other researchers.	"Decellularized matrix from tumorigenic human mesenchymal stem cells promotes neovascularization... http://t.co/aF6TVFIG" (link to an abstract in PubMed)
Discipline-relevant	Tweets that are clearly on topic of the discipline but are not clearly scientific as described in the category above.	"Fri AM in Asia: Asian stocks already heading downward. 50-50 chance of global recession."
Not clear	Both scientific and disciplinary relevance are not clear. Usually because there is not enough information in the tweet for other judgements. The tweets in this category could be fractions of conversations or short answers to earlier questions from another person.	"@[…] Your welcome :)"
Not about science	Tweets that are clearly not scientific nor on the topic of the discipline. This includes personal tweets, links to photos, comments about everyday life in general, and status updates about what they were doing and where they were at the moment.	"The goddamn mice have been at the wiring of my car again. As a bonus the dealership wi-fi blocks twitter and they have no power outlets."

A chi-square test was used to assess whether the disciplines had overall different proportions of tweets falling in each category. Differences in proportions tests at the fixed level p=0.05 were used to test for differences between disciplines for individual categories. These tests were indicative rather than statistically rigorous, however, because we did not have a prior set of hypotheses to test for and so we could not conduct a small enough number of specific tests to control for errors with a Bonferroni correction other than one that compensated for all possible tests.

Results

There were some disciplinary differences in the types of tweets that were sent (Figure 1), confirmed by a chi-square test (p=0.000). In biochemistry 42% of the tweets were retweets in comparison to about 25% in the other disciplines (sig. p <0.05). Conversations were important in astrophysics (31.5% of the tweets), digital humanities (38%) and history of science (28.5%). The proportions of conversations in biochemistry and economics were much lower in both cases at about 16% (difference between the two sets sig. p <0.05). Conversations in general were roughly twice as important in astrophysics, digital humanities and history of science compared to biochemistry and economics. When collecting random tweets only one part of a conversation is available, which makes it difficult to judge whether conversations are about science or not. An example of an unclear tweet is "@[…] Yup! I will indeed keep you posted." It is possible that the conversation is about science, but it could be about something else too. Economics shared most (sig. p <0.05) links (38%), but sharing links was important also in the other disciplines. In history of science 27% of the tweets were shared links, in astrophysics the amount was 23.5% and in biochemistry 21.5%, but in digital humanities only 15.5% of the tweets were links (sig. lower than all the others except biochemistry, p <0.05). Of course some of the retweets and conversations also contained links, however the purpose of sharing the links in these categories can be assumed to be somewhat different than in tweets that are neither forwarded information (retweets) nor part of conversations between two or more persons. When classifying the tweets according to type the inter-coder agreement was very high; only in two cases out of the 250 tweets that two researchers coded had the researchers coded the tweets differently.

A considerable proportion of the retweets contained links. About 75% of retweets in astrophysics contained one or more links, in history of science 70%, in biochemistry about 68%, in economics about 65% and in digital humanities about 62% of retweets contained links. This clearly shows that researchers in these disciplines frequently share web content and forward information and content they have received from people they follow on Twitter.

The remaining tweets made up between one fifth to one fourth of the total tweets in each discipline (*Other* category).

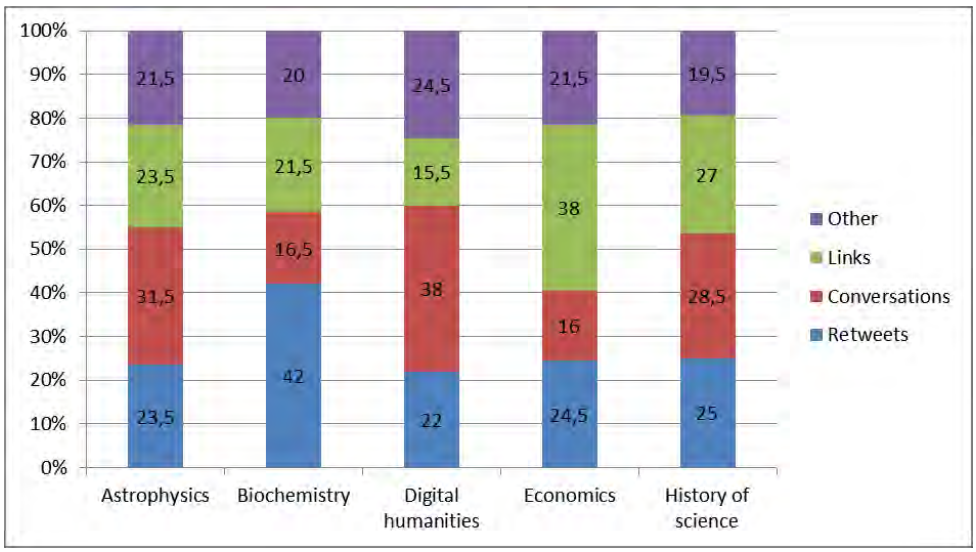


Figure 1. Types of tweets by discipline

There are clear disciplinary differences in the amount of tweets in the scholarly communication category (Figure 2), confirmed by a chi-square test ($p=0.000$). Almost 34% of the tweets in biochemistry were clearly part of scholarly communication (sig. greater than the others, $p < 0.05$), and in astrophysics the number is 23% and in digital humanities 22%. In history of science and economics the number is substantially lower than the others (sig. $p < 0.05$), at 7.5% and 6.5% respectively.

Few economics tweets were clearly for scholarly communication, but many tweets were about economics in general. Some of these may be scholarly communication but it is not clear based just on the tweet. An example of an unclear tweet is the following: "RT @HarvardBiz - Africa's Growth Opportunity - Swaady Martin-Leke and Loic Sadoulet - Harvard Business Review: http://t.co/5WAv7qCJ". The link is to a blog entry in Harvard Business Review from October 2011. The tweet is clearly about economics, but whether the blog entry has scientific value for a researcher is unclear. Economics is a general topic of discussion for citizens and so academics discussing economic issues are not necessarily discussing research, and hence it is difficult to judge whether tweets are about economics or research in economics. Economics had the most tweets that were discipline-relevant (51.5%, sig. $p < 0.05$). The other disciplines had between 22% and 8.5% tweets that were discipline-relevant. While the other disciplines had between 26% and 34% tweets that were not about science nor about the discipline, in history of science 57.5% of tweets were clearly not about science nor about history of science (sig. greater than the others $p < 0.05$). History of science stands out of the group as only 16% of the tweets were for scholarly

communication or discipline-relevant, while the same for other disciplines was substantially higher.

One quarter of the tweets from the random sample were coded twice by two researchers. After the second round of coding the researchers coded the tweets to the same categories in 68.9% of the cases. The standard Cohen's Kappa statistic gave an inter-coder reliability of 0.587, which constitutes as "good" or "moderate" agreement, depending on which interpretation one uses (Fleiss, 1981; Landis & Koch, 1977).

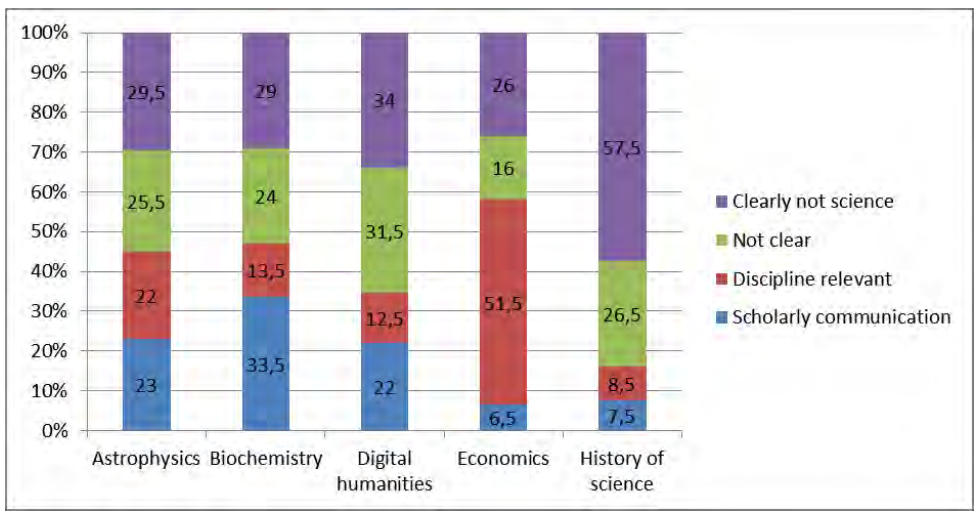


Figure 2. Relevance of tweets by discipline

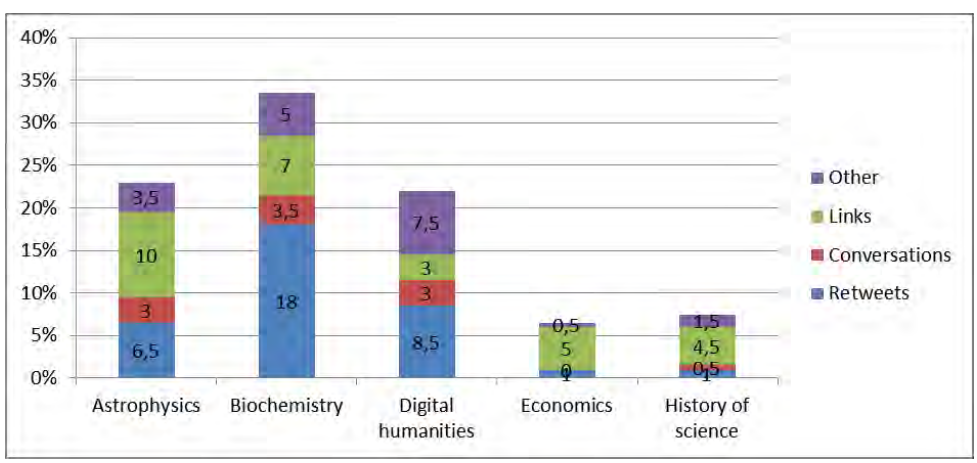


Figure 3. Percentages of scholarly communication tweets by type

All disciplines had retweets for scholarly communication (Figure 3), but especially in biochemistry retweets (18% of all tweets in the discipline) appear to be an important tool to forward scientific information. In economics and history of science the importance of retweets was marginal for scholarly communication. In all disciplines less than 3.5% of the conversations were clearly part of scholarly communication. In fact, none of the conversations in economics and only one conversational tweet in history of science were clearly part of scholarly communication. Both in astrophysics (10%) and in biochemistry (7%) researchers share links to scientific content, while somewhat less scientific links were shared in the other disciplines. Some evidence of scholarly communication was also found in the remaining tweets in the Other category.

An informal content analysis of the tweets from the *Scholarly communication* category showed that the retweets are mainly links to articles in popular science magazines, to blog entries, to newspaper articles, and to promote upcoming events, articles, interviews and radio shows. While almost all of the relevant retweets included links, only four of the retweets contained a link to a scientific paper or to an abstract. In *Conversations* it was not usual to share links, but rather to share opinions, talk science or comment on science facts with colleagues. There were only two tweets with links to scientific papers; one to a publisher's abstract page with a link to full text, and one directly to a pdf file.

In the *Links* category tweets included links to articles in popular science magazines and to blog entries. In addition, 16 tweets contained a link to a scientific paper. Of these four were links directly to the full text files, 5 were to the publishers' page, and 3 were to other online texts that had links to the publishers' page for the article. Of these 16 links to scientific papers 8 were in astrophysics, 4 in biochemistry, 2 in economics, and 1 each in digital humanities and history of science. The remaining links were to an editorial in a scientific journal, a draft of a scientific paper, an abstract in an online database, and to the literature list of an online article. In the *Other* category the tweets were mainly comments and opinions on science facts, promotional or about some workshops or conferences. None of the tweets in this category contained links to scientific articles. A total of 22 links were to scientific papers. This constitutes 2.2% of all tweets, which is somewhat lower than the 6% found by Priem and Hemminger (2010) in their sample.

Discussion and conclusions

In answer to the second research question, the results suggest that there are clear differences in Twitter use between disciplines. Researchers in every discipline retweet, but they do so almost twice as much in biochemistry than in the other disciplines. Researchers forward information substantially more than the average Twitter user does. Boyd et al. (2010) found that only about 3% of tweets were retweets, while in our research we found that on average 27% of the tweets across the five disciplines were retweets. In digital humanities researchers use Twitter more for conversations than in the other disciplines, and substantially more than

in biochemistry and economics. In economics Twitter is used mostly to share links, while this possibility did not seem to be frequently used in digital humanities.

Based on the results it also seems clear that Twitter is used more for scholarly communication in biochemistry and astrophysics (and to some extent in digital humanities) than in economics and history of science. Least evidence of scholarly communication was found among the history of science researchers. Economics proved to be a difficult discipline to evaluate because economics is a common topic of discussions among citizens and because of that researchers discussing economics or sharing news and information about economics, do not necessarily mean that they are involved in scholarly communication.

It seems clear that researchers share more links than the average Twitter users. Both Boyd et al. (2010) and Suh et al. (2010) found that about 20% of tweets contained links, while in our research we discovered that on average 25% of the tweets contained links, and this is excluding the retweets, of which most contained links. The difference between researchers' use of Twitter and the average Twitter user is in particularly clear in the retweets where between 62% and 75% of the researchers forwarded tweets including links to some information resources. In many cases the information shared was related to the discipline, but not necessary to scientific publications. The multitude of different types of information and content shared also shows how researchers are using an abundance of different information sources when keeping themselves up-to-date with news and events in their discipline. How many of these directly benefit their research work is not clear and more qualitative research is needed to fully understand how and why researchers are using social media sites such as Twitter in scholarly communication. In fact, a possible future research direction could be a qualitative investigation about how the researchers themselves in specific disciplines believe that they are using Twitter (and whether that is in correlation with the results discovered in the present study or not) and what kind of possible scholarly benefits they have identified with the microblogging site (for a single discipline, see Priem & Costello, 2010).

Although the biochemistry researchers were the least active Twitter users they were the group that used Twitter most for scholarly communication. Researchers in digital humanities on the other hand used Twitter most actively, but mainly for conversations that were not clearly scientific. Moreover, 57.5% of the tweets by researchers in history of science had nothing to do with science or history of science. These were mainly comments about their everyday lives or status updates about where they were and what they were doing. When analyzing the scholarly communication tweets, few cited research articles directly or indirectly. Only 2.2% of all tweets were like citations in the sense of linking to an academic article. The results suggest that Twitter is for many researchers an important tool in scholarly communication, but it is not frequently used to share information about scientific publications. The results also suggest that disciplinary differences

in the use of Twitter are a fact that has to be taken into account in any future research about scholarly use of Twitter.

Some evidence was discovered that researchers use Twitter to share information about, and links to, scientific articles. However, these were only discovered after the links were manually visited, a procedure that is not reasonable to replicate with a large dataset and for which there are currently no automated procedures for. It is possible to collect all tweets containing specific URLs or top-level domains of links to some publishers article collections, for instance http://www.plosone.org/article/info:doi/ (to PLOS One) or http://www.emeraldinsight.com/journals.htm?issn=0022-0418 (to the Journal of Documentation), but it would not be possible to cover all publishers, online open access journals, institutional repositories and URLs to self-archived papers.

The present research has a number of weaknesses, of which the most significant is in the coding of the tweets. While categorizing the tweets according to type is fairly straight forward, classifying by relevance for scholarly communication is more difficult. Although the Cohen's Kappa value for inter-coder agreement was 0.587 in this research, it is possible that other researchers with background in some of the disciplines in this research might come to a different conclusion regarding the scientific value of some tweets. However, even these tweets should be covered in the first two categories of this research, scholarly communication and disciplinary-relevant, and hence they would have been included as relevant tweets even now. Also, to prevent overestimation of the results we used a conservative approach in the coding, meaning that when in doubt the tweets were coded into a less scientific category. In addition, other fields may have given different results and so, even when the results agree for the five covered here, they cannot be confidently generalized.

Acknowledgements

This manuscript is based upon work supported by the international funding initiative Digging into Data. Specifically, funding comes from the National Science Foundation in the United States (Grant No. 1208804), JISC in the United Kingdom, and the Social Sciences and Humanities Research Council of Canada. The authors wish to thank Andrew Tsou for his help in coding the tweets and the anonymous reviewers for their comments.

References

Ash, T.G. (2011). Tunisia's revolution isn't a product of Twitter or Wikileaks. But they do help. *Guardian*, January 19, 2011. Retrieved March 13, 2011 from http://www.guardian.co.uk/commentisfree/2011/jan/19/tunisia-revolution-twitter-facebook.

Borgman, C.L. (2000). Scholarly communication and bibliometrics revisited. In Cronin, B. & Atkins, H.B. (eds.) *The web of knowledge – A festschrift in honor of Eugene Garfield*. ASIS, 2000.

Borgman, C. & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, vol. 36, no. 1, pp. 2-72.

Boyd, D., Golder, S. & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *Proceedings of the 43rd Hawaii International Conference on System Sciences 2010*. Retrieved March 1, 2011 from http://www.danah.org/papers/TweetTweetRetweet.pdf.

Cole, J.R. (2000). A short history of the use of citations as a measure of the impact of scientific and scholarly work. In Cronin, B. & Atkins, H.B. (eds.) *The web of knowledge – A festschrift in honor of Eugene Garfield*. ASIS, 2000.

Choi, S., Park, J. & Park. H.W. (2012). Using social media data to explore communication processes within South Korean online innovation communities. *Scientometrics*, vol. 90, pp. 43-56.

Earle, P.S., Bowden, D.C. & Guy, M. (2011). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, vol. 54, no. 6, pp. 708-715.

Ediger, D., Jiang, K., Riedy, J., Bader, D.A., Corley, C., Farber, R. & Reynolds, W.N. (2010). Massive social network analysis: Mining Twitter for social good. In *Proceedings of 39th International Conference on Parallel Processing*. Retrieved March 1, 2011 from http://www.cc.gatech.edu/~jriedy/paper-copies/ICPP10-GraphCT.pdf.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley. ISBN 0-471-26370-2.

Forkosh-Baruch, A. & Hershkovitz, A. (2011). A case study of Israeli higher-education institutes sharing scholarly information with the community via social networks. *Internet and Higher Education*, vol. 15, pp. 58-68.

Gruzd, A., Staves, K. & Wilk, A. (2011). Tenure and promotion in the age of online social media. In *Proceedings of the ASIS&T Annual Meeting*, 9.-12.10.2011, New Orleans, USA.

Harlow, S. & Johnson, T.J. (2011). Overthrowing the protest paradigm? How the New York Times, Global Voices and Twitter covered the Egyptian revolution. *International Journal of Communication*, vol. 5, pp. 1359-1374.

Honeycutt, C. & Herring, S.C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In *Proceedings of the 42nd Hawaii International Conference on System Sciences*. Retrieved March 29, 2011 from http://ella.slis.indiana.edu/~herring/honeycutt.herring.2009.pdf.

Hong, S. & Nadler, D. (2012). Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government Information Quarterly*, vol. 29, no. 4, pp. 455-461.

Huberman, B.A., Romero, D.M. & Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday*, vol. 14, no. 1-5, January 2009. Retrieved June 2, 2011 from

http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063.
Jansen, B.J., Zhang, M., Sobel, K. & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2169-2188.
Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*, vol. 33, pp. 159-74.
Letierce, J., Passant, A., Breslin, J. & Decker, S. (2010) Understanding how Twitter is used to spread scientific messages. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, April 26-27, 2010, Raleigh, NC: US. Retrieve January 11, 2013 from http://journal.webscience.org/314/.
Merton, R.K. (1968). The Matthew effect in science. *Science*, vol. 159, no. 3810, pp. 56-63.
Moed, H.F., De Bruin, R.E. & Van Leeuwen, T.N. (1995). New bibliometric tools for the assessment of national research performance – database description, overview of indicators and first applications. *Scientometrics*, vol. 33, no. 3, pp. 381-422.
Priem, J., & Costello, K. (2010). How and why scholars cite on Twitter. In *Proceedings of the 73rd ASIS&T Annual Meeting*. Pittsburgh, PA, USA.
Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, 15(7-5).
Ross, C., Terras, M., Warwick, C. & Welsh, A. (2010). Enabled backchannel: conference Twitter use by digital humanists. *Journal of Documentation*, vol. 67, no. 2, pp. 214-237.
Rowlands, I., Nicholas, D., Russell, B., Canty, N. & Watkinson, A. (2011). Social media use in the research workflow. *Learned Publishing*, vol. 24, no. 3, pp. 183-195.
Scanfeld, D., Scanfeld, M. & Larson, E.L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, vol. 38, no. 3, pp. 182-188.
Suh, B., Hong, L., Pirolli, P. & Chi, E.H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proceedings of IEEE International Conference on Social Computing*, 2010. Retrieved March 1, 2011 from http://web.mac.com/peter.pirolli/Professional/About_Me_files/2010-04-15-retweetability-v18-final.pdf.
Weller, K., Dröge, E., & Puschmann, C. (2011). Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences. In M. Rowe, M. Stankovic, A.-S. Dadzie, & M. Hardey (Eds.), *Making Sense of Microposts* (#MSM2011), Workshop at Extended Semantic Web Conference (ESWC 2011), Crete, Greece (pp. 1–12). CEUR Workshop Proceedings Vol. 718. Retrieved January 17, 2013 from http://ceur-ws.org/Vol-718/paper_04.pdf

Weller, K., & Puschmann, C. (2011). Twitter for Scientific Communication: How Can Citations/References be Identified and Measured? In *Proceedings of the Poster Session at the Web Science Conference 2011* (WebSci11), Koblenz, Germany. Retrieved January 17, 2013 from http://journal.webscience.org/500/1/153_paper.pdf.
Weller, M. (2011). *The digital scholar. How technology is transforming scholarly practice*. Bloomsbury Academic, UK.

THE DISCOVERY OF 'THE UBIQUITIN-MEDIATED PROTEOLYTIC SYSTEM': AN EXAMPLE OF REVOLUTIONARY SCIENCE? (RIP)

Jos J. Winnink[1,a] and Robert J.W. Tijssen[2]

[1] *winninkjj@cwts.leidenuniv.nl,* [2] *tijssen@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies (CWTS),
Wassenaarseweg 62A, 2333 AL Leiden (the Netherlands)

Leiden University Dual PhD Centre The Hague, P.O. Box 13228,
2501 EE The Hague (the Netherlands)

[a]NL Patent Office, P.O. Box 10366, 2501 HJ Den Haag (the Netherlands)

Abstract
In analysing bibliographical information from scholarly publications and from patent publications we try to develop a methodology for pinpointing the stage in which fundamental discoveries occur that later evolve into new technological developments. Our longitudinal bibliometric analyses of these 'breakthrough processes' also aim at obtaining insights into general empirical patterns that may characterise 'revolutionary' R&D dynamics.
In this case study we focus on R&D in the science field that has become known as the 'Ubiquitin System', one of the processes crucial in the functioning of living cells. The discovery of the ubiquitin system is classified by Aaron Ciechanover, one of the Nobel Prize laureates and one of the founding researchers in this area, as a 'challenge' discovery because the ubiquitin system was discovered mostly as a response to scientific challenges in the field. Studying trends in the ubiquitin research domain we attempt to develop an early indicator of breakthroughs, which will enable us to differentiate between 'charge' breakthroughs, which solve problems that are quite obvious, and 'challenge' breakthroughs that are characterised by are a response to an accumulation of facts that are unexplained by, or incongruous, with scientific theories at the time.

Conference Topic
Research Fronts and Emerging Issues (Topic 4) and Technology and Innovation Including Patent Analysis (Topic 5)

Introduction
This case study is one of a series to find bibliometric indicators that can identify a scientific 'breakthrough' at an early stage. Different approaches and tools have been tried in the past to tackle this detection problem, with varying degrees of success (see e.g. Arbesman, 2010; Breiner, Cuhls & Grupp, 1994; Chen et al.,

2009; Julius et al., 1977; Leydesdorff & Rafols, 2011; Martin, 1995; Small, 1977).
Our studies, including the one described in this publication, use examples of well-known and important discoveries that are analysed retrospectively to identify distinctive empirical patterns in the scientific literature and patents. To judge and contextualise the importance of a discovery we rely on the judgement of an scientific expert in the specific field and on argumentation provided by other knowledgeable reviewers, in this case the Nobel prize committee.

Overall research goal

We are searching systematically for structural changes in bibliographical data that signal, at an early as possible stage, the emergence of a scientific development that at a later stage contributed to the development of a new technology. The precise moment a 'breakthrough' occurs often cannot, even in retrospect, be precisely pinpointed on a time scale. Assembling information and combining bibliographical time-series data from scholarly publications and from patent publications enables us to focus on distinctive features or pivotal publications ('signals') that may signify sudden changes in the general pattern of knowledge creation processes.

'Breakthrough' concept

There is no generally accepted description, let alone a universal definition, of the term 'breakthrough' that can count on full support throughout the whole scientific community. This lack of consensus applies to 'breakthrough' scientific discoveries, 'breakthrough inventions', 'radical' technological innovations and other related concepts. For now we rely on the following sources such as Hollingsworth (2008), who defines: *'A major breakthrough or discovery is a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem.'* While (Ahuja & Lampert, 2001) defines 'breakthrough inventions' as *'those foundational inventions that serve as the basis for many subsequent technological developments'*, (Baba & Walsh, 2012) refers to breakthrough inventions in terms of *'... when developing a drug that is the first use of a compound and the first treatment for the disease ... We use the term "breakthrough" innovation for this kind of innovation'.* (Dunlap-Hinkler, Kotabe & Mudambi, 2010) uses a slightly different approach to identify breakthrough innovations: '*Typically, breakthrough innovations start the cycle of technological change'*.

Thomas Kuhn (Kuhn, 1962) distinguishes 'normal' science and 'revolutionary' science. Paradigm shifts are characteristic for revolutionary science. Paradigm shifts[58] change the cognitive structure of science (Andersen, Barker & Chen, 2006) and are visible as new and different concepts become into use. Koshland

[58] With paradigm shift we mean a fundamental change in approach or underlying assumptions

(2007) distinguishes three categories of scientific discoveries for classification. He argues: *'In looking back on centuries of scientific discoveries, however, a pattern emerges which suggests that they fall into three categories— Charge, Challenge, and Chance—that combine into a "Cha-Cha-Cha" Theory of Scientific Discovery.'* Following Kuhn's argumentation 'charge' discoveries can be considered 'normal' science; the 'challenge' discoveries and 'change' discoveries are examples of 'revolutionary' science.

For early identification of scientific discoveries, especially those that evolve into new (patented) technologies, we use Hollingsworth's generic definition as a general analytical framework. Within this setting we apply Koshland's theory and classification scheme focusing on the characterisation of individual discoveries, rather than on cumulative processes of knowledge creation and sequential order of related discoveries. In this paper we focus on 'challenge' discoveries. According to Koshland *'Challenge discoveries are a response to an accumulation of facts or concepts that are unexplained by or incongruous with scientific theories of the time.'*

The Ubiquitin-mediated proteolytic system

The Nobel Prize in Chemistry for 2004 was awarded to Aaron Ciechanover, Avram Hershko and Irwin Rose together for the discovery of 'ubiquitin mediated proteolysis'. The 'ubiquitin process is crucial in the functioning of Eukaryotic cells. Cells in which the genetic material is DNA in the form of chromosomes contained within a distinct nucleus. Eukaryotes include all living organisms other than the eubacteria and archaebacteria.

According to Nobel Prize Committee (2004) *'The breakthrough came in 1980. It was described in two papers that were both communicated on 10 December 1979 to the journal Proceedings of the National Academy of Sciences of the USA.'* The committee continues with '*The unraveling of the ubiquitin proteolytic system is not an exception to the rule that scientific discoveries are based on findings of others and that it can take a long period between the first preliminary findings and the breakthrough discovery*.'

To classify the discovery of the ubiquitin system Ciechanover (2008), using the Koshland classification, concludes that this discovery was a 'challenge' discovery. His argumentation: '*Could the ubiquitin system have been discovered earlier? Possibly yes. This could have happened by chance ... As we now know, the system was not discovered by chance but rather by challenge — mostly as a natural response to developments in the field ... A new system and concept(s) were needed to explain all of these new findings and assumptions, gathering them under a unifying umbrella.'*

Research question and hypothesis

Revolutionary science results in new concepts that can become the basis for new technologies. In our preceding studies on 'charge' breakthroughs, we were able to identify a number of indicators in order to identify breakthroughs in normal science (Winnink & Tijssen, 2011; Winnink, 2012). The ubiquitin system being a 'challenge' breakthrough is an example of revolutionary science. This breakthrough is different in that is not a 'linear' extension of the existing R&D, but also a paradigm shift is needed to correctly interpret the unexpected results. We try to identify this paradigm shift by looking at several structural features of the research field: the degree of multidisciplinarity before and after the discovery, evolution in the terminology used, discontinuities and developments in the networks of researchers and research institutes.

The main hypothesis we want to test using the discovery of the ubiquitin system is *'the occurrence of an identifiable paradigm shift, which is empirically linked to a particular distinct scientific discovery, can be used to detect at an early stage if that discovery is a challenge breakthrough and therefore could lead to a new technology'*.

Preliminary results

Trends in publications

Using the topic Ubiquitin to search for publications in the Thomson-Reuters *Web of Science* database we found 28,778 relevant research publications of the document types 'articles', 'letters' and 'proceedings papers'. Furthermore we searched for patent publications related to ubiquitin in the October 2012 edition of EPO's *Worldwide Patents Statistics* database, also known as PATSTAT, and found 1,522 documents belonging to 682 different patent families. The trends for both data sets are shown in figure 1. In figure 2 we zoom in on the early years of the time-span searching for significant and visible changes after 1979 when the breakthrough papers were published.

Multidisciplinarity of the research field

As a measure of multidisciplinarity we counted the number of different Thomson Reuters' (TR) Journal Categories that correspond to the WoS-indexed publication output per publication year. A Journal Category represents a TR-defined subfield of science. As Ubiquitin-related research becomes more dispersed across related science fields, the number of publications in journals belonging to other fields will increase. This process reflects the pervasive spread of a new 'paradigm' across the sciences, and wider use of the knowledge on the ubiquitin system. The results are shown in figure 3, which shows a fairly linear increase of fields, totalling more than 100 in recent years.

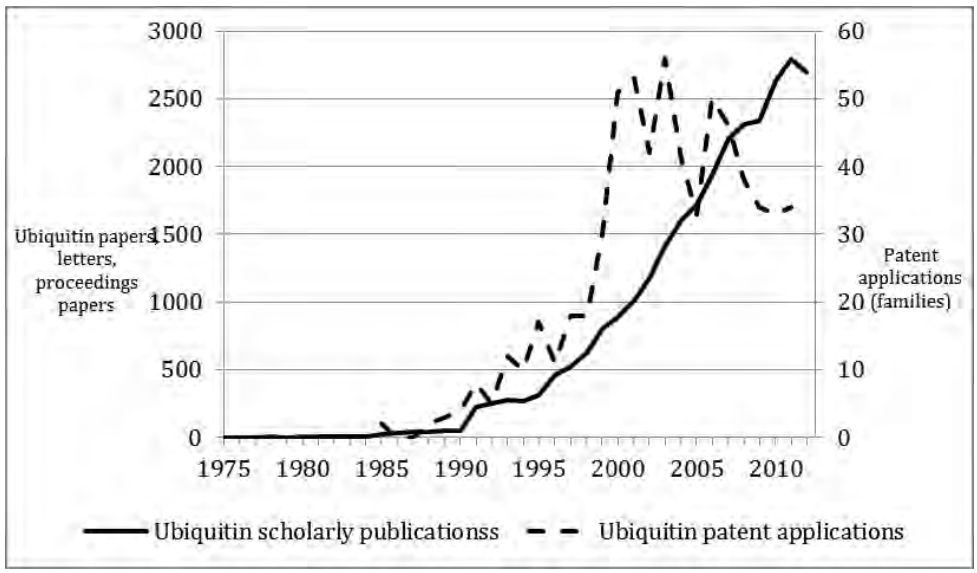


Figure 1 Trend of scholarly publications and patent publications related to Ubiquitin (1975-2012)

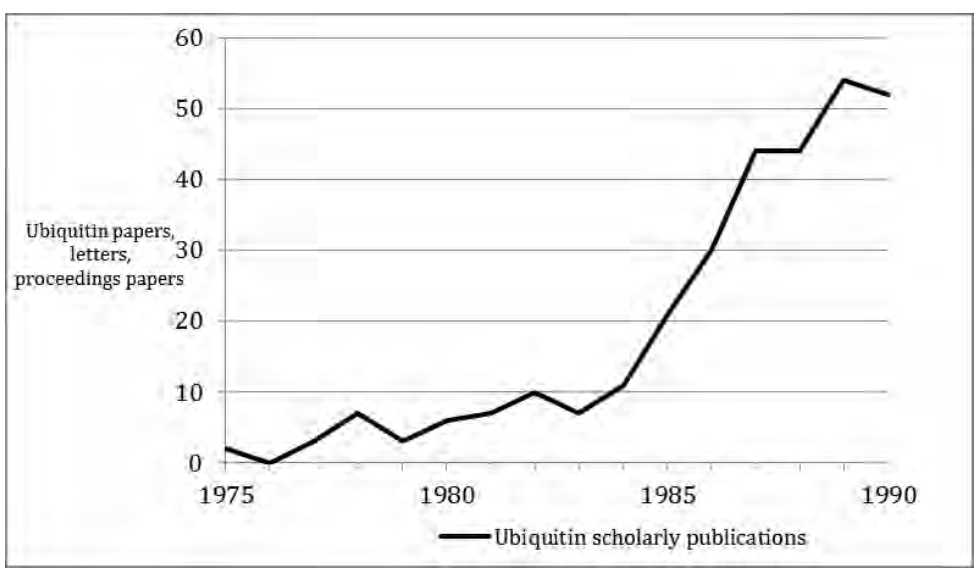


Figure 2 Trend of Ubiquitin scholarly publications (1975 - 1990)

Discussion

The aggregate-level trend data depicted in Figure 1 show several 'development stages' in the number of scholarly publications, notably in 1990-1991 and 2008-2009. These spikes in the time-series suggest prior scientific discoveries that have caused sudden increase in research activity and concomitant rise in the number of

publications. The type of the discovery, and its impact on R&D processes worldwide, however cannot be determined from these trend data.

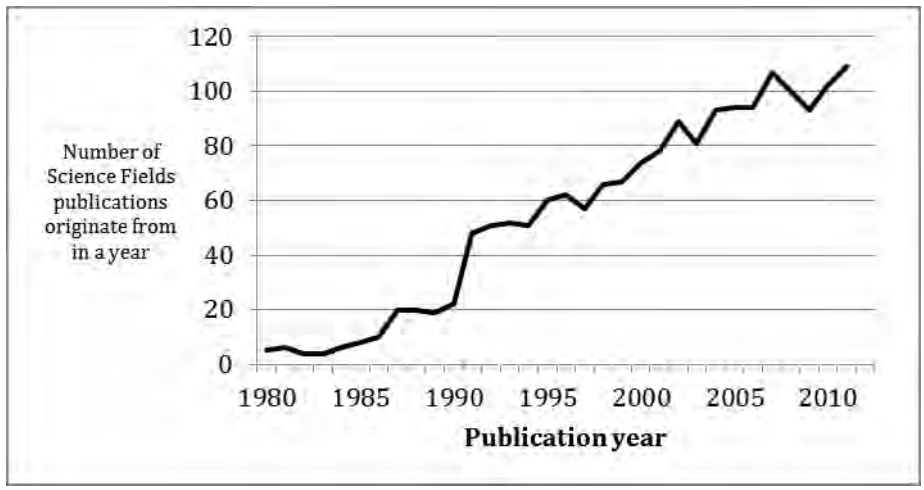


Figure 3 Number of different science fields (1980-2010)

Other information sources and related bibliographical data have to be examined to determine the nature and background of the breakthrough, in order to differentiate between 'charge' and 'challenge' discoveries. The evolution of the number of patent applications, and the time-delay with the upswing of scholarly publications is an indication that from 1985 onwards the scientific knowledge became sufficiently well developed to be transformed into patentable technological applications. In figure 2, which focuses on the period the breakthrough discovery was made; we indeed find a steep rise during the period 1983-1987. This increase might have been be caused by the breakthrough discovery, an assertion which will be examined in the on-going stage of the study. The evolution of networks of researchers and institutions becoming active in this research field should reflect the diffusion of the new paradigm within the scientific community.

Further results of this study, which will focus on the research question and main hypothesis, will be presented at the conference.

References

Ahuja, G. and Morris Lampert, C. (2001). Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, 22(6-7): 521–543.

Andersen, H., Barker, P., & Chen, X. (2006). *The cognitive structure of scientific revolutions*. Cambridge University Press

Arbesman, S. (2010). Quantifying the ease of scientific discovery. *Scientometrics*, 86(2):245–250.

Baba, Y. and Walsh, J. P. (2010). Embeddedness, social epistemology and break-through innovation: The case of the development of statins. *Research Policy*, 39(4): 511–522. Special Section on Innovation and Sustainability Transitions.

Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chávez, C., and Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, 75(3): 495–518.

Breiner, S., Cuhls, K., and Grupp, H. (1994). Technology foresight using a delphi approach - a Japanese-German cooperation. *R & D management*, 24(2): 141–153.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., and Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3): 191–209.

Ciechanover, A. (2009). Tracing the history of the ubiquitin proteolytic system: The pioneering article. *Biochemical and Biophysical Research Communications*, 387(1): 1–10.

Dunlap-Hinkler, D., Kotabe, M., and Mudambi, R. (2010). A story of break-through versus incremental innovation: corporate entrepreneurship in the global pharmaceutical industry. *Strategic Entrepreneurship Journal*, 4(2): 106–127.

Grupp, H. & Schmoch, U. (1992). *Dynamics of Science-Based Innovation*, chapter 9 - At the crossroads in laser medicine and polyimide chemistry: patent assessment of the expansion of knowledge, pages 269–301. Springer-Verlag.

Hollingsworth, J. R. (2008). Scientific discoveries: An institutionalist and path-dependent perspective. In Hannaway, C., editor, *Biomedicine in the Twentieth Century: Practices, Policies, and Politics*, pages 317–353. National Institute of Health.

Isenson, R. S. (ed.) (1969). *Project hindsight (final report)*. Report AD495905, US Dept. of Defense.

Jewkes, J., Sawers, D., and Stillerman, R. (1969). *The Source of Invention*, 2nd edition. MacMillan.

Julius, M., Berkoff, E., C., Strack, A. E., Krasovec, F., and Bender, A. D. (1977). A very early warning system for the rapid identification and transfer of new technology. *American Society for Information Science,* 28(3): 170–174.

Koshland, D. E. (2007). The cha-cha-cha theory of scientific discovery. *Science*, 317(5839): 761–762.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. The Univesity of Chicago Press

Leydesdorff, L. and Rafols, I. (2011). Local emergence and global diffusion of research technologies: An exploration of patterns of network formation. Journal of the American Society for Information Science and Technology, 62(5): 846–860.

Martin, B. R. (1995). Foresight in science and technology. *Technology Analysis & Strategic Management*, 7(2): 139–168.

Meyer-Krahmer, F. and Schmoch, U. (1998). Science-based technologies: university-industry interactions in four fields. *Research Policy*, 27(8): 835–851.

Nobelprize Comitte (2004). The Nobel Prize in Chemistry 2004 - Advanced Information. Nobelprize.org. 16 Jan 2013 http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2004/advanced.html

Perla, R. J., & Carifio, J. (2005). The Nature of Scientific Revolutions from the Vantage Point of Chaos Theory. *Science & Education*, *14*, 263–290

Pavitt, K. (1984). Sectoral patterns of technical change: Towards a taxonomy and a theory. *Research Policy*, 13(6): 343–373.

Sangwal. On the growth of citations of publication output of individual authors. *Journal of Informetrics*, 5(4): 554–564, 2011.

Small, H. (1977). Co-citation model of a scientific specialty - longitudinal-study of collagen research. *Social Studies of Science*, 7(2): 139–166.

Winnink, J. J. and Tijssen, R. J. W. (2011). R&D dynamics in the development of HIV/AIDS drugs. In Noyons, E., Ngulube, P., and Leta, J., editors, *Proceedings of the 13th International Conference of the International Society for Scientometrics & Informatics (ISSI 2011),* pages 855–860.

Winnink, J. J. (2012). Searching for structural shifts in science: Graphene R&D before and after Novoselov et al. (2004). In Archambault, E., Gingras, Y., and Larivière, V., editors, *Proceedings of the 17th International Conference on Science and Technology Indicators (STI 2012),* volume 2, pages 837–846.

THE DISTRIBUTION OF REFERENCES IN SCIENTIFIC PAPERS: AN ANALYSIS OF THE IMRAD STRUCTURE

Marc Bertin[1], Iana Atanassova[1], Vincent Lariviere[2] and Yves Gingras[3]

[1] *marc.bertin@paris-sorbonne.fr; iana.atanassova@paris-sorbonne.fr*
Sens, Texte, Informatique, Histoire (STIH), Paris-Sorbonne University, 1 rue Victor Cousin, 75230 Paris cedex (France) and Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Quebec à Montreal, CP 8888, Succ. Centre-Ville, Montreal, QC. H3C 3P8 (Canada)

[2] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, QC. H3C 3J7 (Canada) and Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Quebec à Montreal, CP 8888, Succ. Centre-Ville, Montreal, QC. H3C 3P8 (Canada)

[3] *gingras.yves@uqam.ca*
Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Quebec à Montreal, CP 8888, Succ. Centre-Ville, Montreal, QC. H3C 3P8 (Canada)

Abstract
The organization of scientific articles typically follows a standardized pattern, the well-known IMRaD structure (Introduction, Methods, Results and Discussion). Using the PLOS series of journals as a case study, this paper looks at how the bibliographic references are distributed along the different sections of papers. We use the section titles of the articles to categorize the sections matching the IMRaD structure. We then identify the variations in the basic IMRaD structure of the different PLOS journals. The results show that, though dominant, the IMRaD structure often changes in some journals and these differences must be taken into account in order to compare the distribution of references along the text using an invariant measure, here the number of sentences in the texts. We examine the different distributions of the references in the articles in different journals and show that these distributions are relatively stable and maybe even invariant when taking into account the inversions of sections identified in some journals.

Introduction
The organization of scientific articles typically follows a standardized pattern, the well-known IMRaD structure (Introduction, Methods, Results and Discussion). This structure has imposed itself in most major scientific journals in the mid-twentieth century, and has become the main standard in the 1970s (Sollaci and Pereira, 2004). Many studies have focused on various aspects of this structure:

automatic classification of sentences in full-text (Agarwal and Yu, 2009), the effects of the use of the IMRaD style (Oriokot et al., 2011), creation of guidelines for scientific writing (Kucer, 1985; Meadows, 1985; Day and Gastel, 2006), providing structured abstracts (Nakayama et al., 2005) and editorial requirements (Barron, 2006).

Research question

This article investigates, from the viewpoint of bibliometrics, the relationships that exist between cited references and the structure of the text. What interests us is the nature of the distribution of references in scientific articles and more precisely, if there exists a typology of scientific writing and referencing practices. These characteristics of scientific papers are studied here using the seven (7) journals published by the Public Library of Science (PLOS), which are peer-reviewed open-access publications covering all disciplines of sciences and social sciences. The free access to full text gives us the opportunity to use the PLOS journals as a test corpus to establish the relation between the distribution of references throughout an article and its structure. Our analysis consists in several steps: categorisation of the sections of the text according to section titles, segmentation into sentences in order to obtain the distribution of the references according to the text progression, reconstruction of the IMRaD structure and the examination of the distribution of the references in the different journals.

Our results provide an overview of the types of articles in the PLOS journals and show some properties of the structure of research articles related to the sections and section titles. We explore the relations between the types of articles and the IMRaD structure, and also the relations between the types of sections and the references in the texts. Finally, we obtain a graphical representation of the distribution of references in an article. The next section presents the corpus of data and its structure. Then, we describe the processing carried out in order to relate the IMRaD structure to the distribution of references in the articles.

Methods

We first categorize the sections which allows us to work with the different types of sections and reorder the sections in a text if necessary. This categorization aims to verify the coherence of the corpus with the IMRaD structure. We then process the text content of all paragraphs in order to segment them into sentences. This segmentation allows us to work with text elements that are smaller than paragraphs so that we can associate the references with a given sentence of the text and obtain their distribution along the text. Finally, our algorithm counts the number of references in each sentence. This task is not trivial, as we will discuss later.

Data source

Founded in 2006, the Public Library of Science (PLOS) is an Open Access publisher of seven peer-reviewed academic journals, mostly in the fields of biology and medicine. PLOS ONE, the publishers' general journal covers, however, all fields of science and social sciences. For this study, we have used the entire PLOS corpus up to September/October 2012. Table 1 presents the number of articles processed for each journal, as well as the average number of sections and sentences per article. More than 47,000 journal articles were analyzed. As these 7 journals follow the same publication model but are in different scientific fields, our aim is to observe the different uses of bibliographic references in these fields and their relation to the structure of the articles. Table 1 show that the average number of sections per article varies between 3.48 and 4.74 according to the journal. We can also observe that the average length of articles is different: 125 sentences on average for PLOS Medicine, compared to 278 sentences on average for PLOS Computational biology. The Table also shows the relative importance of PLOS ONE: papers published in this journal account for more than 71% of all papers in the corpus.

Table 1. Descriptive Statistics on PLOS Journals

Journal	Number of articles	Ave. number of sections	Ave. number of paragraphs	Ave. number of sentences
PLOS ONE	33 782	4,47	40,55	190,67
PLOS Biology	2 965	3,48	37,31	156,23
PLOS Medicine	2 228	4,10	38,51	125,26
PLOS Genetics	2 560	4,73	49,64	230,96
PLOS Computational Biology	2 107	4,69	78,91	278,06
PLOS Pathogens	2 354	4,74	44,88	228,83
PLOS Neglected Tropical Diseases	1 366	4,50	36,74	171,37
Total	47 362	4,43	42,56	192,84

Data structure

PLOS provides access to the articles in the XML format. The set of XML elements and attributes that are used for the representation of journal articles are known as Journal Article Tag Suite (JATS), which is an application of Z39.96-2012 (ANSI, 2012). Some studies (Carter, Funk and Mooney, 2012) give various applications of this standard. Technology evolves quickly and we have to take into consideration that JATS is a continuation of the NLM Archiving and Interchange DTD work by NCBI (http://dtd.nlm.nih.gov/).

The JATS structure of an article consists of three main elements *front – body – back*, where the textual content of the article is in the *body* element, which is further divided into sections and paragraphs. The <*front*> tag contains some traditional fields of metadata (title, authors, etc.) as well as the article type.

Labels and section titles processing

The sections of the texts are categorized automatically by analyzing the section titles in order to match the existing sections with one of the section types in the IMRaD structure. To do this, we have examined the types of articles present in the corpus, where the typology is given in the article's metadata.

Segmentation processing

The first stage of the processing consists in parsing the XML trees and text segmentation into sentences. The JATS structure used by PLOS provides paragraph elements <p> as the finest level of text segments. For our analysis, we needed segmentation into sentences and we parsed the initial JATS trees in order to extract the relevant text segments from the article body, as well as other elements such as sections, section titles, section numbers, paragraphs and the bibliography. These data were stored in the DocBook format that was used as the basis for the further processing.

Each paragraph was segmented into sentences by analysing the punctuation of the text following a set of typographic rules. All the occurrences of symbols denoting sentence boundaries (point, exclamation mark, etc.) were examined and disambiguated. Figure 1 gives some examples which show a few points present in the sentences but which do not finish them. In fact, the occurrence of a point in a text does not necessarily mean a sentence end, because in many cases it can be part of an abbreviation, references, genus species, numeric values, etc.

1. *, SE = 0.44, 0.041); and gene diversity from 0.39 (EMX-4) to 0.69 (LafMS03*
2. *the plastid genome is 0.92±0.03*
3. *an additional 115.0 ml*
4. *(Nyakaana and Arctander 1998; Fernando et al. 2001) and compared them*
5. *HB3 strain of P. falciparum, we demonstrate that at least 60%*
6. *(i.e., the kinase phosphatase*

Figure 1. Examples of occurrences of 'point' that do not signal sentence ends.

We used a set of finite-state automata in order to determine the contexts in which the points signal sentence ends. For this purpose, we have developed a Java application based on the work of Mourad (2001). The algorithm uses a rule-based approach which disambiguates the use of punctuation marks by examining the close context of their occurrences. All punctuation marks in the text are thus labeled as "sentence end" or "no sentence end". Some of the results are presented in table 2.These results synthesize a more general problem in NLP. Once we have

identified the sentence boundaries in the corpus, we can consider the sentences as the finest textual unit and examine the number of references in each sentence. In fact, a sentence can contain one or more references or an enumeration of references, which is rather frequent in the background section or the introduction.

Table 2. Segmentation into sentences according to typographic rules

	Sentence end				No sentence ends
Corpus	point	exclamation mark	question mark	total	point
PLOS Biology	425,255	127	3,137	428,519	431,656
PLOS Computational Biology	508,855	107	2,234	511,196	513,430
PLOS Genetics	559,157	452	2,138	561,747	563,885
PLOS Medicine	246,042	45	2,058	248,145	250,203
PLOS Neglected Tropical Diseases	6,184,077	469	9,353	6,193,899	6,203,252
PLOS ONE	516,277	19	806	517,102	517,908
PLOS Pathogens	516,277	19	806	517,102	517,908
Total	8,676,167	1,237	20,042	8,697,446	8,717,488

Reference processing

Our algorithm examines each sentence and counts the number of references present in the text. In fact, the input data is in the XML format where the references are represented in the <*xref*> tags. However, counting these tags is not a reliable method to obtain the reference counts and could bias the system. As shown in the example on Figure 2, some typographic rules for writing references result in the fact that the XML structure does not render all of the actual references. In this example, three sources are cited (51, 52 and 53), but only two <*xref*> tags are present that delimit a range from 51 to 53. As these cases are rather frequent in the corpus (on average more than once in an article), they must be taken into consideration. Our algorithm covers all possible typographic variations for reference ranges and infers the missing data from the input XML. As a result we obtain the list of sentences in the text, where to each sentence we have associated a reference count as well as a list of reference identifiers corresponding to the bibliography entries.

*"... during differentiation [<xref ref-type="bibr" rid="pbio-0030356-b51">**51**</xref>–<xref ref-type="bibr" rid="pbio-0030356-b53">**53**</xref>]. This prediction ..."*

Figure 2. Example of a reference range rendered in XML

Results

Article Level

Table 3 presents the different article types in the PLOS corpus, exploiting the metadata present in the XML documents. The article types are identified using the contents of the <*article-meta*> tag in the JATS structure. This Table shows, as should be expected, that the 'Research Article' is dominant with 94% of the

papers published. We notice however that *PLOS Medicine* offers a wider variety of article like 'Perspective', 'Correspondence', 'Essay' or 'Policy Forum'.

Table 3. PLOS article typology study

Article type	Journals							
	PLOS ONE	PLOS Biology	PLOS Medicine	PLOS Genetics	PLOS Computational Biology	PLOS Pathgens	PLOS Neglected Tropical Diseases	Percentage
Research Article	33,708	1,552	782	2,373	1,876	2,143	1,154	94.69%
Perspective		35	260	64	44			0.88%
Correspondence		20	252	4	13			0.63%
Review	38		16	56	32	59	48	0.54%
Essay		72	138					0.46%
Policy Forum			206					0.45%
Editorial		21	70	7	24	2	47	0.37%
Primer		156						0.34%
Opinion						80		0.17%
Health in Action			76					0.17%
Pearls						50		0.11%
Research in Translation			49					0.11%
Community Page		48						0.10%
Viewpoints				5			41	0.10%
Other Categories								0.88%
Total								100%

Section Level

We now concentrate our analysis on research articles, which account for the vast majority of documents published by PLOS journals. The number of sections in the texts is particularly important for our study and we first match the section titles with the section position in the four sections of the IMRaD structure. Table 4 presents the results of the categorization of the sections for the seven PLOS journals. We have analyzed all section titles that are present as a separate element in the XML documents. We determine whether the section is part of the IMRaD structure or not by identifying occurrences of "Introduction", "Method", "Result" and "Discussion" with all possible variations, plurals, combinations, etc. Thus, we have created a set of criteria for the categorization that cover the majority of the observed section titles. After normalization, we have considered the subset of titles present in all journals, except for "Supporting information" which was not considered because this type of sections is not part of the scientific argumentation

and serves as complementary information. Finally, to produce tables 5 to 10, we look at the position of titles for each section. We check that Introduction correspond to the section one, Method correspond to the section two, Result correspond to the section three and Discussion to the section four.

Table 4 shows that *PLOS Medicine* and *PLOS Neglected Tropical Diseases* essentially follow the IMRaD structure. The values on the diagonal of the matrix for *PLOS Neglected Tropical Diseases* are well above 85%, which means that virtually all the articles follow the IMRaD standard. In the case of *PLOS Medicine*, the values on the diagonal show that about half of the papers follow the IMRaD structure, while the other half use section titles that did not allow tha automatic categorization of the sections. For both journals, the high values on the diagonals indicate clearly that in almost all of the papers that include sections categorized as Introduction, Method, Result and Discussion, these sections appear in the order defined by the IMRaD structure. Hence, the first column, which corresponds to section one, never includes Method, Results and Discussion. This is coherent with the structure generally presented in the literature.

Table 4. Relation between position of section and title of section for PLOS Medicine and PLOS Neglected Tropical Diseases

PLOS journal	PLOS Medicine				PLOS Neglected Tropical Diseases			
Name of section	Section 1	Section 2	Section 3	Section 4	Section 1	Section 2	Section 3	Section 4
Introduction	**40.10%**	1.30%	0.12%	0.13%	**89.08%**	0.38%	0.00%	0.00%
Method	0.00%	**48.47%**	1.72%	0.91%	0.00%	**89.79%**	0.23%	0.32%
Result	0.00%	0.30%	**48.44%**	0.45%	0.00%	0.23%	**88.73%**	0.32%
Discussion	0.00%	2.13%	1.96%	**52.13%**	0.00%	0.38%	1.00%	**86.06%**
Other	59.90%	47.82%	47.76%	46.38%	10.92%	9.21%	10.04%	13.31%
Total	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 5. Relation between position of section and title of section for PLOS ONE and PLOS Computational Biology

PLOS journal	PLOS ONE				PLOS Computational Biology			
Name of section	Section 1	Section 2	Section 3	Section 4	Section 1	Section 2	Section 3	Section 4
Introduction	**99.90%**	0.07%	0.00%	0.00%	**92.69%**	0.88%	0.00%	0.05%
Method	0.00%	47.82%	5.95%	**46.86%**	0.00%	18.98%	10.63%	**60.13%**
Result	0.00%	**51.48%**	47.90%	0.12%	0.00%	**69.67%**	22.60%	0.15%
Discussion	0.00%	0.08%	**45.79%**	46.78%	0.00%	0.15%	**59.30%**	21.73%
Other	0.09%	0.55%	0.36%	6.24%	7.31%	10.32%	7.47%	17.93%
Total	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 5 shows the relation between the position of the sections and section titles for *PLOS ONE* and *PLOS Computational Biology*. While the first value presented on the diagonal is more than 99%, other values on the diagonal are very low (close to 50%), which indicate that the usual order of sections in IMRaD are in fact changed. the Method section (on line two), can be found not only in section 2 as expected with IMRaD, but also in section 4 usually reserved for Discussion. The standardization proposed for extraction of titles takes into account such variations. This inversion explains that of the Results section often appears in Section 2 instead of 3, and that the methods are presented at the end of the article (Section 4). Of course these papers do not respect completely the IMRaD structure and should present some variations in the distributions of references.

Finally, Table 6 shows the equivalent results for *PLOS Genetics, PLOS Pathogens* and *PLOS Biology*. We note that the distribution of sections and titles for these journals also differs from IMRaD with Methods coming last instead of Second and Discussion third instead of fourth as in the standard IMRaD structure.

Table 6. Relation between position of section and title of section for PLOS Genetics, PLOS Pathogens and PLOS Biology

PLOS journal	PLOS Genetics				PLOS Pathogens			
Name of section	Section 1	Section 2	Section 3	Section 4	Section 1	Section 2	Section 3	Section 4
Introduction	**94.84%**	0.20%	0.00%	0.00%	**93.46%**	0.04%	0.00%	0.00%
Method	0.00%	4.18%	10.41%	**82.16%**	0.00%	7.22%	5.86%	**81.30%**
Result	0.00%	**91.85%**	4.48%	0.04%	0.00%	**86.04%**	7.26%	0.00%
Discussion	0.00%	0.04%	**81.69%**	3.94%	0.04%	0.00%	**80.70%**	7.15%
Other	5.16%	3.73%	3.42%	13.86%	6.50%	6.70%	6.17%	11.56%
Total	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

PLOS journal	PLOS Biology			
Name of section	Section 1	Section 2	Section 3	Section 4
Introduction	**52.89%**	0.55%	0.05%	0.00%
Method	0.00%	0.70%	10.59%	**72.98%**
Result	0.00%	**76.88%**	0.80%	0.06%
Discussion	0.00%	0.00%	**70.69%**	0.83%
Other	47.11%	21.87%	17.87%	26.14%
Total	100.00%	100.00%	100.00%	100.00%

Knowing the structure of the text in terms of section headings – and having reordered the various texts in order to have a consistent order of sections – we can now present the distribution of references along the texts of papers of the different journals. To do this, we have used a subset of the corpus which contains only those research articles that contain the four types of sections of the IMRaD

structure. All the articles in this smaller corpus have at least four sections that correspond to the types Introduction, Method, Result and Discussion but these sections are not necessarily present in the same order in the text. Table 7 shows the number of articles that fulfill these criteria. We can observe that this new corpus represents 82.98% of the corpus.

Distribution of References at the Sentence Level

Figure 3 presents the normalized distributions of the references throughout the texts for two PLOS journals. The horizontal axis presents the text progression from 0 to 100 percent based on the segmentation into sentences. The vertical axis gives the average percentage of the number of references at a given point of the text for each corpus. We can observe that the first 10 percent of the texts in these corpuses contain relatively large amounts of references. The three vertical lines on the graph indicate the average positions of the section boundaries.

Table 7. Research articles containing the four section types of the IMRaD structure

Journal	Number of research article having IMRaD sections	Percentage
PLOS Biology	1,336	45.06%
PLOS Computational Biology	1,548	73.47%
PLOS Genetics	2,096	81.88%
PLOS Medicine	770	34.56%
PLOS Neglected Tropical Diseases	1,079	78.99%
PLOS ONE	30,470	90.20%
PLOS Pathogens	2,003	85.09%
Total	39,302	82.98%

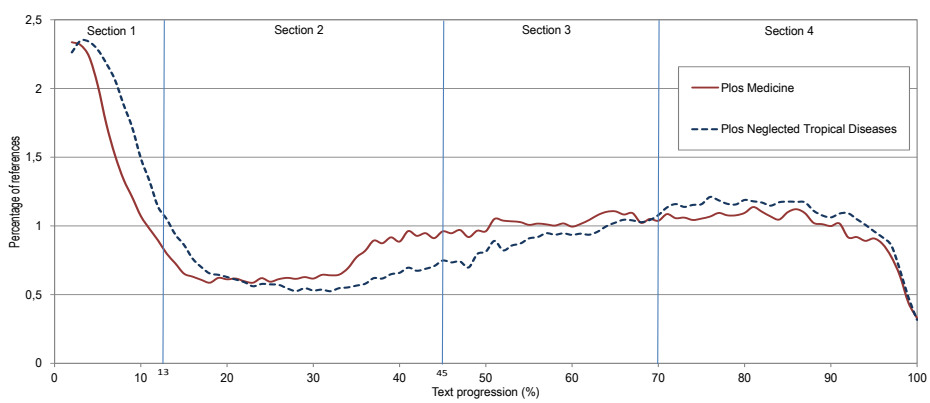


Figure 3. Distribution of References in PLOS Medicine and PLOS Neglected Tropical Diseases

These results are consistent with what might be expected: references are more concentrated in the introduction. The comparison of Tables 4, 5 and 6 with Figures 3, 4 and 5 shows that the distributions of references are similar in the sets of journals having the same structure of section titles. In fact, Figure 3 shows that section 2, which according to Table 4 corresponds to the Method in a majority of articles in these two journals, contains less references that the other sections. On the other hand, Table 6 shows that the Method section tends to be at the end of the articles for three of the journals. This is consistent with the distribution of references on Figure 5 where we can observe that the fourth section contains a smaller number of references that the first three sections. These observations suggest that if we take into account the variations in the positions of sections the distribution of references could be very stable and nearly invariant.

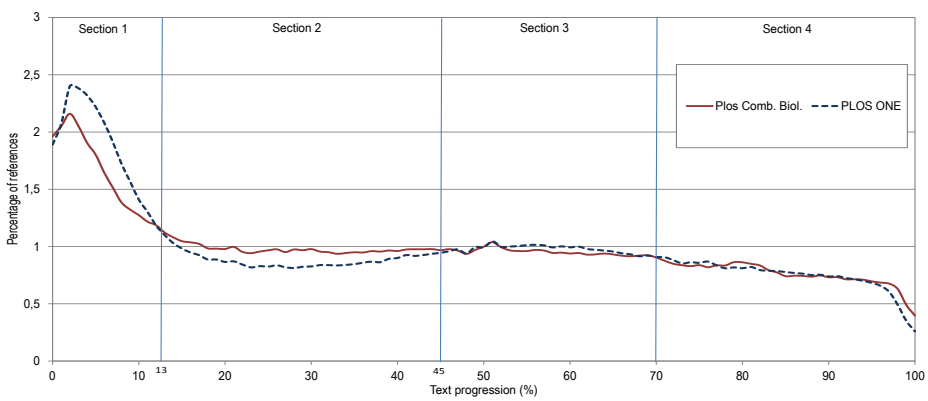


Figure 4. Distribution of References in PLOS Computational Biology and PLOS ONE

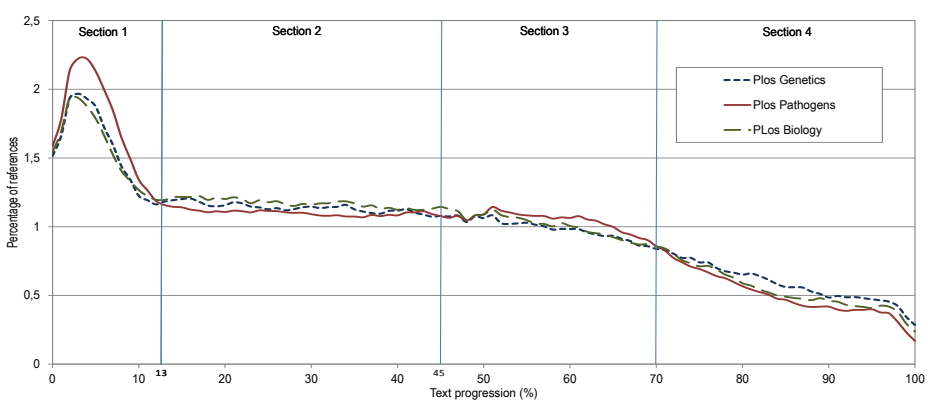


Figure 5. Distribution of Reference in PLOS Genetics, PLOS Pathogens and PLOS Biology

Distribution of References for the ordered IMRaD structure

In order to study the distribution of references independently of the order in which the sections of the IMRaD structure appear in the texts, we have reordered the sections in all articles with respect to the order Introduction, Method, Result, Discussion. The reordered articles were then used to produce the new distribution of references. Figure 6 shows the distributions of references that were obtained for the 7 PLOS journals. We can observe that the distributions for all seven journals share practically the same properties. The Introduction sections contain a relatively large number of references, with a bigger concentration in the first part of the Introduction. The Method section is characterized by a relatively smaller number of references which grows bigger towards the Results and Discussion sections. The "PLOS" curve on this graph corresponds to the distribution of references in the entire corpus.

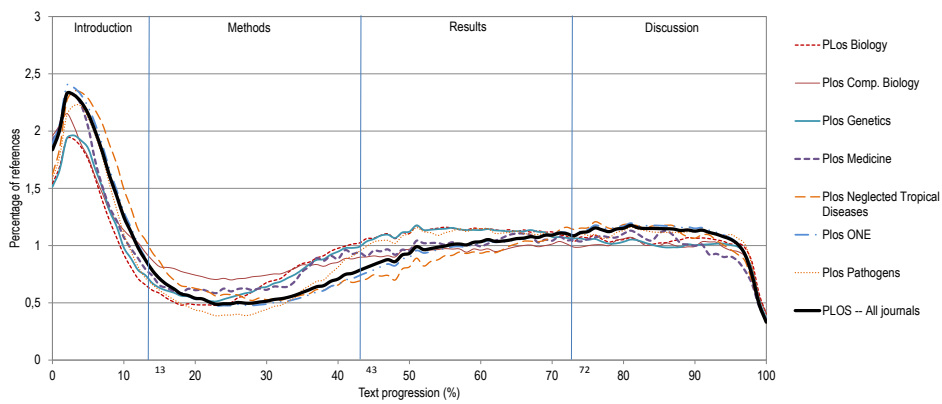


Figure 6. Distribution of References of PLOS journals, following the IMRaD structure

Conclusion

In this paper, we have shown that we can measure the distribution of references along the text of articles using sentences as the counting unit. We have also shown that this distribution seems quite stable and maybe even invariant if we take into account the changes that occur in some journals in the positions of the different sections in the text of the articles. Knowing the structure of the articles, we are now in a position to connect the references with their position in the text in order to better characterize the kinds of references in terms of the nature of the section in which they appear. For it is plausible that the kinds of references present in the introductory section may differ from the ones mentioned in the Method section, for example. While this could be done by hand using a small sample, the methods presented here are applicable to very large data sets.

The results of this study might be of interest for citation context analysis or in case one wants to assign different weights to citations according to their place in

the document (see Bonzi, 1982; Rousseau, 1987). Our future work will focus on citation context analysis, as well as examining the other correlations that might exist between the position in the text and the nature of the references: their publication year or the subject category of the reference journals.

References

Agarwal, S. and Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. Bioinformatics, 25(23): 3174–3180.

American National Standards Institute (2012). JATS: Journal Article Tag Suite. ANSI/NISO Z39.96-2012, 9 August 2012. National Information Standards Organization (NISO). Available at: http://www.niso.org/apps /group_public/download.php/8975/z39.96-2012.pdf

Barron, J. (2006). The Uniform Requirements for Manuscripts Submitted to Biomedical Journals Recommended by the International Committee of Medical Journal Editors. Chest, 129(4): 1098–1099.

Bonzi, S. (1982). Characteristics of a Literature as Predictors of Relatedness Between Cited and Citing Works. Journal of the American Society for Information Science and Technology (JASIST), 33(4): 208–216.

Day, R. A, Gastel, B. (2006) How to Write and Publish Scientific Papers. Cambridge: Cambridge University Press.

Carter, R., Funk, K., and Mooney, R. (2012). The Front Matters: Capturing Journal Front Matter with JATS. Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012. Bethesda (MD): National Center for Biotechnology Information (US); 2012. Available at: http://www.ncbi.nlm.nih.gov/books/NBK100353/

Kucer, S. (1985). The Making of Meaning Reading and Writing as Parallel Processes. Written Communication, 2(3): 317–336.

Meadows, A. (1985). The scientific paper as an archaeological artefact. Journal of information science, 11(1): 27–30.

Mourad G. (2001), Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des Applications informatiques : SegATex et CitaRE, Ph. D. Thesis, Univ. Paris-Sorbonne.

Nakayama, T., Hirai, N., Yamazaki, S., Naito, M. (2005). Adoption of structured abstracts by general medical journals and format for a structured abstract, Journal of the Medical Library Association 93(2), 237.

Oriokot, L., Buwembo, W., Munabi, I., and Kijjambu, S. (2011). The introduction, methods, results and discussion (IMRAD) structure: a Survey of its use in different authoring partnerships in a students' journal. BMC research notes, 4(1): 250.

Rousseau, R. (1987). The Gozinto theorem: Using citations to determine influences on a scientific publication. Scientometrics, 11(3-4): 217–229.

Sollaci, L. and Pereira, M. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. Journal of the Medical Library Association, 92(3): 364.

DO BLOG CITATIONS CORRELATE WITH A HIGHER NUMBER OF FUTURE CITATIONS? (RIP)

Hadas Shema[1], Judit Bar-Ilan[1] and Mike Thelwall[2]

[1] *Judit.Bar-Ilan@.biu.ac.il; dassysh@gmail.com*
Department of Information Science, Bar-Ilan University, Ramat-Gan, 52900 (Israel)

[2] *m.thelwall@wlv.ac.uk*
School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY (UK).

Abstract
Blog posts aggregated at ResearchBlogging.org discuss scientific results and provide full bibliographic references to the reviewed articles. Articles reviewed in these blogs therefore receive "blog citations". We hypothesized that articles receiving blog citations near their publication time become more highly cited later on than the articles in the same journal published in the same year that did not receive such blog citations. Our results for articles published in 2009 support this hypothesis for some journals but not for others.

Conference Topic
Old and New Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Webometrics (Topic 7).

Introduction
Science blogs publish posts related to science and review scientific developments, and have become popular with a section of the scholarly community. Respected scholarly media outlets such as National Geographic, the Nature Group, Scientific American and the PLoS journals all have science blogging networks. A Nature Medicine editorial, discussing blogs and peer review concluded that *"Online science blogs are a valuable forum for commenting on published research, but their present importance lies in complementing rather than replacing the current system of peer review"* (Perfecting peer review?, 2011, p. 1-2).
Citations to academic journal articles from blogs can potentially be used as an alternative source of impact evidence, i.e., an altmetric (Priem, Piwowar, & Hemminger, 2011). Kousha, Thelwall, and Rezaie (2010) have shown that it is possible, at least on a small scale, to calculate blog mentions for a set of published articles by using Google Blog Search. They concluded that, although blog citations were found to be far less common than academic citations, they could still be useful evidence of research impact on wider discussions, especially in the social sciences and humanities. While Kousha et al. considered every mention of scholarly article in blog as a citation, we would like to differentiate between *blog mentions* and *blog citations*. Blog mentions are any sort of reference to scholarly

material in blogs, while blog citations cite scholarly materials in structured, formal styles (e.g., APA, MLA) and appear in blog posts.

ResearchBlogging.org (2008) aggregates blog posts referring specifically to peer-reviewed research. It is a self-selecting aggregator that allows bloggers to cite peer-reviewed research in an academic citation format. Bloggers discussing peer-reviewed research can register with the aggregator and when they mark relevant posts in their blog, these posts appear on the aggregator site, giving one-stop access to a variety of research reviews from different authors. The site has human editors who ensure that blogs submitted to the aggregator follow its guidelines and are of appropriate quality. It also has an altmetric role, since it serves as one of the article level metrics (ALM) displayed for each article in the journal PLOS ONE. Although over 80% of RB blogs are written in English (Shema, Bar Ilan & Thelwall 2012; Fausto et al. 2012) the site also supports blogs in German, Spanish, Portuguese, Chinese, Polish and Italian.

The first ResearchBlogging.org (RB) study was conducted by Groth & Gurney (2010), and focused on 295 aggregated posts tagged "Chemistry." The literature cited in these posts was mostly up-to-date and came from top journals: 70.5% of the cited articles were from the top 20 chemistry journals, and 21% were from the 60 top publications across all disciplines.

Another study (Shema, Bar-Ilan & Thelwall, 2012) focused on established blogs and bloggers that had at least 20 posts aggregated in RB between January 1, 2010 and January 15, 2011. The chosen blogs were non-commercial and written by 1-2 authors. The sample included 126 blogs and 135 bloggers. The most popular blog category was Life Science (39%), followed by Psychology, Psychiatry, Neurosciences & Behavioral Science and Medicine (19%). Blogs about Social Sciences & Humanities and about Computer Science & Engineering were the least popular (5% and 1% respectively). The study found that a majority (59%) of science bloggers were part of the academic community in some capacity. Another survey, of bloggers using the German platform SciLogs, found that 43% were employed in the academy (Puschmann & Mahrt, 2012). In both studies the bloggers were highly educated, with 32% of the RB bloggers and 45% of the SciLogs bloggers having earned a PhD. Shema et al. (2012) and Fausto et al. (2012) confirmed the preference for high-impact journals reported by Groth and Gurney (2010). Both studies indicated that the journals most cited in blog posts were Science, Nature and the Proceedings of the National Academy of Sciences of the United States of America (PNAS). According to Fausto et al. (2012), RB had over 1,230 active blogs in 2012.

In this study, we explore whether the blog citations that articles may receive soon after their publication can predict future “regular” citations (i.e., citations from scholarly articles). To be more precise, we examine whether articles that are published in peer-reviewed journals and are reviewed in blogs aggregated by ResearchBlogging.org soon after their publication are better cited than articles published in the same year and in the same journal but that are not reviewed in the year of their publication in blogs aggregated by ResearchBlogging.org.

This question is of particular interest because previous research has shown a correlation between Twitter mentions ("tweets") about scholarly articles and "regular" citations received by the same articles later on, at least for one open access medical informatics journal (Eysenbach, 2011). In addition, another study found a correlation between twitter mentions of Arxiv pre-prints and subsequent Google Scholar citations (Shuai, Pepe & Bollen, 2012). It is difficult to keep track of tweets because they can disappear from Twitter's search facility if there are too many matching a search or if they are too old, although they can be recovered using third party Twitter data providers. Twitter has donated tweets to the Library of Congress, but a single search can take 24 hours, and only tweets older than six months can be searched (Library of Congress, 2013). Moreover, tweeting seems to be becoming more common for articles and publishers can automatically tweet all published articles, and so tweet counts may at some stage no longer be good predictors of future citations. In comparison, blogs are more sustainable, and their posts are often archived, making the tracking of blog citations easier than that of tweet mentions. Moreover, blog posts seem more difficult to automate and RB posts seem to be mainly written by experts, suggesting that they should have intrinsically more value than tweets. Thus it is especially interesting to find out whether early blog citations can predict future "regular" citations.

Figure 1: Snippet of a blog post aggregated by ResearchBlogging.org

Research setup

ResearchBlogging.org publishes an extended snippet of all the posts aggregated by it. An example of such a snippet can be seen in Figure 1. All the snippets of posts published during 2009 were downloaded using the DownThemAll add-on on to Firefox. Altogether 4880 snippets were downloaded. The following fields were extracted from these snippets: date of publication of the post, number of views of the post, title and URL of the blog post, name of the blogger and of the blog, and for each citation that appeared in the blog post (there are posts that contain several blog citations): author, title, year, source and DOI or URL of the specific publication. This process identified 6927 blog citations.

Since we were interested in blog citations which appeared soon after publication of the article, we considered only blog posts from 2009 reviewing articles

published in 2009. There were 4013 such items out of 6927 blog posts from 2009, indicating that bloggers tend to review newly published items. Next we limited the sample only to journals with 20 or more articles reviewed in ResearchBlogging.org during 2009. Only articles, reviews and proceedings papers were considered, thus editorials and letters and other document types were excluded. Articles which appeared numerous times in the sample were only counted once. A list of journals appears in Table 1. Details of the articles published in these journals during 2009 and the citations they received in 2009, 2010 and 2011 were retrieved from the Web of Science (WoS).

Table 1. Journals with more than 20 articles published in 2009 and reviewed in 2009 in blog posts aggregated by ResearchBlogging.org.

Journal	# articles published by the journal in 2009	# articles reviewed by bloggers in 2009
PLoS One	4403	193
PNAS	3765	166
Science	897	161
Nature	866	119
Psychological Science	234	48
Journal of Neuroscience	1542	40
Journal of the American Chemical Society	3332	34
Current Biology	357	28
PLoS Biology	195	26
New England Journal of Medicine	352	26
Pediatrics	752	23
Nature Neuroscience	208	22

* "articles" includes articles, reviews and proceedings papers

It is well-known that citation distributions are highly skewed, thus it is more reasonable to consider medians instead of averages (Bar-Ilan, 2012). Citations in the sciences typically peak after two to three years. We summed for each article the number of citations it received during 2009, 2010 and 2011 and also counted separately the number of citations it received during 2010 and 2011. Citations received during 2012 were not taken into account, because the data collection was carried out in November 2012, and thus the citation data for 2012 were not complete yet. The median number of citations received during the two periods 2009-2011 and 2010-2011 were computed both for the articles reviewed in blogs aggregated by ResearchBlogging.org and for the articles not reviewed by the bloggers for each of the 12 journals listed in Table 1. In order to test for significance Mann-Whitney non-parametric tests were run for each journal and each period.

Results and Discussion

Table 2. Median number of citations received by the reviewed and the non-reviewed articles in 2009

Journal	Median # citations received during 2009-2011 for 2009 articles reviewed RB in blogs in 2009	Median # citations received during 2009-2011 for 2009 articles not reviewed in RB blogs in 2009	Median # citations received during 2010-2011 for 2009 articles reviewed RB in blogs in 2009	Median # citations received during 2010-2011 for 2009 articles not reviewed in RB blogs in 2009
PLOS One	8	6	8	6
PNAS	20	16	17	14
Science	41	40	37	37
Nature	57	49	52	43
Psychological Science	8	9	7	8
Journal of Neuroscience	22	12	19.5	12
Journal of the American Chemical Society	19	14	18.5	13
Current Biology	13.5	15	13	14
PLOS Biology	18.5	17	16	15
New England Journal of Medicine	172	56	162	50.5
Pediatrics	13	7	12	7
Nature Neuroscience	32.5	24	30	21

Table 2 displays the median values of citations for the reviewed and the non-reviewed items for each journal. We see the medians are higher for the articles that received blog citations except for the journals Psychological Science and Current Biology. The most striking difference is for the New England Journal of Medicine, the median number of citations received by articles that received early blog citations is more than 3 times the median number of citations received by the articles that were not reviewed in 2009 in blog posts aggregated by ResearchBlogging.org. Table 3 shows the p-values of the Mann-Whitney tests for differences between the blogged and non-blogged groups. It can be seen that for 7 out of the 12 journals the differences are significant at $p<0.05$. The results for the Journal of the American Chemical Society are at the edge of significance for the 2009-2011 citation window.

Table 3. Results of Mann-Whitney tests, 2009.

Journal	p-values for the citation period 2009-2011	p-values for the citation period 2010-2011
PLoS One	0.002**	0.006**
PNAS	0.000**	0.000**
Science	0.975	0.865
Nature	0.044*	0.037*
Psychological Science	0.833	0.787
Journal of Neuroscience	0.000**	0.000**
Journal of the American Chemical Society	0.059	0.068
Current Biology	0.253	0.341
PLoS Biology	0.988	0.997
New England Journal of Medicine	0.000**	0.000**
Pediatrics	0.004**	0.003**
Nature Neuroscience	0.003**	0.003**

*p<.05. **p<.01.

Conclusions and Future Research

In this paper we presented preliminary results indicating blog citations as a potential source for alternative metrics. ResearchBlogging.org bloggers chose to cover articles which were significantly better cited than other articles published in 2009 for 7 out of the 12 journals studied. The research is still in-progress and was limited to 2009 posts and articles, but with 2012 ending, we intend to study the 2010 blog posts and articles as well. ResearchBlogging.org is ever-growing, thus the number of blog posts aggregated by it during 2010 is much higher than the number of blog posts in 2009. We expect to have more significant results for this period.

The results show that for some, but not all journals, articles blogged in RB tend to subsequently receive more citations than other articles from the same journal. There are many different possible reasons for the cases of significant differences: bloggers pick better articles to write about and these attract more citations; bloggers sometimes write about articles that they use in their research and perhaps have already decided to cite when they blog about them; bloggers pick articles that are not necessarily better but are more interesting and get more read and hence more cited because of their interest; or the publicity from the blog post generates awareness of an article that leads to more citations. Reasons why the converse could be true include: review articles tend to be highly cited but may be uninteresting to blog about because they contain no new research; and methodological articles may be less interesting to blog about than those with

practical, real world applications whereas the former may tend to be more highly cited and the latter less highly cited. Whatever the reasons, it seems that, on balance, RB bloggers tend to pick articles that go on to become more highly cited than average.

Acknowledgments

This work was funded by the EU FP7 ACUMEN project (Grant agreement: 266632).

References

Bar-Ilan, J. (2012). Journal report card. Scientometrics, 92(2), 249-260.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research,* 13(4). Retrieved February 24, 2012 from http://www.jmir.org/2011/4/e123

Fausto, S., Machado, F. A., Bento, L. F. J., Iamarino, A., Nahas, T. R., & Munger, D. S. (2012). Research Blogging: Indexing and Registering the Change in Science 2.0. PLOS ONE, 7(12), e50109.

Groth, P., & Gurney, T. (2010). Studying scientific discourse on the Web using bibliometrics: A chemistry blogging case study. Presented at the WebSci10: Extending the Frontiers of Society On-Line, Raleigh, NC, USA. Retrieved January 7, 2013 from http://journal.webscience.org/308/2/websci10_submission_48.pdf

Kousha, K., Thelwall, M. & Rezaie, S. (2010). Using the web for research evaluation: The Integrated Online Impact indicator, *Journal of Informetrics*, 4(1), 124-135.

Library of Congress (January 2013). *Update on the Twitter archive at the Library of Congress* [White paper]. Retrieved from http://lib.trinity.edu/research/citing/apaelectronicsources.pdf

Perfecting peer review? [Editorial]. (2011, January 7). *Nature Medicine*, 17, p. 1-2. doi:10.1038/nm0111-1

PLoS One. (n.d.). *Article-Level Metrics Information.* Retrieved January 20, 2013, from PLOS One: http://www.plosone.org/static/almInfo

Priem, J., Piwowar, H., & Hemminger, B. (2011). Altmetrics in the wild: An exploratory study of impact metrics based on social media. Presented at Metrics 2011: Symposium on Informetric and Scientometric Research. New Orleans, LA, USA, October 12. Retrieved January 7, 2013 from http://arxiv.org/abs/1203.4745v1

Puschmann, C. & Mahrt, M. (2012). Scholarly Blogging: A new form of publishing or science journalism 2.0? Presented at the Conference on Science and the Internet 2012, Dusseldorf, Germany.

Shema, H., Bar-Ilan, J.& Thelwall, M. (2012) Research blogs and the discussion of scholarly information. PLoS ONE 7(5): e35869. doi:10.1371/journal.pone.0035869

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: article downloads, twitter mentions, and citations. *PLOS ONE*, *7*(11), e47523.

DO NON-SOURCE ITEMS MAKE A DIFFERENCE IN THE SOCIAL SCIENCES?

Pei-Shan Chi

chi@forschungsinfo.de
Institute for Research Information and Quality Assurance, Schützenstraße 6a, 10117 Berlin (Germany)
Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Unter den Linden 6, Berlin, 10099 (Germany)

Abstract
The publications that are not indexed by well-known citation indices such as WOS or Scopus are named "non-source items," and have been ignored by most bibliometric analyses for a long time. This study explores the effect of the inclusion of non-source items in bibliometric evaluations by WOS in the social sciences, and finds that non-source items increase significantly the number of publications and less so the citations per item and H-indices, for evaluated researchers. The citation rates of non-source items are lower than those of source items, as a result of the limited coverage of WOS partially.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Background
With the large increase of research projects and funding in all countries in recent years, both public and private funding agencies have higher demands than previously to evaluate the effects of their funding and to trace the influence of research results. The quantitative way such as bibliometric methods, or the qualitative way such as interviews or peer reviews, are both frequently adopted in these evaluations. Quantitative methods are becoming widely accepted in evaluations, especially in the natural sciences, for their objective and time-saving nature. However, even though the application of bibliometric methods is more popular and meaningful in the natural sciences, the possibilities of applying bibliometric techniques in the social sciences should be explored (van Leeuwen, 2006).

ISI covers only "top-end" research performance, and provides an easy way for researchers to quickly monitor the publications with the most impact. However, the international orientation and high visibility threshold of WOS causes the loss of a lot of important social sciences literature which is published in local languages or in a locally-oriented channel. The fragmentation of social sciences

literature, which is effected by the heavy emphasis on local audiences and local materials, is another factor making it difficult to cover the research output comprehensively in a single international database (Hicks, 1999). Therefore, although the coverage of journal articles in the natural sciences and life sciences might be relatively high in the SCI (well above 80% or even 90% in many fields), the coverage of social sciences and humanities tends to be considerably less extensive in the SCI, the SSCI and the A&HCI (Hicks, 1999; Nederhof et al., 1989; Schoepflin, 1992).

Language bias is another related cause behind the formation of an incomprehensive database. National social sciences literature published in languages other than English are largely excluded from SSCI. About 93%-95% of the papers contained in the SSCI are published in English, 2%-3% in German, about 1% in French, and 2% in other languages (Nederhof, 2006). In the A&HCI, 70%-72% of the documents are in English, with other major languages being French (11%) and German (8%) (Nederhof, 2006; Nederhof & Noyons, 1990). Another limitation of the SCI, SSCI and A&HCI refers to their non-coverage of non-serial publications, especially for those fields in the social sciences and humanities where books are the most important publishing medium (Nederhof, 2006). The general trend that can be observed: the more books in a field, the less the literature covered by the SSCI. Butler and Visser (2006) found out that the proportion of total output covered in ISI journals ranges from 90% in chemistry to 6% in law.

The problematic publication coverage can also be measured by the reference coverage that shows the insufficiency of coverage by ISI in the social sciences. In a study which analyzed the data of Delft University of Technology from 1994 to 2003 (Van Leeuwen, 2006), the author reported the share of references to ISI covered publications in the social sciences showed similar results across different countries, from the lowest 35% (Germany) to 39% (US). The limited coverage of WOS will certainly lead to errors when applied to these subject fields. The bibliometric indicators which are applied in evaluation procedures in the social sciences therefore need to be considered carefully. Pointing out this feature, Hicks (2004, pp. 492) suggested that SSCI-based bibliometrics will work best if applied to science-like research such as economics and psychology. On the other hand, Nederhof (2006) suggested that one should not rely on ISI source serials only, but also needs to include: non-ISI source serials, monographs, contributions to edited volumes, formal reports, publications directed at a non-scholarly public. Furthermore, Norris and Oppenheim (2007) indicated that any database which covers the social sciences should incorporate to a greater degree the scholarly output found in monographs, reports, articles, and articles appearing in non-English language.

In order to test the actual coverage of WOS in the social sciences and the efficacy of bibliometric indicators, we should find out more about those missing non-source items and their role in bibliometric evaluations. Thus, this study investigates the differences between source items and non-source items in the social sciences in terms of their publication characteristics, citation characteristics, and the results of individual evaluations, to see if the inclusion of non-source items really makes a difference from the normal evaluation results based only on ISI source items.

Data and Methods

This study focuses on Political Science, as it is a field in which there is a relative large number of empirical studies and as a previous study has shown, among the top three fields with the largest increase in citations caused by the inclusion of non-source items (Butler & Visser, 2006). The five year publication output (2003-2007) of two top-ranking German institutions, Department of Political Science at Mannheim University and Institute of Political Science at University of Muenster (CHE, 2010; Hix, 2004), were chosen as research samples. The 1,015 publications of 33 professors in these two institutions were collected from researchers' official websites, institutional repositories, and German Social Science Literature Information System (SOLIS). After data collection, all publications were sent to the professors for verification. References and citations of these items were obtained from March until December 2012 from the WOS in-house database of the Competence Centre for Bibliometrics for the German Science System (Kompetenzzentrum Bibliometrie).

Citations to all items were acquired by matching the corresponding specific search terms to the references in the WOS in-house database following a set of rules formulated particularly for different document types. It takes at least two rounds of SQL queries to identify the references in WOS. Take *Book Chapter* for example, during the first round, we search for *the first word of the title* and *the first author name* (surname, first initial) in the WOS references, then filter the results to include only those items listed with publication year within ±1 year of the target one and try to collect the different abbreviations of journal titles and author names indexed in the database from these results. A specific rule applied to *Book Chapter* here is, the matched references with an accurate first page number may have the ±1 publication year difference from than the original one, but those without accurate first page numbers need to match the publication year exactly. Next, the second round is to repeat the search query for matching *the first page* and *the first author name*, and filter the results again. In the end, duplicates of the combined results from these two rounds will be deleted. For some articles showing no exact first page in their references where there are more than two chapters written by the same author in one book, we manually checked the full-text of the articles to make sure which chapter was cited exactly.

The citations discussed in this study include self-citations. According to Table 1, the general self-citation rate of the publication set (without a fixed citation window) is about 16%. (i.e., about 16% of the citations the authors from these institutions receive in the WOS database are from their own publications indexed in WOS, on average). *Edited Book* has the lowest rate (9.9%). Other types, such as *Others, ISI Journal Article, Book* and *Book Chapter*, have much higher average self-citation rates, which are over or close to 20%. Since the self-citations do not completely dominate the total citations, we tackle the analyses in this study based on total citations.

Table 1. Numbers of publications, citations, and self- citations by document types

	No. of pub.	*No. of Cit.*	*No. of self-Cit.*	*% self-Cit.*	*Ave. Cit.*	*Ave. Cit. (w/o self-cit.)*
ISI Journal Article	70	498	96	19.28%	7.11	5.74
Non-ISI Journal Article	151	189	20	10.58%	1.25	1.12
Book	45	84	16	19.05%	1.87	1.51
Edited Book	76	303	30	9.90%	3.99	3.59
Book Chapter	396	198	36	18.18%	0.5	0.41
Conference Paper	151	26	4	15.38%	0.17	0.15
Others*	126	56	17	30.36%	0.44	0.31
Total	1,015	1,354	219	16.17%	1.33	1.12

*'Others' include Working Paper, Presentation, Report, Lecture/Speech, Discussion Paper, Magazine/Newspaper Article and other types with less than 10 items.

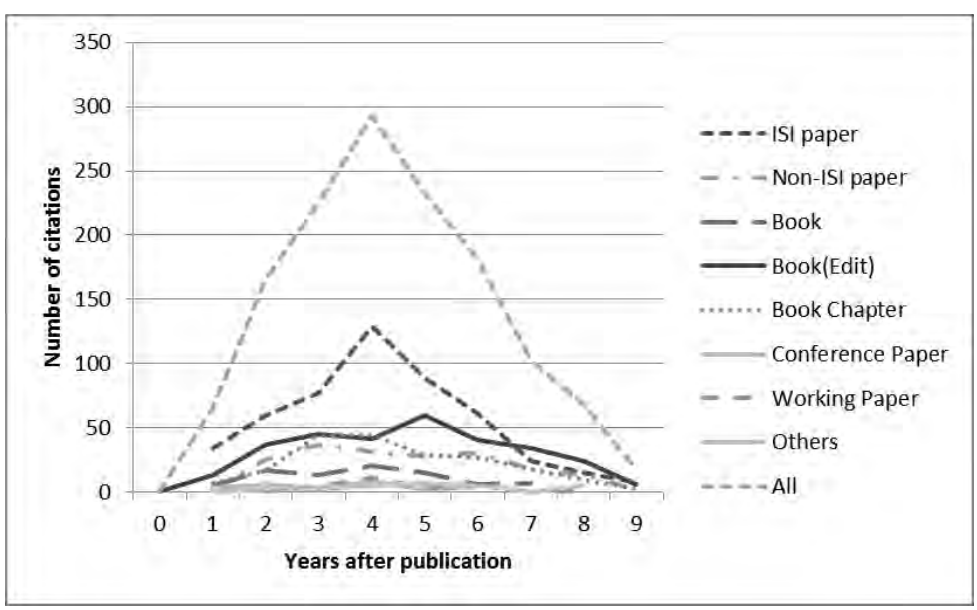


Figure 1. Analysis of the period being cited after published

A citation window of four years (Glänzel, 1997; 2008) is applied in this study. According to Figure 1, most publications were cited within four years after they were published (I.e., an item published in the year 2003 was cited by an article published in the year before 2007). This is the main reason why four-year citation window is applied in this study to calculate citations for all publications. Figure 1 also shows that conference papers are not cited after a long period; the longest citation life of them is six year.

Results

Overall Data

As shown in Table 2, 39% of the 1,015 publications of the selected two institutions are *book chapters* and 22% are *journal articles* (in both peer-reviewed and non peer-reviewed journals). Most of these 221 journal papers are published in peer reviewed journals (73%). Among these 161 peer reviewed journal articles, 70 are indexed by WOS (44%) and 56 (80%) of those indexed are published in English. Therefore, the overall coverage of dataset representing the German political science in WOS is about 7% (70/1015). These 161 *peer reviewed journal articles* receive almost half of all; the second most cited category, *edited books*, receive around one fifth of all citations.

Table 2. Numbers of publications and citations of different document types

Document Types	***Items(%)***	***Total citations up to 2012(%)***	***Citations w/in 4-year Citation Window(%)***
Book Chapter	396(39.0)	198(14.6)	112(14.9)
Journal Article (Peer Reviewed)	161(15.9)	639(47.2)	373(49.7)
Conference Paper	151(14.9)	26(1.9)	20(2.7)
Book (Editor)	76(7.5)	303(22.4)	138(18.4)
Journal Article (non-Peer Reviewed)	60(5.9)	48(3.5)	24(3.2)
Book (Author)	45(4.4)	84(6.2)	56(7.5)
Working Paper	29(2.9)	28(2.1)	17(2.3)
Presentation	16(1.6)	0	0
Report	16(1.6)	3(0.2)	1(0.1)
Lecture/Speech	14(1.4)	0	0
Discussion Paper	10(0.9)	6(0.4)	4(0.5)
Magazine/Newspaper Article	10(0.9)	2(0.2)	1(0.1)
Others	31(3.1)	17(1.3)	4(0.5)
Total	1,015(100)	1,354(100)	750(100)

Note: Types with less than 10 items are combined into 'Others'.

As per Figure 2, it is shown that in a rough 3:1 ratio these German political scientists publish *book chapters* in German vs. in English. In other words, about 70% of the published book chapters are in German. The dominant position of German is also prevalent in other publication types, such as *edited books*, *books*,

and *non peer reviewed journal articles*. However, English is used more often than German in p*eer reviewed journal articles* and *conference papers*. These types obviously serve more international communication purposes and are therefore written in English. Compared to ISI papers which are published mostly in English, the other 91 non-ISI peer reviewed papers are rather published in German (60%) than in English (37%).

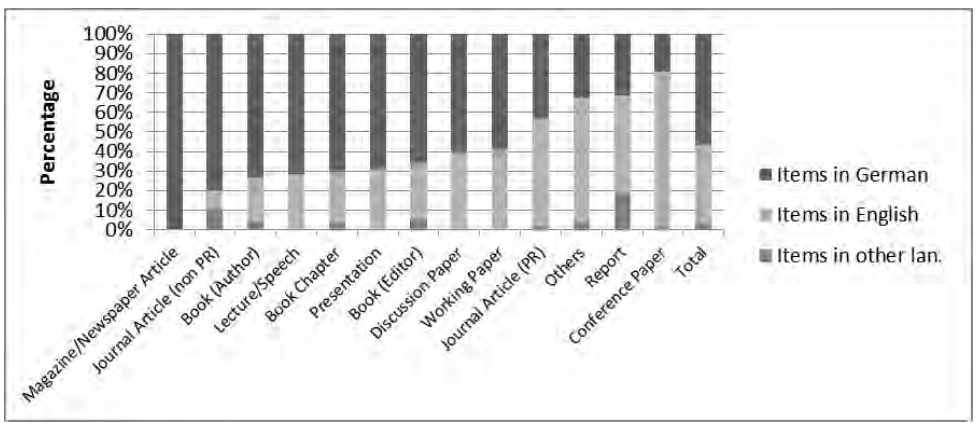


Figure 2. Document types ordered by the share of German language publications

Comparison between source items and non-source items

After the introduction of the dataset as a whole, we now focus on the comparison of source items vs. non-source items. In particular, differences in publication characteristics, citation results, and evaluation characteristics will be analyzed.

1) Publication Characteristics

As mentioned in the overall data section, in German political science only 7% of the items are indexed in WOS. In terms of language, Figure 3 shows that these 70 source items are dominantly published in English (80%), while non-source items are more often written in German (60%) than in English (37%).

Figure 4 shows that non-source items are published as *book chapters* (42%), *journal articles* (16%) and *conference papers* (16%) mostly. *Edited and authored books* account for 13% in total. In Figure 5, ISI journal articles as well as non-ISI journal articles are often published as *articles*. *Reviews* or *editorial materials* do not take a dominant role in composition.

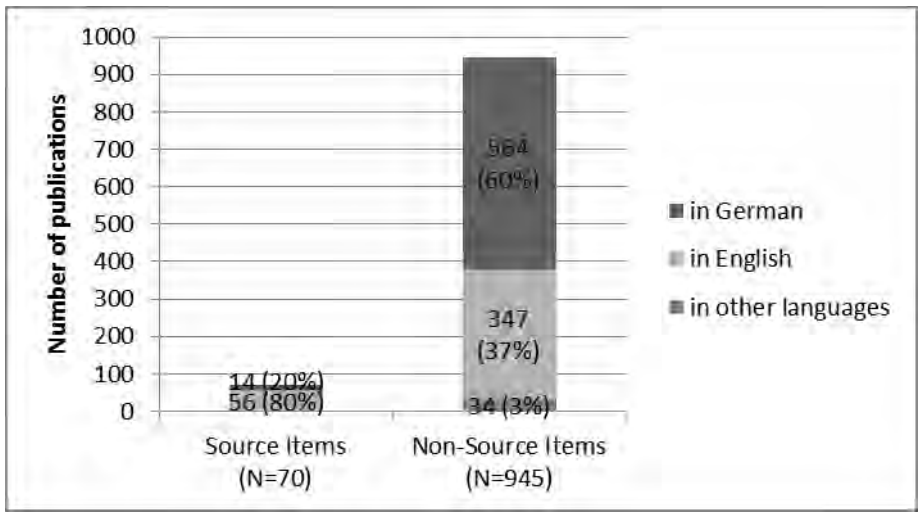


Figure 3. Language analysis of source items and non-source items

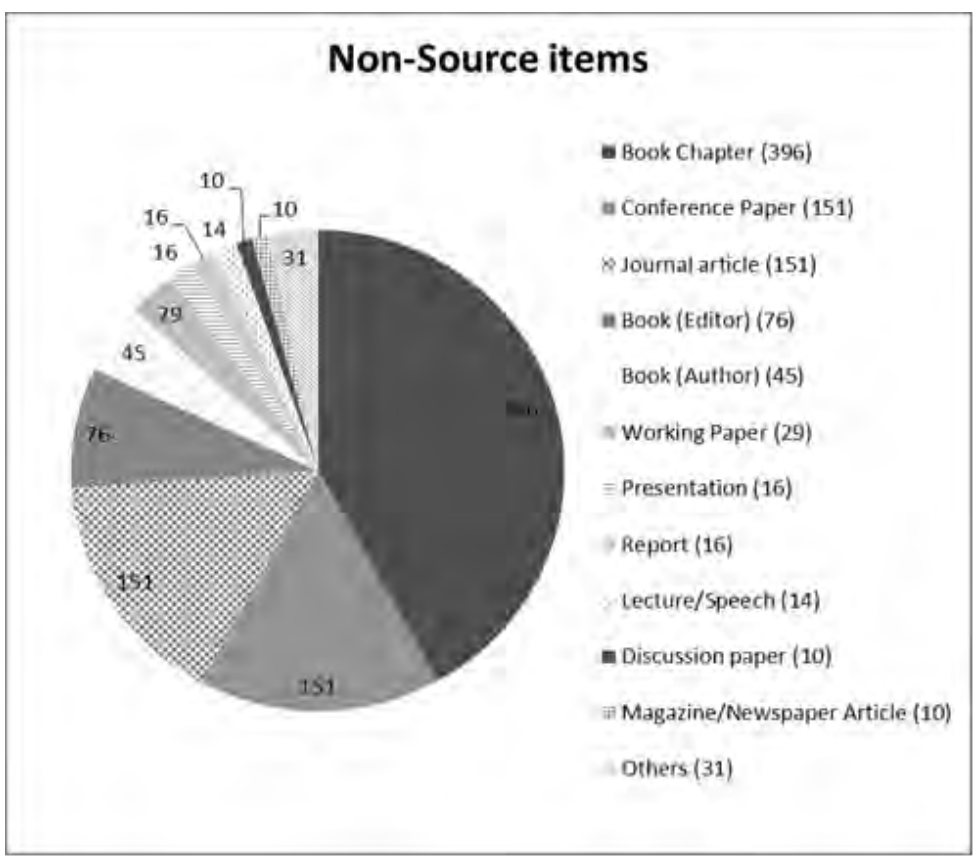


Figure 4. Document type analysis of non-source items

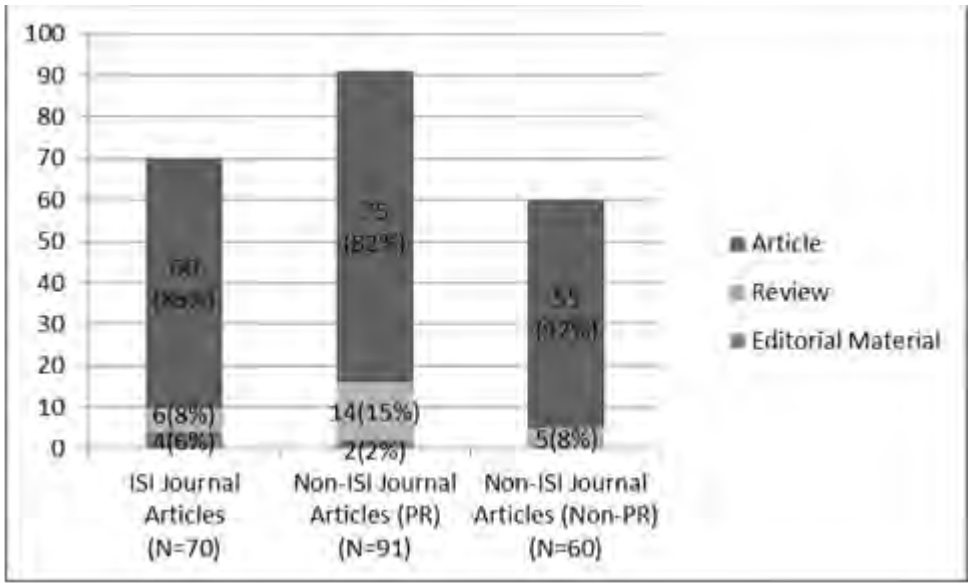


Figure 5. Document type analysis of ISI journal articles and non-ISI journal articles

2) Citation Characteristics

In this study, 70 source items receive a total of 300 citations in WOS within a four-year citation window, while the 945 non-source items receive 450 citations. This means that, the inclusion of non-source items increases the value of the indicator "number of publications" considerably (+1350%), but not so much the number of citations (+150%)[59]. Therefore, the inclusion dilutes the average citation rate of source items (4.29), and generates a lower rate for all items (0.74). Source items, no matter whether in English or German, receive higher average citation rates than non-source items (Fig. 6). From the perspective of language, it is obvious that papers in English are perceived by a broader audience in WOS, resulting in substantially more citations. However, the difference between English and German is not as large as the difference between source items and non-source items.

[59] Although it is important to point out that the citations from non-source items to non-source items cannot be measured with the current methodology. These 'non-source citations' could increase the overall citations, especially of regional publications, considerably.

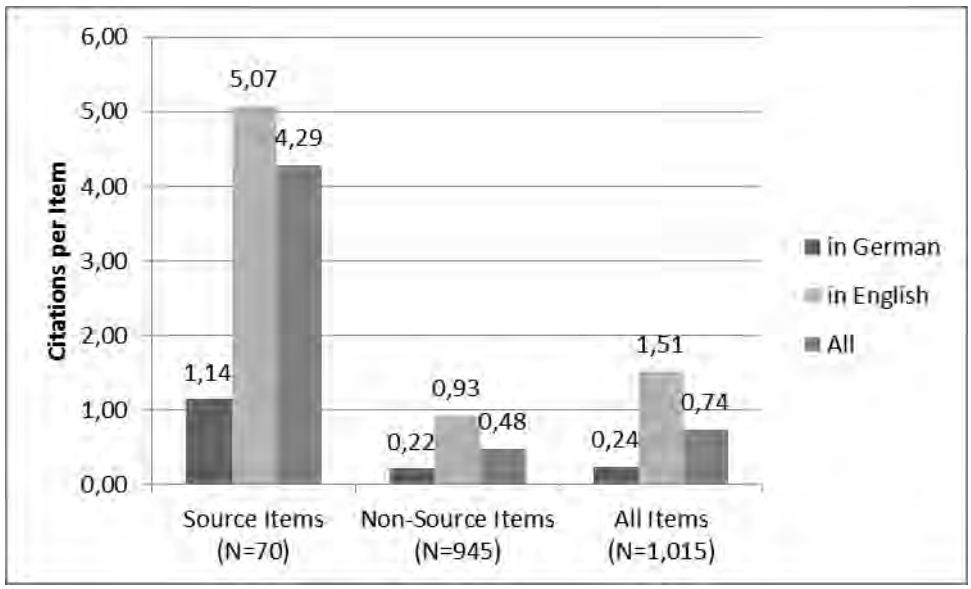


Figure 6. Average citation rate of source items and non-source items

Figure 7 shows that around 56% of items in German are cited by ISI articles in German, but items in English are cited mostly by articles in English. In general, non-source items are cited more by ISI articles in German (23%) than source items (9%).

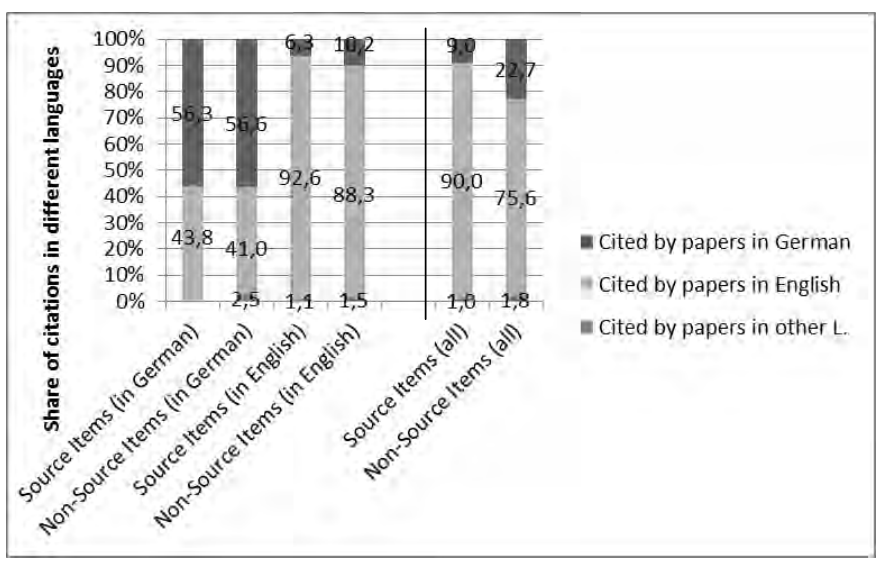


Figure 7. Share of languages of citations of source items and non-source items

Figure 8 shows that source items as well as non-source items are mostly cited in *articles* rather than other types of document. Compared to the share of *articles* in journal articles (86%) shown in Figure 3, they are cited a little bit more by *articles* (85-90%). In other words, about 86% of journal articles are published as *articles*, while around 87 % of items are cited by *articles*. In this study, the document types of source items are assigned by WOS automatically, while the types of non-source items are assigned by their authors.

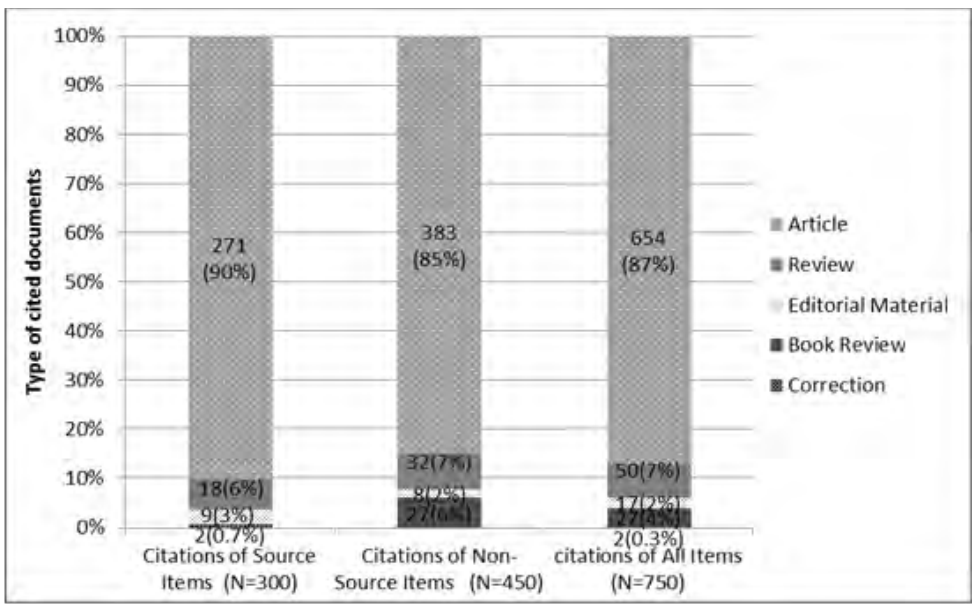


Figure 8. Share of document types of citations of source items and non-source items

3) Individual Evaluation

From the perspective of individual research performance, Figure 9 shows that non-source items increases individual publication output to a much higher degree than it does in total citations (without citation window). The individual number of publications of source items vs. all items has a low correlation (Kendall's tau-b correlation coefficient = .348, $p < 0.01$). It tells us that the inclusion of non-source items does make an enormous difference in the number of publications, predominantly due to many professors having no or only a single source-item publication while showing quite a prolific work which is not indexed. In contrast, the inclusion of non-source items does not change the number of citations as much. The number of citations of source items and all items is in a relatively high correlation (Kendall's tau-b correlation coefficient = .719, $p < 0.01$).

Concerning the H-index, the inclusion of non-source items contributes an increase to the original number generated by source items. In Figure 9, most of the 33 researchers reach a higher H-index when all items are considered. Wilcoxon

signed rank test confirmed that the median of differences between the H-index of source items and the H-index of the inclusion of non-source items is significant different ($p<0.05$). However, the relative change in values by inclusion of non-source items is less prominent (Kendall's tau-b correlation coefficient = .711, $p < 0.01$).

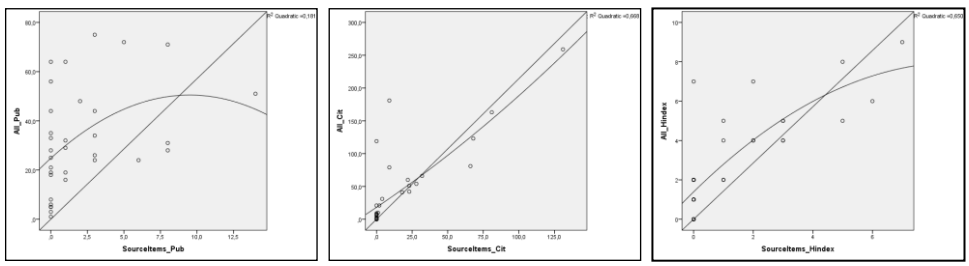


Figure 9. Scatter plot of comparing number of publications, number of citations, and H-indices of each professor based on all their publications vs. their source item publications

Conclusions

The emphasis of this study is to investigate the differences in typical WOS-based bibliometric measures made by the inclusion of non-source items in the social sciences. In this study we see that non-source items increase the number of publications significantly when evaluating these professors. Even though the increase in numbers of publications is massive, the additional publications do not lead to an equivalent increase in the number of citations and H-indices to a concordant amount (however, see footnote 1). The citation rate of non-source items is much smaller than that of source items on average, presumably at least in part as a result of the limited coverage of WOS. Thus, adding them in even lowers the average citation rate per publication. However, speaking from the cited data a further study discussing an addition of the relatively highly cited non-source items such as *edited books* or *authored books*, could be explored.

On the other hand, non-source items do not show a significant difference in the share of the languages of citing papers, or of citation half-time. Both source items and non-source items have similar values in these indicators.

From the data on publications of German political scientists, in Figure 2 it is obvious that they use German in their local communication circles to publish in books and regional oriented journals, but use English as a communication means in more international channels such as peer reviewed journals and international conference papers. Thus, the source items in this study are published in English much more often than in German, while non-source items are in predominantly in German. This is also a hint reminding us that a different but important

communication channel may be ignored if we only focus on source items when evaluating social scientists by applying WOS-based bibliometric indicators.

Concerning the publication types, 51% of all publications in these two political institutions are *books* and *book chapters*, while all *journal articles* combined contribute only 22%. If we take a closer look at non-source items, they are published mostly as *book chapters* (42%), followed by *conference papers* (16%) and *non-ISI journal articles* (16%). *Edited books* have the highest average citation rate per paper among non-source items. The interesting thing is, even though the *journal articles* are not the majority of publications, the share of ISI journal articles among all journal articles (72%) is much higher than non-ISI journal articles (28%). This shows that the main publication channels for these German political scientists are *books* or *book chapters* rather than *journal articles*, but they do prefer to publish in ISI journals when publishing journal articles. Furthermore, non-ISI journals articles are published more in German (68%) than in English (26%), and consequently receive much lower citations due to the more limited audience. We suppose that for the purpose of publishing in local interests they use *books* and *book chapters*, and they publish *journal articles* more for building up an international communication platform. They refer to many WOS articles when they publish ISI journal articles for an 53% internal WOS coverage as shown in a previous study (Chi, 2012). In particular, they reference American journals the most.

The main finding of this study tells us that without including non-source items we may miss on average 93% of publications, and 60% of citations (with four-year citation window; without citation window 63%) belonging to the researchers of these two German institutions. The influence of non-source items can thus not be underestimated. In addition, given the low number of German-language journal publications and the complete lack of two of the main German-language publication types in this field (books and book chapters) in WOS, the actual percentage of citations missed is likely much higher still when counting citations *from* WOS non-source items in addition to those *to* WOS non-source items. What we could suggest for the coverage of publications from an evaluation perspective is that other document types of publications than *journal articles*, especially *monographs,* should be included in bibliometric evaluations in political science since non-ISI journal articles may not take such an important place as other locally oriented document types. It has been shown that there are three main publication types used by German political scientists as a publication venue, namely *monographs*, *conference papers*, and *journal articles*. Thus, further studies could decide which types of publications should be collected for different kinds of evaluations in order to attain a valid assessment.

Acknowledgments

This study is supported by the Competence Centre for Bibliometrics for the German Science System founded by the German Federal Ministry of Education and Research. The author would like to thank her colleagues, Jian Wang, Daniel Sirtes, Andreas Strotmann, William Dinkel and Manuela Zinnbauer, and three reviewers for providing important comments on this study.

References

Butler, L., Visser, M. S. (2006). Extending citation analysis to non-source items. *Scientometrics, 66*(2), pp. 327-343.

CHE University Ranking 2010/11-Political science (2010). Retrieved 10. 08. 2010, 2010, from http://ranking.zeit.de/che2010/en/rankingkompakt?esb=28&hstyp=1

Chi, Pei-Shan. (2012, October). Bibliometric Characteristics of Political Science Research in Germany. , in: *Grove, Andrew (ed):* ASIST 2012: Information, Interaction, Innovation, Proceedings of the 75th ASIS&T Annual Meeting, Vol. 49.

Glänzel, W. (1997). On the possibility and reliability of predictions based on stochastic citation processes. *Scientometrics, 40*(3), pp. 481-492.

Glänzel, W. (2008). *Seven Myths in Bibliometrics. About facts and fiction in quantitative science studies.* Paper presented at the Proceedings of Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting, Berlin.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics, 44*(2), pp. 193-215.

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glänzel, U. Schmoch (Ed.), *Handbook of Quantitative Science and Technology Research*: Springer.

Hix, S. (2004). A global ranking of political science departments. *Political Studies Review, 2*(3), pp. 293-313.

Nederhof, A. J., Zwaan, R. A., De Bruin, R. E., Dekker, P. J. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social and behavioural sciences: A comparative study. *Scientometrics, 15*(5-6), pp. 423-435.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*(1), pp. 81-100.

Nederhof, A. J., Noyons, E. C. M. (1990). *trends in publication and international co-publication activity in the social and behavioral sciences and the humanities(1984–1989)*. Leiden: report CWTS-90-05.

Norris, M., Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics, 1*(2), pp. 161-169.

Schoepelin, U. (1992). Problems of representativity in the social sciences citation index. In P. Weingart, R. Sehringer, M. Winterhager (Ed.), *Representations of Science and Technology: Proceedings of the International Conference on Science and Technology Indicators* (pp. 177-188). Bielefeld: DSWO Press.

van Leeuwen, Th. N. (2006). The application of bibliometric analyses in the evaluation of social science research. Who benefits from it, and why it is still feasible. *Scientometrics, 66* (1), pp. 133-154.

DOWNLOAD VS. CITATION VS. READERSHIP DATA: THE CASE OF AN INFORMATION SYSTEMS JOURNAL

Christian Schlögl[1], Juan Gorraiz[2], Christian Gumpenberger[2], Kris Jack[3] and Peter Kraker[4]

[1]*christian.schloegl@uni-graz.at*
University of Graz, Institute of Information Science and Information Systems, Universitätsstr. 15/F3, A-8010 Graz (Austria)

[2]*(juan.gorraiz|christian.gumpenberger)@univie.ac.at*
University of Vienna, Vienna University Library, Dept of Bibliometrics, Boltzmanngasse 5, A-1090 Vienna (Austria)

[3] *kris.jack@mendeley.com*
Mendeley, 144a Clerkenwell Road, London, EC1R5DF (UK)

[4]*pkraker@know-center.at*
Know-Center, Inffeldgasse 13, A-8010 Graz (Austria)

Abstract
In our article we compare downloads from ScienceDirect, citations from Scopus and readership data from the social reference management system Mendeley for articles from the Journal of Strategic Information Systems (publication years: 2002-2011). Our study shows a medium to high correlation between downloads and readership data (Spearman r=0.73) and between downloads and citations (Spearman r=0.77). However, there is only a medium-sized correlation between readership data and citations (Spearman r=0.51). These results suggest that there is at least "some" difference among the two usage measures and the (citation) impact of the analysed information systems articles. As expected downloads and citations have different obsolescence characteristics. While the highest downloads accrue the first years after publication, it takes several years until the citation maximum is reached.

Conference Topic
Scientometrics Indicators (Topic 1), Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2)

Introduction
There exist already a slew of studies which have compared download and citation data. These studies can be divided to two groups: investigations having been performed at local level and those having been conducted at global level (Bollen and van de Sompel, 2008). While the former are restricted to a specific user population (e.g. a university), global studies are performed on a world-wide

context. Usually they use download data from repositories/preprint archives, open access journals or e-journals from (commercial) publishers as primary data source. Examples for the latter can be found, for instance, in Moed (2005) and in Schloegl and Gorraiz (2010, 2011).
With the advent of the social web and its growing acceptance in academia, alternative metrics seem to be a further source for the measurement of science (Bar-Ilan et al, 2012). In particular, what is called "longer-term metrics" in an editorial of a Nature article (Anonymous, 2012) seems promising. These metrics are based on downloads, readers and user comments. An example is the social reference management system Mendeley. So far social media has not been accepted as part of the measurement of scientific achievement because it has not yet been sufficiently validated. The few investigations which are known to the authors can be found in Bar-Ilan (2012), Bar-Ilan et al. (2012), Kraker et al. (2012), Li, Thelwall and Giustini (2012) and Li and Thelwall (2012). As a consequence, this research in progress paper is to provide one more evidence concerning the potential of social media using the example of Mendeley. In particular, the following issues will be addressed:

- Are most cited articles the most downloaded ones and those which can be found most frequently in user libraries of the collaborative reference management system Mendeley?
- Do citations and downloads have different obsolescence characteristics at publication level?
- Are there other features in which citation, download and readership data differ?

Methodology and data sources

All the following analyses were performed for the Journal of Strategic Information Systems. Both citations and downloads were provided by Elsevier in the framework of the Elsevier Bibliometric Research Program (EBRP). For all documents published between 2002 and 2011 all monthly downloads were made available from ScienceDirect and all monthly citations from Scopus until mid of 2012. Furthermore, we received the total number of occurrences of full length articles in user libraries in Mendeley from 2002 to 2011.

Mendeley provides users with software tools that support them in conducting research (Henning & Reichelt 2008). One of the most popular of these tools is Mendeley Desktop, a cross-platform, freely downloadable PDF and reference management application. It helps users to organize their personal research libraries by storing them in relevant folders and applying tags to them for later retrieval. The articles, provided by users around the world, are then crowd-sourced into a single collection called the Mendeley research catalogue (see Hammerton et al. (2012) for details). At the time of writing, this catalog contains more than 80 million unique articles, crowd-sourced from over 2 million users, making it an interesting source of data for large scale network analysis.

Furthermore, Mendeley enables users to create and maintain a user profile that includes their discipline, research interests, biographical information, contact details, and their own publications. Mendeley then takes this data and automatically generates a profile page for the user that acts as a CV in which they can showcase their expertise. The user's publications are also augmented by readership counts, allowing them to track the popularity of their individual papers within the Mendeley community. These readership counts indicate how many Mendeley users have added the author's article to their personal research library.
To find corresponding articles in the Mendeley catalog, we matched paper titles reported from Elsevier to the titles of articles in the Mendeley database. Since there can be slight differences between article title across the two databases, we employed the Levenshtein distance when matching them up to one another in order to take account of these inconsistencies. We found good matching results of around 99.9% accuracy when employing a Levenshtein ratio of 1/15.83. Nevertheless, we manually verified borderline cases to reduce the likelihood of false positive matches.

Results

Download data

Table 1. Downloads per download type (pdf or HTML)
(publication years: 2002-2011, n=321 docs, download years: <=2011)

Download type	*%*
HTML	39%
Pdf	61%

There are two download types available in ScienceDirect from which pdf was used most (approximately in 61% of all cases between 2002 and 2011 – see Table 1) for the information systems journal under consideration.
As can be seen in Table 2, 94 per cent of all downloads allotted to full length articles (FLAs) which have a proportion of 56 per cent among all document types in ScienceDirect. As a consequence, the number of downloads per document is by far the highest for this document type. Interestingly, documents of other types are also downloaded to some extent, even though several magnitudes lower.

Since the analyzed journal appears in digital form and in print, there is usually a gap between the print publication date and the time when the document is put online. When not considering the one document assigned to the document type "Erratum", FLAs also have an outstanding role here. As is exhibited in Table 3, an electronic "full length article" appeared nearly two months (50 days) before print publication on average.

Table 2. Distribution of document types (n=321 documents) and downloads (publication year: 2002-2011, download year: <=2011) per document type.

Document type	*n*	*% docs*	*% downloads*	*Downloads per doc – relations*[1]
Announcement	5	1.6%	0.4%	5.9
Book review	4	1.2%	0.3%	5.5
Contents list	29	9.0%	0.4%	1.0
Editorial Board	29	9.0%	0.6%	1.5
Editorial	49	15.3%	3.3%	4.6
Erratum	1	0.3%	0.1%	5.7
Full length article	181	**56.4%**	**94.1%**	**35.4**
Index	12	3.7%	0.2%	1.3
Miscellaneous	9	2.8%	0.2%	1.8
Publishers note	2	0.6%	0.2%	7.0
	321	100%	100%	

[1] Since the download numbers are very sensitive, we did not provide the absolute figures but only the relations among them.

Table 3. Average difference between print and online publication date (print publication years: 2002-2011) (n=321 docs)

Document type	n	Online date - print publication date (mean days)
Announcement	5	-13.2
Book review	4	-40.5
Contents list	29	12.9
Editorial Board	29	12.9
Editorial	49	9.0
Erratum	1	-145.0
Full length article	**181**	**-49.8**
Index	12	-4.9
Miscellaneous	9	32.9
Publishers note	2	-13.0
	321	-24.9

Since FLAs are the most interesting type of document from a science perspective, we performed the obsolescence analysis only for this document type. As Table 4 shows, there was a huge increase in the number of downloads between 2002 and 2011. By far the largest proportion of this increase is due to the fact that with each (download) year the range of downloadable documents increased (from 13 in 2002 to 181 in 2011). However, also the general rise in the use of e-journals between 2002 and 2011 might have partly contributed to this increase.

An analysis of the obsolescence characteristics reveals that from the downloads of a certain year, most of them allot to articles either published in the download year or one year earlier (formatted in bold). In six cases articles were already downloaded one year before print publication (in grey) since they were already

available online. Accordingly, it can be concluded that more downloads accrue to recently published articles. However, also older articles are downloaded relatively often. In contrast, in our former studies in the fields of oncology (Schloegl & Gorraiz 2010) and pharmacy (Schloegl & Gorraiz 2011) half of the downloads were already made within the first two years after publication.

Table 4. Yearwise *relation*[1] of downloads per print publication year (2002-2011), (doc type: full length article, download year: <=2011) (n=181)

Pub year	n	Download year											downloads per doc – *relations*[1]
		2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	All	
2002	13	1.0	**2.3**	1.7	1.3	1.2	1.4	2.4	2.8	2.8	2.7	19.6	7.4*x
2003	21	0.0	1.3	**2.2**	1.0	1.0	0.9	1.5	1.3	1.5	1.1	11.9	2.8*x
2004	17			1.7	**2.6**	2.1	2.2	2.4	2.7	2.9	2.3	18.9	5.5*x
2005	18				1.7	2.3	1.8	2.0	2.4	2.6	2.2	15.0	4.1*x
2006	14				0.2	**2.4**	2.1	1.8	2.1	2.0	2.0	12.5	4.4*x
2007	18					0.0	**2.7**	**3.6**	3.4	3.5	2.9	16.1	4.4*x
2008	16						0.0	2.9	**3.5**	3.0	2.4	11.8	3.6*x
2009	14								3.1	**4.0**	3.1	10.2	3.6*x
2010	21									3.9	4.4	8.3	2.0*x
2011	29									0.3	**5.6**	5.9	1.0*x
all	181	1.0	3.7	5.6	6.8	8.9	11.1	16.6	21.4	26.4	29.0	130.4	

[1] Since the download numbers are very sensitive, we did not provide the absolute figures but only the relations among them.

Citation data

Table 5 shows, first of all, that ScienceDirect and Scopus use different document types which are not compatible to each other. The document type "full length article" in ScienceDirect mainly corresponds to the three Scopus document types "article", "conference paper" and "review". As expected, reviews receive more citations per document (20.2) than articles (14.8) whereas conference papers received only very little citations.

Table 5. Distribution of Scopus document types and citations per document type (2002-2011).

Doc type	no. docs	no. uncited	% uncited	Cites	%	Cites per doc type
ar	151	22	15%	2563	86,4%	14.8
cp	13	9	69%	8	0,3%	0.4
ed	33	26	79%	13	0,4%	0.2
re	18	1	6%	383	12,9%	20.2
all	215	58	27%	2967	100%	10.9

ar=article, cp=conference paper, ed=editorial, re=review

One interesting fact is that more than one quarter (27%) of all documents were not cited in the citation window (2002-2011). This is mainly true for editorials (79%) and conference papers (69%). (In contrast, there allotted a certain download volume also for document types like "editorial", "book review" or "announcement" in ScienceDirect.) Also the publication date has a great influence on the citation rate. Usually only a minority of the articles are cited in the year of publication. For instance, 21 articles from 2011 did not receive any citation in the publication year.

Table 6 shows the year-wise citation distribution of articles, reviews and conference papers between 2002 and 2011. As can be seen, in all citation years – from which 2011 is the most interesting one because it has the longest time frame – most citations (formatted in bold) accrue to articles from the publication year 2002. In contrast, as was already mentioned above, only a few documents were cited in the year of publication. This shows a clear difference to downloads which have their maximum either in the year of publication or one year later.

Table 6. Year-wise citations (2002-2011) per publication year (document types: article, review, conference paper), only cited documents (n=150).

Pub year	n	Citation year											cites per doc
		2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	all	
2002	13	2	19	**38**	**69**	**88**	**105**	**158**	**165**	**194**	**199**	1037	79.8
2003	14		1	6	21	27	39	35	41	40	39	249	17.8
2004	17				15	40	56	74	78	88	107	458	26.9
2005	19					16	46	78	76	93	99	408	21.5
2006	14				1	2	14	31	31	53	49	181	12.9
2007	18						1	31	74	92	85	283	15.7
2008	15							3	30	69	83	185	12.3
2009	14								3	34	57	94	6.7
2010	18									5	40	45	2.5
2011	8										14	14	1.8
all	150	2	20	44	106	173	261	410	498	668	772	2954	

Readership data

Since time stamps of the readership data were not available at the date of analysis, we could not perform an obsolescence analysis. Instead, Table 7 displays how many times (full length) articles from the publication years 2002-2011 were mentioned in total in Mendeley user libraries. Contrary to downloads and in particular to citations, the distribution of the occurrences is relatively even. One reason why older articles do not have higher readerships could be that Mendeley started in 2009 and has become popular in 2010.

Another interesting characteristic of Mendeley is its user structure. A preliminary analysis of the readers of the Journal of Strategic Information Systems revealed that by far the majority of them are students, in particular PhD students.

Table 7. Readership data per print publication year (2002-2011), (doc type: full length article, data extracted from Mendeley: October 2012) (n=181)

Publication year	n	Occurrences in user libraries	Occurrences per doc
2002	13	566	43.5
2003	21	344	16.4
2004	17	471	27.7
2005	18	371	20.6
2006	14	382	27.3
2007	18	580	32.2
2008	16	451	28.2
2009	14	416	29.7
2010	21	499	23.8
2011	29	537	18.5
all	181	4617	25.5

Comparison among downloads, citations and readership data

Figure 1 shows a medium to high relation among downloads, citations and readership data which is higher for downloads and citations (Spearman r = 0.77) and for downloads and readership data (Spearman r = 0.73). Among the ten most downloaded articles, seven (not the same) are in the top-10 readership and citation rankings. The correlation was lower between readership data and citations (Spearman r = 0.51) but in line with previous studies. For instance, Bar-Ilan (2012) calculated a correlation between Mendeley and Scopus for articles, reviews and conference papers from the Journal of the American Society of Information Science and Technology (publication years: 2001-2011) of 0.5 (data collection: April 2012). The correlation identified by Li, Thelwall and Giustini (2012) was similar between WoS citations and occurrences in Mendeley user libraries for articles having appeared 2007 in Nature (Spearman r=0.56) and Science (Spearman r=0.54) (data collection: July 2012). Only the analysis by Li and Thelwall (2012) found a higher correlation (Spearman r=0.68) between Mendeley and Scopus for 1397 genomics and genetics articles published in 2008 (data collection: January 2012). One reason for the lower correlation between Mendeley readership and citation data could be that Mendeley users have only been creating their libraries since 2009. Therefore, older articles may have lower occurrences in comparison to downloads in ScienceDirect and, in particular, to citations in Scopus, where there was the possibility to download/cite them already before 2009. Another reason could be that Mendeley users are younger (most are PhD or Master students) who prefer more up-to-date articles. This could in particular be true for computer science. One indication for both arguments could be that there was one article from the publication years 2006, 2008, 2009 and 2010 respectively in the top-10 readership ranking, while the most up-to-date article in the corresponding citation ranking was from 2005.

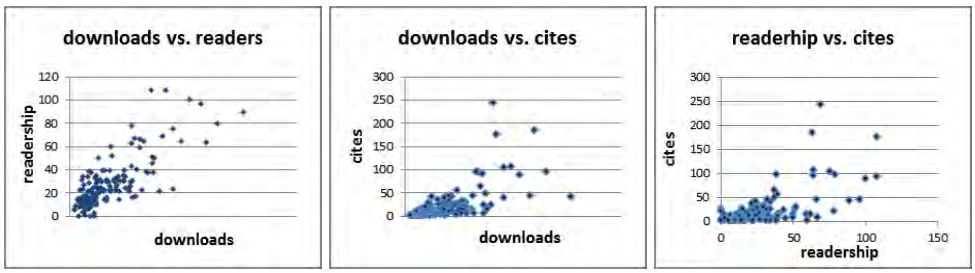


Figure 1. Downloads vs. readers vs. cites, scattergram (publication year: 2002-2011, doc type: full length article, only articles cited at least once) (n=151)

Conclusions and future research

Our analyses revealed both commonalities and differences among citations, downloads and readership data. Citations and downloads have clear differences in their obsolescence characteristics. While it takes several years until articles from the analyzed journal get cited more often, the highest downloads usually happen within the first two years that follow publication. We computed a medium to high correlation among citation, download and readership frequencies. However, a rough analysis of Mendeley users suggests that its user population differs from the one having published (and cited) articles in Scopus. Since this might also be true for the ScienceDirect user community, a perfect relation among these three indicators could not be expected.

As soon as we receive time stamps for the readership data, we will start the obsolescence analyses with them. Since we are aware that the results of our study lack generality due to the small sample, we plan investigations with more journals also from other disciplines (e.g. economics, oncology, linguistics, and history) in the near future.

Acknowledgments

This report is based in part on analysis of anonymous ScienceDirect usage data and/or Scopus citation data provided by Elsevier within the framework of the Elsevier Bibliometric Research Program (EBRP). Readership data were provided by Mendeley. The authors would like to thank both Elsevier and Mendeley for their great support. The Know-Center, which is the affiliation of one co-author, is funded within the Austrian COMET program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth, and the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

Anonymous (2012). Alternative metrics. *Nature Materials,* 11(10), 23 October 2012, 907.

Bar-Ilan, J. (2012). JASIST@mendeley. *ACM Web Science Conference 2012 Workshop*. Retrieved January 23, 2013 from: http://altmetrics.org/altmetrics12/bar-ilan/

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H. & Terliesner, J. (2012). Beyond citations: scholars' visibility on the social web. In *Proceedings of the 17th International Conference on Science and Technology Indicators* (pp. 98-109). Montréal: Science-Metrix and OST.

Bollen, J. & Van de Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1), 136-149.

Hammerton, J., Granitzer, M., Harvey, D., Hristakeva, M. & Jack, K. (2012). On generating large-scale ground truth datasets for the deduplication of bibliographic records. In *International Conference on Web Intelligence, Mining and Semantics 2012* (p. 18). ACM. doi: 10.1145/2254129.2254153

Henning, V. & Reichelt, J. (2012). Mendeley - A Last.fm for research? In *IEEE Fourth International Conference on eScience* (pp. 327–328). IEEE. doi: 10.1109/eScience.

Kraker, P., Körner, C., Jack, K. & Granitzer, M. (2012). Harnessing User Library Statistics for Research Evaluation and Knowledge Domain Visualization. In *Proceedings of the 21st international conference companion on World Wide Web* (WWW 2012 – LSNA'12 Workshop) (pp. 1017-1123). ACM. doi: 10.1145/2187980.2188236.

Li, X., Thelwall, M. & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics*, 91(2), 461-471.

Li, X. & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. In *Proceedings of 17th International Conference on Science and Technology Indicators* (STI 2012), (pp. 541-551). Montréal: Science-Metrix and OST.

Moed, H.F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56(10), 1088-1097.

Schloegl, C. & Gorraiz, J. (2010). Comparison of citation and usage indicators: the case of oncology journals. *Scientometrics*, 82(3), 567-580.

Schloegl, C. & Gorraiz, J. (2011). Global usage versus global citation metrics : The case of pharmacology journals. *Journal of the American Society for Information Science and Technology*, 62(1), 161-170.

DYNAMICS OF SCIENCE AND TECHNOLOGY CATCH-UP BY SELECTED ASIAN ECONOMIES: A COMPOSITE ANALYSIS COMBINING SCIENTIFIC PUBLICATIONS AND PATENTING DATA

Poh-Kam Wong[1], Yuen-Ping Ho[2] and Chan-Yuan Wong[3]

[1] *pohkam@nus.edu.sg*
Entrepreneurship Centre, National University of Singapore, 21 Heng Mui Keng Terrace, Level 5, Singapore 119260

[2] *yuenping@nus.edu.sg*
Entrepreneurship Centre, National University of Singapore, 21 Heng Mui Keng Terrace, Level 5, Singapore 119260

[3] *wongcy111@gmail.com*
Department of Science and Technology Studies, University of Malaya, 50603 Lembah Pantai, Kuala Lumpur, Malaysia

Abstract
The growing competitiveness of Asian economies in science and technology has received increasing attention of science and technology policy researchers. The empirical findings of many studies that unveiled the progress of science and technology in selected Asian economies provide an account of the virtuous cycle of S&T production for these economies. This paper extends these prior analyses in two ways. Firstly, we extended a similar impact study to USPTO patenting data to develop international comparative indicators on national technological output quality. While prior works have tracked the changing share of nations in world total patenting output quantity, we track the growth trend of both the quantity and quality of patenting activities of Asian economies. Secondly, we combine the above measures of quantity and quality of scientific and technological outputs to provide a composite analysis of the temporal dynamics of science and technology catch-up by the East Asian economies. Based on the stylized empirical findings, a dynamic model of phased development involving changing emphasis between science and technology is proposed. The pattern of growth of the 3-NIEs (South Korea, Taiwan and Singapore) appears to be consistent with their industrial development path discussed in literature. Policy implications of our findings for other developing economies and potential extension of the composite analysis approach to a broader range of countries are discussed.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5); and Scientometrics indicators: Relevance to Science and Technology, Social Sciences and Humanities (Topic 1)

Introduction

The growing competitiveness of Asian economies in science and technology has received increasing attention from science and technology policy researchers. Socio-economic development depends on the rate at which new science and technologies are adopted and put into use. In the transition to a knowledge-based economy, the stock of science and technology (both implicit and explicit knowledge) has emerged to be significant. Many Asian emerging economies attempt to raise national investment to develop their science and technological capacity, which can also be the result of virtuous cycle growth in the process of development. Indices on scientific progresses reported by King (2004) and National Science Board (2006) reveal the growing share of selected Asian economies (particularly the four small East Asian NIEs – Korea, Taiwan, Singapore and Hong Kong, as well as the two emerging giants – China and India) not only in total outputs of scientific publications, but also in citation counts of journal articles as well as in the share of the most highly cited journal articles. Patents uptake in the East Asian economies has also experienced remarkable growth in the last 2 decades (Hu and Matthews, 2005). Development models, science and technological learning and national innovation systems of these selected economies are among the major themes or topics of interests in development studies of Asia.

In this paper, we seek to extend these prior analyses to gain insights into the dynamics of Science and Technology catch-up of East Asian economies. In order to cover the different stages of science and technology development, nine Asian economies are selected, namely South Korea, Taiwan, Singapore, Malaysia, Thailand, Indonesia, Philippines, China and India. These economies are selected due to many similarities in catching-up patterns of science and technology (see Lall and Urata, 2003). High-tech services, electronics and semiconductor industrial technologies are particularly essential for advancing the economic growth of these economies. Selected advanced OECD nations, including Japan, are also analysed where appropriate to serve as comparative benchmarks for measuring the catch-up dynamics of the East Asian economies.

We apply similar impact study as used by King (2004) to USPTO patenting data to develop international comparative indicators on national technological output quantity and quality. While prior work (e.g. Khan and Dernis, 2006) has tracked the changing share of nations in world total patenting output quantity, inadequate attention has been paid to tracking changes in the relative quality of patenting among nations. By extending relative quality metrics to patents, we are able to track not only the growth in the volume of technological innovation activities of East Asian economies, but also their dynamics of catching up in terms of quality. In particular, we attempt to address the issue of contention; whether the selected Asian economies emphasize quantity growth first and only pay attention to quality later, or that patent quantity co-evolves with quality.

Review of Literature

Science and Technology Catch-Up in East Asia

The rapid growth of the Asian economies in the last several decades has received extensive treatment in the literature, particularly in the area of development economics. Early studies have attempted to identify the determinants of the East Asian "miracle" (Hughes, 1988; Garnaut, 1989), examining the links between macroeconomic factors and growth in national income. Much of this literature has been concerned with the role of government versus markets in the catching-up process. In the mid-1990s, Krugman (1994) sparked a debate on whether the growth in East Asia economies was driven by growth in total factor productivity (TFP) or largely based on increased inputs. Based on findings by Young (1992) and Kim and Lau (1994), Krugman contended that growth in East Asia was non-sustainable and relied principally on the mobilization of additional resources. The question raised by Krugman focused attention on the question of technological progress or catch-up in the East Asian growth phenomenon. Initially, economists continued in the same vein as the studies cited by Krugman, addressing the question of technological catch-up by estimating growth in TFP. A number of studies in this period estimate the TFP contribution to economic growth in East Asia, with mixed findings (Kawai, 1994; Drysdale and Huang, 1997).

In contrast to the approach of macro-economists, technology-oriented views have focused on exploring how catch-up is achieved through different technological development paths. Traditionally, developing countries were viewed as assimilating and adapting obsolete technologies from advanced countries, consistent with product life cycle theory (Lee et al., 1988). Akamatsu's (1962) Flying Geese model was once a prominent model explaining the economic integration of Asia-Pacific countries. This model was cited to articulate the relocation of manufacturing activities from Japan to first-tier newly industrializing economies (South Korea and Taiwan), then to second-tier NIEs (Malaysia, Thailand and Indonesia), and then to China, India and Vietnam (Kojima, 2000, provides a review). The Flying Geese model portrays Japan as the driving force for economic and technological innovation in the Asia-Pacific region. When wages and other costs in manufacturing rose in Japan, production activities were relocated and technology flowed outward to other Asia-Pacific countries. South Korea, Taiwan and Singapore, the first tier newly industrialized economies (NIEs) have successfully prioritized these pillars in their national strategies and policies, which resulted in remarkable economic development and technological catch-up (Wong, 1999). Subsequent research has suggested that latecomer economies may adopt a stage-skipping catch-up path, consistent with leapfrogging (Perez, 1988) or a path-creating catch-up in which latecomers explore their own path of technological development different from the developed front-runner economies (Lee and Lim, 2001).

Indicators of Science and Technology Catch-Up

To understand the development of science and technology, measures are required to trace the phenomena. This paper adopts the Science and Technology dichotomy established by De Solla Price's distinction between *papyrocentric* science and *papyrophopic* technology (Price, 1965). Science is motivated by peer recognition and visibility within the scientific community, outcomes best achieved through publishing. On the other hand, technology aims to create proprietary products or processes. Hence, the activities in the field of technology lead to patents rather than publications, while science is a publication-directed activity. In many studies, patents are treated as a representation of technology and papers as a representation of science (Meyer, 2002; Wong and Goh, 2012).

However, there are very few studies that have used patents to measure and compare technological catch-up at the country level. Existing work in this area typically focus on specific industries in selected industrializing economies. Among the exceptions are two studies that have examined country-level patents data to draw conclusions about the technological catch-up of East Asian economies. Park and Lee (2006) analyzed US patents data for Taiwan and Korea at the level of the individual technological class, with 376 classes in total. Catch-up in a particular technological class is deemed to occur if the share of the nation's patents in the total patents for the class has increased. The two catching up economies of Taiwan and Korea were found to achieve catch-up in technological classes with shorter cycle time and higher stock of knowledge. Hu and Matthews (2005) applied a framework based on the concept of national innovative capacity to five latecomer economies in East Asia – Taiwan, Korea, Singapore, Hong Kong and China. The outcome measure for innovation output was the number of patents granted to inventors from each of the East Asian economies. The study found some important differences from an earlier study by Furman et al. (2002) using the same framework on 17 OECD countries, suggesting that different strategies are pursued by the latecomer economies to catch up with technological leaders.

While there are studies, albeit few, using patents data to investigate the technological development of latecomer economies, similar work on scientific development are scarce. The use of bibliometric data on scientific publications and citations to evaluate and analyse the outcome of scientific research is well established in the literature (King, 1987; Hicks and Katz 1996), but little has been done using country-level data. An exception is King (2004), who extended on work by May (1997) and applied bibliometrics analysis techniques to measure the quantity and quality of science in 31 selected countries. He established a rank order of nations based on several measures of science citation and noted that there is great disparity between the top ranking countries and the next tiers of nations. The highest ranking non-OECD nation is Israel, in 14th place. Apart from Japan

which placed 4th, the Asian nations rank lowly, with China leading the field in 19th place.

Methodology and Data Sources

Methodology

Following Meyer (2002), Price (1965) and King (2004), we use patents data to represent the technological development of a nation and publications data to characterize its scientific performance. Quantitative indicators of science and technology outputs of a nation are derived from the annual counts of patents and publications generated by that nation. In this study, besides the raw counts of patents and publications, two derived indicators are used – the first measures the number of publications (patents) per million population of a country, while the second measures the relative share of a country's publications (patents) in world total. While the first provides an absolute intensity measure, the second represents a relative measure that takes into account the comparative progress of other countries. To account for the quality of science and technology output of a nation, publication citations data and patent citations data are used respectively to construct appropriate quality measures.

We combine the above measures of scientific and technological output quantity and quality to examine the relative emphasis of nations on science and technology. While prior work have examined the two trends in isolation (for example, King (2004) studied trends in science, Park and Lee (2006) studied trends in technology), our composite analysis allows us to investigate the temporal dynamics of S&T catch-up by the late-comer East Asian economies.

Patents Data

Patents data used in the analysis are utility patents granted annually by the USPTO between 1981 and 2011. Utility patents are used as the reference point as they are also termed "patents for inventions" and are therefore proxies for innovative activities leading to inventions, more so than design and plant patents. Patents are extracted for a number of selected economies and regional groupings, namely 3 East Asian NIEs (Singapore, Taiwan and South Korea), 4 ASEAN economies (Malaysia, Thailand, Indonesia and Philippines) and 2 emerging economies, China and India.

The national affiliation(s) of a patent is determined by the nation(s) of residency of its inventor(s). The convention of "at least one inventor" is adopted. Under this convention, a patent is attributed to an economy if at least one inventor is resident in that economy. A patent may therefore be affiliated to more than one economy if it is co-invented by individuals from different economies. In addition to the number of patents granted to the selected economies, data on the "forward citations" of these patents are also extracted. "Forward Citations" refer to the

number of USPTO patents that refer to the specified patent as "prior arts" of the citing patents. As our database covers the period 1976 to 2011, forward citations for a patent refers to the number citations by subsequent patents granted up to 31th December 2011. Our analysis covered citations received by patents granted in 2006.

Publications Data Source

Data on scientific publications are drawn from the Deluxe version of the National Science indicators (NSI) database developed by Thomson Scientific ®. The NSI database contains counts of the publications and citations taken from around 10,000 peer-reviewed journals indexed by Thomson Scientific. Similar to the approach for extracting patents data, the annual counts for publications and citations are extracted for the selected economies and regional groupings. Annual counts were extracted for the period 1981 to 2011. The country designation of a publication is determined by the address of the publishing author(s). A paper with multiple authors from different countries is equally attributed to all the countries.

One point of departure between the patents and publications databases is the inclusion of Hong Kong in the publications counts for China. In the patents database, data for China are for the People's Republic of China and exclude patents granted to inventions from the Special Administrative Regions of Hong Kong and Macao. In the NSI database, publication counts for China include papers from the mainland as well as Hong Kong and Macao. However, more detailed analysis shows that the bulk of scientific publications from China are from the People's Republic of China, accounting for around 85% to 90% of total papers attributed to China inclusive of the SARs.

The full NSI database contains publications data from the complete range of Thomson Scientific indexed publications, including those in the non-scientific fields of Arts and Humanities and Social Sciences. For this study, we extracted data specifically for publications that appeared in journals that are classified in Science and Engineering related fields. The NSI database categorizes each journal in one of 24 fields or a "multidisciplinary category" using a journal-to-field scheme based on Thomson Scientific's *CC* categories.

Similar to the patents database, the NSI publications database covers papers published up to end of 2011. Citations received by a paper would therefore refer to the number of citations received from subsequent papers published on or before 31th December 2011. Citations made by papers published in 2007 and later are not included in the citations count data.

Findings

Catching Up in Publishing and Patenting Quantity

Table 1 summarizes the average growth rate of publishing and patenting activities by the 3 East Asian NIEs, the four ASEAN economies, China, India and Japan over five-year periods between 1981 and 2011. Scientific publications and patents production of the Asian economies have grown considerably over the decades. In the case of publications, the growth rate for the Asian economies had steadily accelerated from 3.06% p.a. over 1981-85 to nearly 12% p.a. over 2001-05 and about 8% p.a. over 2006-2011. The growth rate for patents increased dramatically from 6.96% p.a. in 1981-85 to over 20% p.a. in 1986-2000, and maintained high double-digit growth over the next 10 years before dipping in 2001-05. Overall, both publications and patenting growth in all three groups of Asian economies were significantly higher than their respective GDP growth rates over the last 20 years.

A clear difference in growth dynamics can be observed between the 3-NIEs and the emerging economies and the 4-ASEAN. The 3-NIEs as a group achieved its peak growth rate in 1991-95 for publications and 1986-90 for patenting, with significant deceleration after the respective peak periods. In contrast, China and India appear to be continuing their growth acceleration for publications right up to 2006-11, although patenting growths rates appear to be reaching a plateau by 2001-05. A similar pattern is observed for the 4-ASEAN economies, although patenting growth within individual economies appears to be more erratic.

Table 2 show the rapid catching-up of selected Asian economies in both publishing and patenting activities, as measured by their growing shares in world totals. From less than 3% of world total publication in 1981-85, the share of late-comers (ex-Japan) increased to over 12% in 2006-11, while the increase of its share in world total patenting was even more dramatic (less than 1% in 1981-85 to 9% in 2006-11).

While China had consistently dominated publications from late-comers, accounting for almost 50% in 2001-05 and 60% in 2006-2011, it is the group of 3-NIEs that dominates patenting from Asian late-comers, accounting for about 80% in 2001-05 and 2006-2011. Although there are positive signs that the growth of publications and patents of the 4-ASEAN are slowly progressing, the production is still quite small and the growth of publications and patents is still weak.

Because of their disparate size in terms of population, **Table 2** may have masked the differences between the group of 3 NIEs and the other of economies. To adjust for this size effect, **Table 3** shows the intensity of publication and patenting output per million of population for the selected of economies. The disparity

between the 3-NIEs and the other two groups is evident – with 1250 publications per million population over 2006-11, the 3-NIEs had almost 13 times the intensity of China and 19 times that of ASEAN4, while the disparity in patenting intensity was even higher (about 100 times). However, it should be noted that the gap has been declining in recent years.

The 3-NIEs have caught-up with Japan in the production of publications and patents. In terms of publication intensity, they have outperformed the production of Japan by 2006-11. In terms of patent intensity, their level is at almost 70% of that of Japan. The gradual growth of other economies could be attributed to the effort in deepening and widening their technological capabilities through assimilation and adaptation of existing S&T knowledge. The effort has recorded positive effects on their publications and patents growth trajectories.

Catching Up in Publishing and Patenting Quality

Table 4 shows the changing average quality level of publishing and patenting activities computed for each of the 5-year periods for the selected economies. While the magnitude of the quality improvement may appear modest, it is important to recognize that this had been achieved in the context of a very rapid increase in quantity.

The group of 3-NIEs shows a consistent pattern of gradual quality improvement for both publishing and patenting over the last 25 years (1985-2006), with a more apparent improvement in more recent years. In contrast, emerging economies, China and India, achieved quality improvement for publications over the 25 years, but not for patenting, which appeared to be declining in quality over the most recent 5 year period. On the other hand, the 4-ASEAN as a group showed a decline in quality of their publications over the last 15 years, but registered an improvement for quality of patents in the same period.

Table 4 also shows the average number of scientific papers cited per patent for selected economies. Almost all the latecomers recorded a significantly high number of science related backward citations per patent. The 2 emerging economies are prolific in science-based patents production with the 3-NIEs at a relative lower range. This pattern of development hints at a co-evolution between science and technology production. The group of ASEAN-4, which is among the most economically advanced of the Southeast Asian countries, was ranked at the bottom in the list of selected economies.

Table 1: Growth Rate in Publications and Patents (compounded in Annual Growth Rate in Period)

	Publication						Patents						GDP Growth Rate					
	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2011	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2011	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2011
Asia Pac	3.06	6.01	8.92	7.46	11.76	7.80	6.96	20.26	16.07	22.77	9.44	10.11	8.43	7.00	7.71	4.82	6.61	4.76
3-NIEs	22.04	22.16	23.60	12.42	10.57	7.25	21.81	37.86	21.21	24.11	7.84	13.73	7.52	9.24	7.52	4.56	4.78	4.63
S'pore	25.84	12.65	24.02	17.23	11.49	6.16	26.98	41.42	35.60	27.10	7.84	13.64	5.55	10.13	9.42	5.96	5.65	6.23
S Korea	25.55	25.24	29.08	18.26	11.10	8.10	26.52	43.66	30.23	22.15	11.70	16.80	8.24	9.40	7.42	3.73	4.74	3.77
Taiwan	18.79	23.72	19.85	5.24	9.49	7.48	20.50	36.21	15.99	25.42	4.97	10.76	6.85	8.68	7.15	5.66	4.58	3.90
Emerging Economies	10.73	9.31	6.82	7.43	14.13	11.01	-3.81	29.74	14.21	30.05	27.33	22.82	8.61	7.10	9.68	6.80	8.65	9.20
China	22.61	18.17	13.01	13.62	19.17	12.57	-9.64	47.63	9.02	27.21	30.90	28.68	12.11	7.62	13.04	8.29	9.90	10.25
India	-1.15	0.45	0.62	1.23	9.09	9.45	2.02	11.85	19.40	32.89	23.76	16.96	5.11	6.58	6.31	5.31	7.40	8.14
4-ASEAN	2.21	7.68	10.22	9.95	12.53	13.41	4.64	0.00	0.00	28.91	15.10	-3.30	2.05	8.38	7.45	0.49	5.36	4.63
Malaysia	0.20	10.13	13.10	8.50	12.21	22.33	41.42	0.00	-8.07	40.87	21.53	16.19	4.65	8.33	9.45	3.53	5.74	4.73
Thailand	5.47	3.79	7.97	11.98	15.88	12.21	-15.91	-15.91	16.36	20.14	6.05	4.97	5.33	11.53	8.64	-0.87	5.80	3.20
Indonesia	9.52	15.62	16.11	10.16	3.37	11.86	0.00	-6.94	13.62	28.78	-4.90	-42.00	1.75	7.90	7.56	-0.87	4.95	5.81
Philippines	-5.88	8.18	4.94	6.24	10.77	7.25	-6.94	25.74	-11.09	14.19	15.90	7.65	-2.42	5.07	2.86	3.46	5.15	4.77
Japan	6.24	5.84	6.24	2.69	1.20	-0.48	11.01	10.26	1.00	7.99	2.46	7.18	3.13	5.27	1.11	0.54	1.57	0.60

Table 2: Share in the World's Publications and Patents (%)

	Publications						Patents					
	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2011	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2011
Asia Pacific	8.415	9.845	11.895	14.655	17.775	18.36%	15.715	21.405	24.49	25.455	27.84	27.62%
3-NIEs	0.3	0.69	1.57	3	4.65	5.17%	0.21	0.73	2.11	4.49	6.26	8.08%
Singapore	0.06	0.11	0.2	0.36	0.61	0.62%	0.01	0.02	0.05	0.13	0.29	0.25%
South Korea	0.08	0.21	0.55	1.39	2.39	2.86%	0.04	0.15	0.79	2.03	2.53	4.29%
Taiwan	0.15	0.36	0.82	1.24	1.64	1.69%	0.16	0.57	1.27	2.33	3.43	3.54%
Emerging Economies	1.685	1.835	2.015	2.485	3.915	5.95%	0.035	0.065	0.09	0.135	0.29	0.75%
China (incl HK)	0.6	1.19	1.71	2.84	5.33	8.93%	0.05	0.11	0.15	0.21	0.47	1.14%
India	2.77	2.48	2.32	2.13	2.5	2.96%	0.02	0.02	0.03	0.06	0.11	0.37%
4-ASEAN	0.21	0.21	0.25	0.34	0.51	0.83%	0.02	0.02	0.03	0.06	0.11	0.12%
Malaysia	0.05	0.05	0.07	0.1	0.14	0.32%	0	0	0.01	0.03	0.06	0.08%
Thailand	0.09	0.09	0.1	0.14	0.25	0.38%	0	0	0.01	0.02	0.02	0.02%
Indonesia	0.02	0.03	0.04	0.05	0.06	0.07%	0	0	0.01	0.01	0.01	0.00%
Philippines	0.05	0.04	0.04	0.05	0.05	0.06%	0.01	0.01	0.01	0.01	0.02	0.02%
Japan	6.22	7.11	8.06	8.83	8.7	6.41%	15.45	20.59	22.26	20.77	21.18	18.67%
N	2,188,812	2,589,316	3,144,518	3,612,976	3,942,141	8,470,755	315,306	412,796	487,805	660,562	776,799	1,249,967

Table 3: Publications per Million Populations and Patents per Million Populations

	Publication per million Population						Patents per million Population					
	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2011	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2011
Asia Pacific	78.27	109.48	164.08	237.20	310.22	523.21	21.27	37.48	51.90	77.19	101.50	128.74
3-NIEs	25.27	65.37	166.54	343.69	569.52	1249.38	2.24	9.45	30.49	84.49	135.12	205.77
Singapore	117.8	242.18	443.48	778.54	1315.45	1879.39	2.54	4.58	15.29	43.95	111.11	113.25
South Korea	11.14	31.35	89.92	247.7	456.6	830.30	0.64	3.03	17.68	59.33	85.09	183.427
Taiwan	42.06	111.49	284.79	471.32	669.17	1038.45	5.59	23.69	60.09	145.4	246.34	320.62
Emerging Economies	11.41	12.74	14.705	18.47	29.98	64.81	0.01	0.035	0.055	0.11	0.375	1.215
China	3.04	6.58	10.45	18.81	38	94.75	0	0.04	0.06	0.1	0.38	1.78
India	19.77	18.89	18.96	18.13	21.96	34.87	0.02	0.03	0.05	0.12	0.37	0.65
4-ASEAN	3.88	4.06	5.21	7.52	11.61	66.43	0.04	0.05	0.1	0.22	0.44	1.91
Malaysia	18.17	19.97	28.24	39.02	56.6	171.25	0.18	0.22	0.71	1.79	4.02	6.69
Thailand	9.55	9.85	12.57	19.64	36.84	81.46	0.03	0.05	0.13	0.37	0.64	0.62
Indonesia	0.74	0.88	1.3	2	2.47	4.28	0.01	0.02	0.03	0.04	0.05	0.02
Philippines	4.3	4.04	4.09	4.93	5.82	8.71	0.11	0.08	0.09	0.18	0.38	0.33
Japan	272.55	355.74	469.86	579.1	629.76	712.22	82.79	140.37	176.95	223.95	270.08	306.07

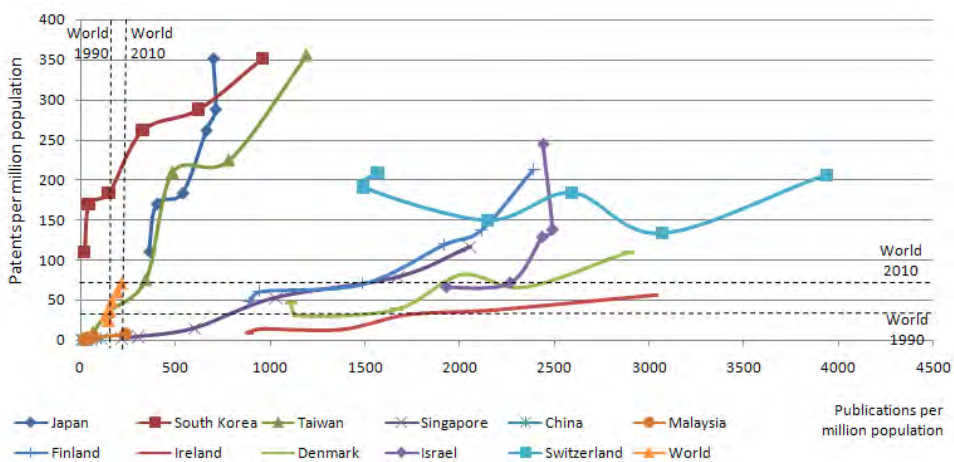
Table 4: Average Citations per Paper and Average Citations per Patent

	Average Citations (within 5 years) per Publication					Average Citations (within 5 years) per Patent					Publications cited (within 5 years) per patent				
Cited Patent Year	1986	1991	1996	2001	2006	1986	1991	1996	2001	2006	1986	1991	1996	2001	2006
Forward Citation period	87-91	92-96	97-01	02-06	06-11	87-91	92-96	97-01	02-06	06-11	na	na	na	na	na
Asia Pacific	3.19	4.10	5.31	7.02	9.81	2.72	3.86	6.17	6.41	4.19	3.15	3.97	5.37	4.22	4.25
3-NIEs	3.15	3.50	5.32	7.76	11.17	2.35	4.06	6.31	6.90	3.12	0.30	0.45	0.65	1.03	1.72
Singapore	2.93	2.12	5.92	7.97	12.33	2.00	4.81	7.60	8.90	3.25	0.33	0.63	0.98	1.86	3.17
South Korea	3.03	4.01	4.84	9.04	11.81	2.69	4.10	5.54	6.32	3.30	0.30	0.50	0.61	0.92	1.39
Taiwan	3.49	4.36	5.21	6.26	9.37	2.37	3.28	5.78	5.47	2.80	0.26	0.23	0.36	0.30	0.60
Emerging Economies	1.93	2.69	3.44	5.81	8.58	2.50	3.90	6.30	5.81	2.90	11.53	13.69	15.34	14.16	12.77
China (incl HK)	2.01	2.93	3.53	6.32	9.31	2.86	3.04	4.36	6.22	3.47	2.86	0.98	3.77	1.52	2.54
India	1.85	2.44	3.34	5.30	7.84	2.13	4.75	8.23	5.39	2.32	20.2	26.4	26.9	26.8	23
4-ASEAN	3.11	4.12	6.28	6.37	8.97	2.42	2.97	6.05	6.92	8.68	0.34	1.08	4.39	0.55	1.02
Malaysia	2.24	3.30	3.95	4.95	6.89	2.50	3.13	6.16	6.92	11.92	0.67	0.22	6.97	0.55	0.81
Thailand	3.10	5.17	7.08	6.50	9.54	-	-	-	-	-	-	-	-	-	-
Indonesia	2.49	3.78	6.56	7.76	9.23	-	-	-	-	-	-	-	-	-	-
Philippines	4.61	4.23	7.53	6.28	10.22	2.33	2.80	5.94	-	5.43	0.00	1.93	1.80	-	1.22
Japan	4.58	6.10	6.21	8.14	10.51	3.61	4.51	6.03	6.00	2.07	0.43	0.65	1.10	1.13	1.50

Growth Dynamics of Science and Technology

While the above analysis has shown that the groups of late-comers' economies achieved relatively fast catching- up in both publications and patenting activities, further insights can be gained by examining the growth pattern of both publications and patenting simultaneously. To do this, it is useful to plot the changing positions of these economies on two dimensions (publications and patenting) over time. **Figure 1** examines the simultaneous growth trajectory of the selected groups of Asian economies in both their changing shares of world publications and patenting simultaneously. To examine the actual positions achieved by the respective Asian economies in recent years, **Figure 1** provides a visualization of their growth trajectories as measured by their level of publishing and patenting per million of populations over time. We showed a period from the beginning (1985) to ending (2010) time points of their trajectories. For comparative analysis, we have also plotted the trajectories for several other OECD countries.

Figure 1: Relationship between Patents per Million Populations and Papers per Million Populations, 1985-2010.

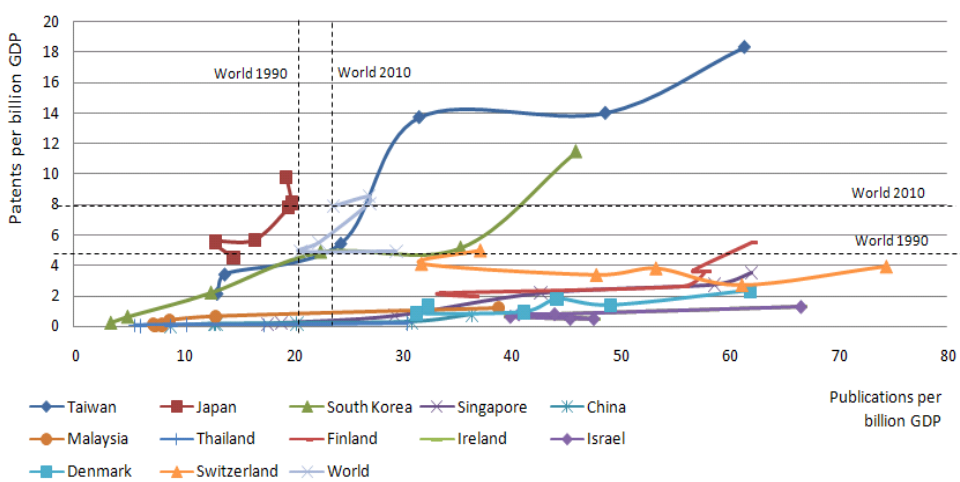


As can be seen, the trajectories of many advanced economies are concentrated at around a 45 degree-angle from the x-axis, suggesting that catching-up in science and technology in respect of economic development had occurred simultaneously. It should also be noted that the trajectories of Japan, South Korea and Taiwan were in the upper left quadrant, indicating that the rate of catching- up had been relatively faster in patenting than in publications. In contrast, the growth trajectories of several European countries (Germany, Switzerland and Sweden) were in the lower right hand quadrant, showing improving share of world publications but declining share of world patenting. Israel, Switzerland, Denmark and Finland appeared to have unchanged publication shares, but rapidly growing

patenting shares. Most of the advanced economies appear to have achieved a level above the world average in publications and patenting. It can be observed that the capacity to generate science and technology knowledge in the developing economies is relatively lower compared to the world standards, suggesting a huge gap to be closed in the catching-up process.

To examine the actual positions achieved with respect to changes in economic performance, **Figure 2** provides a visualization of growth trajectories as measured by each economy's level of publishing and patenting per billion USD (fixed at 2005) of GDP over time. Two salient observations can be made from **Figure 2**. Firstly, the 2 NIEs, South Korea and Taiwan had by 2005 achieved intensities of publishing and patenting approaching and even exceeding the world average and the OECD countries. In particular, Taiwan's economic development appears to have caused scientific knowledge to co-evolve with technological research activities. Secondly, while South Korea and Taiwan had increased their publishing and patenting intensities over the same period, the magnitudes of change in other economies were generally smaller, and they tend to be larger for publications than for patenting. We note that the trajectories for the emerging economies and 4-ASEAN are hardly visible based on the scale of **Figure 2**, as the intensities of these economies were still very much lower than the world average despite their fast catch-up speed.

Figure 2: Relationship between Patents per Billion GDP and Papers per Billion GDP, 1985-2010. (Value fixed at 2005 USD).



Discussion

The Case of 3-NIEs

The 3-NIEs witnessed an emergence of publishing and patenting productivity that can be attributed to significant diffusion of basic and applied R&D. The catching-up process of NIEs appears to starts with technology (patents) capabilities development (Figure 1), followed by technology co-evolving with science (publications). This occurs when the allocation of resources for science increases to support new technology (Wong and Goh, 2012).

We believe that the changing pattern of emphasis on science with technology output over time can be explained in terms of the underlying industrial development strategies of these three economies which have been widely documented elsewhere (e.g. Wong and Singh, 2007 for Singapore; Kim, 1997 for Korea; and Belaguer et. al. 2007 for Taiwan). In particular, we believe that the phase of strong growth in technological outputs that we observed among the Asian NIEs in the period early-1980s to late 1990s coincided with the period of relatively strong focus of the respective governments to promote applied R&D, engineering development and incremental technological innovation as the driver for enhancing the technological competitiveness of their firms during that period. Prior to this phase, R&D and patenting was not that important, as firms could enhance their capabilities primarily by learning to imitate, adapt and exploit technologies transferred from advanced countries (Kim, 1997 and Wong and Singh, 2007). To move beyond this phase of technology imitation/adoption, however, the NIEs needed to start investing in R&D and IP creation to give their firms a competitive edge. While this new push entailed high growth in R&D investment and patent creation, the nature of the R&D was more applied, and the patenting covered primarily incremental innovations. As such, during this phase, while scientific research capabilities were encouraged, they were given lower priorities as they were deemed less critical and relevant to industries than engineering development capabilities. By the late 1990s, however, the 3-NIEs had developed to a point where opportunities for incremental innovation were becoming less available, and the respective governments began to realize the need to invest in more fundamental scientific research capabilities. As the investment in more fundamental scientific research gained momentum in the early 2000s in all three NIEs, the transition towards greater growth of scientific outputs would begin, over-taking the earlier phase's emphasis on incremental innovation output growth.

Besides explaining the observed shift in the science and technology output catch-up rate over time, we believe the industrial development strategies of the 3-NIEs are also consistent with our findings on the shift in quality of science and technology outputs over time. Basically, in the initial stage of the applied R&D/ incremental innovation phase, patenting quality growth was not likely to improve

much or even drop, as the emphasis then was more on quantity. Over time, however, patent quality would be expected to incrementally improve as IP became more important to firms' competitiveness, but substantial improvement in patent quality is likely to occur only when these economies start to invest substantially in more fundamental scientific R&D. Likewise, while quality of scientific publications would be expected to improve gradually as the local universities became upgraded over time, it is also expected to improve more substantially when these economies transit from the incremental innovation phase to the phase of heavy investment in scientific R&D. Moreover, we expect the jump in improvement in publications quality to precede that of patent quality, due to time lag from science to technology. The observed trends in publication and patenting quality improvement (**Table 3**) appear to be consistent with these predictions.

All the three NIEs have achieved phenomenal growth in publishing and patenting activities. However, the quantitative projection also implies a few potential risks that might lead to structural systemic failure. Our observations include:

(i) South Korea's publishing intensity consistently lags behind the other NIEs. The application-oriented science policy of South Korea which was mandated to develop science-based technologies appears to have discouraged the agents of the innovation system to pursue fundamental research activities.[60]

(ii) We found a pattern of structural change in publishing and patenting activities of 3-NIEs. South Korea and Singapore achieved significant progress in production of science-based patents. However, Taiwan, the leading patent producer, is lagging significantly behind in science-based patents. The formation of scientific systems of Taiwan seems have yet incorporated the new technologies that concord with the next wave of innovations.

(iii) The 3-NIEs achieved critical mass for science and technology development. The papers (and patents) per million of population of NIEs indicates their capabilities to develop new science and technology. However, the expansion of economic activities of Singapore has not been co-evolving with the patent production (see Figure 2). This suggests mismatch of scientific development with the progress of technology despite their strong basic research efforts. This could be due to the nature of Singapore's economic structure (that has been hinged on large-scale logistics activities, finance, airlines and other services); the indigenous firms and their technological competencies for economic development are not comparable to those from South Korea and Taiwan.

[60] This is somewhat reflected in Choung and Hwang (2013).

The Case of 4-ASEAN and the Emerging Economies

There is some sense that the science, technology and innovation (STI) policies of ASEAN-4 and the 2 emerging economies, were planned and organized based on the linear model of innovation approach, which assumes basic research activities from universities and research institutions as the core sources of the innovation process. The production trajectory of papers seems has yet to co-evolve with the production of patents (**Table 3**). These economies may face the risk of falling behind due to the limited co-propagating behaviour between science and technology.

Trajectories of Catching Up in Asia: Convergence or Divergence

Our findings refute the suggestion that there is a single converged trajectory for S&T catch up in Asia, as proposed by the Flying Geese model. The Flying Geese model seems to assume that there is a simple linear relationship between the leading economy (Japan) and the followers. However, this assumption fails to foresee the possible existence of impediments to late-comer upgrading. This is particularly evident in many Southeast Asian economies during the 1990s, when they pursued a low-wage policy rather than technological upgrading to counter the emerging threats of China and Vietnam (Wong, 2011). In our view, not all technological trajectories can converge to that of Japan. The 4-ASEAN did not experience the same dynamics as that of South Korea and Taiwan.

The Flying Geese model also assumes that there is a stable hierarchy in development process. In fact, Japan's position at the head of the hierarchy is no longer guaranteed. Many latecomers outperformed Japan in the production of science and technology. The three NIEs outperformed Japan in total publications per million populations. Taiwan outperformed Japan in total patents per million populations. And China and India outperformed Japan and the NIEs in science-based patents production. India attained remarkable performance in science-based patent production, outperforming the average performance of the advanced economies. This could be attributed to the efforts of the Indian government in advancing the capabilities of the public research institutions for development of high-tech service industries.

Conclusion: A Dynamic Catch-Up Model for Late-Industrializing Economies?

In summary, using publishing and patenting data as proxy measures for science and technology outputs, the empirical analysis of this paper provides new findings on the dynamics of the science and technology catch-up process by the fast growing, late-industrializing Asian economies. In particular, we found that catch-up generally occurs simultaneously in both science and technological outputs, but with a stronger emphasis on the latter for at least a certain period of time. Overall, while some individual economies experienced a drop in the quality of science and technology outputs in the process of rapid output growth, Asian late-comers as a

whole actually achieved an improvement in average quality of science and technology outputs while catching up rapidly in output intensity and share of world total.

We believe that the above stylized empirical findings appear novel and it would be interesting for future research to test if a similar pattern can be observed in the science and technology catch-up process by other similar economies. In addition, we believe that the above stylized empirical observations may be suggestive of a more fundamental process dynamics in the science and technology catch-up process of late-industrializing economies. In particular, synthesizing from the empirical observations presented in this paper, and re-examining the growth rates data given in **Table 1** in more detail, we suggest that there may be three successive waves of Asian late-comers' catching-up, with the 3-NIEs constituting the first wave, China and India the second, and the 4-ASEAN the third. Although we only appear to have complete observation for the first wave, we can conjecture that each wave seems to go through a phase of relatively strong emphasis on patenting, which would achieve growth rate surpassing that of publications for a while, before peaking and transiting to a new phase where emphasis would swing to publication outputs.

The 3-NIEs as a group appear to fit this dynamic catch-up model. The three economies had already entered the phase of higher growth rates in patenting than publications by the early 1980s and continued in this phase all the way through the 1990s, and it was only in the first 5 years of the new millennium that growth in patenting had dropped below that of publications, thus transiting to a new phase where emphasis is increasingly shifting to science. The second wave involves China and India, where patenting growth started to exceed publication growth several years later than was the case with the NIEs, and this higher growth rate in patenting had continued right through the first 11 years of the new millennium. It is an open question when the transition to higher emphasis on science would begin. The third wave comprising the 4-ASEAN economies started the same trajectory even later, with patenting growth overtaking publication growth only in the second half of the 1990s and continuing into the first five years of the new millennium.

While admittedly preliminary, the above discussion suggests the potential contribution of this paper's empirical findings towards the development of a more theoretically-based model of the science and technology catching-up process among late-industrializing economies. To the extent that the rapid catch-up process of the Asian NIEs can be shown to follow a well-defined dynamic model that is theoretically linked to the underlying industrial development strategies of these economies, we believe that concrete policy implications can be drawn for other late-industrializing economies based on such a model. For example, policy makers in a developing country may use the time sequencing of prioritization of

investment in science vs. technology by the Asian NIEs as benchmarks for their own S&T investment strategy plan.

Another contribution of this paper is to show the usefulness of the composite analysis approach that combines indicators of publication and patenting quantity and quality indices to discover salient features of the dynamics of science and technology development over time. Although we have focused our attention on the Asian economies in this paper, we believe that our composite analysis method can be used to analyze a broader range of economies, both advanced and emerging, to examine not only the process of science and technology catch-up of the latecomers, but also the process whereby the incumbent leaders seek to sustain their science and technology leadership.

References

Akamatsu K.(1962): A historical pattern of economic growth in developing countries. *Journal of Developing Economies*, 1, 3–25

Balaguer, A. et. al. (2007). The Rise and Growth of a Policy-Driven Economy: Taiwan. In C. Edquist and L. Hommen (Eds.), *Small Economy Innovation Systems: Comparing Globalization, change and Policy in Asia and Europe*. Cheltenham: Edward Elgar

Choung, J-Y. and Hwang, H-R. (2013), The Evolutionary Patterns of Knowledge Production in Korea, Scientometrics, 94, 2, 629-650.

Drysdale, P. and Huang, Y.P. (1997). Technological catch-up and Economic Growth in East Asia and the Pacific. *Economic Record*, 73 (222), 201-11.

Furman, J.L., Porter, M.E. and Stern, S. (2002). The determinants of national innovative capacity. *Research Policy,* 31, 899–933.

Garnaut, R. (1989). *Australia and the Northeast Asian Ascendancy*. Canberra: Australian Government Publishing Service.

Hicks, D.M. and Katz, J.S. (1996). Where is Science going? *Science, Technology & Human Values*, 21(4), 379-406.

Hu, M.C. and Matthews, J.A. (2005). National innovative capacity in East Asia. *Research Policy*, 34, 1322-1349.

Hughes, H. (1988). Achieving Industrialisation in East Asia. Cambridge: Cambridge University Press

Kawai, H. (1994). International comparative analysis of economic growth: trade liberalization and productivity. *Developing Economies*, 32(4), 373-97.

Khan, M. and Dernis, H. (2006). Global Overview of Innovative Activities from the Patent Indicators Perspective. OECD Science, Technology and Industry Working Papers 2006/3, OECD Directorate for Science, Technology and Industry.

Kim, J.I. and Lau, L.J. (1994). The Sources of Growth of the East Asian Newly Industrialised Countries. *Journal of the Japanese and International Economies*, 8(3), 235-71.

Kim, L.S. (1997). *Imitation to Innovation: The Dynamics of Korea's Technological Learning*. Massachusetts: Harvard University Press.
King, D.A. (2004) The Scientific Impact of Nations. *Nature*, 430, 311-316.
King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, 13, 261-276 .
Kojima, K. (2000). The "flying geese" model of Asian economic development: origin, theoretical extensions and regional policy implications. *Journal of Asian Economics*, 11, 375-401.
Krugman, P. (1994). The Myth of Asia's Miracle. *Foreign Affairs*, 6(73), 62-78
Lall, S. and Urata, S. (Ed.) (2003). *Competitiveness, FDI and Technological Activity in East Asia*. Cheltenham: Edward Elgar.
Lee, J., Bae, Z. and Choi, D. (1988). Technology development process: a model for a developing country with a global perspective. *R&D Management*, 18(3), 235-250.
Lee, K. and Lim, C.S. (2001). Technological regimes, catching-up and leapfrogging: findings from the Korea industries. *Research Policy*, 30, 459-83.
May, R.M. (1997). The Scientific Wealth of Nations. *Science*, 275(5301), 793-96.
Meyer, M. (2002). Tracing knowledge flows in innovation systems. Scientometrics, 54(2), 193-212.
National Science Board (2006). *Science and Engineering Indicators 2006. Two volumes*. Arlington, VA: National Science Foundation (volume 1, NSB 06-01; volume 2, NSB 06-01A).
Park, K.H. and Lee, K. (2006). Linking the technological regime to the technological catch-up: analyzing Korea and Taiwan using the US patent data. *Industrial and Corporate Change*,15(4), 715-53.
Perez, C. (1988). New technologies and development. In: Freeman, C., Lundvall, B. (Eds.) *Small Countries Facing the Technological Revolution*. London: Pinter Publishers.
Price, D.S. (1965). Is technology historically independent of science? A study in statistical historiography. *Technology and Culture*, 6, 553-568.
Wong C-Y and Goh K-L (2012). The Pathways for Development: Science and Technology of NIEs and Selected Emerging Countries. *Scientometrics*, 92, 3, 523-548.
Wong, C-Y. (2011). Rent-Seeking, Industrial Policies and National Innovation Systems in Southeast Asian Economies. *Technology in Society*, 33(3/4), 231-243.
Wong, P. K. (1999). National Innovation Systems for Rapid Technological Catch-up: An Analytical Framework and a Comparative Analysis of Korea, Taiwan and Singapore. Paper Presented at DRUID National Innovation System, Industrial Dynamics and Innovation Policy Conference, Rebild, 9-12 June.
Wong, P.K. and A. Singh. (2007). From Technology Adopter to Innovator: The Dynamics of Change in the National System of Innovation in Singapore. in C. Edquist and L. Hommen (eds.). *Small Economy Innovation Systems:*

Comparing Globalization, change and Policy in Asia and Europe. Cheltenham: Edward Elgar

Young, A. (1992). Lessons from the East Asian NICs: A Contrarian View. *European Economic Review*, 38 (3&4), 964-73.

THE EFFECT OF BOOMING COUNTRIES ON CHANGES IN THE RELATIVE SPECIALIZATION INDEX (RSI) ON COUNTRY LEVEL

Dag W. Aksnes,[1] Thed N. van Leeuwen,[2] and Gunnar Sivertsen[3]

[1] *dag.w.aksnes@nifu.no*
NIFU Nordic Institute for Studies in Innovation, Research and Education, PO Box 5183, N-0302 Oslo, Norway

[2] *leeuwen@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, PO Box 9555, 2300 RB Leiden, The Netherlands

[3] *gunnar.sivertsen@nifu.no*
NIFU Nordic Institute for Studies in Innovation, Research and Education, PO Box 5183, N-0302 Oslo, Norway

Abstract
The Relative Specialization Index (RSI) is an indicator that measures the research profile of a country by comparing the share of a given field in the publications of a given country with the share of the same field in the world total of publications. If measured over time, this indicator may be influenced in the world total by the increased representation of certain other countries with different research profiles. As a case, we study the effect on the RSI for the Netherlands of the increased representation of China in the ISI Web of Science. Although the booming of China is visible in the RSI for the Netherlands, especially in the last decade and in fields where the countries have opposite specializations, the basic research profile as measured by the RSI remains the same. We conclude that the indicator is robust with regard to booming countries, and that it may suffice to observe the general changes in the research profile of the database if the RSI for a country is studied over time.

Conference Topic
Topic 1: Scientometrics Indicators: - Criticism and new developments.

Introduction
As is well known from several studies and reports presenting bibliometric indicators on country level, the specialization of a country in a specific science field can be calculated on the basis of publication counts by comparing the relative share of that field in the target country against the relative share of the same field in the world total of publications, the so-called *Relative Specialization Index* (RSI).
The RSI is a further development of the *Activity Index* (AI), which was first introduced by Frame (1977) and further developed by Schubert & Braun (1986)

and by Schubert, Glänzel & Braun (1989) for systematic comparison of countries. The definition of the Activity Index is:

$$\text{AI} = \frac{\text{the world share of the given country in publications in the given field}}{\text{the overall world share of the given country in publications}}$$

or, equivalently,

$$\text{AI} = \frac{\text{the share of the given field in the publications of the given country}}{\text{the share of the given field in the world total of publications}}$$

As introduced in the *Second European Report on S&T Indicators* (1997), see also Glänzel (2000), the Relative Specialization Index (RSI) is then defined as:

$$\text{RSI} = \frac{\text{AI} - 1}{\text{AI} + 1}$$

The position of a country in a specific field is thus benchmarked against the world standard case, where RSI=0. Fields in a country where RSI>0 indicate a relative specialization (at least in terms of production) in that particular field. Note that the overall score for a country should always be 0. This means that positive RSI-values must always be balanced by negative ones. Hence a country cannot have only positive or only negative values.

It has been noted that the RSI may not be statistically reliable when a country only contributes with a small number of publications (Schubert, Glänzel & Braun 1989). Furthermore, there are theoretical problems with the indicator when it comes to extreme values since it is built up from ratios (Rousseau & Yang 2012). We do not regard the examples that we present in the following as affected by these statistical and theoretical problems.

Instead, we investigate another possible problem with the RSI when used for studying developments over time. The specialization of a country will then be benchmarked at different intervals against what we just mentioned as "the world standard case". It follows from the calculation of the RSI that the focus for the interpretation of a possible change in the indicator will be on a given field in a specific country. However, there may be changes in the specialisation profile of the database itself that will influence the calculation. Such changes may result from increasing journal coverage in specific fields, from the growth or decline of specific research areas, or from changes in the representation of the scientific output from certain countries.

In the study we have analysed the specific effects of one country: China. This country has been selected because it by far is the largest booming country. As a contrasting case, we are using the Netherlands because the country is relatively small (would expose effects of the booming country on the RSI) and has a more

stable growth over time. The two countries not only have different growth rates within the database, but different specialization profiles as well.
While the specialization profile of the Netherlands is balanced, but with and orientation towards the Life Sciences, including clinical medicine, the specialization profile of China is more dominated by the Physical Sciences and Engineering Sciences. The profile of the Netherlands resembles those found in other Western European countries and in the USA. All together, these countries contribute to the majority of publications in the database and therefore determine the specialization profile of the database with the largest weight. China, on the other hand, shares its profile with a few other countries, such as South Korea, Taiwan and Singapore (Zhou et al. 2012), which are smaller, but also booming countries. On the other hand, China's profile is distinct from those of some larger booming countries, the so-called BRIC countries (Yang et al. 2012). We therefore found it useful to select only one booming country when studying the effect of its rapid growth over time on another country's RSI.
We recently became aware of this issue when revising the official R&D&I indicator report for the Netherlands (den Hertog et al. 2012): In certain fields, changes in the RSI for the Netherlands since 1981 was partly influenced by the rapidly increasing representation of China in the database, especially in last decade. In this paper, we will discuss the methodological implications of an observation of this type.

Data and methods

The two countries differ widely with regard to growth rates within the database. As seen in *Table 1*, the world share of the Netherlands has been relatively stable with a slight increase after and 2001, which is typical of most Western European countries. China, on the other hand, has had a very rapid increase, especially in the last decade, in which China represented a growth dynamic of its own with an exponential increase in publications in the *Science Citation Index* (Leydesdorff & Zhou 2005). In 2006, at a time when China already was the sixth largest country in terms of scientific production, a continuation of the same growth patterns was expected in the near future (Zhou & Leydesdorff 2006). By 2011, China was the second largest country after the USA with 12.5 per cent of the world's publications.

Table 1. Total number of publications per year and proportion of world total, the Netherlands and China. 1981, 1991, 2001 and 2011.

	The Netherlands		*China*	
Year	*Numb of pub*	*Prop world*	*Numb of pub*	*Prop world*
1981	7345	1.6%	1588	0.3%
1991	13295	2.2%	8696	1.4%
2001	19894	2.5%	34276	4.3%
2011	32975	2.6%	157545	12.5%

Our study is based upon publication data from 1981-2011 as retrieved from the proprietary version of the database Thomson Reuters Web of Science (WoS) at the Centre for Science and Technology Studies (CWTS) at Leiden University. For the analysis of specific fields of research, each source journal within the CWTS/WoS database is attributed to one or more Journal Subject Categories (JSC) defined by Thomson Reuters. In this study we have applied the disciplinary grouping of the JSCs into about 40 main fields of science. Wide-scope journals are often assigned to more than one subfield. The prestigious general journals with broad multidisciplinary scopes, such as Nature and Science, are assigned to a journal category of their own, denoted as 'Multidisciplinary Sciences' and included in the CWTS system under the heading 'Multidisciplinary journals'. A list of all main fields are given in the *Appendix* with the size of each field shown as a percentage of all publications in the database in the four years 1981, 1991, 2001 and 2011. Observe that the Physical Sciences and Engineering Sciences both have increasing shares in the total, while the Life Sciences have decreasing shares. In order to represent all of these main areas with fields of different size, we selected the following eight fields for further analysis:

- Civil engineering and construction
- Computer sciences
- Earth sciences and technology
- Environmental sciences and technology
- Chemistry and chemical engineering
- Physics and materials science
- Clinical medicine
- Health sciences

We selected four years at the beginning of four decades to be studied: 1981, 1991, 2001 and 2011. All calculations and statistics refer to database years – i.e. the year in which Thomson Reuters processed the publications for the WoS database. These measurements differ from those based on publication years, which refer to the publishing date of the journal issue. This issue is, however, not likely to influence on the overall results of the study. Usually there is only a minor indexing delay and most articles have identical publication and indexing years.
Only publications reporting on original research findings are included – i.e. the document types 'normal article', 'letter', and 'review article'. 'Meeting abstracts', 'Corrections', 'Editorials' and other document types are not included. Each publication is attributed by whole counts to each country listed in the author address list of the publication.

Results

Table 2 shows the proportion of the publications by selected disciplines for the Netherlands, China and the world. For the Netherlands there has been a significant decline in the proportion of publications in the two fields Chemistry

and chemical engineering and Physics and material science. In 1981, Chemistry and chemical engineering accounted for 11.5 per cent of the Dutch publications. In 2011, this proportion was only 5.7 per cent. The corresponding figures for Physics and material science are 11.2 and 6.6 per cent, respectively. On the other hand, there is a significant growth for Clinical medicine, from 19.1 per cent in 1981 to 25.8 per cent in 2011. The other disciplines analysed are significantly smaller in terms of publication volume. There is a very strong relative growth for the Health sciences, where the proportion has increased from 0.4 per cent in 1981 to 3.1 per cent in 2011. We also see a strong growth for the Earth sciences and technology and Environmental sciences & technology.

The changes in the publication profile of the Netherlands deviate considerably from the changes in the world average. In fact, Clinical Medicine accounts for a decreasing proportion of the database (20.7 per cent in 1981 to 18.3 per cent in 2011). Moreover, there are relatively small differences for most of the other selected disciplines. In the two fields Environmental sciences & technology and Computer sciences the proportions have increased from 2.3 to 3.7 per cent and from 0.9 to 2.2 per cent, respectively (1981 and 2011 figures).

Table 2. Proportion of publications in selected disciplines per year (1981, 1991, 2001 and 2011) the Netherlands, China and the world, percentage

	The Netherlands				*China*				*World*			
Discipline	*1981*	*1991*	*2001*	*2011*	*1981*	*1991*	*2001*	*2011*	*1981*	*1991*	*2001*	*2011*
Chemistry and chemical eng	11.5	8.8	8.1	5.7	7.8	16.3	22.6	19.4	10.5	10.3	10.9	10.9
Civil engineering and construction	0.3	0.3	0.3	0.4	0.2	0.9	0.8	0.9	0.4	0.5	0.4	0.7
Clinical medicine	19.1	22.3	24.0	25.8	10.6	6.8	6.3	8.1	20.7	20.3	19.5	18.3
Computer sciences	1.3	2.2	2.4	1.9	0.6	1.7	2.4	3.0	0.9	1.7	2.2	2.2
Earth sciences and technology	1.7	2.1	2.8	2.9	11.0	3.5	3.1	2.8	2.2	2.4	2.7	2.7
Environmental sciences & techn	2.5	2.6	3.2	4.2	0.6	1.6	2.1	3.5	2.3	2.4	2.9	3.7
Health sciences	0.4	0.8	1.8	3.1	0.0	0.2	0.5	0.4	1.1	1.2	1.8	2.1
Physics and materials science	11.2	10.2	9.3	6.6	19.4	27.1	25.7	21.0	10.2	11.7	12.6	11.9
N*	8953	16984	26719	45533	1580	9212	42992	210740	544965	751785	1068572	1715531

*) Numbers include double counting of articles that have been assigned to more than one discipline.

China has a scientific profile deviating considerably from the world average. Moreover, there have been major changes in the relative weight of the different disciplines during the 30 year period. In 1981, 7.8 per cent of the Chinese publications were in Chemistry and chemical engineering compared to 19.4 in 2011. This discipline accounts for approximately twice as large proportion as the world average (10.9 per cent in 2011). The latter finding also extends to Physics and materials science. Clinical medicine is a relatively small field in China compared to many other countries with a proportion of 8.1 per cent in 2011.

A question arising in this context is to what extent the tremendous increase in publication number by China can be ascribed to increased database coverage of Chinese journals. Even though the number of Chinese language journals has increased, they still account for a very small proportion of the Chinese output (almost 30 such journals in recent years). Thus, the boosting publication output of China is a "real" phenomenon and methodological factors related to coverage of Chinese journals have marginal importance only.

We will now investigate how the scientific specialization profile of the Netherlands has changed between the years 1981, 1991, 2001 and 2011. In order to illustrate possible influences of the booming of China on the RSI for the Netherlands in these fields, we calculate the RSI both with and without inclusion of China in the world total. The results can be observed in **figures 1a-d**.

There are significant variations in the development of the disciplines, but for most fields the differences between the two indicators are not very large. Both Physics and material science and Chemistry and chemical engineering show a de-specialization during the period. This is related to the fact described above, with significant decreasing national proportions during the time period. The RSI indicator where China has been removed gives slightly higher values for 2011 and accordingly a less strong reduction, but the differences are not large. The pattern for Civil engineering and construction is similar to the latter two fields. This also holds for Computer sciences, but here the RSI shows no uniform trend.
An opposite picture is found for Clinical medicine and Health sciences. Here, the RSI has significantly increased during the period, particularly for Health sciences. However, the latter is a rather small field in terms of publication volume. The RSI indicator with removal of China gives somewhat lower values for 2011. But again, the differences are minor.
In Environmental sciences and technology the RSI indicator is almost stable during the entire period. Interestingly, here the two indicators provide almost identical results. This also holds for Earth sciences and technology, but here the RSI values are increasing from 1981 to 2011.

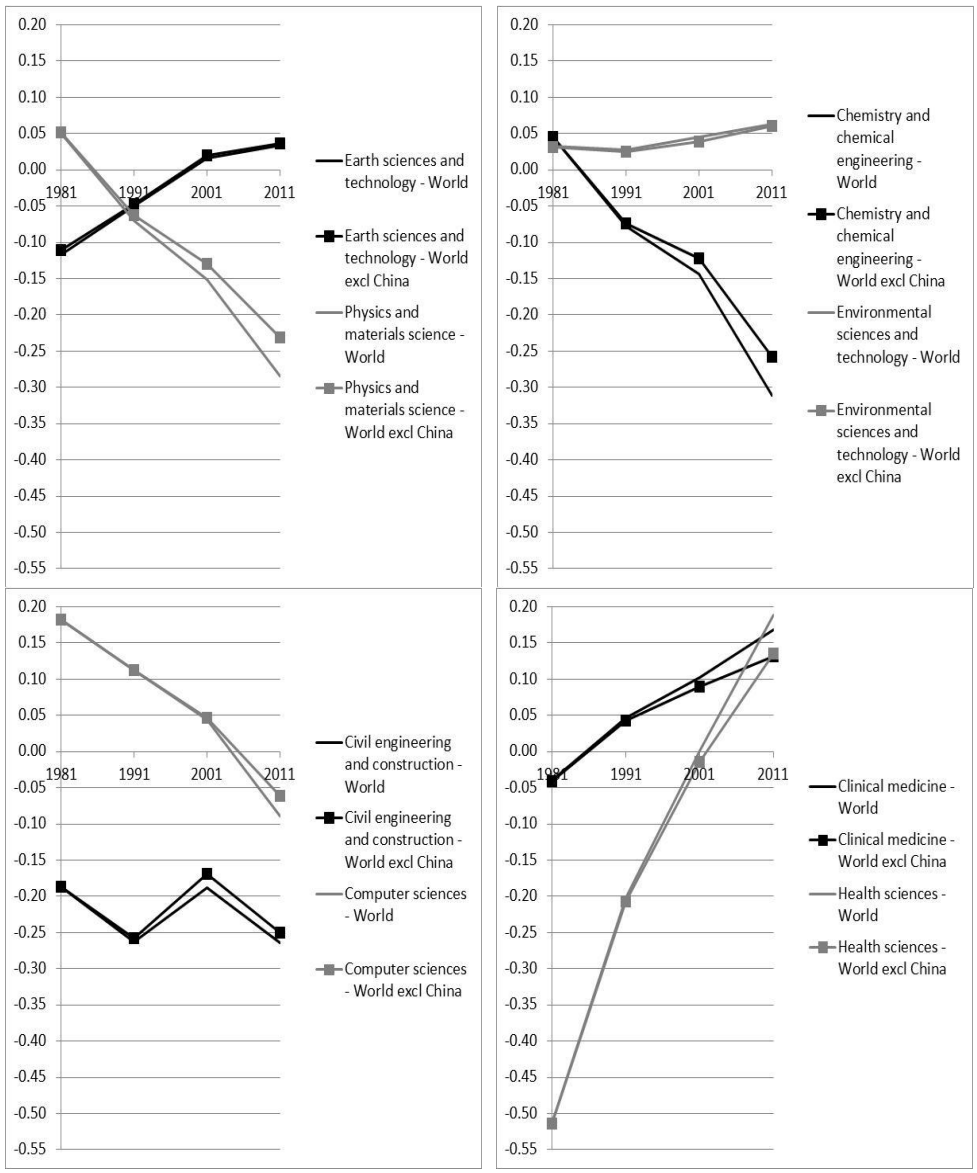


Figure 1a-d. RSI with and without including China in the world total. Selected disciplines for the Netherlands, 1981, 1991, 2001 and 2011.

Discussion & conclusions

The structure of the global science system has changed in the last decade. China and other booming countries now account for a significant proportion of the worlds' scientific efforts. In this paper we have investigated how China influences on the measurement of the specialisation profile of other countries. China has a scientific specialisation which deviates considerably from the world

average, with a strong emphasis on research in engineering and the physical sciences. China's proportion of the Web of Science database has increased from 4.3 per cent in 1981 to 12.5 per cent in 2011.

As expected, our analyses have shown that removing China hardly has any impact on the RSI values for 1981 and 1991. Moreover, the influence on 2001 figures is also very small. For the 2011 figures, we observe larger differences, but also here the differences are relatively modest. In the disciplines investigated, the RSI values change with up to 0.07 points. Basically, the research profile as measured by the RSI remains the same.

The differences between the two measures are largest in the fields where China differs significantly from the world average profile. China has a very high proportion of the world publications in fields such as Physics and material science and Chemistry and chemical engineering. Removing China means that other countries will obtain higher RSI values in these fields. In other fields where China has relatively few publications compared to other countries, such as Clinical medicine and Health sciences, the RSI will be lower.

The study has focused on one booming country only. The effect of other such countries will be analysed in forthcoming studies. As noted in the introduction, the scientific specialisation of these countries differs, and including other countries might also blur effects because they have different and neutralizing specializations.

We conclude that the indicator can be regarded as robust with respect to the main booming country, China, and that it may suffice to observe the general changes in the research profile of the database if the RSI for a country is studied over time (see the appendix table).

References

Frame, J. D. (1977). Mainstream research in Latin America and the Caribbean. *Interciencia*, 2, 143–148.

Glänzel, W. (2000). Science in Scandinavia: A bibliometric approach. *Scientometrics*, 48 (2), 121–150.

den Hertog, P., Jager, C-J., te Velde, R., Veldkamp, J., Aksnes, D.W., Sivertsen, G., van Leeuwen, T. & van Wijk, E. (2012). *Science, Technology & Innovation Indicators 2012*. Dialogic. Utrecht. Retrieved January 21 2013 from: http://www.dialogic.nl/documents/2010.056-1235.pdf

Leydesdorff, L. & Zhou, P. (2005). Are the contributions of China and Korea upsetting the world system of science? *Scientometrics*, 63 (3), 617-630.

Rousseau, R. & Yang, L. (2012). Reflections on the activity index and related indicators. *Journal of Informetrics*, 6, 413– 421.

Schubert, A. & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, 9, 281–291.

Schubert, A., Glänzel, W. & Braun, T. (1989). World flash on basic research: Scientometric datafiles. A comprehensive set of indicators on 2649 journals

and 96 countries in all major science fields and subfields, 1981-85. *Scientometrics*, 16 (1-6), 3-478.
Second European Report on S&T Indicators 1997. Appendix. (1997). EUR 17639. European Commission. Brussels.
Zhou, P. & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. *Research Policy,* 35 (1), 83-104.
Zhou, Q., Rousseau, R., Yang, L., Yue, T. & Yang, G. (2012). A general framework for describing diversity within systems and similarity between systems with applications in informetrics. *Scientometrics*, 93, 787-812.
Yang, L.Y., Yue, T., Ding, J.L., & Han, T.. (2012). A comparison of disciplinary structure in science between the G7 and the BRIC countries by bibliometric methods. *Scientometrics*, 93, 497-516.

Appendix table. Proportion of publications by discipline. World total.

Main area	*Discipline*	*1981*	*1991*	*2001*	*2011*
Natural sciences & mathematics	Astronomy and astrophysics	1.1%	1.1%	1.2%	1.0%
	Chemistry and chemical eng	10.5%	10.3%	10.9%	10.9%
	Earth sciences and technology	2.2%	2.4%	2.7%	2.7%
	Environmental sciences and techn	2.3%	2.4%	2.9%	3.7%
	Mathematics	2.0%	2.1%	2.3%	2.5%
	Physics and materials science	10.2%	11.7%	12.6%	11.9%
	Statistical sciences	0.7%	0.8%	0.8%	0.9%
	Total	29.0%	30.8%	33.4%	33.5%
Life sciences	Agriculture and food science	3.4%	2.7%	2.4%	2.8%
	Basic life sciences	7.8%	8.8%	9.0%	7.8%
	Basic medical sciences	0.7%	0.8%	1.0%	1.5%
	Biological sciences	5.4%	4.8%	4.2%	4.9%
	Biomedical sciences	9.6%	10.1%	9.5%	8.4%
	Clinical medicine	20.7%	20.3%	19.5%	18.3%
	Health sciences	1.1%	1.2%	1.8%	2.1%
	Total	48.7%	48.9%	47.5%	45.8%
Engineering	Civil engineering and construction	0.4%	0.5%	0.4%	0.7%
	Computer sciences	0.9%	1.7%	2.2%	2.2%
	Electrical engineering and telecommunication	2.4%	2.4%	2.7%	3.2%
	Energy science and technology	1.5%	1.5%	1.3%	1.6%
	General and industrial engineering	0.6%	0.7%	0.8%	0.9%
	Instruments and instrumentation	0.8%	0.9%	0.8%	0.8%
	Mechanical engineering and aerospace	1.7%	1.7%	2.0%	1.9%
	Total	8.3%	9.3%	10.3%	11.2%
Social sciences	Economics and business	1.4%	1.2%	1.0%	1.3%
	Educational sciences	1.2%	0.8%	0.6%	0.8%
	Information and communication sciences	0.4%	0.3%	0.3%	0.3%
	Management and planning	0.5%	0.4%	0.4%	0.6%
	Political science and public adm	0.7%	0.5%	0.4%	0.4%
	Psychology	2.3%	1.8%	1.7%	1.7%
	Social and behavioral sci, interdisc	0.7%	0.5%	0.5%	0.5%
	Sociology and anthropology	0.8%	0.7%	0.6%	0.7%
	Total	8.0%	6.2%	5.5%	6.3%
Humanities	Creative arts, culture and music	0.8%	0.7%	0.4%	0.5%
	History, philosophy and religion	1.1%	1.0%	0.9%	0.9%
	Language and linguistics	0.3%	0.3%	0.2%	0.3%
	Law and criminology	0.5%	0.4%	0.4%	0.4%
	Literature	0.9%	0.7%	0.5%	0.3%
	Total	3.7%	3.2%	2.4%	2.4%
	Multidisciplinary journals	2.4%	1.8%	0.9%	0.7%
	Grand total	100%	100%	100%	100%
	N*	544965	751785	1068572	1715531

THE EFFECT OF FUNDING MODES ON THE QUALITY OF KNOWLEDGE PRODUCTION

Peter van den Besselaar[1] and Ulf Sandstrom[2]

[1] *p.a.a.vanden.besselaar@vu.nl*
VU University Amsterdam, Department of Organization Science & Network Institute,
Buitenveldertselaan 3, 1081 HV Amsterdam (The Netherlands)

[2] *ulf.sandstrom@indek.kth.se*
KTH, Indek - Department of Industrial Economics and Management,
Lindstedtsvägen 30, 10044 Stockholm (Sweden)

Abstract
Do funding modes have an effect on the quality of knowledge production? In this paper we develop an approach to investigate this, using the new WoS field on funder data, using climate change research in Sweden and the Netherlands in 2009-2010 as a case. We firstly developed an operational definition of climate change research, and retrieved all WoS records for the countries and years mentioned. We developed a classification scheme for the funding organizations of 13 categories, using dimensions as top-down/bottom-up, large/small research, national/international, and public/private. Then all funding institutions were manually classified in the 13 categories. We then calculated the average impact of the papers for each of the funding categories. The results clearly show differences between the funder types, and also between the countries. The latter indicates that a funding mode may be organized in different ways affecting the effectiveness. Finally, we discuss further research.

Topics
Bibliometric indicators, new developments (topic 1); Science Policy and Research Evaluation (topic 3); Modeling the Science System, Science Dynamics (topic 11).

Introduction
Whereas the research funding landscape in the past was relatively simple, with most funding going as block funds for universities, over the years, the number of funders has grown fast. Of course, national science councils entered the scene, but many other funders in government, private foundations, NGO's and companies are now active, plus many international organizations such as EC, ERC, OECD and so on.
The proliferation of the funding possibilities can be related to the changing relation between science and society, as research has become increasingly important in many realms of society. This changing relation is partly reflected in and constituted by the rise of a variety of new agenda setting arrangements, funding instruments, and new ways of organizing research and the interaction with societal stakeholders.

The development of project funding in its different forms has been studied by Lepori and Van den Besselaar et al [1,2], indicating (i) the growth of project funding in many countries, but (ii) at different levels and paces. A relatively detailed breakdown of the types of funding was developed for the Dutch data [3]. Also the OECD started a project [4,5] to refine the registration of public research funding. The relevance of a better and more detailed classification is obvious, as different types of funding actually may influence the type of research performed, the topical orientation, its relation with societal issues, and the scholarly and societal quality of the output. Only little research has focused on the effect of funding on knowledge production, but the introduction of funding acknowledgements in the Web of Science opens new possibilities. It now becomes possible to investigate the relation between funding mode and research output in more detail. Recent research has shown that the coverage is rather good, although problems of coverage, accuracy and completeness remain, as do problems of identification and disambiguation [6, 7].

In an older study, Cronin found (for information science) no relationship between funding acknowledgements and impact, however without differentiating between the types of funding acknowledged [8]. In a recent study, Rigby studied the relation between the number of funding sources and citation impact within physics and cell biology, and did not find a correlation between the two variables [9]. Costas and Van Leeuwen found that publications with funding acknowledgments present a higher impact as compared to publications without them, again without differentiating between different funders [10]. Wang and Shapira took a different approach and differentiated between funding institutions and types of funding institutions in nanoscience research in several large countries [6]. First of all, they found a predominant national orientation of research funding. But different funding arrangements exist in the different countries. Differentiating between funding modes, they found that the more funding is concentrated to a few recipient organizations, the lower the research impact as measured by citation counts is. Also Van den Besselaar et al focused on differences between research funders, when studying internationalization of research [7, 11].

Over the years, there has been a proliferation of funding (and related agenda setting) arrange-ments. This proliferation is the result of expanding science policy goals, translated by science policy makers into dedicated funding instruments (mechanisms). The more traditional funding modes, such as institutional block funding, and the responsive mode of the research councils are considered insufficient. We have witnessed the emergence of mission oriented, strategic and applied funding schemes, funding schemes for thematic consortia, applied and thematic public research institutes, etc. One may distinguish between four funding modes defined by two dimensions: bottom-up versus top-down, and institutional block grants versus project grants. Each of these four research modes can be organized using a variety of mechanisms, e.g., institutional funding for basic research may go to universities, or to public research insti-

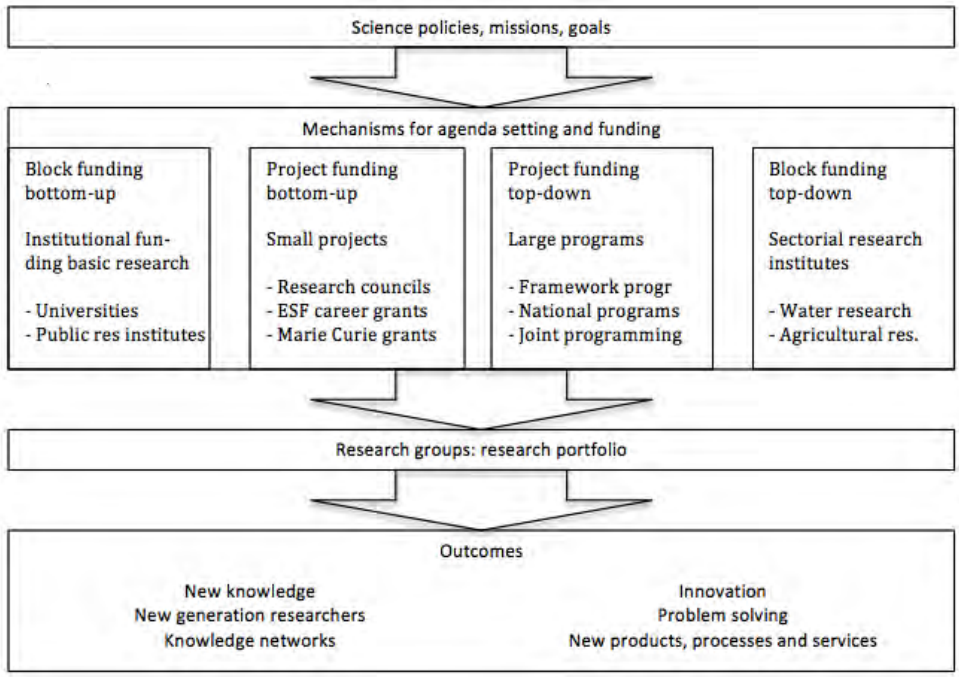


Figure 1: Comparative approach - the effects of funding modes

tutes. Research teams and research performing organizations use one or more of these funding mechanisms to produce their variety of outputs. Figure 1 briefly shows the model.

Several questions come up. Firstly, do these funding arrangements differ in productivity, impact, and originality? These differences can be measured in terms of the numbers of papers acknowledging the funding, and by the numbers of citations these papers receive. Secondly, do these funding arrangements actually fund different types of research and related output (scholarly versus societal output) and/or different topics of research? Are they complementing each other, or are they duplicating (and computing in quality – the first issue) each other? Thirdly, on the systems level, the question can be posed whether the variety of research funding? Is there an optimal variety? In this paper we focus on the first issue: do different funding modes result in different levels of impact?

Of course, we do not only observe differences in funding modes and instruments, but the same type of instrument can be organized in different ways, and this may influence the performance of the instrument. E.g., many ways of organizing applied sectorial research institutes may exist and many different ways of organizing peer review and panel selection processes in councils exist. Furthermore, differences may be related to disciplinary cultures. Therefore, we do not only have to compare the performance of the various funding modes, but also the variety within the modes, also reflecting differences between disciplines.

In order to test the possibilities, we study the different funding practices in one field (climate change) in two countries (Sweden and the Netherlands), in the recent period (2009-2010).

Data and methods

Since 2008 data on acknowledgements to funding units (FU) or funding sources are presented in the Thomson Reuter's database Web of Science as a searchable tag in the database. Data is acquired from the acknowledgements in journal papers (e.g. document types as articles, letters, proceedings and reviews). The indexing procedure copies the spelling mistakes and the different names of organizations presented in the journal papers. This creates a need for disambiguation of names of the funding organizations. One example: The Swedish Research Council can be presented in several different ways:

a.) The Swedish Research Council
b.) Vetenskapsrådet
c.) VR

The first of the above, (a), is the official name in English, (b) is the official name in Swedish, and (c) is the abbreviation of the Swedish name of the organization. There are several more possible versions and combinations of each of these names. Indeed, as we found elsewhere also for other funders, a funding organization may have hundreds of different ways of spelling. Also, there are possible homonyms and synonyms that altogether create a problem that might be solved through a more or less systematic disambiguation of organization names. Although the example above seems quite simple there are many public and private funding sources that can hardly be identified and disambiguated in a correct manner without manual procedures using the Internet or other sources of information.

In our sample, about 70 per cent of papers do have an acknowledgement of funding sources, what is higher than what would be expected as only about half of total Swedish and Dutch papers do have FU-information during the period. Distribution over areas has to be taken into account when we discuss figures of papers with and without acknowledgement of funding.

Classifying funding organizations and funding modes

For Sweden, the ten most frequent funding sources, accounts for more than 20 % of all acknowledgements in the Swedish sample data. The numbers of unique funding sources are about 1,000, which illustrates the problem and the need for disambiguation of funding names. A complete disambiguation of all funding sources is impossible as there are many that only consist of a project of program abbreviation. Under all circumstances, it is necessary to categorize the different funding sources according to the financiers' mission and procedures for evaluation of proposals.

How to account for different funding modes? We started from a two-way matrix based on the distinction between open and thematic mission for a funding

organization on the one hand and the distinction between bottom-up and top-down procedures on the other hand (figure 1). We added the distinction between national and international funding. Basically, we would like to be able to use the distinctions proposed by van Steen [4; 5] between institutional, block grant funding, on the one hand and project funding on the other. Unfortunately that is not possible due to limitations in the FU-data. Therefore, we cannot test hypothesis related to that distinction (although category 9 and 12 can be related to that question). All other categories are dominated by project funding schemes of different sizes and arrangements. Bourke and Butler [15] had more detailed information in their path-breaking study. Heinze [16] focuses more on peer review as mode of funding procedure and concentrates on some main schemes applied. Later on we hope to be able to use that type of granularity. In this investigation we consider the different types of funders that are revealed by the FU tag in the Web of Science. In our understanding, while some of the categories are associated with frequent use of modified peer review in a responsive mode, others are associated with less academic and more open evaluations of proposals (e.g. category 2, 4, 5, 6).

The following categories are used to classify funding organizations:

1. Research Councils bottom-up, open,
2. Organizations, private foundations, NGO's, etc.
3. Foreign
4. Applied funders, county councils, Nordic council
5. Mission-oriented bottom-up
6. Applied research institutes
7. EU framework, Marie Curie etc.
8. University
9. Research Institutes, fundamental research
10. Missing category
11. Companies
12. Large programs, Excellence programs, Research Foundations
13. Societies

After classifying the Swedish data, the Netherlands data were processed in the same way.

Delineating climate change research

We used the three WoS databases SCI expanded, SSCI and AHCI. In order to delineate climate change research, we started with the search ts=climat* and checked for a random set of papers the precision. Clearly quite some papers were retrieved that do not focus on climate change. Then we used a more restricted search, using ts="climat* change*" which led to a much smaller set of papers. Checking the difference between the two sets, we found quite some relevant papers that were not in the second search. Therefore we designed a query that was in between the two tests. A test indicated that the precision and recall were OK. We used the following query:

ts=climat* and (ts=change* or ts=variabilit* or ts=anthropogenic* or ts=model* or ts=strategy* or ts=policy* or ts=regime* or ts=scenario* or ts=carbon* or ts=”integrated assessment” or ts=environment* or ts=reforestati* or ts=deforestati* or ts=desertificati* or ts=”greenhouse gas”* or ts=GHG or ts=ecolog* or ts=environment* or ts=biodiversity or ts=”global change” or ts=”water stress”) or ts=climate-driven or ts=”global warming” or ts="sea level*" and (ts=change* or ts=rising)

We tested whether e.g., papers on climate change mitigation and adaptation were included, even without using the latter two search terms, and this was the case for more than 90%.[61]
The set was refined for publication years (PY=2009-2010), for document type (DT=article or proceeding paper or letter or review), and for country (CU=Sweden or Netherlands). This resulted in 954 Swedish papers and 1293 Netherlands' papers that were used in the analysis.

Analysis

After having classified all mentioned funders, we used a dedicated Swedish tool to estimate the impact of the publications funded by the different sources. We calculate the average field normalized citation impact for each of the funding modes in the two countries, for publications in the field of climate change. We also calculated for each of the funding modes the percentage of papers in the top 1%, the top 5%, and the top 10% in the relevant journal environments [12].
Relative indicators or rebased citation counts, as an index of research impact, are widely used by the scientometrics community. We calculated a weighted NCSf (Field Normalized Citation Score), based on fractional counts based on the number of funders per paper. This gives a weight for the contribution of the funder to the impact of papers. Fractional counting is a way of controlling the effect of collaboration (here between funders) when measuring output and impact. Consequently, figures based on fractional counting show the extent to which the set of papers receives many citations for the collaborative funded papers only, or if the papers that were funded by a single funding agency are cited in the same manner.

Some restrictions

Having FU details does not imply that we have the full information of all funding sources. In some cases universities and university departments are mentioned as one funding source, especially if there is a specific program at the university e.g.

[61] We did not further investigate recall and precision of this search string, as we do not aim to cover climate research completely, but only need a representative sample from climate change research in Sweden and the Netherlands, in order to compare the funding modes. We assume here that the sample is good enough for this.

for climate research, but in the normal case the contributions from the university in the form of faculty funding is not acknowledged by the authors of papers.
We do not have data about the amount of funding per project by funding organizations. One of the organizations might contribute with 1 million Euros and another organization with less than 50,000 Euros. In the same way, it is impossible to know the extent to which different sources have been used for the specific article published. It might be the case that a researcher develops ideas and produces results in a project, but when the article is finally published, he/she may be already involved in a new and completely different project with new funding sources – and consequently acknowledges the *new* funder. All these problems exist, but we have to consider that on the *micro* scale, the systems of input and output are always disconnected (to some extent). However, in the long run there will be a tendency for people to acknowledge funding streams and many of these will rely on sources for several years. In that way there is always a connection established between funding and output.
Climate Research is a growing area, attracting different types of funding. When an area grows, it also attracts interest from researchers relabeling their work in order to fit in to the new funding opportunities. In such arenas there might be a signaling value for the researcher as well as for the funding agency to point out that the respective partners are active within the area of this specific type of research.

Findings

Structure and growth of the field

Clearly, the field is young and grows very fast (fig 2). Is it covered by old journals, changing to climate change research, or new journals focusing on climate change? We list here the 11 most frequent journals (table 1). As the table indicates, the journals are relatively young as seven were founded after 1980, and two more in the 1970s. Apart from the general journal PlosOne, most journals are on climate and global change (6) or on (atmospheric) geophysics (3).

The two countries we focus on both have a faster growth than world average, where we took the year 2000 as 100 (figure 2). The growth of climate change research in the Netherlands has been faster than average since about the year 2000, with growth acceleration around 2003 with the start of the *Klimaat voor Ruimte* program and a second impulse with the *Kennis voor Klimaat* program around 2008. These two programs are (in the classification deployed here) in the "large programs" category.
Sweden has invested heavily in climate research [13] and followed the world growth until 2009, but is strongly speeding up since. We also include two other countries in the graph, for comparison. Switzerland follows a similar fast growth path as the Netherlands, and Germany is following about the world growth rate.

Table 1: Main journals in the field of climate change research

Name or journal	*first volume*	*nr of papers in the set (2009/10)*
Journal of Geophysical Research Atmospheres	1900	1782
Journal of Climate	1988	1724
Geophysical Research Letters	1973	1670
Global Change Biology	1996	1009
Climatic Change	1978	992
Climate Dynamics	1986	940
Atmospheric Chemistry and Physics	2000	927
Palaeogeography Palaeoclimatology Palaeoecology	1965	893
Quaternary Science Reviews	1982	854
PlosOne	2006	852
International Journal of Climatology	1981	698

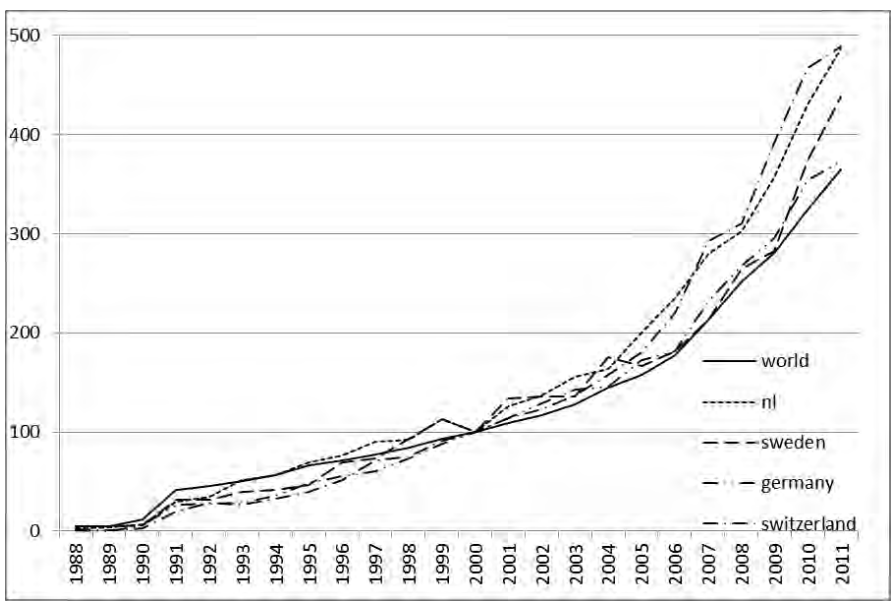


Figure 2: Growth in climate change research 1988-2011 (2000=100)

Funders and impact

First of all, many papers mention more than one funder, and Table 2 shows the number of funders by papers. The table also shows the average field normalized impact (NCSf) for each of the sets. The table suggests that many (more than four) funders are good for the citation rate.

Table 2: Impact by number of funders

Nr FA	Papers	NCSf
1	495	1,84
2	181	1,60
3	114	1,72
4	59	1,64
5	40	2,80
>=6	65	4,09

Sweden, Climate Science, 2009-2010

At the same time, it is not the case that all papers refer to all funders, and therefore we may be able to study the effect of the mode of funding on the impact of papers. Tables 3 (Sweden) and 4 (Netherlands) present the basic findings about the impact of the climate change papers within the several funding modes in the two countries. We report for each of the funding modes the nr of (integer counted) publications, the field normalized citation scores, and the share of papers in the set of top-cited papers.

In both countries, the largest categories are Foreign, EU and the national research council. Also the group of papers without funder is among the largest. Of course, one should take into account that the category "foreign" includes a large number of different funders (and funder types), most of them only funding a few papers. So the high impact of this category is not related to a specific funding mode, but probably to the fact that if a researcher collaborates with foreign researchers that have obtained funding, he/she has a good international team resulting in high impact results.

Table 3: impact of funding types – Sweden, Climate change research, 2009-2010

	# papers	Field normalized citation score	Share in top cited papers 1%	5%	10%
EU	175	2.46	***6.0%**	**19.8%**	**36.5%**
Foreign	322	2.21	**6.2%**	**15.0%**	**26.8%**
No funder mentioned	290	1.92	4.8%	13.1%	24.1%
Mission-oriented Council	98	1.89	**6.1%**	**17.1%**	25.6%
Research Council	142	1.75	3.0%	9.8%	22.0%
Charities, Organizations	70	1.69	**7.8%**	10.5%	17.0%
Corporations	31	1.67	2.6%	4.5%	**29.7%**
Large programs	32	1.59	2.6%	4.8%	10.6%
Societies	35	1.57	3.3%	9.7%	**26.2%**
Universities	107	1.53	3.3%	8.1%	13.0%
Applied Research Institute	21	1.53	0.0%	**13.2%**	25.2%
Applied funder	122	1.48	0.5%	7.6%	17.3%
Total	954	1.96	4.8%	13.1%	24.6%

* bold: belonging to the top 4 performing types in this indicator

Table 4: impact of funding types – Netherlands, Climate change research, 2009-2010

	# papers	Field normalized citation score	Share in top cited papers 1%	5%	10%
Foreign	491	2.56	***6.7%**	**19.9%**	**30.2%**
Large programs (Bsik / FES)	31	2.52	1.2%	**24.8%**	**37.0%**
Corporations	22	2.50	1.7%	10.3%	27.5%
EU	221	2.22	**5.6%**	**19.4%**	**30.1%**
Applied Research Institute	28	2.16	**5.2%**	16.3%	26.7%
Societies	15	2.05	0.0%	18.6%	26.1%
Mission-oriented Council	23	1.96	2.6%	**22.0%**	**36.4%**
Universities	74	1.93	4.7%	17.1%	25.8%
No funder mentioned	486	1.90	**5.1%**	12.6%	22.4%
Research Council (NWO)	208	1.74	2.4%	12.5%	21.8%
Applied funder	56	1.49	3.4%	10.8%	18.9%
Charities, Organizations	8	1.23	0.0%	0.0%	5.2%
Basic research Institute	8	0.91	0.0%	11.4%	11.4%
Total	1293	2.09	5.0%	15.6%	25.7%

* bold: belonging to the top 4 performing types in this indicator

Figure 3 and 4 show the distributions of the normalized citation impact for the main categories of funders in each of the countries. The figures should be read in the following way: We distinguish nine impact classes.[62] For each of the funder types, we calculated the share of papers in each of these nine impact classes. The share of non-cited papers (class 1) is placed to the left in the graph. Then we have the sum of the two lowest scoring classes (1-2), the sum of the three lowest classes (1-3), and so on.

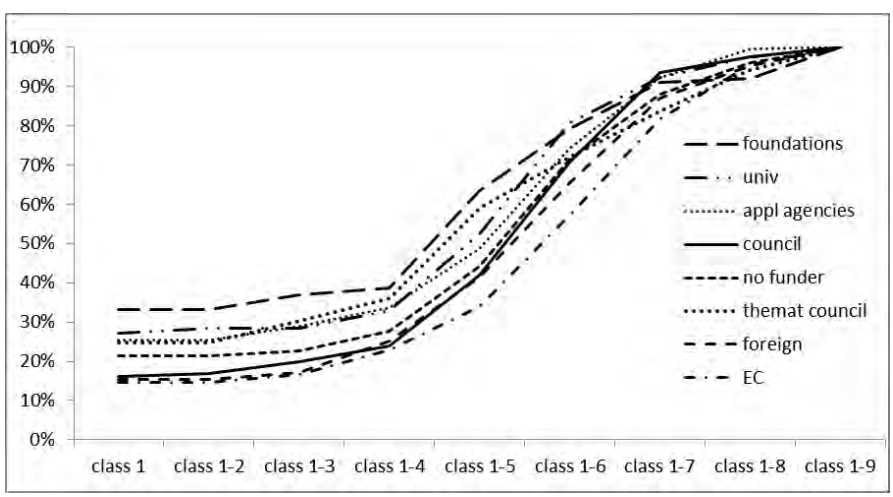


Figure 3: Cumulative distribution of impact by funder type, Sweden (class 1= non-cited, class 1-2 = non-cited plus lowest cited, etc),

[62] The classes of citation impact (NCSf) are defined as follows: 1=0, 2=>0-0,125; 3=>0,125-0,25; 4=>0,25-0,5; 5=>0,5-1,0; 6=1.0-2.0; 7=>2-4; 8=>4-8; 9=>8.

For each of the seven (Netherlands) or eight (Sweden) main funder types, we display the cumulative distribution of impact, and the lower the line in the figure, the larger the share of high impact papers this funder has. E.g., in Sweden, EC funded climate research has the highest (mean and median) impact, whereas the foundations have the lowest.

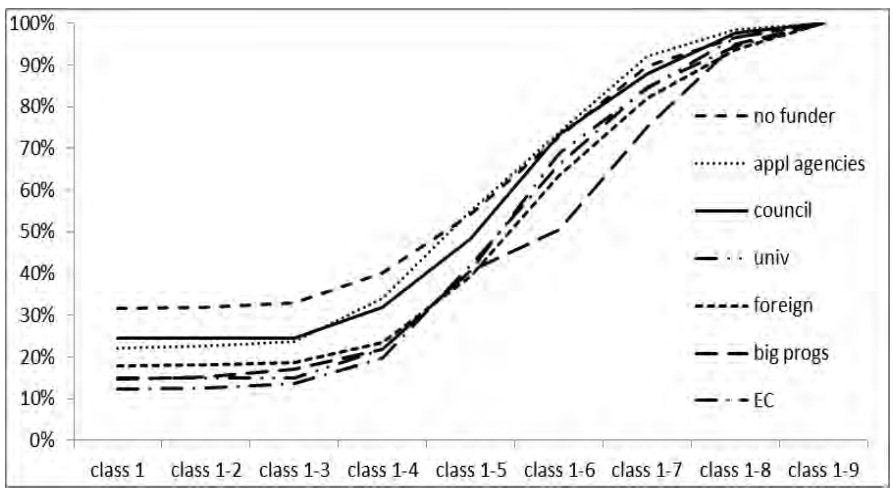


Figure 4: Cumulative distribution of impact by funder type, Netherlands (class 1= non-cited, class 1-2 = non-cited plus lowest cited, etc),

Comparing the tables 3 and 4 shows the following results, consisting of similarities (1-6) and differences (7-9) between the two countries – for the case of climate change research in the recent period.

1. Climate change research started in the early 1980s, and showed a fast growth
2. Dutch and Swedish climate change research has grown fast in the recent period, faster than world average.
3. On average, Swedish and Dutch climate research score about the same, both countries have impact scores 100% above international average.
4. Research councils only perform at an average level, not very strongly contributing to the top output.
5. Output generated together with (funded and therefore high level) international co-authors scores the best in the Netherlands, and almost the best in Sweden. Here we do not see so much an effect of a funding mode, but more a characteristic of researchers: collaborating with foreign researchers that obtain funding for their research seems creating good consortia for high impact research.
6. EU funded work scores very well in both countries, above 100% better than the international average. Here, also the international collaboration effect may play a role.

7. The distribution of papers over types of funding organizations is different in the two countries. E.g., it seems that the charities and other organizations that fund research (such as NGOs) play a much bigger role in Sweden than they do in the Netherlands.
8. The opposite holds for the large programs that score low in Sweden but high in the Netherlands. These are funded out of a specific source and meant for excellent and societal relevant research in targeted fields, such as climate change.
9. Applied research institutes located in the Netherlands score much higher than their Swedish counterparts.

Conclusions

This is a preliminary and first attempt to determine the relation between funding mode and impact of research. The findings suggest that international collaborative and funded research leads to high impact. In the Netherlands, we also find some other high impact funding modes: companies, applied research institutes, and special programs. As these funding modes score lower in Sweden, this poses the question as to whether the organization of funding (next to the type of funding) has an own and independent effect. So special programs can be organized better or poorer, influencing the impact of the funded research. Finally, the impact of papers funded by the national councils is in both countries relatively low. The far majority of these papers do not mention international funding or EC funding, which may be related to this finding.

Further work

In a follow up project, we will apply the approach or a variety of other fields, in different disciplines, and in different stages of development. Not only 'hot' fields as climate change research, but also fields that are less in the focus of science policy makers, and of the general public. We also plan to study different modalities of research funding types, in order to find out how the organization of a type of funding may affect the selection and through this, the impact of the funded research. Thirdly, we intend to compare funding patterns of top-researchers, compared to the average researcher (Verbree et al 2013).

Acknowledgments

The research is supported by the Knowledge for Climate Program (Netherlands) and by the Tercentenary Foundation (Sweden). Agnes Wold (University of Goteborg) provided useful comments on an earlier draft, as did three anonymous reviewers.

References

[1] Lepori, Benedetto, Peter van den Besselaar, Michael Dinges, Bianca Potì, Emanuela Reale, Stig Slipersaeter, Jean Theves, Barend van der Meulen,

Indicators for Comparative Analysis of Public Project Funding. Concepts, Implementation and Evaluation. *Research Evaluation* **16** (2007) 4
[2] Lepori, Benedetto, Peter van den Besselaar, Michael Dinges, Barend van der Meulen, Bianca Potì, Emanuela Reale, Stig Slipersaeter, Jean Theves, Comparing the evolution of national research policies: what patterns of change? *Science & Public Policy* **34** (2007) 5
[3] Versleijen, Anouschka, Barend van der Meulen, Jan van Steen, Penny Kloprogge, Robert Braam, Ruth Mamphuis, Peter van den Besselaar, *Thirty year of public research funding – trends, policy and implications* (in Dutch). Den Haag: Rathenau Instituut 2007.
[4] Van Steen, Jan, *Modes of public funding of R&D: towards internationally comparable indicators*. STI working paper (version 13 January 2012).
[5] Pastor, Elisabeth; van Steen, Jan, C.G. (2012), *Draft guidelines for data collection on modes of public funding of R&D based on GBAORD, OECD, DSTI/EAS/STP/NESTI* (2012)12.
[6] Wang, J. & Shapira, P. (2011). Funding Acknowledgement Analysis: An Enhanced Tool to Investigate Research Sponsorship Impacts: The Case Of Nanotechnology. *Scientometrics*, 87 (3), 563-586.
[7] Van den Besselaar, Peter, Annamaria Inzelt, Emanuela Reale, Measuring internationalization of funding agencies? E. Archambault, Y. Gingras, V. Lariviere (eds.) *Proceedings Science & Technology Indicators 2012*. Montreal, OST & Science Metrix, 2012, 121-130
[8] Cronin, B & D Show, Citation, funding acknowledgement and author nationality relationships in four information science journals. *Journal of Documentation* **55** (1999) 402-408
[9] Costas, R. & van Leeuwen, T.N. (2012). Approaching the 'Reward Triangle: General Analysis of the Presence of Funding Acknowledgements and 'Peer Interactive Communication' in Scientific Publications. *Journal of the American Society for Information Science and Technology* 63 (8), 1647–1661.
[10] Rigby, J. (2011). Systematic Grant and Funding Body Acknowledgment Data for Publications: New Dimensions and New Controversies for Research Policy and Evaluation. *Research Evaluation*, 20 (5), 365-375.
[11] Van den Besselaar, Peter, Annamaria Inzelt, Emanuela Reale, Elisabeth de Turckheim, Valerio Vercesi, *Indicators for internationalization of research institutions*. Strasbourg: European Science Foundation ESF (2012)
[12] Leydesdorff L, and L. Bornmann, Percentile Ranks and the Integrated Impact Indicator (I3), *Journal of the American Society for Information Science and Technology* **63** (9), 1901-1902.
[13] *Svensk klimatforskning – vad kostar den och vad har den gett?* (RiR 2012:2)
[14] Verbree M, Van der Weijden V, Van den Besselaar P, Academic leadership of high performing research groups. In: Hamlin, C.M. Allwood, B. Martin, M.M. Mumford (eds.), *Creativity and leadership in science, technology and innovation*. London: Routledge 2013, pp 113-148.

EFFECTS OF RESEARCH FUNDING, GENDER AND TYPE OF POSITION ON RESEARCH COLLABORATION NETWORKS: A MICRO-LEVEL STUDY OF CANCER RESEARCH AT LUND UNIVERSITY

Fredrik Åström[1], Ingrid Hedenfalk[2], Mikael Graffner[3] and Mef Nilbert[4]

[1] *fredrik.astrom@ub.lu.se*
Lund University Libraries, P.O. Box 134, SE-221 00 Lund (Sweden)

[2] *ingrid.hedenfalk@med.lu.se*
Lund University, Clinical Sciences Lund, Dept of Oncology, BMC C1341a, SE-221 84 Lund (Sweden);
Regional Cancer Centre South, Medicon Village, Scheelevägen 8, Bldg 404, SE-223 63 Lund (Sweden)

[3] *mikael.graffner@ub.lu.se*
Lund University Libraries, P.O. Box 134, SE-221 00 Lund (Sweden)

[4] *mef.nilbert@med.lu.se*
Lund University, Clinical Sciences Lund, Dept of Oncology, BMC C1341a, SE-221 84 Lund (Sweden)
Regional Cancer Centre South, Medicon Village, Scheelevägen 8, Bldg 404, SE-223 63 Lund (Sweden)

Abstract
The aim of this study was to analyse how research funding, gender and research area relate to the size and density of collaborative networks within cancer research. The material consisted of 3,306 publications from scientists in cancer research associated with Lund University, indexed in the Web of Science databases. The author and address fields were analysed, by studying frequencies and distribution of authors and organizations, and by conducting co-authorship analyses on the organizational level. The results showed substantial differences between scientists with and without national funding, defined as research grants from the Swedish Cancer Society (SCS). Collaborative research networks were larger and denser among scientists with national grants and these differences were more pronounced than differences related to sex and research area, i.e. preclinical versus clinical research. The results suggest that the relation between research funding and the size and nature of collaborative research networks is stronger than the relations between gender or research orientation.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6).

Introduction

Research collaboration has been analysed from a wide range of perspectives, most commonly using bibliometrics and in particular through co-authorship analyses (e.g. Katz & Martin, 1997; Melin & Persson, 1996; Newman, 2004). Important areas of investigation include analyses of factors behind research collaboration (e.g. Abbasi et.al., 2011; Birnholtz, J.P., 2007; Hara et.al., 2003; Jeong et.al., 2011; Lewis et.al., 2012), issues related to international research networks, measures of scientific productivity and impact in relation to co-authored research papers (e.g. Lee & Bozeman, 2005; Lukkonen et.al., 1992; Lukkonen et.al., 1993; Narin et.al., 1991; Persson, 2010). Another important issue, in social studies of science, and increasingly also in bibliometrics, is analyses of the impact of gender: on the access to research collaboration networks, but also on the access to research funding, the peer review process and opportunities for building scholarly careers in general, as well as in terms of scientific impact and productivity (e.g. Alcaide et.al, 2009; Kretschmer et.al., 2012; Larivière et.al, 2011; Mählck, 2001; Wennerås & Wold, 1997).

The effects of research funding and research orientation on the construction of, and access to, research networks have been analysed to a lesser extent. The relation between funding and productivity, impact and collaboration has been analysed by e.g. Clark and Llorens (2012), Haiqi (1997) and Heffner (1981), all of whom identified a positive relation between financial support and the size of networks of collaborating scholars; and Zhao (2010), who found that research with grant-funding had a larger impact in library and information science. In relation to type of positions: Bordons et.al. (2003) analysed scientific productivity in relation to gender and professional category; and while there are overall differences between genders, the differences between genders within the professional categories are not significant. When analysing pre-clinical and clinical sciences, Satyanarayana & Ratnakar (1989) found that clinical sciences have a higher average of authors per paper than preclinical basic research areas such as biochemistry and molecular biology.

However, the question remains: can we identify relations between on one hand: type of funding and type of position or orientation, together with gender; and on the other: the character of research networks? The aim of this study is to analyse the extent of which different types of research funding, gender and type of position can be related to the size and density of research collaboration networks. To analyse this issues, micro-level analyses were performed on a group of cancer research scientists associated with Lund University (LU), including researchers at the Skåne University Hospital, and with a particular focus on scientists with or without national funding, defined as research grants from the Swedish Cancer Society (SCS). The reason for focusing on scientists with funding from the SCS is that, in the context of Swedish cancer research, SCS funding/grants can be seen as a proxy indicator on being an established cancer scientist.

Methodology

The first stage of the data selection process was to identify scientists at LU involved in cancer research through identification of scientists responsible for PhD research supervision with projects categorized as cancer research. In total, 93 scientists were identified; 47 with research funds from the SCS and 46 without SCS funding. Using Lund University Publications (LUP), the Web of Science (WoS) 'accession number' for these scientists' publications between the years 2002-2011, were used for retrieving 3,306 publications in WoS. Based on the full dataset, 14 subsets were created in order to perform analyses of cooperation networks among cancer scientists according to the analytical categories selected for this study; a division based on differences in research funding – i.e. with or without research funds from the SCS, gender and work orientation – i.e. those solely having pre-clinical positions and those with clinical or combined positions. To control for effects by the different analytical categories, a further division of sub-sets was done analysing differences between men and women, as well as pre-clinical and clinical/combined scientists, within the SCS and non-SCS document sets.

The analyses were based on the 'author' (AU) and 'address' (CS) field from WoS. Before the analyses were done, author and author address data was cleaned and standardized. The main organization was identified as the name before the first comma in the CS-field, thus: in cases where e.g. both a university and a hospital are part of the same address, only the first mentioned named will occur in the analyses. Also, variant names of organizations were standardized, where e.g. 'Malmo Univ Hosp' was changed into 'Skane Univ Hosp' (Figure 1).

Doc.nr.	Address
1	Lund Univ, Div Clin Chem, Dept Lab Med, Skane Univ Hosp, Malmo, Sweden
1	Malmo Univ Hosp, Wallenberg Lab, Entrance 46, SE-20502 Malmo, Sweden

Figure 1. Example: WoS CS-field, author addresses.

The analyses were performed using the Bibexcel software (Persson, Danell & Schneider, 2009), on both author and organization level. On author level, author frequencies and the distribution of authors per document were investigated; the latter both by looking at the average number of authors per document, and by analysing the distribution of documents according to numbers of authors. The organisation level analyses were performed both by looking at frequencies as well as the distribution of organizations per article; and by co-occurrence analyses (Melin & Persson, 1996) of author addresses, which were visualized using Pajek (de Nooy, Mrvar & Batagelj, 2005).

The analyses were conducted using full counting on both author and organization level. Thus, in cases where there are articles involving e.g. both SCS and non-SCS funded LU scientists, where both are also among the 93 selected, there will

be an overlap of articles distributed between SCS and non-SCS scientists. In cases where there are scientists with more than one affiliation, all organizations were counted if stated in the CS-field.

Results

The results section reports the results in three sub-sections: the first accounts for basic information on the dataset in terms of LU authors laying the foundation for the documents, the second sub-section presents the results of the author level analyses and the third sub-section reports on the number of organizations – as represented by article author addresses – involved in collaboration with LU cancer scientists.

LU Authors and Documents per Analytical Category

The most basic set of results from the analyses was LU author – i.e. the scientists selected as basis for the data collection, the actual number of authors contributing to each article will be reported in the second section of 'Results' – and document frequencies as well as the distribution of documents per author within each analytical category (Table 1). Apart from the number of authors, articles and documents per author, the aforementioned overlap due to the full counting is also reported.

Table 1. Distribution of LU authors and documents per analytical category.

	LU Authors	*No. of Documents*	*Documents/ Author*	*p-value*	*Overlaps*
Full set	93	3,306	35.55		
SCS	47	2,029	43.17	p=0.0033	222
Non-SCS	46	1,499	32.59		
Women	31	993	32.03	p=0.036	350
Men	62	2,663	42.95		
Pre-clinical	43	1,448	33.67	p=0.10	403
Clinical/Combination	50	2,261	45.22		
SCS: Women	12	465	38.75	p=0.16	245
SCS: Men	35	1,809	51.69		
Non-SCS: Women	19	547	28.79	p=0.32	26
Non-SCS: Men	27	978	36.22		
SCS: Pre-clin.	27	1,012	37.48	p=0.011	329
SCS: Clin./Comb.	20	1,346	67.3		
Non-SCS: Pre-clin	16	479	29.94	p=0.56	21
Non-SCS: Clin./Comb.	30	1,041	34.7		

The data contain considerable variations within and between the different analytical categories. Since the main focus of the study was to investigate differences between scientists with and without SCS funding, we designed the

study to have an even distribution of authors between those groups. There were, however, more men than women; and more scientists with a clinical or combined position than with an exclusively pre-clinical position. In terms of documents per LU author, the average number of papers per author was higher among those with SCS funding, men and clinical/combined scientists. It should however be noted that there were also more authors among men and clinical/combined scientists, whereas the distribution of LU authors among SCS and non-SCS funded scientists was relatively even.

As previously mentioned, a further division between sub-categories was made and gender and type of position were analysed in relation to funding from the SCS. Using average values, we found differences between women and men with SCS funding, as well as between pre-clinical and clinical scientists without SCS funding. However, there were some tendencies towards larger differences between e.g. men with or without SCS funding than between men and women without SCS funding. Thus, the rest of the analyses focused on investigating gender and type of position differences in relation to access to funding, rather than as separate entities. At the same time, when looking at p-values, we found significant differences between men and women, as well as between pre-clinical and clinical/combined scientists, together with those with or without SCS funding. However, when looking at the differences in the distribution of papers between men and women, it should be taken into account that the age distribution was also varied, among the LU researchers analysed here: men were typically older than the women; and thus likely to have come further in their careers, as well as having produced more papers (Figure 2).

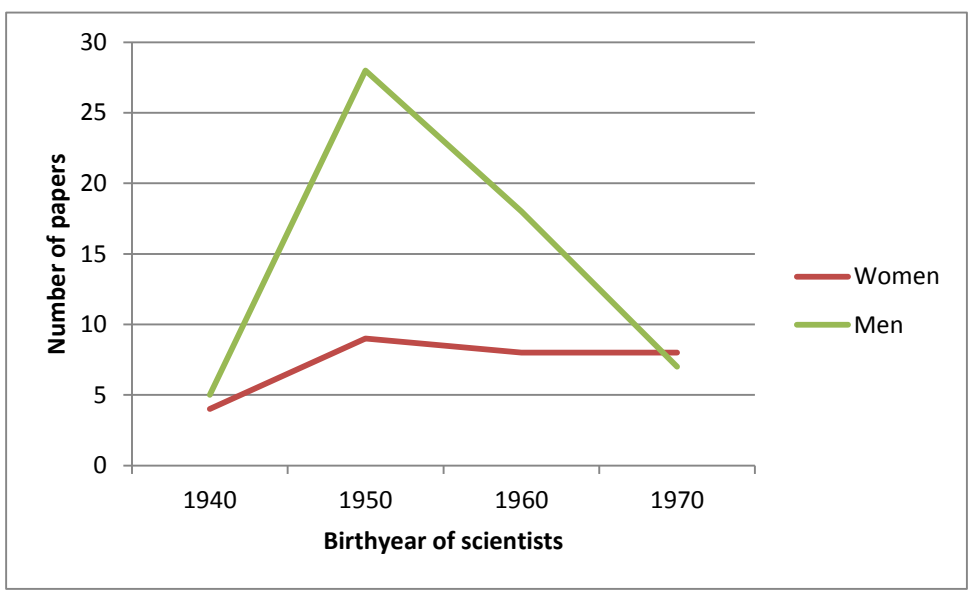


Figure 2. Distribution of papers in relation to age and gender.

Author level collaboration

On author level, investigations were made into the number of authors contributing to LU cancer papers, as well as the average number of authors per article (Table 2). Taken together, the average number of authors per article was almost nine. However, when looking at the different analytical categories, articles by SCS funded LU authors had substantially more authors per article than the ones without SCS funding: whereas the set based on SCS funded authors showed an average around 10, the average for non-SCS funded articles was around six authors per article. In this analysis, the tendency towards differences based more on funding rather than gender or orientation appeared stronger. Average number of authors for papers by male or female LU scientists varied little among e.g. SCS funded papers, while the difference for SCS and non-SCS men or women was larger. The one notable exception was between pre-clinical and clinical/combined scientists without SCS funding.

Table 2. Number of authors and average number of authors per article.

Category (N=no. docs)	*Authors*	*Author/Article*
Full set (N=3,306)	8,930	8.69
SCS (N=2,029)	5,843	10.3
Non-SCS (N=1,499)	3,644	6.38
SCS: Women (N=465)	1,398	10.7
SCS: Men (N=1,809)	5,728	10.25
Non-SCS: Women (N=547)	1163	5.95
Non-SCS: Men (N=978)	2895	6.64
SCS: Pre-clin. (N=1,012)	3,336	9.7
SCS: Clin./Comb. (N=1,346)	3,976	10.86
Non-SCS: Pre-clin (N=479)	1,316	2.23
Non-SCS: Clin./Comb. (N=1,041)	2,556	6.47

Apart from the average number of authors per article, analyses were performed on the distribution of articles over papers with different number of authors (Table 3). In total, the grand majority of the papers had 1-10 authors, regardless of analytical category, and a very small amount of papers with more than 50 authors. However, the largest shares of papers with 1-5 authors were found in the non-SCS set, while the majority of articles with more than 20 authors were primarily found in the set of documents by authors with SCS funding. In addition to the results presented in the table, it is also noteworthy that all papers – albeit being very few – with more than a 100 authors were found among the articles by scientists funded by the SCS.

Both the analyses of average authors per article and the distribution of articles over number of authors showed substantial differences between documents involving LU scientists with or without SCS funding. In terms men and women,

and clinically/combined or pre-clinical scientists, however, there were differences, but to a lesser extent than in relation to funding.

Table 3. Distribution of articles over authors per article.

Category (N=no. docs)	*1-5*	*6-10*	*11-20*	*21-50*	*51-*
Full set (N=3,306)	44 %	38 %	12 %	5 %	0,8 %
SCS (N=2,029)	36 %	41 %	15 %	8 %	1 %
Non-SCS (N=1,499)	52 %	38 %	9 %	1 %	0,3 %
SCS: Women (N=465)	32 %	47 %	10 %	11 %	-
SCS: Men (N=1,809)	36 %	41 %	15 %	8 %	1 %
Non-SCS: Women (N=547)	56 %	32 %	11 %	0,3 %	-
Non-SCS: Men (N=978)	49 %	41 %	7 %	1 %	0,5 %
SCS: Pre-clin. (N=1,012)	37 %	43 %	14 %	5 %	1 %
SCS: Clin./Comb. (N=1,346)	32 %	41 %	15 %	10 %	1 %
Non-SCS: Pre-clin (N=479)	60 %	32 %	8 %	0,6 %	0,4 %
Non-SCS: Clin./Comb. (N=1,041)	49 %	41 %	9 %	1 %	0,3 %

Organization level collaboration

The collaboration networks of LU cancer scientists were also studied on organization level, using the CS-field in WoS. As with the author level analyses, the frequency of organizations were analysed for the different analytical categories, as well as the distribution of organizations per article. In addition to these analyses, collaboration networks were also analysed doing an organization level co-authorship analysis.

In total, 1,385 organizations were identified among the author addresses, with an average of 3.86 organizations per article. And as in the case of collaborating authors, the number of institutions contributing together with LU and Skåne University Hospital was higher among the articles by scientists with SCS funding, while the differences between men and women or clinical and pre-clinical scientists were smaller (Table 4).

To investigate the collaboration networks, a co-authorship analysis was conducted (Melin & Persson, 1996). For each analytical category, the numbers of unique pairs of organizations formed in the co-authorships were identified, together with the number of links between them and the average number of links per pair (Table 5). The latter was in part to adjust for the sheer number of organizations in the different analytical categories, but also as an indicator on the density of the networks.

Table 4. Number of organizations and average number of organizations per article.

	Number of organizations	*Average: Organizations/article*
Full set (N=3,306)	1,385	3.86
SCS (N=2,029)	1,070	4.81
Non-SCS (N=1,499)	733	2.69
SCS: Women (N=465)	260	4.86
SCS: Men (N=1,809)	999	4.76
Non-SCS: Women (N=547)	269	2.59
Non-SCS: Men (N=978)	572	2.72
SCS: Pre-clin. (N=1,012)	665	4.11
SCS: Clin./Comb. (N=1,346)	819	5.22
Non-SCS: Pre-clin (N=479)	313	2.58
Non-SCS: Clin./Comb. (N=1,041)	558	2.72

Table 5. Organization level co-authorship pairs and co-occurrence links.

	Unique pairs	*Link frequency*	*Average links/pair*
Full set	37,293	90,244	2.42
SCS	33,192	81,981	2.47
Non-SCS	8,118	11,372	1.4
SCS: Women	2,844	14,889	5.24
SCS: Men	32,193	74,440	2.31
Non-SCS: Women	1,165	2,373	2.04
Non-SCS: Men	7,027	8,885	1.26
SCS: Pre-clin.	23,770	42,826	1.8
SCS: Clin./Comb.	18,448	51,687	2.8
Non-SCS: Pre-clin.	3,548	4,252	1.2
Non-SCS: Clin./Comb.	4,935	7,264	1.47

In general, the organizational networks for scientists with SCS funding were larger than for those without SCS funding, both in terms of the number of pairs formed as well as links between the individual pairs; and this tendency also remained when comparing men and women as well as pre-clinical and clinical/combined types of positions within the SCS and non-SCS categories. One interesting thing to note, is that women in both the SCS and non-SCS sets had networks with a higher average of links per pairs than men; and in the SCS set, the average for women was more than twice that of the men, thus appearing to have denser collaboration networks.

The co-authorship networks were also visualized using Pajek (de Nooy, Mrvar & Batagelj, 2005) and the Kamada-Kawai (1989) algorithm. The visualization analyses were conducted for LU cancer scientists with and without SCS funding, as well as further breaking down the analyses into men and women with and

without SCS funding, and pre-clinical and clinically/combined scientists with or without SCS funding. In terms of structures discovered in the visualizations of the different analytical categories, there were substantial differences between scientists with or without SCS funding, whereas the differences between e.g. men and women with SCS funding were small. Thus, in this paper, only the visualizations for researchers with and without SCS funding are included (Figure 3a&b).

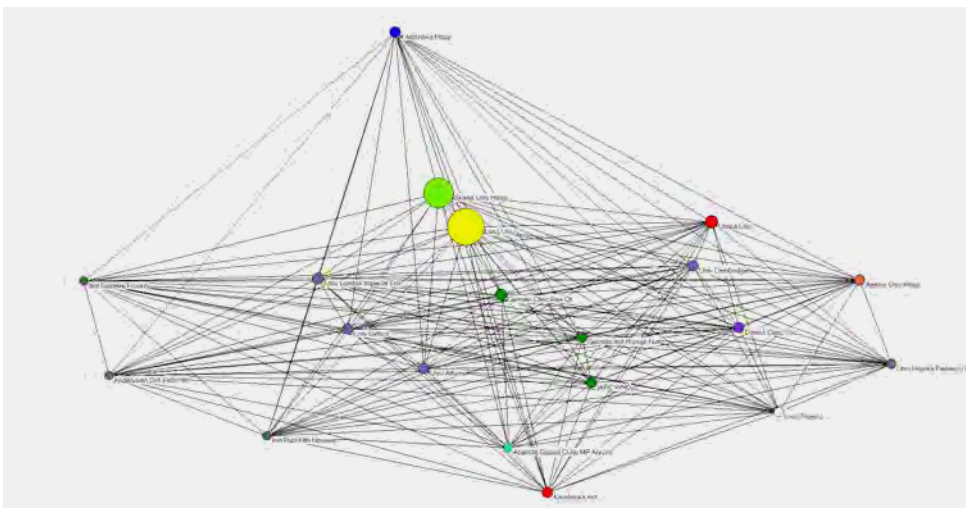


Figure 3a. Organization level co-authorship for LU authors with SCS funding. The 20 most frequently occurring organizations: 80 papers or more. (Unique pairs: 189, Number of links: 12,954, Links/pair: 68.54)

The maps show substantial differences in the structure of the networks. While the number of pairs formed by the 20 organizations included in the analyses was almost twice as high for SCS funded papers, the number of links between these pairs was almost 10 times higher for the SCS papers than for the ones without SCS funding. Both the number of links per se; and the distribution of links per pairs in the analysis, was substantially higher for scientists with SCS funding than for those without. Another difference was the higher frequencies for both non-Nordic organizations as well as organizational types that are not universities or hospitals for scientists with SCS funding; while the share of hospitals and Swedish collaboration partners was higher for the non-SCS papers (Table 6).

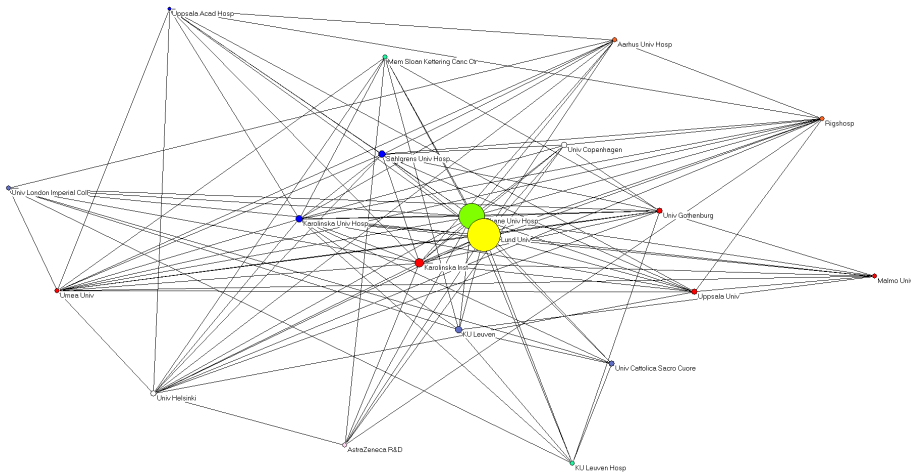


Figure 3b. Organization level co-authorship for LU authors without SCS funding. The 20 most frequently occurring organizations: 20 papers or more. (Unique pairs: 112; Number of links: 1,346; Links/pair: 12.02)

Table 6. Types of organizations in co-authorship maps (Figure 2a&b).

	With SCS funding	*Without SCS funding*
Swedish University	3	6
Swedish Hospital	2	4
Swedish Other Org.	0	1
Nordic University	1	2
Nordic Hospital	1	2
Nordic Other Org.	1	0
Non-Nordic University	6	3
Non-Nordic Hospital	1	2
Non-Nordic Other Org.	5	0

When analysing research collaboration on the organizational level, the differences between men and women, as well as between scientists with pre-clinical or clinical/combined positions, seemed to be even smaller than in the analyses on author level, whereas the differences between cancer researchers with or without SCS funding became even clearer. The collaboration networks for scientists with SCS funding were larger, more international and with higher frequencies of organizations outside academia and the hospital sector; and the density of the networks also seemed higher, with stronger links between the different organizations taking part of the research.

Conclusions

We aimed to study relations between on one hand: gender, research funding and research orientation; and on the other: the formation of research collaboration networks through a micro-level analysis of cancer research scientists at Lund University (LU). The frequency and distribution of contributing authors were investigated both on individual and organizational level and included analyses of co-authorship structures were also analysed.

An initial analysis of the number of LU authors per analytical category and the distribution of papers per LU authors, showed a significant difference between authors with or without funding from the Swedish Cancer Society (SCS), as well as between men and women and SCS funded authors with either a pre-clinical or a clinical/combined type of position. However, when analysing the research collaboration networks, the differences between men and women, as well as between clinical/combined and pre-clinical researchers, became less substantial, while the differences between scientists with or without SCS funding remained or became even more substantial.

That research collaboration networks are larger and more densely populated for scientists with SCS funding was expected and supports previous findings by e.g. Clark & Llorens (2012), Haiqi (1997) and Heffner (1981). In the light of e.g. Alcaide et.al. (2009) and Wennerås & Wold (1997), however, we would expect to see larger differences between men and women. When looking at the distribution of papers per LU author, we found differences between men and women, but as we turned our attention to the collaboration networks and looking at men and women with or without SCS funding respectively, the differences between genders were relatively small. This supports the findings of Bordons et.al. (2003), who did not identify differences in productivity in relation to gender within different professional categories. In our study, women were younger than the men, which imply a shorter research career. Still, even though the results in this study suggests that funding has a stronger relation to collaboration networks than gender, there is obviously the matter of women getting access to e.g. higher academic positions and research funding, as discussed by Wennerås and Wold (1997).

In terms of clinically/combined and pre-clinically oriented scientists, differences seem – as previously mentioned – more related to funding from the SCS than orientation. However, there were a significantly larger amount of papers from clinically/combined scientists among the SCS funded ones; and the networks for clinically/combined authors also seemed larger than for pre-clinical researchers within both the SCS funded and non-SCS-funded papers, which is in line with findings by Satyanarayana and Ratnakar (1989).

In summary, we found differences between men and women as well as between pre-clinical and clinical scientists, with more pronounced results related to number of publications than size and type of research networks. We identified a tendency towards women having fewer publications, although this probably reflects women being at an earlier stage of their careers. Despite smaller networks

among women, they had denser networks with a higher average of links per pair. The most substantial differences, both when analysing number of publications and size and form of networks, however, are related to funding.

References

Abbasi, A., Hossain, L., Uddin, S. & Rasmussen, K.J.R. (2011). Evolutionary dynamics of scientific collaboration networks: Multi-levels and cross-time analysis. *Scientometrics*, 89, 687-710.

Alcaide, G.G., Calatayud, V.A., Valderrama Zurián, J.C. & Benavent, R.A. (2009). Participación de la mujer y redes de coautoría en las revistas españolas de Sociología. *Revista Española de Investigaciones Sociológicas*, 126, 253-166.

Birnholtz, J.P. (2007). *Journal of the American Society for Information Science and Technology*, 58(14), 2226-2239.

Bordons, M., Morillo, F., Fernández, M.T. & Gómez, I. (2003). One step further in the production of bibliometric indicators at the micro level: Differences by gender and professional category of scientists. *Scientometrics*, 57(2), 159-173.

Clark, B.Y & Llorens, J.J. (2012). Investments in scientific research: Examining the funding threshold effects on scientific collaboration and variation by academic discipline. *Policy Studies Journal*, 40(4), 698-729.

de Nooy, W., Mrvar, A. & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. New York: Cambridge University Press.

Haiqi, Z. (1997). More papers, more institutions, and more funding sources: Hot papers in biology from 1991 to 1993. *Journal of the American Society for Information Science*, 48(7), 662-666.

Hara, N., Solomon, P., Kim, S.-L. & Sonnenwald, D.H. (2003). An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information Science and Technology*, 54(10), 952-965.

Heffner, A.G. (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics*, 3(1), 5-12.

Jeong, S., Choi, J.Y. & Kim, J. (2011). The determinants of research collaboration modes: Exploring the effects of research and researcher characteristics on co-authorship. *Scientometrics*, 89, 967-983.

Kamada, T. & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(12), 7-15.

Katz, J.S. & Martin, B.R. (1997). What is research collaboration? *Research Policy*, 26, 1-18.

Kretschmer, H., Pudovkin, A. & Stegmann, J. (2012). Research evaluation. Pert II: Gender effects of evaluation: Are men more productive and more cited than women? *Scientometrics*, 93, 17-30.

Larivière, V., Vignola-Gagné, E., Villeneuve, C., Gélinas, P. & Gingras, Y. (2011). Sex differences in research funding, productivity and impact: An analysis of Québec university professors. *Scientometrics*, 87, 483-498.

Lee, S. & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35(5), 673-702.

Lewis, J.M., Ross, S. & Holden, T. (2012). The how and why of academic collaboration: Disciplinary differences and policy implications. *Higher Education*, 64, 693-708.

Luukkonen, T., Persson, O. & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology and Human Values*, 17(1), 101-126.

Luukkonen, T., Tijssen, R.J.W., Persson, O. & Sivertsen, G. (1992). The measurement of international scientific collaboration. *Scientometrics*, 28(1), 15-36.

Mählck, P. (2001). Mapping gender differences in scientific careers in social and bibliometric space. *Science, Technology and Human Values*, 26(2), 167-190.

Melin, G. & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36(3), 363-377.

Narin, F., Stevens, K. & Whitlow, E.S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics*, 21(3), 313-323.

Newman, M.E.J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America: PNAS*, 101, 5200-5205.

Persson, O. (2010). Are highly cited papers more international? *Scientometrics*, 83, 397-401.

Persson, O., Danell, R. & Schneider, J.W. (2009). How to use Bibexcel for various types of bibliometric analysis. In F. Åström et al (Eds.), *Celebrating Scholarly Communication Studies A Festschrift for Olle Persson at his 60th Birthday*. International Society for Scientometrics and Informetrics. Retrieved December 12, 2012 from: http://www8.umu.se/inforsk/Bibexcel/ollepersson60.pdf.

Satyanarayana, K. & Ratnakar, K.V. (1989). Authorship patterns in life sciences, preclinical basic and clinical research papers. *Scientometrics*, 17(3-4), 363-371.

Wennerås, C. & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature*, 387(6631), 341-343.

Zhao, D. (2010). Characteristics and impact of grant-funded research: A case study of the library and information science field. *Scientometrics*, 84, 293-306.

EVALUATING KNOWLEDGE PRODUCTION SYSTEMS: MULTIDISCIPLINARITY AND HETEROGENEITY IN HEALTH SCIENCES RESEARCH

Johanna E. Steenrod[1], Alla G. Khadka[2] and Abigail R. Stark[3]

[1]*jes189@pitt.edu*
University of Pittsburgh, Graduate School of Public Health, Pittsburgh, PA (United States)

[2]*asg38@pitt.edu*
University of Pittsburgh, Graduate School of Policy and International Affairs, Pittsburgh, PA (United States)

[3]*ars156@pitt.edu*
University of Pittsburgh, Graduate School of Policy and International Affairs, Pittsburgh, PA (United States)

Abstract
Accepted validity and reliability checking mechanisms exist to establish the soundness of research at the construct level of an individual study. However, there is no mechanism to analyze and compare the structural properties of knowledge production systems. Drawing upon Gibbons' et al. (1994) framework, we develop an instrument to test knowledge production systems for multidisciplinarity and heterogeneity. We introduce Citation Network Analysis (CNA) as a method for eliciting and mapping the knowledge production system over time. We employ CNA and centrality measures as a basis for operationalizing multidisciplinarity and heterogeneity as constructs and creating a knowledge production system evaluation instrument. While the instrument is generalizable to any knowledge production system, this framework is applied to evaluate the knowledge system encompassing access to primary health care services in rural United States. Our analysis shows that the knowledge system is incorporating a wider scope of disciplines over time, such as dentistry and mental health, suggesting an increase in multidisciplinarity. Measures of heterogeneity indicate that the knowledge system is becoming geographically concentrated but involving a wider group of organizations. Funding for research that stresses involvement of multiple stakeholders across settings will foster this trend to develop sustainable solutions for this disadvantaged population.

Keywords
knowledge production, network analysis, bibliometrics, primary care, rural health

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3) and Visualization and Science Mapping: Tools, Methods and Applications (Topic 8)

Introduction

In a knowledge society each epistemic community is expected to produce valid, practically applicable and socially accountable research (Machlup, 1962). Accepted validity and reliability checking mechanisms exist to establish the soundness of research at the construct level of an individual study (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). There is no mechanism, however, that enables us to analyze and compare properties of knowledge production systems. Assessing these properties across disciplines is instrumental for scholars and other entities concerned with the production and validity of social science research. In this study we aim to describe and test measures of knowledge systems to evaluate embedded properties that directly affect the quality of research that these systems produce.

Drawing on Gibbons' et al. (1994) Mode 2 model of knowledge production, we focus on *multidisciplinarity* and *heterogeneity* as two indicators that can be used to benchmark current research practices. Multidisciplinarity characterizes a system in which the production of knowledge entails mobilization of a range of theoretical perspectives and practical methodologies (Gibbons, et al., 1994, pp. 4-6; Hessels, 2008, p. 741). As social, political and economic environments become increasingly complex, practitioners and policymakers face issues that demand equally complex solutions, requiring expertise from a variety disciplines. Heterogeneity refers to a system that is diverse in terms of the skills and experience of individuals engaged in knowledge creation. It is marked by an increase in the number of organizations or sites that create high quality knowledge. Heterogeneity describes a process where not only universities are participating in knowledge production, but also organizations like think-tanks and government agencies, each bringing unique resources and connections to relevant stakeholders. The organizations are linked in a variety of ways through networks of communication (Gibbons, et al., 1994, p. 6). We operationalize multidisciplinarity and heterogeneity to quantitatively assess the variety of disciplines and organizations participate in the knowledge system and answer our main research questions:

Question 1: To what extent do cited sources and authors within this knowledge system represent multiple disciplines?

Question 2: To what extent do cited sources and authors within this knowledge system display heterogeneity?

To achieve this end, first we provide a brief overview of the knowledge production literature, beginning with the traditional knowledge production model by Robert Merton (1957, 1973). We then discuss Gibbons' et al., (1994) Mode 2 model of knowledge production with a focus on multidisciplinarity and heterogeneity, which are presented as normative criteria. Second, we introduce

Citation Network Analysis (CNA) as a method for eliciting and mapping the knowledge production system over time; the system is defined as a pool of all publications on a specific topic. We employ CNA and centrality measures to operationalize heterogeneity and multidisciplinarity as constructs and create a knowledge system evaluation instrument. Third, while the evaluation instrument we develop here is generalizable to any knowledge system, this framework is applied to evaluate the knowledge system encompassing access to primary health care services in rural United States

Theoretical Foundation

The knowledge system evaluation instrument created in this study is primarily informed by the Mode 2 model of Gibbons et al. (1994). To gain a fuller understanding of contemporary knowledge production systems and how they evolved to their present state, it is useful to begin our discussion with a historical perspective. Although the idea of studying knowledge dates back to the ancient Greeks, one of the founders of the modern sociology of knowledge is Robert Merton (1945, 1957, 1973). Merton's seminal thesis *Sociology of Knowledge* describes knowledge as a reflection of an existential realm, either social or cultural, and posits that knowledge manifests itself as beliefs, ideas, norms, science, and technology (1945). Merton further describes science as a type of "certified" knowledge, characterized by the methods used to obtain it. As such, the stock of knowledge derived from the methods and cultural values represent reputable knowledge (1973, p. 270). The cultural values, which Merton describes as "institutional imperatives," are universalism, communism, disinterestedness, and organized scepticism (1973, pp. 272-278). Merton's conception of knowledge production was reflective of knowledge creation practices as he saw it, and scholars consider this conception applicable from 1870 through the 1990s (Rip, 2005).

As a result of technological and information revolutions, Merton's version of research practices and values underwent a number of critical changes, particularly its creation and dissemination. Thomas Kuhn explains these changes in his pivotal work, *The Structure of Scientific Revolutions*, in which he contends that science and our perception of science is bounded by existing scientific paradigms (1962, p. 60). These paradigms undergo revolutions of change in response to profound shifts in the external environment, leading to new paradigms and bringing about new frameworks, research methods, and objectives. In this sense, science is responsive to the transformations of the external environment. In the 1980s Jean-Francois Lyotard (1984) observed that technological and economic shifts, forming in the mid twentieth century, were laying the foundation for significant changes in how knowledge is perceived, created and utilized. In line with this prediction, the knowledge production enterprise transformed from a traditional Mertonian mode to a new mode, defined by an interconnected environment with freer flow of information and a wider range of actors involved (Knorr-Cetina,

2007). Gibbons' et al. (1994) framework, presented in *The New Production of Knowledge,* captures the elements of the new knowledge production paradigm, termed Mode 2 (Nowotny, Scott, & Gibbons, 2001; Rip, 2005; P. Scott et al., 1994).

In a follow-up publication, Nowotny, Scott, and Gibbons (2003) identify five properties of a contemporary knowledge production system: (1) Knowledge is created within a *context of application*, meaning that the environment in which the scientific process of problem creation, methodology development, dissemination and use occurs is considered. (2) Knowledge is *transdisciplinary*, and production relies on mobilization of numerous perspectives and methodologies, versus the standard view of limited, incremental knowledge production within a single peer-reviewed discipline. (3) Knowledge is produced within and across a greater *diversity of sites*. (4) Knowledge is *reflexive*, meaning that research is no longer an objective, removed activity but rather is a dialogic process where end users and knowledge producers jointly consider relevant topics, methodologies, and dissemination for their environment. (5) Finally, knowledge production is subject to new forms of *quality control*.

Although Gibbons et al. (1994)/Nowotny et al. (2003) do not explicitly present their model as normative, it can be treated as such. First, the authors make the case that each of the attributes can be viewed as ideal properties of a contemporary knowledge production system. Second, each of the properties can be operationalized and tested. Drawing upon Nowotny's et al. (2003) framework, we develop an instrument that enables us to test a knowledge production system for multidisciplinarity and heterogeneity. While the Mode 2 model of knowledge production suggests that the system is *trans*disciplinary, we chose to evaluate the system for *multi*disciplinarity. While transdisciplinarity presupposes integration of different disciplines and formation of new research areas, we are particularly interested in evaluating a system for its openness to multiple perspectives and sources of information as offered by a variety of disciplines – a property which is captured by multidisciplinarity (Porter, Cohen, Roessner, & Perreault, 2007). Futhermore, we utilize heterogeneity to represent the diversity of sites geographically and by types of organizations involved. We now describe the background of how these constructs are methodologically evaluated.

Method: Citation Network Analysis

Evaluating multidisciplinarity and heterogeneity in knowledge systems requires observation of how knowledge producers create and communicate their knowledge. More specifically, we need to understand which of the knowledge producers are most influential and what organizations, sites, and disciplines those actors represent. We employ Citation Network Analysis (CNA) to elicit the knowledge production system and analyze structural properties of the citation network. CNA research arose from an integration of the fields of information

metrics, citation analysis, and social network analysis. We provide a brief overview of these three fields and how we use their combined result to develop our evaluation instrument.

The progression from information metrics through advances in citation analysis shaped the building blocks for citation network analysis. Information metrics is the mathematical and statistical study of information to understand patterns in documents (journal articles, web documents, etc.) (Milojević & Leydesdorff, 2012). The most basic form analyzes numbers of authors and publications to draw conclusions about scientific fields. Citation analysis builds on this concept by considering the cited sources in documents as well. Early studies by Lotka (1926) and Bernal (1939) included citation analysis concepts, but Garfield (1955; 1964) and de Solla Price (1965) later formalized it. Garfield introduced science citation indexing as a technical means to collect works citing a particular piece in order to understand the scope of science in a subject area. De Solla Price added temporality to this concept and pointed out the "immediacy factor" within citations that suggest "scientific research fronts." Further clarifying this point, Crane noted that innovative changes can be observed as citations of a new seminal work increase and displace previous works (1972). Another development, co-citation analysis, which measures the relationship between documents co-cited by multiple documents, furthered citation indexing by building intellectual connections between co-cited papers and establishing the idea of network structure and clustering (Kessler, 1963; Small, 1973). White, Griffith and McCain (1981; 1998) began mapping author groups and established the importance of key authors and proximity of authors that span group boundaries. Crane also found that subject areas can be observed by dividing researchers into subgroups, connected by communications between highly productive leaders in the groups, forming a network, or "invisible college" (Crane, 1972).

Over the past decade, citation analysis became closely connected with social network analysis, where citation systems were viewed and analyzed explicitly as networks. Social network analysis, in its most basic form, is the study of relationships (links) between entities (nodes) that form networks and has been used extensively in many fields, including health (Freeman, 2004; J. Scott & Carrington, 2011; Valente, 2010; Wasserman & Faust, 1994). The nodes for citation networks are typically authors, documents, and sources while the link is the citation, symbolic of knowledge transfer. Analyzing the links allows us to understand the network as a whole and to quantify communication and influence between actors in the knowledge network versus the descriptive approach used in citation analysis (Crane, 1972). The primary focus in this study is direct citation networks versus co-occurrence citation networks and co-citation networks (Belter, 2012; Weingart, Guo, & Börner, 2010).
We consider measures in social network analysis that allow us to quantify the patterns observed in citation analysis and operationalize our key indicators,

multidisciplinarity and heterogeneity. Many measures exist to assess how influential nodes are in their networks; we identify the most influential authors and journals based on two key measures of influence: in-degree centrality and betweenness centrality. Considering the nodes within a network, an intuitive measure is the degree to which nodes are connected to each other. A highly connected node has greater power for establishing and perpetuating ideas and providing leadership within the network. This concept is measured by degree centrality, or the number of connections to or from a node (Freeman, 1978). In-degree centrality is the measure of connections to a node, in this case the number of times an author or source was cited (J. Scott & Carrington, 2011). High levels of citation of authors located at different sites and across disciplines would suggest that the knowledge production system receives greater influence from these sites and disciplines, indicating multidisciplinarity and heterogeneity.

Table 1: Knowledge Production System Evaluation Instrument (The Scale is for Top 20 Nodes within the Author-Author and Source-Source Networks by Centrality Measures)

Indicators	*Citation Network*	*Scale*
Multidisciplinarity	Author-author network (in-degree and betweenness centrality): gauging ties between the authors' disciplines	Low (≤ 7 disciplines); Med (8-13 disciplines); High (≥ 14 disciplines)
	Source-source network (in-degree centrality): gauging ties between the sources' disciplinary focus	Low (≤ 7 disciplines); Med (8-13 disciplines); High (≥ 14 disciplines)
Heterogeneity	Author-author network (in-degree and betweenness centrality): assessing *regional concentration* by U.S. census regions (9) and international (1)	Low (≤ 3 regions); Med (4-6 regions); High (7-10 regions)
	Author-author network (in-degree centrality): gauging ties between *types of institutions*, e.g., academic, government, think tanks, non-profits	Low (≤ 4 organizations); Med (5-6 organizations); High (≥ 7 organizations)

While in-degree centrality indicates those with the potential to influence, we also consider nodes that connect otherwise disconnected clusters, known as "structural holes" (Burt, 1997). A limitation of in-degree centrality is that it only measures a node's number of connections without considering the importance of these connections. In this sense, a node may have high in-degree centrality within a closed group, limiting its sphere of influence. Nodes that bridge unconnected groups are influential because they have a favourable position to disseminate/access information or resources of the network (Burt, 1997, 1999; Chen et al., 2009). Betweenness centrality is generally thought of as a way to measure these connections and control over communication (Chen, et al., 2009; Freeman, 1978). Leydesdorff (2007) also suggested that betweenness centrality could be used to measure the interdisciplinary nature of a citation network. We, therefore, will consider authors' betweenness centrality as another means to determine the flow of information between disciplines. If authors from many

disciplines are advancing ideas across the network, it will improve multidisciplinarity.

Knowledge Production System Evaluation Instrument

In this section, we present our Knowledge Production Evaluation Index, consisting of two indicators to capture properties of the normative knowledge production model. Each property is assigned measurement scales that we used to evaluate how the system performs.

Knowledge System: Access to Primary Care for Rural United States Populations

We use our instrument to evaluate research in health sciences in this study. Considering the massive breadth of the health sciences field and the variation of research within the field, we focus on the topic of the provision of primary care health services to rural populations in the United States. Primary care is generally the main point of contact between an individual and the health system where clinicians provide a wide variety of services from screening through monitoring across a large range of physical and mental health conditions. The World Health Organization, Healthy People 2020, the United States Institute of Medicine and all leading health organizations agree that access to primary care is of critical importance for the health of populations and many practice and policy interventions have occurred to promote access. Rural populations, which account for 20% of the United States population, have a continuing shortage of primary care providers: half the per capita amount compared to urban populations. Consequently, the health status and health behaviours of rural populations are significantly worse than their urban and suburban counterparts. Given the persistence of this issue and the national shortage of primary care physicians overall, it is of particular interest to explore innovative, multidisciplinary approaches arising from a diversity of sites. Utilizing this topic's stock of knowledge, collected from all publications from 1993 to present, enables us to evaluate the knowledge production system over time and answer key research questions:

Citation Network Analysis of the Knowledge System

To find all potentially relevant publications, we identified all key terms necessary for the search on access to primary health care in rural areas. Using the identified search terms, we performed a search in PubMed and Elsevier's Scopus database. These two databases were selected due to their inclusion of clinical, health sciences, and social sciences journals (Falagas, Pitsouni, Malietzis, & Pappas, 2008). We exported 8,277 resulting publications from the databases into an Excel spreadsheet and reviewed all publications for relevancy. We first removed all duplicate publications, amounting to 2,665. We then reviewed the title, abstract or full text article, as necessary, to determine inclusion or exclusion based on exclusion criteria, as presented in **Figure 1**.

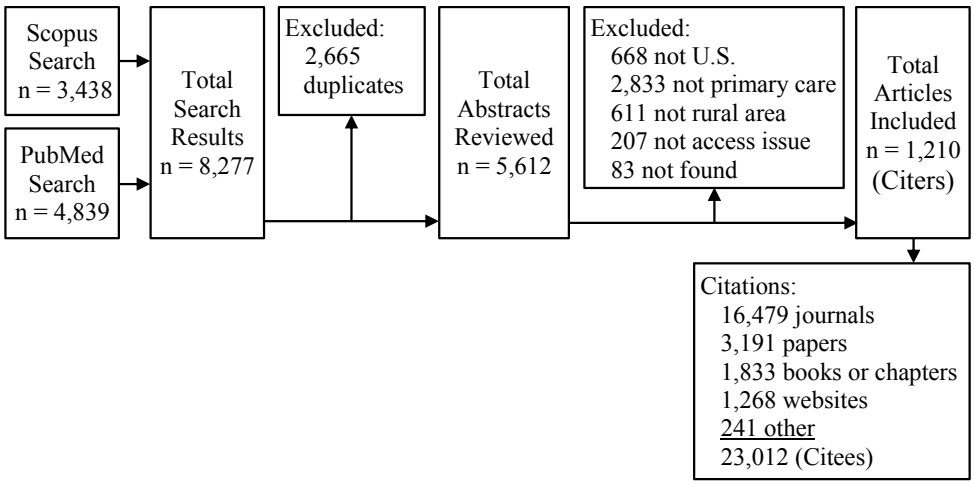


Figure 1: Literature Review and Citation Extraction Results

The first author (JS) reviewed all publications and 30% were double reviewed by the other two authors. This process resulted in 1,210 relevant publications, hereafter referred to as the "citers." The next step was a data generation and cleaning process to ready a "Master File" for CNA. We created a Scopus search for all citer publications and exported a citation Excel file with each citer's detailed information and all "citee" information. From this file we conducted analyses on the citers themselves to analyse publication trends, top citer authors and top citer sources. We obtained information on individual authors from internet searches.

We then imported the Citation Master File into ORA software (Carley, Pfeffer, Reminga, Storrick, & Columbus, 2012). We included two networks in the analysis, author-author and source-source, to elicit the knowledge production network. For this analysis, we only considered first authors as the primary source of knowledge production. Both of these networks were modelled as agent-agent networks because authors and sources appear both as citers and citees. We broke the networks into two subnetworks to analyse the effects of time: 1996-2003 (Period A) and 2004-2012 (Period B). We analysed all six networks as a whole and computed overall network statistics with ORA software. Finally, we analysed in-degree centrality and betweenness centrality and mapped network representations.

Results

We begin with general networks statistics to define the size and character of the networks that we analysed. The author-author network includes 10,696 unique author nodes with 18,639 citation links between them. The source-source network includes 3,944 unique source nodes with 11,298 citation links between them. We

will not explore all of the network statistics displayed in **Table 2** but note that the source-source network, as expected, has a smaller path length, a higher level of connectedness, and a higher level of centralization.

Table 2: Knowledge Production Network Statistics

	Author-Author			*Source-Source*		
Measure	*Whole 1996 - 2012*	*Period A 1996 - 2003*	*Period B 2004 - 2012*	*Whole 1996 - 2012*	*Period A 1996 - 2003*	*Period B 2004 - 2012*
Node count	10,696	4,550	7,431	3,944	1,901	2,815
Link count	18,639	6,730	12,056	11,298	4,252	7,826
Density	0.0002	0.0003	0.0002	0.0007	0.0012	0.0010
Characteristic path length	6.103	5.101	6.869	3.658	4.113	3.771
Network fragmentation	0.007	0.010	0.012	0.000	0.000	0.000
Degree centralization	0.001	0.001	0.001	0.002	0.003	0.003
Betweenness centralization	0.006	0.005	0.005	0.015	0.019	0.014
Closeness centralization	0.000	0.000	0.000	0.000	0.000	0.000
Eigenvector centralization	0.818	0.723	0.956	1.130	1.161	1.081

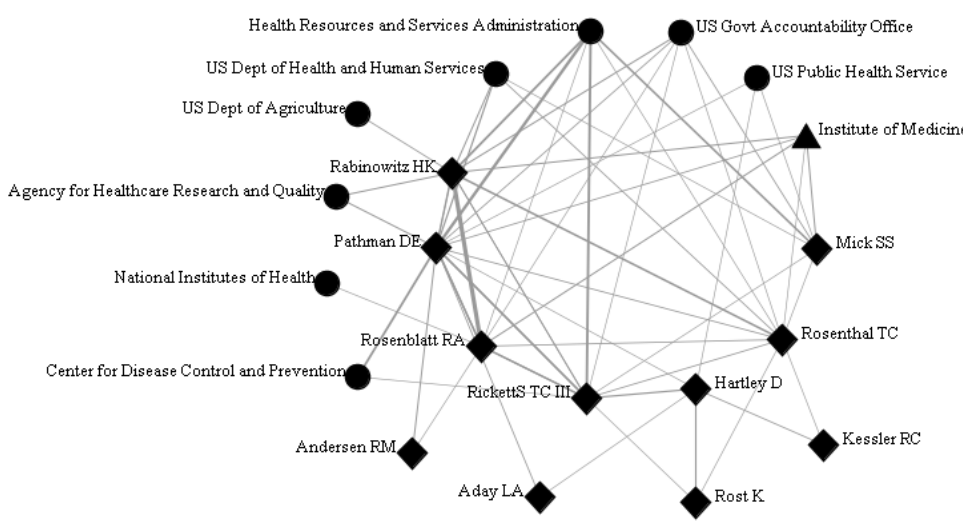


Figure 2: Knowledge Production Top 20 In-Degree Centrality Authors Network Diagram, 1996 – 2012 (Circles: Government Agency, Diamond: Academic, Triangle: Nonprofit Organization)

The analysis of multidisciplinarity within the author-author and source-source networks revealed an increasing trend. For the author-author in-degree centrality network, the top 20 authors are shown in **Figure 2**. This figure indicates that the top authors are very connected to each other, although a definitive "inner ring" of 12 highly interconnected authors shows. The main disciplines within this ring are

family medicine, health policy and management, and health services. The periphery authors, mainly connected to non-core authors, represent the disciplines of epidemiology, mental health, sociology, and health disparities research.

As shown in **Table 3**, the author-author in-degree centrality network increases in number of disciplines between periods A and B. This was due to a decreased concentration in family medicine, health services and sociology, and the introduction of new disciplines, such as health disparities, cancer and dentistry. The count measure moved the knowledge system from a "**low**" to "**medium**" in terms of multidisciplinarity. To reinforce this measure, we also observe that the average, maximum and minimum percentages of unscaled centrality by discipline decreased as expected. For example, the highest concentration by centrality in period A was family medicine at 30.5%, which changed to health services in period B at 22.4%. The number of disciplines by author-author betweenness centrality score decreased from 11 to nine, still within the "**medium**" multidisciplinary range. In period A, this network included mainly family medicine and health policy and management, as well as less conventional disciplines like gerontology, occupational science, economics, and geography. In period B, the latter disciplines were replaced and the network was further concentrated within family medicine, health policy and management, nursing, dentistry, and mental health. We also observe a reversal between health policy and management and family medicine as the top discipline by betweenness centrality. The source-source in-degree centrality network represents a wide range of disciplines, which increases from "**medium**" to "**high**" between periods A and B. The network showed little change in the top disciplines represented, namely medicine, public health, and general health services. The change between periods was due to a reduction in citations of public health reports and primary care journals and an increase in cancer, diabetes, and dentistry related journals. Overall, the results suggest that the knowledge production system has increased in multidisciplinarity.

Table 3: Knowledge Production System Multidisciplinarity Measures (Top 20 Nodes: Count of Disciplines and Percentages of Unscaled Centrality by Discipline)

	Period A: 1996-2003				*Period B: 2004 -2012*			
Network	*Count*	*Avg.*	*Max*	*Min*	*Count*	*Avg.*	*Max*	*Min*
Author-Author, In-degree	7 (Low)	14.3%	30.5%	4.8%	10 (Med)	10.0%	22.4%	3.2%
Author-Author, Betweenness	11 (Med)	9.1%	40.3%	1.8%	9 (Med)	11.1%	38.3%	1.6%
Source-Source, In-degree	12 (Med)	8.3%	27.4%	2.1%	14 (High)	7.1%	26.7%	2.7%

The analysis of the top 20 author-author centrality networks indicates a decrease in regional heterogeneity but an increase in institutional type heterogeneity. As shown in **Table 4**, the in-degree centrality network decreased from five to four census regions, both within the "**medium**" range. The South Atlantic region

dominated both periods with 64.2% of citations in period A and 76.7% in period B. This region includes government agencies in Washington, DC and Maryland, as well as, the Cecil G. Sheps Center for Health Services Research at the University of North Carolina, which has a long standing program dedicated to rural health research. The second region is the Pacific, home of the University of Washington WWAMI Rural Health Research Center, which has also been working for several decades to address rural workforce concerns. The change between period A and B is due to the elimination of the East and West North Central regions and the introduction of the New England region, represented by Harvard University and the University of Southern Maine, which houses the Maine Rural Health Research Center.

Table 4: Knowledge Production Heterogeneity Measures, Region (Top 20 Nodes: Count of Regions and Percentages of Unscaled Centrality by Region)

	Period A: 1996-2003				*Period B: 2004 -2012*			
Network	*Count*	*Avg.*	*Max*	*Min*	*Count*	*Avg.*	*Max*	*Min*
Author-Author, In-degree	5 (Med)	20.0%	64.2%	6.4%	4 (Med)	25.0%	76.7%	6.5%
Author-Author, Betweenness	7 (High)	14.3%	47.9%	1.8%	5 (Med)	20.0%	50.8%	5.1%

Table 5: Knowledge Production System Heterogeneity Measures, Institution Type (Top 20 Nodes: Count of Institution Types and Percentage of Unscaled Centrality by Institution Type)

	Period A: 1996-2003		*Period B: 2004 -2012*	
Institution Type	*Count*	*Concentration*	*Count*	*Concentration*
Academic Institution	13	60.1% (High)	11	46.0% (Med)
Government Agency	6	37.2% (Med)	7	46.2% (Med)
Non-profit Organization	1	2.6% (Low)	2	7.8% (Low)

Both periods have three types of organizations, ranking heterogeneity by types as "**low**." The centrality of government agencies and non-profit organizations in the author-author in-degree network, however, increased in terms of number and rank within the top 20. Government agencies appearing in both periods include the Centers for Disease Control and Prevention, the Health Resources and Services Administration, and the U.S. Department of Health and Human Services. Three government agencies from period A were replaced in period B by the U.S. Department of Agriculture, the Agency for Healthcare Research and Quality and the National Institute of Dental and Craniofacial Research. We also note that the American Academy of Family Physicians non-profit organization from period A was replaced with the American Cancer Society and the Institute of Medicine in period B. Overall, the results for heterogeneity were mixed.

Conclusion

The analysis of the knowledge production system on rural access to primary care yields findings that are informative to the health sciences research community as they further the research agenda. First, the system displayed an increase in multidisciplinarity from both the author-author and source-source in-degree networks. This suggests that while the knowledge system consistently relies on its core disciplines, providing a level of stability, it is open to the influence of ideas from new disciplines, including dentistry, mental health, and chronic disease specialists that can become influential players (Freeman, 1978). Second, while the betweenness centrality measure remained consistent in terms of multidisciplinarity, the change in the top disciplines shows greater alignment with key health care workforces, including family medicine, nursing, dentistry, and mental health. This can promote better communication of ideas and innovations between groups and across structural holes (Burt, 1999). Considering that multidisciplinarity is characterized by multiple perspectives, we can conclude that current research on rural access to primary care benefits from a variety of perspectives and approaches. The measures of heterogeneity suggest that government agencies and other institutions in the South Atlantic region are becoming increasingly influential in the knowledge system. The continued presence of rural health research centers suggests that these institutions have been and continue to be influential in this field. Increased heterogeneity by institution type suggests that as more organizations, including government non-profits, are participating in knowledge production, they may bring new criteria of quality control, enhancing valid research that satisfies not only academic standards but also standards put forward by practitioners. We conclude that focus on and funding for multidisciplinary studies across sites and institution types can enhance the ability of the knowledge system to develop innovative solutions to the ongoing public health issue of rural health care access.

This analysis introduced a unique knowledge production system evaluation instrument, which we believe successfully characterized the multidisciplinarity and heterogeneity of the rural access to primary care knowledge system. The instrument can be used and tested in other studies of knowledge systems to build the normative case for these properties within the framework of Gibbons' et al. (1994) Mode 2 model. Comparing measures of these properties across multiple knowledge systems will also inform the research community of differences between systems. Reiterating our opening point, this assessment will allow us to evaluate embedded properties that directly affect the quality of research that these systems produce.

A limitation of this study is that the articles selected are not based on a full systematic review of all relevant databases. We limited our search to PubMed and Scopus, which we posit represent the majority of relevant medical and social science sources but may exclude some sources. We also limited the author-author

analysis to first authors constraints, which may underestimate the influence of prominent second and last authors. Finally, this analysis relies on the primary appointment of the authors to determine topical focus and does not consider the content of their work, which may be broader ranging. Future work on this subject will expand evaluation instrument to consider other knowledge production characteristics, such as reflexivity. The heterogeneity could also be better understood through mapping the network geographically. We will also conduct semantic network analysis through content analysis of the titles and abstracts to better understand the information flowing within the network (Borgatti & Foster, 2003; van der Gaag & Snijders, 2004). This will allow us to analyze the context of application within the network. Finally, to better understand temporal changes in the network beyond the approach summarized in this paper, we propose considering the evolution of the network as a scale-free network and conducting eco-evolutionary network analysis (Albert & Barabási, 2002).

References

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47-97.

Belter, C. (2012). Visualizing Networks of Scientific Research. *Online, 36*(3), 1-6

Bernal, J. D. (1939). *The Social Function of Science*. London: Routledge & Kegan Ltd.

Borgatti, S. P., & Foster, P. C. (2003). The Network Paradigm in Organizational Research: A Review and Typology. *Journal of Management, 29*(6), 991-1013.

Burt, R. S. (1997). The Contingent Value of Social Capital. *Administrative Science Quarterly, 42*(2), 339-365.

Burt, R. S. (1999). The Social Capital of Opinion Leaders. *Annals of the American Academy of Political and Social Science, 566*(1), 37-54.

Carley, K., Pfeffer, J., Reminga, J., Storrick, J., & Columbus, D. (2012). ORA User's Guide 2012, Technical Report, CMU-ISR-12-105. Carnegie Mellon University, School of Computer Science, Institute for Software Research.

Chen, C. M., Chen, Y., Horowitz, M., Hou, H. Y., Liu, Z. Y., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics, 3*(3), 191-209.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design & analysis issues for field settings*. Chicago: Rand McNally College Pub. Co.

Crane, D. (1972). *Invisible colleges; diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.

De Solla Price, D. J. (1965). Networks of Scientific Papers: The Pattern of Bibliographic References Indicates the Nature of the Scientific Research Front. *Science, 149*(3683), 510-515.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and Weaknesses. *Federation of American Societies for Experimental Biology Journal, 22*(2), 338-342.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215-239.
Freeman, L. C. (2004). *The development of social network analysis: a study in the sociology of science*. North Charleston, S.C: Empirical Press.
Garfield, E. (1955). Citation Indexes for Science. *Science, 122*(3159), 108-111.
Garfield, E., Sher, I. H., & J.Torpie, R. (1964). *The use of citation data in writing the history of science.* Philadelphia, PA: United States Air Force Office of Scientific Research. Institute for Scientific Information.
Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. London: Sage.
Hessels, L. K., and Harro van Lente. (2008). Re-thinking new knowledge production: A literature review and a research agenda. *Research Policy, 37*(4), 740-760.
Kessler, M. M. (1963). Bibliographic Coupling Between Scientific Papers. *American Documentation, 14*(1), 10-25.
Knorr-Cetina, K. (2007). Culture in Global Knowledge Societies: Knowledge Cultures and Epistemic Cultures. *Interdisciplinary Science Reviews, 32*(4), 361-375.
Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology, 58*(9), 1303-1319.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences 16*(12), 317-324.
Lyotard, J. F. o. (1984). *The Postmodern Condition: A Report on Knowledge*. Vol. 10. Minneapolis, Minnesota: University of Minnesota Press.
Machlup, F. (1962). *The Production and Distribution of Knowledge in the United States*. Princeton, NJ: Princeton University Press.
Merton, R. K. (1945). Sociology of Knowledge. In G. Gurvitch & W. E. Moore (Eds.). New York: NY: Philosophical Library.
Merton, R. K. (1957). *Social theory and social structure*. Glencoe, IL: Free Press.
Merton, R. K. (1973). The Normative Structure of Science. In N. W. Storer (Ed.), *The Sociology of Science* (pp. 267-285). Chicago: The University of Chicago Press.
Milojević, S., & Leydesdorff, L. (2012). Information metrics (iMetrics): a research specialty with a socio-cognitive identity? *Scientometrics*, 1-17.
Nowotny, H., Scott, P., & Gibbons, M. (2001). *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty*. New York, NY: Sage Publications Ltd.
Nowotny, H., Scott, P., & Gibbons, M. (2003). Mode 2 Revisited: The New Production of Knowledge - Introduction. *Minerva, 41*(3), 179-194.

Porter, A. L., Cohen, A. S., Roessner, J. D., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*(72), 117-147.
Rip, A. (2005). *Rethinking scientific research: a dynamic perspective*. Paper presented at the Six Countries Programme Conference on The Future of Research: New players, roles and strategies.
Scott, J., & Carrington, P. J. (2011). *The SAGE handbook of social network analysis*. Thousand Oaks, Calif: SAGE.
Scott, P., Gibbons, M., Nowotny, H., Limoges, C., Trow, M., & Scwartzman, S. (1994). *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies* Sage.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
Small, H. G. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *American Society for Information Science, Journal, 24*(4), 265-269.
Valente, T. W. (2010). *Social networks and health: models, methods, and applications*. Oxford: Oxford University Press.
van der Gaag, M., & Snijders, T. (2004). Proposals for the Measurement of Individual Social Capital. In H. D. Flap & B. V lker (Eds.), *Creation and returns of social capital: a new research program* (Vol. 9). New York: Routledge.
Wasserman, S., & Faust, K. (1994). *Social network analysis: methods and applications* (Vol. 8). Cambridge: Cambridge University Press.
Weingart, S., Guo, H., & Börner, K. (2010). Science of Science (Sci2) Tool User Manual, Version Alpha 3
White, H. D., & Griffith, B. C. (1981). Author Cocitation: A Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science (pre-1986), 32*(3), 163-171.
White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327-355.

EVALUATING THE WEB RESEARCH DISSEMINATION OF EU ACADEMICS: A MULTI-DISCIPLINE OUTLINK ANALYSIS OF ONLINE CVS

Kayvan Kousha[1] and Mike Thelwall[2]

[1]*k.kousha@wlv.ac.uk*
University of Wolverhampton, School of Technology, Wulfruna Street, Wolverhampton WV1 1LY (United Kingdom)

[2]*m.thelwall@wlv.ac.uk*
University of Wolverhampton, School of Technology, Wulfruna Street, Wolverhampton WV1 1LY (United Kingdom)

Abstract
Online academic CVs are widely used to publicise qualifications and research achievements, including publications and some people exploit online CVs to find relevant publications by specific academic. It is not clear, however, how commonly academics use their web CVs to direct potential visitors to their research outputs. To assess this, we analysed outlinks from 2,700 web CVs or publication lists in four scientific fields and 15 European countries to examine the extent of linking to open access publications or to publicise research through other means, such as through publishers' websites. Just under half of the online CVs had at least one relevant outlink, and this was lower in public health (35%) and environmental engineering (44%) than astronomy (55%) and philosophy (54%). About a third of the online CVs had at least one outlink to an open access source (e.g., archives, repositories or self-archived PDF files), and this evidence for a kind of '*gold-green web presence*' was considerably higher in astronomy (48%) and philosophy (37%) than in environmental engineering (29%) and public health (21%). The overall findings suggest that, in practice, the majority of researchers are failing to promote to their research optimally, a serious issue for EU online research dissemination.

Conference Topic
Webometrics (Topic 7) and Open Access and Scientometrics (Topic 10).

Introduction
The web has provided new opportunities for academics to publicise and disseminate their research results, helping to facilitate public access to many types of scholarly information. In particular, the open access (OA) publishing of scholarly documents (e.g., journal and conference papers, research reports) through institutional or personal homepages or CVs may help to share scientific information and increase the visibility and impact of research. In contrast, scientific information that is not freely available or which is difficult to find or

afford by academic institutions may hinder future work and may also lead to the duplication of research. Hence, it has been claimed that "Europe is losing almost 50% of the potential return on its research investment until research funders and institutions mandate that all research findings must be made freely accessible to all would-be users, webwide" (Harnad, 2006, p. 12).

Although there are many ways to share research, one logical method for a scientist is to list their publications in an online CV and to embed hyperlinks to online copies of them. Whilst there is much research on OA publishing, this has focused on individual articles or theoretical discussions (see below) rather than online CVs as a dissemination mechanism. This seems to be a serious omission, given the apparently widespread use of online CVs for academics.

Whilst it is impractical to manually analyse the contents of online CVs or publication lists on a large scale to determine how they are used for research dissemination, an outlink analysis of web CVs could help to practically understand the extent of hyperlinking to research texts by scientists. To help fill this gap, the main objective of this paper is to examine how effectively researchers in different scientific fields across European use web CVs or publication lists to give access to their research outputs. Using a web crawler, we automatically generated different types of outlink statistics for evidence of (a) OA publishing, (b) publicising research through publishers' websites, (c) sharing research through social networking tools and (d) distributing scholarly-related multimedia.

Background

Online CVs for scientometrics research

Dietz, Chompalov, Bozeman, Lane and Park (2000) conducted one of the earliest scientometric studies on academic careers and the productivity of scientists and engineers based on coding CVs from *U.S. Department of Energy (DOE)* and *National Science Foundation (NSF)* funded projects, finding a relationship between age, subject areas, and early publication productivity for overall publication output. Since then, the partial shift of many academic CVs and publication lists to the web has made them a more easily accessible source of information for scholarly communication research. Many studies have used academic CVs for scientometric analysis and research assessment because web CVs or publication lists can include multiple publication types (e.g., books, technical reports, preprints) which may not be indexed in major scientific databases. Moreover, CVs may include additional information about grants, awards, jobs, teaching and qualifications which may help scientometric studies. Although several projects have collected and standardised CVs, such as *Europass* (http://europass.cedefop.europa.eu, see also EURO-CV, 2008), many academics either may not have an online CV or publication list or may infrequently update them. A significant drawback for the practical use of CVs in scientometrics, however, is that manually locating online CVs may be difficult and time-

consuming and automatic capturing online CVs by crawlers may not be possible without some manual data cleaning.

Online CVs have been previously used as the main source of evidence for investigations of researcher mobility (e.g., Woolley & Turpin 2009; Cañibano, Otamendi & Solís, 2011), career impact (e.g., Gaughan, 2009; Cox et al., 2011), postdoctoral training and departmental prestige (Su, 2011), maps of scientific fields (Lepori & Probst, 2009) and the grant peer-review process based on applicants CVs (see Cañibano, Otamendi, & Andújar, 2009). Other studies have used a combination of CVs and traditional bibliometric data for research evaluation (e.g., Lee & Bozeman, 2005; Sabatier, Carrere & Mangematin, 2006; Sandström, 2009). Harnad, Carr, Brody and Oppenheim (2003) proposed using online CVs as a rich and cheap data source for the UK Research Assessment Exercise (RAE). Nevertheless, it seems that online CVs have not yet been used for assessing the research dissemination strategies of academics in any way.

OA and self-archiving

OA publishing (the gold road) and self-arching (the green road), have the goal of assisting users to access research. A study of over 10,000 journals revealed that about 10% were freely accessible online and over 90% let authors deposit a preprint or postprint of their articles online through personal homepages or institutional repositories (Harnad et al., 2008). A recent study also reported that the proportion of gold OA journals had risen from about 5% and 7% in 2008 and 2009 respectively to about 12% in 2011 (Van Noorden, 2012). Similarly, in a study of articles published in 1,837 peer reviewed journals in 2008, 8.5% were freely available online at the publishers' sites (gold OA) and an additional 12% could be found by commercial search engines (green OA or illegal copies), making the overall OA percentage about 20% (Björk et al., 2010). Hence, it seems that overall OA publishing and self-archiving is in the minority but is increasing over time.

Despite evidence of the citation advantage of OA publications in different subject areas (e.g., Lawrence, 2001; Antelman, 2004; Kurtz, 2004) and claims that self-archiving can "increase citation impact by a dramatic 50-250%" (Harnad, 2006, p. 12), extent to which individual authors link from online CVs to their publications is unknown. Nevertheless, several researchers have studied the potential of OA in scholarly communication based upon surveys of authors (e.g. Rowlands, Nicholas, & Huntington 2004; Nicholas & Rowlands, 2005), finding that most authors are willing to deposit copies of their articles online and are generally positive towards OA publishing. An international survey of 1,296 authors in different fields and countries from eight years ago showed that a small minority of authors would not (5%) and or would be reluctant (14%) to comply with a requirement from their employer or funder to self-archive their research, whereas 81% were willing to do so and about half of the respondents had self-archived at least one publication during the last three years in some way (Swan & Brown 2005).

OA has been recognised as a key issue by the European Commission (EC) to help the EU to take a leading role in research and innovation in the Horizon 2020 funding programme (European Commission, 2012a). An online survey of '*scientific information in the digital age*' by the EC with 1,140 responses received from 42 EU countries from research funding organisations, universities, libraries, publishers and individual researchers, found that 90% agreed that outcomes from publicly funded research should be freely accessible, whereas about 70% of publishers disagreed (European Commission, 2012b). Another survey of 811 project coordinators participating in an OA pilot in the *EU The Seventh Framework Programme (FP7)* showed that almost half of the respondents (42%) found it easy to have time or personnel to self-archive peer-reviewed articles (European Commission, 2012c). In Denmark a survey with 98 responses and 23 interviews found that more than half of the respondents used OA journal archives or subject repositories at least monthly (Houghton, Swan & Brown, 2011). Whilst most previous findings about OA publishing and self-archiving are based on surveys of scholars, they have not analysed how authors distribute research outcomes from their personal or institutional CVs, which is an important aspect of this issue.

Research questions

Based upon the research questions below, the main aim of this study is to assess how effectively scholars disseminate research in practice from their web CVs.

1. To what extent do EU researchers use web CVs or publication lists to disseminate their research outputs by linking to OA publications (e.g., OA repositories or self-archiving) or other scholarly web sources (e.g., publishers' websites)?
2. Are there significant differences in the extent and nature of outlinking from CVs between disciplines?

Methods

As a practical step, we restricted the sample to be investigated to researchers from four substantially differing subject areas: *astronomy and astrophysics, public environmental and occupational health, environmental engineering,* and *philosophy* and from one large advanced science system: that of the EU.

Research population

There is no definitive register of CVs or URLs for European researchers although there are some partial databases and some countries have comprehensive research information systems. Hence we used ad-hoc methods to generate a sufficiently large and broad sample of CVs from relevant researchers, with an emphasis on active researchers. We located URLs for 2,700 online CVs or publication lists based upon (1) an online email survey of authors publishing articles in relevant journals and (2) Google searches. For the survey, we extracted list of email addresses from published research papers indexed in the Thomson Reuters Web

of Science (WoS) during 2005-2011 in *astronomy & astrophysics, public environmental & occupational health, environmental engineering,* and *philosophy* (including history and philosophy of science). The email addresses were limited to the national domains of 15 EU countries (*Bulgaria, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Israel, Italy, the Netherlands, Poland, Slovenia, Spain,* and *the United Kingdom*) as a practical step (e.g., uk, de, nl, fr, es, ee).

Due to the low WoS coverage of philosophy we devised a program to exact extra emails from publications indexed by Scopus and also included emails ending with non-national domains such as com (e.g., gmail.com, yahoo.com, hotmail.com), org, and net and verified the national locations of the authors from Scopus.

Unfortunately, the response rate from the email survey was low in many countries, ranging from 11.4% in philosophy to 6.6% in environmental engineering. Thus, to increase the sample we collected extra URLs through nation-specific Google searches (e.g., `CV OR resume OR "curriculum vitae" OR "publication list" philosophy site:hu`) and randomly selecting web CVs and publication lists. This was time-consuming and needed extensive data cleaning. Using this method we increased the number of the CVs from 1,110 from the original survey to 2,700 (Table 1). This sample is likely to be biased in several ways. For example, it excludes researchers without a web presence or with a web presence that is difficult to find with Google and disproportionately includes authors that choose to respond to email surveys.

Table 1. General statistics about the CV URLs collected for the outlink analysis

Disciplines	*Email invitations*	*Total (%) responses*	*Response rate*	*CV URLs from the survey*	*CV URLs from Google*	*Total CV URLs*
Astronomy	6,635	528 (24.51)	7.96%	271	386	657
Pub. Health	7,277	534 (25.79)	7.34%	213	407	620
Environ. Eng.	8,686	573 (26.60)	6.60%	284	330	614
Philosophy	4,591	519 (24.09)	11.41%	342	467	809
Total	27,189	2,154 (100%)	7.92%	1,110	1,590	2,700

Webometric analysis of online CVs

Link analysis software was used to analyse outlinks from the 2,700 academics' CVs, homepages or online publication lists. In the reminder of the paper these are collectively referred to as online CVs. *SocSciBot* (socscibot.wlv.ac.uk) was used to crawl the CV URLs and to extract their hyperlinks, if any. Webometric Analyst (lexiurl.wlv.ac.uk) was used to automatically generate different types of outlink statistics for the links in the downloaded pages (e.g., see Thelwall, 2004). We classified the sources of outlinks from web CVs into four broad categories and several sub-classes (see below). This categorisation not only reflects a broad interpretation of web research dissemination (e.g., evidence of outlinking to OA

research), but also shows other specific use of websites (e.g., outlinks from CVs to online videos or blog posts).

Types of outlink analysed

Our broad classification of outlinks reflects the key issues of (1) OA publishing (2) publicising research through publishers' websites (3) communicating through social networking tools and (4) distributing scholarly-related multimedia. Note that the distribution of outlinks was highly skewed in all fields and countries. This is because most outlinks were created by a few prolific researchers and hence the mean was not an appropriate indicators. Nearly all medians were zero and so medians were also not helpful for this data. For this reason only the proportion of researchers with at least one outlink to the selected web sources was used throughout the investigation (for a list of the URLs and file types used see http://cybermetrics.wlv.ac.uk/paperdata/WebResearchDisseminationEUCVs.xlsx) .

Outlinks to OA research (gold-green web presence)

We used four sub-classes for OA publishing and self-archiving of European academics. We did not repeatedly count results that matched multiple sub-classes in broad categories. For instance, a single CV could have outlinks to an individual paper either in an OA sub-class or in PDF format, but we counted these results only once as evidence of at least one outlink to OA research. The lists below were compiled from various online sources as well as from checking commonly linked to URLs in the web CVs.

- Outlinks to major OA repositories: This uses a list of 26 major OA repositories, such as *ArXiv.org* in astronomy and astrophysics.
- National OA repositories or digital libraries: This covers about 770 national or university OA archives listed in the *Directory of Open Access Repositories* (http://opendoar.org/) from 15 European countries in the study (e.g., http://eprints.ucm.es).
- Document sharing sites: This includes sites which are commonly used to share documents online (e.g., PDF or DOC files) such as *DropBox.com* and *DocStoc.com*.
- Document file types: This sub-category covers the four common document file types that are commonly used for either publishing preprint/postprint papers or for other scholarly activities: PDF, Microsoft Word (e.g., file extensions doc, docx, rtf, dotx), presentation slides (e.g., Microsoft and Apple file extensions ppt, pptx and key) and spreadsheet and statistical files (e.g., extensions xls, xlsx, por and sav). We decided not to capture Postscript (PS) files in this study due to the huge amount of duplicate results with PDF files for individual records in several OA repositories, such as *Arxiv.org*.

Outlinks to publishers' websites (including DOI URLs)

This broad category includes outlinks from web CVs to about 70 fee-based publishers' websites (both journal and book publishers). Although this type of outlink doesn't directly indicate OA preprint or postprint publication, it suggests that researchers are publicising their research through linking to bibliographic information and abstract pages of articles or book reviews. Presumably in some cases the publisher prohibits OA archiving but in other cases the researcher may choose not to self-archive, even if it is allowed by the publisher. The academic may expect that some or all readers will have access to the fee-based website and the journal may also be partly or fully OA.

- Outlinks to major academic publishers (mostly journals): This includes a list of 45 major fee-based academic publishers' websites, including *Science Direct*, *Springer*, *IEEEXplore*, *Emerald Insight*, *Nature*, and *Oxford Journals*. We selected URLs of locations where abstracts of articles appear in the publishers' websites such as "journals.cambridge.org" instead of ".cambridge.org" to avoid counting irrelevant outlinks.
- Outlinks to major book publishers, online bookshops and databases: We created a list of 27 major online bookshops, book publishers and databases to assess the extent of publicising research book such as *Amazon.com*, *Google Books*, *National Academic Press*, *Routledge*, *MIT Press* and *Cambridge*. Again we used specific URLs for book sections of publishers, if possible, to prevent double counting of outlinks (e.g., .cambridge.org/gb/knowledge/textbooks and .taylorandfrancis.com/books).
- Outlinks to Digital Object Identifiers (DOIs): DOIs are commonly used to uniquely identify digital objects such as journal or conference papers. Our initial study showed that many academic CVs contain DOI-based hyperlinks that typically redirect to major publishers' websites.

Outlinks to major blogs and social networking sites

Blog posts, online reference mangers and general or professional social networking sites are emerging tools that can be potentially be used to publicise research. We used the sub-classes below to estimate how frequently these tools were linked to by academic CVs to or communicate about research. Although the last two classes are not predominantly for disseminating or discussing research but the first two may be.

- Outlinks to major blogs: We selected a list of about 50 major blogs in different languages based on the *Alexa* top global sites (http://alexa.com/topsites) and directories of blogs (e.g., http://en.wikipedia.org/wiki/List_of_blogs). This class includes not only the most popular blogging sites such as *Blogspot.com* and *WordPress.com* (with Alexa ranks of 12 and 20 respectively), but also

several blog sites in science and philosophy such as *ScienceBlogs.com* and *Philosophyetc.net*.

- Online reference managers: *Mendeley*, *CiteULike*, *Zotero* and *Connotea* allow academics to gather, organise and share bibliographic information about references with others over the Internet. We selected 13 online reference manager sites.
- Outlinks to general social networking sites: We selected about 20 general social networking sites, from *Facebook*, *Twitter* and *MySpace* to other popular European social network sites such as *vk.com*.
- Professional social networking sites: This includes several social networking sites such as *LinkedIn* and *Academia* which have mainly been developed for academic or professional audiences.

Outlinks to multimedia

Previous studies have shown that online videos are increasingly used for research communication for scientific experiments, academic presentations, lectures or artistic outputs (e.g., music and dance) (Kousha & Thelwall 2011; Kousha, Thelwall & Abdoli, 2012; Sugimoto, & Thelwall, in press). Moreover, academics may publicise their research through images (e.g., in visual arts and astronomy) or recorded speeches and lectures (e.g., audio files). We included the classes below to examine this issue.

- Outlinks to online video sharing sites: This used a list of 16 video sharing sites such as *YouTube*, *vimeo*, and *Google Videos* as well as several scholarly-related video databases such as *TedTalks*, *VideoLectures.net*, *PhilosophyTalk.org* and *ScienceStage.com*.
- Outlinks to online image sharing sites: This class includes general image sharing sites such as *Flickr* and *PhotoBucket* as well as scientific image databases in astronomy such as *NASA Images* and *Galaxies.com*.
- Video, audio and image file formats: This category covers outlinks to different multimedia file extensions for video (e.g., mp4, mov and avi), audio (e.g., mp3, wav and ram) and images (e.g., jpeg, gif and tiff).

Results

General outlinking patterns across fields

Table 2 gives an overview of the results. The third and fourth columns of the overall results show that just under half of the academics with online CVs in our sample had at least one outlink to the range of selected web sources or file types in the study, although this proportion was considerably lower in public health (35%) and environmental engineering (44%) than in astronomy (55%) and philosophy (54%). In philosophy and history and philosophy of science a higher proportion of academics tended to create links from their CVs to the file types investigated (about 35%), suggesting that philosophers are more willing to deposit preprint or postprint versions of their publications online through self-archiving

practices (e.g., PDF format). This variation across the subject areas is discussed in more detail below.

Table 2. Web CVs or publication lists with at least one outlink of any analysed type

Disciplines	**Total** *CVs crawled*	*Outlinks to selected websites*	*Outlinks to selected file types*	*Outlinks to any selected URL type*
Astronomy	657	312(47.5%)	186(28.3%)	359(54.6%)
Pub. Health	620	154(24.8%)	123(19.8%)	217(35.0%)
Envir. Eng.	614	182(29.6%)	172(28.0%)	273(44.5%)
Philosophy	809	326(40.3%)	286(35.4%)	439(54.3%)
Total	**2,700**	**974(36.0%)**	**767(28.4%)**	**1,288(47.7%)**

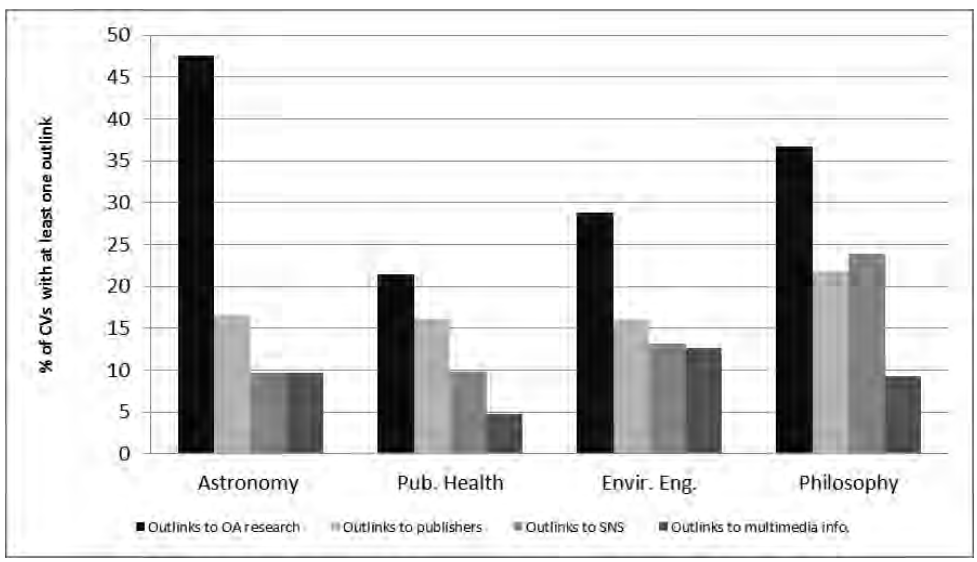


Figure 1. Broad classification of outlink sources from web CVs across subject areas in EU

Classification of outlink targets

Figure 1 reports a broad classification of outlink targets from web CVs and publication lists (see methods). The overall results show that 34% (920 of 2,700) of the online CVs had at least one outlink to OA sources, including OA repositories, digital archives, document file formats (e.g., PDF, doc, rtf), and document sharing sites. Hence, this broad category generally reflects evidence of what could be termed a '*gold-green web presence*' to share full-text research results. Astronomy and astrophysics (48%) and philosophy (37%) had the most outlinks to OA contents, whereas environmental engineering (29%) and public health (21%) had much fewer. One explanation for the difference might be that OA publishing has become the norm in astronomy and astrophysics (e.g., based

on ArXiv), but in public health and environmental engineering there is not an established culture of using specific OA repositories or relatively few publishers permit self-archiving. As shown in Figure 1, in philosophy about 22% and in three other science disciplines about 16% of the web CVs had at least one outlink to the major publishers' websites, academic databases or the DOI site, indicating the extent of using alternative methods to publicise research (e.g., paper abstracts or book reviews). In philosophy the proportion of academics with at least one outlink to social media sites (about 24%) is nearly twice as much as in the three science fields, suggesting that philosophers more commonly use social media tools and blogs in particular to share or discuss science (but see the limitations below). This may partially compensate for the lack of direct OA publishing, which may be particularly difficult for books.

Table 3. Number and percentage of web CVs with at least one outlink based on file types

Disciplines	Document files	Presentation files	Stat. files	Video files	Image files	Audio files	Archive files	Any file type
Astronomy	169 (25.7%)	6 (0.9%)	1 (0.2%)	4 (0.6%)	39 (5.9%)	5 (0.8%)	0 (0%)	186 (28.3%)
Pub. Health	119 (19.2%)	2 (0.3%)	0 (0%)	1 (0.2%)	7 (1.1%)	1 (0.2%)	1 (0.2%)	123 (19.8%)
Envir. Eng.	149 (24.3%)	7 (1.1%)	3 (0.5%)	3 (0.5%)	39 (6.4%)	0 (0%)	6 (1%)	172 (28%)
Philosophy	274 (33.9%)	8 (1%)	0 (0%)	2 (0.2%)	34 (4.2%)	15 (1.9%)	1 (0.1%)	286 (35.4%)
No. (%) of all CVs	**711 (26.3%)**	**23 (0.9%)**	**4 (0.1%)**	**10 (0.4%)**	**119 (4.4%)**	**21 (0.8%)**	**8 (0.3%)**	**767 (28.4%)**

Outlinked file types

Table 3 shows that just over a quarter of the CVs across all fields and EU countries had at least one outlink to a document file type, although this was much higher in philosophy (34%) and lower in public health (19%). Hence, it seems that philosophy researchers are more willing to share preprints/postprints of research through personal or institutional self-archiving, assuming that this is the nature of the linked-to files. Just under three-quarters (73%) of the links are to PDF files, 8% to Microsoft Word files, 2.5% to presentation files (e.g., ppt and pptx) and about 0.5% to Microsoft Excel files. This suggests that the vast majority (about 84%) of outlinks from CVs or publication lists were to either to scholarly publications (preprints or postprints) or other academic contents (e.g., scientific meetings or teaching presentations). Although about 14% of the outlinks were to image files (e.g., jpg, gif and tif), most seem not to be mentioned in the web CVs for scholarly reasons. For instance, our manual checking of 70 outlinks to image files showed that only 16% were created for scholarly reasons (e.g., images of galaxies and nebulas or scanned images of conference papers or posters) and the majority (84%) were from hyperlinked photos of the academics to larger size

photographs. In philosophy we found that about 2% (15) of the outlinks were to mp3 audio files and manual checking showed that all were digitised lectures, speeches and podcasts, suggesting that in the humanities, lectures, speeches and talks by academics may be particularly useful for disseminating and discussing research.

Common outlink types in each discipline

In astronomy and astrophysics the major targets of outlinks were important OA archives: the *ArXiv* e-prints archive and the *SAO/NASA Astrophysics Data System (ADS)* in particular. More than one third (37%) of astronomy and astrophysics academics had at least one outlink from their CV to one or both of these digital libraries, whereas less than 3% of the CVs outlinked to one or more of over 750 national or institutional e-prints archives, showing the importance of subject specific repositories over institutional archives.

In philosophy about 20% and 7% of CVs had at least one outlink to journal publishers' websites (including DOI links) and book publishers or online bookshops respectively, whereas in the three science fields this proportion is lower for journal publishers' websites (about 16%) and much smaller for book publishers (0.7%-1.4%). Thus, it seems that philosophers tend to often direct potential users to publishers' websites where paper abstracts or book reviews can be viewed. This is kind of linking is probably not as useful as linking to OA files, but may still help some visitors to download or buy publications and hence a CV with such links could be termed a blue web presence (in contrast to an OA gold-green web presence).

In public health the majority of outlinks to OA repositories were to *US National Library of Medicine (NLM)* databases and *BiomedCentral.com* (an online publisher of free peer-reviewed scientific articles). However, the *NLM* outlinks may be created either to full-text content in databases such as *PubMed Central* (a free full-text biomedicine archive) or to bibliographic information such as abstract pages in PubMed.

In contrast to the three other disciplines, in environmental engineering there was a relatively high proportion of online CVs with at least one outlink to national or institutional e-prints archives (7%) in comparison to major OA repositories (1%). For instance, there were outlinks to university/institutional digital libraries in Bulgaria (http://eprints.nbu.bg), Denmark (http://orbit.dtu.dk), France (http://irevues.inist.fr), the Netherlands (http://igitur-archive.library.uu.nl/), Italy (http://eprints.biblio.unitn.it), Finland (http://epublications.uef.fi), Germany (http://oops.uni-oldenburg.de) and the UK (e.g., http://eprints.brighton.ac.uk/, http://eprints.soton.ac.uk, http://strathprints.strath.ac.uk). Hence, it seems that researchers in environmental engineering tend to deposit their research through university or institutional OA archives rather than major OA repositories, although neither are used much.

In philosophy, *Springer* (9%), *Wiley* (4%) and *Amazon.com* (4%), in astronomy and astrophysics, *the Institute of Physics* (4%), in public health *Oxford* and *Wiley*

(both about 3%) and in environmental engineering *Elsevier* (about 4%) and *Springer* (3%) were the top publishers and databases in terms of the most outlinking CVs. *DOI* outlinks were also common, from about 4% in philosophy, astronomy and public health to 8.5% in environmental engineering.

Blogs may potentially be useful to discuss the academic's own research or others' research. In philosophy 4% and 3.5% of CVs had at least one outlink to two major blog sites, *WordPress.com* and *BlogSpot.com* (e.g., scientific meeting announcements: http://maureensie.wordpress.com or course lectures: http://logicforlanguage.blogspot.com). Philosophy was the only subject with a substantial amount of linking to blogs. Although the numbers are small, even for philosophy, this is consistent with the humanities having an orientation towards discussing research more informally than through publications. Nevertheless, other evidence shows that there is also a significant community of science bloggers (Shema, Bar-Ilan & Thelwall, 2012).

Outlinks to social media sites are difficult to interpret because they could come from predefined hyperlinked icons to university or department pages in Facebook, Twitter, MySpace, LinkedIn or YouTube and are thus problematic for web research publicity analysis. For instance, manual checking of 60 sampled outlinks to Facebook across fields revealed that all of them were from Facebook hyperlinked images embedded in CVs to universities, schools or departments' pages rather than to the academics' Facebook pages.

Conclusions

Outlink analysis of CVs: This study investigated some aspects of the web research dissemination strategies of academics based on crawling 2,700 CVs and/or publication lists across four fields and 15 EU countries. Just under half of the academics had at least one outlink to at least one recognised source of potentially scholarly information. Moreover, one third of the online CVs had at least one outlink to apparently OA repositories or OA documents (e.g., PDF and doc), a type of *gold-green web presence* to share full-text research and minority of the rest had links instead to non-OA publishers, termed here a *blue web presence*. Although the majority of researchers seem to have been willing to self-archive their research for a long time (Swan & Brown, 2005) and about 90% of journals let authors deposit a preprint or postprint of their papers online (Harnad et al., 2008), he majority of the authors in our study don't seem to have taken advantage of the possibilities for self-archiving, at least through their online CVs (see also the disciplinary differences). This seems alarming from the perspective of improving access to scientific information produced in Europe.

Outlink analyses of academic CVs or publication lists can help to assess the OA publishing and self-archiving behaviour of researchers, especially in a large scale investigation when a survey is not practical or authors' opinions may not fully reflect their actions. Document files and PDF files in particular were the most common, and links to these could therefore be used as an approximate indicator of the open accessibility of research for scientists. This confirms a previous study

which found that about 70% of the OA documents in four science and four social science disciplines citing research papers were in non-HTML format (PDF and DOC) (Kousha, 2009). Outlinks to OA archives, repositories and digital libraries can also be useful indicator for the web research publicity of academics.

Disciplinary differences: there were significant disciplinary differences for outlinking patterns in terms of the methods that scholars in different disciplines use for publicising their own research. For instance, in astronomy and astrophysics, with most outlinks to OA contents (about 50%), two OA repositories, ArXiv and ADS, clearly play a central role for research dissemination. However, in philosophy linking to PDF documents (e.g., self-archived preprint/postprint papers) seems to be more common than depositing research in OA archives. In both environmental engineering and public health, with only about 29% and 21% links to OA contents, outlinks to publishers websites tended to be more frequent. This disciplinary variation may be due to a lack of awareness or interest within some research communities about OA or other factors such as the structure of CVs (e.g., institutional templates or third party platforms such as Academia.edu), copyright issues with journals or different disciplinary communication needs.

Acknowledgements

This publication is part of the Seventh Framework Programme (FP7) EU-funded project *ACUMEN* (http://research-acumen.eu) on assessing the institutional web presence of researchers across EU.

References

Antelman, K. (2004). Do Open-Access articles have a greater research impact? *College & Research Libraries*, 65(5), 372-382. Retrieved January 13, 2013 from http://eprints.rclis.org/5463/1/do_open_access_CRL.pdf

Björk B-C, Welling P, Laakso M, Majlender P, Hedlund T, et al. (2010). Open access to the scientific journal literature: Situation 2009. *PLoS ONE*, 5(6), e11273.

Cañibano, C., Otamendi, F.J. & Solís, F. (2011). International temporary mobility of researchers: A cross-discipline study. *Scientometrics*, 89(2), 653-675.

Cañibano, C., Otamendi, J. & Andújar, I. (2009). An assessment of selection processes among candidates for public research grants: The case of the Ramón y Cajal Programme in Spain. *Research Evaluation*, 18 (2), 153-161.

Cox, M.F., Zhu, J., Ahn, B., London, J.S., Frazier, S., Torres-Ayala, A.T. & Guerra, R.C.C. (2011). Choices for Ph.D.s in engineering: Analyses of career paths in academia and industry. *ASEE Annual Conference and Exposition*, Conference Proceedings, 11-12.

Dietz, J., Chompolov, I., Bozeman, B., Lane, E. & Park, J. (2000). Using the curriculum vita to study the career paths of scientists and engineers. *Scientometrics*, 49(3), 419-442.

EURO-CV: Building new indicators for researchers' careers and mobility based on electronic curriculum (2008). Retrieved January 3, 2013 from: http://www.uam.es/docencia/degin/prime/webprime/documentos/EuroCV/EURO%20CV%20report.pdf

European Commission (2012a). *A Reinforced European Research Area Partnership for Excellence and Growth*. Retrieved January 5, 2013 from: http://ec.europa.eu/euraxess/pdf/research_policies/era-communication_en.pdf

European Commission (2012b). *Online survey on scientific information in the digital age*. Luxembourg: Publications Office. Retrieved January 5, 2013 from: http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf

European Commission (2012c). *Survey on open access in FP7*. Luxembourg: Publications Office of the European Union. Retrieved January 6, 2013 from: http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-open-access-in-fp7_en.pdf

Gaughan, M. (2009). Using the curriculum vitae for policy research: An evaluation of National Institutes of Health center and training support on career trajectories. *Research Evaluation*, 18(2), 117-124.

Harnad, S. (2006). Publish or perish – self-archive to flourish: the green route to open access, *ERCIM News*, 64, 12-13. Retrieved January 13, 2013 from http://eprints.soton.ac.uk/261715/1/harnad-ercim.pdf

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C. & Hilf, E. (2008). The access/impact problem and the green and gold roads to open access: an update. *Serials Review*, 34(1), 36-40.

Harnad, S., Carr, L., Brody, T. & Oppenheim, C. (2003). Mandated online RAE CVs linked to university eprint archives. *Ariadn*e 35. Retrieved December 18, 2012 from: http://www.ariadne.ac.uk/issue35/harnad

Houghton, J.W., Swan, A. & Brown, S. (2011*). Access to research and technical information in Denmark*. Report to The Danish Ministry of Science, Technology and Innovation (FI) and Denmark's Electronic Research Library (DEFF). Retrieved January 6, 2013 from: http://www.deff.dk/uploads/media/Access_to_Research_and_Technical_Information_in_Denmark.pdf

Kousha, K. (2009). Characteristics of open access web citation networks: A multidisciplinary study, *Aslib Proceedings*, 61(4), 394-406.

Kousha, K. & Thelwall. M. (2011). Motivations for citing Youtube videos in the academic publications: A contextual analysis. *17th International Conference on Science and Technology Indicators* (STI), 5-8 September, 2012 in Montreal, Quebec, Canada. Retrieved January 14, 2013 from: http://sticonference.org/Proceedings/vol2/Kousha_Motivations_488.pdf

Kousha, K., Thelwall & Abdoli, M. (2012). The role of online videos in research communication: A content analysis of YouTube videos cited in academic

publications. *Journal of the American Society for Information Science and Technology*, 63(9), 1710–1727.

Kurtz, M.J. (2004). Restrictive access policies cut readership of electronic research journal articles by a factor of two, Harvard-Smithsonian Centre for Astrophysics, Cambridge, MA. Retrieved January 6, 2013 from: http://opcit.eprints.org/feb19oa/kurtz.pdf

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature*, 411, 521.

Lee, S. & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35(5), pp. 673-702.

Lepori, B. & Probst, C. (2009). Using curricula vitae for mapping scientific fields: A small-scale experience for Swiss communication sciences. *Research Evaluation*, 18(2), 125-134.

Nicholas, D. & Rowlands, I. (2005). Open access publishing: the evidence from the authors. *Journal of Academic Librarianship*, (31)3, 179-81.

Rowlands, I., Nicholas, D. & Huntington, P. (2004). *Scholarly communication in the digital environment: what do authors want? Findings of an international survey of author opinion*, CIBER, University College London, London, Retrieved January 6, 2013 from: http://www.homepages.ucl.ac.uk/~uczciro/ciber-pa-report.pdf

Sabatier, M., Carrere, M. & Mangematin, V. (2006). Profiles of academic activities and careers: Does gender matter? An analysis based on French life scientist CVs. *Journal of Technology Transfer*, 31(3), 311-324.

Sandström, U. (2009). Combining curriculum vitae and bibliographic analysis: Mobility, gender, and research performance. *Research Evaluation*, 18, 135-142.

Shema H., Bar-Ilan J. & Thelwall M. (2012). Research blogs and the discussion of scholarly information. *PLoS ONE*, 7(5): e35869. doi:10.1371/journal.pone.0035869

Su, X. (2011). Postdoctoral training, departmental prestige and scientists' research productivity. *Journal of Technology Transfer*, 36 (3), 275-291.

Sugimoto, C.R. & Thelwall, M. (in press). Scholars on soap boxes: Science communication and dissemination in TED videos, *Journal of the American Society for Information Science and Technology*.

Swan, A. & Brown, S. (2005). *Open access self-archiving: an author study*, JISC Technical Report, pp. 1-97. Retrieved December 18, 2012 from: http://eprints.ecs.soton.ac.uk/10999/01/jisc2.pdf

Thelwall, M. (2004). *Link analysis: An information science approach*. San Diego: Academic Press.

Van Noorden, R. (2012) Britain aims for broad open access: But critics claim plan seeks to protect publishers' interests. *Nature*, 486, 302–303

Woolley, R. & Turpin, T. (2009). CV analysis as a complementary methodological approach: Investigating the mobility of Australian scientists. *Research Evaluation*, 18(2), 143-151.

AN EXAMINATION OF THE POSSIBILITIES THAT ALTMETRIC METHODS OFFER IN THE CASE OF THE HUMANITIES (RIP)

Björn Hammarfelt[1]

[1] *bjorn.hammarfelt@abm.uu.se*
Department of ALM, Uppsala University, Sweden

Abstract
The advantages of altmetrics—the diversity of dissemination channels analysed, the speed of getting data, the openness of methods, and the ability to measure impact beyond the 'scholarly realm'—could be seen as especially promising for fields that currently are difficult to study using established bibliometric methods and data sources. This paper reviews the benefits of using altmetric methods to analyse the impact of research in the humanities and tests some of the most common altmetric tools on a small sample of publications and authors. The findings indicate that many of the problems identified in research on the use of bibliometrics on the humanities are also relevant for altmetric approaches. The importance of non-journal publications, the reliance on print as well the limited availability of open access publishers are characteristics that hinder altmetric analysis. However, this study provides only a few examples and further studies are needed in order to examine the possibilities that altmetric methods offer.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Webbometrics (Topic 7).

Introduction
Characteristics of scholarship in the humanities have limited the application of customary bibliometric methods. The reservation against the use of these methods concerns the mixed audience of research in the humanities that includes an international scholarly audience, a national audience as well as a public audience (Nederhof, 2006). The diverse publication channels used by scholars in the humanities—articles, book articles and monographs—are another explanation for the difficulties of applying bibliometric methods to these fields. This as major citation databases, such as *Thomson Reuters Web of Science* (WoS) and *Elsevier Scopus*, foremost index articles in English language journals. Alternative approaches for measuring impact in the humanities have been proposed due to the inadequacy of conventional bibliometric methods (see for example Linmans, 2010; Hammarfelt 2012). 'Altmetrics'—alternative metrics usually based on data from the social web—have been seen as an especially promising approach in the efforts to find appropriate metrics for research fields in the social sciences and the humanities (Tang, et al., 2012). However, few studies have actually examined if

the promises of altmetrics hold true when it comes to research fields in the humanities. In this paper the promises of altmetrics are reviewed, the altmetric coverage for a small sample of publications and researchers is considered and the future prospects of these methods are discussed.

Rapid changes in how research is disseminated have not only challenged established models for publishing, but also these changes have questioned the current methods for measuring scholarly impact. Measures that are not derived from commercial citation indices such as *Web of Science* or *Scopus* have been advocated. These new, 'altmetric' measures, propose not only to solve problems with current methods, but they also appear to open up for the measurement of impact beyond citations in scholarly journals. Thus, altmetrics considers all stages and products of scholarly research from "[…] social literature search via Facebook to discussion of published results via Twitter, including any impact a publication or author may have on other people […] (Bar-Ilan, et al., 2012, p. 2). Altmetrics is not only a growing research area but also a 'movement' with a manifesto (Priem et al., 2010), and a growing market for commercial companies offering altmetric data to researchers and institutions.

Currently, a lot of attention is given to how altmetrics can be used to study and eventually evaluate the impact of scholarly publications. Advocates of this new approach for measuring the impact of research claim that altmetrics have many benefits compared to conventional bibliometric methods. Wouters and Costas (2012) have reviewed the literature on the topic and they identify four arguments in favour of alternative metrics. These are the diversity of dissemination channels analysed, the speed of getting data, the openness of methods, and the ability to measure impact beyond the 'scholarly realm'. Below these four promises are scrutinized with focus on their significance for the humanities.

Diversity

Altmetrics allow for analyses of many different kinds of materials. Scholarly journals, books as well as blogs or 'tweets' can be studied using data available on the (social) web. The range of altmetric methods appears as promising, not the least for many research fields in the humanities, as it opens up for measuring impact beyond English language journals indexed in citation databases. Thus, the humanities with its diverse audience consisting of national and international scholars as well as a large public audience should benefit from an approach that considers many different dissemination forms.

Speed

When using citation data we need to wait a substantial time—usually between two and five years—in order for publications to gather citations for analysis. Additionally, the time it takes for a publication to get cited is often longer in many fields in the humanities, and it has been suggested that lengthier citation

windows should be used in these fields (Nederhof, 2006). However, altmetric data (such as downloads or views) are instantly available and accessible for analysis. The use of altmetric methods could therefore be a viable solution in fields where it takes a substantial time for publications to gather citations.

Openness of methods

In general, altmetric data are readily available for any researchers to download and use. This in contrast to citation databases such as WoS or *Scopus* where an expensive license is needed in order to access the material. The availability of data makes it possible for a larger group of researchers (including scholars in the humanities) to access metrics on the 'impact' of themselves and others. However, as pointed out by Wouters and Costas (2012) many of the services used for altmetric analyses are only partly open, as we know very little about the inner workings of a service such as *Google Scholar*.

Beyond scholarly impact

Altmetric methods are not restricted to the judgements of scholarly authors. Therefore they can be used to measure impact beyond the scholarly world. The possibility of measuring the public or social impact of research appears as encouraging for research fields, such as history or literature, that often target a wide audience stretching outside academia. The potential to measure 'social' impact as well as the ability to study many different dissemination forms appears as two strong arguments for the use of altmetric methods on the humanities.

Method and analysis

A small selection of publications by scholars in the humanities was analysed in order to test the applicability of altmetric measures. The approach adopted here is somewhat similar to the one used by Bar-Ilan and colleagues who studied the web presence and altmetric impact of scholars attending the 11th STI, *Science and Technology Indicators Conference*, in Leiden (Bar-Ilan et al., 2012). The material analysed here consists of publications from researchers at the English department at Uppsala University as well as publications by two influential humanities scholars. The department of English was selected as their publications could, at least in theory, be viewed, downloaded or cited anywhere where English is read. The Scandinavian publication portal, *Academic Archive On-Line* (DiVA), was used in order to identify and select documents for analysis. A relatively recent period was chosen, as one of the alleged advantages of altmetrics is the speed of getting data. Forty-seven publications (the search was limited to journal articles, book chapters, books and proceedings) were registered in the database in the period of 2009-2010. These were searched in *Google Scholar* and those having at least one citation were further used in the analysis of altmeric coverage. Noteworthy is that only eight publications of 47 were cited at all. Beside *Google Scholar* were *Thomson Web of Science* (using cited reference search), *Microsoft*

Academic Search, *Mendeley* and *Library Thing* searched for citations, mentions, readers or members relating to these publications (table 1).

Table 1. Altmetric coverage of the humanities. Most cited (in *Google Scholar*) publications in DiVA from the English department at Uppsala University, 2009-2010 (searches conducted 2013-04-25)

Publication type (identifier)	*Google Scholar*	*WoS*	*Academic search*	*Mendeley*	*Library Thing*	Electronically available / Open access
Monograph (1)	39 cit.	9 cit.	No record	No record	2 members	No/No
Journal article (2)	12 cit.	2 cit.	2 cit.	11 readers	No record	Yes/No
Proceedings (3)	5 cit.	No record	No record	No record	No record	Yes/Yes
Journal article (4)	5 cit.	No record	No record	No record	No record	No/No
Book chapter (5)	3 cit.	No record	No record	No record	No record	Yes/No
Book chapter (6)	3 cit.	No record	No record	No record	No record	No/No
Proceedings (7)	1 cit.	No record	No record	No record	No record	Yes/Yes
Journal article (8)	1 cit.	No record	No record	No record	No record	No/No

The results show that there are few altmetric traces of these eight publications; the most cited publication (1) has two 'members' with the book in their 'library' (*Library Thing*) and the most cited article (2) has eleven readers in *Mendely*. The sample used here is indeed very small and a much larger dataset would be needed in order to substantiate conclusions regarding the application of altmetric methods on the humanities. Yet, the study is illustrative in showing that the different types of publications produced by humanities scholars are an important reason to the limited coverage of the humanities in altmetric data sources. Many of the services used in altmetrics analyses focus foremost on the journal article (*Mendeley*, *Cite U Like* and so forth) and the coverage of other types of documents (proceedings, monographs, book articles) is small. This seriously limits its usability in the humanities as well as in parts of the social sciences. However, one solution to this problem is to use social devices, such as *Library Thing,* that are directed towards books. The findings here can be compared to the field of bibliometrics and research policy where the reference database *Mendeley* covered 82 percent of the sampled outputs of researchers, while 28 percent of the papers had readers in *Cite U Like* (Bar-Ilan et al., 2012).

Previous studies have shown that interdisciplinary scholars in the humanities with an international reputation can be studied using bibliometrics, while the coverage of less prominent researchers is low or very low (Hammarfelt, 2012). Is the same pattern distinguishable when altmetric methods are used? The altmetric record of the historian Karin Johannisson as well as the 'impact' of gender theorist Judith Butler were studied in order to answer this question. Johannisson was selected because she is one of the most famous and prized historians in Sweden, while Butler is one of the most cited contemporary scholars in the humanities (Hammarfelt, 2012). Searches were made using the same tools as in the examples above with the addition of *Google Blog Search*. *Google Blog Search* was added as it could be seen as an indication to impact outside academia.

Table 2. Altmetric coverage of the humanities. The example of 'Judith Butler' and 'Karin Johannisson' (searches conducted 2012-11-24 except for *Library Thing* 2013-04-25)

Data source	*Records for Karin Johannisson*	*Record for Judith Butler*
Thomson WoS (using cited reference search)	104*	5878*
Google Scholar, using *Publish or Perish*, (Harzing, 2007)	700	75, 040
Academic search	No record	5495 cites
Mendeley *Library Thing* *Goggle Blog Search*	No record 108 members 1740 hits	1322 readers† 4880 members 1, 2 million hits

*Cited Author=(butler, Judith) Timespan=All Years. Databases=SCI-EXPANDED, SSCI, A&HCI (2012-12-02); Cited Author=(johannisson, K) Timespan=All Years. Databases=SCI-EXPANDED, SSCI, A&HCI. (2012-12-02)
†Based on 30 documents

The impact of the two scholars is apparent in the records shown above (table 2). Most evident is the number of cites registered for Judith Butler in *Google Scholar* but her altmeric record is also impressive with 1322 readers on *Mendely*, 4880 members on *Library Thing* and her visibility in the blogosphere is huge with 1,2 million mentions. The influence of Judith Butler's work appears as exceptional both in academia and outside the scholarly realm. The many citations to the works of Johannisson are also impressive as she foremost publishes in Swedish. However, Johannisson is not covered in *Academic Search* or in *Mendeley*. This is probably due to the main language (Swedish) and form (monographs) of her publications. The general coverage of altmetric measures appears as meagre and in some cases non-existent when compared to citations registered in WoS and *Google Scholar*. The main explanation for this is that services such as *Mendeley* focus on journal articles, while the most cited publications in fields such as literature, history or gender studies often are monographs. *Library Thing* appears

as one useable service for gauging the impact of books, but it mainly covers publications written in English. Hence, altmetric methods did provide some data on the impact of these authors, especially in the case of Butler, but 1,322 readers on *Mendeley* appears as little compared to 75,400 citations in *Google Scholar* and 5,878 citations in WoS.

Discussion

Altmetric methods seem to solve several problems associated with the use of bibliometrics on the humanities. They allow for a multitude of sources to be analyzed, and altmetric methods make it possible to measure 'instant' impact in fields where citations take long time to gather. It has been argued that the variety of measures available for analysis is one of the great benefits of altmetrics: "Because altmetrics are themselves diverse, they're great for measuring impact in this diverse scholarly ecosystem" (Priem, et al., 2010). However, a majority of altmetric methods focuses on articles in journals as the prime unit of analysis. This works well in fields where (international) journals are the prefered publication channel but it is less effective in research fields where scholars publish in a variety of channels.

Scholarly 'ecosystems' are different also in their search for and use of sources. Research shows that scholars in fields such as history or literature are still more dependent on print material and library resources when compared to scholars in the social sciences and natural sciences (Collins, Bulger and Meyer, 2012). This even if digitalization of books and other sources as well as the emergence of fields such as 'digital humanities' are changing the infrastructure of research in the humanities. The further reliance on print has consequences for the application of altmetric methods as frequent use of web based social devices are often a prequisite for analysing impact. Thus, the scholarly practices of many researchers in the humanities may limit the availablity of altmetric data.

Altmeric methods are dependent on journal usage and access to data. Analysis of usage (downloads/views) demands that the researchers have access to download data from the publishers. This kind of data is often not accessible in the case of commercial publishers, but in the case of open access journals—such as *PloS* used by Priem, Piwowar and Hemminger (2012)—this type of data can be gathered and analysed. However, the available sources for open access publishing are scarce in many fields in the humanities and the social sciences, and one reason is large differences in how research is communicated. Monographs are expensive to produce and compared to research fields in the medical and natural sciences, little has been done to facilitate open access publishing of books (Hall, 2008). Hence, the limited availability of open access publishers is one explanation to why few researchers have applied altmetric methods on the humanities.

The promises that altmetric methods holds have to be examined further as analyses seldom go beyond 'techniques of narcissism' (Wouters & Costas, 2012). Altmetric methods need to be tested on a larger scale; methods for assessing the impact of research units and the structure of research fields have to be developed. Several issues have to be dealt with before altmetric data can be seen as a feasible alternative to traditional metrics: the reliability of data has to be addressed, a theoretical understanding of the units analysed ('downloads', 'view', "hits", "members", "likes" and so forth) has to be developed and, as pointed out in this paper, disciplinary differences in the communication of knowledge must be considered.

The possibilities that altmetric methods offer to the humanities cannot be denied but, as shown in this paper, there are several issues that have to be addressed in order for the promise to be realized. Many reservations against the use of altmeric methods on the humanities relate to problems already discussed in bibliometric literature; the diverse publication channels used by scholars in the humanities, the still large reliance on print in many disciplines as well as the restricted access to data limit the altmetric coverage of these fields. The digitalization of research in the humanities, a general movement for open access across research fields, as well as the further development and diversification of altmetric methods could at least partly solve the issues raised above. Then, altmetrics would be an attractive and in several cases superior alternative to traditional bibliometric methods for analysing and measuring the impact of research fields in the humanities.

References

Bar-Ilan, J. Haustein, S. Peters, I., Priem, J. Shema, H. & Tersliesner, J. (2012). Beyond citations: Scholars' visibility on the social Web. In E. Archambault., Y. Gingras and V. Larivière (Eds.) *Proceedings of 17th International Conference on Science and Technology Indicators*, (pp. 98-109), Montréal: Science-Metrix and OST.

Collins, E., Bulger M. E., Meyer, E. T. (2012). Discipline matters: Technology use in the humanities. *Arts and Humanities in Higher Education, 11(1-2), 76-92.*

Hall, G. (2008). *Digitize this book: The politics of new media or why we need open access now*. Minneapolis: University of Minnesota Press.

Hammarfelt, B. (2012). *Following the Footnotes: A Bibliometric Analysis of Citation Patterns in Literary Studies*. (Diss.). Uppsala: Acta Universitatis Upsaliensis.

Harzing, A.W. (2007) Publish or Perish, Software. Retrieved, 2012-11-20, from www.harzing.com/pop.htm

Linmans, J. A. M. (2010). Why with bibliometrics the humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library holdings and productivity measures. *Scientometrics*, 83(2), 337-354.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66(1), 81-100.
Priem, J. Piwowar, H. A., Hemminger, B. (forthcoming). Altmerics in the wild: Using social media to explore scholarly impact. Retrieved, January 10, 2013 from arXiV.org.
Priem, J., Taraborelli, D., Groth, P. & Neylon, C. (2010). *Altmetrics: A Manifesto*. Retrieved, January 10, 2013, from http://altmetrics.org/manifesto/
Tang, M., Wang, C., Chen, K. & Hsiang, J. (2012). Exploring alternative cyber bibliometrics for evaluation of scholarly performance in the social sciences and humanities in Taiwan. *Proceedings of the ASIS&T Annual Meeting*, Vol. 49. Retrieved, December 12, 2012, from www.asis.org/asist2012/proceedings/openpage.html
Wouters, P. & Costas, R. (2012). *Users, narcissism and control – Tracking the impact of scholarly publications in the 21st century.* SURF-foundation. Utrecht.

EXPLORING QUANTITATIVE CHARACTERISTICS OF PATENTABLE APPLICATIONS USING RANDOM FORESTS

Fuyuki Yoshikane[1], Chizuko Takei[2], Keita Tsuji[3], Atsushi Ikeuchi[4] and Takafumi Suzuki[5]

[1] *fuyuki@slis.tsukuba.ac.jp,* [3] *keita@slis.tsukuba.ac.jp,* [4] *atsushi@slis.tsukuba.ac.jp*
University of Tsukuba, Faculty of Library, Information and Media Science, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

[2] *naoe.chizuko@ynu.ac.jp*
University of Tsukuba, Graduate School of Library, Information and Media Studies, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

[5] *takafumi_s@toyo.jp*
Toyo University, Faculty of Sociology, 5-28-20 Hakusan, Bunkyo-ku, Tokyo (Japan)

Abstract
This study examined Japanese patents in terms of the quantitative characteristics of application documents that resulted in the acquisition of rights. Our purpose is to clarify the relationship between the features and patentability of applications. The groups of approved applications and those that had not been approved were compared for twelve variables: publication time lag; numbers of inventors, classifications, pages, figures, tables, claims, priority claims, countries for priority claims, cited patents, and cited non-patent documents; and median citation age. Furthermore, we carried out the experiments in which patent applications were automatically classified into the two groups by the machine learning method, random forests. As a result, statistically significant differences between the two groups were observed for the following variables ($p < 0.001$): the numbers of inventors, pages, figures, claims, priority claims, and countries for priority claims were significantly larger in the group of approved applications, while the time lag until publication was smaller. In particular, the publication time lag and the numbers of inventors, pages, and figures were variables representing the features that largely contribute to discriminating approved applications in the classification using random forests, which implies that these have relatively strong relationships with patentability.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5).

Introduction
Focusing on patent applications in Japan, this study investigates the tendencies in variables representing features such as the quantity of descriptions in application documents and attempts to discriminate approved applications, which result in the

acquisition of rights, on the basis of those variables. Our purpose is to clarify the relationship between the features and patentability of applications.

Many bibliometric studies have been carried out on patent data from various viewpoints such as citation relationships and market values of patents (e.g., Narin, 1995; Harhoff et al., 1999; Hall, Jaffe & Trajtenberg, 2005). Furthermore, recent years have seen progress in research regarding the adequacy of patent applications as to whether or not they result in the acquisition of rights, i.e., estimating patentability. It is considered that the "quality of patents" reflects their contribution to the entire society involved with them rather than to individual firms (Kashima et al., 2010). It is meaningful to clarify the factors that influence patentability relating to quality in this sense. In Japan, as in many other countries, the establishment of patent rights is based on the substantive examination principle. Recent attention has been drawn to the relationship between the characteristics of descriptions in patent application documents, which form the basis of examination, and whether or not the applications are approved through examination. Several studies on Japanese patents have proposed methods for discriminating approved applications using the numbers of inventors and claims, *tf-idf* (term frequency-inverse document frequency), which represents the degree of importance of terms, and so on (e.g., Kashima et al., 2010; Hido et al., 2012). Classification techniques such as support vector machine (SVM) or logistic regression are applied in these studies.

However, when the primary purpose of the analysis is neither to estimate patentability itself nor to classify whether applications are approved, but rather to calculate each variable's importance in the classification (i.e., influence on patentability), SVM is not suitable. Besides, as for logistic regression, we face multicollinearity, which arises from confounding between variables with a high correlation, and difficulty of dealing with variables whose distribution is skewed and does not follow a normal distribution. Thus, it is desirable for this purpose to adopt a robust method against these issues. Random forests (RF), a machine learning method based on bootstrap samples and decision trees (Breiman, 2001), is not only robust but also showing high performance in some domains such as text classification (e.g., Jin & Murakami, 2007; Suzuki, 2009). In the field of bibliometrics, a few studies have demonstrated the usefulness of RF in classifying authors of academic papers (e.g., Kiyokawa et al., 2011). With this as the background, this study employs RF to classify patent applications into approved ones and those that have not been approved, and to calculate and compare each variable's importance in the classification.

Data

Patent applications published in January 2007 were subjected to our analysis. Data of application documents was extracted from the "patent gazette (publication of unexamined patent applications)" published in Japan. Moreover, we retrieved

the status of the acquisition of rights—i.e., whether applications were approved through examination—from the Industrial Property Digital Library (http://www.ipdl.inpit.go.jp/homepg_e.ipdl), which is provided by the National Center for Industrial Property Information and Training (INPIT), Japan.

NTCIR-8 Patent Translation Test Collection (Fujii et al., 2010) compiled by the National Institute of Informatics (NII), Japan, was our information source for the patent gazette. The patent gazette published before 2004 was recorded in plain text format, thus making it difficult to extract some types of items (e.g., cited patents) in an accurate and comprehensive manner; however, since then, each item has been described with tags in XML format, which makes extraction easier. We extracted the following data from NTCIR-8 for all 20,400 patent applications published in January 2007: dates of application and publication; main and sub-classifications; numbers of inventors, pages, figures, tables, claims, priority claims, countries for priority claims, cited patents, and cited non-patent documents; and publication/application year for each cited patent. In Japan, as well as IPC (International Patent Classification) (WIPO, 2010), FI, the Japanese domestic classification into which IPC is subdivided, is also assigned to application documents. Considering the future possibility of carrying out international comparisons, we extracted and used, not FI, but IPC for the analysis.

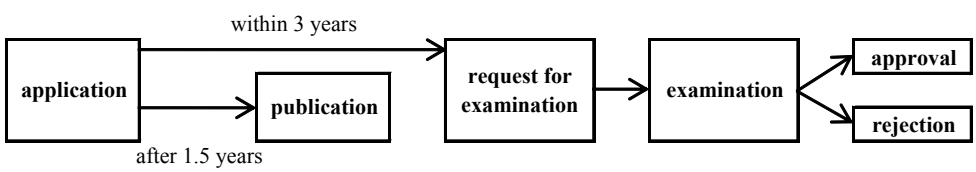


Figure 1. Brief outline of the flow of patent application in Japan.

Figure 1 shows a simplified flow of the steps from application to the acquisition of rights for patents in Japan. As a rule, application documents are published in the patent gazette after a lapse of one and a half years. Request for examination for acquiring rights must be made within three years subsequent to application, and the examination requires several years. This chart implies that, in cases where examination is requested, the results are obtained within a period of around five years from application, that is to say, around four years from publication, though there are also some applications that do not request examination, such as defensive ones whose aim is not acquiring rights but preventing the acquisition of rights by third parties. Therefore, with regard to the applications published in January 2007, it is assumed that we can broadly distinguish between approved ones and rejected ones (or those that have not requested examination) by checking the situation in the latter half of 2012, over five years since their publication. We assume that, although of course some of the applications would not have received the results of examination yet due to prolongation of the examination period,

these cases are in the minority and thus the check in the latter half of 2012 enables us to gain an understanding of trends in the acquisition of patent rights. Utilizing the Industrial Property Digital Library from the end of August until the beginning of September 2012, this study checked the situation of the acquisition of rights for the applications published in January 2007, which were subjected to the analysis.

Methodology

Features

In this study, the following twelve variables were employed to predict whether patent applications were approved: publication time lag (*TL*); numbers of inventors (*IV*), assigned classifications (main and sub-classifications) (*VC*), pages (*PG*), figures (*FG*), tables (*TB*), claims (*CL*), priority claims (*PC*), countries for priority claims (PC_c), cited patents (F_{citing}), and cited non-patent documents (FN_{citing}); and median citation age (*CA*).

Focusing on citations between patents, citation age was calculated for each of the cited patents (including utility models). The values of citation age were derived from the differences between the publication years (or application years) of a subject patent and the patents cited by it, and then the median, *CA*, was calculated for all of these. The median of citation age cannot be calculated in subject patents that have no cited patents. However, it is necessary for handling those patents in RF to assign some value to them. So, after calculating *CA* for each patent with at least one citation, we assigned a median of these values of *CA* (i.e., "median of medians"), as a "neutral" value, to the remaining patents in the analysis using RF. Since the analysis was carried out individually for each field (classification) as described later, a median to be assigned was derived only from the values of *CA* for patents belonging to the same field.

TL was derived from the difference between the dates of application and publication. While, as stated in the previous section, most applications are published in the patent gazette after one and a half years—in other words, the time lag until publication is set at a constant value of around 550 days as a principle—, there are several types of exceptions to this rule. One is the case where an application document needs to be amended because of defects and, therefore, requires time for resubmission; another is an application requested to come under the early publication system through which application documents are published before the lapse of one and a half years. The early publication system allows applicants to move forward the occurrence of the right to demand compensation. The time lag until publication, *TL*, was adopted as a variable in our analysis based on the supposition that such applications as need to be amended tend to have low patentability while those for which the occurrence of the right to demand compensation is moved forward tend to be accompanied by documents having high patentability due to the applicants' profit expectations.

If a patent with higher quality is more often cited, variables correlated with the number of times a patent is cited by others also can be used to predict patentability. This study selected variables based on this assumption. Regarding patents, although there have been few studies on the correlation between the number of forward citations and quantitative characteristics of documents, some variables, e.g., the numbers of inventors, claims, backward citations, scientific linkages, and classifications (*IV*, *CL*, F_{citing}, FN_{citing}, and *VC*), were reported to be correlated with the number of forward citations (Lee et al., 2007; Yoshikane, Suzuki & Tsuji, 2012). As for academic papers, many studies have investigated factors that influence quality (more specifically, quality in respect to impact). For example, the numbers of authors, pages, figures, tables, and references and the Price's index have all been shown to be connected with the number of forward citations (Snizek, Oehler & Mullins, 1991; Peters & van Raan, 1994; Glänzel, 2002; Kostoff, 2007). Broadly speaking, among the variables employed in this study, *IV*, *PG*, *FG*, *TB*, F_{citing} (FN_{citing}), and *CA* correspond to them, respectively. On the other hand, *CL*, *PC*, and PC_c represent information particular to patent data. This study also employed these variables, considering them to reflect the volume and value of inventions.

First, the groups of approved applications and those that had not been approved were compared with regard to the mean value for each of the abovementioned twelve variables. Because these variables exhibit highly skewed distributions deviating from the normal distribution, we tested the significance of the difference between the two groups by the Wilcoxon rank sum test.

Random Forests

Using the machine learning method, random forests (RF) (Breiman, 2001), patent applications were automatically classified into groups of approved applications and those that had not been approved. Furthermore, the error rate of classification and the degree of importance representing contribution to classification for each variable were calculated. The specific processes involved in the experiment were as follows:

(1) Creating 1,000 sets of bootstrap samples by replicating the matrix with i rows for patent applications and j columns for variables.
(2) For each set of bootstrap samples, randomly extracting $m(=\sqrt{j})$ columns from j columns.
(3) For each set of bootstrap samples, constructing an unpruned decision tree using two-thirds of samples on the basis of a decrease in the Gini index. The remaining one-third samples are used for evaluation.
(4) Constructing a new decision tree through a majority vote of the trees constructed in (3).

The two-class classification experiment was carried out individually for each field. The "section" (top layer in IPC) of the main-classification assigned to an application document was deemed to be the field to which it belongs. The value *i* for the number of rows in the above matrix is the number of applications in each field. Meanwhile, the value *j* for the number of columns in the matrix equals the number of variables, which is 12, and *m* variables among them are extracted and used in bootstrap samples. Since the degree of importance of variables is estimated, not by using all of them at the same time in bootstrap samples, but on the basis of average values calculated through repeating random extraction of some of them, RF is robust in respect to confounding between correlated variables.

The degree of importance, that is, contribution to classification, of variables is calculated using the following formula.

$$VI_{acu} = \frac{mean(C_{oob} - C_{per})}{s.e.}$$

where C_{oob} and C_{per} represent the number of applications correctly classified using each variable in the data for evaluation (i.e., one-third of samples) and the number of correctly classified applications when *m* variables are randomly permutated in the data for evaluation, respectively. The standard error is represented by $s.e.$ Based on the degree of importance VI_{acu}, the strength of the connection with patentability was compared among variables.

In RF, because of selecting variables at random when constructing decision trees, the result of classification must depend on a computer-generated random number sequence. Figure 2 shows the fluctuations of the error rate for classification according to the increase in the number of sets of bootstrap samples where variables are randomly extracted, that is, the number of decision trees, taking an instance of the experiment carried out on 2,221 applications that belong to section A (human necessities) in IPC. The graphs at the top and bottom of the figure represent the error rate for the groups of approved applications and those that had not been approved, respectively; the middle graph shows the overall error rate. We can confirm that, although the error rate in classification with a small number of decision trees fluctuates largely, the error rate becomes nearly stable if more than about 200 decision trees are constructed. The results shown in the following section are based on classification experiments in which the number of decision trees was set to 1,000.

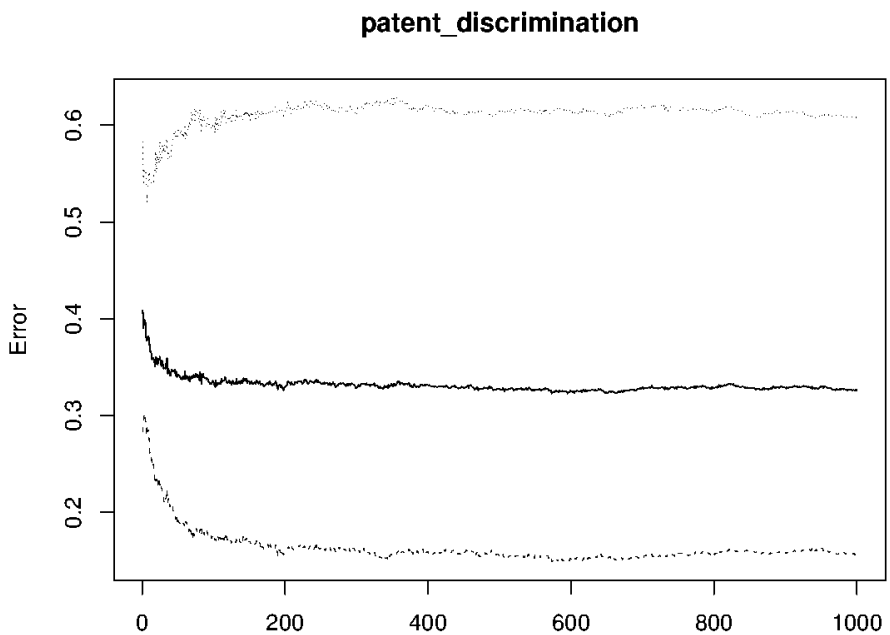


Figure 2. Number of decision trees and error rate.

Table 1. Number of applications and ratio of approved ones to all in each field.

Section	N	N_a	N_{na}	N_a/N
A: human necessities	2221	832	1389	37.46%
B: performing operations; transporting	3305	1413	1892	42.75%
C: chemistry; metallurgy	1597	613	984	38.38%
D: textiles; paper	179	66	113	36.87%
E: fixed constructions	769	392	377	50.98%
F: mechanical engineering, etc.	1779	711	1068	39.97%
G: physics	5345	1950	3395	36.48%
H: electricity	5205	2036	3169	39.12%
Whole data	20400	8013	12387	39.28%

Results and Discussion

Table 1 shows the number of patent applications and ratio of approved ones to all for each field. N represents the number of all applications published in January 2007, while N_a and N_{na} represent the numbers of approved applications and those that had not been approved, respectively. In the period subject to the analysis, there were many applications for sections G and H (more than 5,000 for each) but very few for section D (less than 200). The ratio of approved applications, N_a/N, was about 40% for the whole data. Looking at the ratio for each field, we see that the highest was around 50% in section E while the lowest was around 35% in sections D and G. However, because the data sizes of sections D and E were small

(less than 1,000 for each), the above tendencies in these fields should be interpreted carefully.

Table 2 shows the average values for variables in each field. Because, as mentioned in the previous section, *CA* (median citation age) is incalculable for applications where F_{citing} (number of cited patents) is zero, these were excluded in calculating averages of *CA*. There were 4,762 patent applications in which F_{citing} is 0, accounting for around 20% of all the 20,400 applications.

Table 2. Average for each variable in each field.

	TL	*IV*	*VC*	*PG*	*FG*	*TB*	*CL*	*PC*	PC_c	F_{citing}	FN_{citing}	*CA*
A	504.4	2.18	2.43	17.06	9.98	0.55	6.59	0.15	0.14	1.75	0.24	7.58
B	530.4	2.12	2.69	13.57	8.21	0.37	6.81	0.11	0.10	1.65	0.05	8.24
C	496.4	2.81	4.18	18.69	3.67	2.13	8.63	0.31	0.22	3.04	0.61	9.12
D	501.6	2.29	2.94	12.91	4.57	0.89	6.68	0.26	0.22	2.38	0.06	9.85
E	541.6	2.18	2.02	11.01	8.43	0.11	5.25	0.07	0.06	1.45	0.04	7.83
F	530.4	2.22	2.69	12.30	8.19	0.10	6.38	0.11	0.10	1.48	0.04	8.38
G	513.6	2.16	2.55	17.52	10.66	0.44	8.99	0.15	0.13	1.60	0.17	7.17
H	506.3	2.21	2.72	15.25	9.99	0.26	8.87	0.18	0.15	1.50	0.19	7.11
Whole	514.5	2.23	2.73	15.60	9.12	0.49	7.93	0.16	0.13	1.70	0.18	7.68

Section C (chemistry; metallurgy) was peculiar among the eight fields. The values for *IV*, *VC*, *TB*, F_{citing}, and FN_{citing} in section C were all markedly high compared with those in other fields. In particular, *TB* in C demonstrated a value over twice as high as in other fields and was around 20 times higher than in E and F. In contrast, *FG* was markedly low in section C. It is observed that the content of applications in this field tends to be explained through tables rather than figures. In addition, although not as outstanding as the abovementioned variables, both *PG* and *PC* were highest while *TL* was lowest in C among all the fields.

Although not as marked as section C, D (textiles; paper) also exhibited similar tendencies. For all variables except for *PG*, *CL*, and FN_{citing}, these two fields together occupied either the highest two or lowest two positions. It implies that chemistry (materials chemistry)-related fields have different characteristics than other fields, such as machine or electrical engineering-related fields, in patent applications.

We divided the subject patent applications into two groups, that is, approved applications and those that have not been approved. The latter group includes not only rejected ones but also those that have not requested examination. The average values for variables in each group are presented in Table 3. Looking at the subject applications as a whole, we found that many of the variables were significantly higher in the group of approved applications: *IV*, *PG*, *FG*, *CL*, *PC*, and PC_c ($p < 0.001$); F_{citing} and FN_{citing} ($p < 0.05$). In contrast, *TL* ($p < 0.001$) and *VC* ($p < 0.05$) were significantly lower in this group. That is to say, as an overall

tendency, patentable applications have the following characteristics: (1) those with more inventors and descriptions, (2) those for which early occurrence of the right to demand compensation is requested, and (3) those that consist of content related to limited rather than broad areas.

Table 3. Comparison of variables between approved applications and those that have not been approved.

		TL	*IV*	*VC*	*PG*	*FG*	*TB*	*CL*	*PC*	PC_c	F_{citing}	FN_{citing}	*CA*
A	approved	489.3	2.31 *	2.20	18.38 *	11.67 *	0.47	6.90 *	0.14	0.12	1.91 *	0.22	7.85 +
	not approved	513.5 *	2.10	2.57 +	16.27	8.96	0.60	6.41	0.16	0.14	1.66	0.25	7.39
B	approved	524.9	2.28 *	2.67	14.30 *	8.94 *	0.36	6.86	0.13 +	0.11 +	1.60	0.04	8.18
	not approved	534.5 +	2.00	2.70	13.04	7.67	0.39 +	6.78	0.10	0.09	1.68	0.05	8.29 +
C	accepted	497.2	3.01 *	4.04	17.37	4.06 +	2.03	8.56	0.30	0.19	3.23 +	0.66	9.07
	not approved	496.0	2.69	4.26	19.52	3.42	2.19	8.68	0.31 +	0.24 +	2.93	0.58	9.16
D	approved	484.3	2.56 +	3.02	13.42	4.86	0.80	7.18	0.29	0.26	2.36	0.03	8.74
	not approved	511.8	2.13	2.90	12.60	4.40	0.95	6.38	0.24	0.20	2.39	0.07	10.52
E	approved	531.0	2.35 *	1.98	11.83 *	9.42 *	0.07	5.30	0.08	0.07	1.51 *	0.02	7.74
	not approved	552.6	2.00	2.07	10.15	7.39	0.16 +	5.19	0.05	0.05	1.39	0.06 +	7.93
F	approved	519.4	2.49 *	2.69	13.67 *	9.26 *	0.09	6.82 *	0.12	0.11	1.53	0.05	8.78
	not approved	537.6 +	2.04	2.69	11.38	7.48	0.10 +	6.09	0.10	0.09	1.45	0.03	8.12
G	approved	494.0	2.37 *	2.57	19.44 *	12.19 *	0.60	9.49 *	0.19 *	0.16 *	1.59	0.22 *	7.26
	not approved	524.8 *	2.05	2.54	16.42	9.78	0.35	8.70	0.13	0.11	1.60	0.14	7.12
H	approved	493.2	2.43 *	2.65	16.70 *	11.00 *	0.25	9.41 *	0.20 *	0.17 *	1.48	0.21 +	7.09
	not approved	514.7 *	2.07	2.77 +	14.32	9.34	0.27	8.52	0.16	0.13	1.51	0.18	7.12
Whole	approved	503.0	2.42 *	2.66	16.63 *	10.18 *	0.49	8.20 *	0.18 *	0.14 *	1.72 +	0.19 +	7.76
	not approved	522.0 *	2.10	2.77 +	14.93	8.43	0.49	7.75	0.15	0.12	1.69	0.17	7.62

* Significantly higher ($p < 0.001$)

+ Significantly higher ($p < 0.05$)

Roughly speaking, a situation similar to this overall tendency was also observed for each individual field. However, the tendency of section C was opposite to that of the whole data in that the two variables related to priority claims, i.e., *PC* and PC_c, were significantly lower in the group of approved applications ($p < 0.05$). With regard to section D, statistically significant differences were not observed between both groups for any variables except for *IV*, which could be a result of small amount of data. *TB*, contrary to *FG*, was significantly lower in the group of approved applications for sections B, E, and F ($p < 0.05$), while the values of *TB* in both groups were almost equal and showed no significant difference for the whole data. As for *CA*, the difference between the two groups was not statistically significant in most of the fields as well as in the whole data. Although the difference in *CA* was significant for sections A and B ($p < 0.05$), the results were contradictory, that is, the value of *CA* in the group of approved applications tends to be higher than that in the other group for A but lower for B.

The following is the result of experiments using RF though which patent applications were automatically classified into the groups of approved applications and those that have not been approved for each field. Figure 3 shows the proportion of incorrectly classified ones to all applications, that is, the error

rate. In the figure, two types of errors are illustrated separately: the first is a case in which an application that has not been approved was incorrectly classified as belonging to the group of approved applications (not approved-approved); the second is a case in which an approved application was incorrectly classified as belonging to the group of applications that have not been approved (approved-not approved). The error rate for classification was in the region of 35%. While the error rate was comparatively low in section A (around 30%), it was comparatively high in sections B and E (around 40%). In most fields, errors of the latter type occurred more often than did those of the former type; in other words, it was difficult to classify approved applications with high recall. This suggests that the characteristics shared by approved applications are few and weak. However, there was only one exception: in section E, errors of the former type occurred more often.

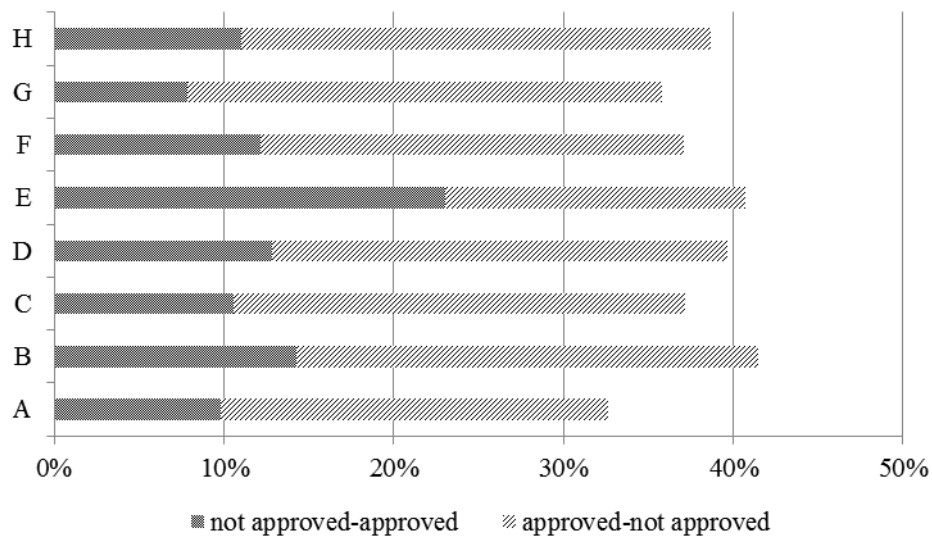
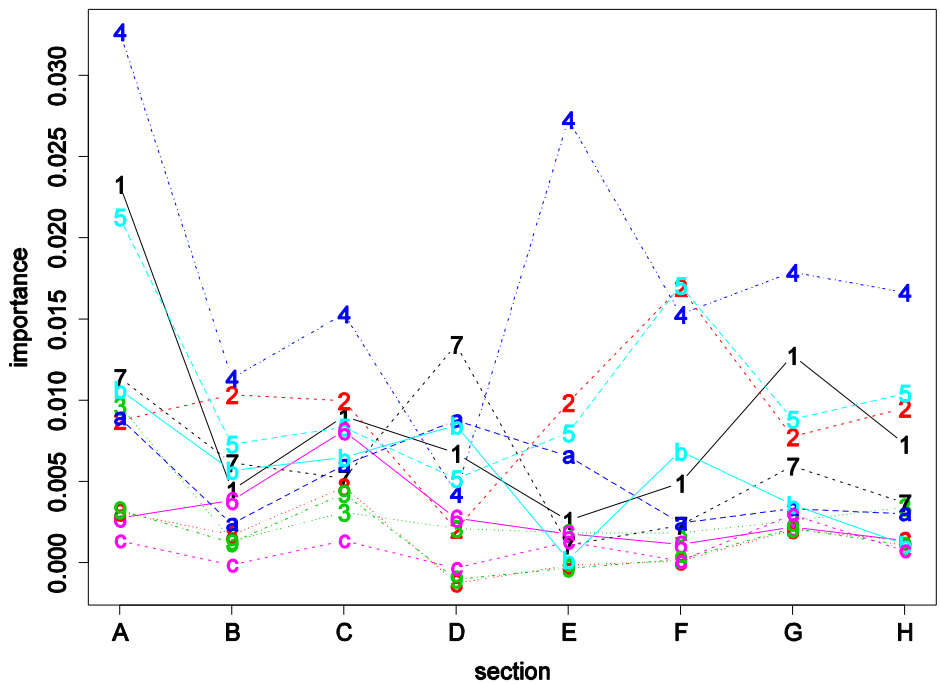


Figure 3. Error rate of the discrimination using random forests.

For each field, Fig. 4 plots the degree of importance of variables that represents contribution to classification. Results showed that an upper group of variables with remarkably high degree of importance is constituted by some variables from among *TL*, *IV*, *PG*, and *FG* in most of the fields. In particular, *PG* stands out for high degree of importance in many fields. Variables other than the above four are largely concentrated at the bottom of the figure.

Table 4 arranges the variables in order of degree of importance. In all fields other than section D (textiles; paper), the top three ranks were occupied by *TL*, *IV*, *PG*, and *FG*. Section D was unique in that none of these variables was found in the top

three ranks; instead, *CL* and the two variables relating to patent citations (i.e., F_{citing} and *CA*) occupied the top positions. The degree of importance of *CL*, in particular, was markedly higher than the other variables in D as shown in Fig. 4. In addition to the two variables relating to priority claims (i.e., *PC* and PC_c), FN_{citing} was stable with a rank not higher than eight. This result means that these variables do not effectively function as features that contribute to the discrimination of approved applications in any of the fields.



1: *TL*, 2: *IV*, 3: *VC*, 4: *PG*, 5: *FG*, 6: *TB*, 7: *CL*, 8: *PC*, 9: PC_c, a: F_{citing}, b: *CA*, c: FN_{citing}

Figure 4. Importance of variables in the discrimination using random forests.

Conclusions

This study examined Japanese patents with respect to the quantitative characteristics of application documents that resulted in the acquisition of rights. The groups of approved applications and those that had not been approved were compared for twelve variables, including the numbers of inventors, classifications, and pages. Furthermore, we carried out the experiments in which patent applications were automatically classified into the two groups by the machine learning method, random forests. As a result, statistically significant differences between the two groups were observed for the following variables ($p < 0.001$): the numbers of inventors, pages, figures, claims, priority claims, and countries for priority claims were significantly larger in the group of approved

applications, while the time lag until publication was smaller. In particular, the publication time lag and the numbers of inventors, pages, and figures were variables representing the features that contributed to the discrimination of approved applications largely, which implies that these have relatively strong relationships with patentability.

The aim of this study was to assess the influence of indicators on patentability rather than to obtain high performance in automatically predicting approved applications. Thus, indicators that are assumed to have the influence on patentability and easily available from application documents have comprehensively been included and compared in the analysis. In future research, aiming to achieve better performance, we will select appropriate indicators not only from bibliographic information, such as the number of inventors or classifications, but also from features regarding the main text of patent specifications, such as the diversity of vocabulary.

Table 4. Ranking of variables in terms of their importance in the discrimination using random forests.

	A	B	C	D	E	F	G	H
1	*PG*	*PG*	*PG*	*CL*	*PG*	*FG*	*PG*	*PG*
2	*TL*	*IV*	*IV*	F_{citing}	*IV*	*IV*	*TL*	*FG*
3	*FG*	*FG*	*TL*	*CA*	*FG*	*PG*	*FG*	*IV*
4	*CL*	*CL*	*FG*	*TL*	F_{citing}	*CA*	*IV*	*TL*
5	*CA*	*CA*	*TB*	*FG*	*TL*	*TL*	*CL*	*CL*
6	*VC*	*TL*	*CA*	*PG*	*VC*	F_{citing}	*CA*	*VC*
7	F_{citing}	*TB*	F_{citing}	*TB*	*TB*	*CL*	F_{citing}	F_{citing}
8	*IV*	F_{citing}	*CL*	*VC*	FN_{citing}	*VC*	FN_{citing}	*TB*
9	PC_c	*PC*	*PC*	*IV*	*CL*	*TB*	*VC*	*PC*
10	*PC*	*VC*	PC_c	FN_{citing}	*CA*	PC_c	*TB*	*CA*
11	*TB*	PC_c	*VC*	PC_c	*PC*	*PC*	PC_c	PC_c
12	FN_{citing}	FN_{citing}	FN_{citing}	*PC*	PC_c	FN_{citing}	*PC*	FN_{citing}

Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research (C) 23500294 (2012) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and we would like to show our gratitude to the support.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T., Ehara, T., Echizen-ya, H. & Shimohata, S. (2010). Overview of the patent translation task at the NTCIR-8 workshop. In *Proceedings of NTCIR-8 Workshop Meeting* (pp. 371-376). Tokyo: National Institute of Informatics.

Glänzel, W. (2002). Co-authorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies. *Library Trends*, 50(3), 461-473.
Hall, B.H., Jaffe, A. & Trajtenberg, M. (2005). Market value and patent citations. *The RAND Journal of Economics*, 36(1), 16-38.
Harhoff, D., Narin, F., Scherer, F.M. & Vopel, K. (1999). Citation frequency and the value of patented inventions. *The Review of Economics and Statistics*, 81(3), 511-515.
Hido, S., Suzuki, S., Nishiyama, R., Imamichi, T., Takahashi, R., Nasukawa, T., Idé, T., Kanehira, Y., Yohda, R., Ueno, T., Tajima, A. & Watanabe, T. (2012). Modeling patent quality: A system for large-scale patentability analysis using text mining. *Journal of Information Processing*, 20(3), 655-666.
Jin, M. & Murakami, M. (2007). Authorship identification using random forests. *Proceedings of the Institute of Statistical Mathematics*, 55(2), 255-268.
Kashima, H., Hido, S., Tsuboi, Y., Tajima, A., Ueno, T., Shibata, N., Sakata, I. & Watanabe, T. (2010). Predictive modeling of patent quality by using text mining. In *Proceedings of the 19th International Conference for Management of Technology* (IAMOT 2010).
Kiyokawa, A., Yoshikane, F., Kawamura, S. & Suzuki, T. (2011). How activity of a researcher is influenced by conducting interdisciplinary research. In E. Noyons, P. Ngulube & J. Leta (Eds.), *Proceedings of the 13th International Conference of the International Society for Scientometrics & Informetrics* (ISSI 2011) (pp. 1005-1007). Leuven: ISSI.
Kostoff, R.N. (2007). The difference between highly and poorly cited medical articles in the journal Lancet. *Scientometrics*, 72(3), 513-520.
Lee, Y.-G., Lee, J.-D., Song, Y.-I. & Lee, S.-J. (2007). An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST. *Scientometrics*, 70(1), 27-39.
Narin, F. (1995). Patents as indicators for the evaluation of industrial research output. *Scientometrics*, 34(3), 489-496.
Peters, H.P.F. & van Raan, A.F.J. (1994). On determinants of citation scores: A case study in chemical engineering. *Journal of the American Society for Information Science*, 45(1), 39-49.
Snizek, W.E., Oehler, K. & Mullins, N.C. (1991). Textual and nontextual characteristics of scientific papers: Neglected science indicators. *Scientometrics*, 20(1), 25-35.
Suzuki, T. (2009). Extracting speaker-specific functional expressions from political speeches using random forests in order to investigate speakers' political styles. *Journal of the American Society for Information Science and Technology*, 60(8), 1596-1606.
WIPO (World Intellectual Property Organization) (2010). *International Patent Classification* (IPC). Retrieved October 13, 2012 from: http://www.wipo.int/classifications/ipc/en/.

Yoshikane, F., Suzuki, Y. & Tsuji, K. (2012). Analysis of the relationship between citation frequency of patents and diversity of their backward citations for Japanese patents. *Scientometrics*, 92(3), 721-733.

EXTENDING AUTHOR CO-CITATION ANALYSIS TO USER INTERACTION ANALYSIS: A CASE STUDY ON INSTANT MESSAGING GROUPS

Rongying Zhao [1] and Bikun Chen [2]

[1] *zhaory@whu.edu.cn*
Wuhan University, School of Information Management, Research Center for China Science Evaluation, The Center for the Studies of Information Resources, Luojia Hill, 430072Wuhan (China)

[2] *chenbikun2011@whu.edu.cn*
Wuhan University, School of Information Management, Research Center for China Science Evaluation, The Center for the Studies of Information Resources, Luojia Hill, 430072Wuhan (China)

Abstract
Author co-citation analysis (ACA) was an important method for discovering the intellectual structure of a given scientific field. While traditional ACA was mainly confined to the data of scientific literatures, such as ISI Web of Science, Google Scholar and so on. In this study, the idea and method of ACA was extended to web user information interaction research. Firstly, the development of ACA was briefly introduced. Then the sample data and method used in this study were given. Three QQ groups' instant messages of a Chinese company were selected as the raw data and the concepts and model of user interaction analysis (UIA) were proposed based on the data. Social network analysis method was used to measure the intensity of user information interaction. Operatively, Excel, Ucinet, Pajek and VOSviewer software were combined to analyze user information interaction intensity quantitatively and visually. Finally, it concluded that UIA model was relatively reasonable and was applicable to the web user research.

Conference Topic
Webometrics (Topic 7) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction
ACA (author co-citation analysis) was firstly introduced by White & Griffith（1981）. Different researchers applied it to detect intellectual structure of a given scientific field. For example, White & McCain (1998) used traditional techniques (multidimensional scaling, hierarchical clustering and factor analysis) to display the specialty groupings of 120 highly-cited information scientists. White (2003) used another kind of technique–Pathfinder Networks (PFNETs) to remap the paradigmatic information scientists with White & McCain's raw data from 1998. Jevremov et al. (2007) mapped the personality psychology as a

research field. Osareh & McCain (2008) studied the structure of Iranian chemistry research. Then some researchers extended ACA from the traditional citation databases to the Web environment. Leydesdorff & Vaughan (2006) started an exploratory research by selecting 24 authors of information science under web environment with Google Scholar. Qiu & Ma (2009); Ma et al. (2009) conducted studies of information science scholars in China with the Chinese Google Scholar. Obviously, data sources of the researches above are scientific literatures, such as ISI Web of Knowledge, Google Scholar, CNKI (China National Knowledge Infrastructure) and CSSCI (Chinese Social Sciences Citation Index). Besides, ACA was also applied in Webometrics in recent years. Zuccala (2006) compared ACA and Web Colink Analysis (WCA) by taking mathematics as the subject. He stated that although the practice of ACA might be used to inform a WCA, the two techniques did not share many elements in common. The most important departure between them existed at the interpretive stage when ACA maps became meaningful in light of citation theory, and WCA maps required interpretation based on hyperlink theory. Vaughan & You (2010) proposed a new Webometrics concept-Web co-word analysis to measure the relatedness of organizations by using the data from Google and Google Blogs. Wang et al. (2011) studied songs/singers co-collection relationship of online music web users by referring the co-citation analysis theory.

Previous researches on the analysis and practice of ACA were meaningful and have covered traditional citation databases, Google Scholar, Google search engine, Google Blogs, online music web and so on. But most of the research relied on the data of scientific literatures. In this study, it aimed to extend the idea and method of ACA to web user research and provided a new view to re-think the traditional bibliometric and scientometric method. So, a new kind of web users' data-QQ group instant messages of a Chinese company was selected as the raw data. In China, Tencent QQ is the most popular Instant Messaging product (detailed information about Tencent Inc. can be acquired in this portal: http://www.tencent.com/en-us/index.shtml). QQ group is one of typical applications launched by Tencent QQ. QQ group allows a group of people with the same interests, same job, same company or same department to instantaneously chat with certain topics. It also provides the users with other services: group BBS, group albums, shared files, group homepages and so on. Based on the raw data, UIA (User Interaction Analysis) model was proposed to measure the user information interaction intensity by referring the ACA theory. Social network analysis method and mapping and clustering techniques were applied to detect the user information interaction intensity.

Data and Method

Data

The sample data were derived from Tencent QQ groups in a company of China. The company focuses on the development and maintenance of computer hardware

and software, broadband network, web sites, telephone networks and television networks (detailed information can be get in this portal: http://www.pmcc.com.cn/). It owns about 300 employees and four departments: software department, system integration department, system security department and marketing department. In software department, there are 20 employees, including one manager, one deputy manager, three technical directors and fifteen ordinary staff.

In the enterprise, there are a variety of network relationships, which can be classified into formal network and informal network. Formal network refers to the network driven and formed by enterprise task and can be managed by enterprise, which is the specific reflection of the organizational structure. Informal network refers to the network formed spontaneously by the employees, not constrained by enterprise task, which is loose, unorganized, various and difficult to maintain (Xu, 2011). In addition, there are also some semi-formal networks in the enterprise, existing between the formal one and the informal one. This study has conducted the interview survey, finding that there were three main kinds of QQ groups in the company: department group, project team group in certain department, new employees group per year. Therefore, a simple stratified sampling method was applied to select sample data in terms of the three kinds above: software department group (group A), group of a project team in software department (group B), group of 2011 new employees (group C). According to the theories above, group A is formed by formal organization, group B is formed by semi-formal organization and group C is formed by informal organization. In the end, instant messages of group A, B and C were selected from October 1st, 2011 to February 29th, 2012 and provided by several instant messaging group users in the company. Then clean the sample data by deleting the invalid and redundant messages. The final sample data were counted as follows (in order to protect the privacy of the enterprise members, each member was identified by a number).

Table 1 Basic Statistics of Three Groups

QQ Group	*Message Count*	*Topic Count*	*Member Count*	*Member ID*
A	258	41	18	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
B	2184	131	6	14, 15, 16, 17, 18, 19
C	452	41	21	13, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36

From Table 1, there are certain relationships among the three instant messaging groups: Group A and B own five common members: 14, 15, 16, 17 and 18. Group B and C have three common members: 16, 17 and 18. Group A and C have four common members: 13, 16, 17 and 18. It is worth mentioning that member 1 is the software department manager; member 2 is deputy manager; member 3, 4 and 5 are technical directors; member 19 is the technical guide from a professional software company in China.

In the perspective of user interaction contents, QQ group instant messages are comprised by different kinds of topics, as well as the conversations in our daily life. How do we recognize the topics in QQ group instant messages? As we know, every topic has a time span. So, it is reasonable to cut instant messages into topics in terms of the messages' date and time. In this study, if the time interval between one message and the next message is 30 minutes or more, then cut them off. The segmentation method above stems from the hypothesis below: within half an hour or more, if no member in the instant messaging group speaks a word, a topic is over. In terms of this segmentation method, instant messages of group A, B and C are cut into 41, 131 and 41 pieces of topics respectively (shown in Table 1). In addition, member 6 has only two pieces of messages and has no contact with other member in instant messaging group. Considering member 6 is one of the only two females in group A and the further study below, this study perceived that member 6 has joined only one topic.

Concepts of UIA

White & Griffith (1981) summarized that the mapping of a particular area of science can be done using authors as units of analysis and the co-citations of pairs of authors as the variable that indicates their "distances" from each other. The analysis assumes that the more two authors are cited together, the closer the relationship between them. Co-citation of authors results when someone cites any work by any author along with any work by any other author in a new document of his own. Based on the descriptions above, the concepts of UIA are proposed. Specifically, the concepts of UIA rely on the hypothesis below: different members in an instant messaging group participate in a certain topic because they are interested in the topic or they are familiar with each other and willing to exchange their information. So, in this study, a piece of topic cut from QQ group instant messages can be seen as a journal article, any users included by the topic can be perceived as the authors of cited references (shown in Figure 1). Since ACA uses author co-citation count as a measure of the relatedness of authors' research, the concepts of UIA proposed in this study can be viewed as an extension of the concepts of ACA. However, the most important difference between them exists at the interpretive stage when ACA becomes meaningful in terms of citation theory and UIA requires interpretation based on user information behavior theory and social network theory.

Standard formula of user interaction intensity: user interaction intensity is defined as the relations between one member and any other member and is set as ϕ, Ψ is set as a certain topic, i and j are set as any two members in the instant messaging group. The intensity ϕ between member i and j is the sum of every minimum number of member i and j co-occurring frequency in any topic Ψ (Wang, 2011). Essentially, its mathematical idea is consistent with author co-citation.

$$\phi_{ij} = \sum_{\Psi} \min(\Psi i, \Psi j)$$

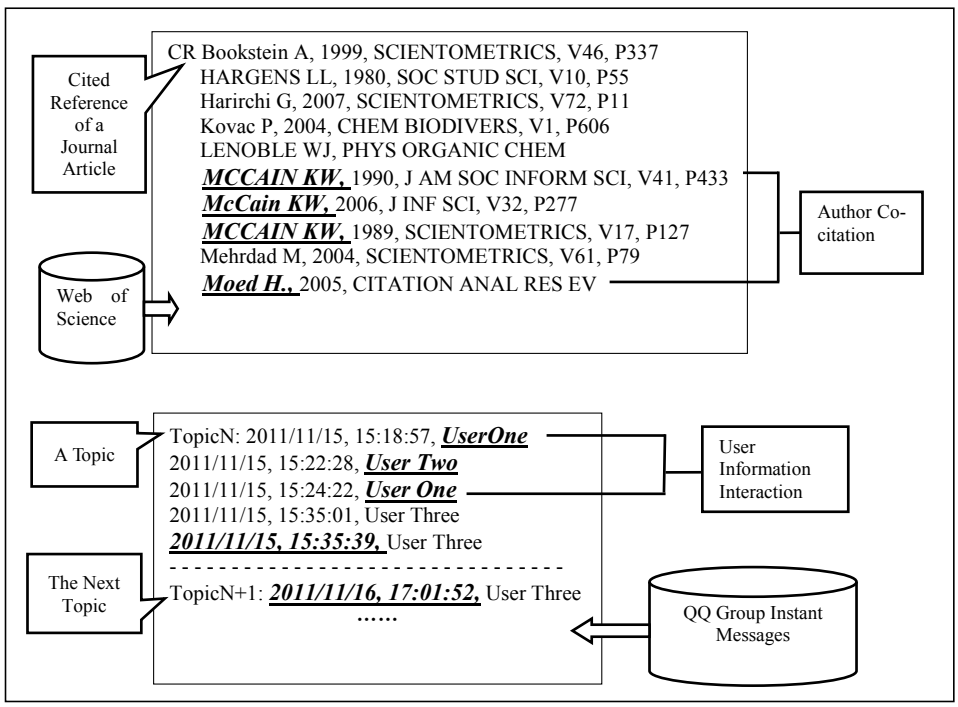


Figure 1 Author Co-citation and User Interaction

McCain (1990) summarized the steps in author co-citation analysis: selection of the author set, retrieval of co-cited author counts, compilation of raw co-citation matrix, conversion to correlation matrix, multivariate analysis of correlation matrix and interpretation & validation, which is called traditional model. The basic steps of the model are comprised by compilation of raw co-citation matrix, conversion to correlation matrix and multivariate analysis of correlation matrix. Social network analysis (SNA) method was also verified to be applicable to the co-citation research (e.g., Xu & Zhu, 2008; Groh & Fuchs, 2011). Therefore, in this study, the UIA steps included: compilation of raw co-citation matrix, conversion to correlation matrix, social network analysis and visualization.

Tools

In bibliometric and scientometric research, a lot of attention is paid to the analysis of networks of, for example, documents, keywords, authors or journals. Mapping and clustering techniques are frequently used to study such networks. Waltman et al. (2010) firstly presented their proposal for a unified approach to mapping and clustering. In the bibliometric and scientometric literature, the most commonly used combination of a mapping and a clustering technique is the combination of

multidimensional scaling and hierarchical clustering by SPSS software (for early examples, see White & Griffith, 1981; Small et al., 1985; McCain, 1990; Peters & Van Raan, 1993). However, various alternatives to multidimensional scaling and hierarchical clustering have been introduced in the literature, especially in more recent work, and these alternatives are also often used in a combined fashion. A popular alternative to multidimensional scaling is the mapping technique of Kamada and Kawai algorithm (1989) by Pajek software; (e.g. Leydesdorff & Rafols, 2009; Noyons & Calero-Medina, 2009; Leydesdorff, Kushnir & Rafols, 2012), which is sometimes used together with the pathfinder network technique (e.g. Schvaneveldt, Dearholt & Durso, 1988; Chen, 1999; White, 2003; de Moya-Anegon et al., 2007). Two other alternatives to multidimensional scaling are the VxOrd mapping technique (e.g., Boyacketal., 2005; Klavans & Boyack, 2006) and VOSmapping technique of VOSviewer software (e.g., Van Eck et al., 2010). Factor analysis, which has been used in a large number of studies (e.g., de Moya-Anegon et al., 2007; Zhao & Strotmann, 2008; Leydesdorff & Rafols, 2009), may be seen as a kind of clustering technique and, consequently, as an alternative to hierarchical clustering. Another alternative to hierarchical clustering is clustering based on the modularity function of Newman and Girvan (2004); (e.g.Wallace, Gingras & Duhon, 2009; Zhang et al., 2010). As to the mapping and clustering software, Leydesdorff et al. argued that Gephi and VOSviewer offer superior visualization techniques (Leydesdorff, Kushnir & Rafols, 2011), while Gephi and Pajek/ Ucinet offer network statistics. However, the comparison made us realize that with little effort we could also make our outputs compatible with Pajek, and via Pajek also for Gephi (which read Pajek files). This offers additional flexibilities such as using algorithms for community detection among a host of other network statistics which are available in Pajek and Gephi, but not in VOSviewer (Leydesdorff et al., 2012).

According to the theories and practices above, in this study, Excel, Ucinet, Pajek and VOSviewer software were combined to analyze user interaction intensity quantitatively and visually. Excel VBA programming was used to construct user interaction matrix in terms of standard formula of user interaction intensity above. Ucinet was applied to read the matrix and generate .net file. Pajek was used to load the .net file to draw user interaction figure. VOSviewer was further applied to visualize the user interaction figure with its own clustering algorithm based on modularity optimization. Ucinet and Pajek were combined to offer network statistics.

Results

In Pajek, user interaction network was visualized with the spring-based algorithm of Kamada and Kawai (1989). This algorithm reduces the stress in the representation in terms of seeking to minimize the energy content of the spring system. In the user interaction figure, every node signifies a member, the size of every node means its degree centrality in the network, the position of every node in the network (in the center or in the edge) signifies its importance, and the

thickness of the line between two nodes signifies its interaction intensity, the distance between one node and any other node signifies its closeness. In addition, different color signifies different groups (obtained by K-core analysis). A subset of vertices is called a k-core if every vertex from the subset is connected to at least k vertices from the same subset. Cores in Pajek can be computed using "Network/ Create Partition/ k-Core/ All". Result is a partition: for every vertex its core number is given (shown in Table 2).

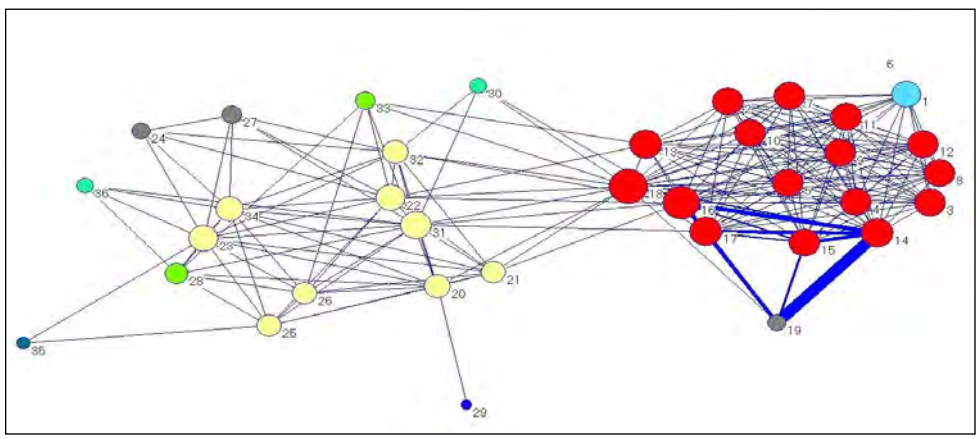


Figure 2 User Interaction Network by Pajek

Table 2 Clusters of User Interaction Network

Cluster	*Member*	*Freq%*
1	1	2.7778
2	2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18	44.4444
3	6	2.7778
4	19, 24, 27	8.3333
5	20, 21, 22, 23, 25, 26, 30,31, 32, 34	25.0000
6	28, 33	5.5556
7	29	2.7778
8	30, 36	5.5556
9	35	2.7778

In Figure 2, member 16 and 18 have the highest degree centrality mainly because they are the common member of Group A, B and C and they contact frequently and broadly with other members. On the whole, group A and B have a higher density and size than group C, which indicates that group A and B have the higher information interaction intensity than group C. Member 19 is special, scattering in the edge of group A and B. In reality, member 19 is the technical guide from a professional software company in China. In addition, there are nine groups (clusters) in the network in terms of their color and member 1, 6, 29 and 35 forms

a single group (clusters) respectively (shown in Table 2). In the company, member 1 is the software department manager but he belongs to a single group (clusters). It is advisable for him to contact more with others members in the department to promote user interaction. Member 6 is one of the only two females in group A and she has few contacts with other members in the software department. Besides, member 19, 24 and 27 belong to a cluster but they belong to different instant messaging groups, which is inconsistent with the reality. The reason lies in that the user interaction network is the combined network of the three groups, the k-core analysis is applied to the combined network.

Network Density and User Interaction Intensity

In the density view, items are indicated by a label. Each point in a map has a color that depends on the density of items at that point. That is, the color of a point in a map depends on the number of items in the neighborhood of the point and on the importance of the neighboring items. By default, VOSviewer uses a red-green-blue color scheme (see Fig. 3). In this color scheme, red corresponds with the highest item density and blue corresponds with the lowest item density. The density view is particularly useful to get an overview of the general structure of a map and to draw attention to the most important areas in a map (Van Eck et al., 2010). In Figure 3, areas of member 1 to 18 (member 6 excluded) and member 22 & 31 turn out to be important. These areas are very dense, which indicates that overall the information interaction intensity among these members are highest. It can also be seen that there is a clear separation between the areas of group A and C.

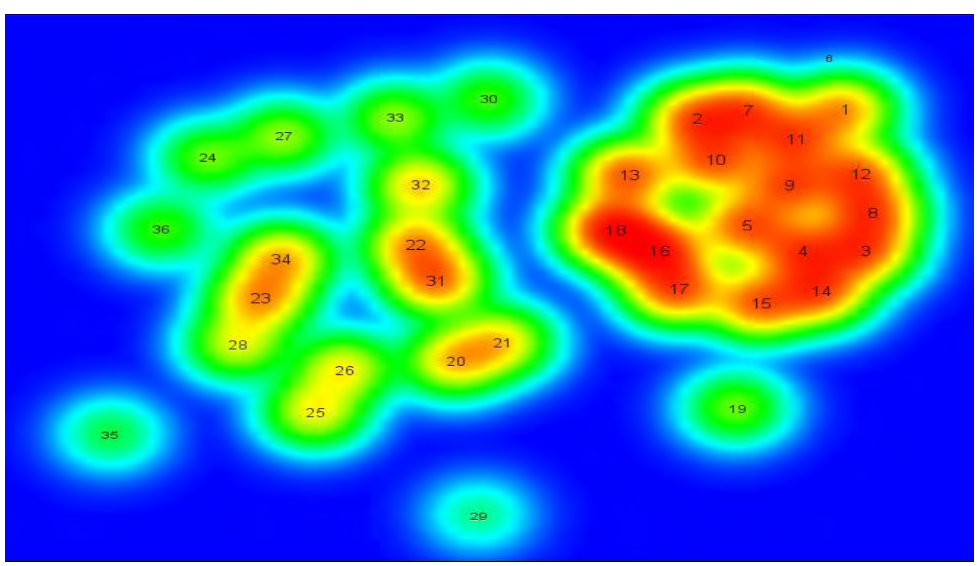


Figure 3 Density View by VOSviewer

Centrality and User Interaction Intensity

In Ucinet software, centrality is measured by "Network-Centrality-Closeness, Degree, Betweenness, Eigenvector and so on", in this study, only "Closeness, Degree and Betweenness" are discussed.

Table 3 Centrality of User Interaction Network (Top 10)

Member ID	*nBetweenness*	*Member ID*	*nCloseness*	*Member ID*	*NrmDegree*
18	22.697	18	43.21	14	6.133
16	13.42	16	41.667	19	5.248
31	12.998	31	39.326	16	2.845
22	8.7	17	38.889	17	2.346
20	7.791	13	38.889	15	2.24
23	4.545	22	38.043	18	0.958
32	4.323	32	37.634	31	0.371
13	4.03	20	37.634	23	0.281
17	3.606	21	37.234	32	0.273
21	2.944	14	35.714	7	0.186

Betweenness centrality of an actor is the extent to which an actor serves as a potential "go-between" for other pairs of actors in the network by occupying an intermediary position on the shortest paths connecting other actors. Closeness centrality of an actor is the extent to which the most direct paths connecting an actor to each of the actors in a network are short rather than long. Degree centrality is the number of connections that an actor has in a network (Kilduff & Tsai, 2003). As a word, they show the importance of the member in the user interaction network. In Table 3, member 18, 16, 31, 32 and 17 enter the centrality of combined network Top 10, which show that they occupy a central position in the combined network. As the important members in the network, they control the most information flow and have the advantages to contact more with other members. So, it is advisable for them to contact more with others members, improving the information interaction atmosphere and environment.

Structural Holes and User Interaction Intensity

In Ucinet software there are two ways to detect structural holes: "Network-Centrality-Freeman Betweenness-Node Betweenness" and "Network-Ego Networks-Structural Holes".

Table 4 Structural Holes of User Interaction Network

Method	*Structural Holes*
Node Betweenness	18, 16, 31, 17, 13
Structural Holes	18, 16, 13, 14, 17

From Table 4, the results of the two detection methods are consistent: their common members are 18, 16, 13 and 17. Structural Holes shows the situation of an actor as the middleman, controlling the enterprise information flow, which play a vital role in the user interaction. So the four members act as the middleman of the overall user interaction network. In reality, the four members are new staff in 2011 and they are also the common members in group A and B. In short, the reality and the detection results are consistent. Besides, member 14 and 31 are the structural holes of group A and B respectively.

Conclusions and Discussions

In a word, for user interaction network (shown in Figure 2), it is easy to distinguish group A, group B and group C. The three groups are linked together by member 18, 16, 13 and 17, which are the information interaction middleman of the overall network. Also, member 14 and 31 are the information interaction middleman of group A and group C respectively. If the six members keep on communicating frequently with others, the atmosphere of information interaction will be better. Besides, member 6 is isolated in the overall network (shown in Figure 2), who contacts less with others. So, the business managers should pay more attentions the core node, isolated node and structural holes in the overall network and take certain measures to tackle the problems and promote user information interaction.

In this study, the idea and method of ACA in bibliometric and scientometric research are extended to web user research and the results are consistent to the company reality, which proves that the UIA model is relatively reasonable and UIA is applicable to the web user research. Also, social network analysis method can quantitatively and visually diagnose user information interaction status and guide the managers to deal with the problems in their company.

Although the study focused on enterprise entities, the UIA could potentially be applied to other types of organizations such as universities or governments. A limitation of the study is that it only tested the UIA in a particular company. More studies in other areas are needed to determine the applicability of the UIA. More qualitative studies are also needed to enrich this quantitative study and to gain a deeper understanding on the relative pros and cons of traditional bibliometric and scientometric methods.

Acknowledgments

This paper is supported by Major Program of National Social Science Foundation in China (11&ZD152) and High-level International Journal Program of Wuhan University (2012GSP062).

References

Boyack, K.W., Klavans, R. & Borner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351–374.

Chen, C. (1999).Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, 35(3), 401–420.
de Moya-Anegon, F., Vargas-Quesada, B., Chinchilla-Rodriguez, Z., Corera-Alvarez, E., Munoz-Fernandez, F.J. & Herrero-Solana, V.(2007).Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology*, 58(14), 2167–2179.
Groh, G. & Fuchs, C. (2011). Multi-modal social networks for modeling scientific fields. *Scientometrics*, 89(2), 569-590.
Jevremov, T., Pajic, D. & Sipka, P. (2007).Structure of personality psychology based on cocitation analysis of prominent authors. *Psihologija*, 40(2), 329-343.
Kamada, T. & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1), 7–15.
Kilduff, M. & Tsai, W. (2003). *Social Networks and Organizations*. London, England: Sage.
Klavans, R. & Boyack, K.W. (2006). Quantitative evaluation of large maps of science. *Scientometrics*, 68(3), 475–499.
Leydesdorff, L., Kushnir, D. & Rafols, I. (in press). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics*.
Leydesdorff, L., Hammarfelt, B. & Salah,A.A.A. (2011). The structure of the Arts & Humanities Citation Index: A mapping on the basis of aggregated citations among 1,157 journals. *Journal of the American Society for Information Science and Technology*, 62(1), 2414-2426.
Leydesdorff, L & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, 60(2), 348–362.
Leydesdorff, L. &Vaughan, L. (2006).Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and Technology*, 57(12), 1616-1628.
Ma, R.M., Dai, Q.B., Ni, C.Q. & Li X.L. (2009).An author co-citation analysis of information science in China with Chinese Google Scholar search engine, 2004-2006.*Scientometrics*, 81(1), 33-46.
McCain, K.W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433-443.
Newman, M.E.J. & Girvan, M. (2004).Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 026113.
Noyons, E.C.M. & Calero-Medina,C.(2009).Applying bibliometric mapping in a high level science policy context. *Scientometrics*, 79(2), 261–275.
Osareh, F. & McCain, K.W. (2008). The Structure of Iranian Chemistry Research, 1990-2006: An Author Cocitation Analysis. *Journal of the American Society for Information Science and Technology*, 59(13), 2146-2155.

Peters, H.P.F. &Van Raan, A.F.J. (1993).Co-word-based science maps of chemical engineering. Part II: Representations by combined clustering and multidimensional scaling. *Research Policy*, 22(1), 47–71.
Qiu, J.P. & Ma, R.M. (2009). The application of ACA method in web environment. *Library and Information Service*, 52(2): 85-87. (in Chinese).
Schvaneveldt, R.W., Dearholt, D.W. & Durso, F.T. (1988).Graph theoretic foundations of pathfinder networks. *Computers and Mathematics with Applications*, 15(4),337–345.
Small, H., Sweeney, E. & Greenlee, E. (1985).Clustering the Science Citation Index using co-citations. II. Mapping science. *Scientometrics*, 8(5-6), 321–340.
VanEck, N.J. &Waltman,L.(2010).Softwares urvey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.
Vaughan, L. & You, J. (2010). Word co-occurrences on Webpages as a measure of the relatedness of organizations: A new Webometrics concept. *Journal of Informetrics*, 4(4), 483-491.
Wallace, M.L., Gingras,Y. & Duhon, R.(2009).A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the American Society for Information Science and Technology*, 60(2), 240–246.
Waltman, L., Van, Eck N.J. & Noyons, E.C.M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4), 629-635.
Wang, X.W., Hu, Z.G., Ding, K. & Liu, Z.Y. (2011). Research on classification of singers in online music websites based on co-citation theory. *Journal of the China Society for Scientific and Technical Information*, 30(5), 471-478. (in Chinese).
Wang, Z.F. (2011). Use and Management of Internet Communication Tools in Community Constructions- Take L Community in Hangzhou for Example. Unpublished master's thesis, Zhejiang Gongshang University, Hangzhou, China. (in Chinese).
White, H.D. & Griffith, B. (1981). Author cocitation: A literature measure of intellectual structures. *Journal of the American Society for Information Science*, 32(3), 163-171.
White, H.D. & McCain, K. (1998). Visualizing a discipline: An author cocitation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, 49(4), 327–355.
White, H.D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54(5), 423-434.
Xu, L.M. (2011). Research on Mechanism of Enterprise Internal Knowledge Sharing Based on Social Network. Unpublished master's thesis, Wuhan University, Wuhan, China. (in Chinese).

Xu, Y.Y. & Zhu, Q.H. (2008). Demonstration study of social network analysis method in citation analysis. *Information Studies: Theory & Application*, 31(2), 184-188. (in Chinese).

Zhang, L., Liu, X., Janssens, F., Liang, L. & Glanzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, 4(2), 185–193.

Zhao, D. & Strotmann, A. (2008). Information science during the first decade of the Web: An enriched author cocitation analysis. *Journal o f the American Society for Information Science and Technology*, 59(6), 916–937.

Zuccala, A. (2006). Author cocitation analysis is to intellectual structure as web colink analysis is to ... ?. *Journal of the American Society for Information Science and Technology*, 57(11), 1487-1502.

EXTENDING CITER-BASED ANALYSIS TO JOURNAL IMPACT EVALUATION

Kun Lu[1], Isola Ajiferuke[2] and Dietmar Wolfram[3]

[1] *kunlu@whu.edu.cn*
School of Information Management, Wuhan University, Luojiashan Road No. 1, Wuhan, Hubei, China, 430072

[2] *iajiferu@uwo.ca*
Faculty of Information and Media Studies, University of Western Ontario, London, ON, Canada N6A 5B7

[3] *dwolfram@uwm.edu*
School of Information Studies, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI U.S.A. 53201

Abstract
The concept of citer analysis investigated earlier by Ajiferuke and Wolfram (2009, 2010) is extended to journals where different citing units (citers, citing articles, citing journals) are compared with the journal impact factor and each other to determine if differences in ranking arise from different measures. The citer measures for the 31 high impact journals studied are significantly correlated, even more so than the earlier citer analysis findings, indicating that there is a close relationship among the different units of measure. Still, notable differences in rankings for the journals examined were evident for the different measures used, indicating that a journal's impact can be relative depending on the measure used. Overall, citer analysis at the journal level appears to offer less distinctive results than at the author level.

Conference Topic
Scientometrics Indicators - Criticism and new developments (Topic 1); Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3)

Introduction and Previous Research
Academic journal standing and prestige are determined at least in part by assessment measures based on citations. The most well-known is the journal impact factor (Garfield & Sher, 1963), which has long served as a benchmark by which the significance of journals has been assessed. The stakes can be high in this assessment exercise. How journals are ranked can have consequences for journal publishers. Libraries with limited budgets may base their purchase decisions in part on the perceived prestige of journals as determined by impact factors. Similarly, authors' decisions on where to submit the outcomes of their research may be based on the standing of a given journal. The journals in which an author publishes, in turn, play a role in how the authors themselves are assessed, in particular for promotion and tenure in academe. Journal impact

assessment has increasingly become a controversial topic, with greater research investigation of impact factors, particularly since the mid-1990s (Archambault & Larivière, 2009). Measures developed to assess journal impact are argued to be misused (Pendlebury, 2009) or have shortcomings (Glänzel & Moed, 2002). The ongoing debate is recently evident in Vanclay (2012), who calls for a major overhaul of the traditional impact factor based on an analysis of its weaknesses. Rousseau (2012) echoes this sentiment by recognizing that improvements are needed, but with no clear solutions at present. In response to Vanclay, Moed et al. (2012) address the value of journal assessment and outline several measures that may serve as complementary to the existing journal impact factor employed by Thomson Reuters in its *Journal Citation Reports*. Other measures for assessing journal impact and quality have been proposed, as outlined by Rousseau (2002). Bollen, Van de Sompel, Smith and Luce (2005), for example, outline metrics based on author/reader and frequency/structure dimensions using download counts of journal contents as well as social network metrics to rank journals as alternative measures.

The study of author impact has been equally longstanding, with equal controversy. Citation counts and indices such as the h-index (Hirsch, 2005) and its variants have been developed to assess and compare the influence of authors. Issues of citer motivation, self-citation, how citations are counted, to name a few, have been perennial issues discussed in citation analysis (MacRoberts & MacRoberts, 1989). In previous studies, the present authors have promoted the use of citer-based measures to assess impact because citation counts on their own do not take into account the origin of the citations--aside from self-citations--and do not reflect the reach of an author or a work (Ajiferuke & Wolfram, 2009; Ajiferuke & Wolfram, 2010; Ajiferuke, Lu, & Wolfram, 2010). This idea of counting citers is not new, going back at least to the 1970s (Dieks & Chang, 1976), but has not been widely studied or implemented to date. In the more recent citer analysis studies, the authors found that there is a strong correlation between citer and citation-based measures, but that some authors' rankings among their peers could vary widely using citation-based or citer-based measures. Ajiferuke and Wolfram observed that the influence of some the issues associated with citation analysis may be reduced. For example, their proposed citer-based h index (ch index) provided a means of assessing author impact or reach by excluding self-citations and recurrent citers (i.e., those who cite the same work multiple times). Franceschini et al. (2010) further explored the ch index concluding that it offered a complementary measure to the h index. Egghe (2012) noted that there is a linear relationship between the proposed citer h index and the more traditional citation-based h index.

Ajiferuke and Wolfram (2009, 2010) found that there were some notable differences in the ranking of authors when comparing citation and citer-based counts. Does the same apply to citer analysis in the context of journals? With journals, there are additional measures at different levels of granularity that could

be used to count impact or reach based on the number of citers, citing articles and citing journals. In this study we explore the idea of citer-based measures for the ranking of journals to determine if these measures notably change rankings by relying on a different perspective of the citing process. More specifically, this study asks:

1) Do citer-based measures of journal impact provide alternative or complementary measures to traditional citation-based approaches such as the journal impact factor?
2) Does the level of granularity of the citer-based measure (citer/author, article, journal) influence journal ranking outcomes when compared with other measures and, if so, to what extent?

Method

Data were collected from Thomson Reuters Web of Science (WoS). Top journals with impact factors of greater than 0.5 were selected from the subject category Information Science & Library Science from the 2010 Journal Citation Reports (JCR) Social Sciences Edition for this initial exploration. The impact factor of 0.5 was selected to provide a sufficient body of citations. The inclusion of lower impact journals could result in spurious outcomes for other measures. Journals associated with allied subject areas such as Management Information Systems and Medical Informatics were excluded. Thirty one journals were included in the present study. A list of the journals and abbreviations is provided in Appendix 1. A focus on journals from a familiar field to the authors provides the opportunity to explore the feasibility and outcomes of this explored area for further study in broader areas.

Searches were conducted in WoS for the publications in these journals between 2007 and 2011. Only three types of documents were kept: articles, reviews and conference proceedings. The other document types were considered less likely to represent research contributions. Using the "Create Citation Report" function provided by WoS, we obtained the citing articles of each journal on the list. It should be noted that citing articles from the journal itself were included here as they are still considered as the citations to the journal. Next, we used the "Analyze Results" function on these citing articles to collect the citers for each journal. We relied on WoS to produce the list of citers in the study. The problem of author name disambiguation has been widely discussed in the literature (Smallheiser & Torvik, 2009). To determine the impact of the ambiguous author names on our study, we implemented a simple but effective author name disambiguation algorithm proposed by Strotmann, Zhao and Bubela (2009) and compared the results with the ones produced by WoS. Only slight differences were found between them, with no more than a few percent difference. Therefore, we decided to stick to the WoS outputs for the citer data. Impact factors and five years impact factors of the journals were also collected from JCR 2010 for further analysis.

Comparative analyses were conducted on the collected data using several available and derived measures including: number of publications, number of citing articles, number of citers, number of citing journals, journal impact factor, and 5-year impact factor. Correlation analyses were carried out and differences in rankings based on each measure were tabulated for comparison.

Results

Appendix 2 summarizes the number of publications indexed by WOS over the 5-year period for each journal as well as the citing figures for these journals.

- The number of publications indexed per journal varies from 61 to 937
- The number of citing articles ranges from 37 to 3340
- The number of citers varies from 74 to 5536
- The number of citing journals ranges from 24 to 934

Of note, the maximum values for all these variables are for the journal *JASIST* while *ARIST* has the minimum number of publications, *Online* with the minimum number of citing articles as well as the minimum number citers but *Law Library Journal* has the minimum number of citing journals. The median values for the number of indexed publications, number of citing articles, number of citers, and number of citing journals are 162, 271, 521, and 116 respectively. Given the varying number of publications indexed, the citing values needed to be normalized by the number of publications for a meaningful comparison to be made among the journals. The normalized values along with the impact factor and 5-year impact factor can be found in Appendix 3. Rankings appearing in the tables below are based on the corresponding values from this appendix.

We next examined the correlation between the three citing indices and the two popular journal impact indices. The correlation coefficients are shown in Table 1. (Note: Although not strictly a random sample, the data collected by WoS do represent subsets of the overall population.) Looking at the correlation coefficients between any of the citing indices and either of the journal impact indices, we observed that the highest correlation exists between the 5-year impact factor and the number of citing articles per publication. This is not surprising given that the definitions for both are quite similar except that one value was calculated from our data while the other was obtained from *Journal Citation Reports*.

Although the correlations are quite high, a comparison of the change in a journal's ranking between the 5-year impact factor and each of the citer-based measures reveals that there can be sizeable differences between the ranks (Table 2). Three journals experience a difference of more than five places for citing articles per publication, four for number of citers per publication, and twelve for number of citing journals per publication. In the case of the number of citing journals per publication, *Information Research*, *Portal: Libraries and the Academy*, and

Scientometrics saw the largest drop in their rankings, indicating that the number of citing journals was relatively smaller than for other journals with lower impact factors. Conversely, *Health Library and Information Journal*, *Library Collections Acquisitions & Technical Services* and *Social Science Information* showed the greatest gain, indicating that although they receive relatively fewer citations, they are cited proportionately by a larger number of journals.

Table 1. Spearman Correlation coefficients between citing indices and popular journal impact indices

	Impact Factor	*5-year Impact Factor*	*# of Citing Articles per Publication*	*# of Citers per Publication*	*# of Citing Journals per Publication*
# of Citing Articles per Publication	.818 (.000)*	.910 (.000)	-	.957 (.000)	.875 (.000)
# of Citers per Publication	.734 (.000)	.860 (.000)	.957 (.000)	-	.890 (.000)
# of Citing Journals per Publication	.652 (.000)	.764 (.000)	.875 (.000)	.890 (.000)	-

* Significance level in parentheses

Table 2. Change in Journal Rank Based on 5-year Journal Impact Factor and Citer Measures

# of Citing Articles per Publication		*# of Citers per Publication*		*# of Citing Journals per Publication*	
Change in Rank	# of Journals	Change in Rank	# of Journals	Change in Rank	# of Journals
+5	1	+12	1	+10	3
+4	5	+5	2	+8	1
+3	3	+4	4	+6	2
+2	4	+3	3	+5	2
+1	5	+2	2	+4	2
-1	4	+1	3	+3	2
-2	2	0	3	+2	2
-3	2	-1	3	+1	2
-4	2	-2	2	-2	5
-6	1	-3	4	-3	3
-7	1	-5	1	-5	1
-12	1	-6	1	-6	1
		-12	2	-8	2
				-10	1
				-11	1
				-13	1

* 5-year Journal Impact Factor Rank Minus Citing Measure Rank

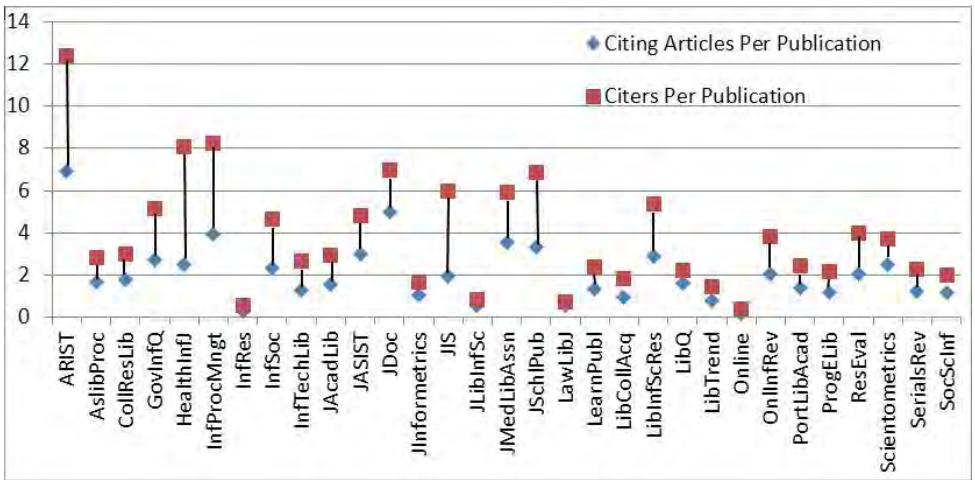


Figure 1. Comparison of number of citing articles per publication and number of citers per publication

Table 3. Change in journal rank between number of citing articles per publication and number of citers per publication

*Change**	*Number of Journals*
+8	1
+7	1
+3	1
+2	1
+1	4
0	13
-1	5
-2	1
-3	1
-4	1
-5	2

* Number of citing articles per publication Rank – Number of citers per publication Rank

We next used the number of citing articles per publication as the usual journal impact index, and then correlated it with the other two citing indices. The correlation between the number of citing articles per publication and the number of citers per publication is very high (see also Figure 1), and in fact if we examine the change in journal ranks from one index to another, we noticed that 22 out of the 31 journals (i.e. about 71%) either did not change position or moved only one place up or down (see Table 3). There were fewer more dramatic changes than observed for the impact factor comparison in Table 2 above. What this means is that for most of these journals neither were there many citers responsible for a lot of the citations nor was the overlap in the authors of the citing articles very limited. The first scenario is observed

with *Scientometrics*. As with the number of citing journals per publication, it is one of the two journals with the largest drop in rank (see Table 4) while the second scenario applies to the *Journal of the Medical Library Association* that has the highest rise in rank.

Table 4: Citer concentration for Scientometrics

Number of Citer Occurrences	*Number of Citers*	*Number of Citer Occurrences*	*Number of Citers*
47	1	15	3
42	1	14	2
41	1	13	5
40	1	12	6
32	1	11	6
30	2	10	8
29	1	9	18
28	1	8	21
27	1	7	8
23	1	6	21
21	1	5	42
19	3	4	70
18	2	3	142
17	1	2	430
16	4	1	2513

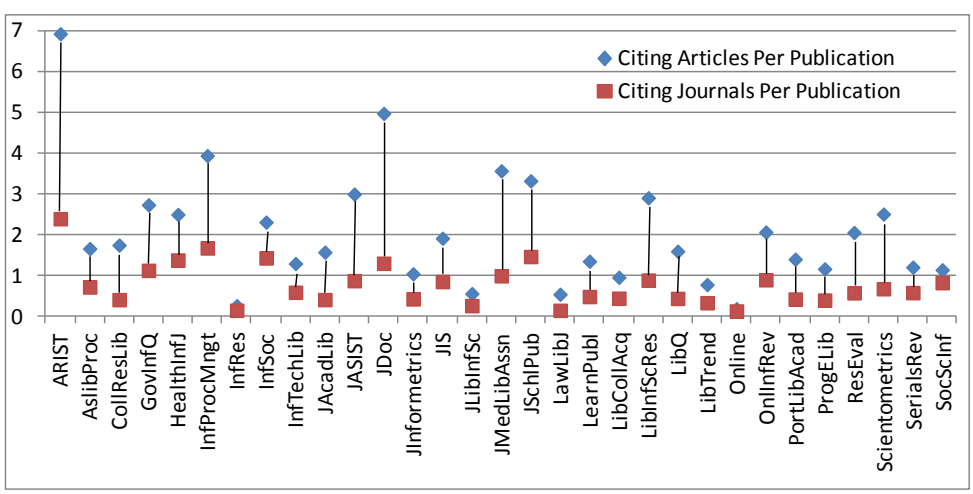


Figure 2. Comparison of number of citing articles per publication and number of citing journals per publication

The correlation between the number of citing articles per publication and the number of citing journals per publication is also very high (see Figure 2), though not as high as for the number of citers per publication. The change in ranks from

the number of citing articles per publication to the number of citing journals per publication is also more varied (see Table 5). Here, we have *Social Science Information* moving up 10 places in rank while *College & Research Libraries* moved down 11 places. *Social Science Information* ranked much higher in terms of the number of citing journals because most of the citations received were not concentrated in a few journals while *College & Research Libraries* ranked much lower because about 50% of its citations came from seven journals (see Table 6).

Table 5: Change in journal rank between number of citing articles per publication and number of citing journals per publication

*Change**	*Number of Journals*
+ 11	1
+ 7	1
+ 6	1
+ 5	3
+ 3	2
+ 2	3
+ 1	4
0	4
-1	1
-2	0
-3	2
-4	4
-5	2
-6	2
-10	1

* Number of citing articles per publication Rank Minus Number of citing journals per publication Rank

Table 6: Citing journal concentration for College & Research Libraries

Number of Citing Journal Occurrences	*Number of Citing Journals*
38	1
31	1
20	1
16	1
15	1
11	1
9	1
8	1
6	4
5	2
4	8
3	4
2	7
1	31

Discussions & Conclusions

As the findings demonstrate, journal ranking can be largely dependent on the assessment measures used. Unlike the citer analysis measures for authors discussed earlier where a few non-significant correlations were found between selected measures, citer-based measures for journals are even more highly correlated, whether examined at level of citer, article or journal level. Despite the high correlations, notable differences in the ranking of journals can be found for citer-based measures. The observed large differences in ranks between the 5-year journal impact factor and number of citing journals per publication demonstrate that some journals may attract a more modest number of citations than other journals, but those citations represent a broader array of journals. The range of 5-year impact factor values for those journals with large ranking differences between the impact factor and citing journals per publication indicates that these differences for specific journals are not tied to whether a journal is highly cited or not. The number of citing journals per publication, surely, also represents a measure of the reach of a cited journal that may not be evident in the number of citers alone or other singular measure of impact. Just as citer tallies take into account the origin of the citations and do not provide additional credit for repeated citations by the same individual (Ajiferuke, Lu, & Wolfram, 2011), examining citer patterns at the journal level can provide a higher level and less granular indication of the reach of a journal. Large differences in rankings when comparing citing articles per publication and citing journals per publication provide an indication that the citing practice for some journals may be very different, favouring some journals over others for given measures. This is demonstrated by other measures gaining popularity for journal assessment such as the Eigenfactor (Bergstrom, West, & Wiseman, 2008) or SJR indicator (Gonzalez-Pereira, Guerrero-Bote, & Moya-Anegón, 2008). The purpose of the current research was not to compare the citer outcomes with these newer measures--because they assess journals in different ways--but to look at how a different perspective on traditional citations may provide additional insights into journal impact or reach. The assessment of journal impact or reach is a multi-dimensional concept with relative points of view for assessment.

One limitation of the present study arises from the focus on a single discipline. As an exploratory study, it's natural to focus on a subject area of expertise. Results for library and information science would indicate that there is not much difference between citer-based and more traditional journal assessment measures, and therefore may not be worth further study. Data could also be collected for other disciplines where levels of co-authorship may vary and, which could then influence individual citer outcomes but may not influence the number of citing articles or citing journals. Also, the observed relationships among the different citation and citer assessment measures may change over time. This study examined a recent snapshot of publications. With the growth in the number of journals and researchers contributing to those journals, the currently observed

differences based on citers and journals may only grow, much in the same way that journal impact factors continue to rise over time (Althouse, West, Bergstrom, & Bergstrom, 2008).

The ongoing debates over journal impact measures will undoubtedly continue. The stakes for recognition can be high from an academic perspective, where editors vie to attract the best research to increase the impact of their journals, and authors compete to be published in the most prestigious journals in their fields. Citer-based measures for journals may not offer substantial differences than more traditional citation-based measures, but they can provide complementary assessment outcomes, or confirmatory measures that strengthen the journal assessment process.

References

Ajiferuke, I., Lu, K., & Wolfram, D. (2010). A comparison of citer and citation-based measure outcomes for multiple disciplines. *Journal of the American Society for Information Science and Technology*. 61(10), 2086-2096.

Ajiferuke, I., Lu, K., & Wolfram, D. (2011). Who are the disciples of an author? Examining recitation and oeuvre citation exahaustivity. *Journal of Informetrics*, 5(2), 292-302.

Ajiferuke, I., & Wolfram, D. (2009). Citer analysis as a measure of research impact: Library and information science as a case study. In B. Larsen & J. Leta (Eds.), *Proceedings of the 12th international conference of the international society for scientometrics and informetrics* (ISSI) (pp. 798–808). Rio de Janeiro.

Ajiferuke, I., & Wolfram, D. (2010). Citer analysis as a measure of research impact: Library and information science as a case study. *Scientometrics*, 83(3), 623-638.

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2008). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, 60(1), 27-34.

Archambault, E., & Larivière, V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics*, 79, 635–649.

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The Eigenfactor metrics. *The Journal of Neuroscience,* 28(45), 11433-11434.

Bollen, J., Van de Sompel, H., Smith, J.A., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management*, 41(6), 1419-1440.

Dieks, D. and Chang, H. (1976). Differences in impact of scientific publications: Some indices derived from a citation analysis. *Social Studies of Science*, 6, 247-267.

Egghe, L. (2012). A rationale for the relation between the citer h-index and the classical h-index of a researcher. *Scientometrics*, 1-4.

Franceschini, F., Maisano, D., Perotti, A., and Proto, A. (2010). Analysis of the ch-index: an indicator to evaluate the diffusion of scientific research output by citers. *Scientometrics*, 85(1), 203-217.

Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14(3), 195-201.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics*, 53(2), 171-193.

Gonzalez-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, 4(3), 379-391.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102, 16569-16572.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5), 342–349.

Moed, H. F., Colledge, L., Reedijk, J., Moya-Anegon, F., Guerrero-Bote, V., Plume, A., & Amin, M. (2012). Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way. *Scientometrics*, 92(2), 367-376.

Pendlebury, D.A. (2009). The use and misuse of journal metrics and other citation indicators, *Archivum Immunologiae et Therapiae Experimentalis*, 57(1), 1-11.

Rousseau, R. (2002). Journal evaluation: Technical and practical issues. *Library Trends*, 50(3), 418-439.

Rousseau, R. (2012). Updating the journal impact factor or total overhaul? *Scientometrics*, 92, 413–417.

Smalheiser, N.R., & Torvik, V.I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, 43(1), 1-43.

Strotmann, A., Zhao, D., & Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. In *Proceedings of the American Society for Information Science and Technology*, 46(1), 1-20.

Vanclay, J.K. (2012). Impact factor: outdated artefact or stepping-stone to journal certification? *Scientometrics*, 92, 211–238.

Appendix 1

List of journals studied

Journal Abbreviation	*Journal Name*
ARIST	Annual Review of Information Science and Technology (ARIST)
AslibProc	Aslib Proceedings
CollResLib	College & Research Libraries
GovInfQ	Government Information Quarterly
HealthInfJ	Health Library and Information Journal
InfProcMngt	Information Processing & Management
InfRes	Information Research
InfSoc	Information Society
InfTechLib	Information Technology and Libraries
JAcadLib	Journal of Academic Librarianship
JASIST	Journal of the American Society for Information Science and Technology (JASIST)
JDoc	Journal of Documentation
JInformetrics	Journal of Informetrics
JIS	Journal of Information Science
JLibInfSc	Journal of Library and Information Science
JMedLibAssn	Journal of the Medical Library Association
JSchlPub	Journal of Scholarly Publishing
LawLibJ	Law Library Journal
LearnPubl	Learned Publishing
LibCollAcq	Library Collections Acquisitions & Technical Services
LibInfScRes	Library and Information Science Research
LibQ	Library Quarterly
LibTrend	Library Trends
Online	Online
OnlInfRev	Online Information Review
PortLibAcad	Portal: Libraries and the Academy
ProgELib	Program-Electronic Library and Information Systems
ResEval	Research Evaluation
Scientometrics	Scientometrics
SerialsRev	Serials Review
SocScInf	Social Science Information

Appendix 2

Journal citing indices

Journal	*# of Publications*	*# of Citing Articles*	*# of Citers*	*# of Citing Journals*
ARIST	61	422	756	146
AslibProc	184	306	521	133
CollResLib	155	271	463	64
GovInfQ	247	675	1269	279
HealthInfJ	183	457	1474	253
InfProcMngt	394	1551	3251	661
InfRes	282	75	157	43
InfSoc	113	261	523	162
InfTechLib	91	118	240	54
JAcadLib	294	463	857	122
JASIST	937	3340	5536	934
JDoc	213	638	1022	187
JInformetrics	224	1113	1557	292
JIS	250	830	1716	367
JLibInfSc	88	92	148	38
JMedLibAss	236	452	1414	202
JSchlPub	105	59	86	28
LawLibJ	158	86	112	24
LearnPubl	139	188	330	68
LibCollAcq	76	73	138	34
LibInfScRes	148	430	794	132
LibQ	97	155	213	43
LibTrend	228	178	323	77
Online	192	37	74	26
OnlInfRev	260	537	989	234
PortLibAcad	124	174	299	53
ProgELib	132	154	287	53
ResEval	162	333	646	94
Scientometrics	889	2229	3317	605
SerialsRev	106	128	242	62
SocScInf	140	160	279	116

Appendix 3

Impact factor, 5-year impact factor, and normalized citing values

Journal	*Impact Factor*		*5-year Impact Factor*		*# of Citing Articles per Publication*		*# of Citers per Publication*		*# of Citing Journals per Publication*	
	Num.	*Rank*	*Num.*	*Rank*	*Num.*	*Rank*	*Num.*	*Rank*	*Num.*	*Rank*
ARIST	2.00	3	2.35	3	6.92	1	12.39	1	2.39	1
AslibProc	0.60	25	0.72	20	1.66	16	2.83	17	0.72	14
CollResLib	0.68	21	0.90	17	1.75	15	2.99	15	0.41	25
GovInfQ	1.88	5	2.18	4	2.73	8	5.14	9	1.13	7
HealthInfJ	0.76	19	0.94	15	2.50	10	8.05	3	1.38	5
InfProcMngt	1.67	6	1.79	7	3.94	3	8.25	2	1.68	2
InfRes	0.82	18	0.86	18	0.27	30	0.56	30	0.15	29
InfSoc	1.24	10	1.71	8	2.31	11	4.63	11	1.43	4
InfTechLib	0.53	29	0.64	22	1.30	21	2.64	18	0.59	16
JAcadLib	0.87	15	0.91	16	1.57	18	2.91	16	0.41	24
JASIST	2.14	2	2.11	5	3.56	4	5.91	7	1.00	8
JDoc	1.45	7	1.41	9	3.00	6	4.80	10	088	11
JInformetrics	3.12	1	3.59	1	4.97	2	6.95	4	1.30	6
JIS	1.41	8	1.86	6	3.32	5	6.86	5	1.47	3
JLibInfSc	0.64	24	0.54	26	1.05	25	1.68	26	0.43	22
JMedLibAss	0.84	17	1.28	10	1.92	14	5.99	6	0.86	12
JSchlPub	0.52	31	0.39	31	0.56	28	0.82	28	0.27	28
LawLibJ	0.90	14	0.50	28	0.54	29	0.71	29	0.15	30
LearnPubl	1.04	11	0.67	21	1.35	20	2.37	20	0.49	19
LibCollAcq	0.53	28	0.39	30	0.96	26	1.82	25	0.45	20
LibInfScRes	1.36	9	1.25	11	2.91	7	5.36	8	0.89	10
LibQ	0.65	23	0.74	19	1.60	17	2.20	22	0.44	21
LibTrend	0.67	22	0.59	24	0.78	27	1.42	27	0.34	27
Online	0.52	30	0.45	29	0.19	31	0.39	31	0.14	31
OnlInfRev	0.99	12	0.98	14	2.07	12	3.80	13	0.90	9
PortLibAcad	0.87	16	1.01	13	1.40	19	2.41	19	0.43	23
ProgELib	0.60	26	0.52	27	1.17	23	2.17	23	0.40	26
ResEval	0.94	13	1.07	12	2.06	13	3.99	12	0.58	18
Scientometrics	1.91	4	2.42	2	2.51	9	3.73	14	0.68	15
SerialsRev	0.71	20	0.58	25	1.21	22	2.28	21	0.58	17
SocScInf	0.55	27	0.63	23	1.14	24	1.99	24	0.83	13

FIELD-NORMALIZATION OF IMPACT FACTORS: RESCALING *VERSUS* FRACTIONALLY COUNTED

Loet Leydesdorff,[1] Filippo Radicchi,[2] Lutz Bornmann,[3] Claudio Castellano,[4] and Wouter de Nooy[5]

[1] *loet@leydesdorff.net*
Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam (The Netherlands)

[2] *f.radicchi@gmail.com*
Universitat Rovira i Virigili, Av. Països Catalans 26, 43007 Tarragona (Spain)

[3] *Lutz.Bornmann@gv.mpg.de*
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, D-80539 Munich (Germany)

[4] *claudio.castellano@roma1.infn.it*
Istituto dei Sistemi Complessi (ISI-CNR). Via dei Taurini 19, I-00185 Rome, Italy, and Dipartimento di Fisica, Sapienza Università di Roma, P. le A. Moro 2, 00185 Rome (Italy)

[5] *W.deNooy@uva.nl*
Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam (The Netherlands)

Abstract
Two methods for comparing impact factors and citation rates across fields of science are tested against each other using citations to the 3,705 journals in the *Science Citation Index* 2010 (CD-Rom version of *SCI*) and the 13 field categories used for the *Science and Engineering Indicators* of the US National Science Board. We compare (*i*) normalization by counting citations in proportion to the length of the reference list (1/*N* of references) with (*ii*) rescaling by dividing citation scores by the arithmetic mean of the citation rate of the cluster. Rescaling is analytical and therefore independent of the quality of the attribution to the sets, whereas fractional counting provides an empirical strategy for normalization among sets (by evaluating the between-group variance). By the fairness test of Radicchi & Castellano (2012a), rescaling outperforms fractional counting of citations for reasons that we consider.

Conference Topic
Scientometric Indicators – Criticism and New Developments (Topic 1a)

Introduction

The use of journal impact factors (IFs) for evaluative comparisons across fields of science cannot be justified because fields of science differ in terms of citation practices. In mathematics, for example, reference lists are often short (< 10), while in the bio-sciences reference lists with more than 40 cited references are common. Thus, the *citation potentials* across fields of science are different for purely statistical reasons (Garfield, 1979). Apart from this scale effect, citation distributions have specific characteristics (Albarrán *et al*., 2011; Glänzel & Schubert, 1988) and thus one may hope to find ways to make them comparable, but only after appropriate normalization. This question of normalization is urgent for the evaluation process because institutional units are rarely monodisciplinary, and thus at the level of institutional units, one can hardly avoid the conundrum of comparing apples with oranges (Rafols *et al*., 2012).

Small & Sweeney (1985) first proposed using "fractional citation counting," that is, the attribution of citation credit to the cited paper in proportion to the length of the reference list in the citing paper. Zitt & Small (2008) used the audiences of the citing papers as the reference sets for developing Audience Factors of journals—as an alternative to Impact Factors—and Moed (2010) proposed to combine these two ideas when developing SNIP ("Source-Normalized Impact per Paper") as a journal indicator for the Scopus database. Leydesdorff & Opthof (2010) radicalized the idea of fractional counting at the paper level and proposed abandoning normalization in terms of relevant fields that are defined in terms of journal sets, and to use the citing *papers* as the reference sets across fields and journals, and then to attribute citations fractionally from this perspective (cf. Waltman & Van Eck, forthcoming). Using papers as units of analysis allows for fractional counting of the citations across institutional units with different portfolios (Leydesdorff & Shin, 2011; Zhou & Leydesdorff, 2011). Furthermore, the change to the level of papers for the evaluation allows for statistical decomposition in terms of percentile ranks and hence the use of nonparametric statistics (Bornmann & Mutz, 2011; Leydesdorff *et al*., 2011)

Leydesdorff & Bornmann (2011) decomposed the journal set of the *Science Citation Index* 2008 at the paper level and reconstructed a fractionally counted impact factor. Using numerators from the 3,853 journals included in the CD-Rom version of this database and denominators from the *Journal Citations Report* 2008, these authors found an 81.3% reduction of the between-group variance across 13 major fields distinguished by PatentBoards™ and the US National Science Foundation (NSF) for the biannual evaluation in *Science and Engineering Indicators* (NSB, 2012). The remaining between-group variance was no longer statistically significant. Leydesdorff, Zhou & Bornmann (2013) repeated this analysis using 2010 data, but with more strict criteria, improved statistical methods, and time horizons other than the two-year citation window of the standard impact factor (IF2). As before, the reduction of the between-group

variance was 79.2% (as against 81.3% in 2008), but the IF5 further improved on this reduction to 91.7%. The latter result was statistically significant, whereas the former in this case was not.

In the final paragraphs, Leydesdorff, Zhou, & Bornmann (2013) raised the question of how their results would compare to the universal normalization procedure for citation distributions proposed by Radicchi, Fortunato, & Castellano (2008). In this study, we compare the two normalization schemes using the fairness test proposed by Radicchi & Castellano (2012a). This is a statistical test specifically designed for measuring the effectiveness of normalized indicators aimed at the removal of disproportions among fields of science. Radicchi & Castellano (2012a) used this test to show that the rescaled indicator introduced by Radicchi *et al*. (2008) outperformed the fractional indicator proposed by Leydesdorff & Opthof (2010) in the analysis of the citations received by papers published in the journals of the American Physics Society (APS).

Radicchi *et al.*'s (2008) normalization can be applied to any comparison among subsets. The attribution of the cases to the subsets can even be random. The normalized (field-specific) citation count is $c_f = c / c_0$, in which c is the raw citation count and c_0 is the average number of citations per unit (article, journal, etc.) for this field—or more generally—this subset. The normalization sets the mean of the scores in each group equal to 1. Consequently, the between-group variance of the rescaled scores is necessarily zero independently of the attribution of the units to the groups.

Whereas the reasoning of Radicchi and his coauthors (2008, 2012a, 2012b) is analytical and focuses on the homogeneity of the set after normalization, Leydesdorff & Bornmann (2011) studied whether the statistical significance of the dividedness between the groups is reduced by fractionalization as an empirical strategy using the so-called variance components model: in addition to papers being organized in journals (at level 1), journals are at a next level 2 intellectually organized in fields of science. This second-level effect can be measured independently of the first-level effect using multi-level analysis. If the between-group variance is statistically significantly different from zero, the sets' citation impact can be considered as heterogeneous. In other words, the multi-level model (of generalized linear mixed models) enables us to quantify and statistically test the effects of fractional counting in the comparison among sets, whereas rescaling sets the between-group variance by definition equal to zero.

In this study, we use the same data as in Leydesdorff *et al*. (2013), and compare the fractionally normalized values with the results of normalization based on dividing by the arithmetic mean of the parameter under study (e.g., the IF5) at the level of each cluster, using Radicchi *et al*.'s (2008) rescaling. We rescaled the integer-counted impact factors and their numerators (total citations), and

additionally the numerators of IF2 and IF5 as provided by the Journal Citation Reports 2010 of the Science Citation Index-Expanded, but for this same set of 3,705 journals. The project was done in June-August 2012, and at that time the data for 2010 were the most recent data available.

Materials and Methods

Data

Data was harvested from the CD-Rom version of the *Science Citation Index* 2010. This version contains a core set of 3,705 journals contained in the *Science Citation Index-Expanded,* but selected as most representative and used for policy purposes. The U.S. firm PatentBoard™—previously named CHI Inc.—has for several decades been under contract of the U.S. National Science Foundation to add 13 categories to the journal list that is used for the biannual updates of the *Science and Engineering Indicators* of the National Science Board (2012). We use these 13 categories from the 2010 list as the second level, but two categories are not used in the analysis because they are poorly populated in this subset of 3,705 journals: cluster 8 ("Humanities") contained only two journals, and cluster 11 ("Professional fields") only eight journals. Thus, we work with 3,695 journals organized in eleven broad fields of science. The reader is referred to Leydesdorff, Zhou & Bornmann (2013) for further details about the data processing and the distinction of 23 possible variables (including the two- and five-year impact factors).

Table 1: Variables considered for rescaling. TC=total cites; IC=integer counting; IF=impact factor; JCR=Journal Citation Reports

	Variable	
1.	ISI-IF2	IF2 from JCR 2010
2.	ISI-IF5	IF5 from JCR 2010
3.	IF2-IC	IF2 integer counted from CD-Rom
4.	IF5-IC	IF5 integer counted from CD-Rom
5.	ISI-TC	Times cited, JCR 2010
6.	TC-IC	Times cited, integer counted from CD-Rom
7.	TC-IC2	IF2 numerator from CD-Rom
8.	TC-IC5	IF5 numerator from CD-Rom
9.	IF2-Num	IF2 numerator from JCR 2010
10.	IF5-Num	IF5 numerator from JCR 2010
11.	c/p 2010	c/p ratio: variable 5 / Citable items 2010 (JCR)

Among the 23 variables used by Leydesdorff *et al*. (2013), we use the variables listed in Table 1 for the rescaling procedure in this study. We do not rescale any of the fractionally counted analogues of these integer-counted indicators—IF2-FC, IF5-FC, TC-FC, TC-FC2, and TC-FC5—because the objective of the study is

to compare the effects of fractionalization *versus* rescaling as normalization strategies.

Variables 1 and 2—taken from JCR—are different from the corresponding values of variables 3 and 4 because the ISI-IF includes all citations in the larger set of JCR 2010 in the numerator (N = 10,196 journals), whereas variables 3 and 4 are based on counting only in the domain of the 3,705 journals included in the CD-Rom version. (The denominators are the same, that is, the sum of citable items in the previous two years as provided by JCR.) The various numerators are separately studied as variables 5 to 10. Finally, variable 11 adds a value derived from JCR: the total cites of each journal (variable 5) divided by the number of this journal's citable items (articles, reviews, and proceedings papers) in the current year 2010.

Methods

Radicchi & Castellano (2012a) provide a fairness test that can be applied to differently normalized datasets in order to compare whether fractional counting of the citations or rescaling of the citation counts leads to a better result. Note that this is not a trivial question despite the analytical character of rescaling. Different normalizations may have different effects on the distributions of the variables in the various subsets so that variable proportions may belong, for example, to the top-10% of most-highly-cited journals. According to the notion of a fair indicator, the probability of finding a journal with a particular value of this indicator should not depend on the subset of research (e.g., discipline) to which this journal is attributed. The "fairness" of a citation indicator is therefore directly quantifiable by looking at the ability of the indicator to suppress any potential citation bias related to the classification of journals in disciplines or topics of research.

The fairness test was previously used for the assessment of indicators devoted to the suppression of disproportions in citation counts among papers belonging to different sets, but it can straightforwardly be extended in the present case to test the performances of indicators aiming at the suppression of discipline-dependent bias in journal evaluation. In this study, we analyze a set of N = 3,695 journals divided into G=11 different categories. We indicate with N_g the number of journals within category g. Each journal in the entire set has associated a score calculated according to the rules of the particular indicator we want to test (Table 1). Imagine sorting all journals depending on this indicator and then extracting the top z% of journals from the sorted list. The list of top z% journal is composed of $n^{(z)} = \lfloor {zN}/{100} \rfloor$ journals (where $\lfloor x \rfloor$ indicates the integer part of x).

If the indicator is fair, the presence in the top z% should not depend on the particular category to which the journal belongs. That is, the presence of a journal of category g in the top z% should depend only on N_g and not on the fact that category g is privileged for some other reason. Under these conditions, the

number of journals $m_g^{(z)}$ of category g that are part of the top z% is a random variable that obeys the hypergeometric distribution:

$$P\left(m_g^{(z)}\middle|n^{(z)},N,N_g\right)=\binom{N_g}{m_g^{(z)}}\binom{N-N_g}{n^z-m_g^{(z)}}/\binom{N}{n^z} \quad (1)$$

where $\binom{x}{y}=x!/[y!\,(x-y)!]$ is the binomial coefficient (Radicchi & Castellano, 2012a, at p. 126). By using this distribution, it is therefore possible to estimate the confidence intervals of an ideal fair indicator, and thus one can statistically judge the "fairness" of any other indicator.

Results

a. Rescaling versus fractional counting of the impact factors

We examined rescaled versions of all the indicators listed in Table 1. Figure 1 shows graphically the outcomes of analyses using the fairness test for the comparison of rescaled values of IF2 and IF5 *versus* their fractionally counted equivalents. In the left column of the first row, the deviations from the 10% expectation are shown for the rescaling of IF2-s and in the right column for fractionally counted IFs-2. The second row repeats the analysis for the case of five-year IFs. Vertically, the graphs are somewhat similar, but horizontally the differences are considerable.

Table 2: Percentages of journals belonging to the top-10% set under the different conditions.

Cluster	IF2 Rescaled	IF5 Rescaled	IF2 Fractionally counted	IF5 Fractionally counted
1. Biology	9.46	9.25	5.57	5.78
2. Biomedical Research	11.35	11.15	17.70	16.73
3. Chemistry	10.78	11.11	11.75	12.70
4. Clinical Medicine	9.68	9.77	12.33	11.55
5. Earth & Space	6.27	5.54	7.01	7.01
6. Engineering & Tech	6.53	7.88	3.15	4.27
7. Health Sciences	12.50	12.50	9.38	9.38
9. Mathematics	17.92	15.03	4.05	5.20
10. Physics	11.93	12.76	10.61	11.43
12. Psychology	19.05	16.67	9.52	11.90
13. Social Sciences	9.68	9.68	0.00	0.00
Mean (± st.dev.)	11.38 (± 4.03)	11.03 (± 3.16)	8.28 (± 4.97)	8.72 (± 4.75)
$\Sigma_i \lVert x_i - 10 \rVert$	31.91	27.11	43.72	42.68

Rescaling outperforms fractional counting: both the summed and average deviances from the 10% score, as well as the standard deviations, are smaller in the case of rescaling (Table 2). Furthermore, the rescaled values passed the test of the 90% confidence interval (assuming a hypergeometric distribution) while the fractionally counted values did not. Thus, the differences in the distributions among scientific fields are effectively removed when one uses the rescaled versions of these indicators.

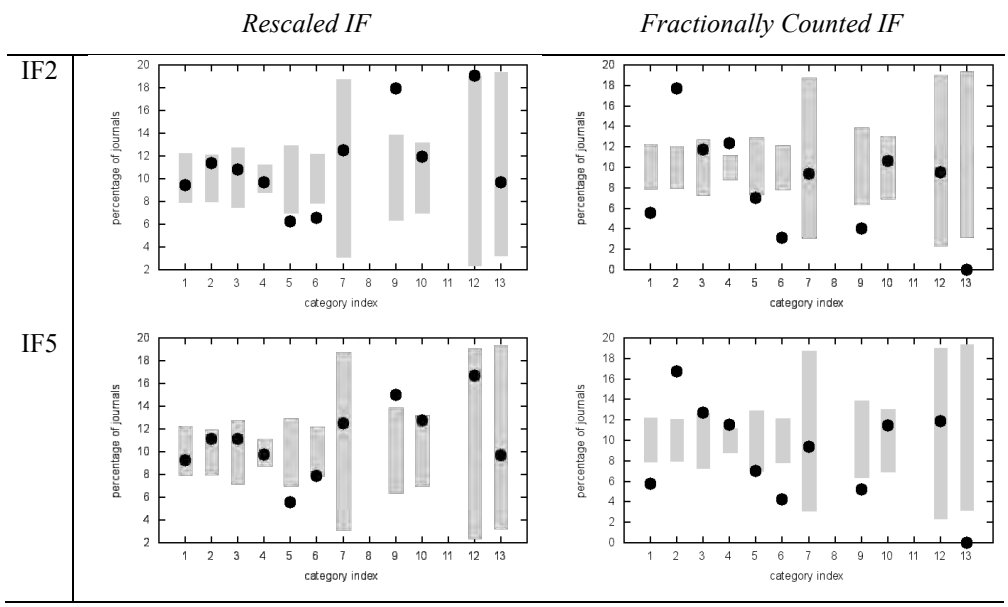


Figure 1: Percentages of journals belonging to the top-10% of 3,695 IFs 2010 in eleven different groups, normalized by rescaling and fractional counting of the citations, respectively. Grey areas bound the 90% confidence intervals and are calculated using Eq. (1)

At the level of the individual clusters, fractional counting completely fails the fairness test (with a success rate of 0%) in the case of cluster 13, that is, the "Social Sciences." These are 31 journals attributed to the social sciences, within the domain of the Science Citation Index (and not the Social Science Citation Index). The highest-ranked journal in this deviant set is the *Journal of Human Evolution* which ranks at the 673rd position with ISI-IF2 = 3.843 or 579th position with ISI-IF5 = 4.290. The corresponding ranks are 713th and 556th in the more restricted SCI set under study. Fractionally counted, however, these rankings are worsened to the 726th and 600th positions, respectively. All these values are far outside the domain of the top-10% group of 370 journals (N = 3695/10 = 369.5). In the social sciences, referencing is relatively high and citation low in comparison with the natural and life sciences so that fractional counting cannot be expected to improve on the relative standing of these journals in the rankings. By

using the arithmetic means of the group as the reference points—the mean values are 0.576 (IF2) and 0.721 (IF5), respectively—rescaling of this set of 31 journals provides *Journal of Human Evolution* with the 127th and 116th positions, respectively, within the set of top-10% highest-ranked journals. However, the number of observations is small in this case.

For another example, let us turn to cluster 9, which is composed of 173 journals designated "Mathematics". Mathematics is the well-known exception in terms of exceptionally low referencing behavior. Might this explain the low value of 5.20% of these journals among the top-10% when using a fractionally counted IF5? The highest IF5 in this group is attributed to *Siam Review* with a value of 3.373. This value ranks the journal at the 428th position and therefore outside the domain of the top-10% of 369 most-highly-ranked journals.[63] Fractionally counted, however, the IF5 of *Siam Review* is upgraded to the 179th position. Three other journals in this group (*Annals of Mathematics* – 115th position; *J American Mathematics Society* – 133rd position; *Acta Mathematica [Djursholm]* – 135th position) are ranked higher than *Siam Review* after fractional counting, among nine journals in total belonging to the top-10% group. Thus, fractional counting in this case corrects for between-field differences. Rescaling brings the fairness test to a value of 15.03%, that is, rather far at the opposite side of the reference standard of ten percent. In the case of "Physics" (cluster 10 with 245 journals), the correction of fractional counting even outperforms rescaling; but this is the exception rather than the rule.

Further statistical analysis taught us that the arithmetic means of the fractionally counted citations per cluster correlate significantly with the results of the corresponding parameters on the fairness test ($r = .91$, $p < .01$ for IF2; $r = .92$, $p < .01$ for IF5). This indicates that the fractionally counted IFs still reflect between-field differences. Furthermore, the normalization in terms of fractional counting has uncontrolled effects on the shape of the distributions in terms of standard deviations, skeweness, and kurtosis when comparing across the clusters, whereas rescaling (as a linear transformation) behaves neutrally in this respect.

Rescaling ISI-IFs

Whereas for the construction of fractionally counted IFs, Leydesdorff, Zhou & Bornmann (2013) needed individual journal-journal citations and where therefore limited to the set of 3,695 journals contained in the CD-Rom/DVD version of the *Science Citation Index* 2010, rescaling can be applied to any set. For Table 3, we use the same 3,695 journals for comparing ISI-IFs (both for two and five years) with the same values divided by the arithmetic means of each of these 11 subsets.

[63] The ISI-IF5 of this journal is 5.73; this leads to the 325th position in the ranking, i.e., within the top-10%. (See the discussion about Table 3 below).

Note that in Table 3 the rescaled values of ISI-IF2 outperform the normalization when compared with the rescaled values of ISI-IF5.

Table 3: Percentages of journals belonging to the top-10% set when comparing the ISI-IFs of 3,695 journals with their rescaled equivalents

Cluster	ISI-IF2	ISI-IF5	ISI-IF2 Rescaled	ISI-IF5 Rescaled
1. Biology	4.50	6.64	9.46	9.03
2. Biomedical Research	20.62	19.46	12.33	12.52
3. Chemistry	10.79	10.79	11.44	11.76
4. Clinical Medicine	13.53	12.41	9.33	8.90
5. Earth & Space	4.80	6.64	8.49	8.86
6. Engineering & Tech	2.47	3.37	9.23	9.68
7. Health Sciences	9.38	12.50	12.50	12.50
9. Mathematics	0.58	0.58	10.98	12.72
10. Physics	6.94	6.53	8.64	8.23
12. Psychology	16.67	19.05	11.90	11.90
13. Social Sciences	0.00	0.00	16.13	16.13
Mean (± st.dev.)	8.21 (± 6.69)	8.91 (± 6.62)	10.95 (± 2.27)	11.11 (± 2.39)
$\Sigma_i \lVert x_i - 10 \rVert$	62.96	60.45	20.12	22.83

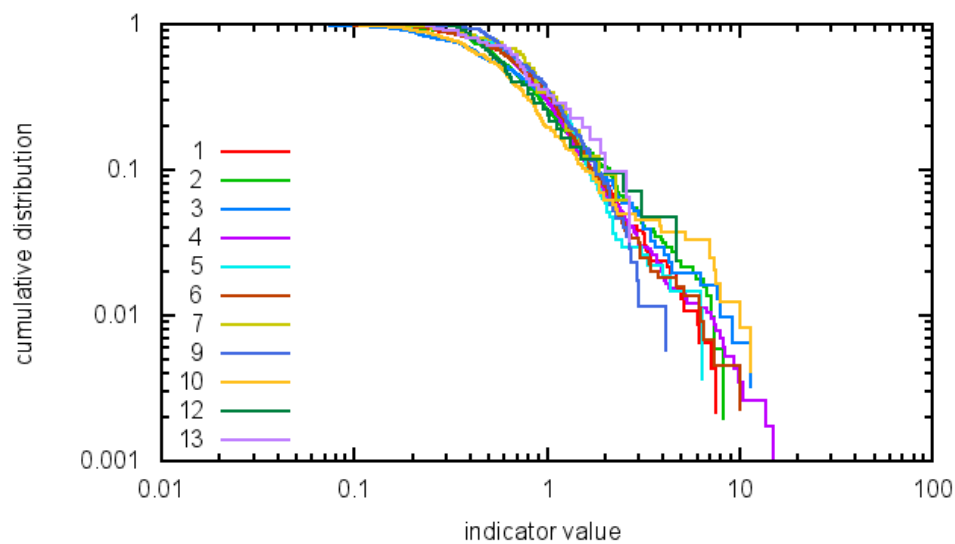


Figure 2: Cumulative distribution of rescaled ISI-IFs for eleven groups among 3,695 journals for IF2. (A perfectly fair indicator would have produced a precise collapse of the various curves.)

As in the previous comparison, Cluster 13 ("Social Sciences") is not included in the top-10% set when using either ISI-IF2 or ISI-IF5, and only the journal *Siam Review* is within this domain among the 173 mathematics journals (0.58%). Using

rescaling, however, the percentages in Table 3 can meaningfully be compared with the reference standard of ten percent.

Figure 2 shows the cumulative distributions of the rescaled ISI-IF2s. The distributions have a similar shape for IF2 and IF5. Differences are small but the curves do not coincide perfectly. Hence the universality that has been claimed for the distributions of article citations within fields (Radicchi *et al*., 2008) is valid only approximately when journal impact factors are considered as citation scores (cf. Waltman, van Eck, & van Raan, 2012).

Normalized Impact Factors

Since the tests indicate that the rescaled impact factors can be compared across fields of science, one can proceed with the comparison. Table 4 lists the top-10 thus normalized ISI-IFs 2010 sorted on the rescaled values of ISI-IF2.

Table 4: 25 journals (abbreviated titles) ordered in terms of the ISI-IF2 after rescaling.

Rank	Abbreviated journal title	Rescaled ISI-IF2 2010	Rescaled ISI-IF5 2010
1	CA-CANCER J CLIN	26.211	19.283
2	REV MOD PHYS	19.514	18.155
3	ACTA CRYSTALLOGR A	18.881	8.576
4	NAT MATER	17.979	16.951
5	NEW ENGL J MED	14.872	14.380
6	ANNU REV PLANT BIOL	14.063	12.002
7	ANNU REV IMMUNOL	13.700	12.822
8	CHEM REV	11.479	12.641
9	ANNU REV ASTRON ASTR	11.452	10.508
10	NAT NANOTECHNOL	11.440	11.684

Thus normalized, *Science*, for example, follows only at 36th position with a rescaled IF2 = 7.130 and IF5 = 6.985, whereas it held the 16th position (ISI-IF2 = 31.364) behind *Nature* at the 9th (ISI-IF2 = 36.101) in the JCR 2010. Using fractional counting, *Science* would rank at the 16th (IF2fc = 2.696) and *Nature* at the 13th position (IF2fc = 2.888). However, it should be noted that for the rescaling based on the classification of PatentBoard/NSF, these two journals are not considered as "multidisciplinary science," but as two of 514 journals in the cluster "Biomedical Research," and accordingly rescaled using the arithmetic mean of this subset as denominator. In the case of fractional counting, the attribution to predefined disciplinary groups does not play a role in the normalization because fractional counting is performed at the level of papers and across groups.

Table 5 provides the results of a correlation analysis among the different indicators for the comparable sets of 3,695 journals (11 groups).

Table 5: Spearman's rank-order correlations ρ organized in the upper triangle and Pearson correlations r in the lower triangle. The N of journals varies between 3,675 (for the rescaled values) and 3,695 because of missing values. All correlations are statistically significant at the .01 level.

	ISI-IF2	Rescaled IF2	Fractionally counted IF2	ISI-IF5	Rescaled IF5	Fractionally counted IF5
ISI-IF2		.859	.857	.973	.835	.815
Rescaled IF2	.935		.778	.860	.972	.763
Fraction. IF2	.933	.909		.826	.834	.958
ISI-IF5	.977	.919	.922		.883	.824
Rescaled IF5	.913	.976	.941	.941		.778
Fraction. IF5	.906	.896	.973	.932	.896	

Table 5 shows that the rescaled IF2 and IF5 correlate across the file precisely as high with each other (Spearman's $\rho = 0.97$ and Pearson's $r = 0.98$) as the unscaled ISI-IF2 and ISI-IF5.[64] The ISI-IFs correlate slightly less with the corresponding normalized IFs, but the rank-order correlations between rescaled and fractionally counted IFs-2 and IFs-5 are only 0.76 and 0.78, respectively. For details about the correlations between fractionally and integer-counted impact factors, the reader is further referred to Leydesdorff, Zhou, and Bornmann (2013: Table 3).

The full sets of both rescaled and fractionally counted impact factors 2010 are available online at http://www.leydesdorff.net/if2010/index.htm and at http://www.leydesdorff.net/if2010/normalized_ifs_2010.xlsx , respectively. Further exploration taught us that correlations are higher in the top and bottom deciles than in the middle range where different normalizations may have large effects on the ranking (Leydesdorff *et al*., in press).

Conclusions and Discussion

In this study, our two teams joined forces to address the question raised by Radicchi & Castellano (2012a) about comparing the fractional counting of citations with rescaling by dividing by the arithmetic mean of each subset, using the complete set of journals studied by Leydesdorff, Zhou & Bornmann (2013) to generate quasi-IFs. The original idea was to apply the multi-level method used in the latter study also to the set of rescaled values so that the variance components

[64] When the sets of journals are equal, one would expect the Pearson correlations between IF2 and IF5 to be the same for the original and rescaled IFs because rescaling extracts the between groups variation both from the numerator (covariance between the two variables) and from the denominator (product of the standard deviations of the two variables). If so, the equality among the correlations is analytical. In our case, however, the numbers of journals are different because they were taken in the one case from the Web of Science and in the other from the CD-Rom version.

could be specified and made comparable. However, rescaling annihilates between-group variance because all the arithmetic means of the groups are set at unity. The two sets of values could therefore not be compared using this method.

Radicchi & Castellano (2012a) proposed a "fairness test" that was applied to APS publications and showed that rescaling outperformed fractional counting in this case. Our results confirm this conclusion. The fairness test was even more convincing when applied to the ISI-IFs provided by the *JCR 2010* of the Web of Science (based on 10,196 journals) then to the integer-counted citations to 3,695 journals which provided the basis for our study of fractionally counted citations. However, the correlation in the top-10% among non-normalized and (differently) normalized values of IFs is high (Figure 4).

Rescaling makes it possible to compare across differently grouped sets because the resulting distributions are, at least approximately, "universal" (Figure 3). The distributions are highly comparable (at least within this set of journals; cf. Waltman *et al*. [2012]). The law of cumulative advantages as specified by Price (1976) or other mechanisms dictating the shape of citation distributions thus seem to operate field-independently; that is, the log-log distribution remains after correction for the differences among fields by using rescaling. At the top- and bottom-ends of the distributions, however, considerable deviance from this "universal" regularity is also visible (Leydesdorff & Bensman, 2006).

The different objective of the multi-level approach remains that one can specify the reduction of between-group variance and test the remaining between-group variance on its deviance from zero. In other words, rescaling is insensitive to the quality of the clustering, whereas the variance decomposition based on fractional counting can also be quantified among alternative groupings. Fractional counting can further be improved (and tested!) using methods recently specified by Waltman & Van Eck (forthcoming).

In this study, the different forms of normalizations were applied to journal impact factors (Garfield, 1972). Criticism of this measure for the evaluation of journals (e.g., Seglen, 1997) and *a fortiori* for the evaluation of papers within journals should in this context be mentioned (Braun, 2012; Lonzano *et al*., 2012; Leydesdorff, 2012; Vanclay, 2012). More recently, however, book citations (Kousha *et al*., 2011; Leydesdorff & Felt, 2012) have been added to the potential candidates for impact evaluation. The reasoning here above is not confined to journal evaluation.

When one compares across heterogeneous sets—for example, in the case of evaluating composed sets such as universities with departments and/or when it is difficult to distinguish crisp sets—one can be advised to use rescaling because the quality of the attribution of cases to clusters cannot invalidate this method. Note

that one can rescale any variable that differs systematically across sets (e.g., publication rates). One pragmatic advantage in the case of citations, however, is that citation analysis of the citing papers is not needed before rescaling, while the full audience set is required for computation in the case of fractional counting.

Acknowledgment

We thank Thomson-Reuters for access to the data. We thank Kim Hamilton for providing the journal list of Patent-Board, and NSF for giving permission. A full version of this paper (Leydesdorff *et al*., in press) was in the meantime accepted for publication in the *Journal of the American Society for Information Science and Technology*.

References

Albarrán, P., & Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology, 62*(1), 40-49.

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics, 5*(1), 228-230.

Braun, T. (2012). Editorial to a Special Issue on Journal Impact Factors. *Scientometrics, 92*(2), 207-208.

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation. *Science 178*(Number 4060), 471-479.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359-375.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science, 14*(2), 123-127.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*.

Leydesdorff, L. (2012). Alternatives to the journal impact factor: I3 and the top-10%(or top-25%?) of the most-highly cited papers. *Scientometrics, 92*(2), 355-365.

Leydesdorff, L., & Bensman, S. J. (2006). Classification and Powerlaws: The logarithmic transformation. *Journal of the American Society for Information Science and Technology, 57*(11), 1470-1486.

Leydesdorff, L., & Bornmann, L. (2011). How fractional counting affects the Impact Factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology, 62*(2), 217-229.

Leydesdorff, L., & Bornmann, L. (2012). Testing Differences Statistically with the Leiden Ranking. *Scientometrics, 92*(3), 781-783.

Leydesdorff, L., & Felt, U. (2012). Edited Volumes, Monographs, and Book Chapters in the Book Citation Index (BKCI) and Science Citation Index (SCI, SoSCI, A&HCI). *Journal of Scientometric Research, 1*(1), 28-34.
Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: fractional counting of citations. *Journal of Informetrics, 4*(4), 644-646.
Leydesdorff, L., & Shin, J. C. (2011). How to evaluate universities in terms of their relative citation impacts: Fractional counting of citations and the normalization of differences among disciplines. *Journal of the American Society for Information Science and Technology, 62*(6), 1146-1155.
Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables in citation analysis one more time: Principles for comparing sets of documents *Journal of the American Society for Information Science and Technology, 62*(7), 1370-1381.
Leydesdorff, L., Zhou, P., & Bornmann, L. (2013). How Can Impact Factors be Normalized Across Fields of Science? An Assessment in terms of Percentile Ranks and Fractional Counts. *Journal of the American Society for Information Science and Technology,* 64(1), 96-107.
Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., & de Nooy, W. (in press). Field-normalized Impact Factors: A Comparison of Rescaling versus Fractionally Counted IFs. *Journal of the American Society for Information Science and Technology*.
Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*(3), 265-277.
National Science Board. (2012). *Science and Engineering Indicators*. Washington DC: National Science Foundation; available at http://www.nsf.gov/statistics/seind12/.
Price, D. J. de Solla (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292-306.
Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.
Radicchi, F., & Castellano, C. (2012a). Testing the fairness of citation indicators for comparison across scientific domains: the case of fractional citation counts. *Journal of Informetrics, 6*(1), 121-130.
Radicchi, F., & Castellano, C. (2012b). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE, 7*(3), e33833.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences, 105*(45), 17268-17272.
Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy, 41*(7), 1262-1282.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal, 314*, 498-502.
Small, H., & Sweeney, E. (1985). Clustering the Science Citation Index Using Co-Citations I. A Comparison of Methods. *Scientometrics 7*(3-6), 391-409.
Vanclay, J. K. (2012). Impact Factor: Outdated artefact or stepping-stone to journal certification? *Scientometrics, 92*(2), 211-238.
Waltman, L., & Van Eck, N. J. (forthcoming). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison; available at http://arxiv.org/abs/1208.6122.
Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E., Tijssen, R. J. W., van Eck, N. J., . . . Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419-2432.
Waltman, L., van Eck, N. J., & van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology, 63*(1), 72-77.
Zhou, P., & Leydesdorff, L. (2011). Fractional counting of citations in research evaluation: A cross- and interdisciplinary assessment of the Tsinghua University in Beijing. *Journal of Informetrics, 5*(3), 360-368.
Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology, 59*(11), 1856-1860.

FUNDING ACKNOWLEDGEMENTS FOR THE GERMAN RESEARCH FOUNDATION (DFG). THE DIRTY DATA OF THE WEB OF SCIENCE DATABASE AND HOW TO CLEAN IT UP.

Daniel Sirtes[1]

[1] *sirtes@forschungsinfo.de*
iFQ Institute for Research Information and Quality Assurance, Schützenstrasse 6a, D-10117 Berlin (Germany)

Abstract
Since August 2008 the Web of Science database includes funding acknowledgements information. To date no study has been conducted concerning the data quality of these entries. In this paper, we show the vast array of problems emerging if one wishes to unify all funding organization entries of a large and diverse funding body such as the German Research Foundation (DFG). After enumerating all possible sources of error found by manual sifting through all funding acknowledgement entries of German publications, we introduce a new semi-automated method, in order to facilitate the same cleaning task for future years. The method which uses regular expressions and Levenshtein distance algorithms as building blocks shows a rather good result with precision and recall of 96% and 94%, respectively. With the cleaned data set, two examples are shown of the new possibilities emerging of this kind of bibliometric data. Connecting this information with financial funding data opens up the path to new kind of input-output analysis in the realm of scientific research while corroborating the validity of the funding acknowledgement data.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction

Structure and content of the funding acknowledgement fields in the Web of Science database

Since August 2008 the Web of Science database (WoS) includes funding acknowledgements. Thomson Reuters is extracting this information from the journal articles and fills the fields of funding organization and grant number. Additionally, it includes the raw extracted acknowledgement text in a grant text field. In the relational database developed on the basis of the raw WoS database by the Competence Centre for Bibliometrics for the German Science System

(http://bibliometrie.info/en/home.html) the structure of these fields is as depicted in Figure 1.

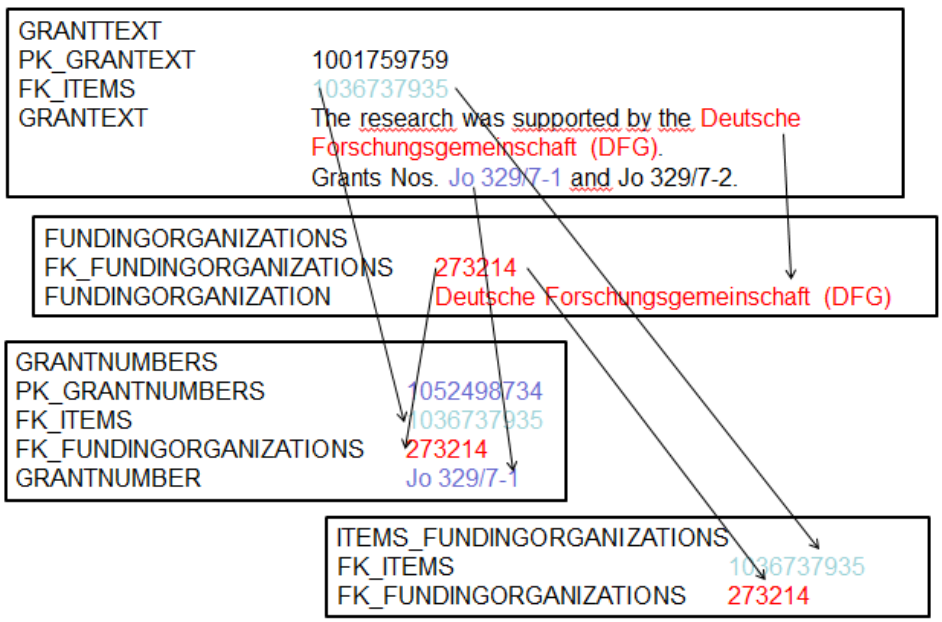


Figure 1. Structure and connections of the funding acknowledgement fields in the database of the Competence Centre for Bibliometrics for the German Science System.

Coverage of funding acknowledgements in the database

As the Competence Centre's database is frozen in week 17 of each year, it is possible to document the dynamics of the inclusion of the funding acknowledgement since its inception. From this information one can see that the amount of items with funding acknowledgements is growing far faster than the growth of the database for the most recent year, suggesting that the extraction methodology of Thomson Reuters is still changing substantially, although the journals' more standardized formatting of the acknowledgement field and more funding acknowledgements in general may also contribute to this growth. Figure 2 shows the count and percentage of journal articles with funding acknowledgements for all three full years of the funding field according to the past two years of the competence centre's database (called WOS2010 and WOS2011, respectively).

The overall coverage depicted above is only an average figure that does not represent the immense diversity in coverage in different disciplines. Table 1

shows that in certain disciplines (assigned by the WoS subject categories (SC)) the share of articles with funding acknowledgements (FA) is very high while in others it is only moderate or even hits zero. The worldwide coverage in these subject categories is juxtaposed with the coverage of articles with German contributions.

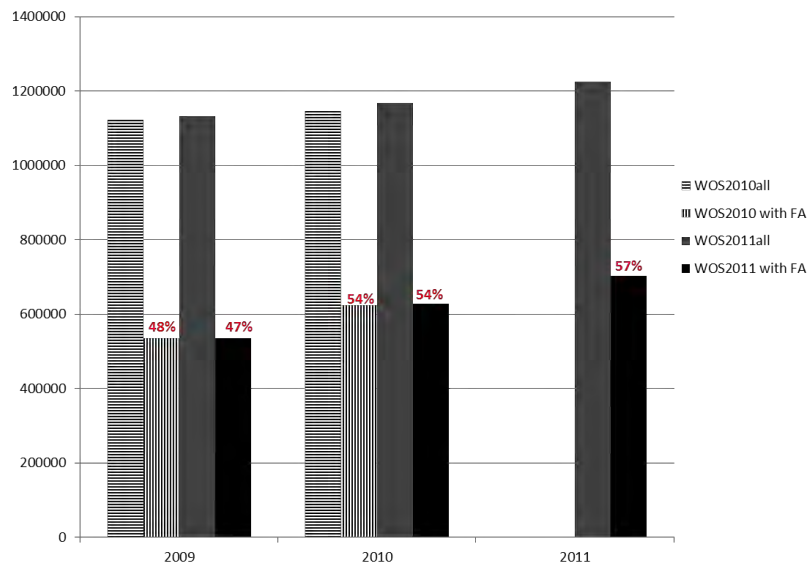


Figure 2. Count of all journal articles and those with funding acknowledgments and their share 2009-2011

Table 1. Coverage of articles in WOS2011 with funding acknowledgements (FA) worldwide and articles with German affiliation (representative selection)

WOS Subj. Cat.	All Articles	with FA	Percent of articles with FA	German articles	German articles with FA	Percent of of German articles with FA
Biology	14065	11524	82%	1082	972	90%
Biochemistry & Molecular Biology	44764	37517	84%	3734	3160	85%
Cell Biology	19558	16298	83%	1962	1669	85%
Ecology	14162	11332	80%	1106	893	81%
Physics, Atomic, Molecular & Chemical	15850	12065	76%	1958	1554	79%
Chemistry, Physical	42967	32165	75%	3459	2678	77%
Materials Science, Multidisciplinary	53242	35790	67%	3753	2542	68%

Physics, Applied	41464	25362	61%	3239	1982	61%
Mathematics	20450	11433	56%	1359	669	49%
Engineering, Chemical	21635	11513	53%	1159	469	40%
Medicine, General & Internal	16481	5777	35%	720	254	35%
Psychology, Experimental	5390	1665	31%	564	224	40%
Economics	14373	1081	8%	1147	67	6%
Humanities, Multidisciplinary	3037	0	0%	54	0	0%
Political Science	4908	0	0%	286	0	0%

This skewed distribution of articles with funding acknowledgements could be contributing to problems of data extraction, but is also consistent with an interpretation that certain disciplines do not have as much external funding as others. This is clearly the case when comparing biological sciences with humanities in general.

Finding Publications funded by the German Research Foundation (DFG)

A simple search and its problems

Finding all the publications funded by the German Research Foundation (DFG) is not a simple task. Thomson Reuters does not unify any of the entries in their funding organization field, which means that every different entry, even if it is only a one letter typo, gets its own identification number as a different funding organization[65]. This problem is multiplied enormously by the following problems.

a. The German Research Foundation has many funding programs (like *Sonderforschungsbereich*, *Emmy-Noether-Programm*, *Exzellenzinitiative*, etc. (for a full list see http://dfg.de/foerderung/index.html)). Very often these funding programs are entered in the grant text and therefore also into the funding organization field and thus is not subsumed under the DFG.
b. Not even the funding program, but rather the funded research facility or network are mentioned (e.g. 'Nanosystems Initiative Munich' or 'Ruhr University Research School').
c. As the name of the German Research Foundation and of its funding programs are originally in German, but many articles translate their name into English (sometimes with their official name, but to a substantial amount also with a creative translation) there are several name variants

[65] The problems of unification for a funding organization has been pointed out in (Rigby 2011) and exemplified for the Swiss National Science Foundation by (Van den Besselaar et al. 2012). However, the complexity of the problem, especially for such a big organization without a standardized system for funding acknowledgements in place, seems to be more daunting than expected (see footnote 3).

even for the same funding program. (Examples of 'creative' translations include 'German Society for the Advancement of Scientific Research' and 'German Academic Research Society').

d. There are a substantial amount of extraction errors which include:
    a. Substitution of the grant number for funding organization (i.e. funding organization 'SFB760')
    b. Co-funded papers appear in the database as a single funding organization (i.e. funding organization 'DFG and NIH')
    c. Severely incorrect extractions of funding organization from the grant text (e.g. from the grant text "…and funding by the GSC 203 for Carolin Schwarz" (which is a graduate school funded by the DFG) the funding organization assigned was 'Carolin Schwarz').

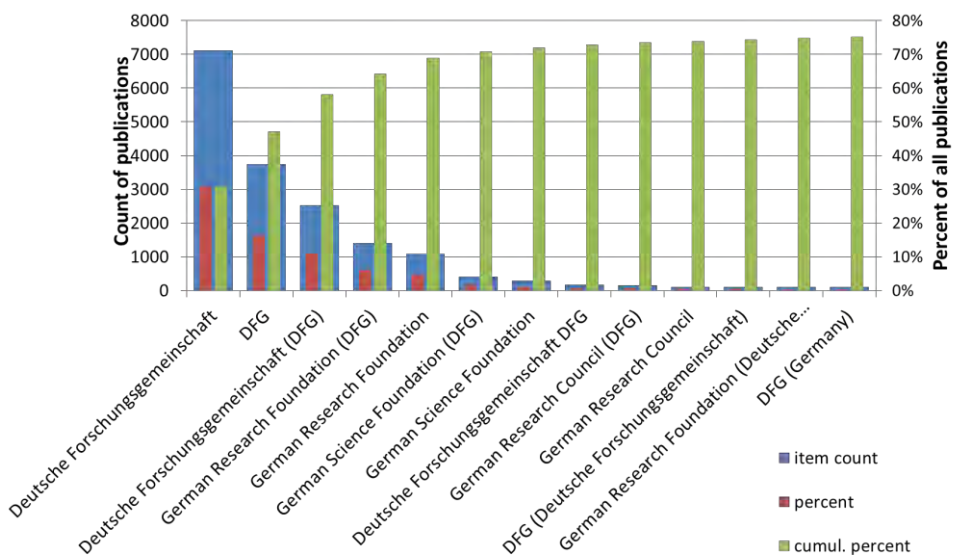


Figure 3. The 13 most common aliases for the German Research Foundation in the 2010 version of the database. Absolute item count, percentage of all publications and cumulative percentage of all publications are shown.

Manual sifting through all German publication

Because of these problems, a first step in finding the DFG funded publications cannot avoid sifting manually through all of German publications for entries in the funding acknowledgement field. (Although some DFG funded publications do not have contributions with a German affiliation, this methodology (restricting the publications to German ones) seems the only feasible one). Many hours of manually comparing the entries in the database with the list of programs funded by the DFG, in harder cases with the help of the grant text and wider internet

searches, has been executed (We would like to thank Simone Falk for her meticulous and excellent work conducting this laborious tasks).
Figure 3 shows the 13 most common entries for DFG funded publications and illustrates the problem with finding all of them. The first six aliases for the DFG cover around 60% of all publications. However, the additional amount of publications per alias flats out very fast and displays a typical power law distribution: Only the first 13 shown here have more than 100 publications per alias. Not more than 87 aliases have at least 10 publications each. Finally, 5747 aliases are associated with only one publication. Thus, the total number of DFG aliases amounts to an astonishing 6370 for the 2010 version of our database.

Development of a semi-automated method for finding aliases in subsequent years

In order to facilitate the search for DFG aliases in the database for subsequent years, a semi-automated method has been developed. With the help of a visual basic script, the results of the manual search has been reproduced. (We would like to thank Mathias Riechert for his help writing the script).
The method has three main components:

a. Regular expressions for the aliases found.
b. Calculation and definition of acceptable levels in Levenshtein distance in order to accommodate orthographical mistakes.
c. A false positive list of aliases that cannot be excluded with regular expressions.

Thus, the first step included finding appropriate regular expressions that are implemented in Oracle SQL in order to capture the aliases found in the manual search. (http://docs.oracle.com/cd/B12037_01/appdev.101/b10795/adfns_re.htm). Examples for these regular expressions can vary in their complexity from ‘for.*gr.*’ for ‘Forschergruppe’ to ‘em.*no\w+.([^ir]\w+)’ for ‘Emmy Noether’.
In a second step, the database entries found with these regular expressions are compared according to a Levenshtein distance algorithm (Levenshtein 1966) in order to calculate the amount of deletions, insertions and substitutions (single-character edits) needed in order to arrive from the found entry to the correct original alias. For example, ‘Forschargruppe’ would have a Levenshtein distance of 1 from ‘Forschergruppe’ as the first e was substituted for an a. In order to achieve uniformity in the algorithm, the Levenshtein distance was calculated as a share of the number of possible substitutions of a string of the same length as the correct entry (The so called ‘Hamming distance’). Thus, the relative Levenshtein distance of the above example is 1/14=0.07, as one out of 14 letters were substituted. The upper bound of acceptable Levenshtein distance was set relatively high with 0.4.
As some of the false positive results of this method were not eliminable with better regular expressions, a list of those entries was compiled in order to subtract

it automatically from the list of the entries found. For example, the California Department of Fish and Game (CDFG or California DFG) will appear in any searches for the DFG. Another example is the Austrian equivalent of the German 'Sonderforschungsbereich' (SFB) (collaborative research center), which uses the same name and abbreviation (e.g. 'Austrian SFB project IR-ON' or 'Austrian Science Fund (FWF) SFB17'). However, we maintained the goal of keeping this false positive list as short as possible which has reached 521 entries. Finally, at some point it did not seem viable to invent new regular expressions for singular entries; therefore 84 aliases were not included into the list for the reproduction of the manual results.
The lists and algorithm was then applied to the 2011 version of the database and yielded the results shown in table 2.

Table 2. Results of the semi-automated method for searching DFG-Aliases

Results		WOS2010	WOS2011
a.	**Levenshtein all**	6807	9550
b.	**Levenshtein *true positive***	6286	8655
c.	**total *false positive***	521	895
d.	**2010 false positive list**	521	521
e.	**false positive not in 2010 list (*new false positive*)**	0	374
f.	**Total true positives with method**	6370	8739
g.	**2010 *false negative* list**	84	84
h.	**Non-Levenshtein (*false negative*)**	84	659
i.	**Non-Levenshtein without 2010 (*new false negative*)**	0	575
j.	**Total DFG aliases**	**6370**	**9314**

Thus, the result of our 2010 method is composed by three lists

a. Levenshtein-list (all results obtained with the regular expression/ Levenshtein script).
c. False positive list (the list obtained by the script resulting in incorrect entries).
g. False negative list (The list of entries not entered into regular expressions).

The resulting list is therefore $a-c+g = f$= 6807-521+84= 6370. As the two false lists could be used for the 2011 application of the method the calculation of precision and recall of the method includes those lists as obtained by the method itself: True positive = $f = b+g$=8739, new false positive = e = 374, and new false negative = i= 575. The precision is therefore 8739/(8739+374) = 96% and the recall is 8739/(8739+575) = 94%. However, as 6370 entries were already set from 2010 one could alternatively calculate the precision and recall of the new entries

in the 2011 database. This yielded the following results: $\text{precision}_{(new)}$ = (8739-6370)/(8739-6370+374) = 86% and $\text{recall}_{(new)}$ = (8739-6370)/(8739-6370+575) = 80%. Considering that the method found 37% more entries in 2011 than in 2010, these results are quite promising.

Portrayal of the cleaned publications set with funding acknowledgements for the German Research Foundation

In order to exemplify the new possibilities of portrayal of the research funded by the DFG and in order to corroborate the validity of the funding acknowledgements data, two preliminary results are presented in the following:

Share of DFG funding by discipline

With the publication set obtained by our method it is now possible to study in which disciplines the German Research Foundation is more or less active. Figure 4 shows a selection of disciplines and the share of German publications with funding acknowledgements and with DFG funding in particular.

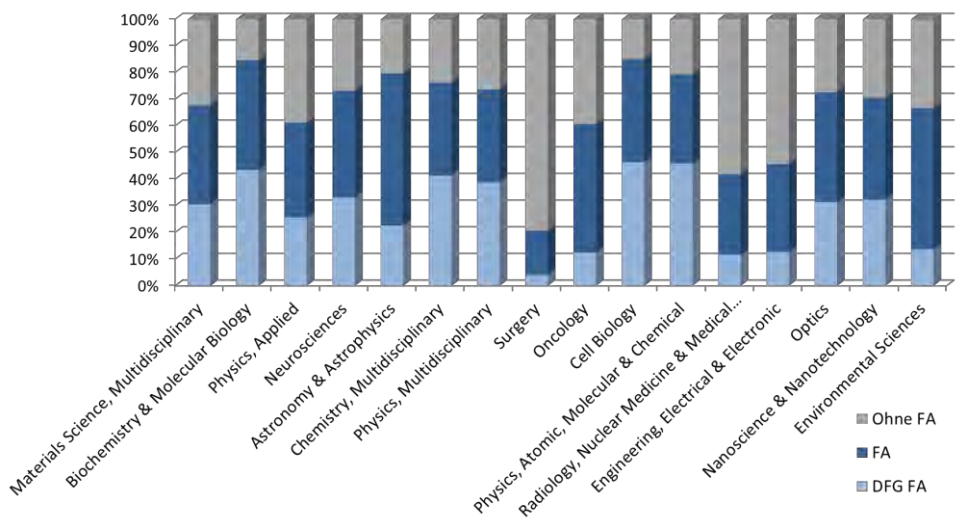


Figure 4. Share of 2010 German publications without, with no DFG, and with DFG, funding acknowledgements, accordingly.

The tendency of the German Research Foundation to fund basic and not applied research which is funded by other means can be directly observed.

Connecting DFG funding acknowledgements with DFG funding amounts

A more elaborate use of the cleaned data set can be obtained by connecting funding acknowledgments with other sources. With the data contained in the DFG issued 'Funding Atlas 2012' (http://www.dfg.de/en/dfg_profile/evaluation_

statistics/funding_atlas/index.html) the financial funding per university and discipline can be inferred. The amount of publications per discipline and German university in 2010 (subsumed in EFI SC super-categories) can then be compared to the funding received from the DFG in the years 2008-2010. Figure 5 shows all publications and all of funding in large German universities[66], while Figure 6 only shows publication and funding in the natural and life sciences. A remarkable correlation can be observed between the two. Although this cannot be considered conclusive evidence as other variables like the size of the universities were not controlled for, it is however noteworthy that in the natural and life science 83% of the variation can be explained by amount of funding received. The lower correlation in the overall picture (R^2=80%) could also be due to different coverage in different disciplines. A hint in this direction is the comparatively low output of Aachen TH, a technical university and the known lower coverage in technology and engineering publications in the WoS database.

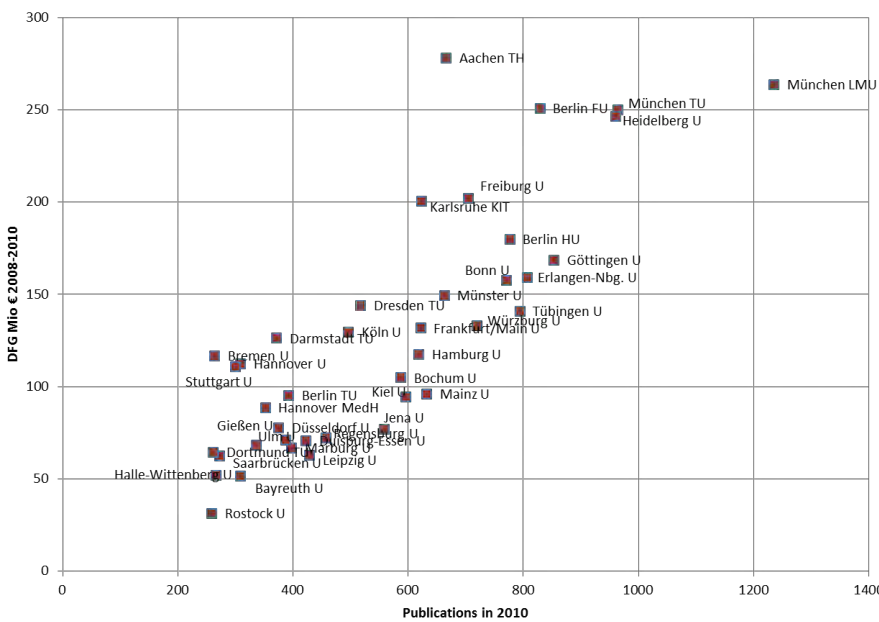


Figure 5. Comparison of all publications funded by the German Research Foundation in 2010 with the amount of funding by the DFG for the same university in the years 2008-2010.

[66] In Figure 5 and 6 only universities with at least 250 and 230 publications in the year 2010 are shown, respectively. However, the coefficient of determination is calculated with all universities that have received DFG funding.

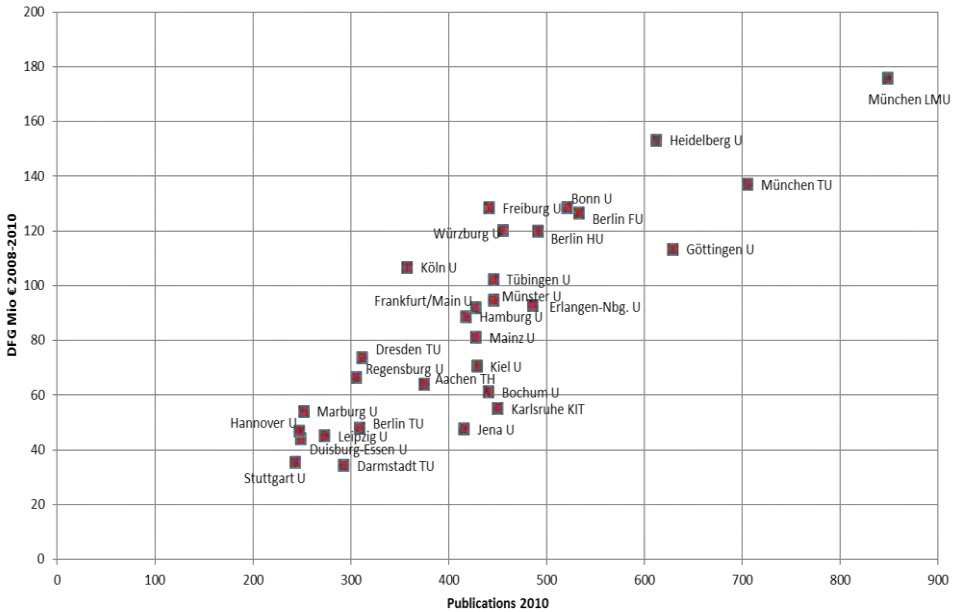


Figure 6. Comparison of all publications in the natural and life science funded by the German Research Foundation in 2010 with the amount of funding by the DFG for the same university in the years 2008-2010.

Discussion

Following the introduction of funding acknowledgements information in the Web of Science database in August 2008, this paper shows the necessary steps needed in order to make this information useful for further study. The growth of publications with funding acknowledgement between the years 2009, 2010 and 2011 shows that 2010 is probably the first year that can be used for further analysis. An analysis of the share of publications with funding acknowledgements in different disciplines shows that in some, like the life sciences the share is that high, that one could assume that most acknowledgements are processed in the database. Although in other disciplines the share is far lower, it is yet unclear whether this is due to less third party funding in these disciplines or due to problems with the extractions of the funding information in certain journals. However, the overall share of 57% for the 2011 shows that this information is usable for a new kind of analysis of the science system. The far more problematic part of this new information is the data quality. In this study we have looked at the German Research Foundation, a particularly large and diverse funding body with many different funding programs. Both on the side of the original funding text in the articles and in their extractions by Thomson Reuters immense problems emerge. Especially, the issue of funding programs being mistaken for funding organizations is particularly pressing and needs of a lot of man-hours in order to

be corrected. Further problems include many variations in translation of the German names of the funding organization and funding programs. In addition to the many orthographic mistakes occurring before and by the data extraction, more severe data extraction errors are apparent. Grant numbers are included in the funding organization field and several funding organization are treated as one combined one on several occasion. In conclusion, a first manual data cleaning step is unavoidable. This array of problems can however be sorted out if enough work is invested. The astonishing result is several thousand entries synonymous with funding given by the DFG[67]. In order to reduce this manual procedure for the subsequent years a new semi-automated method has been employed that uses the regular expression possibilities of the Oracle SQL and a visual basic script implementing a tolerance to typos with a Levenshtein distance algorithm. Using the replicated 2010 results with this method in order to identify new, but similar aliases the 6370 results for the 2010 version of the database could be expanded to include 8739 aliases in the 2011 version. Precision and recall of the method show promising results with 96% and 94%, respectively. In order to exemplify the potential of this cleaned data set two ways to use it in a broader context have been shown. First, with this data the amount of publications in different disciplines funded by the German Research Foundation can be demonstrated. This can be used to assess the disciplines in which the funding body is especially active and in which ones other funding organizations have a higher input. Second, putting the funding acknowledgement data in relation to the funding amounts given by the DFG, as they are included in the DFG Funding Atlas 2012, one can show an input-output relationship in funding. The high correlation between these two data sources shows on one side the validity of the funding acknowledgement information, on the other side opens up possibilities of assessment of funding result not known before. As said, this is only the beginning. The laborious task of data cleaning has now been completed for the German Research Foundation. Once all the major funding organizations are cleaned and unified, a new kind of bibliometric research is possible. Its limits are only set by our own imagination.

Acknowledgments

I would like to thank Simone Falk and Mathias Riechert for their help with the data cleaning and the writing of the visual basic script, respectively.

References

Levenshtein, V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 707–10.

Rigby, John. (2011) Systematic Grant and Funding Body Acknowledgement Data for Publications: New Dimensions and New Controversies for Research Policy and Evaluation. *Research Evaluation* 20, 365 -375.

[67] Grant Lewison, a pioneer in the study of funding acknowledgements was completely incredulous and flabbergasted, confronted with this finding (personal communication at the STI 2012)

Van den Besselaar, P., Inzelt A. & Reale, E. (2012) Measuring Internationalization of Funding Agencies. In: Archambault, Éric / Gingras, Yves / Larivière, Vincent (eds): Proceedings of 17th International Conference on Science and Technology Indicators, Montréal: Science-Metrix and OST, Volume 1, 121-130.

GENDER AND ACADEMIC ROLES IN GRADUATE PROGRAMS: ANALYSES OF BRAZILIAN GOVERNMENT DATA

Jacqueline Leta[1], Gilda Olinto [2], Pablo Diniz Batista[3] & Elinielle Pinto Borges[2]

[1] *jleta@bioqmed.ufrj.br*
Universidade Federal do Rio de Janeiro (UFRJ), Av. Brigadeiro Trompowisky s/ nº, Prédio do CCS, Bloco B – sala 39, CEP 21941-590, Rio de Janeiro (Brazil)

[2] *gilda@ibict.br; elinielle@yahoo.com.br*
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Rua Lauro Muller, 455 - 4° andar, CEP 22290 – 160, Rio de Janeiro (Brazil)

[3] *batista@cbpf.br*
Centro Brasileiro de Pesquisas Físicas (CBPF), Rua Dr. Xavier Sigaud, 150 , CEP 22290-180, Rio de Janeiro (Brazil)

Abstract
Brazilian researchers affiliated to graduate programs play a major role in scientific and technological development in the country. Hence, they have to cope with various academic roles, especially the core one: teaching. Considering Boudieu's concept of 'scientific capital', this study aims at investigating whether female and male teacher-researchers have an equal share of academic activities with low 'scientific capital' return, such as teaching undergraduate courses, and high 'scientific capital' return, such as publication of journal articles. Official data from Brazilian Ministry of Education with information about the population of more than 52 thousand teacher-researchers, and available in 6,741 pdf files, were gathered in a single data matrix. Indicators of academic tasks and productivity were generated. Data analyses for the whole population did not show differences between gender and tasks. Nevertheless, analyses focusing on specific scientific areas indicated that in some fields women are having a larger share of less valued tasks than men: a burden that might help to explain the difference in productivity favoring men in these fields. As gender in science and technology is still an issue of high interest, our study aims at contributing to a better understanding of gender boundaries in this field.

Conference Topic
Scientometrics Indicators: Relevance to Science and Technology, Social Sciences and Humanities (Topic 1)

Introduction
Recently, one of the most prestigious scientific journals, *Science*, published an extensive article about the outstanding growth of Brazilian share in one of the largest scientific international database (Regalado, 2010). However, such strong

growth has been primarily a result of the inclusion of dozens of new Brazilian scientific titles in the database collection, as reported by Leta (2011).

Although this growth is explained by the dynamics utilized for periodical indexing in databases, it is undeniable that Brazilian science has been growing significantly. In fact, during the last two decades, not only scientific outputs (publications) increased notably but also Brazilian scientific inputs, as, for example, human resources for science and academia (Olinto & Leta, 2011). Most of these changes have occurred in public universities and research institutes, which became the main locus of research as well the locus of training and qualifying researchers and university professors.

Accompanying the growth in scientific production, the number of graduate programs increased significantly, spreading all over the country and in several fields of knowledge. Today, there is a strong inter-relationship and mutual contribution between graduate programs and research. Thus, being a researcher in Brazil is practically synonymous of being a teacher of a graduate program from a public research institute or from a public university.

Since the beginning of the 1990's, graduate programs are regularly evaluated by CAPES, the agency of the Ministry of Education responsible for control and evaluation of graduate programs. The number of publications (mainly with international visibility) is one of the criteria especially emphasized in the model adopted by CAPES for the evaluation of graduate programs. In this scenario, Brazilian teacher-researchers have to cope with publication pressure and other academic roles, especially the core one: teaching undergraduate and graduate courses. Hence, do Brazilian teacher-researchers, men and women, combine equally the multiple academic-scientific roles including teaching in different academic levels? In other words, are the different academic tasks evenly distributed between men and women in Brazilian academia?

Considering the above research questions in the context of the concept "scientific capital" developed by Bourdieu (2003), discussed below, and of the extensive literature about gender differences in science and technology, we expect that Brazilian female teacher-researchers show higher burdens of time consuming academic activities with low return in 'scientific capital' while Brazilian male teacher-researchers show higher participation in tasks which at the same time promote and reflect higher levels of 'scientific capital".

Comparison between these roles seems to be a difficult task, especially if we have in mind analyses of large data sets. In Brazil, however, due to governmental procedures conceived to register and evaluate Graduate programs, as specified below, we were able to generate a large database containing information about the performance of academic roles for the population of Brazilian teacher-researchers. The identification of their gender, however, was a difficult challenge that we able to face, as described below in the methodological section. This database, including exactly 52,294 cases, is the source of the analyses presented in this study.

Scientific capital and academic roles

Bourdieu's ideas about the 'Scientific Field' and his concept of 'scientific capital' give theoretical support to the questions formulated above. To this author the understanding of the scientific field and of scientific production should not be focused only on its epistemological aspects or on successive new contribution of individual or groups of scientists. According to Bourdieu, the scientific field is a social space in which a variety of conflicts of interests and a constant struggle for legitimization take place among its members. Likewise, subtler processes, including rules of reciprocity as well as academic rituals, are regularly activated to determine who will get rewards for his or her activities, who will be accepted as an authority and receive scientific credit, and who will remain in a submissive and less prestigious position. The characteristics of this social environment would have profound implications for scientific outcomes.

Describing the scientific field, Bourdieu (2003) introduces the concept of 'scientific capital', a symbolic resource that can be defined as the recognition or prestige attributed to the members of a specific scientific field. Two kinds of scientific capitals should be distinguished, according to Bourdieu: 1) the specific or pure and 2) the institutional. The first one refers to personal prestige resulting from more objective scientific products which can be expressed in publications, citations – a tentative measure of prestige –, rewards, etc. The other kind of scientific capital – the institutional – has a bureaucratic and political dimension. It can contribute and at the same time be the consequence of occupation of prominent positions – as, for example, being Head of Department or member of academic committees. This institutional dimension of the scientific capital also refers to personal abilities related to contacts and other types of influences that can be strategic to scientists and institutions, as it is the case of access to research funding. The distinction of these two types of scientific capitals suggests that merit and recognition in science imply several aspects that cannot be solely attributed to fulfillment of academic tasks. However, these two kinds of capitals are not clearly distinguished, a fault that can be partially attributed to Bourdieu and partially to the mutual contribution that occur between the institutional and the pure or specific capital.

Although the subordination of women in society is one of the highlights of Bourdieu's work, he has not dedicated himself to the study of gender relations in science. However, his ideas about the formation of gender dispositions and habits, the specificity of feminine occupations and the social undervaluation of women in the labor market can be brought to the study of women participation in science. The strong gendered differences in scientific disciplines seem to be related to deeply and socially acquired dispositions and habits. The difficulties faced by women in their scientific careers, the 'glass ceilings' they encounter in their work environment, making it difficult for them to have access to the more valued academic tasks, as well as to attain prestigious and better paid positions, indicate that the 'institutional scientific capital' is probably at work in the science field favoring male teacher-researchers.

Gender in science is a well documented issue, and empirical evidences of gender differences in the scientific and technological (S&T) environment have been accumulated through works by social scientists, as well as by international organizations aiming at promotion of equality of opportunities, as OECD and UNESCO since the 1990s (for example, Lamont *et al.*, 2004; Fox, 2001: Long, 1990, 1992; OECD, 2006; UNESCO, 2007;). The production of gender boundaries at work, determining gender differences in opportunities and productivity, has been analyzed by these authors, who consider that disadvantages for women in the workplace and their consequences for career development would be the result of several discriminatory processes. Focusing on differences in productivity by gender as a consequence of different opportunities of collaboration, Long (1990, p 1297) points out to the several small differences that occur between genders since the start of the scientific career: "In addition to differences in the process of collaboration, many small differences that disadvantage women and advantage men are found in the levels of resources affecting productivity at the start of the career".

In the Scientometrics field, studies about gender differences in S&T also accumulate, focusing mostly on productivity, although the inputs of science – presence and characteristics of S&T human resources - are also the focus of some studies (for example, Prpic 2002; Bornmann & Enders, 2004; Mauleon & Bordons, 2010; Penas & Willett, 2006; Symonds *et al*. 2006; van Arensbergen et al., 2012). The Scientometrics field has also been receiving new contributions of authors bringing about Brazilian evidences of women participation in S&T and their productivity (for example, Batista & Leta, 2009; Leta & Lewison, 2003; Olinto, 2009).

In the present study, we try to bring a less frequent approach to the mentioned theme with the particular characteristic of considering productivity indicators together with indicators of other academic tasks that are assumed, or should be assumed, by teacher-researchers in graduate programs. This approach has already been taken by Izquierdo *et al* (2004). These authors found that women from the Universidad Autónoma de Barcelona tend to dedicate more time in teaching and other 'invisible' academic tasks. Another recent study found that this scenery varies according to academic area (NAS, 2009).

Bringing Bourdieu's ideas to the data we have available, we assume that participation in more prestigious and visible tasks – as doctoral advising, banking participation and publication of journal articles – can represent more solid gains in scientific capital, and, as a consequence, better career chances to teacher-researchers. The opposite is expected from those that dedicate themselves more intensely to activities of lower prestige, as teaching and tutoring at the undergraduate level. Our expectation is, therefore, that women would receive higher share of these less valued activities.

Methodology

With a quantitative approach, this study uses the documental analysis technique. The basic and unique documental source of information is a collection of pre-established forms – 11 in total - elaborated by CAPES. These forms have to be annually filled by each Brazilian graduate program recognized by this Ministry of Education agency and all data contained in these forms are also used for evaluation of these programs. Hence, these filled forms are official documents with information about various characteristics, as well as quantitative and qualitative data, of each teacher-researcher who participate in graduate programs. All this information is accessible to anyone and published in *Cadernos de Indicadores* available through the URL: http://conteudoweb.capes.gov.br/conteudoweb/CadernoAvaliacaoServlet.

For the present study, we selected three of the above mentioned forms: *CD - Corpo Docente, Vínculo Formação*, *DP - Docente Produção* and *DA - Docente Atuação*. These three forms were downloaded for each of the 2,247 graduate programs evaluated in 2009. A total of 6,741 archives were downloaded in a *pdf* format and converted to a Microsoft Excel format with the help of Cogniview Software.

A particular difficulty for studies about women in science is the availability of information about the scientist's gender, since most published articles do not identify the author's gender and this variable is not available in most databases. Therefore, bibliometric studies tend to deal with small data sources or have to face the difficult task of identifying sex in large data sources. This problem has been pointed out by Mauleon & Bordons (2010). Hence, aware that the gender of the 52,294 teacher-researchers affiliated to these graduate programs was not included in CAPES' forms, we developed a series of strategies to allow for this classification. Firstly, we developed a software to confront these 52,294 teacher-researchers' names, which appeared in the forms with the corresponding names catalogued in Lattes Curriculum - a curriculum database organized and supported by CNPq (National Council for Scientific and Technological Research) (MCT, 2012). All Brazilian scientists are required to have their curriculum in the Lattes database. As this database furnishes information about the scientist's gender, we could identify the gender of a large portion of Brazilian teacher-researchers included in the forms. However, due to a few inconsistencies, and to the dynamic aspect of the Lattes database, some of the teacher-researchers' names were not found. Therefore, a new strategy was then developed: a list of the more frequent first names was elaborated and, based on this list, the corresponding gender was attributed to each name. Still, this strategy was not sufficient for gender identification of a few lasting names that were checked, one by one, at the Lattes database and/or at the internet. Of the total population of teacher-researchers mentioned above, only 79 cases, or 0,001%, could not receive sex identification.

Once the problem of gender classification was solved, data in Excel, extracted from the three above mentioned forms, were then processed by a program developed by Pablo Batista to generate a single archive. Eventual inconsistencies

were once again checked and corrected manually. Finally, it was possible to generate a data matrix, referring to the information contained in the three forms for each teacher-researcher. The last methodological step taken to prepare data for analysis was the conversion of this matrix to SPSS (Statistical Package for the Social Sciences), version 12.

The population of the study represented in this matrix can be so defined: teacher-researchers that participated in graduate programs in Brazil in 2009 (N=52,294). The initial moment of data analysis was concentrated in the selection and reclassification of variables that are taken into account for the present study. Among the variables that characterize the unit of analysis – each teacher-researcher – two groups were focused in the present study: (a) basic characteristics of the teacher-researcher (gender, S&T area, year of doctoral title and institution of affiliation) and (b) academic roles performed by each teacher-researcher, or, in other words, responsibilities that could contribute to his or her scientific capital (teaching undergraduate and graduate courses; undergraduate and graduate advising; banking participation, project leadership and types of publications).

For the classification of S&T area of the graduate programs, we utilized the grouping of nine categories considered by CNPq, which is available through the URL: http://memoria.cnpq.br/areasconhecimento/index.htm.

With regard to institutional affiliation, due to the diversity of the institutions in which the teacher-researcher develops his or her activities, we have ordered them from the highest to the lowest frequency and have selected the first 81 institutions that absorbed just above 80% of teacher-researchers. This set of institutions was, then, manually classified as public or private.

Academic roles performed by teacher-researchers were here considered as having different values with regard to the amount of 'scientific capital' that could be attributed to them. Some were considered as having low scientific capital or prestige. These are typically responsibilities with undergraduate students: teaching undergraduate courses and undergraduate tutoring. Some other activities were considered as having low level of scientific capital but could be placed in a higher position than the first ones mentioned. This is the case of teaching graduate courses and graduate level advising. A last group of four activities are here taken as indicators of high scientific capital. Research project leadership and banking participation are in this group and at the same time suggest the presence of the institutional dimension of the scientific capital. The highest value of scientific capital was attributed to publication of full articles in Annals and in Academic Journals. These two last activities were considered as the closest to the idea of pure scientific capital.

Results

Two blocks of results are presented here: 1) a table containing some basic characteristics teacher-researcher of Brazilian graduate programs considering

their gender; 2) a series of figures focusing on the relation between gender and academic responsibilities assumed by teacher-researchers.

Basic characteristics of male and female teacher-researchers

The increase in the incorporation of qualified women in Brazilian academia is especially remarkable, as demonstrated by a survey undertaken by MEC - the Brazilian Ministry of Education (MEC, 2010). According to this study, the proportion of women as teachers in public universities has increased from 39.7% in 2000 to 42.6% in 2005, and the proportion of women in this work sector with PhD has also increased from 32% to 38.6% in the same period.

Table 1 shows information obtained from data generated for this study: some characteristics of the population of 52,294 teacher-researches members of graduate level courses in the country in 2009. From these results we can point out that the percentage of women in graduate programs is similar to their participation in undergraduate courses, as observed by MEC.

Focusing on S&T areas, it is possible to perceive in Table 1, as expected, that women are still a clear minority in the Hard Sciences and Engineering whereas in Language/Letters and Humanities they overpass 50% forming a majority group. This type of gender segregation in S&T, is known as "horizontal segregation" or "territorial segregation" (Schienbinger, 2001), revealing the process through which, professional, scientific and technological fields become clearly "gendered", with women tending to predominate in less prestigious fields, receiving lower salaries. This is not a specificity of Brazilian teacher-researchers; international studies reveal a similar tendency. An example is the study conducted by the European Commission, showing large gender gaps between scientific fields but also showing large differences among countries, which point to the subtle socio-cultural processes that are at the origin of these gender gaps (EU, 2009).

With the information contained in the three last lines of Table 1 we try to detect important aspects related to the formation and institutional allocation of teacher-researchers in Brazil. The objectives are: 1) describe how public institutions contribute to graduate programs and their equal absorption of male and female teacher-researchers; 2) identify the absorption of new faculty members in graduate programs in the last decade (2000-2009), showing the relative participation of men and women among these new faculty members; 3) indentify the endogenous versus exogenous character of PhD training in Brazil and the share of men and women in these different types of training. The percent values that appear for each gender were calculated from the totals of each gender group.

As the results in Table 1 clearly show, public institutions - universities and research institutes - are the main locus of graduate teaching and research, with both genders represented in similar proportions. This is also the case in practically all countries (EU, 2009). Universities are also the institutions that absorb greater and increasing percentages of women in S&T, as pointed out by a study of The European Commission (2009). According to this study, in Europe "the average

annual growth rate for female researchers has stood at 4.8%, compared with 2.0% for male researchers" (p.23).

Table 1 - Distribution of teacher-researchers in Brazilian graduate programs by scientific area, type of institution, period of doctoral title and gender. Brazil, 2009.

Characteristic	Men (%)	Women (%)	Men (N)	Women (N)
All programs	59.4%	40.6%	31,033	21,182
By main área				
Agrarian Sciences	67.1%	32.9%	1,317	645
Applied Social Sciences	65.4%	34.6%	3,806	2,017
Biological Sciences	54.8%	45.2%	3,484	2,872
Engineering	79.3%	20.7%	5,117	1,333
Exact Sciences	76.3%	23.7%	4,947	1,540
Health Sciences	48.6%	51.4%	4,763	5,037
Human Sciences	48.5%	51.5%	4,349	4,609
Interdisciplinary	59.7%	40.3%	2,406	1,623
Language, Letters	35.9%	64,1%	844	1,506
Public Institutions *#	92.4%	92,0%	23,395	16,346
Doctoral titles between 2000-2009#	41.6%	44.6%	12,899	9,434
Doctoral titles in Brazil#	75.8%	83.5%	23,516	17,689

Notes: Classification of S&T areas was based in the nine categories considered by CNPq (http://memoria.cnpq.br/areasconhecimento/index.htm).
*The Categories "public" and "private" institutions were calculated for the 81 institutions with highest frequencies, which absorb more than 80% of teacher-researchers in graduate programs.
Percent values are calculated from totals in each gender.

With respect to the absorption of new members in recent periods, data indicate that a reasonably large portion of graduate personnel was titled in the last decade. This result is due to a major increase of graduate programs in the country, as already mentioned. It is also noticeable that women are in greater proportion among those more recently titled. This seems to be is a good indicator; it shows a relative increase in the participation of women in those programs in the last decade considered (from 2000 to 2009).
Regarding exogenous versus endogenous training, data point out to a massive training in the country for both genders, but women show even greater values for training in Brazil. To explain this difference we can resort to the argument that due to accumulation of house tasks and responsibilities women would have lower mobility, their academic networks would have fewer connections and they would tend to establish fewer collaboration ties abroad. In fact, the burden of housework in daily routines of female faculty members, and their potential effects on academic performance, has been the subject of various studies and also the object of political demands posed in the context research institutions, as Harvard and

MIT (Fox, 2001, 2010; MIT, 1999; Lamont, 2004). As Fox (2010, p.997) stated, "Faculty in academic science and engineering are a strategic group for consideration of work and family/household interference because of particularly high demands in their work time, workload, work commitment, and scheduled benchmarks for performance".

Gender and academic roles of Brazilian teacher-researchers

Figure 1 presents the average number of the different types of academic activities in which teacher-researchers of Brazilian graduate programs are involved. Taking into account the idea of 'scientific capital' - suggesting differences in the symbolic value that could be attributed to them – we considered that the identified activities could be positioned in a rank order of return in 'scientific capital'. We considered teaching undergraduate courses with the lowest scientific capital return and publication of academic articles as the activity with the highest return, the one that would come closest to the idea of 'pure scientific capital'. The other activities are presented in Figure 1 according to their lower or higher proximity to these two extreme points.

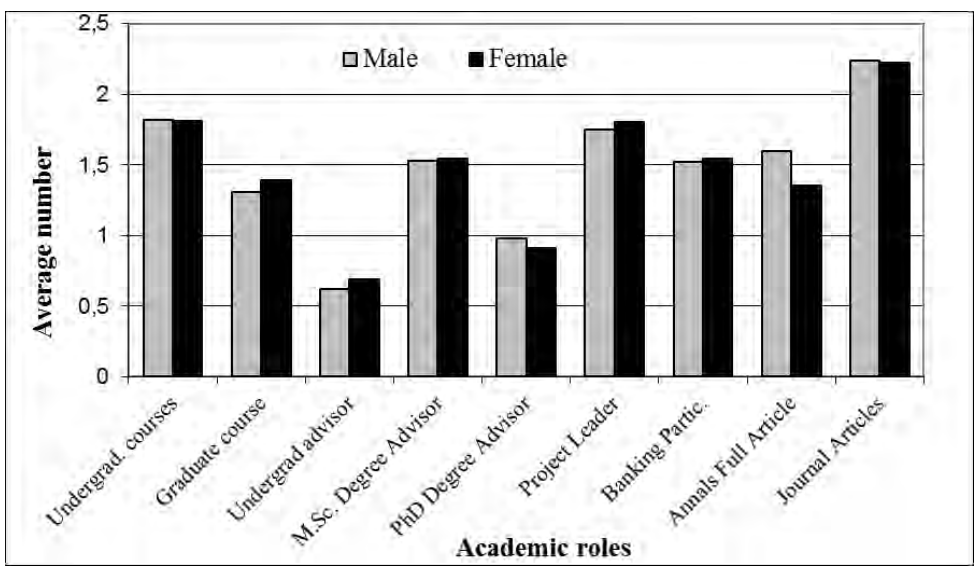


Figure 1: Average number of participation in different academic roles by gender. Brazilian teacher-researchers in Graduate programs in 2009.

Our expectation was that female teacher-researchers would be proportionally more represented in the lower valued activities, whereas male teacher-researchers would be proportionally more concentrated in the higher valued activities. However, the results for the studied population indicate that male and female teacher-researchers seem to be having an equal share of these activities, regardless of their symbolic value.

Aiming at a more detailed look at gender differences in academic roles, we selected two variables which represent the participation in activities at the two extremes of the symbolic value ladder: number of undergraduate courses given (with the lowest value) and number of articles published in academic periodicals (with highest value). From figure 2, which shows relative frequencies men and women participation in these two activities, it is possible to conclude that both genders have similar presence with similar performance.

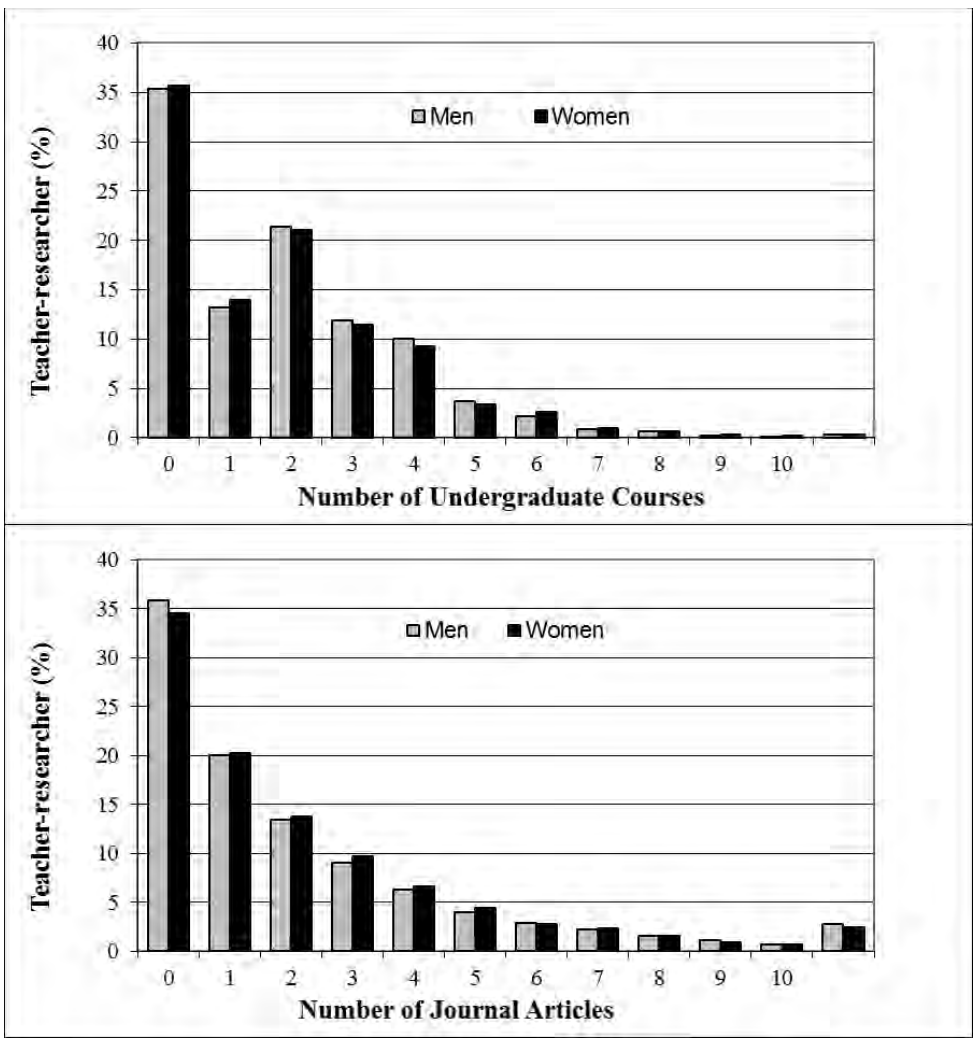


Figure 2: Participation of Brazilian teacher-researchers in (A) number of undergraduate courses given by gender and (B) number of articles in periodicals by gender, 2009.

An interesting result of both sections of Figures 2 is that increase in frequency of course teaching and publishing articles does not indicate substantial gender differences.. However, with regard to publishing, extreme values seem to favor men – zero publication and more than six articles -, whereas the relative high proportion of those who publish from one to five articles is better represented by women faculty members.
What is also outstanding in this figure is that about 35% of both male and female graduate faculty members did not teach undergraduate courses and about 30% of them did not publish at all during the academic year of 2009.

Gender and academic roles of Brazilian teacher-researchers: comparison between scientific areas

The analysis of table 1 showed us a clear segmentation of male and female teacher-researchers by academic field, indicating that, as it occurs in most other countries, a process of 'territorial gender segregation' characterizes S&T in Brazil.

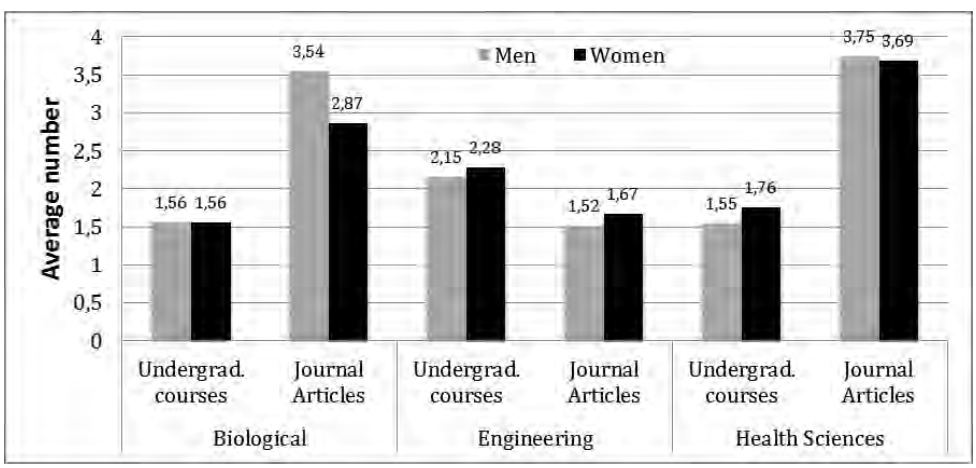


Figure 3: Average number of participation in different academic roles by gender and main area. Brazilian teacher-researchers in Graduate programs in 2009.

Focusing in a comparison between academic areas, the question we pose here refers to the different dynamics related to the performance of academic roles that might occur inside scientific fields with different gender imbalances. Would male or female predominance in a specific scientific field contribute to a different in gender behavior with regard to involvement in lower and higher valued academic tasks? To answer this question we selected three of the inclusive academic areas shown in Table 1: Heath Sciences in which women form a majority (51,4% female); Engineering, a predominantly masculine field (79,3% male) and Biological Sciences in which men predominate but not much above the majority (54,8% male). Data presented in Figure 3, below, describes participation of male

and female teacher-researchers in the least and the most valued academic tasks considered in Figure 2: teaching undergraduate courses and publishing journal articles.

Figure 3 indicates that some differences occur between the areas here considered. Focusing on the Biological sciences it is possible to perceive that there is a gender balance in the less prestigious task – teaching - but male predominance in the most prestigious one – journal article publishing. In Engineering, the most masculine area, women are prominent in both types of roles, which might reflect a compensatory mechanism: fearing competition or gender discrimination in a field where they are a small minority, they maintain higher productivity levels than their male colleagues and, at the same time, they are more involved in other heavy academic tasks, which is the case of undergraduate teaching. The Health Sciences, an area in which men are slightly predominant, we observe a tendency of gender task segregation that seems to go in the direction of the theoretical discussion and previous evidence, as considered at the beginning of this article: in comparison to their male colleagues, women show a relatively higher presence in lower valued tasks – teaching - and at the same time a lower presence in the most valued task - journal articles production. These results suggest that gendered behavior and segregation might be sensitive to the differences in gender balance found in the diverse scientific fields.

Concluding remarks

In this study, we tried to identify evidences about differences between men and women with respect to their main academic roles in the context of graduate level programs in Brazil. Comparison between less visible and less prestigious tasks with and more visible and prestigious ones were undertaken. Bourdieu's concept of 'scientific capital' was the main theoretical source considered.

Initial evidences, taking into account the population of more than 50 thousand faculty members of these programs, have not shown substantial differences between genders. This general result seems positive: Brazilian women who are graduate faculty members are keeping up with their male colleagues, with similar publishing outcomes and sharing equally heavy and lower valued tasks, as it is the case of teaching undergraduate courses.

However, changing our focus – adjusting our zoom lenses – we tried to identify gender differences between scientific fields. We took into consideration areas which include women in different proportions: Biology, Engineering and Health Sciences. The results indicate that in some areas men and women have similar attribution of lower and higher valued tasks. In others, however, notably in Health Sciences, women are having larger share of less valued talks than men: a burden that might help to explain the difference in productivity favoring men in this field. These results support Izquierdo *et al.* (2004) and NAS (2009) findings. These results also suggest that attention should be given to the specificities of 'scientific fields' and their possible effect on gender discrimination.

The distinctive results obtained when one passes from macro analyses– describing the whole population of teacher-researchers - to a closer look that takes into consideration specific fields seem to point at the same time to the strength and the weakness of the analyses presented here. The strength of our study is the fact that it deals with this huge amount of data covering all Brazilian graduate programs; its weakness is the fact that the macro analyses might not be sufficient to the identification of gendered processes. Different types of contextual effects on gender discrimination might be counter balanced when this macro perspective is undertaken.

Brazil is a country with enormous social and cultural contrasts. Besides analyses of gender differences related to scientific field, several other aspects can still be taken into account in further studies with the same database, in order to better understand gender differences in science. Regional differences seem to be an interesting focus of analysis to observe the fulfillment of academic roles by gender. Some other variables that might also be taken into consideration in other studies are those related to the program's institution which varies from large and traditional public institutions to small private and recent ones. Another important variable seems to be the program evaluation by the Ministry of Education. Are men and women fulfilling similar academic roles in those different contexts?

The gender question is still a very present subject. This study does not intend to be exhaustive but to contribute to the relevant discussion of women participation in science.

Acknowledgments

Authors are grateful to CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for financial support.

References

Batista, P.D. & Leta, J. Brazilian Authors' Scientific Performance: Does Gender Matter?. In: 12th International Conference of the International Society for Scientometrics and Informetrics. Rio de Janeiro: BIREME/PAHO/WHO and Federal University of Rio de Janeiro, 2009. v. 1, p. 343-353

Bornmann, L. & Enders, J. (2004). Social origin and gender of doctoral degree holders. Impact of particularistic attributes in access to and in later career attainment after achieving the doctoral degree in Germany. *Scientometrics, 61*(1), 19-41.

Bourdieu, P. *Os_usos sociais da ciência*. São Paulo: UNESP, 2003.

EU, European Commission. She Figures 2009. Statistics and Indicators on Gender Equality in Science – available at: http://ec.europa.eu/research/science-society/document_library/pdf_06/she_figures_2009_en.pdf. Access: January, 2013.

Fox, MF (2001) Women, Science, and Academia: Graduate Education and Careers, *Gender & Society* 15: 654- 666.

Fox, M.F. (2010) Women and Men Faculty in Academic Science and Engineering Social-Organizational Indicators and Implications**.** *American Behavioral Scientist*, v. 53, n. 7, p. 997-1012.

Izquierdo, M.J., Mora, E., Duarte, L. & Le.n, F. J. 2004, El sexisme a la UAB: propostes d'actuaci. In guia per a un diagnostic , Universitat Autonoma de Barcelona, Servei de Publicacions, Bellaterra.

Lamont et al. (2004) *Recruiting, promoting, and retaining women academics: lessons from the literature*. A report of the Standing Committee for the Status of Women, Faculty of Arts and Science, Harvard University. Available at: < http://www.wjh.harvard.edu/~mlamont/lessons.pdf>. Access: September, 2012.

Leta, J. (2011) Growth of Brazilian Science: a real internalization or a matter of databases' coverage? In: 13th International Conference of the International Society for Scientometrics and Informetrics, 2011, Durban, Africa do Sul. Proceedings of the ISSI 2011 Conference. Leiden, Zululand: Leiden Univ & Zululand Univ, v. 1. p. 392-397.

Leta, J. & Lewison, G. (2003) The contribution of women in Brazilian Science: a case study in astronomy, immunology and oceanography. *Scientometrics*, v. 57, n. 3, p. 339-353.

Long, J. S. (1990). The origins of sex differences in science. *Social Forces*, 68, 4, 1297-1315.

Long, J.S. (1992) Measures of sex differences in scientific productivity. *Social Forces*, v.71, n.1, p.159-178.

Mauleón, E. & Bordons, M. (2010). Male and female involvement in patenting activity in Spain. *Scientometrics*, 83, 605-621.

MCT, CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). Plataforma Lattes, 2012. Available at: http://lattes.cnpq.br

MEC, INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira). Resumo técnico censo da educação superior de 2010. Brasília: INEP, 2012. Available at: <http://download.inep.gov.br/educacao_superior/censo_superior/resumo_tecnico/resumo_tecnico_censo_educacao_superior_2010.pdf> Access: September, 2012.

NAS, National Academy of Sciences 2009. Gender Differences at Critical Transitions in the Careers of Science, Engineering and Mathematics Faculty, The National Academies Press, Washington, D.C.

OECD. Measuring gender (ine)quality: introducing gender institutions and development data base (GID): DEV/DOC(2006)1, 2006. Available at: <http://www.oecd.org/dev/36228820.pdf>. Access: September, 2012.

Olinto, G. & Leta, J. (2011) Gender (im)balances in teaching and research activities in Brazil. In: 13th International Conference of the International Society for Scientometrics and Informetrics, 2011, Durban, Africa do Sul. Proceedings of the ISSI 2011 Conference. Leiden, Zululand: Leiden Univ & Zululand Univ, v. 2. p. 618-637.

Olinto G. (2009) Human Resources in Science and Technology Indicators: longitudinal evidence from Brazil. In: 12th International Conference of the International Society for Scientometrics and Informetrics. Rio de Janeiro: BIREME/PAHO/WHO and Federal University of Rio de Janeiro, v. 1. p. 359-368.

Penas, C. S., & Willett, P. (2006). Gender differences in publication and citation counts in librarianship and information science research. *Journal of Information Science*, 32(5), 480–485.

Prpic, K. (2002). Gender and productivity differentials in science. *Scientometrics*, 55(1), 27–58.

Regalado, A. (2010). Brazilian Science: Riding a Gusher. *Science*, 330 (6009), 1306-12.

Schienbinger, L. O feminismo mudou a ciência? Bauru: EDUSC, 2001.

Symonds MR, Gemmell NJ, Braisher TL, Gorringe KL, Elgar MA (2006) Gender Differences in Publication Output: Towards an Unbiased Metric of Research Performance. PLoS ONE 1(1): e127. doi:10.1371/journal.pone.0000127. available at: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0000127;jsessionid=9494355C2878F51D266E02CFD87D6E9E

UNESCO. Science, technology and gender: an international report, 2007. Available at: <http://unesdoc.unesco.org/images/0015/001540/154027e.pdf>. Access: September, 2012.

van Arensbergen P, van der Weijden I & van den Besselaar P (2012) Gender differences in scientific productivity: a persisting phenomenon? *Scientometrics*, vol 93, 857–868.

GENDER INEQUALITY IN SCIENTIFIC PRODUCTION (RIP)

Maite Barrios[1], Anna Villarroya[2], Candela Ollé[3], Lidia Ortega[4]

[1] *mbarrios@ub.edu*
Department of Behavioral Sciences Methods, University of Barcelona, Spain

[2] *annavillarroya@ub.edu*
Department of Public Economy, Political Economy and Spanish Economy, University of Barcelona, Spain

[3] *candela.olle@gmail.com*
Faculty of Information and Communication. Open University of Catalonia, Spain

[4] *lortegca8@gmail.com*
Faculty of Psychology, University of Barcelona, Spain

Abstract
This study aims to identify possible gender imbalances in the scholarly output of Spanish researchers in several areas of science through a study of documents published in 2007 extracted from the Thomson Reuters Web of Science (WoS). The researchers' scientific output was studied through two indicators: publication practices (signing order of authors in the documents and proportion of females per article) and impact of the articles (measured through the number of citations received). The results show that in all scientific areas the average proportion of female authors per article and presence of women signing as first author were lower than for men. Moreover, in some disciplines, there was a statistically significant relationship between the gender of the first author and the proportion of female authors in the article, which was higher when the first author was female. However, no significant differences were found in the number of citations received between articles signed by a male or a female or in the proportion of females present in the documents. In summary, the study highlights a gender bias towards the presence of females in different areas of science, although the impact of the research conducted by the two genders seems to be similar.

Conference Topic
Scientometrics Indicators: Relevance to Science and Technology, Social Sciences and Humanities (Topic 1).

Introduction
Many studies have shown that women are under-represented in the scientific profession science, especially in the higher echelons. Scholarly publishing is central to academic success since the quantity and quality of publications affects tenure and promotions, access to funding, and salaries. As a way of gauging the

performance of women in the academic environment the present study aims to examine the publication patterns of male and female researchers.

At international level, many gender studies have focused on the quantitative analysis of productivity. The results have differed depending on the countries, disciplines and variables assessed. While most of these studies have reported lower productivity among women in terms of the publication of academic research, others have found no relationship between gender and productivity (for a review of the literature, see Borrego et al., 2010).

Besides productivity, the impact of the articles published, measured through the number of citations received, is another important indicator in the evaluation process. Studies on this topic have also shown contrasting results. While some studies have noted no differences (Cole and Zuckerman, 1984; Lewinson, 2001; Ledin, Bornmann, Gannon and Wallon, 2007; Mauléon et al. (2008); Copenheaver, Goldbeck and Cherubini, 2010), others have shown a higher average number of citations per paper in the case of women (Long, 1992; Symonds et al., 2006; Borrego et al., 2010) and still others a lower number of citations (Hunter and Leahey, 2010; Larivière et al., 2011). The differences reported may be attributable to factors such as the country where the study was performed, the time period analysed, the type of institution evaluated, the methodology used and, most importantly, the scientific area analysed. Since the participation of women is not equal in all areas of science the present study takes into account a range of scientific disciplines (e.g., Computer Science, Economy, Engineering, Neuroscience, Pharmacology, Physics and Chemistry and Psychology) in order to assess potential gender imbalances in the scientific production published by Spanish researchers. We focus on the authorship of publications, patterns of collaboration and the impact of scientific output in each area.

Method

A serious methodological problem facing gender studies is the fact that most bibliographical databases – including the Thomson Reuters ISI Web of Science (WoS), which is usually the main source of bibliometric data – identify authors by their initials alone. However, in 2007 the WoS began to incorporate authors' full names, thus facilitating the identification of the gender of the researchers.

Using the WoS, we selected all articles and reviews published in journals classified in any of the seven scientific disciplines (Computer Science, Economics, Engineering, Neuroscience, Pharmacology, Physics and Chemistry and Psychology) in 2007 containing Spain in the affiliation field of the authors. Among the documents retrieved, we eliminated those in which the gender of the authors could not be identified and those in which the corresponding author's affiliation was not in Spain. As the rules for the order of co-authors vary significantly between disciplines, the analysis of the authorship of publications, patterns of collaboration and impact were carried out using only the documents that did not list the co-authors in alphabetical order.

We recorded the gender of the first author and we calculated the proportion of females per article, as well as the proportion of female co-authors depending on the gender of the first author.
The impact of the publications was measured through the number of citations received by the article from its publication until December 2012. Log transformation was applied in order to improve the normality of this variable.

Results

The discipline with the highest number of publications in which the authors are ordered alphabetically was Economics, while the disciplines that showed the lowest use of alphabetical order were Neuroscience and Pharmacology (Table 1).

Table 1. Documents with authors listed in alphabetical order and non alphabetical order by discipline

	Non- alphabetical order	*Alphabetical order*
	n (%)	*n (%)*
Computer Science	*371 (67.6%)*	*178 (32.4%)*
Economics	*56 (25.1%)*	*167 (74.9%)*
Engineering	*431 (83.5%)*	*85 (16.5%)*
Neuroscience	*427 (92.4%)*	*35 (7.6%)*
Pharmacology	*435 (91.4%)*	*41 (8.6%)*
Physics and Chemistry	*513 (86.1%)*	*83 (13.9%)*
Psychology	*373 (78.5%)*	*102 (21.5%)*

Table 2. Gender of the first author by discipline in alphabetical and non alphabetical order

	Gender of first author	*Non alphabetical order*	*Alphabetical order*	χ^2 *(p-value)*
Computer Science	*Men*	*309 (57.3%)*	*138 (25.1%)*	*2.638 (p =.104)*
	Women	*62 (11.3%)*	*40 (7.3%)*	
Economics	*Men*	*33 (14.8%)*	*132 (59.2%)*	*8.816 (p =.003)*
	Women	*23 (10.3%)*	*35 (15.7%)*	
Engineering	*Men*	*360 (69.8%)*	*68 (13.2%)*	*0.624 (p =.429)*
	Women	*71 (13.8%)*	*17 (3.3%)*	
Neuroscience	*Men*	*221 (47.8%)*	*14 (3%)*	*1.789 (p =.181)*
	Women	*206 (44.6%)*	*21 (4.5%)*	
Pharmacology	*Men*	*219 (46%)*	*20 (4.2%)*	*0.037 (p =.848)*
	Women	*216 (45.4%)*	*21 (4.4%)*	
Physics and Chemistry	*Men*	*304 (51%)*	*57 (9.6%)*	*2.652 (p =.103)*
	Women	*209 (35.1%)*	*26 (4.4%)*	
Psychology	*Men*	*205 (43.2%)*	*66 (13.9%)*	*3.105 (p =.078)*
	Women	*168 (35.4%)*	*36 (7.6%)*	

Economics was the only discipline in which, when alphabetical order was used, the proportion of males signing as first author was significantly higher and consequently the proportion of females was lower (Table 2).

Participation of women per article

In order to study the participation of women in multi-authored papers, single-authored papers were excluded from the analysis. The percentage of women signing as first author was lower than expected on the basis of population data (Spanish National Statistics Institute –INE-, 2012) in Computer Science and in Psychology (Table 3). All other scientific disciplines showed similar percentages to the population data with a confidence interval of 95%. The average proportion of women per paper (co-authors) was lower than expected on the basis of the INE data in five scientific disciplines (i.e., Computer Science, Engineering, Pharmacology, Physics and Chemistry and Psychology) (Table 3).

Table 3. Presence of women in publications

Disciplines (n of documents)	*% of women according to INE*	*% of women as first author*	*Average proportion of women per paper*
Computer Science (n=371)	*21.6%*	*16.7% (n=62) [CI: 13.0%, 20.5%]*	*0.14 (SD: 0.24) [CI:0.12, 0.18]*
Economics (n=56)	*36.3%*	*41.1% (n=23) [CI: 28.2%, 54.0%]*	*0.40 (SD: 0.43) [CI: 0.28, 0.51]*
Engineering (n=431)	*18.1%*	*16.5% (n=71) [CI: 13.0%, 20.0%]*	*0.12 (SD: 0.22) [CI: 0.10, 0.14]*
Neuroscience (n= 427)	*	*48.2% (n=206) [CI: 43.5%, 53.0%]*	*0.41 (SD: 0.27) [CI:0.38, 0.43]*
Pharmacology (n=435)	*51.5%*	*49.7% (n=216) [CI: 45.0%, 54.4%]*	*0.43 (SD: 0.27) [CI : 0.41, 0.47)*
Physics and Chemistry (n=513)	*41.4%*	*40.7% (n=209) [CI: 36.4%, 45.0%]*	*0.30 (SD: 0.26) [CI:0.28, 0.33]*
Psychology (n= 373)	*52.5%*	*45.0% (n=168) [CI: 40.0%, 50.0%]*	*0.44 (SD: 0.37) [CI: 0.40, 0.48]*

INE: Spanish National Statistics Institute. SD: Standard deviation, CI: Confidence interval at 95%. Note*: Data not available.

Additionally, in Pharmacology, Physics and Chemistry, and Psychology, the data presented statistically significant differences in the proportion of female authors depending on the gender of the first author (Table 4). That is, when the first author was female, the average proportion of female co-authors per paper was higher, whereas when the first author was male, the average proportion of women was lower.

Table 4. Average proportion of women depending on the gender of the first author.

Disciplines		*Mean (SD, [95% CI]*	*student t (d.f.)*	*p-value*
Computer Science	*Men (n=309)*	*0.13 (SD: 0.24), [CI: 0.11, 0.16]*	*1.550 (369)*	*.122*
	Women (n=62)	*0.19 (SD: 0.22), [CI: 0.13, 0.24]*		
Economics	*Men (n=33)*	*0.41 (SD: 0.44), [CI: 0.26, 0.57]*	*0.440 (54)*	*.662*
	Women (n=23)	*0.36 (SD: 0.43), [CI: 0.16, 0.55]*		
Engineering	*Men (n=360)*	*0.11 (SD: 0.22), [CI: 0.09, 0.14]*	*1.887 (429)*	*.060*
	Women (n=71)	*0.17 (SD: 0.21), [CI: 0.12, 0.22]*		
Neuroscience	*Men (n=221)*	*0.40 (SD: 0.27), [CI: 0.36, 0.43]*	*0.895 (425)*	*.371*
	Women (n=206)	*0.42 (SD: 0.26), [CI: 0.38, 0.46]*		
Pharmacology	*Men (n=219)*	*0.40 (SD: 0.27), [CI: 0.37, 0.44]*	*2.186 (433)*	*.029*
	Women (n=216)	*0.46 (SD: 0.26), [CI: 0.43, 0.49]*		
Physics and Chemistry	*Men (n=304)*	*0.28 (SD: 0.25), [CI: 0.25, 0.31])*	*2.676 (511)*	*.008*
	Women (n=209)	*0.34 (SD: 0.26), [CI:0.31, 0.38]*		
Psychology	*Men (n=205)*	*0.40 (SD: 0.38), [CI: 0.35, 0.45]*	*2.282 (371)*	*.023*
	Women (n=168)	*0.49 (SD: 0.36), [CI: 0.43, 0.54]*		

SD: Standard deviation, CI: Confidence interval at 95%, d.f.: degrees of freedom.

Impact

Analysis of covariance showed that after controlling for the number of authors signing the article, number of pages and references, there were no differences in the number of citations depending on the gender of the first author in any of the seven scientific disciplines (Table 5). Neither did partial correlation controlling for the same extraneous variables show a significant relationship between the proportion of women by article and the number of the citations that the document received (Table 6).

Table 5. Number of citations depending on the gender of the first author.

Disciplines	*Gender*	*mean citations (SD)*	*F value (d. f.)*	*p-value*
Computer Science (n=387)	*Men (n=327)* *Women (n=60)*	*9.89 (14.95)* *9.00 (10.10)*	*0.294 (1, 382)*	*.588*
Economics (n=103)	*Men (n=73)* *Women (n=30)*	*6.14 (6.63)* *5.23 (6,47)*	*2.167(1, 98)*	*.144*
Engineering (n=402)	*Men (n=340)* *Women (n=62)*	*11.49 (14.29)* *13.42 (20.21)*	*0.182 (1, 397)*	*.670*
Neuroscience (n=425)	*Men (n=225)* *Women (n=200)*	*17.86 (21.25)* *19.62 (30.10)*	*1.249 (1, 420)*	*.264*
Pharmacology (n=426)	*Men (n=221)* *Women (n=205)*	*16.18 (15.40)* *13.77 (13.74)*	*1.489 (1 ,421)*	*.223*
Physics and Chemistry (n=527)	*Men (n=319)* *Women (n=208)*	*16.46 (19.74)* *16.73 (16.66)*	*0.044 (1, 522)*	*.834*
Psychology (n=367)	*Men (n=204)* *Women (n=163)*	*8.80 (10.31)* *8.10 (9.47)*	*0.315 (1, 362)*	*.575*

Table 6. Relationship between the proportion of women by article and the number of citations

Disciplines (n)	$r_{xy.abc}$*(d.f.)*	*p-value*
Economics (n = 103)	*-0.148 (32)*	*.405*
Pharmacology (n = 426)	*0.014 (421)*	*.772*
Computer Science (n = 387)	*-0.087 (340)*	*.110*
Engineering (n = 402)	*0.026 (384)*	*.613*
Physics and Chemistry (n = 547)	*-0.075 (502)*	*.091*
Neuroscience (n = 425)	*0.047 (407)*	*.342*
Psychology (n = 367)	*-0.059 (322)*	*.287*

$r_{xy.abc}$*: partial correlation coefficient. d. f.: degree of freedom*

Conclusions

The data showed that women were the first authors of a lower proportion of papers than expected in all scientific disciplines. The same imbalance was observed in the proportion of female authors in general. Our results also indicate that in some areas there is a bias in the presence of women depending on the gender of the first author: the proportion of female authors is higher when the first author is a woman. These results have been attributed to the greater propensity of researchers to work with partners of the same gender (Ferber and Teiman 1980; McDowell and Smith 1992; Bentley 2003; Villarroya, Barrios, Borrego and Frías 2008; Barrios et al. 2012) and evidence a gender imbalance favouring men that makes it more difficult for women researchers to pursue their career.

Regarding the impact of publications, the analysis of the number of citations did not show gender differences depending on the gender of the first author or the

proportion of women signing the article. This result indicates that there is an equal recognition of scientific production within the disciplines.

Acknowledgments

The research reported here was supported by the Spanish Ministry of Education and Science (EA2008-022).

References

Barrios, M., Villarroya, A., & Borrego, A. (2012). Scientific production in psychology: a gender analysis. Scientometrics. Advance online publication. doi: 10.1007/s11192-012-0816-4.

Bentley, J. T. (2003). Gender differences in the careers of academic scientists and engineers: A literature review. Arlington, VA: National Science Foundation.

Borrego, A., Barrios, M., Villarroya, A., & Ollé, C. (2010). Scientific output and impact of postdoctoral scientists: a gender perspective. Scientometrics, 83(1), 93–101.

Cole, J., & Zuckerman, H. (1984). The productivity puzzle: persistence and change in patterns of publication of men and women scientists. Advances in Motivation and Achievement, 2, 217–258.

Copenheaver, C. A., Goldbeck, K., & Cherubini, P. (2010). Lack of gender bias in citation rates of publications by dendrochronologists: What is unique about this discipline? Tree-Ring Research, 66(2), 127–133.

Ferber, M. A., & Teiman, M. (1980). Are women economists at a disadvantage in publishing journal articles?. Eastern Economics Journal, 6(3-4), 189-193.

Hunter, L. A., & Leahey, E. (2010). Parenting and research productivity: new evidence and methods. Social Studies of Science, 40(3), 433–451.

Larivière, V., Vignola-Gagné, E., Villeneuve, E., Gélinas, P., & Gingras, Y. (2011). Sex differences in research funding, productivity and impact: an analysis of Québec university professors. Scientometrics, 87(3), 483–498.

Ledin, A., Bornmann, L., Gannon, F., & Wallon, G. (2007). A persistent problem. EMBO Reports, 8(11), 982–987.

Lewison, G. (2001). The quantity and quality of female researchers: a bibliometric study of Iceland. Scientometrics, 52(1), 29–43.

Long, J. S. (1992). Measures of sex differences in scientific productivity. Social Forces, 71(1), 159–178.

Mauleón, E., Bordons, M., & Oppenheim, C. (2008). The effect of gender on research staff success in life sciences in the Spanish National Research Council. Research Evaluation, 17(3), 213–225.

McDowell, J. M., & Smith, J. K. (1992). The effect of gender-sorting on propensity of coauthor: Implications for academic promotion. Economic Inquiry, 30(1), 68-82.

Symonds, M., Gemmell, N., Braisher, T., Gorringe, K., & Elgar, M. (2006). Gender differences in publication output: towards an unbiased metric of research performance. PLoS One, 1(1), e127.

Villarroya, A., Barrios, M., Borrego, A., & Frías, A. (2008). PhD theses in Spain: a gender study covering the years 1990–2004. Scientometrics, 77(3), 469–483.

GENETICALLY MODIFIED FOOD RESEARCH IN CHINA: INTERACTIONS BETWEEN AUTHORS FROM SOCIAL SCIENCES AND NATURAL SCIENCES

Yuxian Liu[1], Raf Guns[2], Wenjie Wei[3], Jiahui Yin[4]

[1] *yxliu@tongji.edu.cn*
Tongji University, Tongji University Library, Siping Street 1239, 200092 Shanghai, China

[2] *raf.guns@ua.ac.be*
University of Antwerp, IBW, Venusstraat 35, 2000 Antwerp, Belgium

[3] *wjwei@lib.tongji.edu.cn*
Tongji University, Tongji University Library, Siping Street 1239, 200092 Shanghai, China

[4] *jhyin@tongji.edu.cn*
Tongji University, Tongji University Library, Siping Street 1239, 200092 Shanghai, China

Abstract
In order to understand how Genetically Modified Food (GMF) research is developing in China under the interaction between authors from social sciences and natural sciences, we analyze the distribution of Chinese articles on GMF among different fields and study how the authors of these articles collaborate with each other. We construct a co-author network using Chinese articles on GMF and divide this network into the sub-networks of different fields. The fields are defined according to the Chinese Library Classification system (CLC). The fields of these articles on GMF comprise almost all fields in the CLC. Q-measures are used to characterize how the authors collaborate among these fields. Authors from most fields in the social sciences and humanities (SSH) occupy a peripheral position in the network. Authors from Economics are more central and bridge between authors from the natural sciences. However, although Economics is the largest field in our dataset in terms of number of articles, only a few authors have Q-measures which are larger than zero and the Q-measures of these authors are relatively small. In this sense SSH authors play a limited role in interdisciplinary collaboration among authors engaged in GMF research. The fields that ranked in the top by Q-measure values are the fields of Light industry (including Food industry), Agricultural Science, Medicine and hygiene and Environmental science. The authors from these natural science fields act as brokers between different disciplines pertaining to GMF research.

Conference
Topic 6 Collaboration Studies and Network Analysis

Introduction

Genetically modified food (GMF) is derived from genetically modified organisms (GMOs). The genetic material (DNA) in these organisms has been altered in a way that does not occur naturally. Genetic modification involves the insertion or deletion of genes. It allows selected individual genes to be transferred from one organism to another, also between non-related species (Genetically modified food, 2013).

GMF is a hot and controversial topic. The development of GMF has significant implications for the future of mankind. GMF is harvested from plants that have been modified to enhance desired traits such as increased resistance to herbicides or improved nutritional content. Comparing with mutation breeding, genetic modification can introduce needed genes more precisely and within a much shorter period. Plant scientists, backed by results of modern comprehensive profiling of crop composition, point out that crops modified using GM techniques are less likely to have unintended changes than conventionally bred crops. So it may benefit humanity.

Since genetic modification left the laboratory, fruits, vegetables and industrial crops have been harvested and some have reached our tables as food. Concerns on environmental hazards, human health risks and economic implications arouse a lot of criticism against GMF. The European Union, and especially the UK, fears the use of GMF.

As the technology of genetic modifications holds more and more promises for mankind, more and more Chinese scientists are engaging in the research and publishing their results in science journals. Because of their effort and great achievement, Genetic modification technology (GMT) makes a great progress in China. GMT is one of the sixteen key projects in the *Outline of the National Plan for the Development of Science and Technology* (The PRC State Council, 2006). A security certificate for genetically modified rice was issued by the Ministry of Agriculture in 2009 (Ministry of Agriculture of PRC, 2009). In 2012, the national rural working conference announced that the Chinese will continue to implement the GMT strategy (Liu, 2012).

As a staple food, rice can be genetically modified and planted in China since 2009. The Chinese government is drafting a policy to decide whether staple food can be genetically modified or not. Chinese researchers argue how the policy should be made based on their knowledge and the benefit to their respective groups. These arguments are published in Chinese journals.

How strong are these arguments? To what extent do these researchers work together across disciplinary boundaries to strengthen their arguments? In this contribution, we will use scientometric and network analysis methods to analyze how these arguments happen among different fields in China.

Data collection

We searched for all articles about GMF from China National Knowledge Infrastructure (CNKI). CNKI is an e-publishing project which began in 1996. It

contains e-journals, newspapers, dissertations, proceedings, yearbooks, reference works, etc. CNKI is the world's most comprehensive online resource for accessing China's intellectual output. With more than 36 million articles and thousands added every day, CNKI is a major player in the Chinese academic environment. CNKI comprises four databases including Chinese Academic Journals Database (CAJ), China Core Newspaper Database (CCND), China Dissertations (CDMD), and China Conference Proceedings (CPCD).

We collected data on 22 January 2012, searching for the Chinese equivalents to the terms *Genetic Modification* and *Genetically Modified Food* within keywords, thesaurus, titles and full text, both precisely and vaguely. "Precisely" means that the articles recalled must have the same Chinese characters in the same order as the key terms, "vaguely" means that the retrieved article contains the same Chinese characters as the key terms, but maybe in different order.

Table 1. Number of articles on Genetic Modification (GM) and GMF in CNKI with different retrieval strategies

		Thesaurus	*Title*	*Keywords*	*Full-text*
Precisely	GM	57,472	24,614	19,333	292,309
	GMF	4,707	1,390	4,303	15,311
Ratio of GMF to GM (%)		8.2	5.6	22	5.2
Vaguely	GM	66,349	26,892	18,622	688,960
	GMF	4,355	1,601	2,239	142,222
Ratio of GMF to GM (%)		6.6	6	12	20.6

Obviously, the number of publications dealing with GM is much larger than GMF. Most of the GM related articles do not mention food. The ratio of GMF to GM ranges from 5% to 22%. So GMF does not dominate GM research. GM is a biological technique leading in the first applications in medicine. One can see that if we search for GM or GMF using a controlled vocabulary (thesaurus or keywords), we recall more articles with a precise retrieval strategy. If we try to retrieve them from free text fields, we recall more articles with a vague retrieval strategy.

Next we focus on GMF. On March 23rd, 2012, we searched for '转基因食品' (GMF) as a keyword for the period 1994 to 2012. We did not focus on any specific food such as rice or tomoatoes. These vegetables or crops might only be used as the test object or materials. We got 4612 articles in total. We downloaded the bibliometric data including title, author, institute, source, and published time. We also developed a program to download the classification code of every article (see the next section).

The code system for newspapers is different; also, the year book and the Sci-Tech Report have no code, and some other articles were not encoded. These data were deleted, leaving us with 3549 articles published in Chinese journals.

Constructing and subdividing the network

Using these articles on GMF, we constructed a collaboration network. The nodes in the network are the authors of the articles on GMF; two authors are linked if they have co-authored one or more articles. This network is unweighted. There are 2734 nodes, 4 of which are isolated. There are 3928 edges. The average degree is 2.9. There are 790 weakly connected components (4 isolates). The largest connected component consists of 179 nodes. The density is 0.00105. This network is very sparse, as can be seen in Figure 1.

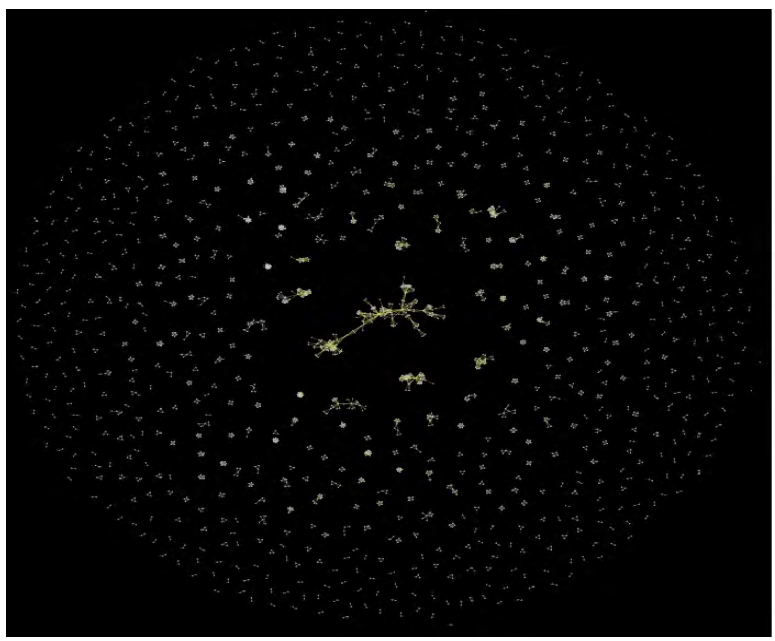


Figure 1. The sparse co-author network

We then used the classification code of every article to subdivide the network. The classification code is encoded according to the Chinese Library Classification System (CLC). It is a comprehensive knowledge classification system used widely in China. Chinese journals normally require the authors to provide the classification codes for their articles. CNKI use the codes in their bibliometric data.

According to the CLC system there are 22 main categories, shown in Table 2 (Liu & Rousseau, 2007).

These main categories are further subdivided into subcategories. These subcategories are encoded in digital numbers except subcategories of T (Industrial technology), whose first subcategories are encoded with an English letter. As GMT is also a kind of technology, we will also use the subcategories of category T. These subcategories are shown in Table 3.

Table2. Chinese Library Classification System

A	Marxism, Leninism and Chinese communism
B	Philosophy and religion
C	Social sciences
D	Politics and law
E	Military sciences
F	Economics
G	Culture, science (of sciences), education
H	Languages (incl. linguistics)
I	Literature
J	Arts
K	History and geography
N	Natural sciences (general)
O	Mathematics, physics and chemistry
P	Astronomy and geosciences (incl. marine sciences)
Q	Bioscience
R	Medicine and hygiene
S	Agricultural science (including forestry)
T	Industrial technology
U	Transportation
V	Aviation and spaceflight
X	Environmental sciences
Z	Others

Table 3. Subcategories of category T: Industrial technology

TB	Fundamental engineering technology
TD	Mining engineering
TE	Oil and natural gas industry
TG	Metal industry
TH	Mechanics
TJ	Weapon industry
TK	Dynamics and sources of energy
TL	Atomic energy technology
TM	Electrotechnics
TN	Wireless electronics and telecommunication technology
TP	Computer science
TQ	Industrial chemistry
TS	Light industry
TU	Architecture and urban planning
TV	Water conservation and irrigation

We assigned an attribute to every author according to the classification of his/her articles. If an author has several articles, we assigned an attribute according to the classification of her/his articles in which he/she acts as the first author or a single author. We then got a network that consists of several sub-networks of different fields.

Analyzing the interaction of these factors based on the structure of the network

The distribution of classifications

Normally every article has one classification code. But if the authors think their articles concern two or three topics, they may give several codes. The first code is the most important code which indicates the main content of the articles.

All the articles on GMF belong to 20 fields. Most articles are in the fields of Economics, Agricultural science, Light industry (including food industry), Bioscience, Politics and law, and Medicine and hygiene. 90% of articles on GMF are in these fields (see Table 4).

Table 4. Fields distribution of articles on GMF

Code	*Fields*	*Number of articles*	*% of articles*
F	Economics	1013	28.54
S	Agricultural science	894	25.19
TS	Light industry	663	18.68
Q	Bioscience	280	7.89
D	Politics and law	246	6.93
R	Medicine and hygiene	193	5.44
G	Culture, science (of sciences), education	96	2.71
X	Environmental sciences	72	2.03
B	Philosophy and religion	27	0.76
C	Social sciences	13	0.37
Z	Comprehensive	12	0.34
N	Natural sciences (general)	10	0.28
K	History and geography	8	0.22
TP	Computer science	6	0.17
I	Literature	4	0.11
O	Mathematics, physics and chemistry	4	0.11
TN	Wireless electronics and telecommunication technology	4	0.11
TH	Mechanics	2	0.06
E	Military sciences	1	0.03
J	Arts	1	0.03

Many articles have more than two classification codes, but most of these codes belong to one field. Only 71 out of these 3549 articles are assigned to two fields. Table 5 shows the fields involved: 11 out of 20 fields have articles that can be assigned to other fields. The two fields F (Economics) and TS (Light industry) have articles that are assigned to 7 other fields by their authors. The field of Economics contains articles that are assigned to B (Philosophy and religion), D (Politics and law), G (Culture, science (of sciences), education), R (Medicine and hygiene), TS (Light industry (including Food industry)) and X (Environmental sciences). The field of TS (Light industry) has articles that are also assigned to the

fields of D (Politics and law), F (Economics), O (Mathematics, physics and chemistry), Q (Bioscience), R (Medicine and hygiene), S (Agricultural science) and TP (Computer science). We can see that the fields of Economics and Food industry influence the research on GMF.

Table 5. Fields that have articles which are assigned to the other fields by authors

Field code	*The other fields that are assigned by authors*	*The number of the assigned fields*
F	B,D,G,R,S,TS,X	7
TS	D,F,O,Q,R,S,Tp	7
Q	B,R,S,TS	4
S	F,Q,R,TS	4
R	F,S,TS,	3
B	F,Q	2
D	F,TS	2
O	TP,TS	2
TP	TS,O	2
G	F,	1
X	F	1

Table 6 is the upper triangular matrix that shows the numbers of articles that are assigned to two fields. It may indicate how intensely these two fields are connected in the context of GMF. For instance, there are 20 articles that can be assigned both to F (Economics) and D (Politics and law), there are 9 articles that can be assigned into the fields TS (Light industry) and O (Mathematics, physics and chemistry), 6 articles that can be assigned into the fields S (Agricultural science) and F (Economics), 5 article that can be assigned into the fields G (Culture, science (of sciences), education) and F (Economics) at the same time. We can see that Economics is a very important field that is connected to several other fields in the context of GMF.

Table 6. the numbers of articles that are assigned into two fields

	B	*D*	*F*	*G*	*O*	*Q*	*R*	*S*	*TP*	*TS*	*X*
B	*	0	2	0	0	1	0	0	0	0	0
D		*	20	0	0	0	0	0	0	1	0
F			*	5	0	0	3	6	0	3	2
G				*	0	0	0	0	0	0	0
O					*	0	0	0	1	9	0
Q						*	1	5	0	7	0
R							*	1	0	1	0
S								*	0	2	0
TP									*	1	0
TS										*	0
X											*

Connections between the sub-networks

How do these sub-networks connect with each other? Do the authors in one field influence the authors from other fields? We use global and local Q-measures and betweenness centrality to answer these questions.

Global and local Q-measures and betweenness are concepts related to centrality. Centrality refers to the position or importance of a node relative to the entire network. Betweenness is one of these centrality measures. Betweenness centrality is a sophisticated measure that characterizes the importance of a given node for establishing short pathways between other nodes (Freeman, 1977). Mathematically, we assume we have a undirected network $G=(V,E)$, consisting of a set V of nodes and a set E of links between them. Betweenness is defined as:

$$C_B(a)=\frac{2}{(n-1)(n-2)}\sum_{g,h\in V}\frac{p_{g,h}(a)}{p_{g,h}} \tag{1}$$

Where $p_{g,h}$ is the number of shortest paths between nodes *g* and *h*, and $p_{g,h}(a)$ is the number of shortest paths between nodes *g* and *h* that pass through node *a*.

Flom et al. (2004) extended this measure to a linkage between two sub-networks. Rousseau (2005) introduced this concept into the field of informetrics.

Global and local Q-measures are introduced in (Guns & Rousseau 2009) as complementary measures to gauge the bridging function of nodes in a network with any finite number of subgroups.

The global Q-measure of *a*, denoted as $Q_G(a)$, is defined as formula (2). The local Q-measure of *a*, denoted as $Q_L(a)$, is defined as formula (3).

$$Q_G(a)=\frac{2}{S(S-1)}\sum_{k,l}\left(\frac{1}{TP_{k,l}}\sum_{\substack{g\in G_k\\ h\in G_l}}\frac{p_{g,h}(a)}{p_{g,h}}\right) \tag{2}$$

$$Q_L(a)=\frac{1}{S-1}\sum_{l\neq k}\left(\frac{1}{(m_k-1)\cdot m_l}\sum_{\substack{g\in G_k\\ h\in G_l}}\frac{p_{g,h}(a)}{p_{g,h}}\right) \tag{3}$$

We previously used global and local Q-measures and betweenness to measure how international collaboration happened in the fields of informetrics (Guns & Liu, 2010; Guns, Liu and Mahbuba 2011). We distinguish between a global Q-measure, which characterizes the extent to which node *a* belongs to shortest paths between nodes from different groups, and a local Q-measure, which characterizes the extent to which *a* belongs to shortest path between nodes from its own group and those from other groups. The local Q-measure gauges the bilateral relationship, whereas the global Q-measure gauges the multilateral relationship.

Here we will use global and local Q-measures and betweenness centrality to analyze how these authors are connected through the different categories.
Most Q-measure values are zero, only 172 authors have Q-measures that are larger than 0. All values of global Q-measures are smaller than the value of the same node's local Q-measure. This implies that no author is positioned in a shortest path between nodes of the other sub-networks. It also means that no collaboration happened among three fields or more than three fields. The other 68 authors have a value of betweenness larger than 0 while their Q-measures are zero, implying that these authors just collaborate with authors from the same field.

Table 7. Global and local Q-measures and betweenness

Name	*Classification*	Q_G	Q_L	*Betweenness*
Li ning	TS	0.000995	0.001836	0.002497
Peng Yufa	X	0.000897	0.002343	0.000839
Yang Xiaoguang	TS	0.000772	0.001637	0.002713
Zhao Jing	X	0.000336	0.000462	0.000156
Liu Xiumin	TS	0.000328	0.001171	0.001735
Wang Hongxin	TS	0.000308	0.001014	0.001649
Luo Yunbo	TS	0.000167	0.000519	0.000912
Dong Fengshou	X	0.000167	0.000179	1.54E-05
Yao Jianren	X	0.000167	0.000179	1.54E-05
Huang kunlun	TS	0.000118	0.000458	0.000773
Yang Yuexin	TS	0.000113	0.000117	0.000258
Zhang Wei	TS	0.000108	9.89E-05	0.000666
Zhang Kewei	TS	5.84E-05	0.000134	0.000413
Men Jianhua	R	5.17E-05	0.000414	0.00011
Wang Zhu	R	4.14E-05	0.000331	8.79E-05
Zhang Hongfu	S	4.10E-05	4.63E-05	9.43E-05
Zhu Benzhong	TS	3.13E-05	2.78E-05	0.000147
Piao Jianhua	S	2.91E-05	5.97E-05	7.07E-05
Wang Wei	TS	2.88E-05	0.000231	0.000233
Chen Xiaoping	R	2.30E-05	0.000158	4.21E-05
Jia Shirong	F	2.08E-05	0.000166	9.43E-05

Table 7 shows the first 21 authors that ranked by global Q-measure. These authors are normally positioned in the centrality of the largest component of the network. In figure 2 we mark the first three authors. These three authors are all important members of the Security Council of the National Agricultural genetically modified biology.
Table 8 shows the fields that have the authors whose centrality measures are larger than zero. We can see that only the authors in the fields of Light industry (including Food industry), Agricultural science, Medicine and hygiene, Environmental science, Economics, and Computer science have authors whose Q-measures are larger than zero. The authors in these fields are positioned more

central than the authors from the other fields. The authors from the field of Culture, science (of sciences), education only have two authors whose betweenness are larger than zero, but the Q-measures of all authors in this field are zero. This indicates that the authors in this field do not promote collaboration among the other fields, but two of them promote collaboration within this field.

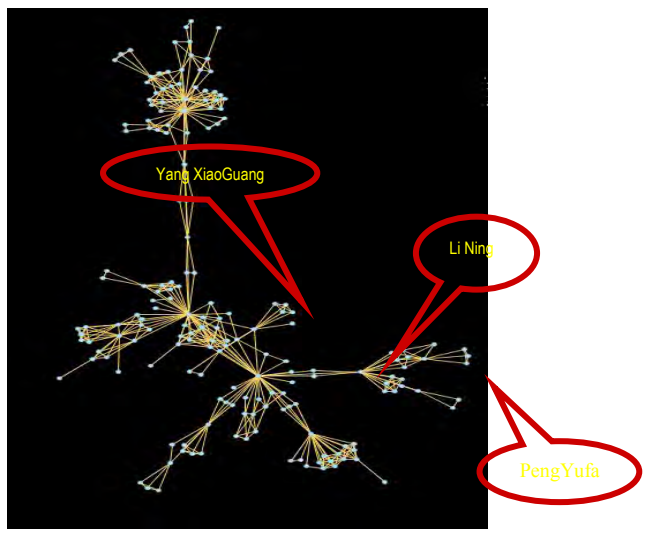


Figure 2: The largest component and the most central authors

GMF is discussed across a broad range of fields. This is witnessed by the fact that the GMF articles are classified by their authors into diverse fields of social science such as Culture, science (of sciences), Education; Politics and law; Philosophy and religion; History and geography; Literature; Military sciences, Arts and Economics. The authors in most fields of social science do not have Q-measures larger than zero. Only authors in Economics are connected to authors from the natural sciences. However, only 5 authors in Economics have Q-measures larger than zero. Even these authors are not ranked among the highest positions according to their Q-measures. The first author from Economics ranks in the 21st position (see Table 7), even though no field contains more articles on GMF than Economics.

Authors from the fields TS (Light industry (including Food industry)), S (Agricultural Science), R (Medicine and hygiene), X (Environmental science) obtain the top ranks. Furthermore, the number of authors whose Q-measures are not zero are larger in these fields than in the other fields. Hence, the authors from these natural science fields act as brokers between different disciplines pertaining to GMF research.

Why are authors from the natural sciences more likely to form interdisciplinary bridges in research collaboration on the topic of GMF? GM technology involves many aspects. Some knowledge might be beyond the expertise of a specific expert. Hence, these experts have to collaborate with each other when they meet a problem that they cannot solve by their own knowledge or expertise. Moreover, experts on GM technology often branch out a company which also promotes collaboration with the others.

Table 8. Number of authors per field with non-zero Q-measures and betweenness centrality

Code	*Field*	*Q-measure*	*Betweenness*
TS	Light industry	99	121
S	Agricultural science	37	57
R	Medicine and hygiene	12	13
X	Environmental sciences	8	12
F	Economics	5	21
TP	Computer science	3	3
G	Culture, science (of sciences), education	0	2

Conclusion

In order to understand how Genetically Modified Food (GMF) research is developing in China under the interaction between authors from social sciences and natural sciences, we have analyzed the distribution of Chinese articles on GMF among different fields and studied how the authors of these articles collaborate with each other. The problem was studied from a network perspective, focusing on the co-authorship network of research on GMF.

Q-measures were used to characterize collaboration among different scientific fields. Only 172 authors have Q-measures that are larger than 0. All global Q-measure values are smaller than the same node's local Q-measure. This implies that no author is part of a shortest path between two authors from different fields. It also means that no collaboration took place among three or more fields. Authors from most fields in the social sciences and humanities (SSH) occupy a peripheral position in the network. These SSH fields include Culture, science (of sciences), education; Politics and law; Philosophy and religion; History and geography; Literature; Military sciences; and Arts. Compared to these SSH fields, authors from Economics are more central and bridge between authors from the natural sciences. However, although more articles in the dataset are from Economics than any other fields, only few of their authors have Q-measures larger than zero, and the Q-measures of these authors are relatively small. In this sense SSH authors play a limited role in interdisciplinary collaboration among authors engaged in GMF research. The fields that ranked in the top by Q-measure values are the fields of Light industry (including Food industry), Agricultural Science, Medicine and hygiene and Environmental science. The authors from these natural science fields act as brokers between different disciplines pertaining to GMF research.

Acknowledgments

We thank Ronald Rousseau for his inspirational discussions when we began this work. This work is supported by the National Natural Science Foundation of China (NSFC grant No. 71173154)

References

Freeman, L. C. (1977). A set of measures of centrality based upon betweenness. *Sociometry*, 40, 35–41.

Flom, P. L., Friedman, S. R., Strauss, S., & Neaigus, A. (2004). A new measure of linkage between two subnetworks. *Connections*, 26(1), 62–70.

Genetically modified food Retrieved January 22, 2013 from: http://en.wikipedia.org/wiki/Genetically_modified_food

Guns, R., & Liu, Y. X. (2010). Scientometric research in China in the context of international collaboration. *Geomatics and Information Science of Wuhan University*, 35, 112–115. (ICSUE 2010 special).

Guns, R. & Rousseau, R. (2009). Gauging the bridging function of nodes in a network: Q-measures for networks with a finite number of subgroups. In B. Larsen & J. Leta (Eds.), Proceedings of ISSI 2009—the 12th international conference on scientometrics and informetrics (pp. 131–142).

Guns, R., Liu, YX. and Mahbuba D. (2011). Q-measures and betweenness centrality in a collaboration network: a case study of informetrics. *Scientometrics*, 87,133-147.

Liu, D., (2012). Genetic modified technology will not cease to develop, *Chinese Science newspaper*, 2012,1,17. Retrieved December 21, 2001 from: http://news.sciencenet.cn/sbhtmlnews/2012/1/253522.shtm

Liu, YX., ROUSSEAU, R. (2007), Hirsch-type indices and library management: the case of Tongji University Library. In: D. TORRES-SALINAS, H. F. MOED (Eds), *Proceedings of ISSI 2007*. Madrid: CINDOC-CSIC, pp. 514–22.

Ministry of Agriculture of PRC (2009). The second approved list of security certificate sfor agriculturally genetically modified biology in 2009. [in Chinese] http://www.stee.agri.gov.cn/biosafety/spxx/P020091127591594596689.pdf

The PRC State Council. (2006). *Outline of the National Plan for the Development of Science and Technology (2006-2020).* Beijing:People's publishing house.

Rousseau, R. (2005). Q-measures for binary divided networks: An investigation within the field of informetrics.In *Proceedings of the 68th ASIST conference* (Vol. 42, pp. 675–696)

A GLOBAL OVERVIEW OF COMPLEX NETWORKS RESEARCH ACTIVITIES

Fei-cheng Ma[68], Peng-hui Lyu[1] and Xiang Liu[2]

[1] *lvph@whu.edu.cn*
Wuhan University, Center for the Studies of Information Resources, Wuhan 430072
(People's Republic of China)

[2] *xiangliu@whu.edu.cn*
Huazhong Normal University, School of Information Management, Wuhan 430079
(People's Republic of China)

Abstract
Complex network research publications have increased rapidly over last decade, most notably in the past four years. This paper attempts to visualise the research outputs of complex network research in a global context for the purpose of evaluating the world research progress and quantitative assessment of current research trends. The scientometric methods and knowledge visualization technologies were employed with a focus on global production, main subject categories, core journals, top productive countries, leading research institutes, publications' most used keywords and the papers with top citations. The keywords cluster analysis was used to trace the hot topics from all papers, which is also the hot point in scientometrics and informetrics researches.
Research output descriptors have suggested that the research in this field has mainly focused on dynamics, model and systems for complex networks. All the publications have been concentrated in two journals such as *Physical Review E* and *Physica A*. The USA is the leading country in complex network research and the world research centre is located there and it has the best scientists in the world. The research trend in complex network research seems to involve complex routing strategy, models complex networks social as well as scale free percolation efficiency. Complex networks, dynamics, model and small-world networks are highly used keywords in the publications from the main scientific database.

Introduction
Complex networks, as an effective reflection contacting the real world with theoretical research initially attracted the evolved attention slowly from the impact of chaos theory and fractal studies on a small set of computer scientists, biologists, mathematicians and physicists, are thoroughly studied in many fields now. Two pioneering works, small world network and scale-free network, encouraged a wave of international research concerning complex networks by the end of the 20th century. Small-world networks explored by Watts and Strogatz, which can be highly clustered and have small characteristic path lengths (Watts and Strogatz 1998), can portray biological, technological and social networks

To whom correspondence should be addressed to *ph@nimte.ac.cn*

better than the networks completely regular or completely random. In many large networks it was found that the property that the vertex connectivity followed a scale-free power-law distribution (Barabasi and Albert 1999) by Barabási A.L and Albert R. Counting from this emergence, complex networks have gone through its first decade.

In the early 21th century, the discovery of small world effect and scale-free property in the real network largely provoked the publications boom of complex networks. Initial research on complex networks focused on the analysis and modelling of network structure at large, such as degree exponents (Dorogovtsev, Goltsev et al. 2002), dynamical processes (Yang, Zhou et al. 2008), network growth (Gagen and Mattick 2005), link prediction (Zhou, Lu et al. 2009) and so on. Then Strogatz S.H tried to unravel the structure and dynamics of complex networks from the perspective of nonlinear dynamics (Strogatz 2001). The statistical mechanics of network as topology and dynamics of the main models as well as analytical tools were discussed (Albert and Barabasi 2002), the theory of evolving networks was introduced in Albert R and Barabasi A.L's work.

The developments of complex networks, including several major concepts, models of network growth, as well as dynamical processes (Newman 2003) were discussed in Newman MEJ's paper. The basic concepts as well as the results achieved in the study of the structure and dynamics of complex networks (Boccaletti, Latora et al. 2006) were summarized. The error tolerance was displayed only in scale-free networks, and it showed an unexpected degree of robustness (Albert, Jeong et al. 2000). Network motifs and patterns of interconnections to uncover the structural design principles of complex networks was defined (Milo, Shen-Orr et al. 2002). The way in which self-organized networks grows into scale-free structures, and the role of the mechanism of preferential linking were investigated (Dorogovtsev and Mendes 2002). A number of models demonstrating the main features of evolving networks were also presented. Mixing patterns in a variety of networks were measured (Newman 2003) and technological as well as biological networks were found disproportionally mixed, while social networks tend to be assorted. It was pointed out that scale-free networks catalysed the emergence of network science (Barabasi and Oltvai 2004). The number of driver nodes is determined primarily by the network's degree distribution was also found, and the driver nodes tend to avoid the high-degree nodes (Liu, Slotine et al. 2011). The control of degrees on complex networks was carefully studied later (Egerstedt 2011). The fragility of interdependency on complex networks was also studied hence (Vespignani 2010).

With the continuous development of complex networks, in addition to the theoretical and technical research on the complex network itself, scholars have also focused on the network function. Barabasi A.L and Oltvai Z.N indicated that cellular networks offer a new conceptual framework for biology and disease pathologies (Barabasi 2009), which could potentially revolutionize the traditional view. An approach which not only stresses the systemic complexity of economic networks was pointed out (Schweitzer, Fagiolo et al. 2009), it can be used to

revise and extend traditional paradigms in economic theory which is urgently needed. A biologically complex multistring network model was designed to observe the evolution and transmission dynamics of ARV resistance (Smith, Okano et al. 2010).

The current situation is that the complex network research was not only limited to the study of the theory and methods, but has become a new research direction of multi-disciplinary and a powerful tool in multi-disciplinary research. Nowadays complex network have been applied in many different areas including spread (Yang, Zhou et al. 2008), network synchronization (Motter, Zhou et al. 2005), transports (Wang, Wang et al. 2006), game theory (Perc and Szolnoki 2010), physics (Newman 2002), computer science (Guimera and Amaral 2005), biochemistry or molecular biology (Jeong, Tombor et al. 2000), mathematics (Guimera and Amaral 2005), engineering (Olfati-Saber, Fax et al. 2007), cell biology (Rosen and MacDougald 2006). These research directions took us more and more productions and publications in recent years.

Most important it was known to all that the methods of complex networks are used more and more for scientomtrics and informetrics research in information science. For example the complex networks analysis was employed for co-citation or co-occurrence network to get the knowledge structure as well as scientific cooperation performance for a specific filed. While in these studies, the metric data is the base of all complex networks analysis. Traditional bibliometrics research was widely applied to acquaint information from the scientific or technical literatures, and for further study the complex networks method could also help.

In this study the records of literature were analysed with scientometric methods via several aspects. This effort will provide a current view of the mainstream research on complex networks as well as clues to the impact of this hot topic. In addition, this study also attempted to analyse the significance of the complex networks production patterns, especially in the way of co-authors and authors' keywords study originally acted from WoS database. The main body of this article includes scientometric analyses in production, subject category, and geographical distribution of WoS data. Moreover, appropriate statistical tests were used in the authors' keyword yearly to predict the developing trend of complex networks research.

Data and Method

This study is based on the metadata analysis of the articles from the authoritative scientific and technical literature indexing databases such as SCIE, SSCI and CPCI. The impact factor of SCI & SSCI journals with the latest data available in 2011 was determined by Journal Citation Reports (JCR) of Thomson Reuters, which was operated by Thomson Scientific, Philadelphia, PA now (Proudfoot, McAuley et al. 2011). The statistical analysis tool is Thomson Data Analyser (TDA) and the drawing tool is Aureka and MS Office Excel 2010.

Date Retrieved

The metadata source came from WoS Web offered by Thomson Reuters, and the publishing time span was last updated in Dec 31st, 2012. Data in this study was required on Jan 3nd, 2013 using the topic= "complex network*" selecting "all the years" within the metadata including publication's title, keywords and abstract. In total, 10,707 articles were retrieved from the database of Web of Science (WoS). Precision retrieval strategy used in this paper make the ability of the search term to minimize the number of irrelevant records retrieved. Due to WoS is an abstract database, only the metadata can be extracted from it, certainly if given the chance to extract information from the full text of all paper the results may be more accurate.

Analysis Methods

In this scientometrics study, the annual numbers, top subjects, core journals, productive countries, fruitful institutes, main authors and funding agencies of the papers was deeply studied using the methods of quantitative analysis. In this study comparative analysis was also used to analyse the data by putting the SCI and SSCI data into the same figure so that a direct and vivid result can be gotten from the figures and as much as possible information obtained.

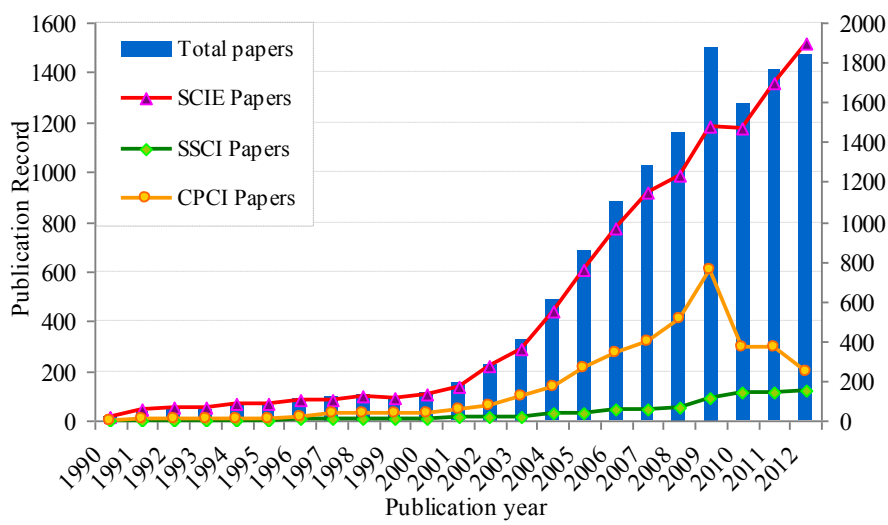


Figure 1. Papers record indexed in WoS from 1990 to 2012

Result Analysis

In this section, figures and tables are used to describe the production and the development trends of complex networks research in both science and social science fields. Publications (as indicator for scientific performance) are commonly accepted indicators for quantitative analysis on innovation research

performance (Garfield 1970). Papers from SCI as well as SSCI were studied together in this paper with scientometric analysis.

Productivity Analysis

As seen in Figure 1 the complex networks publications dramatically increased in the last two decades. From 1990 to 2001, the complex networks' research were just begun and its publications were relatively low and there were not more than 200 in the WoS database. After 2001, the outputs increased rapidly from less than 200 in 2001 to more than 2000 in 2009 and then stabilized changed recently. The complex networks research came into its fast growth stage in 21st century and may enter the mature period of its publish cycle in the next decade.

Subjects Analysis

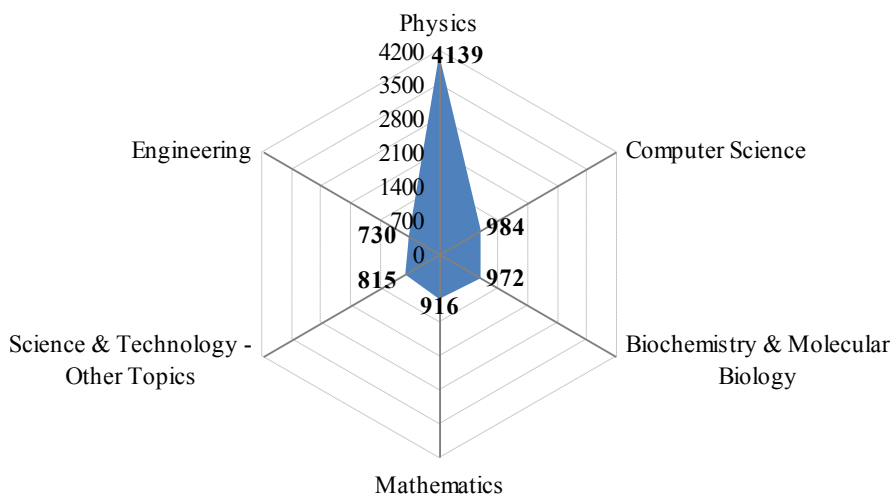


Figure 2. Subjects distribution of SCI&SSCI papers

The complex networks related research was distributed in the subjects of physics, computer science, biochemistry & molecular biology, mathematics and engineering, which is shown in Figure 2. Most complex networks outputs were produced under the subject of physics due to it is a branch of theoretical physics originally. As time went by, this approach was used in bioscience or engineering to solve many problems as a migrating concept, which proved its superiority for many disciplines from the metadata of SCI&SSCI papers.

Journals Analysis

See from Figure 3 the complex networks research was published mainly in physics related journals such as *Physical Review E* and *Physica A,* which published most complex networks papers in all journals from the SCI&SSCI

database. *PLoS One* produced 217 papers and ranked third, *European physical journal B* with 205 papers in forth and *Chaos* with 205 papers in fifth for publications in the complex networks research. The American journals *PNAS* and *Physical Review Letters* were two journals with the highest impact factor in 2011.

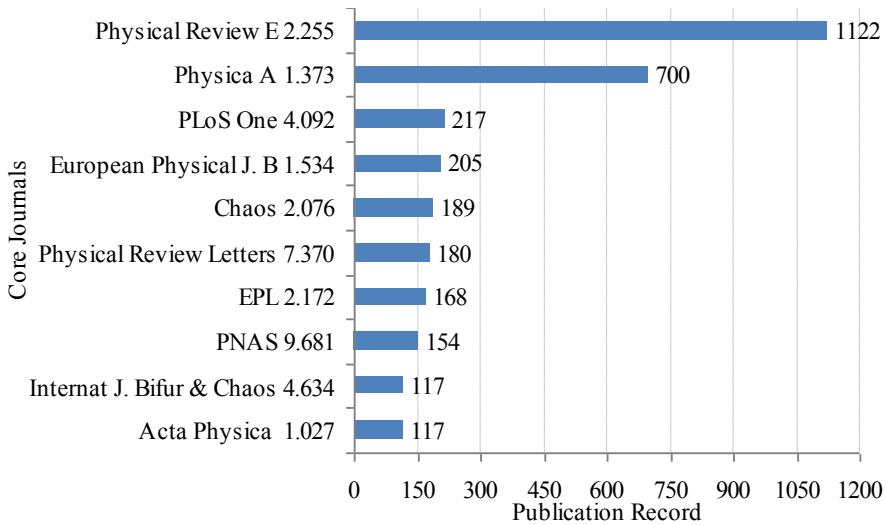


Figure 3. Top productive SCI&SSCI journals with its IF in 2011

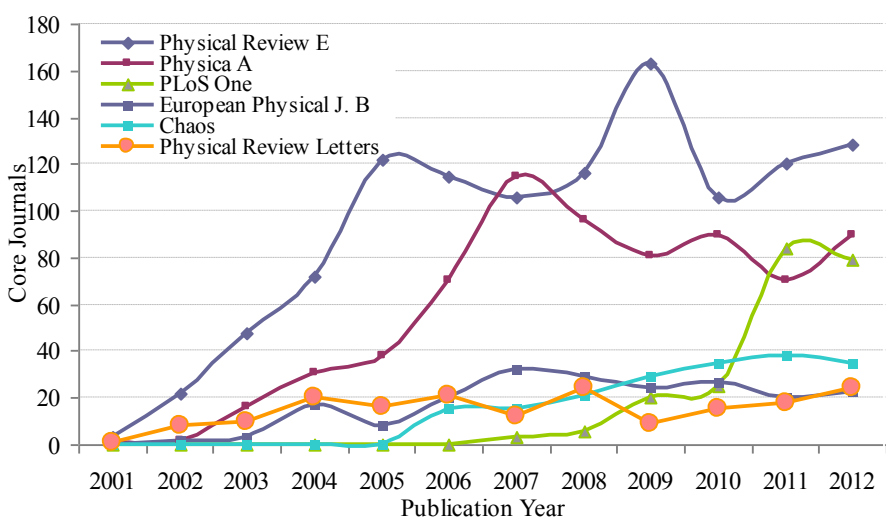


Figure 4. Annual SCI&SSCI journals outputs distribution during 2001-2012

The annual publications distribution about complex networks papers are shown in Figure 4. The research in this field attracted the most attentions from scientists far in the year 2001. *Physical Review E* was the main publisher of complex networks in the last decade, while *Physica A* reached the publication level of *Physical Review E* in 2007 once. Other journals kept a stable publication state in the past decade with about 30 papers per year in SCI&SSCI database; *PLoS One* (Full name of *Public Library of Science One*) was the only exception with a dramatically increasing rate in recent three years.

Countries Analysis

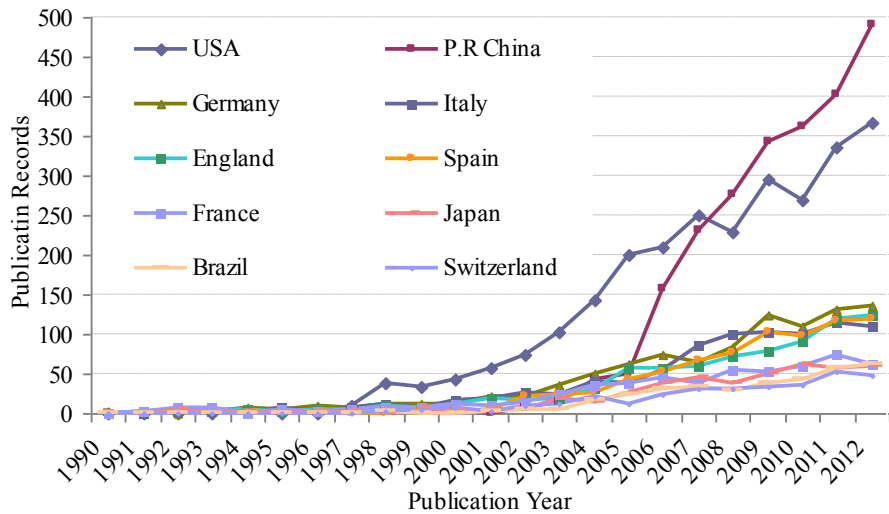


Figure 5. Annual country record distribution in the complex networks study

In all complex networks publications, the United States of America and the People's Republic of China contributed the most parts as shown in Figure 5. Hence the research centre was located in these two countries at present. However, the USA started complex networks research as early in 1997 but dropped behind P.R. China in productions after 2007. Other countries such as Germany, Italy, England and Spain produced less outputs with a stable increasing rate in complex networks related publications. While in total these European countries published more papers than former other countries.

Active degree is defined as the outputs number in recent three years to all years' publication number in general bibliometric research. P.R. China had the highest active degree of 52.3% in all countries in the world, indicating that the research of complex networks was treasured much and in fact such activity as the Conference for Chinese Complex Networks (short for CCCN) was held for eight times already in recent years in P.R. China.

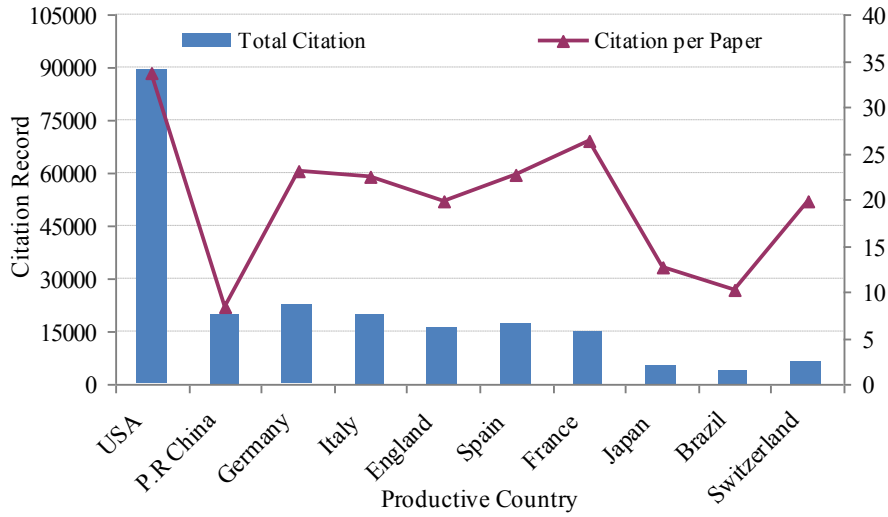


Figure 6. Citation distribution of global SCI&SSCI papers

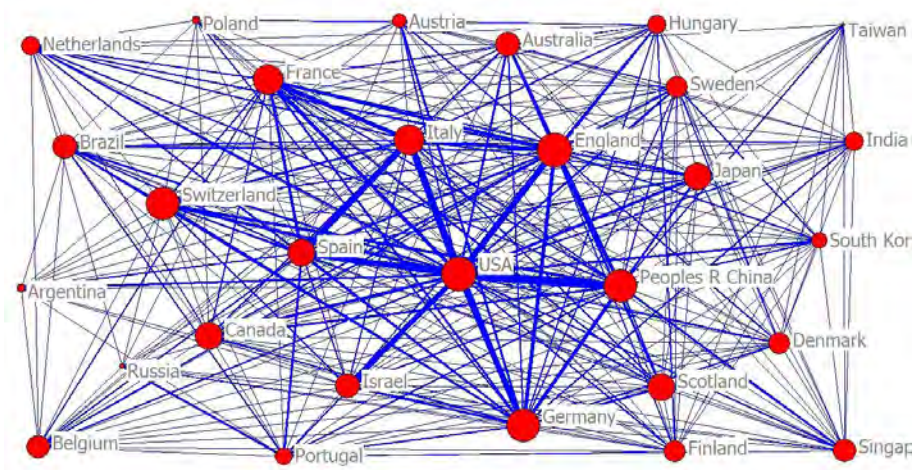


Figure 7. The international collaboration network of complex networks

The SCI&SSCI papers' citations results were shown in Figure 6. From this figure it can be found that the USA obtained the most citations, which attested its high level in the field of complex network research. The highest citation per paper was from European countries such as England and Germany. P.R. China's average citation was relatively lower than most European countries and Brazil or Japan, but not far behind the USA with less total citations.

In the international collaboration of papers of complex networks, the USA, Germany and P.R. China are located in the central positions which can be seen in

Figure 7. It is also clear that USA is in the centre of collaborating activities. Other countries such as England, France, Italy, Spain and Switzerland had less cooperation in complex networks research in SCI&SSCI publications. The cooperation network between top productive countries reflected the knowledge transmission in the field of complex networks research in the world.

Institutes Analysis

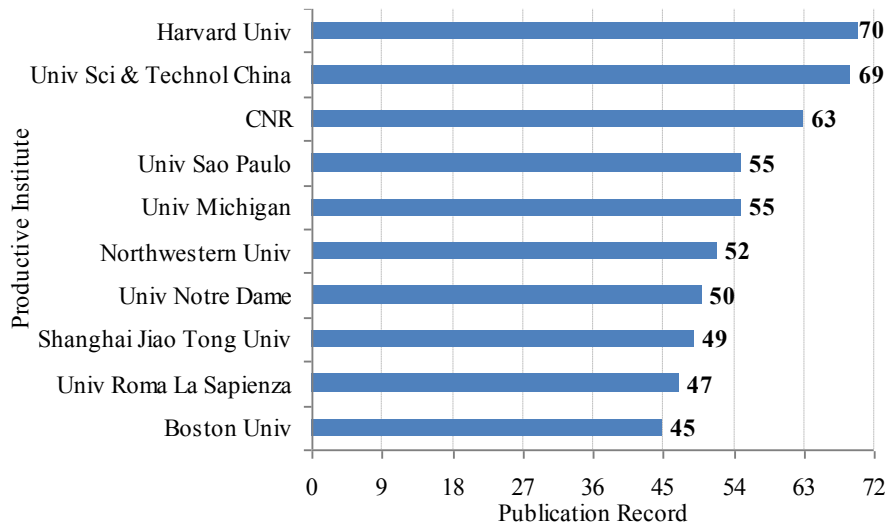


Figure 8. Top productive institutes of SCI&SSCI papers published

The top productive institutes with an accumulative paper quantity of more than 40 are ranked in Figure 8. The Harvard University published 70 papers in total, ranking first, followed by University of Science and Technology of China (USTC, 69) and CNR from France with an output of 63 papers. Other institutes produced many papers as the former ones did in complex networks, which reflecting their overall strength like these American and Chinese agencies.

For most productive institutes, the time span during 2005 to 2007 was the best years with most publications as shown in Figure 9. The Harvard University from USA as well as University of Science and Technology of China produced the most complex networks research papers accumulatively before, far more than all the institutes in world organizations. After 2008, all top productive institutes published less paper for about five years while Boston University not. The production came into a former maturity stage in its publish cycle then.

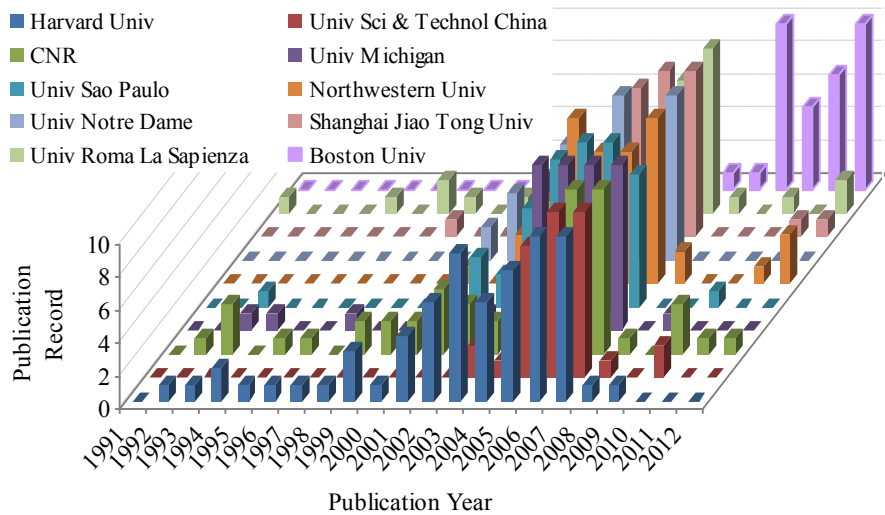


Figure 9. Annual distribution of SCI&SSCI papers produced by top institutes

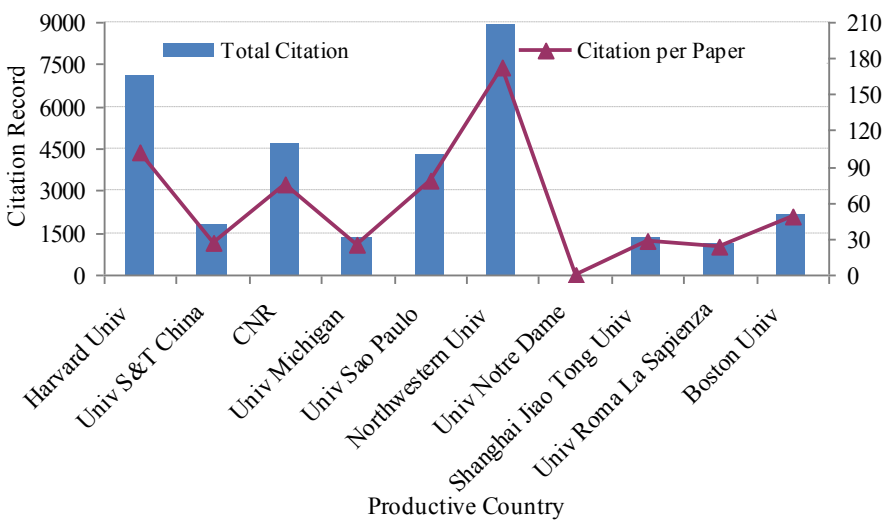


Figure 10. Citation of top productive institutions in SCI&SSCI papers published

It can be seen from Figure 10 that the Northwestern University has the most total citations as well as citations per paper in the world, which proved their priority in complex networks research. The Harvard University had the second most total citations and citations per paper. Compared with the University of Science and Technology of China, North western University and Harvard University has the

highest citations per paper and most total citations. US agencies have the best research in complex networks research in the world.

Keywords Analysis

All the high frequency keywords plus more than 200 are listed in 0. The complex networks related hotspots were mainly distributed in the dynamics, model and systems research as we can see from Table 1.What's more, the research of Small World networks, the internet and evolution were also the high frequency key words that emerged in research papers. Scale-Free Networks as well as organizations became the hot words only less than words plus listed above.

Table 1. Keywords plus distributions of SCI&SSCI Publications

Keywords Plus	Complex Networks	Dynamics	Model	Systems	Small-World Networks
Records	4004	1428	823	785	680
Keywords Plus	Internet	Evolution	Organization	Scale-Free Networks	Synchronization
Records	507	481	453	409	341
Keywords Plus	Topology	Stability	Expression	Community Structure	Gene-Expression
Records	331	324	303	285	283
Keywords Plus	Graphs	Metabolic Networks	Web	Escherichia-Coli	Models
Records	283	257	249	236	234

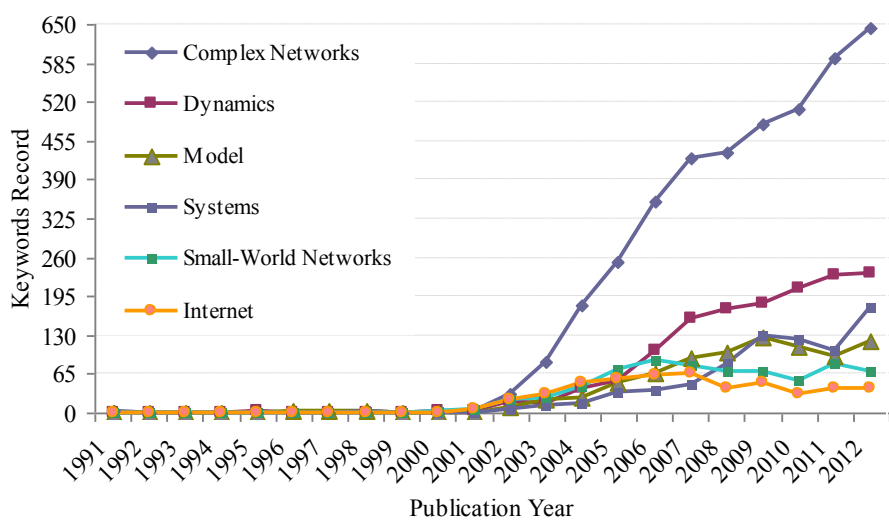


Figure 11. Annual number of keywords plus of SCI&SSCI papers published

The annual keywords plus distribution was drawn in Figure 11. The main retrieval word of complex networks was turned up in 2001 and with a fast increasing trend in the past decade. Scholars paid little attentions in the dynamics research before 2000, while they were interested in it during 2001 to 2007 so that the number of this word increased sharply from then. In the year 2006, almost all the research of systems, internet and Small-World Networks maintained a fast increase in papers production.

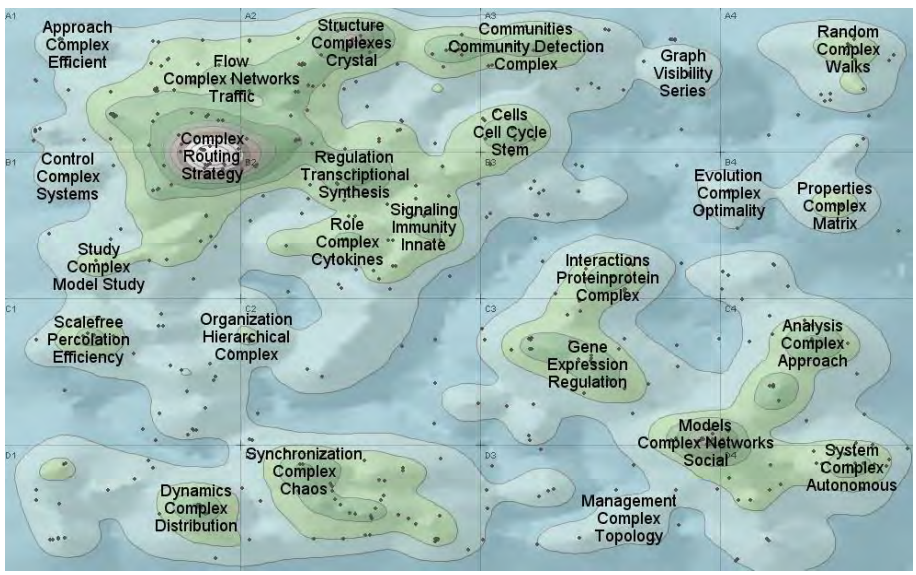


Figure 12. Cluster and co-words map of SCI&SSCI papers published

The complex networks research co-words map was drawn for the hotspot analysis in this paper as Figure 12. All the words were extracted from the title, author keywords and abstracts of the publications automatically by the Aureka software and then clustered in the knowledge map to trace the research trends. The complex routing strategy is the most popular research domain in all outputs for complex networks research. And such fields as models complex networks social as well as scale free percolation efficiency were less popular than the complex routing strategy research in recent years.

Citation Analysis

The most frequently cited papers are key literatures link the research of complex networks for years so the top SCI&SSCI papers with most citations are listed in Table 2. In the top 10 high influence research papers, six of which came from the USA and the remainders were produced by European countries. The paper "Statistical mechanics of complex networks" written by Albert R and Barabasi A.L from Notre Dame University was the most frequently cited paper in the

world. These two famous scientists also wrote other three most cited papers under the topics of error and attack tolerance, network biology and metabolic networks before.

Table 2Top 10 SCI&SSCI papers with Most Citations.

No	*Time Cited*	*Authors*	*Title*	*Journal*	*Institute*	*Country*	*Year*
1	5775	Albert, R; Barabasi, AL	Statistical mechanics of complex networks	Reviews of Modern Physics	Univ Notre Dame	USA	2002
2	4117	Newman, MEJ	The structure and function of complex networks	Siam Review	Univ Michigan	USA	2003
3	2390	Strogatz, SH	Exploring complex networks	Nature	Cornell Univ	USA	2001
4	2048	Boccaletti, S; Latora, V; Moreno, Y; Chavez, M; Hwang, DU	complex networks: Structure and dynamics	Physics Reports-Review Section of Physics Letters	CNR	Italy	2006
5	1999	Albert, R; Jeong, H; Barabasi, AL	Error and attack tolerance of complex networks	Nature	Univ Notre Dame	USA	2000
6	1990	Barabasi, AL; Oltvai, ZN	Network biology: Understanding the cell's functional organization	Nature Reviews Genetics	Univ Notre Dame	USA	2004
7	1948	Jeong, H; Tombor, B; Albert, R; Oltvai, ZN; Barabasi, AL	The large-scale organization of metabolic networks	Nature	Univ Notre Dame	USA	2000
8	1660	Milo, R ; Shen-Orr, S; Itzkovitz, S; Kashtan, N; Chklovskii, D; Alon, U	Network motifs: Simple building blocks of complex networks	Science	Weizmann Inst Sci	Israel	2002
9	1315	Dorogovtsev, SN; Mendes, JFF	Evolution of networks	Advances In Physics	Univ Porto	Portugal	2002
10	1039	Thiery, JP; Sleeman, JP	complex networks orchestrate epithelial-mesenchymal transitions	Nature Reviews Molecular Cell Biology	CNRS	France	2006

Conclusion

As a strictly selected academic thesis abstract database, Web of Science (WoS, including SCI and SSCI) has been long recognized as the useful tool that can cover the most important science & technology, social science research productivities. SCI & SSCI citation search systems are unique and significant, not only from the perspective of literature cited but also from the academic

assessment of the value in articles or from cooperation networks to research references. So all the papers published in this database were carefully studied to get the publication pattern and production orderliness.

The methods of complex networks are used more and more in other research such as in information science. The complex networks analysis was employed for co-citation or co-occurrence network to get the knowledge structure as well as scientific cooperation performance for a specific scientific filed often. Hence traditional scientometrics research was widely applied to acquaint information from the scientific or technical literatures, this will lead us a new direction for complex networks method in future as this research do.

Hence in this study, the impact of global complex networks literature has been studied with scientometric methods and the research history has been recalled firstly according to the complex networks research literatures. The publications history started from 1990 and boosted in recent four or five years. From 1990 to about 2001, the complex networks research stepped into its infancy stage and then began a fast increasing stage in growth, and now in the former stage of maturity in its life cycle. In near future the publications in this field will still keep going larger and larger for quite a long time as can be predicted.

Complex networks research are mainly in the subjects of physics. All the output concentrated in two journals such as *Physical Review E* and *Physica A* in SCI&SSCI database. The research papers were mainly completed by several authors according to network theory aggregation nodes in a power law correlation, and the multiple-authors made up an increasingly larger ratio to form a group size measured using papers. So the co-authored papers in the complex networks research were the mainstream of complex networks research and it formed a complex collaboration networks about complex networks research.

Complex networks related papers were distributed unevenly over all countries. The USA, China and Germany were the top productive countries of SCI&SSCI papers. Some Europe countries such as Italy and Germany published top influence paper than those productive countries. The complex networks research centre was located in the USA in the last few decades according to the metadata from countries and institutes analysis. Harvard University and USTC produced most SCI papers and some USA institutes such as University of Michigan and University of Notre Dame contributed most influence SCI articles.

Research on the fields of complex networks research focused on complex routing strategy, models complex networks social as well as scale free percolation efficiency. From the analysis of author keywords, except “complex networks”, “dynamics”, “model” and “small-world networks” were highly used key words plus in the scientific database. It is clear that complex networks research will be a hot spot in the complexity science field in the future. With scientometric and informetric method, the findings of this study can help scientific researchers understand the performance and central trends of complex networks research in the world, and therefore suggest directions for further research.

Acknowledgments

This work is supported by the Natural Science Foundation of China (Grant No. 71173249: Research on Formation Mechanism and Evolution Laws of Knowledge Networks). The authors are grateful to Min-xuan Liu, Ya-ting Li, Bai-tong Chen and Gerard Joseph White for their helpful discussions and suggestions. And our special thanks also dedicate to those anonymous reviewers for their valuable comments for this research work.

References

Albert, R. and A. L. Barabasi (2002). "Statistical mechanics of complex networks." Reviews of Modern Physics 74(1): 47-97.

Albert, R., H. Jeong, et al. (2000). "Error and attack tolerance of complex networks." Nature 406(6794): 378-382.

Barabasi, A. L. (2009). "Scale-Free Networks: A Decade and Beyond." Science 325(5939): 412-413.

Barabasi, A. L. and R. Albert (1999). "Emergence of scaling in random networks." Science 286(5439): 509-512.

Barabasi, A. L. and Z. N. Oltvai (2004). "Network biology: Understanding the cell's functional organization." Nature Reviews Genetics 5(2): 101-U115.

Boccaletti, S., V. Latora, et al. (2006). "complex networks: Structure and dynamics." Physics Reports-Review Section of Physics Letters 424(4-5): 175-308.

Dorogovtsev, S. N., A. V. Goltsev, et al. (2002). "Pseudofractal scale-free web." Physical Review E 65(6).

Dorogovtsev, S. N. and J. F. F. Mendes (2002). "Evolution of networks." Advances in Physics 51(4): 1079-1187.

Egerstedt, M. (2011). "Complex Networks Degrees of control." Nature 473(7346): 158-159.

Gagen, M. J. and J. S. Mattick (2005). "Accelerating, hyperaccelerating, and decelerating networks." Physical Review E 72(1).

Garfield, E. (1970). "Citation indexing for studying science." Nature 227(5259): 669-671.

Guimera, R. and L. A. N. Amaral (2005). "Functional cartography of complex metabolic networks." Nature 433(7028): 895-900.

Jeong, H., B. Tombor, et al. (2000). "The large-scale organization of metabolic networks." Nature 407(6804): 651-654.

Liu, Y. Y., J. J. Slotine, et al. (2011). "Controllability of complex networks." Nature 473(7346): 167-173.

Milo, R., S. Shen-Orr, et al. (2002). "Network motifs: Simple building blocks of complex networks." Science 298(5594): 824-827.

Motter, A. E., C. S. Zhou, et al. (2005). "Network synchronization, diffusion, and the paradox of heterogeneity." Physical Review E 71(1).

Newman, M. E. J. (2002). "Assortative mixing in networks." Physical Review Letters 89(20).

Newman, M. E. J. (2003). "The structure and function of complex networks." Siam Review 45(2): 167-256.
Olfati-Saber, R., J. A. Fax, et al. (2007). "Consensus and cooperation in networked multi-agent systems." Proceedings of the Ieee 95(1): 215-233.
Perc, M. and A. Szolnoki (2010). "Coevolutionary games-A mini review." Biosystems 99(2): 109-125.
Proudfoot, A. G., D. F. McAuley, et al. (2011). "Translational research: what does it mean, what has it delivered and what might it deliver?" Current Opinion in Critical Care 17(5): 495-503.
Rosen, E. D. and O. A. MacDougald (2006). "Adipocyte differentiation from the inside out." Nature Reviews Molecular Cell Biology 7(12): 885-896.
Schweitzer, F., G. Fagiolo, et al. (2009). "Economic Networks: The New Challenges." Science 325(5939): 422-425.
Smith, R. J., J. T. Okano, et al. (2010). "Evolutionary Dynamics of Complex Networks of HIV Drug-Resistant Strains: The Case of San Francisco." Science 327(5966): 697-701.
Strogatz, S. H. (2001). "Exploring complex networks." Nature 410(6825): 268-276.
Vespignani, A. (2010). "Complex Networks The fragility of interdependency." Nature 464(7291): 984-985.
Wang, W. X., B. H. Wang, et al. (2006). "Traffic dynamics based on local routing protocol on a scale-free network." Physical Review E 73(2).
Watts, D. J. and S. H. Strogatz (1998). "Collective dynamics of 'small-world' networks." Nature 393(6684): 440-442.
Yang, R., T. Zhou, et al. (2008). "Optimal contact process on complex networks." Physical Review E 78(6).
Zhou, T., L. Lu, et al. (2009). "Predicting missing links via local information." European Physical Journal B 71(4): 623-630.

HOW ARE COLLABORATION AND PRODUCTIVITY CORRELATED AT VARIOUS CAREER STAGES OF SCIENTISTS?

Zhigang Hu[1] Chaomei Chen[2] and Zeyuan Liu[3]

[1] *huzhigang@mail.dlut.edu.cn*
WISE Lab, Dalian University of Technology, Dalian (China)
Joint-Institute for the Study of Knowledge Visualization and Scientific Discovery, Dalian University of Technology(China)-Drexel University(USA), Dalian(China) and Philadelphia(USA)

[2] *cc345@drexel.edu*
The iSchool, Drexel University, Philadelphia (USA)
Joint-Institute for the Study of Knowledge Visualization and Scientific Discovery, Dalian University of Technology(China)-Drexel University(USA), Dalian(China) and Philadelphia(USA)

[3] *liuzy@vip.163.com*
WISE Lab, Dalian University of Technology, Dalian (China)
Joint-Institute for the Study of Knowledge Visualization and Scientific Discovery, Dalian University of Technology(China)-Drexel University(USA), Dalian(China) and Philadelphia(USA)

Abstract
Collaboration is believed to be influential on researchers' productivity. However, the impact of collaboration relies on latent factors such as disciplines, collaboration patterns, and collaborators' characters. Moreover, at different career stages, such as the novice stage, the experienced stage, etc., collaboration is different in scale and breadth, and its effect on productivity varies. In this paper, we study collaborative relationships in four disciplines, Organic Chemistry, Virology, Mathematics and Computer Science. We find that the productivity is correlated with collaboration in general, but the correlation could be positive or negative on the basis of which aspect of collaboration to measure, the collaboration scale or scope. The correlation becomes stronger as individual scientists progress through various stages of their career. Furthermore, experimental disciplines, such as Organic Chemistry and Virology, have shown stronger correlation coefficients than theoretical ones such as Mathematics and Computer Science.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6) and Scientometrics Indicators (Topic 1).

Introduction

In recent decades, collaborative research has been considered as essential for scientific research. For individual researchers, scientific collaboration is desirable because there are many perceived benefits if they collaborate. Some of the most common motivations for collaboration include accessing to expertise or unavailable resources (Link, Paton, & Donald, 2002), learning tacit knowledge (Jansen, Görtz, & Heidler, 2009), and developing Science and technology human capital (Bozeman & E. Corley, 2004). After all, enhancing the productivity is one of the primary objectives for collaboration (Landry, 1998).

Scientific collaboration has been studied across a broad range of collaborative activities, including co-authorship in publications, joint patent applications, joint grant proposals, to name a few. In this study, we focus on collaboration in terms of co-authorship in publications. Productivity has been quantified in the literature in terms of the number of publications, the number of patents, the number of doctoral students graduated, and other indicators. We choose to measure the productivity of a scientist in terms of the number of publications. Although many previous studies reported a positive correlation between collaboration and productivity (Beaver, 1979; E. a Corley & Sabharwal, 2010; Ding, Levin, Stephan, & Winkler, 2010; Duque, 2005; Hsu & Huang, 2010; Katz & Martin, 1997; Lee & Bozeman, 2005), several papers reported that the relationship was not found to be statistically significant, and in some cases, collaboration even has a negative impact on productivity. The contradictory findings about collaboration are in part due to the diversity of collaboration patterns and the complexity of the correlation between collaboration and productivity. On the one hand, the effects of collaboration depend on academic environment (Birnholtz, 2007), discipline features (Porac, 2004; Yoshikane, Nozawa, & Tsuji, 2006), specific type of collaborations (Perkmann & Walsh, 2009), and individual scientist characteristics (Birnholtz, 2007). On the other hand, this correlation may vary at different stages of a scientist's academic career, followed by the change of the motives for collaboration. The stages of academic career are different phases and status of a scientist. At the beginning stage of one's academic career, one may collaborate for accessing equipment or other research resource; while at a later stage, one collaborates with others as mentors, or he/she is just too busy to carry out research without collaboration.

However, these hypotheses need to be investigated to clarify the current findings in the literature. There is a lack of quantitative studies that specifically address how scientists collaborate at various stages of their career. So in this article, we aim to examine the effects of the collaboration on scientists' productivity over a 30-year career path in four disciplines. We focus on two main research questions : 1) Is there positive correlation between collaboration and productivity? 2) What are the role of career stage in relation to collaboration and productivity?

Related Work

Many researchers have argued that collaboration has a positive impact on researchers' productivity and performance (E. a Corley & Sabharwal, 2010; Defazio, Lockett, & Wright, 2009; He, Geng, & Campbell-Hunt, 2009; Lee & Bozeman, 2005). The study of Beaver & Rosen was among the earliest empirical researches about the effects of collaboration on productivity. They found collaboration could increase the opportunity to publish papers for individual researchers. Pravdi & Olui-Vukovi (Pravdi & Olui -Vukovi, 1986) also found productivity correlate of collaboration in research. In addition, Lee and Bozeman(2005) analysed impacts of collaboration on productivity at the individual level, and the conclusion is mixed: If measured by normal count of one's publications, the productivity of a researcher is statistically correlated with the number of collaborators. If measured by fractional count, the correlation is not statistically significant.

Furthermore, many recent studies found that the influence of collaboration relies on the latent variables such as collaboration pattern, discipline characteristic, and a scientist's research ability. He et al. (He, Geng, & Campbell-Hunt, 2009) examined the influence of the collaboration of different levels, namely, international, domestic and within-university collaboration. Sooryamoorthy (Sooryamoorthy, 2009) and Yoshikane (Yoshikane, Nozawa, Shibui, & Suzuki, 2008) investigated the differences of collaboration at different disciplines, and how it result in different research output. Bozeman & Corley (Bozeman & E. Corley, 2004) revealed that the effect of collaboration depends on scientists. For the senior researchers, collaboration may reduce the productivity because collaborating with graduate students and postdoctoral researchers for the mentoring purpose is not an effective choice for them. Yoshikane (Yoshikane et al., 2006) found that the impact of collaboration is associated with the collaborators, and they conclude that collaborating with productive authors is more helpful for increasing productivity.

Previous studies of career age have shown that the career age is influential on productivity (Dietz & Bozeman, 2005). According to the analysis of Costas and Leeuwen (Costas & Leeuwen, 2010), The productivity in the middle of career is higher than the beginning and the end of the career. Bonaccorsi (Bonaccorsi & Daraio, 2003) also argued that aging may decrease productivity. Despite many concerns with the impacts of scientists' age on productivity, few studies have research its effect on collaboration. With the aim to fill this gap, in this article, we explore how collaboration has developed over various stages of career.

Data and Methodology

Data collection

As mentioned above, patterns of collaboration may vary in different scientific fields. We identified four disciplines based on Journal Citation Report (JCR) of Thomson Reuters. Two of them, Mathematics and Computer Science (theory and

methods), represent theoretical disciplines, while the other two, Organic Chemistry and Virology, represent experimental ones. We choose these two types of disciplines because we want to explore how collaboration and its impact differ between theoretical disciplines and experimental disciplines. Besides, we chose these four subjects also because they are relatively independent - most of the journals in these subjects belong to a single subject category. In contrast, interdisciplinary subjects tend to have many authors publishing in a diverse range of subject matters.

All the documents published during the period of 1970-2010 in journals[69] listed under these four subject categories (JCR) were downloaded from the Web of Science. Authors were classified by the year when they published their first publications. For example, those who published there first publications in 1981[70] were defined as newcomers in 1981.

We choose newcomers in 1981 as the target population of the study, and study their performance from 1981 to 2010. We expect that the 30-year time window is long enough to cover the scientific career of most scientists. The number of selected authors and their publications from 1981 to 2010 is shown in Table 1.

Table 1. Target scientists and their publications in the period of 1981 to 2010

	Computer Science	*Mathematics*	*Organic Chemistry*	*Virology*
Scientists	676	1308	4211	1600
Publications	2651	7733	25725	8078

Finally, the 30-year research career window was divided into 5 equal-length periods. Each 6-year period is considered as a representation of one stage of career. In comparison to the 3-year or 10-year divisions, the 6-year division gives more stable results and therefore is used in this study. In the results, the change of relationships between collaboration and productivity over time across these stages will be shown.

Measurement

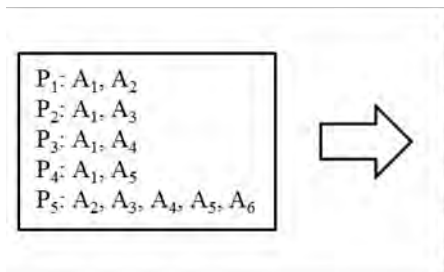
- Collaboration

In this study, we use two relevant yet distinct measures of collaboration at each stage for each scientist. One is the average number of authors per paper (COPP) at each stage. It's a widely used indicator of collaboration, which is in fact known as

[69] For there are too many journals in organic chemistry subfield, we selected 11 most import journals hereby. In the other three disciplines, all the journals are involved.

[70] It's reasonable to assume that the authors who did not publish from 1970 to 1980 are a newcomer of 1980. According to a researcher by YOSHIKANE, 7 years publication interval is long enough, and thus 10 years interval is justifiable.

Collaboration Degree in previous literature (Savanur & Srikanth, 2009). COPP is calculated stage by stage. COPP($stage_i$),i=1 to 5, is calculated based on the publications at stage i. The other measure is the number of unique co-authors at each stage (COPS). It aggregates the total number of unique collaborators. Unlike COPP, COPS emphasizes the breadth of collaboration at each stage. A scientist may have a large COPS score but a small COPP score if he/she collaborated with many different researchers but works with them separately, as illustrated in Figure 1.



A	COPP	COPS
A_1	2.0	4
A_2	3.5	5
A_3	3.5	5
A_4	3.5	5
A_5	3.5	5
A_6	5.0	4

Figure 1 An example about the calculating of the COPP and COPS (P: Publication, A: Author)

- Productivity

We use three measures of scientific productivity of each scientist per stage. They are the number of his/her publications (FULL) at each stage; the fractional count of the publications (FRAC), which is normalized by the number of authors; and the number of first-authored publications (FIRST), which only consider the first-authored publications. These three measures were adopted by some previous studies (Yoshikane et al., 2008).

Results and Discussions

Trends in collaboration

Before examining the change of correlation between productivity and collaboration over career, we explored the trend of the collaboration and productivity first. Figure 2(a) presents data of the mean COPP of newcomers of 1981 at each stage. As expected, COPP increases in the past 30 years in all the four scientific disciplines, although the increase is not so evident in the field of mathematics and computer science as that in the fields of virology and chemistry. It's not surprised to attain such a result since many previous studies has proved that in experimental disciplines scientists is more likely to collaborate than ones in theoretical disciplines.

However, it's noteworthy that the trend of COPP is almost completely consistent with the average value at the period. In chemistry, for example, COPP at the first stage (the first 6 years in their scientific career or the years from 1981 to 1986) is

around 4, while at the same time the average number of co-authors is also about 4; when the COPP became 8 at the final stage (from 2005 to 2010), the average number of co-authors at this period remains the same. The increase of COPP cannot be simply explained by the argument that scientists are more likely to collaboration at the later stage. COPP stays steady during the research career, after excluding the effects of average trend. In other words, one scientist's collaboration scale per paper did not increase or diseases at different stage. This result is somewhat beyond expectations.

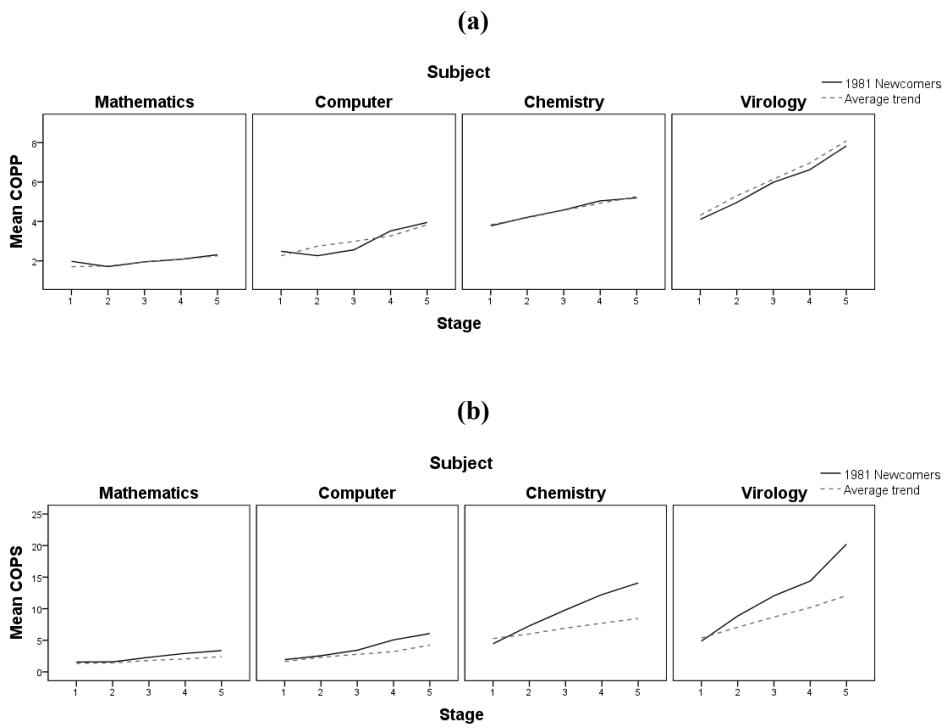


Figure 2. Trends of collaboration over career stage: (a) COPP; (b) COPS

Figure 2(b) shows the growth curve of another measure for collaboration -- COPS. In general, COPS go through a similar increase as COPP over the past 30 years. Furthermore, in comparison with scientists in general, COPS grows more dramatically. At the first stage, target researchers' COPS equal the average value; while at the later stages, it exceeded the average level of collaboration. It suggests that, scientists tend to collaborate with more scientists in later stages than early stages, both absolutely and relatively. Especially in the experimental disciplines such as chemistry and virology, we can see a distinct increasing trend of collaboration in terms of COPS.

Trends in Productivity

The distributions productivity by stage are shown in figure 3(a) (FULL), Figure 3(b) (FRAC) and Figure 3(c) (FIRST). In Figure 3(a), an increase of productivity is evident in all four disciplines. In mathematics science particularly, an inverted U-shape curve can be observed, and the productivity began to fall gently at the last stage. Relatively, the researchers in experimental disciplines have a longer growth trend. Within the 30-year time window, their productivity is still increasing.

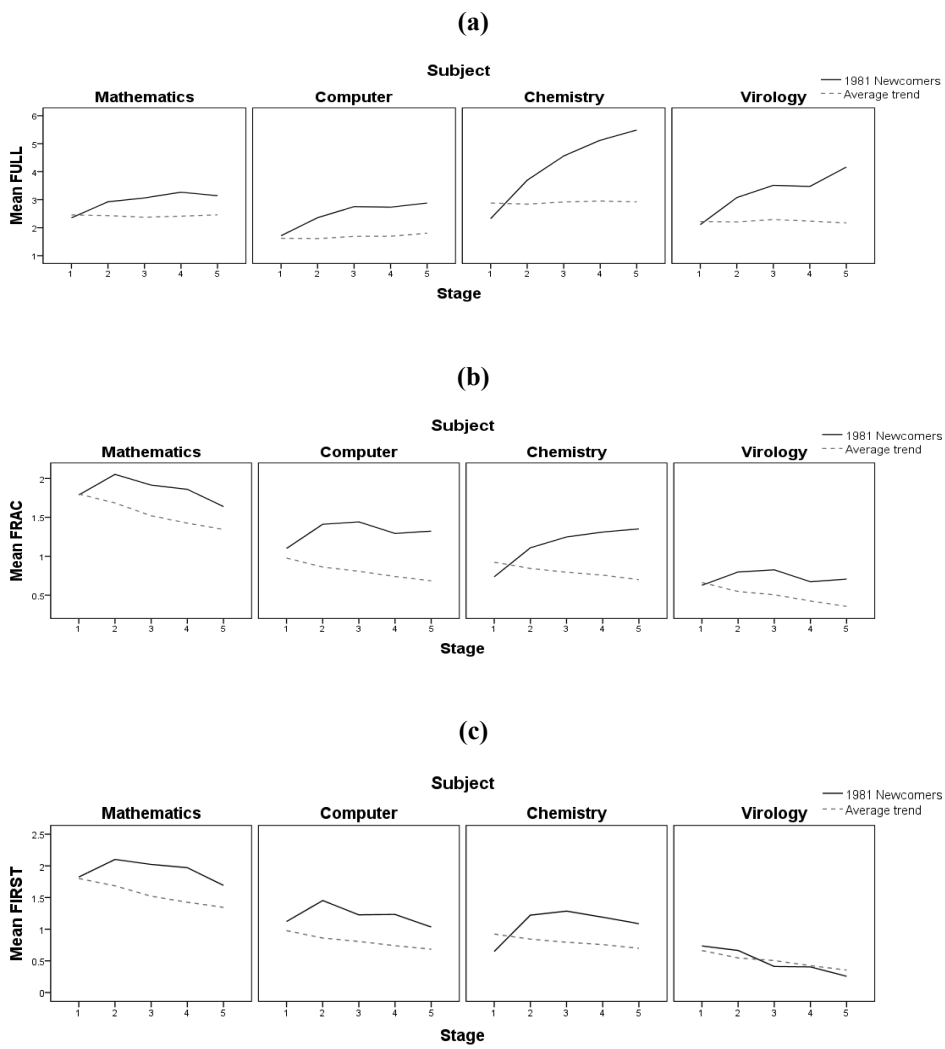


Figure 3. Trends of productivity over career stage: (a) FULL; (b) FRAC; (c) FIRST

To understand the effect of the average change in productivity over the past 30 years, we calculated the average productivity at each period. Surprisingly, this

indicator didn't increase with the development of science and extension of journals. In all of the disciplines we selected, this value remains constant in past 30 years. As showed in Figure 2(a), on average, the mean productivity per stage (or 6 years) remain 3 (Chemistry), 2 (Computer Science), or 2.5(Mathematics and Virology) from early 1980s to recent years. A possible explanation for this phenomenon is that, in spite of more journals or conferences published now, the number of scientists grows accordingly.

A more obvious bell-shaped curve can be seen from Figure 3(b), which shows the trend of productivity measured by fractional count (FRAC). In general, FRAC peaked at an early stage. For example, in computer science and virology, FRAC reached its maximum value (about 1.5) at the third stage, while in mathematics the peak stage is even earlier, which was reached at the second stage. Interestingly, measured by FRAC, the productivity of scientists in theoretical disciplines seems higher than those in experimental disciplines; while measured by FULL, the result is just the opposite: the productivity in experimental disciplines is higher and it is lower in theoretical disciplines.

The similar trend is observed in FIRST. FIRST is used to measure the productivity of scientists only considered first-authored papers. Like FRAC, FRIST start to decline after a short growth. Additionally, in theoretical disciplines, scientists have a higher FIRST than those in experimental fields.

The effect of collaboration on productivity

This article focuses mainly on the relationship between collaboration and productivity at different stages. We calculated their correlations to explore how this relationship changed over career. With 3 measures of productivity and 2 measures of collaboration, a total of 6 sets of correlations will be presented here. The impact of COPP and COPS on productivity will be reported respectively.

- Correlation between productivity and COPP

Figure 4 displays the correlation between COPP and productivity over career stages of 30 years. In the case of the full count, we find no significant correlation between collaboration and productivity except for the chemistry field. In chemistry, a weak negative correlation (no more than 0.1) is found during the second to five career stages. In other disciplines, negative correlations are also observed. However, the correlation is not statistically significant. A negative correlation between COPP and FULL means that: first, collaborating in one publication does not help the productivity; secondly, authors at the lower end of productivity are likely to involve more co-authors than high productive ones on average. One possible interpretation is that high productivity scientists are more capable of publishing a paper with fewer co-authors. In other words, they don't have to collaborate with many co-authors.

Similarly, we also examined how COPP is associated with FRAC and FIRST. For these two measures, negative correlations are significant at most stages in each discipline, although the values of correlation coefficients are not very high. Again,

a negative correlation suggests that higher co-authors per paper is companied with lower productivity measured in FRAC and First, and vice versa.

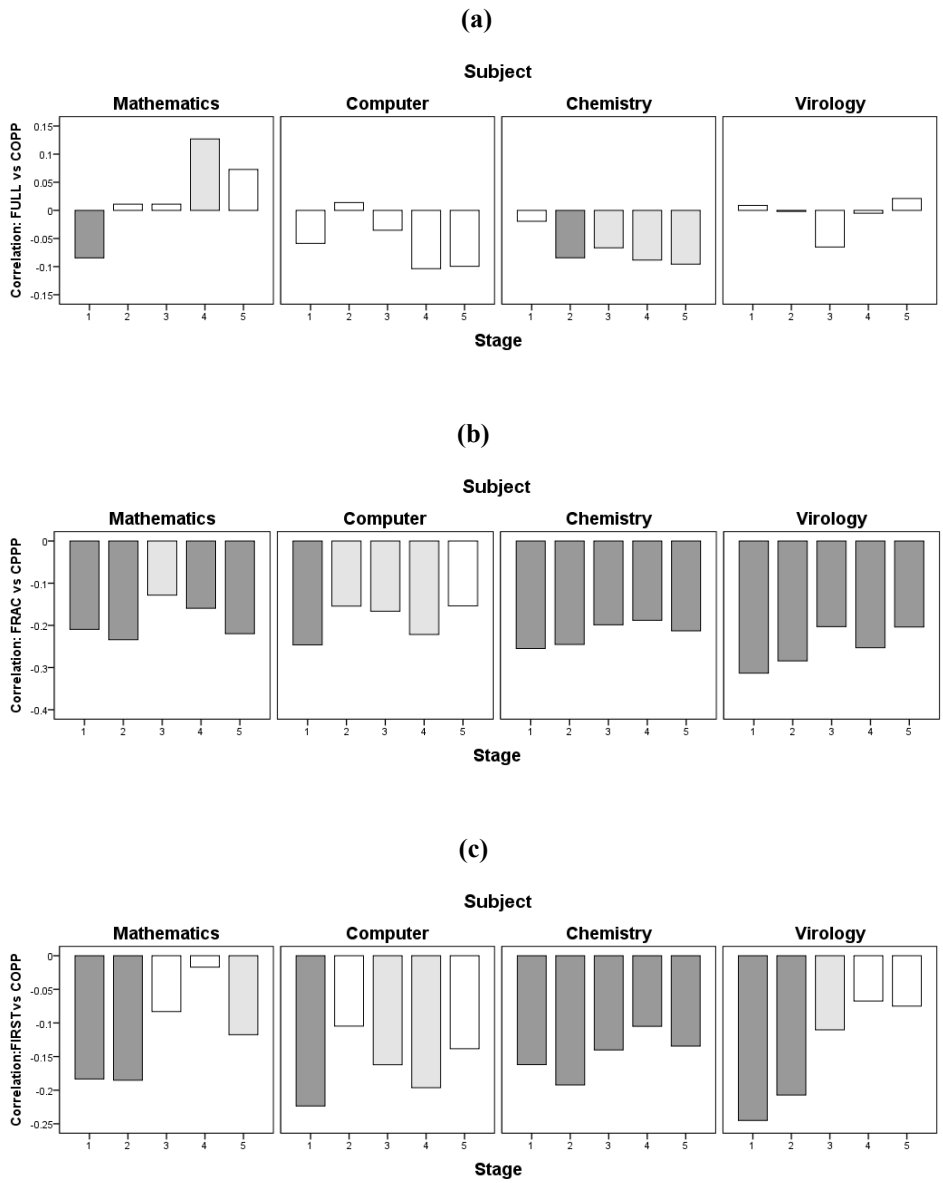


Figure 4. Correlation between COPP and productivity: (a) FULL; (b) FRAC; (c) FIRST

It's important to note that the negative relationship is evident at the early stages. This result suggests that the COPP has a larger "adverse" effect on productivity at

early stages. At early stages, collaboration is a primary way of mentoring, and thus a low value of COPP, which means a small group, is probably better for young scientists to be mentored.

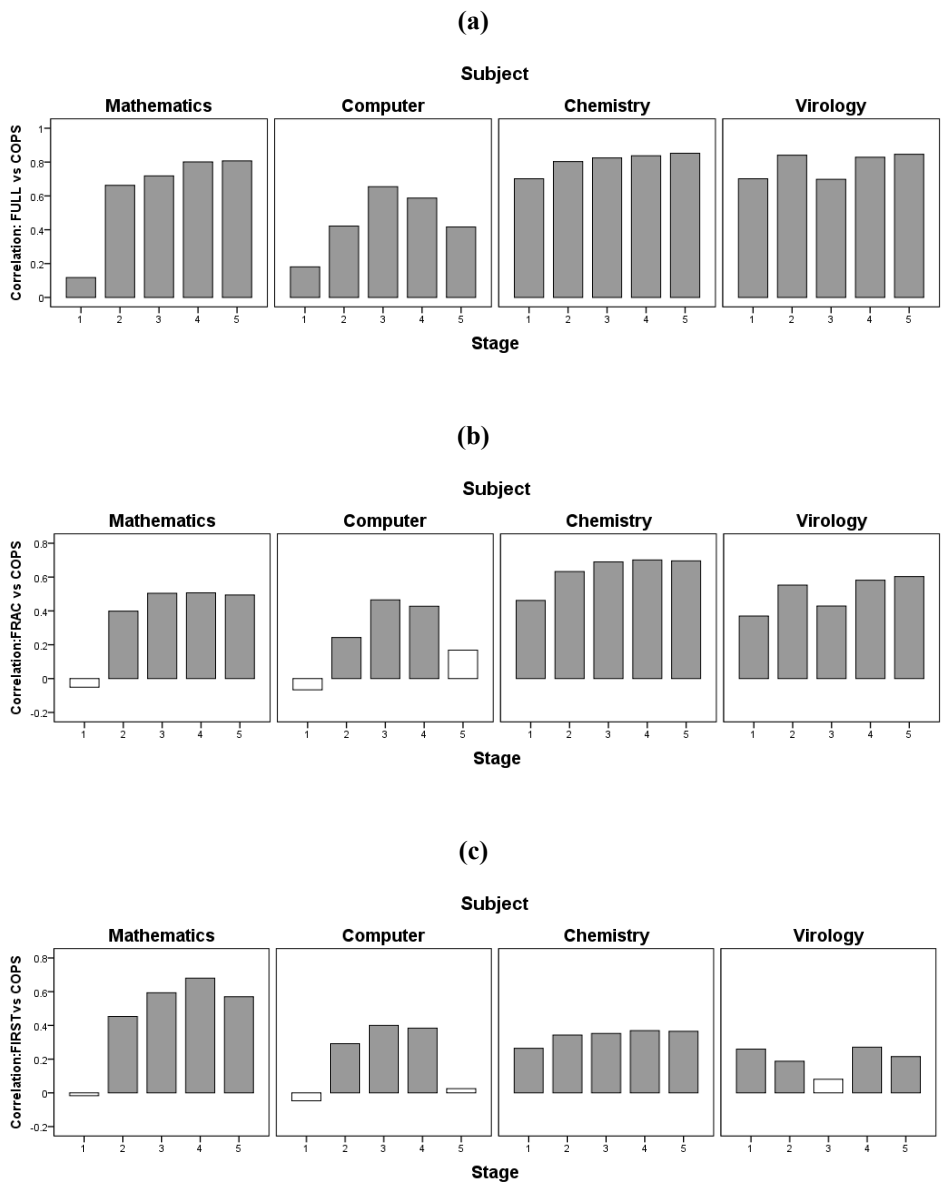


Figure 5. Correlation between COPS and productivity: (a) FULL; (b) FRAC; (c) FIRST

- Correlation between productivity and COPS

In this part, we examined the relationship between COPS and productivity. As shown in Figure 5, there is an obvious positive relationship between the two, especially when the productivity is measured in terms of full count. On average, the correlation coefficient each stage is between 0.6 and 0.8 in all the fields except computer science (between 0.2 and 0.6). We expected to see a positive relationship between COPS and productivity but didn't expect to see the correlation coefficient so strong. It states that the collaboration, in terms of the total number of collaborators at each stage, has an important effect on the productivity. The more different collaborators one collaborates have, the higher his productivity tends to be.

We also observed that productivity is more likely to be influenced by the collaboration scope in middle or late career. According to this result, for scientists, it's more important to look for collaborators at middle stages (computer science) or later (the other three disciplines). This collaboration strategy is consisted with general characteristics of senior scientists, who are actively seeking partners for research. .

A similar trend, albeit with a lower correlation coefficient is observed when productivity measured by FRAC. With regard to this measure is originally used to offset some effect of collaboration, the lower correlation is not difficult to understand here. The relationship tends to more weak when productivity is measured by FIRST. Collaboration has less influence on publishing first-authored papers. But obviously, even in this case, collaboration remains helpful and necessary, since the influences are positive at most stages.

- Independence between COPP and COPS

An important question is to what extent the two metrics of collaboration, COPP and COPS, are correlated. The results above revealed an interesting paradox: the productivity has a negative correlation with collaboration measured by COPP, but a significant positive relationship when collaboration is measured by COPS. The seemingly inconsistent results, however, has a reasonable explanation. COPP and COPS represent two different aspects of collaboration. COPP reflects the scale of research group when publishing one paper, while COPS focuses on the openness of scientists, it measures the total breadth of collaborators at one stage. For example, one scientist could have a large COPS and a small COPP, if he has many different collaborators at a same time, but produces papers separately with them in small group. Based on this result, scientists with large COPS and small COPP tend to have a high productivity.

To reveal the relationship between COPP and COPS, we reviewed the correlation between them in Figure 6. At different stages, the correlations are different. In mathematics or computer science, the correlations in the middle stages are lower than 0.4, which means in middle career stages the correlation between COPP and COPS is very weak. While in Chemistry and Virology, the correlation in the last

one or two stages are even lower than 0.2, which means at those stages COPP and COPS are essentially independent of one another. Actually, low correlations between COPP and COPS justify their uses in this study.

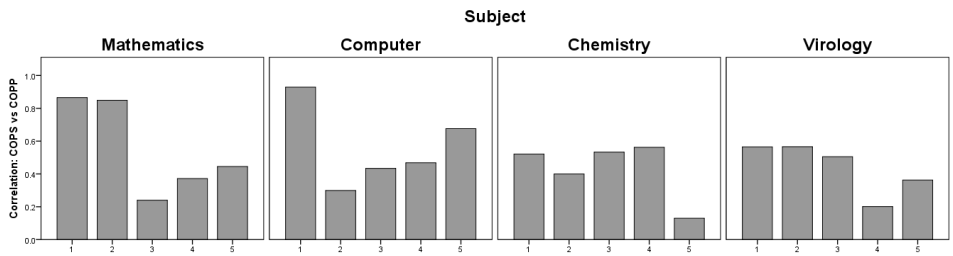


Figure 6. Correlation between COPS and COPP

Conclusions

The main purpose of this study is to examine the changing of correlation between collaboration and productivity over career. We hypothesize that collaboration has more influence on productivity at the early stage. Somewhat surprisingly, the results give no evidence to that. In contrast, at early stages, not only COPP is proved to be more negative associated with a high productivity, but also COPS has a low positive correlation with it. So collaboration should not be emphasized especially at early career. It's better to cooperate with a few partners and improve experience gradually. Instead, we find that, scientists at middle and later stages could benefit more from collaboration than those at early stages. From a collaboration strategy standpoint, the result suggests that scientists should give more emphases on collaboration at later stages rather than at early stages. This suggestion seems reasonable regarding the change of research patterns. At early stages, scientists usually work on a narrowly defined research field, and collaboration may not be critical; while at later stages, collaboration is essential because the research topics become wider and more technical.

Author disambiguation is an inevitable problem in a study of this kind. Besides COPP, all the other indices, including COPS, FULL, FRAC, FIRST, would be affected by this issue. Romanised Chinese and Korean author names usually cause most of the author disambiguation issue. Fortunately, since we choose the authors who began their academic career in 1981, when Chinese and Korean authors have not yet flourish. It avoids the problem to a certain extent.

References

Beaver, D. B. (1979). Studies in scientific collaboration: part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-1830. *Scientometrics*, *1*(2), 133-149.

Birnholtz, J. P. (2007). When Do Researchers Collaborate ? Toward a Model of Collaboration Propensity. *Journal of the American Society for Information Science*, *58*(14), 2226-2239. doi: 10.1002/asi.

Bonaccorsi, A., & Daraio, C. (2003). Age effects in scientific productivity The case of the Italian National Research Council (CNR). *Scientometrics*, *58*(1), 49-90.

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: implications for scientific and technical human capital. *Research Policy*, *33*(4), 599-616. Elsevier.

Corley, E. a, & Sabharwal, M. (2010). Scholarly Collaboration and Productivity Patterns in Public Administration: Analysing Recent Trends. *Public Administration*, *88*(3), 627-648.

Costas, R., & Leeuwen, T. N. V. (2010). A Bibliometric Classificatory Approach for the Study and Assessment of Research Performance at the Individual Level : The Effects of Age on Productivity and Impact. *Journal of the American Society for Information Science*, *61*(April), 1564-1581.

Dietz, J., & Bozeman, B. (2005). Academic careers, patents, and productivity: industry experience as scientific and technical human capital. *Research Policy*, *34*(3), 349-367.

Ding, W. W., Levin, S. G., Stephan, P. E., & Winkler, A. E. (2010). The impact of information technology on academic scientists' productivity and collaboration patterns. *Management Sci*, *56*(9), 1439–1461.

Duque, R. B. (2005). Collaboration Paradox: Scientific Productivity, the Internet, and Problems of Research in Developing Areas. *Social Studies of Science*, *35*(5), 755-785.

He, Z.-L., Geng, X.-S., & Campbell-Hunt, C. (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university. *Research Policy*, *38*(2), 306-317. doi: 10.1016/j.respol.2008.11.011.

Hsu, J.-wien, & Huang, D.-wei. (2010). Correlation between impact and collaboration. *Scientometrics*, (July 2010), 317-324. doi: 10.1007/s11192-010-0265-x.

Jansen, D., Görtz, R., & Heidler, R. (2009). Knowledge production and the structure of collaboration networks in two scientific fields. *Scientometrics*, *83*(1), 219-241.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research policy*, *26*(1), 1-18. Elsevier.

Landry, R. (1998). The impact of transaction costs on the institutional structuration of collaborative academic research. *Research Policy*, *27*(9), 901-913.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, *35*(5), 673-702.

Link, A. N., Paton, D., & Donald, S. S. (2002). An analysis of policy initiatives to promote strategic research partnerships. *Research Policy*, *31*(8-9), 1459-1466.

Perkmann, M., & Walsh, K. (2009). The two faces of collaboration: impacts of university-industry relations on public research. *Industrial and Corporate Change*, *18*(6), 1033. Oxford Univ Press.
Porac, J. (2004). Human capital heterogeneity, collaborative relationships, and publication patterns in a multidisciplinary scientific alliance: a comparative case study of two scientific teams. *Research Policy*, *33*(4), 661-678.
Pravdi, N., & Olui -Vukovi, V. (1986). Dual approach to multiple authorship in the study of collaboration/scientific output relationship. *Scientometrics*, *10*(5), 259-280. Springer.
Savanur, K., & Srikanth, R. (2009). Modified collaborative coefficient: a new measure for quantifying the degree of research collaboration. *Scientometrics*, *84*(2), 365-371.
Sooryamoorthy, R. (2009). Collaboration and publication: How collaborative are scientists in South Africa? *Scientometrics*, *80*(2), 419-439.
Yoshikane, F., Nozawa, T., Shibui, S., & Suzuki, T. (2008). An analysis of the connection between researchers' productivity and their co-authors' past attributions, including the importance in collaboration networks. *Scientometrics*, *79*(2), 435-449.
Yoshikane, F., Nozawa, T., & Tsuji, K. (2006). Comparative analysis of co-authorship networks considering authors' roles in collaboration: Differences between the theoretical and application areas. *Scientometrics*, *68*(3), 643-655.

HOW TO COMBINE TERM CLUMPING AND TECHNOLOGY ROADMAPPING FOR NEWLY EMERGING SCIENCE & TECHNOLOGY COMPETITIVE INTELLIGENCE: THE SEMANTIC TRIZ TOOL AND CASE STUDY

Yi Zhang, [1]Xiao Zhou, [1]Alan L. Porter[2] and Jose M. Vicente Gomila[3]

[1]*yi.zhang.bit@gmail.com; belinda1214@126.com*
School of Management and Economics, Beijing Institute of Technology, Beijing, 100081(China)

[2] *alan.porter@isye.gatech.edu*
Technology Policy & Assessment Center, Georgia Institute of Technology, Atlanta, GA; and Search Technology, Inc. Norcross, GA (USA)

[3]*vicentegomila@gmail.com*
Universitat Politecnica de Valencia, Valencia, Spain

Abstract
Competitive Technical Intelligence (CTI) addresses the landscape of both opportunities and competition for emerging technologies as the boom of Newly Emerging Science & Technology (NEST) – characterized by a challenging combination of great uncertainty and great potential – has become a significant feature of the globalized world. We have been focusing on the construction of a "NEST Competitive Intelligence" methodology, which blends bibliometric and text mining methods to explore key technological system components, current R&D emphases, and key players for a particular NEST. As an important part of these studies, this paper emphasizes the semantic TRIZ approach as a useful tool to represent "Term Clumping" results and apply them to Technology Roadmapping (TRM), with the help of semantic Problem & Solution (P&S) patterns. A greater challenge lies in the attempt to extend our approach into NEST Competitive Intelligence studies by using both inductive and purposive bibliometric approaches. Finally, an empirical study for Dye-Sensitized Solar Cells (DSSCs) is used to demonstrate these analyses.

Keyword
Semantic TRIZ; Text Mining; Technology Roadmapping; DSSCs

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5) and Visualization and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction

Newly Emerging Science & Technology (NEST), related to "disruptive technology," can either be a new combination of existing technologies or entirely new technologies whose application to problem areas or new commercialization challenges (e.g., systems or operations) can cause major paradigm shifts in regard to certain technology products or create entirely new products (Kostoff et al. 2004). There are two views of disruptive technologies: one emphasizes the "different" nature of technology, whereas the other emphasizes the emergence of a new high technology (Walsh 2004). Obviously, the definition of NEST involves both. NESTs create growth in the industries they penetrate or create entirely new industries through the introduction of products and services that are dramatically cheaper, better, and more convenient (Walsh and Linton 2000). The apparent acceleration in NESTs is not only an obvious global feature of today's Science, Technology & Innovation (ST&I) scene, but is also becoming a key factor in national R&D programs (Zhang et al. 2012a). Globally, NEST industries have played increasingly important roles in R&D strategy and management. Governments and multinational corporations are paying attention to the fundamental methodology research on NESTs and NEST industries.

The Society of Competitive Intelligence Professionals defined Competitive Intelligence (CI) as "Timely and fact-based data on which management may rely in decision-making and strategy development. It is obtained through industry analysis, which means understanding all the players in an industry; competitive analysis, which is understanding the strengths and weaknesses of competitors." On the other hand, Technical Intelligence (TI) is "the capture and delivery of technological information as part of the process whereby an organization develops an awareness of technological threats and opportunities" (Kerr et al. 2006; Yoon and Kim 2012).

Over the past twenty years, the Georgia Institute of Technology's (Georgia Tech) Technology Policy and Assessment Centre (TPAC) has been pursuing variants of the "Tech Mining" approach (Porter and Cunningham 2005) for retrieving usable information on the prospects of particular technological innovations from ST&I resources (Porter and Detampel 1995; Zhu and Porter 2002). Based on bibliometrics and text mining techniques, we have contributed to Competitive Technical Intelligence (CTI): we illuminate current R&D emphases and key players, defining our approaches as "NEST Competitive Intelligence." Our research aids in determining how to realize the CTI-NEST combination effectively. On the one hand, we aim to handle challenging analyses of millions of phrases and terms derived from ST&I datasets by Natural Language Processing (NLP). We focus on "Term Clumping" steps, to improve the cleaning and consolidation of phrases and terms (Zhang et al. 2012b). We also combine qualitative and quantitative methodologies to compose a Technology

RoadMapping (TRM) model, a useful instrument to visually indicate technology development trends (Zhang et al. 2012a).

As defined by Altshuller several decades ago, the underlying concept of TRIZ indicates that potential logical rules and principles lead invention from problems to solutions (Rizzi 2011). Currently, TRIZ has become an extremely powerful methodology that enhances creativity in the engineering fields (Savransky 2000) and is rapidly extending to other innovation activities. Verbitsky (2004) introduced bibliometric techniques into TRIZ theory and realized the new methodology – semantic TRIZ – with the help of the software GoldFire Innovator (see Reference). Aiming to bridge our term clumping results and TRM approaches effectively, in this paper, we emphasize the concept and methodology studies and concentrate on the stepwise processes of semantic TRIZ. Moreover, "Subject – Action – Object (SAO)" structures are engaged, which represent the key concepts and expertise of inventors by defining the explicit relations among components (Choi et al. 2011). Especially, we have recently explored such an approach for the case of a Triple Helix innovation study (Zhang et al. 2013).

The structure of this paper includes the following parts: In Section 2, we summarize key literature on TRIZ and semantic TRIZ. Section 3 describes our Dye-Sensitized Solar Cells (DSSCs) data and elaborates on our methodology for using semantic TRIZ to combine term clumping results with TRM effectively. Finally, we draw conclusions in Section 4.

Literature Review

TRIZ

In the 1940s, G.S. Altshuller started to extract similar ideas and analogous solutions from massive numbers of patents, summarize the common patterns of "original" and "creative" inventions, and then, the "theory of inventive problem solving" was named (Nakagawa 2001; Rantanen and Domb 2008). The idea of TRIZ depends on two major ideas: contradiction and ideality. Contradiction is the basic law of materialist dialectics while ideality is the essence of idealism (Savransky 2000). TRIZ theory indicates that fundamental performance limits arise when one or more unsolved contradictions exist in a system, which also enables system engineers to identify requirement contradictions that inhibit desired performance levels (Shahbazpour and Seidel 2007; Blackburn et al. 2012). Savransky (2000) defined the change of technical systems and processes as "gradual – consequent – or breakthrough – revolutionary – development" and claimed that TRIZ recognized these long-term changes as the trajectory of a technique's evolution, which came into being and progressed because of human activities in research, design and development. Moreover, two kinds of technological forecasting methods of TRIZ evolutionary theory are summarized: 1) TRIZ includes the technical system's evolutionary S-curve, a determination

tool for a technology's degree of maturity and system operator; 2) the second draws out, as a natural outgrowth of the TRIZ research, the evolutionary patterns of technological systems (Stephen 2002).

After conducting a survey on 43 TRIZ applications in Europe and North America, Moehrle (2005) noticed that (1) the whole toolset of TRIZ is not always necessary to solve inventive problems; (2) a certain combination of several techniques is often applied; (3) different tools do not appear to use techniques with great variance from each other; and (4) constructing TRIZ training and implementation is helpful.

Semantic TRIZ and Subject – Action – Object (SAO) Analysis

We have dabbled in TRIZ theory and its related researches for nearly ten years and have tried to combine TRIZ research with patent analysis to explore the potential significance for technology innovation management (Porter and Cunningham 2005). Even more relevant, Kim et al. (2009) identified a kind of "problem & solution" pattern, which was similar to the contradiction concept in TRIZ (although the authors never mentioned "TRIZ" in this paper). He retrieved these patterns from patent records by automatic semantic analysis and located them in the time dimension. His work described the technology trend, just as TRM does, according to "when problems occurred" and "when and how problems are solved."

"Subject – Action – Object" analysis emphasizes the "key concepts" and provides information on their semantic relationships (Cascini et al. 2004; Choi et al. 2011). This approach can also be organized according to a problem & solution pattern, where Action – Object (AO) states the problem and the Subject (S) forms the solution (Moehrle et al. 2005). Not to conflict with SAO model, Verbistsky (2004) applied semantic analysis to traditional TRIZ theory by identifying the meaning of text on the basis of its semantic items and also extended the "Problem & Solution" patterns in his new "Semantic TRIZ" concept. Several efforts were added by the related software, GoldFire Innovator, involving (1) Natural Language Processing (NLP) techniques to extract semantic items (e.g., system components, actions, and solutions); (2) Machine Learning techniques to "teach" computers to understand semantic items and memorize them; and (3) matching algorithms to provide an exact solution to the problem. In addition, we have provided another literature review on TRIZ and semantic TRIZ in one of our previous papers where we focused on Triple Helix innovation studies in combination with Term Clumping, semantic TRIZ and TRM (Zhang et al. 2013).

Data and Methods

Data

We have concentrated on Dye-Sensitized Solar Cells (DSSCs) since 2008. Addressing DSSCs, we built a combined ST&I dataset searching in two global databases – the Science Citation Index Expanded of Web of Science (WoS) and EI Compendex. With our continuous studies, we have gathered rich experiences and expert networks, and also, have continued to update new records. In this paper, we still focus on our old DSSC dataset, including 5784 publication records and covering the time span from 1991 (DSSC was first announced in *Nature* by the two Swiss scientists O'Regan and Gratzel [1991] in a seminal paper) through 2011 (not complete for this last year).

Term Clumping

Various kinds of ST&I text analyses emphasize the terms derived from NLP techniques; however, these phrases and terms, which indicate potentially valuable information and significant topics, can easily reach more than 100,000 items. Aiming to solve this challenge for the sake of further topical analyses, we defined "term clumping" as the steps used to clean, consolidate and cluster rich sets of topical phrases and terms in a collection of documents. Currently, all "term clumping" processing for this dataset have been completed and we have obtained review by seven experts (Georgia Tech, Tsinghua University, Dalian University of Technology, IBM, and Booz Allen Hamilton, Inc.) of the Top 300 valuable terms and 8 topical factors from 90,980 phrases and terms (Zhang et al. 2012b). We present the term clumping stepwise results in Table 1, and the 8 topical factors and their related items in Table 2 – fundamental material for further semantic TRIZ study here.

Comments on the term clumping steps are given here; we have discussed these issues in detail elsewhere (Zhang et al. 2012b):

(1) The "Term_Clustering.vpm" (marked as *) was a macro only fit for a small dataset. Here, we only used it to classify the terms and selected those that contained 2, 3 or 4 words;

(2) It is critical to remove all the terms (marked as **) that only appear in one record, because several of the latest emerging technological terms should be ignored. Thus, this step depends on any particular case. If a researcher aims to explore more NESTs, skipping over it could be a better option;

(3) We ran TFIDF analysis (marked as ***) and evaluated the TFIDF value for each term; we then removed the 100 highest TFIDF terms, and use the remaining 14740 terms for the next steps. The number "100" is not a strict range for TFIDF analysis; we eliminate these terms based on our previous experience, because the

top 1% of total TFIDF terms (e.g., "photovoltaic performance," "electron transport," etc.) should be overly common terms in the DSSC domain.

Table 1. Term Clumping Stepwise Results (Zhang et al. 2012b)

Field selection	*Number of Phrases and Terms*
Phrases with which we begin	90980
Step a. Applying Thesauri for Common Term Removal	63812
Step b. Fuzzy Matching	53718
Step c. Combining	
Combine_Terms_Network.vpm (Optional)	Not Applied Here
Term_Clustering.vpm	52161 to 37928*
Step d. Pruning	
Remove Single terms**	15299
Fuzzy Matching	14840
Step e. Screening	
Term Frequency Inverse Document Frequency (TFIDF)	14840 (with the Sequence of TFIDF) to 14740***
Combine_Terms_Network.vpm (Optional)	8038
Step f. Clustering	
Principal Components Analysis (PCA)	11 Topical Clusters

Table2.Topical Clusters and Related Items

Number	*Topical Factors*	*Related Items*	*Rank*
1	Photoelectric property	Photoelectric property, Hydrothermal method, Higher *conversion efficiency*	*6*
2	Sol gel	Sol gel, Sol gel process	*3*
3	Electron donor	Electron donor, Electron acceptor, Molecular design	*7*
4	Ruthenium sensitizers	Ruthenium sensitizers, Ruthenium complex, Efficient sensitizers, Absorption spectrum, Charge transfer sensitizer, Red shift, Density functional theory, High molar extinction coefficient	*4*
5	Electric resistance	Electric resistance, Sheet resistance, Internal resistance	*2*
6	Modulated photocurrent spectroscopy	Modulated photocurrent spectroscopy, Electron diffusion coefficient, Electron traps, Recombination kinetics, Photo-injected electron, Electron diffusion length	*5*
7	Photo-induced electron transfer	Photo-induced electron transfer, Electrons transit, Interfacial electron transfer, Rate constant	*1*
8	Ultraviolet spectroscopy	Ultraviolet spectroscopy, UV vis spectroscopy	*8*

Drawing on expert opinions and literature reviews, we determined before we dived into the DSSCs studies that there are four subsystems in the DSSC domain

– “photo anode,” “sensitizer,” “electrolyte,” and “counter electrode.” However, we did not find it easy to classify the eight topical factors in Table 2 into the four subsystems. We tried to reveal the reasons:

(1) The original purpose for the “term clumping” was to explore more detailed techniques; thus, we applied Term Frequency Inverse Document Frequency Analysis (TFIDF) and removed all Top 100 high TFIDF – value terms, where several typical but common terms (e.g., “counter electrode,” “electrolyte,” and etc) were included in the removed term list.

(2) The inaccuracy of the Principle Components Analysis (PCA) functions of our software VantagePoint (see Reference) and the research backgrounds and experiences of the experts also influenced this classification. This is one motivation to introduce semantic TRIZ to the process of term clumping, which we will discuss in the next section.

Among these interesting topics, we chose “conversion efficiency” as the system component for further semantic TRIZ studies for the following reasons: (1) Undoubtedly, “conversion efficiency” is one of the core evaluation indexes for batteries; thus, it is an important topic in the DSSC domain; (2) In contrast to the “material” topics, the range of its research covers all four subsystems, which probably helps us to draw relations between the items of the 4 subsystems; 3) This topic seems to be easier to understand for us and those who don’t have a specific background in the DSSC field; therefore we would not need to refer to experts for detailed technical questions as much.

Semantic TRIZ

Term clumping studies emphasize the “phrases and terms,” but misunderstanding would likely occur if we simply focus on these isolated ones. After the term clumping steps, the terms “organic dye” and “secondary dye-absorption” are retrieved. They are key terms in the field of DSSCs. However, analysts like us, who lack the background knowledge in this domain, do not know how the two dye-related terms are used in the field. At this time, if we take note of the sentence containing these terms and then analyze the SAO structures, the meaning becomes clearer. We present an example below – the subject “secondary dye-absorption” (double-underline) and the object “organic dye” (blacked) are easily connected by the action “using” (underline). In this instance, we derive the idea to extend the term analysis to SAO structures and related sentences from this sequence.

Fabrication of working electrode for dye-sensitized solar battery, involves… secondary dye-absorption using ***organic dye****…*

As discussed, SAO analysis solves some shortcomings of keyword-based ST&I text analysis and addresses more complete semantic understandings. In this paper,

we combined SAO analysis and semantic TRIZ in the Problem & Solution (P&S) patterns. The resulting mapping with topical factors, SAO models, and P&S patterns is shown in Figure 1:

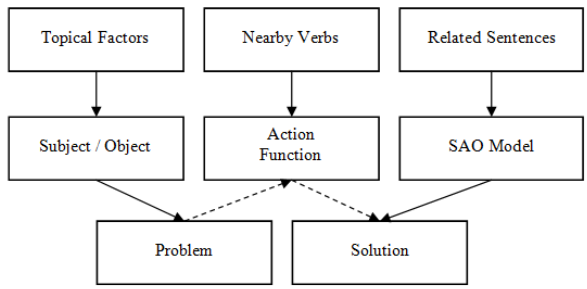


Figure 1.Mapping with Topical Factors, SAO Model and P&S Patterns

(1) We focus on the topical factors/clusters in the term clumping steps, and the nearby "Verbs" are absolute "Actions," which connect their related topical factors as the "Subjects" or "Objects;"

(2) Considering the connection between the SAO model and P&S patterning, it is easy to map the "Subject/Object" to the "Problem," while transferring the whole SAO model to the "Solution" with its "Action" or "Function";

(3) As described in TRIZ theory, the "contradiction between object and tool" and the "ideal final result" could be considered respectively as the Problem and its Solution.

In our earlier research, on the one hand, we ignored the mass of information that could potentially be garnered from verb-related phrases or short sentences and thus experienced some difficulties with understanding the term clumping results. On the other hand, we located the phrases and terms on the TRM directly and paid more attention to the "system components" themselves, but not to the relationships between problems and their solutions. Therefore, this paper emphasizes how to figure out the connection between term clumping results and TRM using the semantic TRIZ tool (shown as Figure 2).

In contrast to current bibliometric research on publication abstracts, a growing number of semantic analyses on patents attempt to draw on the raw patent records, involving both traditional titles/abstracts and the full text. In the past, we ignored this. The semantic TRIZ tool (e.g., GoldFire Innovator, see References) follows these "new" ideas of semantic analyses and improves their ability to

search, calculate, and retrieves topics from the raw records. Usually, semantic TRIZ focuses on the following two parts:

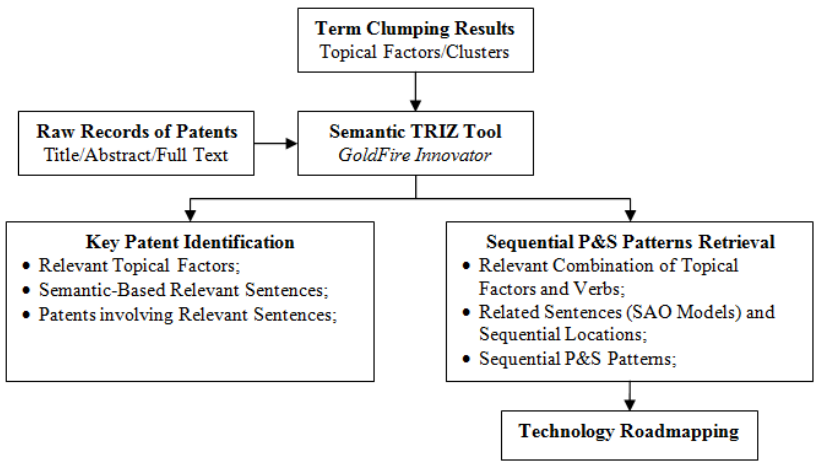


Figure 2.Framework to Apply Term Clumping Results to TRM using Semantic TRIZ Tools

(1) Key patent identification: this is clearly not a new direction for patent analysis but goes back several decades. With semantic TRIZ, we are able to use topic factors/clusters deriving from the term clumping steps to retrieve their relevant sentences. Of course, we could also establish evaluation requirements to rank these relevant sentences (e.g., weighting coefficients or special emphases). In this context, we expect it to be much easier to identify "key patents."

(2) Sequential P&S pattern retrieval: according to the P&S patterns, we can dive into the patent records and understand "what problems occurred" and "what solutions were used to solve these problems." Furthermore, we link the P&S patterns with the time (patent's publication year) to add a chronological dimension when the problem or solution occurred. Also, the sequential P&S patterns afford excellent material for TRM.

Besides on the obvious relationship, that the "solution" solves the "problem," we also listed 3 other types of relations between "problem and problem," "solution and problem" and "solution and solution." These relations help us to understand the potential information contained in the P&S patterns, which might become particularly useful for constructing the TRM. Relations in the P&S patterns are shown in Table 3.

Table 3. Relations in Problem & Solution (P&S) Patterns (Zhang et al. 2013)

Number	*Relations*	*Relations*
1	Problem to Problem	Relate: relations between two problems, e.g., they share subsystem
2	Problem to Solution	Solve: problem is solved by solution indicated
3	Solution to Solution	Relate/Upgrade: relations between two solutions, e.g., they share subsystem; or "next" solution upgrades previous one
4	Solution to Problem	Evolve: solution evolves to new problems (we mark this relation as S/P in TRM)

Once we decided to choose the "conversion efficiency" topic as the system component for example semantic TRIZ studies, we began processing the content of the 5784 records relating to DSSCs with GoldFire Innovator (we used both the title and abstract). As discussed, GoldFire Innovator retrieved related "verbs" near the system component and combined them as SAO structures, which were the same as "P&S" patterns. Samples of the results generated by GoldFire are shown in Table 4.

Table 4. P&S Patterns of Semantic TRIZ on "Conversion Efficiency"

P&S Patterns	*Year*	*Type*	*Level*
The 1-mu m-thick mesoporous film, made by the superposition of three layers, showed *enhanced* solar **conversion efficiency** by about 50% compared to that of traditional films of the same thickness made from randomly oriented anatase nanocrystals.	2005	Solution	M
The dye-sensitized solar cells, comprised of TiO2 photo electrode, deposited at substrate temperature of 200C, show maximum photoelectric **conversion efficiency**; however, further enhancement of *sputtering temperature* drastically *reduces* the **efficiency**.	2002	Problem	T&C
As a result, the *tandem structured cell exhibited* higher photocurrent and **conversion efficiency** than each single DSC mainly caused from its extended spectral response.	2004	Solution	P
The photo-to-electricity energy **conversion efficiencies** of ruthenium-dye-sensitized solar cells (DSC) *are measured* under *a solar simulator*.	2004	Solution	T&C

*M = Materials; T&C = Techniques & Components; P = Products

As an example, we might know some relations between "mesoporous film" and "conversion efficiency" exist, because we have noticed that they both appeared in the high-frequency term list after the term clumping steps. However, we don't know "how they related" and "which kind of influences they had on each other." We show some details in Table 4.

There are 2 additional columns in Table 4: "Type" and "Level." The "type" column is self-explanatory – problem or solution – but for the classification for the "Level" column, we needed to engage experts. The 3 levels are the general phases of technology development as we will discuss in the TRM part.

Technology Roadmapping

Based on the previous analyses, depicting the P&S patterns in Technology Roadmapping became feasible. Just as in the topic selection step, we asked several Ph.D. students who focused on DSSC research at the Beijing Institute of Technology for informed suggestions. They helped us to interpret the empirical findings, and divided up about 100 P&S patterns from GoldFire Innovator among the 3 levels – materials, techniques/components, and products. They also cut out some useless patterns and modified them to make the description succinct and fit for TRM rendition. Finally, 62 P&S patterns were used as objects in TRM. The TRM for "conversion efficiency" research in DSSCs is shown as Figure 3.

Conclusions

We have focused on NEST, CTI, and ST&I research for decades, emphasizing the combination of qualitative and quantitative methodologies for technology management and innovation. Our research focused on the special characteristics of NESTs and applied Tech Mining approaches for CTI and ST&I studies. Currently, on the one hand, we have continued to apply bibliometric and text mining techniques to publication records. On the other hand, we have also started to explore the balance between publication and patent records. While we formerly emphasized the theoretical research, we now pay more attention to practical applications.

This paper proposed a bibliometric method of associating term clumping results with TRM byusing semantic TRIZ tools. Term clumping techniques help us to consolidate the topical information to obtain fruitful conceptual units for further analyses. This provides a novel extension for applying these techniques not just to the analysis of research publications but also to patent records. Semantic TRIZ methods help to relate "clumped" terms to purposive meaning – the Problem & Solution analyses. We think TRM visualization enriches understanding by providing perspective on the evolution of topical R&D emphases, which can, in turn, be associated with the key actors engaged with each and can be ordered chronologically.

We have constructed a framework for NEST Competitive Intelligence studies in which semantic TRIZ is an important tool to bridge term clumping results with TRM. It is also challenging to relate specific P&S actions to emerging technology sub-systems and, from there, to potential applications. We continue to try different ways to combine empirical analyses with a diverse set of multi-disciplinary expertise. We also see potential in enhancing TRM visualizations to

highlight key developmental trends and the changing involvement of different players in those trends.

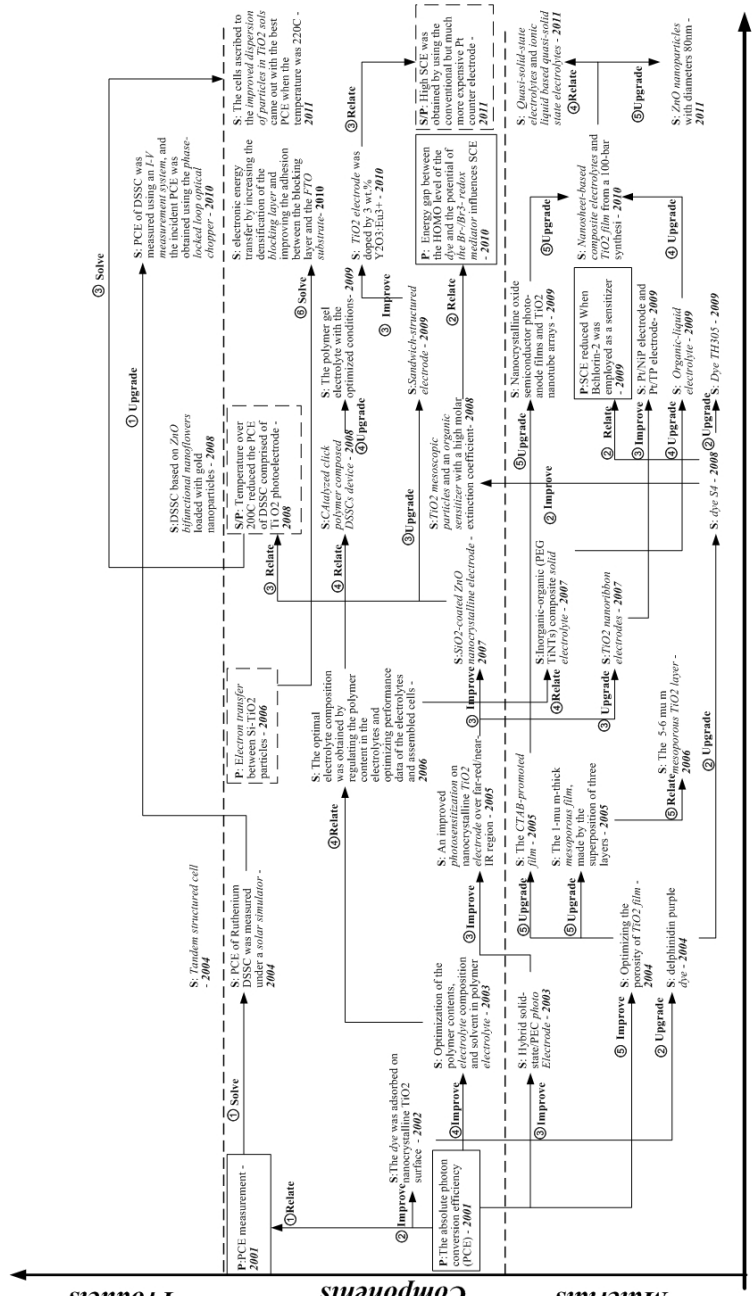


Figure 3.Technology RoadMapping for Conversion Efficiency in DSSCs

Acknowledgments

We acknowledge support from the US National Science Foundation (Award #1064146 – "Revealing Innovation Pathways: Hybrid Science Maps for Technology Assessment and Foresight"). The findings and observations contained in this paper are those of the authors and did not necessarily reflect the views of the National Science Foundation.

References

Blackburn, T. D., Mazzuchi, T. A., Sarkani, S. (2012). Using a TRIZ Framework for System Engineering Trade Studies. *Systems Engineering*, 15(3), 355-367.

Cascini, G., Fantechi, A., Spinicci, E. (2004). Natural Language Processing of Patents and Technical Documentation. In *Document Analysis Systems VI* (pp. 508–520).

Choi, S., Yoon, J., Kim, K., Lee, J., Kim, C. (2011).SAO Network Analysis of Patents for Technology Trends Identification: A Case Study of Polymer Electrolyte Membrane Technology in Proton Exchange Membrane Fuel Cells. *Scientometrics*, 88, 863–883.

Goldfire Innovator, http://inventionmachine.com/products-and-services/innovation-software/goldfire-innovator/(accessed January 1, 2013).

Kerr, C., Mortara, L., Phaal, R., Probert, D. (2006). A Conceptual Model for Technology Intelligence. *International Journal of Technology Intelligence and Planning*, 2(1), 73-93.

Kim, Y., Tian, Y., Jeong, Y., Ryu, J., Myaeng, S. (2009).Automatic discovery of technology trends from patent text. *Proceedings of the 2009 ACM symposium on Applied Computing*, Hawaii, USA.

Kostoff, R. N., Boylan, R., Simons, G. R. (2004). Disruptive Technology Roadmaps. *Technological Forecasting & Social Change*, 71, 141-159.

Kremer, G. O., Chiu, M., Lin, C., Gupta, S., Claudio, D., Thevenot, H. (2012). Application of Axiomatic Design, TRIZ, and Mixed Integer Programming to Develop Innovative Designs: A Locomotive Ballast Arrangement Case Study. *International Journal of Advanced Manufacturing Technology*, 61, 827-824.

Moehrle, M. G. (2005). How Combinations of TRIZ Tools are used in Companies – Results of a Cluster Analysis. *R&D Management*, 35(3), 285-296.

Moehrle, M. G., Walter, L., Geritz, A., Müller, S. (2005). Patent-based Inventor Profiles as a Basis for Human Resource Decisions in Research and Development. *R&D Management*, 35(5), 513–524. doi:10.1111/j.1467-9310.2005.00408.x.

Nakagawa, T. (2001). Introduction to TRIZ: A Technological Philosophy for Creative Problem Solving. *The 23rd Annual Symposium of Japan Creativity Society*, Tokyo, Japan.

O'Regan, B., Grätzel, M. (1991).A Low-cost, High Efficiency Solar-cell based on Dye-sensitized Colloidal TiO2 Films. *Nature*, 353(6346), 737–740.

Porter, A.L., Cunningham, S.W. (2005).*Tech Mining: Exploiting New Technologies for Competitive Advantage*. New York: Wiley.
Porter, A. L., Detampel, M. J. (1995).Technology Opportunity Analysis. *Technological Forecasting & Social Change*, 49, 237-255.
Rantanen, K., Domb, E. (2008).*Simplified TRIZ: New Problem Solving Applications for Engineers and Manufacturing Professionals* (Second Edition). New York: Auerbach Publications, Taylor & Francis Group.
Rizzi, C. (2011). TRIZ and Intelligence Property Management. *Research in Interactive Design*, 3, 123-128.
Savransky, S. D. (2000). *Engineering of Creativity (Introduction to TRIZ Methodology of Inventive Problem Solving)*. Florida: CRC Press LLC.
Shahbazpour, M., Seidel, R. (2007). Strategic Manufacturing System and Process Innovation Through Elimination of Trade-offs. *International Journal of Computer Integrated Manufacturing*, 20(5), 413-422.
Stephen, R.L. (2002).*A Conceptual Design Tool for Engineer: An Amalgamation of Theory of Constraints, Theory of Inventive Problem Solving and Logic*. Virginia: Old Dominion University.
VantagePoint, www.theVantagePoint.com (accessed January 1, 2013).
Verbitsky, M. (2004). *Semantic TRIZ*.triz-journal.com. http://www.triz-journal.com/archives/2004/(accessed January 18, 2013).
Walsh, S. T., Linton, J. (2000). Infrastructure for Emerging Markets Based on Discontinuous Innovations. *Engineering Management Journal*, 12 (2), 23-31.
Walsh, S. T. (2004). Roadmapping a Disruptive Technology: A Case Study - The Emerging Microsystems and Top-down Nanosystems Industry. *Technological Forecasting & Social Change*, 71, 161-185.
Yoon, J., Kim, K. (2012). Trend Perceptor: A Property – Function based Technology Intelligence System for Identifying Technology Trends from Patents. *Expert Systems with Applications*, 39, 2927-2938.
Zanasi, A. (2004). Knowledge Advantage through Online Text Mining. Research Trends in Competitive Intelligence and Virtual Communities Applications. In Sirmakessis, S (Eds.), *Text Mining and Its Applications* (pp. 151-157).New York: Springer-Verlag Berlin Heidelberg.
Zhang, Y., Guo, Y., Wang, X., Zhu, D., Porter, A. L. (2012a). A Hybrid Visualization Model for Technology Roadmapping: Bibliometrics, Qualitative Methodology, and Empirical Study. *Technology Analysis & Strategic Management*, to appear.
Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., Newman, N. C. (2012b). "Term Clumping" for Technical Intelligence: A Case Study on Dye-Sensitized Solar Cells. *Technological Forecasting & Social Change*, to appear.
Zhang, Y., Zhou, X., Porter, A. L., Vicente-Gomila, J. M. (2013). Triple Helix Innovation in China's Dye-Sensitized Solar Cell Industry: Hybrid Methods with Semantic TRIZ and Technology Roadmapping. *Scientometrics*, to appear.

Zhu, D., Porter, A. L. (2002). Automated Extraction and Visualization of Information for Technological Intelligence and Forecasting. *Technological Forecasting & Social Change*, 69, 495-506.

HOW WELL DEVELOPED ARE ALTMETRICS? CROSS-DISCIPLINARY ANALYSIS OF THE PRESENCE OF 'ALTERNATIVE METRICS' IN SCIENTIFIC PUBLICATIONS (RIP)

Zohreh Zahedi[1], Rodrigo Costas[2] and Paul Wouters[3]

[1] *z.zahedi.2@ cwts.leidenuniv.nl,* [2] *rcostas@cwts.leidenuniv.nl,*
[3] *p.f.wouters@cwts.leidenuniv.nl,*
Centre For Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands

Abstract
In this paper an analysis of the presence and possibilities of altmetrics for bibliometric and performance analysis is carried out. Using the web based tool Impact Story, we have collected metrics for 20,000 random publications from the Web of Science. We studied the presence and frequency of altmetrics in the set of publications, across fields, document types and also through the years. The main result of the study is that less than 50% of the publications have some kind of altmetrics. The source that provides most metrics is Mendeley, with metrics on readerships for around 37% of all the publications studied. Other sources only provide marginal information. Possibilities and limitations of these indicators are discussed and future research lines are outlined. We also assessed the accuracy of the data retrieved through Impact Story by focusing on the analysis of the accuracy of data from Mendeley; in a follow up study, the accuracy and validity of other data sources not included here will be assessed.

Conference Topic
Topic 1 Scientometrics Indicators: criticism and new developments; Topic 2 Old and New Data Sources for Scientometrics Studies: Coverage, Accuracy and Reliability and Topic 7 Webometrics

Introduction
Social media are increasingly investigated by information scientists and will remain an important research theme in the near future (Wang, Wang & Xu, 2012). The development and increasing use of the tools has created new challenges for research and many scholars have begun to investigate the impact of social-networking sites on scholarly communication. There is a growing interest in tracking and measuring scholar's activities on the web, through the use, development and combination of new methods and indicators of research with other more traditional impact metrics and web-based alternatives such as webometrics, cybermetrics, and recently social web analysis or Altmetrics (Priem et al., 2010; Wouters & Costas, 2012). Citation analysis is a popular and useful

measurement tool in the context of science policy and research management. Citations are usually considered as a proxy for 'scientific impact' (Moed, 2005). However, citation analysis is not free of limitations, and the need for alternative metrics to complement previous indicators has become an object of many studies. Researchers have explored and made use of other metrics (such as log analysis, usage counts, download and view counts, webometrics analysis, etc.) (Haustein, 2012, Thelwall, 2008 & Thelwall, 2012) to overcome the weakness of traditional impact measurement.

An important approach is "altmetrics" which was introduced in 2010 (Priem, et al., 2010) as a novel way of "assessing and tracking scholarly impact on social web", to enhance the process of measuring scholarly performances. In recent years, there has been a growth in the diversity of tools (and also companies) that aim to track 'real-time impact' of scientific outputs by exploring the sharing, reviews, discussions, bookmarking, etc. of scientific publications and sources. Among these tools and companies are F1000 (http://f1000.com/), PLoS article-level-metrics (ALM) (http://article-level-metrics.plos.org/), Altmetric.com (http://altmetric.com/), Plum Analytics (http://www.plumanalytics.com/), Impact Story[71] (http://impactstory.org/), CiteULike (http://www.citeulike.org/), and Mendeley (http://www.mendeley.com/).

Objectives

This paper builds upon Wouters & Costas (2012). Our general research objective is to explore whether the new metrics allow for the analysis of more dimensions of impact than is currently possible through citation analysis and what kind of dimensions of scientific activity or performance might be represented by the new web based impact monitors. In exploring these issues, we pursue the following research questions:

1) What is the accuracy and validity of the data retrieved by Impact Story (IS) from Mendeley? Are there any limitations to take into account when using this tool?

2) What is the presence of altmetrics across scientific fields and document types?

3) What is the potential of altmetrics in measuring research performance? What are the relationships between altmetrics and citation indicators?

Research design and methodology

We have focused on IS. Although still at an early stage ('beta version'), IS is one of the current web based tools with more potential for research assessment purposes (Wouters & Costas, 2012). IS aggregates "impact data from many sources and displays it in a single report making it quick and easy to view the impact of a wide range of research output" (http://impactstory.org/faq). It takes as

[71] Previously known as Total Impact. For a review of tools for tracking scientific impact see Wouters & Costas (2012), we use IS in this study to refer to Impact Story.

input different types of publication identifiers (e.g. DOIs[72], PubMed[73] ids, URLs[74], etc.); which are run through different external services to collect the metrics associated with a given 'artifact' (e.g. a publication); thus a final report is created by IS and shows the impact of the 'artifacts' in different indicators. Using NEW ID () query in SQL, a random sample of 20,000 publications with DOIs (published from 2005 to 2011) from all the disciplines covered by the Web of Science (WoS) has been collected. Using IS, these DOIs were entered into the system and the metrics were collected and saved in CSV format for further analysis[75]. The result table was matched with the CWTS in-house version of the Web of Science on the DOIs (and their altmetric values) to be able to add other bibliometric data to them. Given some mistakes in the table (i.e. missing DOIs from the output coming from IS and also some documents that changed in the meantime in the WoS database) the final list of publications resulted in 19,722 DOIs. Based on this table, we studied the distributions of altmetrics across fields and document types. Citation and collaboration indicators were calculated and the final files were imported in IBM SPSS Statistics 19 for further analysis.

Analysis of the accuracy of the data retrieved by IS

In this section we present the result of a manual check on the altmetrics provided by IS, particularly regarding their accuracy with the data from Mendeley. Thus, according to Krejcie, & Morgan, (1970), with 95% confidence level, the minimum required sample size for 19722, is 377 observations; therefore, 377 DOIs[76] were selected for manually checking in order to see whether each DOI retrieved refers to the same publication in Mendeley and the same metrics are collected. We found that 208 items had exactly the same scores as before, 154 presented an increase in readerships (which can be explained by the time lag between the download of the data from IS and the manual check) and 4 with

[72] DOI (Digital Object Identifier) is a unique alphanumeric string assigned by the International DOI Foundation to identify content and provide a persistent link to its location on the Internet (http://www.doi.org/)

[73] PubMed comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books (http://www.ncbi.nlm.nih.gov/pubmed)

[74] Uniform resource locator (URL) is a specific character string that constitutes a reference to an Internet resource (http://en.wikipedia.org/wiki/Uniform_resource_locator)

[75] Another important element of the data downloaded from IS is that only the first 3 columns (TIID: Total Impact Identifier, Title and DOI) of the CSV files are the same and in the same position, while the other columns are different depending on the values/metrics that they contain (e.g. if in a set of publications only Mendeley and CiteULike metrics are present for the items, only these two columns of metrics would appear, while if in a second search other metrics appear like for example Twitter, then a third column would be added to the field). This situation created the problem that the different files presented a different column distribution, making the merging of all the files in one final table more problematic. The CSV files were uploaded into Google Spread sheets and downloaded back as Plain Text. A SAS program was used to merge all the files together and put them in the template made previously.

76 Altmetrics retrieved by IS contains two parts: metrics found and not found. We decided to select a small sample (two sets of 5 items from each part) to check for the accuracy of data retrieved; therefore both DOIs with and DOIs without metrics were checked

decrease in readership counts, 2 items were not found, for 6 items it wasn't possible to get the readership scores and 3 mistaken items were found[77].
Since most of information is entered by users in Mendeley and not all items have DOIs or some may have incorrect DOIs; the title searches of the 377 DOIs were also done in Mendeley in order to see if there are any differences between the DOI and Title search regarding each publication. The result showed that only 10 items can't be found by their titles although they are saved in Mendeley and can be retrieved by their DOIs/Pub Med IDs through IS and only for 2 cases there were metrics through their titles but not through their DOIs. In general, these results suggest that for this sample, the data from Mendeley retrieved through IS is quite reliable although there are some limitations in Mendeley (see Bar-Ilan, 2012)[78] which have to be taken into accounts when checking the data.

Table 1. Presence of IS altmetrics from all data sources across publications

Data Source	*papers with metrics*	*%*	*papers without metrics*	*%*
Mendeley readers	7235	**36.7**	12487	**63.3**
PubMed pmc citations	2593	13.1	17129	86.9
CiteULike bookmarks	1638	**8.3**	18084	91.7
PubMed pmc citations reviews	929	4.7	18793	95.3
Wikipedia Mentions	270	1.4	19452	98.6
Facebook likes	142	0.7	19580	99.3
Topsy Tweets	95	0.5	19627	99.5
PubMed pmc citations editorials	55	0.3	19667	99.7
Facebook shares	57	0.3	19665	99.7
Facebook comments	42	0.2	19680	99.8
Delicious bookmarks	33	0.2	19689	99.8
Topsy influential tweets	18	0.1	19704	99.9
PlosAlm_pmc_full_text	1	0.0	19721	99.9
PlosAlm _pmc_abstract	1	0.0	19721	99.9
PlosAlm_pubmed_central	1	0.0	19721	99.9
PlosAlm _pmc_pdf	1	0.0	19721	99.9
PlosAlm_pmc_supp_data	1	0.0	19721	99.9
PlosAlm _pmc_unique_ip	1	0.0	19721	99.9
PlosAlm _pmc_figure	1	0.0	19721	99.9
PlosAlm _html_views	1	0.0	19721	99.9
PlosAlm _pdf_views	1	0.0	19721	99.9
PlosAlm _scopus	1	0.0	19721	99.9
PlosAlm _crossref	1	0.0	19721	99.9
PlosAlm	1	0.0	19721	99.9
Facebook clicks	16	0.1	19706	99.9

[77] The DOIs retrieved for these 3 were different from the DOIs entered; thus pointing to different articles.
[78] Sometimes, publications can not be found in Mendeley because the titles are not entered correctly by the users; and there are also some duplicates records for a single publication with different number of readerships in Mendeley.

Results and main findings

Table 1 shows the frequencies and percentage of all altmetrics data retrieved by IS (with the only exception of F1000 that has been left out of this study as they are only available for medical journals and with a yes/no value). Most of the metrics present a very low frequency in our sample, mainly all the PlosAlm indicators as they are only available for the PLoS journals and their presence in our sample is negligible.

Considering table 1, our main finding is that, with the exception of Mendeley, the presence of metrics across publications and fields is very low. Clearly, their potential use for the assessment of the impact of scientific publications is still limited. Based on Table 1, we decided to remove some of the metrics from our study: PlosAlm due to their low frequency and PubMed-based indicators because they are limited only to the Health Sciences and they refer to citations, which we will calculate directly. We also decided to sum the metrics coming from Facebook (i.e. Facebook likes, shares, comments, clicks) given their high correlation (Priem, Piwowar, & Hemminger, 2012) and their relatively low frequency and due to exceeding the downloading limit of IS at the time of data collection, we excluded data from Twitter since it was not reliable.

The presence of altmetrics across fields

Figures 1 and 2 show the distributions of altmetrics across major fields of science and document types. The altmetrics presence did not vary much by publication year. Multidisciplinary journals ranked highest in almost all metrics. The major source for altmetrics data in our sample is **Mendeley** with the highest readership from Multidisciplinary fields (55.1% of the publications in this field have at least one Mendeley reader). In **Wikipedia,** Multidisciplinary fields (6.5%) ranked the highest as well.

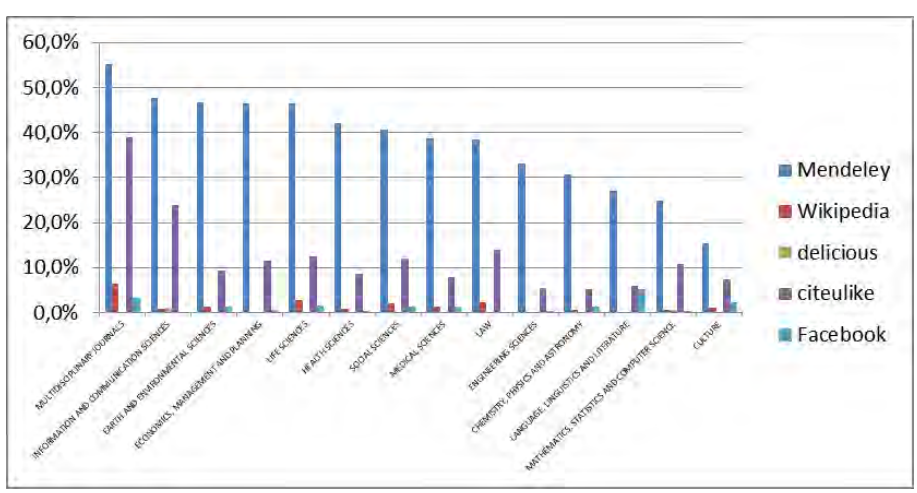


Figure 1. Distribution of altmetrics across fields

The presence of altmetrics across document types

Regarding document type, there are 16888 (84.6%) articles, 946 (4.79) review papers, 488 (2.47%) letters and 1600 (8.11%) non-citable[79] items in the sample. According to figure 2, around half of (49.6%) the review papers and 40% of articles in the sample have readerships in the Mendeley. With the exception of Delicious, which has a negligible presence, review papers have proportionally attracted more metrics than other document types in our sample, although the number of review papers in our sample is smaller than the number of articles.

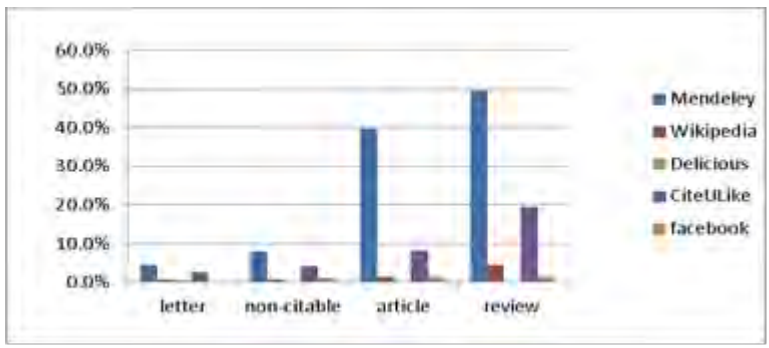


Figure 2. Distribution of altmetrics across document types

Table 2. Spearman's Correlation among variables

	Mendeley readers	*Facebook*	*delicious bookmarks*	*CiteULike bookmarks*	*pub year*	*Number of references*	*Number of authors*	*Number of institutes*	*Citations*	*self Citations*
Wikipedia mentions	0.05	0.021	0.017	0.089	-0.041	0.063	0.019	0.016	0.097	0.073
Mendeley readers		0.033	0.007	0.171	-0.003	0.298	0.111	0.098	0.307	0.195
Facebook			0.098	0.009	0.058	0.019	0.015	0.008	-0.002	-0.005
Delicious bookmarks				0.024	0.005	0.011	-0.01	-0.002	0.006	0.002
CiteULike bookmarks					-0.015	0.152	0.003	0.033	0.185	0.119
pub_year						0.045	0.034	0.034	-0.431	-0.268
n_refs							0.142	0.149	0.407	0.313
n_authors								0.467	0.251	0.24
n_institutes									0.142	0.154
Citations										0.692

Relationships between altmetrics and bibliometric and citation indicators

In this section we studied the relationship between the main altmetric indicators and citation indicators, publication year, and collaboration indicators (the number of authors and institutions in the papers). For the calculation of the total number of citations we have used a variable citation window (i.e. citations up to the date). Self-citations have been identified for the different publications and introduced in

[79] non-citable document type corresponds with all WOS document types except article, letter and review

the study as a separate variable. The relationships among altmetric and bibliometric indicators were investigated using Spearman correlation coefficient since the data were skewed (table 2). Concerning citations, we found moderate (r=.3) and small (r=.18) correlations with Mendeley and CiteULike. It is remarkable that Facebook is the source with the lowest correlations with all the other indicators, thus suggesting that this indicator could be related with other types of impact not related to scholarly impact (i.e. measured through citations).

Conclusions and Discussions

This study shows that IS, although being in an initial stage of development (it is still in a 'Beta' version), is an interesting source for aggregating altmetrics from different sources. However, we also see important limitations particularly regarding the speed and capacity of data collection and formatting of the data. Out of 19,722 publications 7235 (36.7%) had at least one reader in Mendeley, which is considerably a lower share of Mendeley coverage as compared to previous studies such as 97.2% for JASIST articles published between 2001 and 2011 (Bar-Ilan, 2012); 82% coverage of articles published by researchers in Scientometrics (Bar-Ilan et al., 2012); 94% and 93% of articles published in Nature and Science journals in 2007 (Li, Thelwall and Giustini, 2012); and more than 80% of PLoS ONE publications (Priem et al 2012), followed by 1638 (8.3%) publications bookmarked in CiteULike. Previous studies also showed that Mendeley is the most exhaustive altmetrics data source (Bar-Ilan et al., 2012, Priem et al., 2012). Correlation of Mendeley readerships with citation counts showed moderate correlation (r=.30) between the two variables which is also found in other previous studies (Bar-Ilan, 2012; Priem et al., 2012; and Shuai, Pepe & Bollen, 2012). This indicates that reading and citing are different scientific activities. Multidisciplinary fields (i.e. the field where journals such as *Nature*, *Science* or the *PNAS* are included) attracted more readerships. Review articles were proportionally the most read, shared, liked or bookmarked format compared to articles, non-citable and letters in Mendeley. This may be evidence for the specific role of this document type in dissemination of scientific knowledge.

The main result of this study is that the presence of altmetrics is not yet prevalent enough for research evaluation purposes. As indicated in table 1, in our sample, except in Mendeley (63% of publications without metric), in all other data sources more than 90% of the publications are without any metric; thus less than 50% of all publications in this study showed some altmetrics. The amount of altmetrics is still quite low, and given these low numbers problems of validity and reliability could appear when used for real and broad research assessment purposes. For this reason, it is still too soon to consider altmetrics for robust research evaluation purposes, although they already present an interesting informative role. Previous studies also discussed that altmetrics may be useful for the research impact measurement but not proven yet (Li, Thelwall & Giustini, 2012) and in order to

be regarded in this context, they need to meet the necessary requirements for data quality and indicator reliability and validity (Wouters & Costas, 2012).

Acknowledgements

Special thanks to Erik van Wijk from CWTS for doing the programming part and Jason Priem and Heather Piwowar from Impact Story who gave us very practical and useful hints during the data collection, and special gratitude to Professor Mike Thelwall from Wolverhampton University for his valuable comments on this paper.

References

Bar-Ilan, J. 2012. *JASIST@ Mendeley*. Presented at ACM Web Science Conference Workshop on Altmetrics, Evanston, IL, 21 June 2012. Retrieved January 5, 2013 from: http://altmetrics.org/altmetrics12/bar-ilan/

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H., & Terliesner, J. (2012). Beyond citations: Scholars' visibility on the social Web. *In Proceedings of the 17th International Conference on Science and Technology Indicators,* Montreal, Quebec. Retrieved December 5, 2013 from http://arxiv.org/abs/1205.5611/

Haustein, S. (2010). Multidimensional journal evaluation. *Proceedings of the 11th International Conference on Science and Technology Indicators* (pp. 120–122), Leiden, the Netherlands.

Krejcie, R.V. & Morgan, D.W. (1970) Determining sample size for research activities. Educational and Psychological Measurements, 30, 607-610.

Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics*, 91(2), 461–471.

Moed, H.F. (2005). Citation analysis in research evaluation. *Berlin/ Heidelberg/New York: Springer.*

Priem, J., Taraborelli, D., Groth, P., and Neylon, C. (2010). *Altmetrics: a manifesto*. Retrieved December 10, 2012 from: http://altmetrics.org/manifesto/

Priem, J., Piwowar, H., & Hemminger, B. H., (2012). *Altmetrics in the wild: Using social media to explore scholarly impact.* ArXiv: 1203.4745v1

Shuai X, Pepe A, Bollen J (2012). *How the scientific community reacts to newly submitted preprints: article downloads Twitter mentions, and citations*. Retrieved 25 December 2012 from: ArXiv: 1202.2461v1201

Thelwall, M. (2008). Bibliometrics to Webometrics, *Journal of Information Science*, 34(4), 605-621.

Thelwall, M. (2012). Journal impact evaluation: A webometric perspective, *Scientometrics*, 92(2), 429-441.

Wang, X. W., Wang, Z., Xu, S. M. (2012). Tracing scientist's research trends realtimely. *Scientometrics*, Retrieved 5 January 2013 from: http://arxiv.org/abs/1208.1349

Wouters, P., Costas, R. (2012). *Users, narcissism and control – Tracking the impact of scholarly publications in the 21st century*. Utrecht: SURF foundation. Retrieved September 20, 2012 from: http://www.surffoundation.nl/nl/publicaties/Documents/Users%20narcissism%20and%20control.pdf

INTERMEDIATE-CLASS UNIVERSITY RANKING SYSTEM: APPLICATION TO MAGHREB UNIVERSITIES (RIP)

Hamid Bouabid, Mohamed Dalimi, Mohammed Cherraj

Faculty of Science, Mohammed V - Agdal University
4, Avenue Ibn Battouta B.P. 1014 RP, Rabat, Morocco.

Abstract
Most of the world-class university ranking systems are still criticized due to their heterogeneous approach in ranking universities from quite different ecosystems and countries. On the other side, it is obvious that ranking systems of national-class do not bring any competition or contest among universities toward excellence and outstanding quality. This paper shows that an intermediate-class ranking, which is above a national-class and below a world-class, is appropriate to ensure coherence and similarity of a national class and to attenuate bias that raise through heterogeneous and diverse universities in world-class rankings. Furthermore, intermediate-class ranking reasonably ensures competing for excellence and quality that may vanish in a national-class. A hybrid ranking system is built to satisfy intermediate-class requirements. The hybrid system is composed of mixed indicators out of which some are taken from existing and proven world-class ranking systems (adopted or adapted), and others are introduced to fairly take into account the level of development of higher education and research of the countries considered, offering more stability and objectiveness of the ranking system.

Keywords
Ranking; university; country; intermediate; hybrid; world-class; national-class; Maghreb.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3) and Management and Measurement of Bibliometric Data within Scientific Organizations (Topic 9).

Introduction
Several studies since the beginning of the seventies have been done to build university rankings. Their main objective was to provide students, particularly future students, and their parents with information to be able to choose a university and curricula. Since then ranking systems have proliferated and targeted to attract more talented students and academic staff. In spite of their wide use, university ranking systems continue to be criticized (Enserink, 2007, Stolz, 2010, Taylor and Braddock, 2007). Enserink concluded by quoting A. Einstein '*Not everything that counts can be counted, and not everything that can be*

counted counts'. In their analysis of correlation of university ranking and excellence, Taylor and Braddock suggested rankings not far than an informed approach to look carefully at its criteria and then consider them as a tool of achievement with respect of those criteria. Even though, the exploitation and use of universities rankings have been exacerbated by globalization and are more and more used by decision makers as well as scientists themselves. Finally, ranking systems serve as a tool in teaching or research evaluations and accountability requirements.

University ranking systems are of two categories: national-class and world-class. Examples for national-level are: US News & World Report and Princeton in USA, McLean's in Canada, CHE in Germany, Excelencia in Spain, La Repubblica in Italy, Education 18 in China, League Table, Daily Telegraph, The Guardian, in UK, Good Universities Guide and Melbourne Institute in Australia, SwissUp in Switzerland, etc. For world-class the most known ranking systems are: Times Higher Education, Leiden, Webometrics, US News & World Report (world's best universities) and Academic Ranking of World Universities. These ranking systems consist of ranking 500 or 1000 (or more) of world universities using a common set of indicators. W*orld-class* ranking systems are of a greater quality, outstanding excellence and highly demanding. These requirements may be so to fulfill the challenges of an internationally open space of research and higher mobility of scientists and human resources. Should a national-class ranking system be rational, since the competition is among universities of the same ecosystem and under the same rules, similar objectives and common language, a world-class is not as appropriate since it ranks universities from heterogeneous ecosystems that are ruled differently and mostly have divergent objectives. Even if these shortcomings might be mitigated by globalization and other bilateral or multilateral agreements, they remain some of the major weaknesses of world-class ranking systems and of a high concern in presenting their results.

Research activity is believed to be one of the world-class concerns. However, some researchers have found that even national-class rankings are dominated by research as do world-class rankings. In fact, Van Dyke (2005), when comparing academic institutions on a national basis, found that 75% of examined national rankings affect almost all the weight to research quality (research/prestige) rather than teaching. Van Raan *et al.* (2011) found a severe effect of language - non English - on universities ranks particularly those of bibliometric-based indicators. However, ranking according to the same indicators both English and non English speaking universities (or countries) is not the only bias for the world-class ranking systems. Buela-Casal *et al.* (2007) pointed out the methodology of the ranking beside the choice of the indicators weights, which lead to huge variability of ranking. Florian (2007) has shown that the results of the *ARWU* ranking cannot be reproduced given raw data and its public methodology. Compiling scores for universities from several countries lacks reproducibility, which is essential for example to compute score for any university that is not in top 500 or to build up a

policy to aspire entering the top 500. *ARWU* in spite of its world recognition does not qualify as a tool for quality of academic institutions (Billaut *et al.* 2010).
In *THE* ranking system, reputational survey counts for 34,5% of the total score of the ranking (15% for teaching, 19,5% for research). Beyond its qualitative scoring, the reputational surveys are senseless in a highly heterogeneous set of countries, because the perception of 'reputation' may widely differ from a country's community to another and also because the priorities and the strategies differ from a higher education system to another. In fact, it may be considered as a reputational research if it addresses tropical diseases in some countries and not issues of aging society and vice-versa. Technology transfer may be of a high reputation in research in some countries and oppositely may be of no recognition where even patenting is almost absent. The same different perception of reputation may be toward green research where research itself is used for economy industrialization in some countries. In teaching reputation, expectations for only employment may be as much important in some countries as university graduates being able to find a job immediately after graduation. In other countries, priorities are even lesser or higher than these expectations, for example to serve international market and mobility - as does the *THE* ranking system- or gain awards and prizes, which substantially influence opinions and then scores attributed to each university.
For the *Webometrics* ranking system, the main objective is to improve information technology use within universities and higher education system. However, ranking of universities depends fundamentally of the level of IT infrastructure of their respective countries. Universities are part of the countries IT ecosystems in which they evolve and consequently their ranking is not really that of the university itself but that of the country.
In spite of the *Berlin Principles* (BPs) that serve as a charter and deontology in building and producing ranking systems, particularly in world-class ranking systems, these would not overcome weakness of world-class heterogeneous characteristics. In their interesting work to rank 25 European ranking systems, almost all national-class rankings, according to the BPs, Stolz *et al.* (2010) have shown that no ranking system achieves good overall congruence with the BPs and almost all failed in methodology relevance to BPs. As a result, national ranking systems may be perceived as unqualified ranking systems perhaps of their confinement to a national level where there are fewer constraints of competition and less pressure of research, educational, political contexts than there may be on higher education institutions in a world -class. Even though the BPs insist in their 3rd and 5th criteria to take both *recognizing the diversity, different missions and goals of institutions, the national higher education background and context*, into account in a ranking system, Cheng and Liu (2008) have found that the two criteria could not be well implemented by ranking practices.
In spite of the world-ranking systems' weaknesses, several researchers and specialists agree that these rankings are a useful quantitative tool in analyzing and improving higher education and research (Dalsheimer et Despreaux, 2008). The

objective is then how to take benefit of these rankings while reducing as much as possible inherent bias. Aguillo *et al.* (2010) have compared some world-class ranking systems: *ARWU*, *Webometrics*, *THE*, *Taiwan-Ranking* and *Leiden*, using three techniques. They found that similarities between these world-class ranking systems increased when the comparison is limited to the European universities regional level rather than world level of these ranking systems. It is obviously because of many shared factors by European countries such as higher education regulatory, history of the research and education, programs of mobility among academic staff and students of European universities (Erasmus Mundus, FPs, EUREKA, etc) that should result in a convergence of the countries higher education systems in spite of their differences.

This finding supports our proposal of an intermediate-class where diversity is preserved as well as reasonable and rationale ranking. It then allows to avoid any divergence of methodology or object covering from one side, and to allow useful ranking that serve for positive competing for excellence and quality that goes inevitably beyond national level. A ranking in a meso-level guarantees both diversity which is strongly required for a competing culture and a coherent ranking that is really useful and objective. An intermediate-class ranking system is then required to satisfy the meso-level requisites.

Intermediate-class

In this article, the Intermediate-class will be the Maghreb universities. These universities share a lot of determinants strengthened by the several countries common factors. In their work on research activities in Africa from 2004 and 2008, Adams et al. (2010) have used co-publications between African countries and observed 4 emerging clusters. One of the clusters is composed of Maghreb countries (Morocco, Algeria, Tunisia and Libya) and Egypt. In addition, Toivanen and Ponomariov (2011) analyzed African innovation systems and their *intra-muros* collaborations and noticed the existence of 3 networks: South-East, North and West, which have quite different patterns and research collaboration characteristics as well as high distinctive internal dynamics. Even if Egypt is part of the Northern network, its internal research collaboration dynamics are less intensive with the other countries of the North network: Tunisia, Algeria and Morocco.

Furthermore, the Maghreb countries share other factors like the history, the language in their higher education systems (French language), geography, educational background (ancient Arabo-Islamic Empire and later French colonization). Higher education history in the Maghreb countries is centuries deep in time. Alqaraouine University was founded in Fez, Morocco in the 9th century. It is the oldest degree-awarding institution in the world operating until now (Adams *et al.*, 2010). The modern universities in the Maghreb were founded under the French colonization or just after the independence under cooperative programs and benefited from academic staff expatriation.

Hybrid ranking system

The systemic and arbitrary application of a world-class, assumed to be of high criterion of 'excellence' and 'quality', to Maghreb universities will completely be biased at the moment. In fact, we have applied *ARWU* ranking system to well research-advanced universities in Maghreb (since *ARWU* is almost fully research-based ranking). These universities were chosen by taking one university from each country: Houari Boumediene from Algeria, Cadi Ayyad from Morocco and Sfax from Tunisia. To prove the finding we have compared these universities to the last one in the top 500 *ARWU* ranking, namely York University in Canada. Table 1 shows clearly the gap yet existing between the York University and the best Maghreb universities, that is to say the unreadiness of Maghreb universities to compete for a world-class ranking system such as *ARWU*. Further, publishing and citing patterns and characteristics have been found to be distinctively different from developed and developing countries (Bouabid and Larivière, 2013). Table 1 shows also that selected Maghreb universities and York University do compete only for 70% of *ARWU* ranking system because none of them has a score in the first (Alumni of the institution winning Nobel Prizes or Fields Medals) and second (Awarded Staff of an institution Nobel Prizes or Fields Medals) indicators that weigh together 30% of the ranking.

Table 1: ARWU raking applied to Maghreb Universities and York University

University	Alumni	Award	HiCi	N&S	Pub	PCP	Total score
York (CA)	0,00	0,00	0,00	100,00	100,00	100,00	44,00
Cadi Ayyad (MA)	0,00	0,00	100,00	2,02	6,53	47,11	26,42
Sfax (TN)	0,00	0,00	100,00	0,00	12,36	24,09	24,88
STHB (DZ)	0,00	0,00	0,00	0,00	7,71	2,62	1,80

Maghreb universities do not qualify for these two indicators for which York University does with even other universities like San Fransisco university ranked 18th which obtained a zero score in the first indicator. However, one could notice from Table 1 that the gap for other indicators between York University and Maghreb universities is also wider even if interestingly two of the Maghreb universities have each a Highly-Cited (*Hi-Ci*) researcher unlike York University. Results presented in Table 2 demonstrate again that a world-class ranking is also far to be used to rank universities - and in general higher education - inputs and outputs or provide accurate comparative scale for universities in the region due may be to their performance and not to the ranking system. Maghreb universities has a paper in *Nature* and *Science* Journals during the period from 2005 to 2009. Further, the scores for published papers in Web of Knowledge are much fewer (12,36 in best case) than that obtained by York university. It is the same case for the per capita academic performance (obtained scores scaled to academic staff:

PCP). For the *PCP* indicator, the score of York University is twice the best score gained by Maghreb universities (47,11).
Again, to prove that a world-class ranking is still far to be fairly used for ranking Maghreb universities, we have applied the *Leiden* ranking system. Indeed, since Tunisia has the highest increase rate of published papers and impact in the region (using *SCI* data, *WoS*), average 47% per year from 1997 (535 papers) to 2009 (3534 papers), we apply the *Leiden* ranking system to assess the difference in score between Tunisian universities and European (western) universities. With a very fewer researchers than Morocco, Tunisia got more than 70 projects of the 7th call of FPRD promoted by the EU with an amount of 7,4 Million Euros, compared to Morocco with 73 projects totalizing an amount of 8 Million Euros.
Only universities with more than 100 papers during the period of 2007 to 2009 are considered in the ranking. According to Leiden ranking (orange) the scores of Tunisian universities[80] are yet far from those obtained by European universities, due may be to their overall research productivity and quality than to the Leiden ranking system itself.

Table 2: Leiden ranking (c/p) applied to Tunisian universities

University	*Medicine and pharmacy*	*Chemistry*	*Physics and Maths*	*Geology*	*Biolog and Agriculture*	*Engineering and technology*	*Social Science and Humanities*	*Total score*
Tunis	2,46	0,73	0,48	0,37	0,69	0,29	0,19	0,74
Sfax	0,70	0,64	0,33	0,60	0,64	0,73	0,58	0,60
Carthage	0,46	0,49	0,33	0,56	0,53	0,47	1,09	0,56
Elmanar	0,50	0,57	0,38	0,47	0,44	0,65	0,49	0,50
Manouba	0,69	0,54	0,11	0,70	0,58	0,23	0,61	0,49
Sousse	0,32	0,62	0,38	0,0	0,65	0,79	0,38	0,45
Monastir	0,62	0,52	0,43	0,0	0,55	0,52	0,29	0,42

The last European (western) university is Tech University-Madrid (Spain) getting a score of 0,80 (2011/2012 *Leiden* Table). It is worth reminding that the difference between universities in *Leiden* ranking is in most cases less than 0,01.
The hybrid ranking system to be built aims at contributing to providing an information tool, stimulating competition among Maghreb universities and contributing to prepare Maghreb universities to world-class visibility. The ranking system built up is hybrid in the sense that it comprises indicators which properly reflect the level of development of higher education in these countries and also other selected indicators from existing world-class ranking systems to help in rising university's quality standards by intra competition and also improving their scores in these global rankings. This mix is to contribute to enhancing research and teaching activities within the Maghreb universities aspiring *in-fine* to be visible in world-class ranking tables in the future. The hybrid ranking system preserves as much as possible similarity within the set of universities of the intermediate-class to be ranked and prepares them for world-class in the future.

[80] for the ranking we aggregate the scientific fields into 7 broad categories (see Table 2).

Indeed, the hybrid ranking system is composed of two major criteria: Teaching and Research. Teaching counts for 35% of the whole ranking score and Research for 65%. The indicators and weighing rates are given in the Table 4.
Research and knowledge production is unavoidably the main field of competition and university recognition, which is why a higher weight was attributed to research criterion and that's also why a special indicator for papers is introduced.

Table 4: Weighting scheme of the Intermediate-class ranking system

Criteria	Indicator	Code	Weight
Research	Number of papers in WoS	Pub	20%
	Number of citations/paper	Cit/Pub	15%
	% papers/academic and research staff	Pub/Staf	20%
	Number of papers in *Nature* and *Science*	N&S	10%
Teaching	% academic and research staff/total students	Staf/Stud	10%
	% of PhD graduates/academic and research staff	PhD/Staf	15%
	% foreign students/total students	For/Stud	10%
Total			**100%**

The indicators are:

Pub: Number of papers indexed in WoS from 2007 to 2009. Three-year period is done to reduce fluctuation effect from a year to another on a score and obtain more stable score.

Cit/Pub: Ratio of the number of citations to the number of papers published between 2007 and 2009. The time window of 4 years is considered to count citations: 2007-2010.

Pub/Staff: Ratio of published papers in WoS scaled against the number of academic staff. The indicator gives an idea of the normalized research outputs to the size of the university.

N&S: Number of papers in Nature and Science between 2005 and 2009. The same methodology for counting in ARWU is used here.

Staff/stud: Ratio of the number of permanent academic staff to the total number of students during the previous year. It is believed that teaching quality is correlated to higher ratios.

PhD/Staff: Number of PhDs awarded by a university the year before scaled against the number of academic staff, to obtain score unbiased by the size. The ratio is a sign of a university's attractiveness to high level research graduation offered by its academic staff.

For/Stud: Ratio of the number of foreign students to total number of students of a university. Since the enrollment of foreign students in Maghreb universities is some times government-ruled under quota, the indicator weights just 10%. This indicator aims to contribute to raising awareness regarding the openness of a university to international knowledge space.

Table 5: Type of indicators in the hybrib ranking system

Indicator	Origin	Weight
Number of papers in WoS	proper	50%
% papers/academic and research staff	proper	
% academic and research staff/total students	proper	
Number of papers in *Nature* and *Science*	Adapted (ARWU)	25%
% of PhD graduates/academic and research staff	Adapted (THE)	
Number of citations/paper	Adopted (Leiden)	25%
% foreign students/total students	Adopted (THE)	

From Table 5, one could find that the proper indicators cumulate half of the weight to reflect properly the Maghreb context and its higher education and research ecosystem. These weights were chosen after several simulations and prior tests. The other half of the ranking weight is equally divided between adapted and adopted indicators from known and world-class ranking systems. These indicators are more to contribute to promoting and raising Maghreb universities to world-class visibility.

Ranking results

When a datum is not available its weight is fairly affected to other indicators (which happens in a very few cases). A university was excluded from the hybrid ranking table if the number of its papers is less than an average of 20 per year. Data sources (Staff, students, PhD graduates, etc) are mainly from Ministries and Universities and retrieved from their respective websites. The numbers of papers and citations are extracted from the *WoS* database.

Table 6 shows that 6 Tunisian universities are ranked in top 10 on the hybrid ranking. This result is not unexpected since the hybrid ranking system is at least 65% research-based weighing and that Tunisia is the most dynamic country in research activities in the Maghreb. Algerian universities scores are on the whole the lowest among Maghreb universities.

Table 6: Ranking scores of the Intermediate-class ranking system of Maghreb universities

Rank	University	Country	Pub	Cit/Pub	Staff/Stud	For/Stud	PhD/Staff	N&S	Pub/Staff	Total
1	Sfax	TN	100,0	78,4	47,6	10,2	75,2	0,0	100,0	68,8
2	Houari Boumediène	DZ	78,3	44,2	62,7	0,0	76,2	0,0	72,4	60,1
3	Mohammed V-Agdal	MA	44,9	47,8	61,4	80,1	100,0	0,0	67,6	58,8
4	Carthage	TN	95,5	58,9	52,8	6,6	17,7	40,0	81,9	56,9
5	Tunis el Manar	TN	76,9	69,6	87,9	11,5	57,6	40,0	38,7	56,1
6	Cadi Ayyad	MA	62,9	61,5	49,0	21,0	26,0	40,0	76,4	52,0
7	Monastir	TN	65,9	67,1	90,5	0,0	30,9	0,0	48,9	46,7
8	Sousse	TN	70,4	55,6	77,2	6,3	26,5	0,0	52,8	45,3
9	Chouaib Eddoukali	MA	14,1	80,5	77,8	59,0	19,2	0,0	46,2	40,7
10	Tunis	TN	29,0	71,3	53,4	3,2	64,3	0,0	35,1	38,8
11	Ibn Tofail	MA	15,1	40,8	46,7	46,3	55,8	0,0	52,9	37,4
12	Abdelmalek Essaadi	MA	16,7	58,4	39,8	45,5	55,7	0,0	34,7	35,9
13	Hassan II Aîn Chock	MA	21,8	68,3	52,5	45,5	29,6	0,0	30,8	35,0
14	Essenia	DZ	20,8	37,2	45,4	19,5	22,3	100,0	17,6	33,1
15	Djillali Liabes	DZ	28,4	55,8	40,0	21,7	25,3	0,0	43,9	32,8
16	Mohammed V-Souissi	MA	6,6	44,2	100,0	100,0	15,0	0,0	7,9	31,8
17	Mohammed Premier	MA	16,6	51,5	33,3	33,5	44,0	0,0	37,1	31,8
18	Abdelhak Benhamouda	DZ	15,6	59,4	38,9	0,0	28,3	0,0	36,4	30,5
19	Ferhat Abbas	DZ	34,7	49,0	36,1	8,4	0,0	0,0	35,3	29,9
20	Mohamed Boudiaf	DZ	17,3	53,2	82,2	8,0	7,2	0,0	33,0	28,2
21	Sidi Med Ben Abdellah	MA	14,2	47,3	28,4	31,6	57,2	0,0	17,8	28,1
22	Mentouri	DZ	49,7	32,6	38,1	0,0	40,9	0,0	30,2	27,2
23	Hassan II Mohammadia	MA	8,9	58,0	42,6	33,4	34,6	0,0	17,9	26,9
24	Abderrahmane Mira	DZ	20,9	46,9	36,3	19,4	0,0	0,0	29,6	26,5
25	Moulay Ismail	MA	12,6	43,6	39,4	39,5	29,3	0,0	25,9	26,5
26	Saad Dahlab	DZ	14,6	43,9	39,2	12,9	60,1	0,0	14,0	26,5
27	Abou Bekr Belkaid	DZ	21,4	37,1	48,8	12,4	9,3	40,0	22,7	25,9
28	Gabes	TN	9,2	100,0	58,3	1,9	2,4	0,0	13,0	25,8
29	08-mai-45	DZ	6,4	41,1	54,5	25,6	46,5	0,0	13,6	25,2
30	Badji Mokhtar	DZ	25,6	27,8	43,8	0,0	74,6	0,0	23,1	20,4
31	Manouba	TN	11,8	43,1	73,9	7,5	5,8	0,0	11,8	20,2
32	Mouloud Maameri	DZ	14,4	35,2	36,4	0,0	0,0	0,0	16,2	20,2
33	El Hadj Lakhdar	DZ	15,6	36,6	33,1	0,0	0,0	0,0	15,7	20,1
34	Ibnou Zohr	MA	12,6	39,4	18,6	7,0	18,3	0,0	30,5	19,9
35	Abdelhamid Ibn Badis	DZ	10,0	41,1	37,0	9,8	10,0	0,0	18,3	18,0
36	Mohamed Khider	DZ	9,0	30,0	36,9	9,3	21,3	0,0	15,3	17,2
37	M'hamed Bougara	DZ	8,5	23,3	59,9	8,0	5,7	0,0	10,5	15,0

A test of correlation was brought on the hybrid ranking system and shows no correlation between the 7 indicators. The only partial correlation was observed between the *Pub* indicator and *Pub/Staf* indicator (Figure 1) which is likely expected. The independence of the 7 indicators puts forward the fact that they cover different missions (teaching and research) and objectives of Maghreb universities without being redundant.

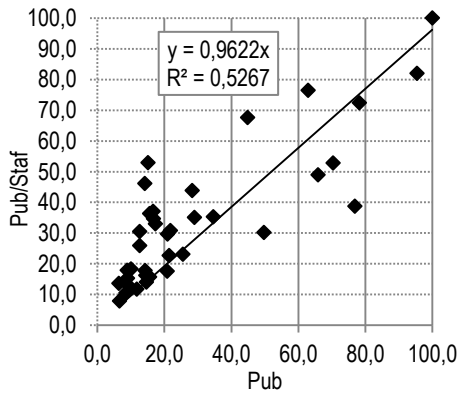


Figure 1: Correlation factor of *Pub* and *Pub/Staf* indicators

Concluding remarks

The dichotomy of national-class against world-class ranking could substantially be attenuated using an intermediate-class. Indeed, Intermediate-class is chosen for a set of countries or region where universities share many factors avoiding controversial heterogeneous world-class ranking while offering a meso-level of positively competing for excellence that goes beyond national level. The Maghreb region is a good case as an intermediate-class. We have proven that universities of this region are yet far to fairly and rationally compete for a world-class ranking system such as *ARWU* or *Lieden*, unattributed to the ranking system itself. In the case of *ARWU* the best score gained in Maghreb region is less than half the last university score in *ARWU* table. As a result, a hybrid ranking system is constructed to satisfy intermediate-class requisites and offer a reasonable and rational ranking. It is hybrid as it comprises proper indicators to reflect the meso-level's development of higher education in these countries, and other indicators adopted or adapted from existing world-class ranking systems to contribute to enhancing and rising up research and teaching standards within Maghreb universities, and to aspire being visible in world-class rankings in the future.
Intermediate-class is proposed to conciliate between national-class approach where similarity and homogeneity are present and world-class approach where competition for excellence and quality are fulfilled. Hybrid ranking system fairly satisfies Intermediate-class requirements. A sensitive mix of proper indicators and others taken from existing and proven world-class ranking systems forms the

hybrid ranking one. A correlation test among its indicators shows that they all are independent and then cover different missions and goals of universities.

References

Adams J., King C., Hook D. (2010), Global Research Report of Africa, Thomson Reuters.

Aguillo IF., Bar-Ilan J., Levene M., Ortega JL. (2010), Comparing university rankings, Scientometrics, 85:243-256.

Billaut, J. C., Bouyssou, D., & Vincke, P. (2010). Should you believe in the Shanghai ranking: An MCDM view, Scientometrics, 84(1), 237-263.

Bouabid H., Larivière V. (2013), The lengthening of papers' life expectancy: a diachronous analysis, Scientometrics (forthcoming).

Bouabid H., Martin B.R. (2009), Evaluation of Moroccan research using a bibliometric-based approach: investigation of the validity of the h-index, Scientometrics, 78, 2, 203-217.

Buela-Casal, G., Gutierez-Martınez, O., Bermudez-Sanchez, M. P., & Vadillo-Munoz, O. (2007), Comparative study of international academic rankings of universities. Scientometrics, 71(3), 349-365.

Cheng, Y., and Liu, N. C. (2008), Examining major rankings according to the Berlin Principles. Higher Education in Europe, 33, (2/3), 201-208.

Dalsheimer N. et Despréaux D. (2008), Analyse des classements internationaux des établissements d'enseignement supérieurs, Education & Formation, n° 78.

Enserink M., (2007), Who Ranks the University Rankers?, Science, 24, Vol 317, 1026 - 1028.

Florian R. V. (2007), Irreproducibility of the results of the Shanghai academic ranking of world universities, Scientometrics, 72(1), 25-32.

Stolz I., Hendel D. D., Horn A. S. (2010), Ranking of rankings: benchmarking twenty-five higher education ranking systems in Europe, Higher Education, 60, 507-528.

Taylor P., Braddock R.,(2007), International University Ranking Systems and the Idea of University Excellence, Higher Education Policy and Management, 29 (3), 245-260

Toivanen H., Ponomariov B. (2011), African regional innovation systems: Bibliometric analysis of research collaboration patterns 2005-2009, Scientometrics, 88, 471-493.

Van Dyke N., (2005), Twenty years of university report cards, Hi. Educ. Europe, 30, 103-125.

Van Raan A. F. J., Van Leeuwen T. N., Visser M. S. (2011), Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings, Scientometrics, 88:495-498.

IDENTIFYING EMERGING RESEARCH FIELDS WITH PRACTICAL APPLICATIONS VIA ANALYSIS OF SCIENTIFIC AND TECHNICAL DOCUMENTS

Patrick Thomas[1], Olga Babko-Malaya[2], Daniel Hunter[3], Adam Meyers[4] and Marc Verhagen[5]

[1] *pthomas@1790analytics.com*
1790 Analytics LLC, 130 North Haddon Avenue, Haddonfield, NJ 08033, USA

[2] *olga.babko-malaya@baesystems.com*
BAE Systems AIT; 6 New England Executive Park, Burlington, MA 01803, USA

[3] *daniel.hunter@baesystems.com*
BAE Systems AIT; 6 New England Executive Park, Burlington, MA 01803, USA

[4] *meyers@cs.nyu.edu*
New York University, Computer Science Dept, 719 Broadway, New York, NY 10003, USA

[5] *marc@cs.brandeis.edu*
Brandeis University, Computer Science Dept, MS 018, Waltham, MA 02454, USA

Abstract

This paper outlines a system designed to determine whether practical applications exist for research fields, particularly emerging research fields. The system uses indicator patterns, based on features extracted from the metadata and full text of scientific papers and patents, to assess different characteristics that point to the existence of practical applications for research fields. The system may thus help determine whether a particular research field has moved beyond the early, conceptual phase towards a more applied, practical phase. It may also help to classify emerging research fields as being more 'technological' or more 'scientific' in nature. The system is tested on data from a number of research fields across a range of time periods, and the outputs are compared to responses from subject matter experts. The results suggest that the system shows promise, albeit based on a relatively small data sample, in terms of determining whether practical applications exist for given research fields. The system also shows promise in detecting the transition from absence to existence of practical applications over time, which may be of particular value in evaluating emerging technologies.

Approved for Public Release; Distribution Unlimited

Conference Topic

Technology and Innovation Including Patent Analysis (Topic 5) and Modeling the Science System, Science Dynamics and Complex System Science (Topic 11)

Introduction

Emerging technologies are of great interest to a wide range of stakeholders. These include government agencies looking to fund promising new ideas, corporations hoping to gain a foothold in a rapidly emerging field, and investment institutions seeking returns from early investments in key innovators. Emerging technologies have also been a focus of academic research ever since Schumpeter (1912) coined the term 'creative destruction' to describe the emergence of new technologies, which spawned new industries while destroying old ones. More recently, Christiansen & Bower (1996) used the term 'disruptive technology' to describe a new development that disrupts the status quo in an existing technology.

Despite this widespread interest in emerging technologies, identifying such technologies remains problematic. The problems are both theoretical and practical. The theoretical issue is how to recognize an emerging technology, without a clear definition of what constitutes such a technology. As noted by Goldstein (1999), there is a lack of precision in the meaning of emergence, and even greater ambiguity about how it occurs. The practical issues result from the sheer scale of information available, especially in the electronic age. Researchers and analysts searching for interesting new technologies face the unenviable task of locating meaningful signals among this mass of information. Their task is not aided by the fact that the number of truly emergent technologies is dwarfed by the number of mature, mundane, or failed technologies.

In an effort to address both the theoretical and practical issues associated with locating and characterizing emerging technologies, the authors are pursuing an automated system directed to these issues. The system processes very large collections of scientific publications and patents, and extracts features from the full text and metadata of these publications and patents (rather than solely from metadata, as is the case with products such as Thomson's ESI and InCites, or Elsevier's SciVal). It then constructs quantitative indicators based upon these extracted features. These indicators are designed specifically to locate and characterize emerging scientific and technological fields.

The system thus has similar goals to the European PromTech project, which also endeavours to locate emerging technologies via analysis of scientific literature (Roche et al, 2010; Schiebel et al, 2010). It is also similar to recent work by Érdi et al (2013), who used a graph-based approach to locate new connections between disparate technologies via citation links. These new connections, which can be identified at the point when patents issue, are regarded as being indicative of a possible emerging field or technology. This finding is similar to that reported in earlier research by Schoenmakers and Duysters (2010), who found that radical inventions often resulted from the combination of knowledge from domains that are not otherwise extensively connected.

In this paper, we use the system to examine one specific aspect of emerging technologies – the extent to which they have exhibited the potential for practical application. The existence of a practical application may be a reflection of the increasing maturity of a particular technology, as it moves beyond the purely conceptual stage. Conversely, the lack of a practical application may signal a technology that is still in its early, pre-emergent stage. The existence of a practical application, or the lack thereof, may also be a reflection of the research field itself. Applied research fields (such as mechanical technologies or information technologies) may inherently have greater potential for near-term practical application than theoretical and basic science research fields (such as theoretical physics or mathematics). The latter may attract increasing interest from researchers over time, and thus be defined as having emerged, without demonstrating a near-term practical application. In simplistic terms, the presence of a practical application may thus point to an emerging research field that is more 'technological' rather than 'scientific' or, in traditional terms, more 'applied' rather than more 'basic' (Stokes, 1997).

This paper contains four further sections. In the first of these sections, we outline the theoretical foundation for our system. We then describe how, guided by this foundation, we construct indicators designed to determine whether emerging technologies have demonstrated a practical application. These indicators are based on features extracted from the metadata and full text of scientific publications and patents. Having outlined the indicators, we then demonstrate how they are combined via Bayesian networks to optimize the model of practical application. Finally, we show the results of applying this model in practice to sets of scientific publications and patents associated with eight sample technologies, both emerging and non-emerging. These document sets are referred to as Related Document Groups (RDGs).

The results demonstrate the extent to which the outputs of the model concur with the opinions of subject matter experts (SMEs) regarding the presence or absence of practical applications for the eight sample technologies. The analysis is carried out across six time periods. This temporal aspect is necessary, since a practical application may not exist for a given a technology at one point in time, only to develop in a later time period.

Theoretical Background

The theoretical foundation for our system is provided by actant network theory (Latour, 2005). This theory provides a vision of science and technology as constituted by networks of heterogeneous elements, interconnected by disparate relationships. These networks do not just contain individuals, but also institutions, instruments, practices, terminology, materials, funders, meetings, government organizations, laws, journals, patents, publications, and so on. The membership of elements within such a network, and the nature and extent of the relationships between these elements, is dynamic and constantly changing.

In the idiom of actant network theory, the task facing our system is to identify, characterize, and evaluate over time the actant networks that comprise emerging research fields. More specifically, this task is to use indicators from the metadata and full-text of publications and patents associated with these fields in order to identify, characterize, and evaluate over time the actant networks of science and technology.

In this paper, we use indicators based on actant network theory to address the question of whether a practical application exists for a given research field at a particular point in time. As noted above, the presence of a practical application may suggest that a research field has experienced a certain degree of emergence, and has moved beyond the early, conceptual stage. This emergence may be reflected in a certain level of activity, maturity and robustness in the actant network associated with the research field.

The presence of a practical application may also help determine whether a research field is more 'technological' or more 'scientific' in nature. The processes of scientific and technological emergence are not distinct or linearly related, but are often intertwined. They also have many similar properties, and thus often have many actants in common, for example trained scientists funded and tasked to work on a given problem. However, there may be differences in the actant network that point to research fields being more (but not exclusively) technological or more (but not exclusively) scientific. For instance, commercial enterprises, particularly small companies without access to extensive funding, may be more active in the actant networks associated with technological fields that offer a greater prospect of financial return in the near term. Also, the commercial marketplace may play a greater role in the actant networks associated with technological fields. The presence of the marketplace actant in the network may in turn affect the types of outputs required from scientists. In particular, there may be a greater interest in patents (which offer the prospect of a monopoly over a given technology) rather than papers (which offer no such monopoly), especially on the part of potential investors in a given technology (Häussler, Harhoff & Müller, 2012).

Indicators and Indicator Patterns

Based on the theoretical foundation discussed in the previous section, we define three characteristics of actant networks that may point to a practical application existing for a given research field in a particular time period: (1) the extent of commercial involvement, (2) the presence of significant patenting, and (3) the degree of maturity of the field.

Each of these characteristics is addressed by a different set of indicators, referred to hereafter as an 'indicator pattern'. These indicator patterns are based on features extracted from full text and metadata features of both scientific papers and patents. Scientific paper collections we currently process include Elsevier (full text articles from 438 journals over 1980-2011, ~4M records), Thomson Reuters' Web-Of-Science® (abstracts of journals and conference proceedings for

the same time period, ~40M records) and Pub-Med Central (full text articles from biomedical journals, ~250k records, dominantly 2008-present). We also processed Lexis-Nexis Patent data which includes granted patents and published patent applications from multiple national patent offices. Each of the indicator patterns, and the indicators contained therein, are described below.

Indicator Pattern 1: Commercial Involvement

This pattern measures the extent to which commercial organizations are active in the actant network associated with a given research area, relative to academic, government and non-profit organizations. The rationale for this pattern is that research areas with near-term practical applications are particularly likely to attract the attention of commercial organizations. Meanwhile, research areas for which practical applications are far in the future, or have yet to be defined, may be characterized by a lower level of activity from commercial organizations. This is not to say that commercial organizations will be entirely absent from these research areas, since many large corporations have specific departments devoted to blue-sky research. However, their activity may be at lower than in research areas that already have practical applications. The indicators in the Commercial Involvement pattern are:

Indicator 1.1 - Percentage of Researchers Affiliated with Commercial Organizations - this indicator is calculated by dividing the number of distinct authors affiliated with commercial organizations by the total number of authors publishing in a given Related Document Group (RDG) and time period. In order to count individual researchers accurately, all author references are disambiguated. The rationale for this indicator is that industrial researchers may be particularly likely to focus on technologies with near-term practical applications. The presence of a large number of such researchers in a given research area, relative to the overall number of researchers active in the area, may indicate that the field has demonstrated promise in terms of practical application.

Indicator 1.2 - Percentage of Publishing Organizations defined as Commercial – this indicator is similar to the industrial researchers indicator described above, but is calculated at the level of organizations, rather than individual researchers. Specifically, this indicator counts the number of organizations that published at least one scientific article in a given RDG and time period, and then calculates the percentage of these organizations that are commercial, rather than academic, non-profit or government. In order to count organizations accurately, all organization references are disambiguated.

Indicator 1.3 - Percentage of Funding Organizations defined as Commercial - this indicator starts by collating all funding organizations acknowledged in publications in a given RDG and time period. It then determines the percentage of these organizations that are defined as commercial, rather than academic, non-profit or government. In order to count funding organizations accurately, all organization references are disambiguated. The rationale for this indicator is that,

while commercial organizations do fund basic research, they may be especially interested in research efforts related to more applied technologies. The presence of extensive funding from commercial organizations may thus indicate that a technology is more applied, and has potential near-term practical applications.

Indicator 1.4 - Percentage of Funding from Non-Academic Organizations - this indicator is similar to Indicator 1.3, but divides the organization types differently. Specifically, rather than isolating commercial funders, it instead isolates academic funders, and counts the percentage of funding organizations that are defined as non-academic. The rationale for this indicator is that, if a large proportion of funders are academic, the field may be more likely to be at a basic science stage of development, rather than an applied technology stage. Conversely, if a large proportion of funders are non-academic, the field may be more applied, and closer to near-term practical application. These non-academic funders may be corporations, government agencies, or non-profit organizations.

Indicator 1.5 - Percentage of Patents with Company Names in their Background Section – this indicator computes the percentage of patents that include company names within the Background section of their Specification. The Specification section of a patent often (but by no means always) contains a description of the background of the claimed invention, including a discussion of prior research. If this Background section contains a reference to a company name, this may be because that company is responsible for an element of this prior research. The existence of numerous patents with such references to companies suggests that a research area may be close to practical application, or has already exhibited such an application.

Indicator 1.6 Percentage of Patents with Company Names in the Example Section - this indicator computes the percentage of patents that include company names within the Example section of the Specification. In order to demonstrate the usefulness of their invention, patent applicants may list examples describing practical experiments or applications of the invention. These examples are contained in the Specification section of the patent, typically under the heading 'Example' or 'Examples'. These examples may contain references to company names, for example suppliers of testing equipment, raw materials, or diagnostic tools. The presence of numerous patents with such references to company names may be indicative of a technology in which researchers are undertaking practical experiments using equipment and supplies sourced externally. In turn, this suggests that they foresee near-term applications for the invention that would justify this expense.

Indicator Pattern 2: Significant Patenting

The presence of the marketplace in the actant networks associated with more 'technological' research fields may affect the types of outputs required of the scientists working in these fields. In particular, there may be a greater interest in patents (which offer the potential of a monopoly over a given technology) rather

than papers (which offer no such monopoly), although it should be recognised that the propensity to patent may to vary across fields. For example, patents may play a greater role in fields such as semiconductor and pharmaceuticals than they do in fields such as retail and finance. This pattern includes the following indicators:

Indicator 2.1 Percentage of Patents - the percentage of patents is determined by dividing the number of patents by the combined number of patents and papers published in a given RDG and time period. If a technology demonstrates a near-term practical application, particularly an application with possible commercial value, researchers may be more likely to protect their innovations via patents, which offer monopoly rights over these innovations. Conversely, if a field is at a more basic, theoretical stage, the possibilities for patenting may be reduced, and researchers may be more likely to publish their findings in scientific journals. A high percentage of patents among the documents in an RDG may thus be indicative of a technology that is demonstrating potential near-term practical application.

Indicator 2.2 Extent of Highly Cited Patents - if a technology is shown to have near-term practical application, it may attract the attention of researchers interested in developing improved versions of the technology. The patents from these researchers will often cite key early patents describing the technology (Breitzman & Mogee, 2002). Hence the presence of numerous highly cited patents suggests that a technology has been the subject of extensive research, which may be due to its potential practical applications. Highly cited patents are defined as those with a Citation Index greater than one. The Citation Index is derived by dividing the number of citations received by a patent by the mean number of citations received by all patents from the same Patent Office Classification (POC) and issue year. The expected Citation Index for an individual patent is one, and a Citation Index greater than one shows that a patent has been cited more often than peer patents from the same year and POC.

Indicator 2.3 Extent of Citation in Highly Cited Patents - this indicator measures the soft max of the Citation Index values of the most highly cited patents in a given RDG and time period. As noted above, the presence of highly cited patents suggests that a technology has been the subject of extensive research, which may be due to its potential practical applications. If the most highly cited patents within a technology have extremely high citation rates associated with them, this may be a particularly strong indicator that a practical application has been identified for the technology.

Indicator 2.4 Percentage of Commercial Patent Assignees - this indicator counts the percentage of patent assignees (i.e. owners) in a given RDG and time period that are classified as commercial, rather than individual, academic, government or non-profit. The rationale for this indicator is that, while commercial organizations do fund basic research, they may be especially interested in research efforts

related to more applied technologies. The presence of a high percentage of commercial patent assignees may thus indicate that a research field has potential near-term practical applications.

Indicator 2.5 Percentage of Commercial and Individual Patent Assignees – this indicator is similar to Indicator 2.4, but counts both commercial and individual patent assignees, rather than just the former. The rationale for counting commercial patent assignees as a possible signal for practical application is the same as in Indicator 2.4 – i.e. companies may be especially interested in research efforts related to more applied technologies. Meanwhile, patents assigned to individuals often describe specific practical applications, such that these individuals see sufficient near-term commercial potential in their invention to justify the expense of filing and prosecuting the patent. The presence of a high percentage of commercial and individual patent assignees - relative to academic, government and non-profit assignees - may thus indicate that a research field has potential near-term practical applications.

Indicator 2.6 Average Generality Index – the Generality Index measures the extent to which the patents citing a given starting patent are dispersed across technologies (Hall, Jaffe & Trajtenberg, 2001), as defined by Patent Office Classifications (POCs). The rationale for this indicator is that, once an innovation has been shown to have practical use, it may draw the attention of researchers in other disciplines, who consider its potential application elsewhere. This indicator is normalized in the same way as the Citation Index (i.e. by POC and year) and has an expected value of one.

Indicator 2.7 Extent of patents with high Generality Index scores - This indicator computes the number of patents in a given RDG and time period with Generality Index scores greater than 1.5 (i.e. they are at least 50% more generally applicable than peer patents from the same issue year and technology). The indicator is similar to Indicator 2.6, but focuses on individual, high scoring patents, rather than the mean Generality Index across all patents in a given RDG and time period.

Indicator Pattern 3: Maturity of Field

This pattern analyzes the maturity of a given research field, with the purpose of identifying fields that have moved beyond the initial, pre-emergent phase. The indicators in this pattern analyze the types of documents present in the RDG (e.g. the existence of product reviews), as well as the availability of technologies cited by documents in the RDG (i.e. whether the technologies are only available on an experimental basis or are readily available for use). Other indicators measure the extent of characteristic terminology in the claims section of patents, which tells us whether the terminology of the RDG has been legally accepted; as well as the extent of specific lexical phrases often used to talk about practical applications, which may be indicative of a relatively mature field.

Indicator 3.1 Extent of product review articles - we created an ontology of genres (interpreted as the term for any category of scientific literature characterized by a particular style and form), and applied it to the documents within a given RDG and time period. The ontology includes 21 types that are arranged in a shallow hierarchy. Some of the most frequent genres are Research Article, Review Article, Report, Book Review, Product Review, Commentary, Abstract, Letter, Correction, Case Report, Editorial, Short Communication, and Discussion. For this indicator, we computed the number of Product Reviews in a given RDG and time period. By their nature, product review articles are related to a product that is available, or will soon be available, in the commercial marketplace. As such, the product has – or is purported to have – a practical application. The presence of numerous product review articles suggests a research field is connected to existing, or forthcoming, products.

Indicator 3.2 Maturity of technologies referenced in patents – this indicator measures the maturity of the technologies referred to in patents, using a classification scheme in which these referenced technologies are defined as unavailable, immature or mature. The classification system uses an ontology that starts with a set of seed patterns and seed technologies, and employs machine learning tools and other heuristics to create the classifications from these seeds. The rationale for this indicator is that research fields with a practical application will tend to reference more mature technologies than research fields that are still at a conceptual stage.

Indicator 3.3 Density of Manufacture relations – the Manufacture relation links arg1 (person, organization or document) with arg2 (document or term), such that arg1 is the manufacturer of arg2. For example, in their *Materials and Supplies* sections of scientific articles, authors may provide details of items they used and where they obtained them. Similarly, patents may refer to practical experiments using particular tools or equipment. This indicator computes the density of Manufacture relations, i.e. the average per document in a given RDG and time period. The presence of Manufacture relations suggests that researchers in a research field are discussing relevant manufacturing tools and techniques. In turn, this suggests that the research field has moved beyond the purely conceptual or theoretical stage, and has entered a more applied stage where specific methods of manufacture are being considered. A high density of Manufacture relations may thus be indicative of an applied technology with potential near-term practical application.

Indicator 3.4 Density of Practical relations – the Practical relation refers to a particular item either being used, or being useful in some way. It often involves patterns in which trigger words (verbs like *use* or *utilize* or instrumental prepositions such as *by*, *via* or *with)* are in close proximity to either citations or terminology describing tools, methods or other descriptors of technology. This indicator computes the density of Practical relations identified in a given RDG and time period. A high density of such relations may be indicative of a research

field in which practical applications are a major focus, and are being actively discussed. In turn, this suggests that such practical applications may exist, or are forthcoming.

Indicator 3.5 Percentage of patents with Examples in the Description of patents - this indicator computes the percentage of patents that include Example headings within their specification section. To demonstrate the usefulness of their invention, patent applicants may list examples describing practical experiments or applications of the invention. These examples are contained in the Specification section of the patent, typically under the heading 'Example' or 'Examples'. The presence of numerous patents containing such an 'Example' section may be indicative a research field where it has been possible for many researchers to report the results of practical experiments and applications.

Indicator 3.6 Percentage of patents that include references to trademarks - this indicator computes the percentage of patents that include references to trademarks in their Specification sections. While the purpose of patents is to protect innovations, the purpose of trademarks is to protect products or services. Patents may make reference to trademarks in their Specifications, for example when referring to a potential application for an invention, a component used in the invention, or a piece of equipment used to test the invention. The presence of numerous patents with references to trademarks may be indicative of a research field that is closely related to existing products, or is being tested in practical experiments, and may thus have near-term practical application.

Model Description

As outlined above, there are three indicator patterns directed to the existence of practical applications for given research fields in particular time periods. These patterns are combined in our model, which then produces an output in the form of Yes/No response to the question: 'Was there a practical application for <concept> during <time period>?'. In order to test our model, the responses are compared to those from subject matter experts (SMEs) as to whether practical applications existed for given research fields in particular time periods. There are a total of eight such technologies (DNA Microarrays; Genetic Algorithms; Cold Fusion; Steganography; RF Metamaterials; Horizontal Gene Transfer; Tissue Engineering; and RNA Interference) and six time periods (1981-1985; 1986-1990; 1991-1995; 1996-2000; 2001-2005; 2006-2010). For example, the SMEs might be asked whether a practical application existed for Genetic Algorithms in 1996-2000, or Tissue Engineering in 2001-2005.

Before reporting the results from comparing the outputs of our model with the responses from the SMEs, it is first useful to outline details of the model. Our model is based on Bayesian networks, which are probabilistic graphical models. Probabilistic relationships among variables are captured by a directed acyclic graph. In this graph, the nodes are variables together with a specification of the probability distribution for each variable conditional on values for its parent

variables (i.e. variables having edges to the given variable). Such a representation allows for a decomposition of the joint probability distribution over the variables that supports efficient computation of probabilities.

The model is hierarchical, with indicator variables linked to pattern variables, which are in turn directly linked to the question variable. The pattern variables represent abstractions or summaries of related indicator variables. Such a hierarchical structure has a number of advantages. First, it allows the use of many correlated variables without danger of over-counting. For example, percentages of commercial researchers and percentages of commercial organizations will be highly correlated, so that treating them as independent pieces of evidence would overstate their influence. If, however, we take these two variables to be manifestations of a more general commercial involvement variable, and link this variable to the question variable, then the correlation between the commercial researcher and commercial organization variables will be properly accounted for.

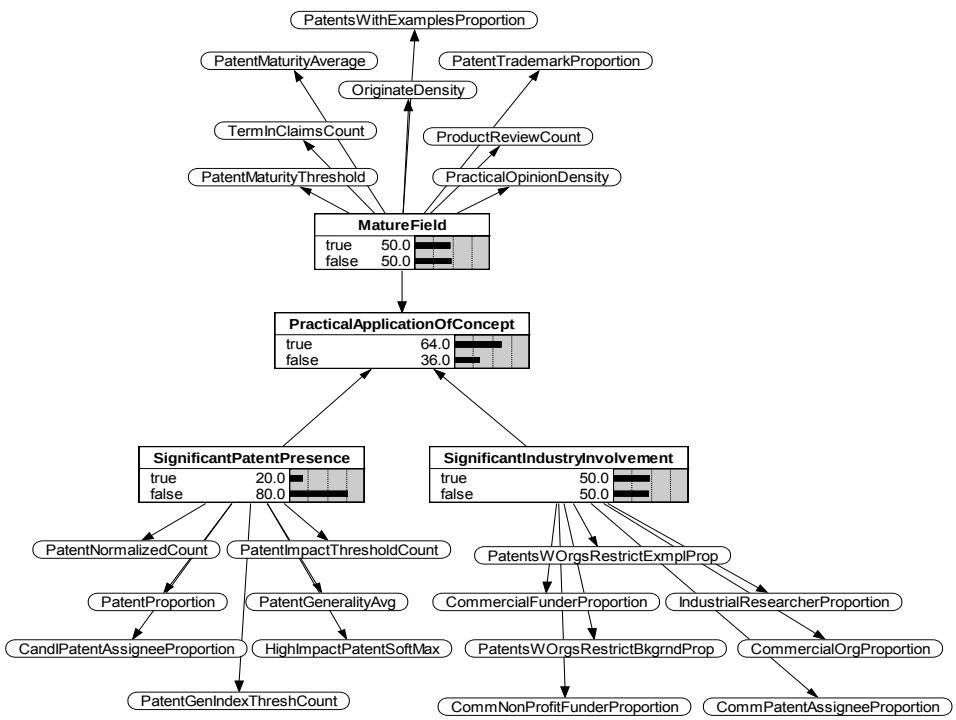


Figure 1 – Model for Existence of Practical Application

A second advantage of the hierarchical approach is that it enables more accurate modeling when training data is sparse. With sparse data, estimating the correct, or even an approximately correct, probability distribution for each indicator variable conditional on the question variable is very error prone. With intermediate pattern variables, theoretical considerations can be brought to bear to help shape the

distribution in conjunction with whatever ground truth data is available. A third advantage to a hierarchical structure is that explanations for answers are more comprehensible when framed in terms of meaningful patterns above the level of individual indicators. Finally, the intermediate patterns will often be of interest in and of themselves, independently of how the question is answered.

Figure 1 shows the model for addressing the question 'Was there a practical application for <concept> during <time period>?'. This model includes the three indicator patterns outlined in the previous section – i.e. commercial activity, extensive patenting, and field maturity. It should be noted that other indicator patterns may also have informational value regarding this question. However, the additional patterns we considered were not used, either because analysis showed their information value to be low, or because modeling their relationship to the existence of a practical application proved difficult and error-prone.

Each subgraph in Figure 1, consisting of a pattern node and its children, is a naïve Bayes model. Updating in a naïve Bayes model is particularly simple. Evidence takes the form of an assignment of a value to a child variable and each piece of evidence contributes independently to the posterior distribution over the parent variable.

For our model of practical application, we used the following Boolean conditions over the pattern variables as conditions for inferring the existence of a practical application:

1. Commercial Involvement is TRUE.
2. Patenting is TRUE.
3. Maturity is TRUE.

Each of these conditions is considered 'sufficient to some degree' for a practical application to exist. The Boolean conditions are combined using a weak disjunctive model known as the "noisy-or" distribution. The probability of the existence of a practical application is greater than 0.5 when at any one of the three conditions is satisfied, but the conditions differ in their strength of support for a practical application. Finding a significant patent presence provides very strong support (greater than 0.9) for the existence of a practical application independently of the other pattern variables while maturity provides only weak support (0.55) when the other two pattern variables are false. When all pattern variables are false, the probability of a practical application is low (0.1).

Results

We ran the model on the RDGs for each of the eight sample technologies in each of the six time periods listed above. Where the model output a probability value greater than 0.5 for the existence of a practical application, this was considered a positive answer to the question (i.e. the model states that a practical application existed in that research field and time period). Conversely, a value lower than 0.5 was considered a negative output (i.e. the model states that a practical application did not exist for the field during the time period). These outputs were then

compared against the responses from the SMEs, and the results are shown in Table 1.

Table 1. Results Comparing Model Outputs with SME Responses

	All	*RF Metamaterials*	*Tissue Engineering*
True positives	27	2	5
False positives	5	1	0
True negatives	11	3	1
False negatives	3	0	0
Recall	0.90	1	1
Precision	0.84	0.66	1
Accuracy	0.82	0.83	1

There are a total of 46 answers in the 'All' column of Table 1. This represents eight technologies times six time periods, minus two time period/technology pairs for which the data were too sparse for the model to run (both pairs were from the earliest time period, 1981-1985). The results in Table 1 show that the model worked consistently well with respect to the SME responses. Recall is 0.90 (i.e. 90% of time period/technology pairs with positive SME responses to the question of whether a practical application existed were also given positive answers by the model). Meanwhile, Precision is 0.84 (i.e. 84% of time period/technology pairs marked with a positive answer by the model were also marked positive by the SMEs); and Accuracy is 0.82 (i.e. in 82% of cases, the responses from the model and the SMEs matched, whether these responses were positive or negative).
Tables 2 and 3 provide more detailed views of the different indicator patterns incorporated in the model, and how they contribute to the responses generated by the model. Table 2 shows results for Tissue Engineering, while Table 3 shows results for RF Metamaterials.

Table 2. Results for Tissue Engineering Related Document Group (RDG)

	SMEs	*Model*	*Commercial Involvement*	*Patents*	*Maturity*
1981 - 1985	NO	NO	FALSE	FALSE	FALSE
1986 - 1990	YES	YES	TRUE	TRUE	FALSE
1991 - 1995	YES	YES	TRUE	TRUE	FALSE
1996 - 2000	YES	YES	TRUE	TRUE	TRUE
2001 - 2005	YES	YES	TRUE	TRUE	TRUE
2006 - 2010	YES	YES	TRUE	TRUE	TRUE

In Table 2, it is possible to see the transition from entirely negative answers to the practical application question from both the model and the SMEs in the earliest time period (1981-1985), to entirely positive responses in the more recent time periods. This reflects the widespread application of tissue engineering techniques

in recent years. The responses from the SMEs and the model match for each time period, so precision, recall and accuracy are all one for this RDG.
Table 3 shows the SME responses and model outputs for the RF Metamaterials RDG. This table reveals that a practical application for RF Metamaterials has been identified much more recently, according to both the SMEs and the model. For the first four time periods, covering 1981 to 2000, the SMEs gave a negative response to the question of whether a practical application existed for RF Metamaterials. They gave a positive response to this question for both 2001-2005 and 2006-2010. The model, meanwhile, suggested that there was a practical application for RF Metamaterials earlier than the SMEs, and switched from a negative to a positive response in the 1996-2000 time period. This is largely due to the presence of a high percentage of industrial researchers during this period. The positive response from the model in 1996-2000 results in a false positive for this period, reflected in the figures for precision (0.66) and accuracy (0.83).

Table 3. Results for RF Metamaterials Related Document Group (RDG)

	SMEs	*Model*	*Commercial Involvement*	*Patents*	*Maturity*
1981 - 1985	NO	NO	FALSE	FALSE	FALSE
1986 - 1990	NO	NO	FALSE	FALSE	FALSE
1991 - 1995	NO	NO	FALSE	FALSE	FALSE
1996 - 2000	NO	YES	TRUE	FALSE	FALSE
2001 - 2005	YES	YES	FALSE	TRUE	TRUE
2006 - 2010	YES	YES	FALSE	TRUE	TRUE

The results in the tables above suggest that our model shows promise in terms of determining the existence of practical applications for given research fields. This determination is based solely on the content of scientific and technical documents, and without access to product or market data. Given that product and market data are often difficult to collate, or are expensive to source, a model such as this that does not require access to such data may be a useful tool.
Practical applications do not only exist for emerging research fields, but also for mature fields. Such fields may also exhibit the characteristics covered by the indicator patterns included in our analysis – i.e. commercial involvement, significant patenting, and maturity. Hence, it is not necessarily the existence of a practical application that is interesting from an emerging technology standpoint, but the existence of such an application for the first time. As a result, the promise shown by the model in recognizing the transition from absence to existence of practical applications for given research fields may be of particular interest when evaluating emerging technologies.
Although the results are promising, it should be noted that they are based on a very small data set, largely due to the time-consuming nature of surveying SMEs. Additional research using more extensive data sets may thus be instructive, in

order to determine whether the promising results reported here are repeated for a larger sample of technologies. It also needs to be recognized that, as with any human-based response data, there is the possibility of bias or misunderstanding in the responses from the SMEs. It is thus possible that the model outputs are being compared against erroneous responses from the SMEs.

Conclusions

This paper outlines a system directed to the determination of whether practical applications exist for research fields, particularly those that are emerging. The system uses indicator patterns - based on features extracted from the metadata and full text of scientific papers and patents - to assess different characteristics that point to the existence of practical applications. This system may help determine whether a particular field has moved beyond the early, conceptual phase towards a more applied, practical phase. It may also help to classify emerging research fields as more 'technological' or more 'scientific' in nature.

The results reported in this paper suggest that the system shows promise in determining whether practical applications exist for given research fields in particular time periods, based on comparisons with responses to the same question from subject matter experts. Perhaps more interestingly, the system shows promise in detecting the transition from absence to existence of practical applications over time, which may be of particular value in evaluating emerging technologies. It should be noted, however, that the results are based on a small sample of research fields, due to the time consuming nature of obtaining responses from subject matter experts. More extensive research using a larger sample of research fields may be worthwhile, in order to determine whether the results are repeated for this larger sample. Also, in this paper, the system is applied to pre-determined research fields. It may be interesting to apply the system to a broader corpus of documents covering multiple research fields, to determine the extent to which it is able to locate those fields that are emerging, and also appear to have a practical application.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Breitzman, A. & Mogee, M. (2002). The many applications of patent analysis. *Journal of Information Science*, 28, 187-205.

Christensen, C. & Bower, J. (1996). Customer power, strategic investment, and the failure of leading firms. *Strategic Management Journal*, 17(3), 197-218.
Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P. & Zalányi, L. (2013). Prediction of emerging technologies based on analysis of the US patent citation network, *Scientometrics*, 95(1), 225-242.
Goldstein, J. (1999). Emergence as a Construct: History and Issues. *Emergence: Complexity and Organization*, 1, 49-72.
Hall, B., Jaffe, A. & Trajtenberg, M. (2001). The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools, *CEPR Discussion Paper No. 3094*.
Häussler, C., Harhoff, D. & Müller, E. (2012). To be financed or not - the role of patents for venture capital financing. *ZEW - Centre for European Economic Research Discussion Paper No. 09-003.*
Latour B. (2005). *Reassembling the social: An introduction to actor-network theory*. Oxford, UK: Oxford University Press.
Roche, I., Besagni, D., François, C., Hörlesberger, M. & Schiebel, E. (2010). Identification and characterisation of technological topics in the field of Molecular Biology. Scientometrics, 82(3), 663-676.
Schiebel, E., Hörlesberger, M., Roche, I., François, C. & Besagni, D. (2010). An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices. Scientometrics, 83(3), 765-781.
Schoenmakers, W. & Duysters, G. (2010). The technological origins of radical inventions. *Research Policy*, 39, 1051-1059.
Schumpeter, J. (1912). *Theorie der Wirtschaftlichen entwicklung*, Leipzig: Duncker & Humboldt.
Stokes, D. (1997). *Pasteur's Quadrant: Basic Science and Technological Innovation*. Washington, DC: Brookings Institution Press.

IDENTIFYING EMERGING TECHNOLOGIES: AN APPLICATION TO NANOTECHNOLOGY

Mickael Pero

mickael.pero@isi.fraunhofer.de
Fraunhofer ISI, Breslauer Straße 48, 76139 Karlsruhe (Germany)

Abstract
Research and inventive activities represent core elements of science-based companies' comparative advantage. In a competitive environment, this requires methods to identify at an early stage the most promising technologies within the scope of the companies' business model. Generally, they frame their technological strategies on personal expertise, intuition or gut feelings but this intangible decision process faces possible adverse effects like imperfect information or tunnel vision. The technology scanning literature usually tackles this limitation by analyzing patent trends in absolute terms. However, this approach disregards both the relative nature of technology emergence, as well as the scientific dynamics behind technologies. This paper proposes an alternative decision support tool which identifies technologies with relative emerging patterns based on science and technology data, and connected by adequate –and expert reviewed- keyword strategies. Emerging technologies are identified from Sharpe ratios in a two dimensional S&T framework. An empirical test is conducted in the field of nanotechnology where emerging technologies are found to belong to diverse material types, although the largest dynamic is observed for carbon based technologies. This method appears as adequate to support the decision process regarding companies' technological choices by providing insightful information on ongoing emerging technologies.

Introduction
Science-based companies in a competitive environment need to assess and decide at a fast pace which technologies they should adopt or develop (Chesbrough and Appleyard, 2007). This is necessary to improve existing comparative advantages, or merely to "stay in the race" of an area before the knowledge gap between own competencies and rival ones is too wide (Cohen and Levinthal 1989). Companies' technological decision-making ultimately relies on vision, intuition or gut feeling (Hayashi 2001). Although crucial, this process can be improved by accounting for tangible evidence which can help to avoid possible biases such as tunnel visions or lock-in adverse effects (Arthur 1989).
The economic and management literature addresses this issue in topics called technology intelligence, foresight, or forecasting. The objective is to identify opportunities –e.g. emerging technologies- lying outside the company's boundaries that would not have been identified otherwise. In turn, this additional knowledge supports technological decision making by providing new or enriched evidence on specific technologies (i.e. technology monitoring) or the current technology landscape (i.e. technology scanning). There are two main channels for

companies to source information. On the one hand, knowledge can be inferred from technology scouts, who are "explorative" experts retrieving information inside as well as outside the boundaries of the company. Their role is to scan ongoing technological developments to identify which technologies are emerging within the globalised scientific and technological communities (Rohrbeck, 2006). On the other hand, knowledge can be induced from publication and patent databases, a technique also known as "tech mining" (Porter and Cunningham 2005). Although showing some delays due to database updates, the strength of this method is to rely on the statistical significance of technology developments and patterns.

Until now, the "tech mining" literature proposed to identify the emerging phase of a technology in comparison to other development phases of the same technology. The dominant framework using this approach is the technology life cycle (TLC) framework. It provides a way to interpret the evolution of a technology over time. The TLC describes the stages of technology evolution namely: introduction, growth, maturity, and decline (Popper and Buskirk, 1992) or emerging, rapid growth, maturity, decline (Roper, Cunnigham, Porter, Mason, Rossini, and Banks, 2011). In theory, this sequence is illustrated by an S-curve which depicts the evolutionary path of a technology over time (Andersen, 1999).

Although adequate as a starting point, the research described above is based on two strong assumptions which appear fragile when considering the evolution of science and technology. The first assumption is that technology fields or technologies can be defined as emerging when they fulfil certain conditions in absolute terms. However, technologies are interacting and thus the emergence of one is not only determined by endogenous factors but also by exogenous factors such as competition or market forces. The second assumption is the restriction to invention data (i.e. patents). However, in the context of increasingly science based technologies, science dynamics should also be taken into account for identifying technologies in emerging stages.

The method developed in this paper adopts the point of view that the emergence of technologies is a relative concept, namely that exogenous factors affect a technology dynamic. Thus, the performance of a technology is measured more adequately when compared to others. Although this approach has been applied for evaluating the dynamics of fields (Reiß, Hartig, and Schmoch, 2009), it has not been applied to technologies. Also, science dynamics behind technologies will be accounted for by identifying at the technology level the related S&T literature using common keyword strategies. This approach has been used in order to define science or technology fields (Noyons, et al., 2003; Schmoch and Thielmann, 2012), but has not been used to connect the underlying S&T activities of a given technology.

Conceptual framework: emerging technologies and S&T dynamic

The central theoretical claim of this study is the existence of intrinsic scientific and technological dynamics behind science based technologies. Since those

dynamics differ among technologies, an emerging stage is identified by the co-occurrence of stronger S&T growth dynamics with respect to a reference (e.g. a sample of technologies). The framework therefore considers the magnitude of the interactive process between research and inventive activities as key determinant for technology emergence (c.f. the TLC only looks at the inventive dynamic). The following sketch summarises the concept.

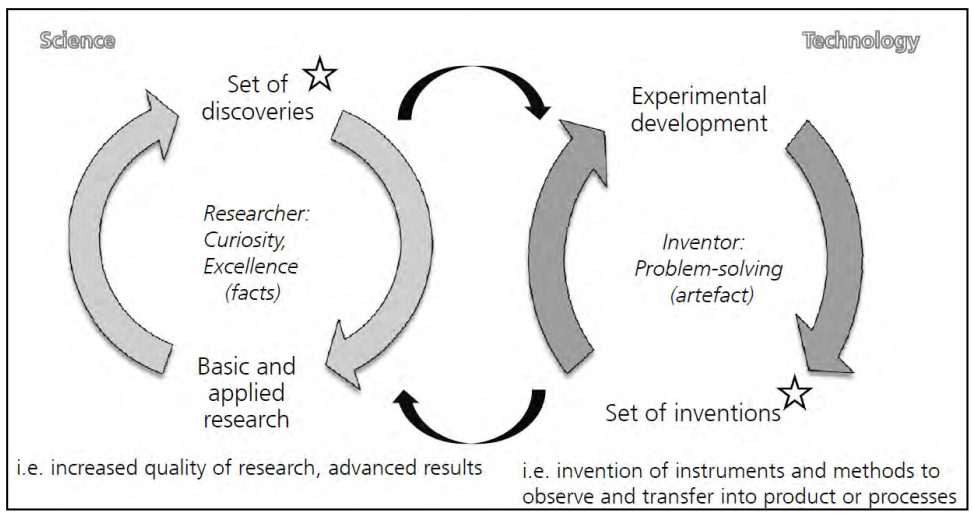


Figure 8: Conceptual Framework

Science and technology constitute two parallel and interacting dimensions in this framework (OECD, 1993). On the one hand, the scientific process consists in fundamental and applied science activities carried out by researchers. Their activities continuously feed the set of discoveries (the R in R&D). The nature of their discoveries is usually information which can be quickly transmitted and shared with the scientific community. In turn, the pool of created knowledge (i.e. discoveries) interacts with experimental developments which can ultimately lead to the development of prototypes, new methods or more generally inventions (the D in R&D). On the other hand, the transition from development to technology occurs when an idea is concretized into an artefact or invention. A "(…) first occurrence of an idea for a new product or process" (Fagerberg, 2005, S. 4) which is both strongly science related and depends on the skills and talent of inventors. This phase accounts for intermediary inputs which "take the form of new materials, new machines, new components, or technical processes that never show up in conventional measures of final product for the simple reason that they are not final products" (Rosenberg, 1982, S. 71-72). Indeed, a significant amount of these inputs are used back as support for science.

The emergence of new science based technologies is therefore driven by a constant interaction between scientific and technological developments: "science

often is dependent, in an absolute sense, on technological products and processes for its advances. Over the course of history thus far, it is moot whether science has depended more on technological processes and products than innovation has depended on science" (Kline and Rosenberg, 1986, S. 287). This non linear process has been conceptualized by Kline and Rosenberg's chain-linked model which is particularly relevant for emerging technologies (Kline and Rosenberg, 1986, S. 303): on the one hand "potentiation of wholly new devices or processes from research"; and on the other hand "much essential support of science itself from the products of innovative activities, i.e. through the tools and instruments made available by technology". This applies to current dynamic fields such as biotechnology or nanotechnology which tend to be more closely linked to science than traditional fields (Mansfield, 1990; Schmoch, 1997; Järvenpää, Mäkinen, & Seppänen, 2011). This is due to the fact that a large part of today's bottlenecks can be solved at the infinitely small where both science and technological activity become increasingly intertwined. As such, the object of a research (potential discovery) and its observation (potential invention) happen closely together in the emerging process. One example of this evolution is the increase in scope of patentable areas to include scientific advancements in the case of genes or new materials seen as inventions rather than discoveries due to their "novel" nature. This can be seen in the data (Eurostat 2012) where academic patenting in Europe almost tripled in ten years (i.e. from 486 thousands patents in 1998 to almost 1.266 thousands patents in 2008)

Measures

The measure used for scientific activity, and more precisely discoveries made from research activities are counts of publications in journals, trade journals, book series and conference material that have an ISSN (International Standard Serial Number) assigned to them, as well as conference papers. All these serial publications are taken into account as a measure for research activity in order not to disregard emerging topics which might be exploratory and therefore not necessarily present only in peer-reviewed scientific articles.

Reliable and often used proxies of inventions are patents which describe the technological inventive dynamics of a technology. Patent counts describe the technology side of the conceptual framework in that they are an "indicator of invention rather than innovation: they mark the emergence of a new technical principle, not a commercial innovation" (Smith, 2005, S. 160).

It is to note that although these bibliometric measures do not distinguish within variations of quality, practices across fields and organisations, and are limited by delays or secrecy, they are still providing key information on research and inventive activity (Watts & Porter, 1997).

Indicator

The indicator that will be used to interpret the research and inventive dynamics of technologies needs to possess two characteristics. First, it should provide a way to

compare different technologies. This means that the method should identify technologies which are emerging with respect to a given reference. Second, the indicator should be measured by the average growth over a given time period. Indeed, when the size of the activities are limited –as in the case of emerging technologies – relying solely on year to year growth is deeply affected by "noise" due to exogenous event or mistakes coming from the database or data treatment. It is therefore safer to rely on the average of longer periods to assess the growth of activities. Designed to take into account the elements mentioned above, the Sharpe ratio appear as an adequate indicator to identify emerging technologies (Reiß, Hartig, and Schmoch, 2009; Sharpe, 1994).

$$\boldsymbol{Sharpe\_Ratio} \quad (\textbf{when } \sigma_a > 0, I_a \neq 0) \qquad S_a = \frac{(\bar{I}_a - \bar{I})}{\sigma_a}$$

Where $\bar{I}_a = \frac{1}{n}\sum_{t=1}^{n} I_a^t$ and $\bar{I} = \frac{1}{n*N}\sum_{a}^{N}\sum_{t=1}^{n} I_a^t$

Where a denotes the technology, I_a^t year t technology growth, $\bar{I}_a$ the mean of year to year technology growth over the period, $\bar{I}$ the mean growth of the reference (e.g. selected technologies) over the period and σ_a the standard deviation of technology growth over the period.

The Sharpe ratio has characteristics that suit the identification strategy. Looking at the numerator, it compares the average growth of a given technology to the overall average growth of the reference sample. A positive numerator would depict a relatively higher growth and vice versa for a negative numerator. This difference is normalized by the standard deviation of the annual technology growth over the period under scrutiny. This accounts for the variability of the growth rates in the way that technologies showing stability in their growth rates are less penalized in their score than more volatile fields (which also tend to be smaller in size).

For completeness, both the absolute and relative growth measures are proposed to identify emerging technologies. The reason to use two growth measures for each activities (i.e. research and inventive) is that they both identify emerging phases but with different emphasis.

Considering absolute growth where ${I_a}^t = X_a^t - X_a^{t-1}$ (X_t being the literature count –publications or patents- in year t), technologies experiencing more sustained growth are advantaged. It is to mention that this measure would face computational limitations in the following cases:

If $\bar{I}_a - \bar{I} \neq 0$ **and** $\sigma_a = 0$ $\qquad S_a = \infty$

If $\bar{I}_a - \bar{I} = 0$ **and** $\sigma_a = 0$ $\qquad S_a$ **is indeterminate**

In order to correct for this singularity, cases where the Sharpe indicator is not computable are discarded. Thus the following condition is imposed: $_a$>0. In contrast, considering relative growth where $I_a{}^t = (X_a^t - X_a^{t-1})/X_a^{t-1}$ is a measure which advantages technologies which begin to show some activities during the observed period (i.e. higher growth rates). However, this proxy is non-measurable when no research or inventive activity can be observed. Additional non-computable and disregarded cases are:

If $\boldsymbol{X_a{}^{t-1} = 0}$ ***and*** $\boldsymbol{X_a{}^t} > 0$	$\boldsymbol{I_a{}^t = infinite}$
If $\boldsymbol{X_a{}^{t-1} = 0}$ ***and*** $\boldsymbol{X_a{}^t = 0}$	$\boldsymbol{I_a{}^t = indeterminate}$
If $\boldsymbol{X_a{}^{t-1}} > 0$ *and* $\boldsymbol{X_a{}^t = 0}$	***not consistent with*** $\boldsymbol{I_a{}^t = infinite}$

These cases are caused by a lack of data for computing the relative growth values for a given technology. For example, this problem arises when the number of entries among the period is only one (indefinite standard deviation), or the data is constant over time (0 standard deviation). In the data there are also series with 0's and 1's. In that case, the problem faced is that the growth from 0 to 1 is infinite, whereas the passage from 1 to 0 is -1. This may lead to biases in the results. In order to solve this limitation, the immeasurable cases mentioned are disregarded and replaced by missing values (c.f. the third case where $X_a{}^{t-1} > 0 \; and \; X_a{}^t = 0$).

A last condition is that for each year, both research and inventive growth measures by technology should be computable in order to be accounted in the analysis. In the case one "scientific" Sharpe indicator is not computable for a given year; the "technology" Sharpe indicator is not included to avoid bias. Connecting the two dimensional framework with the Sharpe ratio indicator gives the following framework.

Table 4: Emergence typology

Technology status during period p		Technological Sharpe ratio	
		Negative	Positive
Scientific Sharpe ratio	Positive	*Basic (2)* *Strong growth of publications;* *Weak growth / decline of patents* *"too" new / Maturing (4)* *Weak growth / decline of publications; weak growth / decline of patents*	*Emerging (1)* *Strong growth of publications; strong growth of patents* *Applied (3)* *Weak growth / decline of publications; strong growth of patents*
	Negative		

From the above typology, it can be noted that the key pattern for emergence is the co-occurrence of growth activities (i.e. Quadrant 1). Quadrant 2 defines basic

technologies which are only experiencing high growth in research activities. Quadrant 3 defines applied technologies which are only subject to high inventive activities. Finally the last quadrant is the combination of low to null research and inventive activities. Depending on the novelty of the reference field, technologies appearing in that quadrant can either be “too” new or “mature”.

Data collection

The empirical part of this paper focuses on nanotechnologies, which represents a new and promising science based field strongly subject to both scientific and technological activities. Estimation of the market size of nanotechnology varies greatly: from 27 billion dollars by the European Commission (EC, 2011) for the nanotechnology field to an estimate of nano enabled product market of 1 trillion dollars by 2015 for the National Science Foundation in the US to 2.5 trillion dollars for the OECD (OECD, 2009). Technically, nanoscience can be defined as the study of nanostructures and nanotechnology as the discipline which uses nanostructures to create useful nanoscale devices. Nanotechnology have size (1nm<n<100 nm or a billionth meter) and novelty characteristics. Any definition of nanotechnology should include: the size of the structure, the ability to work at that scale, and exploitation of properties and functions specific for the nanoscale (Malanowski, Heimer, Luther, and Werner, 2006).

The “nano” technology data is collected using the following procedure:

1. Identify nanoscience and nanotechnology field using retained measures
2. Identify key nanomaterials
3. Define keyword strategies to link science and technology
4. Validate keywords (back to step 2 if not passed)

Nanoscience and nanotech fields

The dataset is based on information retrieved from both scientific publications and patent databases from 1996 to 2008. The lower and upper years are subject to limitations faced by publication and patent database access respectively.

The nanoscience publication set belongs to Scopus (Elsevier) and spans from 1996 to 2009. The set of nanoscience publications has been identified using the keyword strategy by Noyons et. al. (2003). It is to note that other search strategies have been proposed by several studies. As reported by Huang et. al., (2008), main references concerning nanoscience and nanotechnology search strategies are Glänzel et. al (2003), Noyons et al (2003), Porter et. al. (2008) and Schmoch and Thielmann (2012). Following Schmoch (1997), this publication time series will be anticipated by one year in order to represent the time lag between the submission year and publication. The available dataset therefore contains all scientific publications with submission year from 1996 to 2008. In total, the extracted information concerning “nano” research activity in nanoscience from 1996 to 2008 is close to half a million documents (480,417).

Concerning the nanotechnology patents, the retrieved set spans from 1995 to 2008 where the year represents the priority year (year of first application). The set of

documents is identified using the EPO tag "Y01N", a nano-related category. Alternatives exist such as EPO's new category called "B82" implemented in October 2011. However, this new category does not accurately incorporate all patent documents yet. Other keyword strategies to identify patents in nanotechnology can be mentioned such as the one proposed by Noyons et. al. (2003) or Schmoch and Thielmann (2012). The selected patents follow the concept of transnational patents from Frietsch and Schmoch (2010). It consists in EPO and PCT patent applications excluding EURO-PCT patents. It is to note that both have delays between patent filling and application of 18 months. This means that to have a full dataset, the accessed Patstat schema (April 2011) contains full information up to September 2009. Since we use yearly information, the dataset will therefore span until 2008 for completeness. In total, the total "nano" inventive activity for the 1996 to 2008 period is close to sixty thousands patent applications (57,269).

Identify key nanomaterials

This paper aims at identifying those "nano" technologies that are at an emerging stage of development. For this, a nanomaterial level of analysis is adopted. Nanomaterials are considered here as technologies in that they enables new product functions and capacities. To identify the set of promising materials a literature review is conducted to select at a fine grained level the most promising nanomaterials from different expert sources: VDI technologiezentrum (VDI, 2010), Ratner and Ratner (2004), EPO nanotechnology categories (2012), EAG (2009), LUX Research, and BCC Research (BCC, 2010). Overall, 34 nanomaterials were retained as being relevant for nanoscience and nanotechnology. These can be grouped in 7 nanomaterial types: bio based, carbon based, ceramic nanoparticles and nanostructures, nanocomposites, metal nanostructures and alloys, polymers, and semiconductor nanostructures.

Keywords as link between science and technology

The methodological challenge is to link a given technology to both its related research and inventive activities. Existing studies have used many different strategies to link science to technology fields, but which usually do not focus on specific technologies and give priority to field generality. The usual strategy is either to identify the degree of correlation between science and technology related fields (Coward and Franklin, 1989; Hullmann and Meyer, 2003; Reiß & Thielmann, 2010; Wydra, Haas, Jungmittag, Reiss, & Thielmann, 2012) scientists who are both authors and inventors (Hullmann and Meyer, 2003; Noyons, et al., 2003). However, the objective of this study aims at identifying and comparing research and inventive activity related to specific technologies within a given field, namely nanotechnology. One way to proceed is to use adequate keywords associated with a given technology that could be used to measure both its scientific and technological activity. This strategy appears relevant when investigating emerging technologies for comparative purposes. Therefore for each

technology in the retained technology set, a keyword strategy is applied. The keyword strategy is defined in such a way that each technology will have the same keyword(s) for both patents as well as publications queries. This strategy appears as accurate in making sure to capture technologically related research and inventive subsets. The reasonable assumption is that although the research questions and objectives are different, the same key terms in science and technology are used. Keyword terms for each technology are identified based on the specialized nanoscience and technology literature (BCC, 2010; EAG, 2009; VDI, 2010), as well as expert check from the Fraunhofer Institute for System and Innovation Research (ISI). The search is then performed with both specific search terms and nanoscience and technology fields search strategies from the first step.

Validate keywords

An additional challenge is to make sure that what is detected with the selected search terms measures the activity behind the technology of interest. This is the objective of the validation step. In order to perform this task, a representative subset of the extracted data for each technology is randomly selected using the Yamane formulae (1967:886). Each representative subset is verified and keyword revised if non related entries are found.

The end result is a dataset exclusively related to technologies in the nanoscience and technology field. Although restrictive, this strategy captures the dynamics between technologies and not necessarily their absolute number. Finally, at the end of this process a panel data of 442 observations (34 technologies * 13 years) is compiled.

Descriptive statistics

The dataset is composed of 34 nano related technologies grouped in seven types of nanomaterials: bio based, carbon based, ceramic nanoparticles and nanostructures, nanocomposites, metal nanostructures and alloys, polymers, and semiconductor nanostructures. This large spectrum of nanomaterials in the sample aims at accounting for the most promising technologies (i.e. candidate emerging technologies) in the nanoscience and technology field in terms of potential application in different industrial sectors.

A first inspection at the data shows the diversity in research and innovative activities within the sample. For instance, the range of publications and patents largely differ among materials and generally materials undergoing relatively high number of publications will similarly undergo relatively high number of patents. Extreme examples concern on the one hand Carbon Nanotubes with more than 40,000 publications and about 1000 patents and on the other hand Carbon Aerogels, with less than 300 publications and 10 patents for the entire period. In general, the correlation between research and inventive activity is positive and significant in the period under scrutiny (with Pearson correlation ranging from 0.48 to 0.89) which confirms the relevance of looking both at science and technologic dynamics for each technology. The sample shows an average of 695

publications and 8 patents produced across time and materials with large dispersion across materials as summarised in the Table 2

Table 5: Between and Within standard deviations

Variable		*Mean*	*Std. Dev.*	*Min*	*Max*	*Observations*	
publications	overall	694.767	994.8248	2	7318	N =	442
	between		732.2371	22.46154	3209.077	n =	34
	within		684.1747	-2287.31	4803.69	T =	13
patents	overall	7.642534	16.36361	0	147	N =	442
	between		13.21169	.2307692	76.76923	n =	34
	within		9.897909	-64.1267	77.8733	T =	13

Across materials, between and within standard deviations do not strongly differ (732 vs 684), which means that the variation of publication activity in the past ten years across materials is similar to that observed within a material over time. In other words, nanoscience has experienced a large variation in its activities both in time and across technologies. This is shown by the average variation in publications between materials between 22 and 3209 and by the variation over time between -2287 and 4803. The large negative value means that with respect to the overall average, a technology produced 2287 publications less than the average, whereas another generated 4803 publications on top of it. Concerning patents, between standard deviations shows a figure of 13 and within standard deviations about 10, but an important point concerns the large range of patent produced, especially by the inspection of minimum and maximum figures for between variance (i.e. ~0 to 77) and within variance (i.e. -64 and 78).
Overall, although at a lower magnitude the dispersion of material "performances" in terms of patents mirrors the one of publications. Another interesting aspect of the sample concerns the overlapping research and inventive activity of a technology across material types. Indeed, it may be that some materials are not isolated but interact and can be combined with others to create new, more complex nanostructures. The data shows only about 17% of the sample of publications belonging to two different types of materials, and 8% of patents that belong at least to two material types. These are low figures that suggest the relative independence of material types. Note however the significant variability across materials with a maximum reached by polymer with composites for publications and composite with carbon for patents. This can be explained by the fact that composites are a combination of different nanomaterials, which also suggests the nature of the future generations of nanoscience and technology.

Results

The Sharpe ratio is a relative indicator where the performance of a technology is measured with respect to a reference for a given time period. Indeed, the

emerging technologies that are potentially identified are both reference and period dependent. Three references are used in the 1996-2008 period to compute the absolute and relative Sharpe ratios and provide some insight on the technology dynamics at different levels.

The first reference concerns the average growth of databases from where the data was retrieved, namely Scopus and Patstat. Both the relative and absolute Sharpe ratios are computed. Starting with the absolute growth Sharpe ratios, using the entire databases' growth as reference assigns all technologies from the sample as new (c.f. quadrant 4 in Table 1). This is coherent since no single material activity can do better than databases in absolute terms. On the opposite, the relative growth Sharpe ratios assigns all materials as emerging material (c.f. quadrant 1 in Table 1). This means that most of these materials outperform the database in terms of percentage growth, which is evidence that all materials are indeed experiencing significant dynamics. The second reference uses the average growth of nano related publications and patents in the Scopus and Patstat databases. As the set of materials are also a subset of nanoscience and technology, the same observation can be made in absolute terms with respect to the nano field reference: single nano materials do not outperform the nano database subset in absolute growth terms. In relative terms however, when technologies are compared to the nano subset of the database, several materials move from emerging to basic technology (c.f. quadrant 1 to 2 in Table 1). This means that all selected materials grow faster than the nano field in terms of publications but not necessarily in terms of patents. The last –and most informative- reference is computed from the average growth of the selected set of sampled materials and provides a comparative view of each nanomaterial dynamics. Sharpe ratios for each material are computed and emerging trends identified.

The figures below illustrate the Sharpe ratio results –with the average growth of the sample as reference- for each material. Note that the figures were centred to focus on the most important elements. Looking at the Figure 2, emerging technologies are identified as Carbon Nanotubes, Polymer Nanotubes, Polymer Particle, Core Shell, Silica and Silver. Those materials are subject to both a higher absolute growth in research and inventive activities than the sample average. The second quadrant defines more basic materials which can be illustrated by Gold which is relatively more important in research activities than inventive activities. This may be due to cost constraints which are more relevant for inventions than discoveries. The third quadrant concerns the more applied materials such as Graphene or Titania which appear to grow more concerning inventive than research purposes. The last and fourth quadrants are new materials which are undergoing relatively fewer research or inventive activities. Additional light can be shed by using the relative growth performance of each material (still with the sample as reference). In Figure 3, most materials end up in quadrant 4 due to the relatively high growth performance of Graphene, Polymer Nanotubes and Core Shell. All three are small to average sized areas.

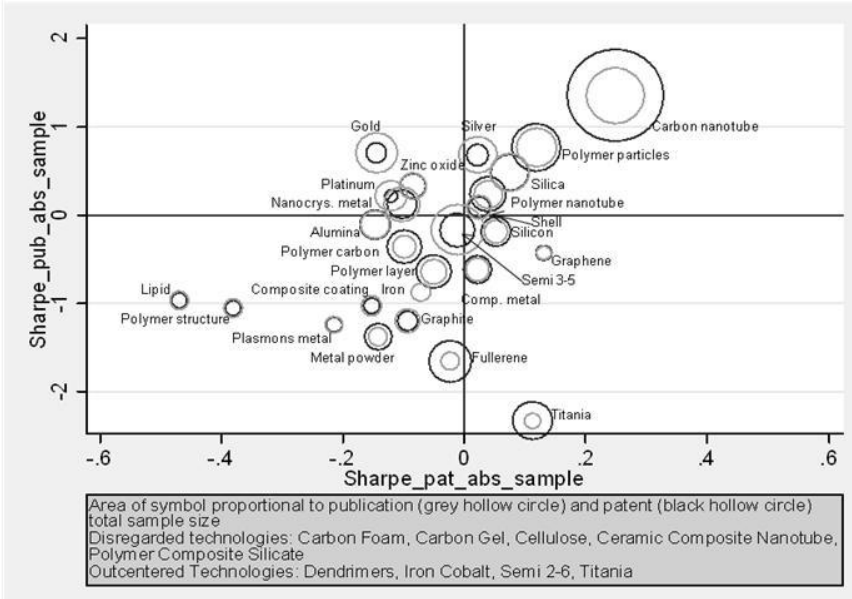


Figure 9: Absolute Sharpe Ratio

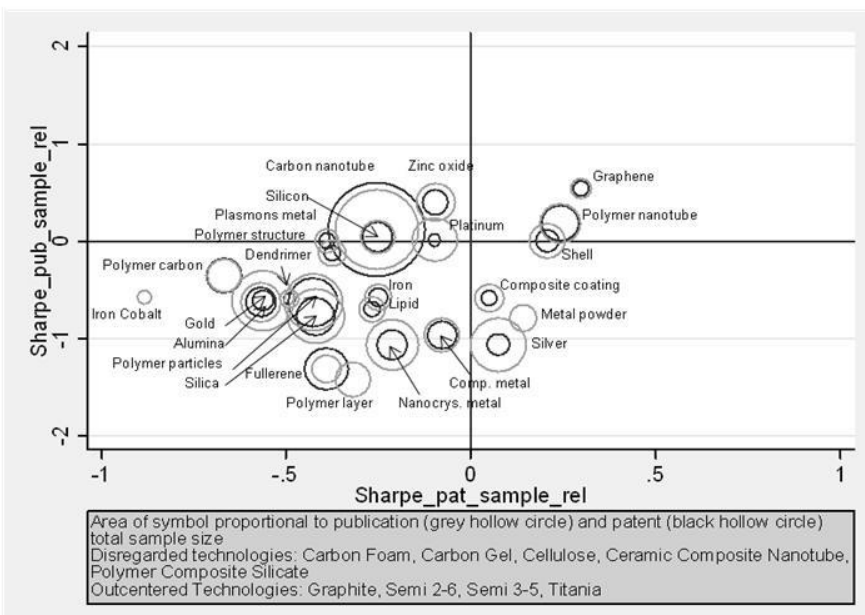


Figure 10: Relative Sharpe Ratio

The two latter technologies are quite stable which means that they are areas which are both getting larger (i.e. absolute performance) and at a faster pace (i.e. relative performance) than the average of the sample. For core shells, this can be explained by the large potential application in biotechnology and medicine. As compared to the absolute results, Graphene moves from the third to the first quadrant, highlighting the high growth of related research activities in percentage, certainly due to the proximity with Carbon Nanotube research. The latter in turn moves from the first to the second quadrant. The interpretation is that although this technology is subject to a large amount of activities, it experiences a slow down of patent growth in relative terms. This could be explained by the challenge faced by nanotechnology and its key technologies to be transferred to mass commercialisation (Schmoch and Thielmann, 2012). Keeping in mind the reference and time dependency of the outcome, results shed light on promising technologies in a tangible, consistent and systematic fashion.

Conclusion

Identifying emerging technologies at an early stage is strategic to better assess the coming S&T challenges, and how to address them best. This is particularly relevant for companies where decisions on their technological trajectories are crucial for their comparative advantage. In the majority of cases, companies' technology choices rely on instinct, personal expertise or vision. This is an uncertain process which can face tunnel vision or lock in effects. This paper aims at supporting this intangible decision-making by providing a novel informative tool which captures the dynamics of key technologies using scientific and technological data.

The novelty of the method is twofold: first, technologies' S&T performances are measured by their underlying publication and patent dynamics connected together by the use of –expert reviewed- keywords; second, to identify which ones are emerging by using the Sharpe ratio both in absolute and relative terms. This method responds and contributes to the technology scanning literature by integrating in the identification strategy the scientific dynamic of technologies as well as the relative nature of emerging stages. With respect to the application of the Sharpe indicator, it appears that using absolute growth favours materials' depicting large activity size; whereas relative growth favours materials undergoing a fast pace of their activities irrespective of their size. The methodology is tested on the field of nanoscience and technology, where the science based nature of the field makes the approach proposed in this paper relevant. Results concerning this empirical case depict no clustering per material types which suggests that the nanoscience and technology field evolves in multiple directions, supporting the "nano" relevance across domains. However, among all these technologies, carbon based materials appear to undergo the largest change, from Carbon Nanotubes in terms of absolute growth to Graphene in terms of relative growth. This can be related to the attractive properties offered by carbon nanostructures and the relative facility in which carbon based materials

can be manipulated and designed. Overall, the method presented in this paper can be easily adapted to any technology sample which makes it a promising tool for technology scanning and ultimately improves companies' technological decisions.

Acknowledgements

Many thanks to Dr. Reiß, Dr Thielmann, Dr. Wydra, D. Van Doren and C. Michels and three anonymous reviewers.

References

Andersen, B. (1999). The hunt for S-shaped growth paths in technological innovations: a patent study. *Journal of Evolutionary Economics* (9), 487-526.

Arthur, W Brian. 1989. Competing Technologies, Increasing Returns, and Lock-In by Historical Events. *The Economic Journal* 99(394):116–131.

BCC. (2010). Nanotechnology: *A Realistic Market Assessment*. Report Code NAN031D

Berger, M. (2007). *Debunking the trillion dollar nanotechnology market size hype*. Retrieved Mars 1, 2012, from Nanowerk: http://www.nanowerk.com/spotlight/spotid=1792.php

Chesbrough, H. W., and Appleyard, M.M. (2007). Open Innovation and Strategy. *California Management Review* 50(1):57–77.

Cohen, W.M, and Levinthal D.A. (1989). Innovation and learning: The two faces of R&D. *The Economic Journal* 99(397):569–596.

Coward, H., and Franklin, J. (1989). Identifying the Science-Technology Interface: Matching Patent Data to a Bibliometric Model. *Science, Technology and Human Values* , 14 (50).

De Sola Price, D. (1984). The science / technology relationship, the craft of experimental science, and policy for the improvement of high technology innovation. *Research Policy*, 13, 3-20.

EAG, N. E. (2009). *Position Paper on Future RTD of NMP for the period 2010-2015*.

Eurostat (2012). Table : Demandes de brevets déposées auprès de l'OEB par année de priorité (pat_ep_nic). http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database

Fagerberg, J. (2005). *Innovation: a guide to the literature*. In J. Fagerberg, D. Mowery, and R. Nelson, The oxford handbook of innovation (pp. 1-26). Oxford University Press.

Frietsch, R., and Schmoch, U. (2010). Transnational patents and international markets. *Scientometrics* , 82, 185-200.

Glänzel, W, and Meyer, M. (2003). Patents cited in the scientific literature: an exploratory study of "reverse" citation relations, Open Access publications from KUL, Katholieke Universiteit Leuven.

Hayashi, A M. 2001. When to trust your gut. *Harvard Business Review* 79(2):58–65, 155.

Huang, C., et. al. (2008). Nanotechnology Publications and Patents: A Review of Social Science Studies and Search Strategies. Working Paper No. 2008-058 , 58.

Hullmann, A., and Meyer, M. (2003). Publications and patents in nanotechnology: An overview of previous studies and the state of the art. *Scientometics* , 58 (3), 507-257.

Järvenpää, H. M., Mäkinen, S. J., & Seppänen, M. (2011). Patent and publishing activity sequence over a technology's life cycle. Technological Forecasting and Social Change, 78(2), 283–293.

Kline, S., and Rosenberg, N. (1986). *An Overview of Innovation*. In N. R. Council, The Positive Sum Strategy: Harnessing Technology for Economic Growth (pp. 275-305). National Academy Press.

Malanowski, N., Heimer, T., Luther, W., and Werner, M. (2006). *Growth Market Nanotechnology*. Weinheim: WILEY-VCH Verlag GmbH & Co. KGaA.

Noyons, E., et. al. (2003). *Mapping Excellence in Science and Technology Accross Europe: Nanoscience and Nanotechnology*. Report to the European Commission.

OECD. (1993). *Frascati Manual* (Fifth edition ed.).

OECD. (2009). *Nanotechnology: An Overview Based on Indicators and Statistics*. OECD.

Popper, E., and Buskirk, B. (1992). Technology life cycles in industrial markets. *Industrial Marketing Management* , 23-31.

Porter, A., Youtie, J., Shapira, P., and Schoeneck, D. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research* , 10, 715-728.

Porter, A..L., and Cunningham S.W. 2005. *Tech Mining Exploiting New Technologies for Competitive Advantage*.Wiley & Sons.

Ratner M., and Ratner D. (2004). *Nanotechnology*. Pearson Education.

Reiß, T., and Thielmann, A. (2010). Nanotechnology Research in Russia - An Analysis of Scientific Publications and Patent Applications. *Nanotechnology Law & Business* , 7, 387.

Reiß, T., Hartig, J., and Schmoch, U. (2009). *ERACEP: Emerging Research Areas and their Coverage by ERC-supported Projects*. EU project research paper.

Rohrbeck, R. 2006. Technology Scouting – Harnessing a Network of Experts for Competitive Advantage. *Technology Management*:1–15.

Roper, A., Cunnigham, S., Porter, A., Mason, T., Rossini, F., and Banks, J. (2011). *Forecasting and management of technology*. Wiley & Sons.

Rosenberg, N. (1982). *Inside the Black Box*. Cambridge University Press.

Schmoch, U. (1997). Indicators and the relations between science and technology. *Scientometrics* , 38 (1), 103-116.

Schmoch, U., and Thielmann, A. (2012). Cyclical long-term development of complex technologies - premature expectations in nanotechnology? *Innovation Systems and Policy Analysis*, 31.

Sharpe, W. F. (1994). The Sharpe Ratio. *Journal of Portfolio Management*..21 (1) pp. 49-58

Smith, K. (2005). *Measuring Innovation*. In J. Fagerberg, D. Mowery, and R. Nelson, The oxford handbook of innovation (pp. 148-177).

VDI. (2010). *Meta-Roadmap Nanomaterialien: Zukünftige Entwicklunden und Anwendungen*.

Watts, R. J., & Porter, A. L. Innovation forecasting. , 56 Innovation in Technology Management The Key to Global Leadership PICMET 97 25–47 (1997). IEEE. doi:10.1109/PICMET.1997.653329

Wydra, S., Haas, K.-H., Jungmittag, A., Reiss, T., and Thielmann, A. (2012). *Economic foresight study on industrial trends* Draft of the final report, Karlsruhe.

Yamane, T. (1967). *Statistics: An Introductory Analysis*, 2nd Ed., New York: Harper and Row

IDENTIFYING EMERGING TOPICS BY COMBINING DIRECT CITATION AND CO-CITATION

Henry Small[1], Kevin W. Boyack[2] and Richard Klavans[3]

1 hsmall@mapofscience.com
SciTech Strategies, Inc., Bala Cynwyd, PA 19004 USA

2 kboyack@mapofscience.com
SciTech Strategies, Inc., Albuquerque, NM 87122 USA

3 rklavans@mapofscience.com
SciTech Strategies, Inc., Berwyn, PA 19312 USA

Abstract
We present a novel approach to identifying emerging topics in science and technology. An existing co-citation cluster model is combined with a new method for clustering based on direct citation links. Both methods are run across multiple years of Scopus data, and emergent co-citation threads in a specific year are matched against the direct citation clusters to obtain the emergent topics ranked by a difference function. The topics are classified and characterized in various ways in order to understand the motive forces behind their emergence, whether scientific discovery, technological innovation, or exogenous events. Cross-sectional analysis of citation links and paper age are used to study the process of emergence for discovery based science topics.

Conference Topic
Research Fronts and Emerging Issues (Topic 4); Modeling the Science System, Science Dynamics and Complex System Science (Topic 11)

Introduction
Researchers in information science have long pondered how and why scientific topics emerge. Derek Price famously analyzed the emergence of the topic of N-rays using a citation network represented as a matrix (1965). Eugene Garfield studied the development of genetics by constructing a node and link citation network that he called a historiography (Garfield, Sher, & Torpie, 1964). Later on co-citation clusters were used to detect emergence (Small, 1977), and more recently co-authorship networks (Bettencourt, Kaiser, Kaur, Castillo-Chavez, & Wojick, 2008) and direct citations (Shibata, Kajikawa, Takeda, & Matsushima, 2008) have been used for the same purpose.
Methods differ in the degree of foreknowledge used. Most rely on a case study approach where a literature search is conducted for a specific topic expected to be emergent, and then methods are used to verify that, in fact, emergence has

occurred. These might be termed local methods because only a literature local to the targeted topic is used. More *a priori* or global approaches, in contrast, make no assumptions about what new areas might have emerged. Global approaches are based on a comprehensive analysis of an entire literature database by methods such as cluster analysis using co-citation, bibliographic coupling (Boyack & Klavans, 2010), or other methods such as topic modelling (Blei & Lafferty, 2007). An important new methodology which uses simple citation links has recently been developed (Waltman & Van Eck, 2012) which uses a variant of modularity clustering and takes normalized direct citation links as input. The method arrives at an assignment of papers to clusters by maximizing a function that rewards linked papers if they are in the same cluster and penalizes them if the papers in the same cluster are not linked. An optimization algorithm is used to maximize the function. Interestingly this new method turns the original local methods of Price and Garfield into global methods with the ability to automatically break up huge multiyear citation link databases into what are, in effect, separate historiographs. In this paper we will use a unique marrying of two global methodologies, direct citation clustering and co-citation clustering, for the purpose of identifying emerging topics in science and technology.

Methods

The co-citation method forms clusters of cited papers based on their joint citation in an annual slice of a citation database, and assigns current papers from that annual slice to one or more of the clusters based on their referencing patterns. The resulting clusters tend to be small and narrowly focused at the scientific problem level. The annual solutions are then merged to form threads which connect clusters in adjacent year slices based on shared cited papers (Klavans & Boyack, 2011). This merges the yearly cluster slices into a longitudinal picture. The resulting threads can be classified by their duration. For example, possibly emergent threads for a given year are considered to be those that begin in the previous or current year, that is, are only one or two years old. It is then possible to identify all papers from a given year that belong to potentially emergent threads.

Unlike co-citation which relies on the joint citation of earlier papers, the direct citation clusters are based simply on the citation of individual papers by each other and finds local concentrations of citation links by maximizing a modularity criterion. The process generates clusters that are much larger and more broadly focused than the co-citation model. The resulting direct citation networks, like the co-citation threads, are of varying duration and involve different numbers of papers per year.

Once the co-citation threads and direct citation clusters are in hand, the task is to select those direct citation clusters that are the most emergent in specific years. The approach used is to count the papers in the direct citation clusters that belong to emergent threads (one or two years old) in the co-citation model. This is done on a year by year basis, so the direct citation clusters having the highest emergent

counts in a given year can be identified. In addition, the number of papers in a matching direct citation cluster in a set of prior years (greater than two years prior to the emergent year) is subtracted from the emergent year counts to avoid selecting areas with high publication activity in prior years. This ensures that the emergent topics are increasing in size in addition to containing many papers belonging to emergent threads. There are of course numerous variations of selection criteria that could be attempted, but by combining evidence from both forms of analysis we can take advantage of the high precision of the co-citation model and the stronger growth characteristics of the direct citation model. The difference between the emergent year counts and the prior year counts provides a metric on which to rank the emergent topics in a given year. We call this the emergence differential.

Figure 1 is an example of how a direct citation cluster is matched with emerging co-citation threads. The topic is computed tomography angiography and the year of emergence is 2007. The graph shows the growth in number of citing papers by year in the direct citation cluster, superimposed on which are matching co-citation threads which start in 2007 and hence are considered emergent. The numbers of papers in emergent threads that match the direct citation cluster are given in the thread boxes. Only some of the matching threads are shown. The sum of the matching papers minus papers prior to 2005 in the direct citation cluster gives the emergence differential.

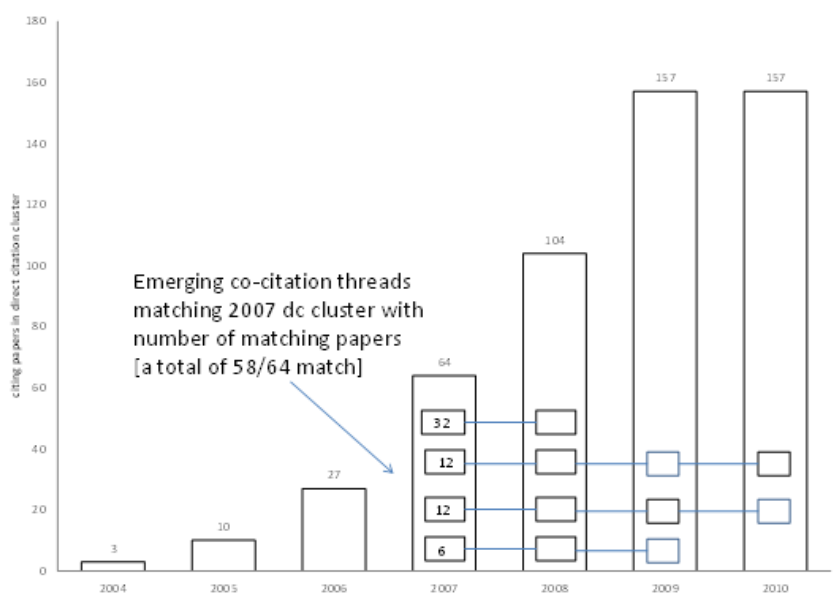


Figure 1. Matching a direct citation cluster and emerging co-citation threads on the topic of computed tomography angiography. The matching papers in 2007 are given in the thread boxes. The number of papers in the direct citation cluster is above each bar.

The data set used is a 15 year Scopus database (1996–2010) under a special arrangement with Elsevier. Direct citation clustering was carried out on this compilation using CWTS open access software (Waltman & Van Eck, 2012). Existing co-citation clusters and threads were also used covering the same time period. The years 2007-2010 were selected for identification of the top 25 emerging topics. The emerging threads (one or two years old) were identified for each year and their papers matched against the direct citation cluster papers for the same year. The number of matching papers minus the papers in the direct citation cluster greater than two years prior to the emergent year gave the emergence differential which was used to rank the topics in each year. A total of 71 distinct topics were selected across the four years, 50 of which appeared in only one year, and the remaining 21 in two or more years. Six topics were in the top 25 for three years, but none appeared in all four years. We will focus here on the topics for 2010 which are listed in Table 1.

Results

The first column of table 1 gives the rank number of the direct citation cluster determined by sorting the emergence differential. A topic name is given in the second column which is based on a manual analysis of the titles and abstracts of 2010 papers in the intersection of the direct citation and emerging co-citation clusters. The third column labelled "type" is a categorization of the type of event mainly responsible for the emergence. We consider three types of events: discovery, innovation and exogenous. The categorization was made by examination of the 2010 papers in the topic and the papers they cited. "Discovery" refers to scientific areas where an unexpected finding is made or fundamental knowledge is gained. An example is the first topic on the list, iron-based high temperature superconductivity, which was a discovery of superconductivity in a new class of materials not previously thought to be a good candidate for superconductivity.

The "innovation" category refers to areas of technology where existing science or technology is used to create new devices or capabilities that serve specific purposes. An example is cognitive radio which takes a new approach to assigning radio spectrum. The third category "exogenous" refers to factors external to science and technology, such as natural disasters, health threats, or societal events with major impacts such as the launch of a new web product or a government standard. An example is the second topic on the list, the swine flu pandemic of 2009, in which the global spread of a virus mobilized the health care community to understand and combat the disease. If an innovation or discovery topic also involves an exogenous event, a combined code is used. For example, the flu pandemic is considered both a discovery and exogenous because a new virus was discovered and it was a worldwide health event. Another example is topic 18 on crystallographic evaluation where a new software service was introduced to validate crystal structures. It should also be clear that discovery topics can also

involve elements of technological innovation and vice versa. What is sought here is the main catalyst of emergence.

Table 1. 2010 top 25 emerging topics. Abbreviations: r = rank; dis = discovery; inn = innovation; exo = exogenous; year Ev = year of event; year HC = year of most cited paper; year Em = year of first emergence; Ev to HC = time lag from event to most cited paper; Ev to Em = time lag from event to first emergence; H = H index.

r	*label*	*type*	*year Ev*	*year HC*	*year Em*	*Ev to HC*	*Ev to Em*	*H*
1	iron-based superconductors	dis	2008	2008	2008	0	0	48
2	swine flu (H1N1) pandemic	dis/exo	2009	2009	2009	0	0	22
3	spectrum sensing in cognitive radio	inn	2005	2005	2007	0	2	26
4	graphene nanosheets and nanocomposites	dis	2006	2004	2010	-2	4	30
5	Horava-Lifshitz quantum gravity	dis	2009	2009	2010	0	1	24
6	graphene oxide nanosheets	dis	2008	2004	2010	-4	2	22
7	induced pluripotent stem-cells	dis	2006	2006	2008	0	2	27
8	MapReduce framework	inn/exo	2007	2008	2010	1	3	13
9	signal recovery from compressed sensing	inn	2006	2006	2009	0	3	27
10	graphene transistors and optical devices	dis	2005	2004	2010	-1	5	15
11	zigzag graphene nanoribbons	dis	2006	2004	2009	-2	3	22
12	cardiovascular events in type 2 diabetes	dis/exo	2008	2008	2008	0	0	14
13	transformative optics	dis	2006	2006	2009	0	3	26
14	spectrum allocation in cognitive radio	inn	2005	2005	2010	0	5	11
15	IDH1 and IDH2 mutations in cancer	dis	2009	2009	2010	0	1	16
16	epitaxial graphene	dis	2006	2004	2010	-2	4	23
17	H1N1 pandemic and seasonal flu	dis/exo	2009	2009	2010	0	1	10
18	crytallographic validation	inn/exo	2009	2009	2010	0	1	10
19	social tagging	inn/exo	2004	2006	2007	2	3	15
20	mechanical properties of graphene	dis	2008	2008	2010	0	2	16
21	online social networking	inn/exo	2006	2007	2010	1	4	7
22	gold nanocrystals	dis	2007	2007	2009	0	2	14
23	cloud computing	inn/exo	2006	2009	2010	3	4	10
24	cognitive radio networks	inn/exo	2003	2006	2010	3	7	8
25	metal-organic frameworks	dis/exo	2009	2009	2009	0	0	16

"Discovery" was the most common category with 12 topics. The combination of "discovery/exogenous" had four topics, and these were mostly medical such as the flu virus or a drug trial (topic 12). "Innovation" had only three topics, for example, a new mathematical approach to signal compression (topic 9). The

combination "innovation/exogenous" had, however, six instances, suggesting that technology areas often have an exogenous component. Many of these combinations were computer science oriented involving, for example, a new programming system (topic 8) or launch of a new web service (topic 21) that stimulates research. Overall "discovery" applied to about two-thirds of topics, "innovation" to one-third, and about 40 percent of topics had "exogenous" influences.

A more detailed analysis of the causative factors for emergence suggests that in most cases the publication of a new idea is what sets the stage for the emergence. Fifteen of the 25 topics follow this pattern. In other cases the causative event was the launch of a technology such as cloud computing services (topic 23) or a new data management framework from Google (topic 8). Also government actions such as DARPA's architecture for cognitive radio (topic 24), or the failure of a clinical trial (topic 12) can spark new research.

The fourth column labelled "year Ev" gives the year of the event. In cases where a specific paper is driving emergence, this is the publication year of the paper. This year may or may not correspond to the year of the most cited paper given in the fifth column labelled "year HC". Citation counts are determined by collecting all references from the 2010 papers that are in the intersection of the direct citation cluster and the emerging co-citation threads. Hence, this count is local to a specific set of 2010 papers and differs from the global citation count found in Scopus. Local citation counts are used because we want to assess the importance of the paper to the specific topic. Examples of where the most cited paper differs from the paper that appears to have directly stimulated the topic are some of the graphene related areas. The most cited paper for these topics is usually the original graphene discovery paper by Novoselov and Geim (2004), while the paper most germane to the specific graphene topic often corresponds to a less cited paper, but usually within the top three or four.

The sixth column labelled "year Em" is the year in which the topic was observed to emerge in the top 25 going back to 2007. Because we have generated top 25 lists for each year from 2007 to 2010, it is possible that a given topic will be in the top 25 for multiple prior years. This is illustrated in Figure 2 which plots the rank of topics which have appeared in the top 25 in three consecutive years from 2007 to 2010. For example, the iron-based superconductor topic was ranked first for three consecutive years from 2008-2010, while induced pluripotent stem-cells rose from rank 19 in 2008 to rank 7 in 2010, and social tagging fell from rank 1 in 2007 to rank 19 in 2010. Fourteen of the 25 topics in 2010 appeared in the ranking for the first time in 2010, and it is likely that several of these topics will fall out of the top 25 ranking in 2011.

The seventh and eighth columns labelled "Ev to HC" and "Ev to Em" give two time lags of interest: the time lag from the emergence event to publication of the most cited paper, and the lag from the event to the year of first emergence. In the former, lags will be positive if the most cited paper is published after the emergence event and negative if the most cited paper precedes the key event. The

negative time lags are due to the graphene discovery paper being published prior to the highly cited paper closest to the topic in content. Positive time lags tend to be associated with exogenous stimuli, such as a software system, web products, or government standards that stimulate research and result in highly cited papers at later dates. Across all topics, the average lag from event to most cited paper is near zero. The second type of lag shown in the column labelled "Ev to Em" is more a measure of our system's ability to detect emergence at an early stage. Large positive lags indicate a delay in detection, and there are no negative lags. The average delay in detection across the 25 topics is 2.5 years, and the largest lags include both discovery and innovation cases where delays may be due to technical or conceptual problems, as was possibly the case with some of the graphene topics which were technically difficult.

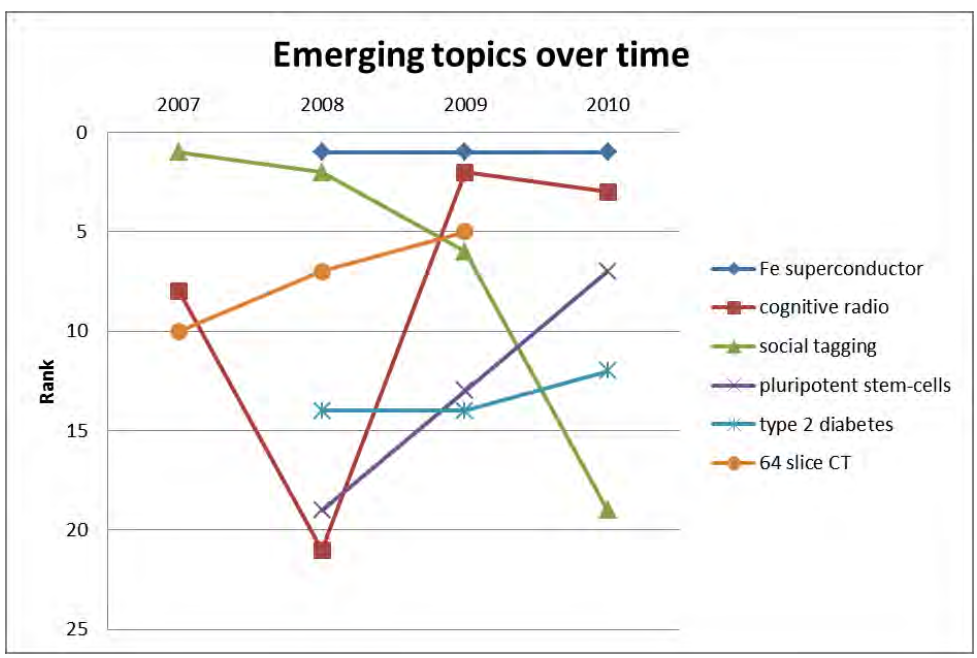


Figure 2. Change in rank of topics in top 25 that appear in three or more years 2007 -2010.

The last column labelled "H" gives the H index, the number of papers N cited at or above N times. This indicates the number and citedness of highly cited papers in the topic. The data suggest that low H values are associated with topics which are driven by exogenous events, such as swine flu, cloud computing, and social tagging. As one would expect, the H indexes are higher for topics associated with specific discovery or innovation papers. The highest H index is for iron-based superconductivity (topic 1), clearly a discovery based topic, while the lowest is

online social networking (topic 21) which is focused on analyses of data from social network services such as Twitter and Facebook.
The topics were also coded for indications of any practical applications that researchers hoped to achieve. Interestingly all of the topics, with the exception of quantum gravity (topic 5), foresaw some type of practical application. About half the topics envisioned specific devices or physical products, while the other half anticipated improvements in services, for example, health care or software.

Validation

In the absence of a definitive list of emerging topics against which to evaluate this list, we fall back on other types of evidence to corroborate that the topics are of current importance, such as awards to authors of most cited papers or recognition in the science press. The awards should be relevant to the topics and post-date the highly cited work in question. Two Nobel Prizes were related to the topics, one for graphene awarded to Novoselov and Geim in 2010 (topics 4, 6, 10, 11, 16, 20), and another to Shinya Yamanaka in 2012 for induced pluripotent stem-cells (topic 7). Graphene was also named a runner-up to "Breakthrough of the Year" by Science in 2009. Both graphene and induced pluripotent stem-cells have been the object of recent bibliometric studies (Chen, Hu, Liu, & Tseng, 2012; Shapira, Youtie, & Arora, 2012; Shibata, Kajikawa, Takeda, Sakata, & Matsushima, 2010).
Other highly cited authors also received recognition. In 2009 Hideo Hosono received the Bernd T. Matthias Prize for his discovery of iron-based high temperature superconductivity (topic 1), and in 2008 the topic was named a runner up to "Breakthrough of the Year" by Science. Sir John Pendry was awarded the UNESCO-Niels Bohr gold medal in 2009 and the 2010 Willis E. Lamb Award for Laser Science and Quantum Optics for his work on transformative optics and meta-materials (topic 13). In 2008 David Dohono received the IEEE Information Theory Society Paper Award for his work on compressed sensing (topic 9), an award he shared with the author of the second most cited paper in the topic Emmanuel Candes. In 2010 Anthony Spek received the Kenneth Trueblood award for his work in chemical crystallography and crystallographic computing (topic 18). In addition, the swine flu virus (topics 2 and 17) was named "virus of the year" by Science in 2009, and in 2008 IDH1 and IDH2 mutations in cancer (topic 15) was named a runner up to "Breakthrough of the Year" by Science (topic 15).
While this search for awards is necessarily incomplete, it provides evidence that at least some of the topics and their highly cited authors have received recent recognition for work that has topical relevance.

Citations during emergence

To gain a better understanding of the process of emergence, the pattern of citations was examined during the period of emergence for the first ranked topic – iron-based superconductivity. The analysis is based on all citation links extracted

from the direct citation cluster for this topic. In this case a specific discovery paper had appeared in 2008 which was critical to the topic. The procedure was to make annual time slices into the citation network and compute the most cited papers in each year.

Table 2 gives the ten most cited papers for each of three years, 2007-2009 which spans the year of emergence 2008. We use letter codes to identify the papers and also show the age of the cited papers with respect to the citing year. The discovery paper is indicated by an asterisk, and the letter code for the paper is underlined if the paper continues from the prior year.

First we observe a dramatic increase in the H index across the time slices coinciding with the appearance of the discovery paper at the top of the ranking in 2008 when H goes from 3 to 30. Of course, this goes hand in hand with a rapid increase in the number of papers and citations in the direct citation cluster. Second we see a decrease in the age of the cited papers. In the year of emergence the top seven papers have an age of 0, that is, were published in the citing year. Third we see a low continuity of cited papers prior to emergence and a high continuity of cited papers following emergence. Of course, high post-emergence continuity leads to an aging of the highly cited work, which will continue unless new papers become highly cited.

Table 2. Iron-based superconductivity top 10 papers by year during emergence showing paper age, citations and continuity.

Cited paper	*2007 age*	*#cites*	*Cited paper*	*2008 age*	*#cites*	*Cited paper*	*2009 age*	*#cites*
A	1	4	K*	0	277	K*	1	517
B	12	3	L	0	140	T	1	275
C	1	3	M	0	132	L	1	258
D	4	2	N	0	106	M	1	235
E	12	2	O	0	104	U	14	202
F	12	2	P	0	96	O	1	193
G	6	2	Q	0	93	N	1	169
H	6	2	R	13	84	Q	1	166
I	5	2	S	13	79	P	1	143
J	5	2	C	2	79	V	14	131
		H=3			H=30			H=51

_ underline – continuing from previous year
* discovery paper

This suggests that the discovery event was sufficiently persuasive to immediately dominate the community, stimulate a new crop of compelling findings and carry this interest forward in time. We do not know yet whether this pattern holds for other topics in the list, particularly those that are not so clearly associated with specific discovery papers. Nevertheless the results suggest a general pattern which might hold for discovery-based science where the combined factors of citedness, age, and continuity are important indicators.

Discussion

Despite the fact that citation data are often regarded as biased toward science, we are struck by how strongly technology-based topics are represented. These topics were generally categorized as innovation. Eight of the topics are clearly technology-based, and a number of other more science-based areas such as epitaxial graphene, metal-organic frameworks and transformative optics have important technological components. Five of the technology topics are oriented toward computer science, and their appearance possibly reflects the strong representation of this subject in the Scopus database.

Since one factor in our detection methodology is growth in the direct citation network, we could ask whether the topics identified are prone to bandwagon effects. Such a tendency could be the result of an availability of a large pool of researchers with adequate support to be able to rapidly exploit a new finding. Such might be the case, for example, with the high temperature superconductivity community within materials science and applied physics. Another way to pose this question is to ask why we do not see more topics in basic physics, chemistry, and biology, and whether such topics may have less dramatic growth characteristics? Perhaps varying the selection parameters for matching direct citation clusters and co-citation threads would give a stronger representation of these disciplines.

Another feature of the list that requires further research is the repetition of topics within the top 25, such as the appearance of six graphene related topics and three on cognitive radio. It is perhaps not surprising that a material of such practical and theoretical interest as graphene should have such a strong representation. It is usually possible to draw subtle distinctions between the various subtopics dealing with graphene, and these distinctions are usually apparent in the citing papers as well as a different mix of highly cited papers. The most likely explanation for this repetition is an overly granular setting of the underlying direct citation clustering parameters, or perhaps also the proneness of citation data to fragmentation.

A more fundamental question regarding the methodology we have used to identify emerging topics is whether alternative methodologies would perform equally well, or whether known cases of emergence during the 2007-2010 period were missed. For example, could either the direct citation clusters or co-citation threads be used on their own to detect emergence? Direct citation clusters have measurable growth properties so a slope analysis looking for inflection points might be possible. Alternatively, emergent co-citation threads could be grouped using some alternative bibliometric measure independent of the direct citation clustering and used as an emergence indicator. These possibilities remain to be explored, but what we can say now is that the two methods, based on different citation metrics and algorithms, can be used in a complementary manner that takes advantage of the longitudinal and cross-sectional strengths of the respective methods. Lacking any definitive list of emerging topics for the period, we cannot say whether areas have been missed, but a good source of intelligence on this

question can be obtained from the Breakthrough of the Year listings in *Science*, where we have seen some confirmation of our selections, but not a one-to-one match.

Conclusions

It seems clear that specific highly cited papers have played a key role in emergence in 17 of 25 topics, including technological areas such as cognitive radio and compressed sensing. It is likely that most of these discoveries and innovations could not have been anticipated, even though with hindsight we might be able to identify precursor papers in the direct citation network that might foretell possible forthcoming breakthroughs. One task for future research will be to use this list of topics and similar lists from other years to see if common preconditions to discovery and innovation can be found. It is also of interest to study the fate of these emerging topics in later years. Did work continue, decline or disappear? We would not be surprised if some were proved to be errors, dead ends, or continued under their own inertia until well past their prime. Having a reasonably certain inventory of emergent topics as a quasi-gold standard opens up many new research possibilities, for example, studies of sentiment words changes during emergence, or correlated social network or institutional factors.

The role of exogenous events, which was a factor in 40 percent of topics, also deserves further attention. Previous bibliometric case studies have been carried out on topics such as the 9/11 and anthrax terrorist attacks (Chen, 2006; Morris, Yen, Wu, & Asnake, 2003), but perhaps more common exogenous events are disease or natural disaster-related. We do not know how pervasive such influences are or in general the role that extra-scientific factors have in emergence. As we delve more deeply into other topics, we may find further evidence of exogenous stimuli. For example, in metal-organic frameworks (topic 25) it was not immediately obvious that the DOE had issued new targets for hydrogen storage.

Regarding our methodology, we do not know whether we can reduce the average time lag of 2.5 years from the so-called emergence event to our detection of emergence. This may depend on our ability to identify emergent co-citation threads earlier perhaps by adjusting our threading threshold, since we know that the slope of the direct citation cluster growth curve will not be steep at earlier stages. Perhaps an indicator of network structure can also be devised.

In modelling the emergence process at the paper level we need to further investigate the factors of citedness, paper age, and continuity of the highly cited papers. These variables might eventually be part of an emergence index, in conjunction with the topic growth rate. Obviously the precision of topic paper identification is critical in such an analysis, and the combination of direct citation and co-citation methods used here has probably contributed to this accuracy.

Clearly at this stage we are engaged in detection and not prediction of emergence. Perhaps the most important implication of the present work is that detection by citation-based methods is broadly feasible using a global approach to data

analysis rather than a local or case study approach which up to now has been the predominant approach. Whether detection can be enhanced by a deeper analysis of full texts, or application, for example, of word-based methods remains to be seen.

Acknowledgments

Scopus data from 1996 to 2010 were generously provided by Elsevier under an agreement with SciTech Strategies, Inc. We would like to thank Ludo Waltman and Nees Jan van Eck and CWTS for use of the direct citation clustering software. This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20152. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Bettencourt, L. M. A., Kaiser, D. I., Kaur, J., Castillo-Chavez, C., & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics, 75*(3), 495-518.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics, 1*(1), 17-35.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Chen, C., Hu, Z., Liu, S., & Tseng, H. (2012). Emerging topics in regenerative medicine: A scientometric analysis in CiteSpace. *Expert Opinion on Biological Therapy, 12*(5), 593-608.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia: Institute for Scientific Information.

Klavans, R., & Boyack, K. W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology, 62*(1), 1-18.

Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology, 54*(5), 413-422.

Novoselov, K. S., Geim, A. K., Morozov, S. V., Jiang, D., Zhang, Y., Dubonos, S. V., et al. (2004). Electric field effect in atomically thin carbon films. *Science, 306*(5696), 666-669.

Price, D. J. D. (1965). Networks of scientific papers. *Science, 149*, 510-515.

Shapira, P., Youtie, J., & Arora, S. (2012). Early patterns of commercial activity in graphene. *Journal of Nanoparticle Research, 14*(4), art. num. 811.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation, 28*, 758-775.

Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., & Matsushima, K. (2010). Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting & Social Change, 78*(2), 274-282.

Small, H. (1977). A co-citation model of a scientific specialty: A longitudinal study of collagen research. *Social Studies of Science, 7*(139-166).

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378-2392.

IDENTIFYING LONGITUDINAL DEVELOPMENT AND EMERGING TOPICS IN WIND ENERGY FIELD

Ssu-Han Chen[1], Mu-Hsuan Huang[2], Dar-Zen Chen[3*]

[1] *ssuhanchen@ntu.edu.tw*
Department of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan (R.O.C)

[2] *mhhuang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, Taipei, Taiwan, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan (R.O.C)

[3*] *Corresponding Author: dzchen@ntu.edu.tw*
Department of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan (R.O.C)

Abstract
To manage strategic deployment timely, investors are looking for visualization of the chronological development and potential technologies. A methodology combines patentometrics, social network analysis, clustering algorithm and text mining is proposed to achieve the task specified in this study. This method divides a field into tight-knit technology communities over time and their inter-year continuity is tracked. Following seven statements are examined as indicators: pace of technological progress, patent age, citation of scientific literatures, pending for patents, frequency of interdisciplinary phenomenon in the cited references, context cohesiveness, and public sector participation. Recently, wind energy has attracted significant attention in the wake of the implementation of global energy policies and greater awareness amongst people of the importance of renewable energy. A set of wind energy patents were retrieved from the database of United States Patent & Trademark Office (USPTO) in this study. These patents defined a number of main evolving technology trajectories. Technological trajectories found include control systems of wind power generator, transmission systems, vertical-axis wind turbines, design of airfoil, style and materials, steering control equipment of blade, and connection methods of grids. Furthermore, these major emerging topics can be divided into two categories: rotor blades with variable angle and speed, and super-grid connection.

Conference Topic
Research Fronts and Emerging Issues (Topic 4) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction

Patentometrics has been used to construct social networks using patents. This paper aims to understand underlying community-level structures of a network, particularly with patents that expand field expertise (Newman, 2004b). Patents collected from early research captured static properties of a given snapshot, ignoring the fact that most real-world communities are quite dynamic. Recent researches use dynamics to display evolutions of technology communities, which is a crucial element that depicts technology development across particular time span. The analysis is conducted by sliding window under a sequence of time-points, observing the dynamics of patent and tie communities. The evolution of technology community at given snapshots can be depicted by temporal information, showing that monitoring the technology evolution can effectively categorize and track the changes of trajectories powered by the technology dynamics.

It is also important for different stakeholders to identify emerging topics among various evolving technologies so as to manage their portfolios. Emerging topic is denoted as a technology in its infant or adolescence stage that some observers may consider as candidates for partially or a substitution of legacy technologies (wiseGEEK, 2012). Areas of intensive innovation are useful for piloting government policies and as intermediaries to speed industrial growth through incentive programs and funding subsidies. In addition, the practice helps industries to seek innovation breakthroughs, investment allocations and build competitive advantage in cooperation with scholars securing research momentum of promising topics.

As many subject areas become fields of interest, it has become important for many to measure its popularity. Different approaches to the study results in different results. Indicators such as the changing of cluster size (Small, 2003), the currency index (Small, 2006), the average age (Kajikawa Yoshikawa, Takeda & Matsushima, 2008) have been used to measure the benchmark of the emerging topics. However, single indicator is insufficient to describe the complex task of identifying emerging topics. Multi-dimensional viewpoints are therefore proposed from various aspects. Chang and Breitzman (2009) argued that clusters with higher public sector participation, science linkage, and originality are more likely to signal emerging and higher risk areas. Upham and Small (2010) explored a community emergence model using linear regression technique, concluding that the coefficients for both endogeneity and multi-disciplinarity are positively significant. Guo, Weingart and Börner (2011) combined three hypotheses to describe features of emerging areas: first the rapid increase of usages of specific terms, increase number of new authors, and last interdisciplinized references. Multiple indicators are employed in this study to represent the emerging topics from different insights.

This study explores the technology evolution of wind energy and its emerging topics. The related patents issued from 2001 to 2011 are collected and analysed. The patent citation network of each snapshot is built from bibliometric coupling

(BC) analysis and the corresponding technology communities are detected using clustering algorithm. The temporal information is used to track the technology community at a time when it is evolved into certain community in the next snapshot. Through experts' summarizations based on a series of key terms, technology topics over successive snapshots are identified. We ultimately identify the emerging topics in the last snapshot by introducing the multiple indicators. The following voting panel is introduced: (1) the pace of technological progress for emerging topics is fast, (2) emerging topics become more current, (3) emerging topics cite more scientific literatures, (4) a longer pending time occurs in the emerging topics, (5) emerging topics cite interdisciplinary references, (6) emerging topics create cohesiveness quickly, and (7) the public sector's participation in emerging topics is higher. A community would be regarded as emerging if a community obtains four or more of the above-mentioned votes.
The rest of this paper is organized in the following structure; First of all, we explain and justify our research methodology. Then the experimental environment is delineated. Following the discussion of results, we make concluding remarks and further suggestions.

Research Methodology

Identification of technology communities over time

After relevant patents were retrieved, five steps are required to generate timeline plot of technology evolution (Chen, Huang & Chen, 2012; Chen, Huang, Chen & Lin, 2012). The detailed process of the proposed method is shown as follows:

Step 1 Selection of high-impact patents

Community detection is conducted based on core document analysis (Glänzel & Czerwon, 1996), with significant amount of citations of greater technical impact and technical quality. Since the absolute thresholds have disciplinary bias on average citation frequencies (Aksnes, 2003), the study takes relative threshold to filter high-impact patents. High-impact patents are defined as the documents with cited times that are above the average plus one standard deviation derived by the issued patents cited at least once in each annual cohort of the same issue year.

Step 2 Determination of sliding window length

Researchers observed a phenomenon of truncation bias in citation windows, referring to difficulties encountered when deciding upon appropriate window length to evaluate the patent performances in different technology fields (Narin & Hamilton, 1996). Technology cycle time (TCT), the median age of patent backward citation in a particular technological field, has been employed to determine appropriate window length for different technology domains (Chen, Huang & Chen, 2012). The time window is defined by splitting the citation network into equidistant slices (Falkowski, 2009) in overlapped mode to simulate dynamic movement of the patents over time.

Step 3 Selection of the bibliographic coupling pairs

BC is one of the commonly used approaches to measure the similarity of documents, as it provides current and immediate information about patent relationships and for its reinforcing of regions of dense citation. BC measures the similarity between patents by examining the number of references two patents share in common. However, coupling strength is too rough as a measure of similarity, for there is a need to consider the coupling strength as well as the strength of each patent (Persson, 1994). Therefore, coupling strength of the document pairs should be normalized based on Salton's cosine. The number of co-occurrence of references for each document pair is divided by the square root of its number of references. After coupling strengths are normalized, strong Salton's cosine are selected to solve the problem that partial ones are extremely weak. Similarly, a relative threshold is adopted to acquire strong BC pairs whose coupling strength are above the average plus one standard deviation of the Salton's cosine for each snapshot.

Step 4 Detection and identification of technology communities over time

With the information of vertices and ties in a given snapshot, patent citation network can be composed by adjacency matrices. In network analysis, communities are detected using the weighted Girvan-Newman (GN) algorithm (Newman, 2004a) owing to its non-involvement of human judgment to set a *priori* for the number of communities and its suitability for detecting community structure in an undirected and weighted network.

After clustering procedure is carried out, a thematic topic for each community using natural language processing (NLP) is identified so that analysts can better interpret the results of technology communities. First, patents titles and abstracts are collected as a corpus; a purging and cleaning process is undertaken on the corpus by lower case conversion, punctuation and number removal, multiple whitespace stripping, and singularization. Each word is then is tagged as a part of speech (POS) depending on its context in the text (Mitchell, Santorini & Marcinkiewicz, 1993). Three linguistic filters shown in Equation (1) are applied since most meaningful terms consist of nouns, adjectives, and sometimes prepositions (Frantzi, Ananiadou & Mima, 2000). Undesirable words would be excluded with such filters.

$$\begin{aligned} &\text{Noun} + \text{Noun} \\ &(\text{Adj} \mid \text{Noun}) + \text{Noun} \\ &((\text{Adj} \mid \text{Noun})+ \mid ((\text{Adj} \mid \text{Noun})*(\text{NounPrep})?)(\text{Adj} \mid \text{Noun})*)\text{Noun} \end{aligned} \quad (1)$$

Finally, these terms are weighted by term frequency-inverse document frequency (*tf-idf*) to measure the frequency and features of terms in a specific community compared to the other communities. In this study, the terms associated with the top *tf-idf* values in each community are regarded as the characteristic terms (Chen,

Huang & Chen, 2013). This automatic procedure paves the way for identifying a thematic topic in technology community.

Step 5 Presentation of community continuity and timeline plot

To determine the development patterns of continuing high-impact patents from one snapshot to the next, overlapping the successive time slices of data described above. Community strings are formed when two communities of two successive snapshots share at least one common document (Small, 2006). Research or technology evolution is visualized on a timeline plot where communities are drawn as function of their size and average age against time. The communities are plotted two-dimensionally according to the analytical time point of the sliding window and average age. The number of documents in each community is represented by the size of a circle. In such timeline plots, each research or technology trajectory is isolated and consists of at least two successive years of communities linked by a string, enabling us to visualize technology development and trends.

Specifications of the multiple indicators at the emerging stage

An indicator shows when the communities come into being is necessary for this study. This study introduces a multiple indicator model which is instrumental to the diagnosis of recently emerging topics. We specify a series of indicators, explain the rationale about why they were chosen, and then consult to literature and industrial practice for supporting statements. The chosen indicators are available after a patent is granted and have the time-invariance characteristics.

- Technology cycle time (TCT): The pace of technological progress for an emerging topic is faster than non-emerging one. Kayal (1996) proposed. TCT is defined as the average value of median age gaps between the subject patent and other cited patents within community's innovations. In general, TCT is considered the speed of invention, which is a sign of development in the technology (Kayal & Waters, 1999).
- Currency index (CI): Scientific and technological developmentA new development is likely to quickly attract attentions, and then it expands as inventors create innovations based on the patents with original invention. Small (2006) proposed currency index to be defined as the average age of documents relative to a specific time frame, suggesting that an area grows more rapid if there are more recent documents in the same specific area.
- Science linkage (SL): A topics emerges when the number of literature citations increases. SL reveals the contribution of science to technology. This indicator is represented by the average number of scientific papers referenced in a community's patents (Carpenter & Narin, 1983). The literature citations increase with innovative development. (Haupt, Kloyer & Lange, 2007).
- Pending duration (PD): Examination process is more time-consuming in an emerging topic. This indicator is the average time duration of the successful

patents of a community during its application-grant process (Xie & Giles, 2011). Haupt et al. (2007), stated that the examination processes take longer duration for the start-up innovations. This is because original invention of a technological development tends to characterize broader claim range to limit the chances for subsequent inventions in the beginning stage.

- Originality index (OI): An emerging topic builds on its invention from previous technologies. This indicator measures the extent in which patents combine aspects of technology inventions. This is because emerging areas that explore beyond its existing research or technology are more likely to synthesize knowledge across a wide variety of disciplines in its patents citation (Breitzman & Thomas, 2007; Guo, Weingart & Börner, 2011).
- Endogeneity index (EI): Citing and cited actively involves in an emerging topic. A cited patent can be a citing patent at the same time in a specific field. As noted by Upham and Small (2010), the extent to which the patents build on each other for the community may be important for its potential growth. Assignees or inventors in such community are more likely to build on each other's innovation quickly and to create a cohesive paradigm.
- Public sector participation (PSP): Higher public sector participation signifies new trends. Government policy pilots the beginning of innovative activities through funding subsidies. As Breitzman and Thomas (2007) noted, the technology communities containing patents from academic or governmental laboratories suggest a higher scientific content that are more likely to describe early-stage technologies.

The voting panel consists of multi-indicators that select qualified emerging topics. The multiple indicators of communities are extracted by the last two successive snapshots. Then the value of changes between the last two successive snapshots is calculated and is used for identifying emerging topics. For example, one point is given to the community if its value of change of TCT is in decreasing manner and of other indicators in increasing manner. When a community received four or above points, it is regarded as an emerging topic.

Experimental Results

Case profile

To prove the feasibility of the research methodology, the wind energy field was chosen as a case study. In the wake of climate change and global warming, a large amount of investment is expected to flow to the market of wind energy in the coming decade to combat rising oil and gas prices. The rapid development of this field has attracted attention from both inventors and funding bodies. Since patents are viewed as the valid document source for monitoring the development of a technology, the target technology chosen in the study is from the parts of current USPC class 307 (Electrical transmission or interconnection systems), 415 (Rotary kinetic fluid motors or pumps), and 416 (Fluid reaction surfaces). There are a total of 6,149 patents granted during 2001 and 2011 from the database of the United

States Patent & Trademark Office (USPTO). Both attribute data (*i.e.* application date, issue date, title, abstract, assignee type, and USPC class) and relational data (*i.e.* patent or literature citations) of the selected patents were recorded. As shown in Figure 1, the number of issued patents on this subject was stable and steady each year till 2009. Then it increased to that of over one thousand in the year of 2011.

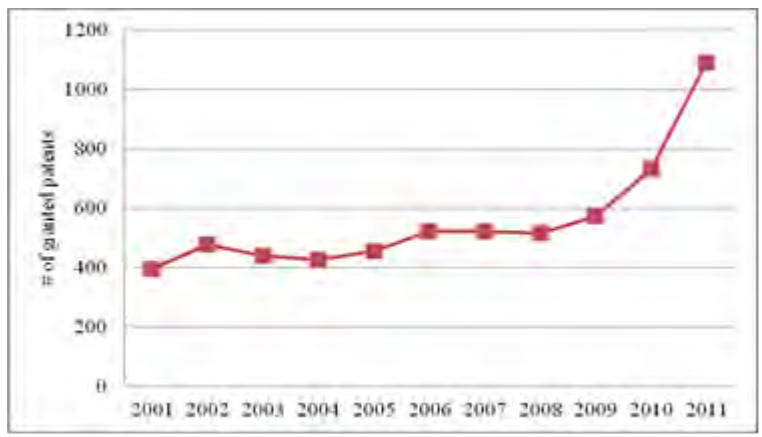


Figure 1. Development of the number of granted patents concerning the wind energy field.

Technology communities over time in wind energy field

After collecting the patents, we analysed the data by a self-programming toolkit under the 'R' environment with the igraph, tm, RWeka, stringr, openNLP, wordnet, and gdata packages (see http://cran.r-project.org/). The date range for selecting high-impact patents is from 2001 to 2011, and the high-impact patents were selected annually. We calculated the average and the standard deviation of cited counts annually, and then the patents which are cited at least above the average plus one standard deviation of cited counts in each annual cohort of patents with the same issued year were selected. There are a total of 6,149 patents and 648 high-impact patents which account for 10.54%, were selected. To decide the length of the sliding window, we calculated the average time lag of the patent inventions upon which a new invention was based at, which yielded a TCT value of 5.01. Consequently, the length of the time span for each citation window in this study is 5. This implies that wind energy field is a fast-developing technology. The criterion of a fifth of window length is used to determine the window step size, which is one year (Moody, Farland, & Bender-deMoll, 2005), reducing the impact of fluctuations on the rolling clustering. The temporal overlap ensures consistency in community composition while allowing new communities to emerge and existing communities to merge, split, or die away (Kandylas, Upham, & Ungar, 2010). After the sliding window were specified, all high-impact patents and relatively strong normalized BC strengths that occurred in this window were

aggregated into a patent citation network for each time point. The relatively strong normalized BC strengths were preserved by selecting the patent pairs that have strengths above the average plus one standard deviation of the normalized BC strengths. There are a total of 19,971 BC pairs of patents and 1,956 strong patents, which account for 9.79% of the total patents. Related patents were assembled as communities through a GN clustering operation, identifying dominant communities to prepare for later size and average age calculation as well as topic detection.

The presentation of community continuity is given in Figure 2, which shows the evolution of a community in the wind energy field. Among these seven trajectories, three of them persist across seven time periods with their mainstream, longer than the other trajectories. Three trajectories appeared respectively in 2007, 2008, and 2010, and have lasted till 2011. Only one trajectory died in the year of 2008.

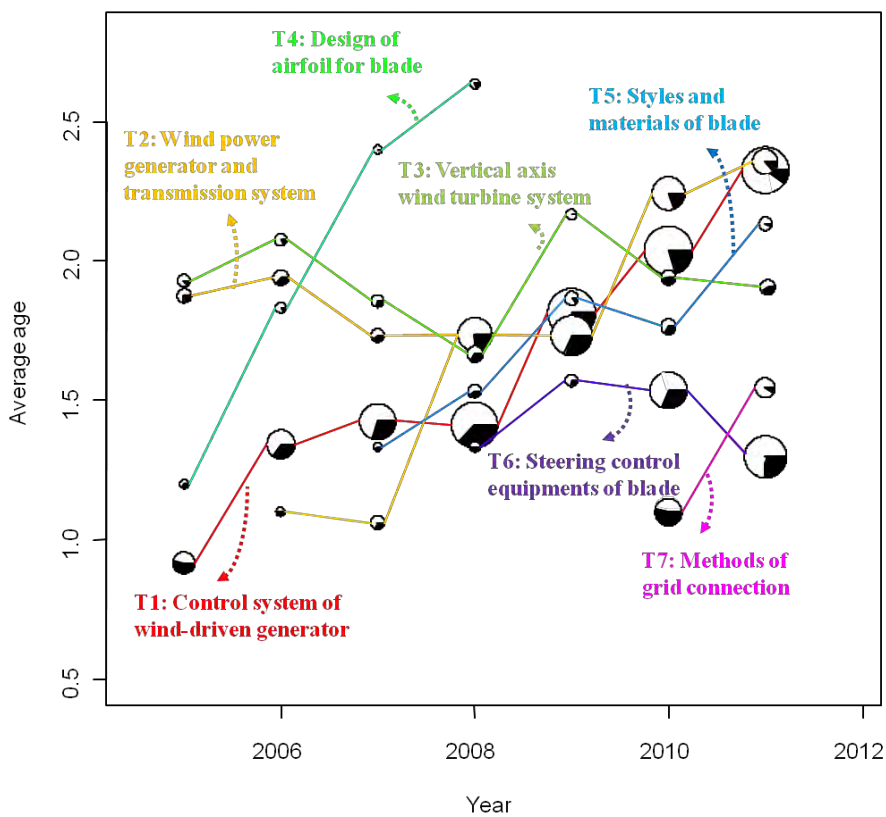


Figure 2. Presentation of technology trajectories in wind energy field.

The evolvement of the major trajectories is discusses as follows:

- Trajectory T1 is about the control system of wind-driven generator. The control system is responsible for the start-stop, shutter, power, loading, speed that have tremendous impact on the operation process and the efficiency of the generator. Components of control system of variable-speed generator such as grid-side rectifier, matrix switcher, power switcher, loading control, aerodynamic control, had received many attentions during 2005 and 2006. Peripheral equipment related to variable-speed generator such as variable-speed driver, pitch controller, rotor icing detector are disclosed. Since 2008, the idea of automatic parameter setting based on measurable data and historical experience has been integrated into power control to optimize the operation. In 2011, the control or prevention mechanisms of doubly-fed induction machine facilitate the operation of machine's activity and enhance the conversion efficiency.
- Trajectory T2 mentions the wind power generator and transmission system. Wind power generates the blade rotating system, and the momentum can be transferred to the generator through the acceleration of the gearbox. In 2005, the research and development of wind power generator and wind energy transmission system were used to reduce the manufacturing and maintenance costs. In 2006-2007, two significant traces appeared. One trace focused on the improvements of concentric gearbox, the method and equipment of the air gap control, and direct-drive wind power generator, in order to reduce the transmission system as well as the size of the cabin and its cooling circuit. The other trace researched and developed the different units, such as layer flow, enhanced diffusion of wind power generator, and multi-impeller generator. In 2008, the method for removable bearing and the improved propeller transmission were proposed for a higher efficiency and stability. In 2009-2011, the technology emphasized the additional methods for power transmission and productive equipment, paying more attention on the development of off-shore wind power generator to solve the friction problem of depletion.
- Trajectory T3 refers to vertical axis wind turbine system, which is a small, low-cost, low-maintenance alternative to horizontal axis currently available on the market. The advantage of this arrangement includes generators and gearboxes can be placed close to the ground, which makes these general usage and maintenance of the components easier. Recently, many kinds of vertical axis wind turbine system are being proposed continuously, covering omni-directional, coupled vortex, imaginary, aerodynamic-hybrid, Savonius, propeller, or pneumatic mixing vertical-axis wind turbines.
- Trajectory T4 is related to the airfoil for wind turbine. Airfoils adopts kinetic energy from wind and push forward generator to produce power. Numerical simulation on the aerodynamic analysis of different kinds of aerofoil of blade was conducted so as to realize the generator performance. Designing an airfoil optimal geometric are taken into account in 2007 and 2008. The

factors include the cutting tooth form, reducing blade, angle of torsion, size, number of blade, etc.

- Trajectory T5 is associated with the styles and materials of blade. Blades are considered as consumption because when exposed outside may result in corrosions, deformations, cracks, attachments, or being struck by lightning. Such disasters caused wear outs and failures, and increase maintenance costs. Therefore, the styles and materials are crucial? The production of blade. In 2007, inventors focused on the design of blades, and the separable blades are the most popular. Then materials such as carbon fibres, glass fibres, or hybrid fibre composites are used to make blades in order to enhance their adjustability. In 2009, blade modules were tested according to multi-dimensional evaluations of weather or environment conditions, which help to reduce damages. Lately, blades are improved throughout adding spar caps, reducing loading, or multi-step/multi-plate types.
- Trajectory T6 is associated with the steering control equipment of blade. The change of wind speed and direction has significant influence on the speed and direction rotation of blade such that power generation changes accordingly. Since 2008, wind speed and direction indicators were invented to provide accurate measurements for critical weather conditions. In order to maintain the reliability of power generator, it tended to install oscillation dampers on wind turbine blades. Lately, rotor blades with variable angle and speed are currently evolving, continuous rated revolution and output power could be expected.
- Trajectory T7 is related to the methods of grid connection. Conventionally, power is distributed from high to low voltage. However, the power distribution of renewable energy is the opposite to the need for micro-grid infrastructure. Much reactive power is absorbed from electric system when wind generators connect to the grids when loading and the instability of voltage increases. The voltage monitoring system such as power control and scheduling schemes, voltage stabilizers, harmonic detections, are then developed. Recently, it tends towards the super-grid connection in which the renewable energy array is devised to allocate distributed generations.

Emerging topics in the wind energy field

Another task of this study is to identify recent emerging topic among communities through a proposed multi-indicators analysis. The multiple indicators are applied to communities at the last two successive snapshots extracted. The results are shown in Figure 3. The values of changes of indicators between the last two successive snapshots are calculated. One point is given to the community if its value of change of TCT is in decreasing manner and of other indicators is in increasing manner. The result of the voting panel is shown in Table 1, where potential emerging topics that received four or above points are identified. Note that public sectors seldom participated in highly cited patents.

PSP indicators of each community come to approximately zero. In sum, we have two emerging topics in T6 and T7, explicated respectively as follows:

- Variable-speed wind turbine: Fixed-speed wind turbines with induction generators were commonly used in the 1980s. Now the trend has shifted toward wind turbines of variable-speed as these turbines generate more energy in a given wind speed regime, and the active and reactive power generated can be easily controlled (Zinger & Muljadi, 1997). There is less drive train mechanical stress, lower aerodynamic noise, and more smooth power fluctuations as the rotor acts as a flywheel. Such kind of system is much more 'grid-friendly'. Although the drawbacks of variable speed are more expensive, the use of complexity has been increased in off-shore applications due to the advantages mentioned above.
- Super-grid connection: A super grid is in a wide range transmission network that creates long distance transmission lines to take advantage of renewable sources that are distantly located. Recently, inventors have wrestled with the problem of taking wind energy from the periphery to a central position in fulfilling the expected increasing electricity demand in the future. While such grids cover great distances, the capacity to transmit large volumes of electricity remains limited due to congestion and control issues. Besides, in order to detect the imbalances caused by fluctuating wind energy and other renewable sources, and to reroute, reduce load, or reduce generation for network disturbances, the inventors have tried to solve the mentioned issues.

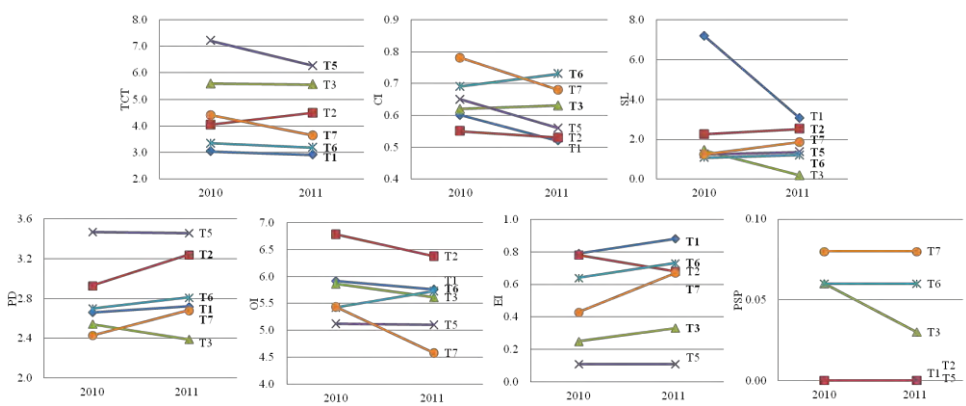


Figure 3. Indicator of communities in last two snapshots in wind energy field.

Table 1. The result of the voting panel.

Trajectory / Indicator	T1	T2	T3	T4	T5	T6	T7
TCT	1	0	0	--	1	1	1
CI	0	0	1	--	1	1	0
SL	0	1	0	--	1	1	1
PD	1	1	0	--	0	1	1
OI	0	0	0	--	0	1	0
EI	1	0	1	--	0	0	1
PSP	0	0	0	--	0	0	0
Total scores	**3**	**2**	**2**	--	**3**	**5**	**4**

Conclusions and Discussion

This study explores the technology evolution and identifies the emerging topics in the wind energy field. Techniques were borrowed from patentometrics, social network analysis, clustering algorithm, and text mining to analyse a set of USPTO-issued patents of the wind energy field longitudinally. The basic idea is to divide a given field into strongly connected communities and to track their technology topics over time in terms of overlapping snapshots. Then multi-indicators are calculated to detect the recently emerging topics and visualize them. The main results are as follows:

The wind energy field encompasses seven major evolving trajectories. The control or transmission systems of power generator and the generator itself (T1 and T2) have attracted the highest interest in the struggle of worldwide patent portfolio. Those two trajectories continued with dominant but aging sizes. Blade (T4 to T6) is another large issue where the design of airfoil, styles, materials, steering control are concerned. Among these three trajectories, steering control equipment of blade is relative dominant and young. More new issued patents have joined in. The vertical axis wind turbine system (T3) and methods of grid connection (T7) clustered significantly in the recent snapshot. The former becomes younger and the latter is the youngest. Among a wide variety of technology communities, variable-speed wind turbine and super-grid connection now have been extensively focused. They have emerged as one of the potential systems, which not only provide renewable energy but also offer good commercial viability in the future.

All in all, the proposed methodology provides knowledge and insight into recent discussion about emerging topic detection by its contribution of multi-indicators analyses. Such research may assist policymakers to decide which innovation is a worthy investment. It could also potentially aid would-be researchers, government officials, or enterprisers in gaining a landscape of worldwide inventions, keeping abreast of current trends, selecting appropriate sub-domains, and making strategic timing of road-mapping. Finally, it is suggested that future researches design

more characteristics concerning emerging topic, improve the simple binary scoring and voting mechanism, and apply them to other potential energy fields.

Acknowledgements

This research is partially supported by the National Science Council, Taiwan, under contract No. NSC102-3113-P-002-029-.

References

Aksnes, D.W. (2003). Characteristics of highly cited papers. *Research Evaluation*, 12(3), 159-170.

Breitzman, A. & Thomas, P. (2007). The emerging clusters project final report. 1790 Analytics LLC Working Paper. Retrieved July 20 2012, from: http://www.ntis.gov/pdf/Report-EmergingClusters.pdf.

Chang, C.K.N. & Breitzman, A. (2009). Using patents prospectively to identify emerging, high-impact technological clusters. *Research Evaluation*, 18(5), 357-364.

Carpenter, M.P. & Narin, F. (1983). Validation study: Patent citations as indicators of science and foreign dependence. *World Patent Information*, 5(3), 180-185.

Chen, S.H., Huang, M.H. & Chen, D.Z. (2012). Identifying and visualizing technology evolution: A case study of smart grid technology. *Technological Forecasting and Social Change*, 79(6), 1099-1110.

Chen, S.H., Huang, M.H., Chen, D.Z. & Lin, S.Z. (2012). Detecting the temporal gaps of technology fronts: A case study of smart grid field. *Technological Forecasting and Social Change*, 79(9), 1705-1719.

Chen, S.H., Huang, M.H. & Chen, D.Z. (2013). Exploring technology evolution and transition characteristics of leading countries: A case of fuel cell field, *Advanced Engineering Informatics*. DOI: 10.1016/j.aei.2013.02.001.

Falkowski, T. (2009). *Community analysis in dynamic social networks*, Dissertation, University Magdeburg.

Frantzi, K.T., Ananiadou, S. & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), 115-130.

Glänzel, W. & Czerwon, H.J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2), 195-221.

Guo, H., Weingart, S. & Börner, K. (2011). Mixed-indicators model for identifying emerging research areas. *Scientometrics*, 89(1), 421-435.

Haupt, R., Kloyer, M. & Lange, M. (2007). Patent indicators for the technology life cycle development. *Research Policy*, 36(3), 387-398.

Kajikawa, Y., Yoshikawa, J., Takeda, Y. & Matsushima, K. (2008). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change*, 75(6), 771-782.

Kandylas, V., Upham, S.P. & Ungar, L.H. (2010). Analyzing knowledge communities using foreground and background clusters. *ACM Transactions on Knowledge Discovery from Data*, 4(2), Article 7.

Kayal, A.A. (1996). *An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in the superconductor technology*. Dissertation, The George Washington University, Washington, DC.

Kayal, A.A. & Waters, R.C. (1999). An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in superconductor technology. *IEEE Transactions on Engineering Management*, 46(2), 127-131

Mitchell, M., Santorini, B. & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The penn treebank. *Computational linguistics*, 19(2), 313-330.

Moody, J., Farland, D. & Bender-deMoll, S. (2005). Dynamic network visualization. *American Journal of Sociology*, 110(4), 1206-1241.

Narin, F. & Hamilton, K. (1996). Bibliometric performance measures. *Scientometrics*, 36(3), 293-310.

Newman, M.E.J. (2004a). Analysis of weighted networks. *Physical Review E*, 70(5), 056131.

Newman, M.E.J. (2004b). Detecting community structure in networks. *The European Physical Journal B*, 38(2), 321-330.

Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science*, 45(1), 31-38.

Small, H.G. (2003). Paradigms, citations, and maps of science: A personal history. *Journal of the American Society for Information Science and Technology*, 54(5), 394-399.

Small, H.G. (2006). Tracking and predicting growth areas in science. Scientometrics, 63(3), 595-610.

Upham, S. & Small, H. (2010). Emerging research fronts in science and technology: Patterns of new knowledge development. *Scientometrics*, 83(1), 15-38.

wiseGEEK. What are some emerging technologies. Retrieved June 29 2012, from: http://www.wisegeek.com/what-are-some-emerging-technologies.htm.

Xie, Y. & Giles, D.E. (2011). A survival analysis of the approval of U.S. patent applications. *Applied Economics*, 43(11), 1375-1384.

Zinger, D.S. & Muljadi, E. (1997). Annualized wind energy improvement using variable speeds. *IEEE Transactions on Industry Applications*, 33(6), 1444-1447.

THE IMPACT OF CORE DOCUMENTS: A CITATION ANALYSIS OF THE 2003 SCIENCE CITATION INDEX CORE-DOCUMENT POPULATION

Bo Jarneving[1]

[1] *bo.jarneving@ub.gu.se*
University of Gothenburg, Gothenburg University Library, Renströmsgatan 4, 40530 Gothenburg (Sweden)

Abstract
In the 1960s Kessler introduced bibliographic coupling as a method for grouping research papers, facilitating scientific information provision. Later research has verified the applicability of this method in various information science contexts, such as information retrieval and science mapping. In this study the impact of so called 'core-documents', previously highlighted in the context of research front mapping, was elaborated applying state of the art impact indicators and varying citation windows. Due to limited resources, a random sample from the 2003 core-document population from the *Science Citation Index* was applied for statistical inference. Results were analyzed at the 95 % confidence level, applying confidence intervals for the arithmetic mean, proportions and the regression line. Findings indicated that core-documents were well cited above baselines and that a large share belonged to the top-cited papers of the world. Findings, not contradicting previous results, but providing with considerably more detail, lay ground for a more nuanced interpretation of core-documents' citation impact, where previous claims of key-positions in the science communication system were moderated. Findings also indicated that core-documents may have a rate of obsolescence notably deviating from the world average.

Introduction
Bibliographic coupling (BC) was introduced by Kessler through a number of reports and research articles in the 60s' (Kessler 1960; 1962; 1963a; 1963b; 1965). A bibliographic coupling unit was defined as: "[a] single item of reference shared by two documents…" (1962). BC was basically presented as a method for grouping technical and scientific documents which would facilitate scientific information retrieval. The original experiments performed by Kessler were based on small data sets from the journal *Physical Review*, why only limited conclusions of the method's applicability could be drawn. It took about two decades before a large scale experiment in a multidisciplinary environment took place (Vladutz and Cook, 1984). Findings showed that strong bibliographic coupling links generally implied strong subject relatedness. About the same time, Sen and Gan (1983) elaborating on the relation between subject relatedness and BC from a theoretical point of view, suggested a measure of coupling strength, the Coupling Angle (CA). With the point of departure in a hypothetical Boolean matrix where

elements indicated presence or absence of a relationship between citing documents (rows) and cited documents (columns), the CA was expressed as:

$$CA = \frac{(D_{oj} \bullet D_{ok})}{\sqrt{(D_{oj} \cdot D_{oj})(D_{ok} \cdot D_{ok})}} \ ,$$

where

D_{oj} and D_{ok} are the binary vectors of document *j* and *k*.

Specifically, the CA corresponds to the cosine of the angle for two vectors, *j* and *k*. The range is [0,1] where a cosine of 0 corresponds to an angle of 90° and a cosine of 1 to an angle of 0°. Using the CA as a measure of similarity between two documents, the minimum value (0) implies no common references whereas the maximum value (1) implies identical reference lists.

A more convenient way to express the same relation between document *j* and *k* is to calculate the ratio between the number of common references for *j* and *k* and the geometric mean of the number of references for *j* and *k*:

$$\frac{r_{jk}}{\left(n_j n_k\right)^{½}}$$

where

r_{jk} is the number of references common to both *j* and *k*

and

n is the number of references in document *j* or *k*.

Lacking empirical evidence of document-document similarity based on BC, Sen and Gan suggested a preliminary threshold of CA = 0.5 which corresponds to an angle $\theta = 60°$.

The relation between document-document similarity and BC was further elaborated by Peters, Braam and van Raan (1995) where the cognitive resemblance within groups of documents, bibliographically coupled by one and the same highly cited item, was explored using publications from the field of Chemical Engineering. Measuring word-profile similarities between the citing documents, it was found that word profile similarity within groups sharing a citation to a highly cited publication was significantly higher than between documents without such a relationship.

It may be concluded that empirical evidence of this method's ability to group similar papers, enough to warrant further investigations of plausible bibliometric areas of application, had been gathered at this point in time. In 1995 Glänzel and Czerwon presented a method applying BC for the identification of so called "hot research topics". Their method was based on the concept of "core-documents" which implied established thresholds for both the CA and the number of papers connected at the same set CA. Hence, a core-document would be defined as a paper connected with at least ten other papers with a minimum coupling strength of CA = 0.25. A limitation of document types was also done so that only articles, notes and reviews were included. In their empirical study, the whole annual accumulation of the 1992 volume of SCI was applied and about one percent of all publications of the preferred document types were identified as core-documents. In a sequel (1996) the same set of core-documents was analyzed with regard to the distribution of core-documents over journals, subfields and corporate addresses. A citation analysis was performed at the national level, applying a two-year citation window for all indicators.

Three main citation indicators were applied:

- *The relative citation rate* (RCR) which is the ratio of the mean observed citation rate (MOCR) to the mean expected citation rate (MECR). With regard to MECR, actual citations were substituted with journal impact factors.
- *Percentage of documents cited above average*. This indicator sums up the number of core-documents cited at least as many times as the corresponding journal impact factor and calculates the share.
- *Number of highly cited papers*. A core-document is considered "highly cited" if it has received at least 5 times as many citations as the corresponding journal impact factor.

Findings showed that core-documents, as defined, generally reflected "hot" research front topics, though the method seemed to have a bias towards the life sciences as most core documents were found in biomedical sub-fields. It was concluded that core-documents hold a key position in science communication on grounds of their high citation impact.

Research rationale

Later research on core-documents based on BC has involved cluster analytical approaches and network analysis (Jarneving 2007a, b; Glänzel and Thijs, 2011, 2012) as well as the combined application of textual information and citation data (Glänzel and Thijs, 2011, 2012). However, the citation impact of core-documents has not yet been exhaustively elaborated. Previous research on citations of core-

documents has been limited to the use of journal impact factors for the expected citation rates, with a focus on geographical distributions. Hence, there is a need of investigating the citation of core-documents using current citation based performance indicators. In addition, a wider citation window would complement previous findings where a two year window was applied. In particular, the claim that core-document may be considered keys for the identification of outstanding research performance (Czerwon and Glänzel, 1996) should be elaborated on.

Data and methods

From the SCI volume 2003 on CDROM, 619,570 records of the document type article were downloaded. A delimitation of document types to genuine research articles was made on grounds that this document type best mirrors empirical research. This population is referred to as the 2003 SCI core-document population, though ten percent of the core-documents had a publication year other than 2003. A total of 17,674,944 references were processed and 6,060 core-documents identified, which is approximately one percent of the total population of articles. Limited resources implied that citation indicators could not be generated for the total population of core-documents, why a random sample substituted the population of core-documents and estimates were applied. In order to be representative of the population, the sample was based on proportionate stratified sampling where strata were constructed on basis of major fields of science as defined in *Essential Science Indicators* (Thomson Reuters). The appropriate sample size was computed with a point of departure in the standard error of a proportion:

$$0.05 = 2 \cdot 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}$$

where

n = the sample size

p = the share of papers in the sample.

This means that a width of the confidence interval of 0.05 at the confidence level of 95 % was accepted. However, as we do not know the different shares, p must be guessed. Substituting p with 0.5 (which gives the largest value for n) gives the following equation after squaring and simplifying:

$$n = (2 \cdot 1.96)^2 \cdot \frac{0{,}5^2}{0{,}05^2}$$

Conclusively, a sample of 1,500 papers would probably work well. This means that approximately a quarter of all papers should be randomly drawn from the

population of 6,060 core documents. This rather large share of a *finite* population as well as the fact that the sampling was performed without replacement requires a correction factor (Isserlis, 1918) for both proportions and means when computing the standard error,

$$\sqrt{\frac{N-n}{N-1}}$$

where N = the population size,

and n = the sample size.

A total of three citation based indicators was decided on:

The average field normalized citation score ($\bar{C}_f$), where the expected number of citations (e) was computed as the average number citations to publications of the same type, with the same publication year and from the same field. It is defined as:

$$\frac{\sum_{i=1}^{n} \frac{c_i}{e_i}}{n}$$

where

c_i = number of citations to publication i
e_i = the expected number of citations to publication i
n = number of publications

This indicator was presented by Lundberg (2007) as the "Item oriented field normalized citation score average".

The average journal normalized citation score $\bar{C}_j$ is calculated analogously but the expected citation frequency is calculated as the average citation frequency of the corresponding journal, considering document type and publication year.

Top *n %* is the percentage core-documents that belong to the *n %* most cited papers in the world, where papers are matched with regard to publication year, field and document type. In this study *n* assumes the values 5, 10 and 20.

The expected citation frequencies as well as the top *n %* indicator values were matched with each individual publication of the random sample by CWTS,

Leiden University, using data from Thomson Scientific/ISI. All indicator values were computed with self citations excluded.

Findings

Before presenting the results from the citation analysis, some descriptive statistics should be commented on. Considering the distribution of core-documents over journals, 995 distinct journal titles out of 3,567 contained at least one core document, which means that 72 % of all journals in the 2003 SCI volume did not contain any core documents. The corresponding figure in Glänzel & Czerwon (1996) was 75 %. Another distribution of interest concerns co-authorships. The mean number of authors of a core-document was 6.0 and the maximum number of authors 255. The corresponding figures for the 1992 volume were 4.5 and 104 (Glänzel & Czerwon, 1996). For the whole 2003 volume the mean number of authors was 4.4.

Impact

For each core-document in the sample, citation data for 7 years was assembled. Counting whole publication years, the maximal error with regard to the publication date was thus less than but approximately one year. The first whole year after the publication year was considered to correspond to a (minimum) citation window of one year. In this way, three citation windows were applied:

- 2 years: three years after the publication year
- 4 years: five years after the publication year
- 6 years: seven years after the publication year

In Table 1, the arithmetic mean for $\bar{C}_f$ and $\bar{C}_j$ are displayed with confidence intervals at the 95 % confidence level. As can bee seen, both $\bar{C}_f$ and $\bar{C}_j$ decrease over time and $\bar{C}_f$ is notably higher.

Table 1. The $\bar{C}_f$ and $\bar{C}_j$ for three citation windows with confidence intervals at the 95 % confidence level.

Citation window	$\bar{C}_f$	CI	$\bar{C}_j$	CI
2 years	2.90	± 0.19	2.23	± 0.14
4 years	2.67	± 0.18	2.13	± 0.14
6 years	2.52	± 0.18	2.03	± 0.13

Considering the impact of core-documents on fields, the top *n* % indicators show the share of core-documents that belong to the world's top *n* %. Here, *n* assumes values of 5, 10 and 20, which are displayed over three citation windows along with confidence intervals at the 95 % confidence level (Table 2).

Table 2. The share of core-documents belonging to *n %* top-cited papers: 5, 10 and 20 percent levels are displayed for three citation windows with confidence intervals at the 95 % confidence level.

Citation window	top 5 %	CI	top 10 %	CI	top 20 %	CI
2 years	0.25	± 0.02	0.36	± 0.02	0.52	± 0.02
4 years	0.24	± 0.02	0.35	± 0.02	0.51	± 0.02
6 years	0.22	± 0.02	0.32	± 0.02	0.48	± 0.02

The impact profile for the sampled core-documents is displayed in Figure 1, providing with comprehensible class intervals (cf. Adams, Gurney & Marshall, 2007) for $\bar{C}_f$ over a 6-year citation window. Note that the upper bounds increase by a factor of two for each new class interval. The 6-year citation window was chosen in order to exhaustively assess the influence of the category "uncited".

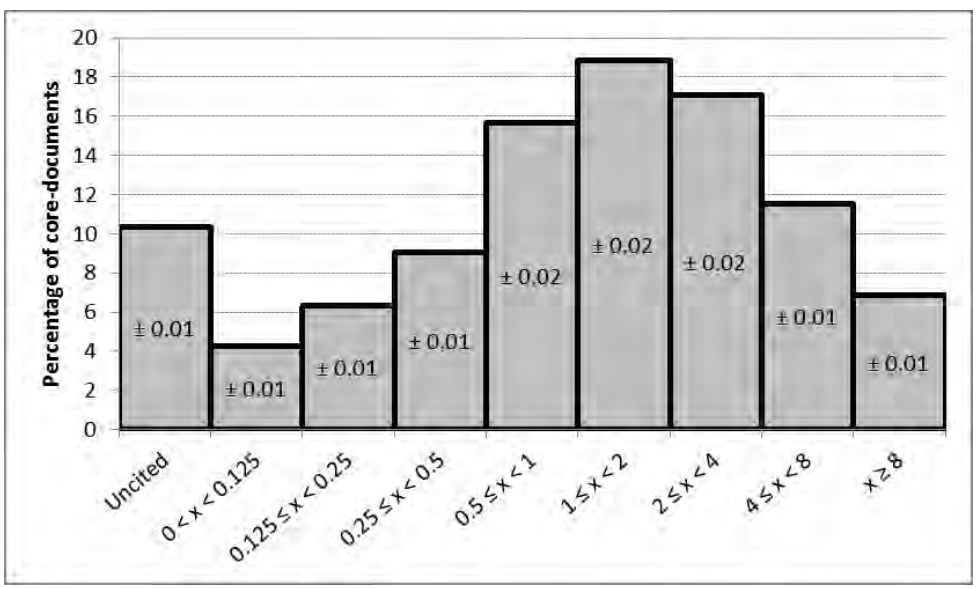


Figure 1. The 6-year impact-profile for 1500 core-documents: class intervals where $\bar{C}_f = x$ marked on the x-axis. Confidence intervals at the 95 % confidence level displayed within bars.

Growth of citations

Focusing on the relation between the length of the citation window and the number of observed citations, a regression analysis was performed. The graph in Figure 2 illustrates a near perfect linear relationship with confidence intervals at the 95 % confidence level for the number of observed citations. Given this growth model, the set of sampled core-documents receives an annual contribution of 8,618 citations, while the lower bound was 8,192 citations and the upper 9,045. The ratio of the annual number of expected citations to the lower bound of the

observed citations was 1 to 2.5, reflecting a much faster accumulation of citations to the sampled core documents (Figure 2).

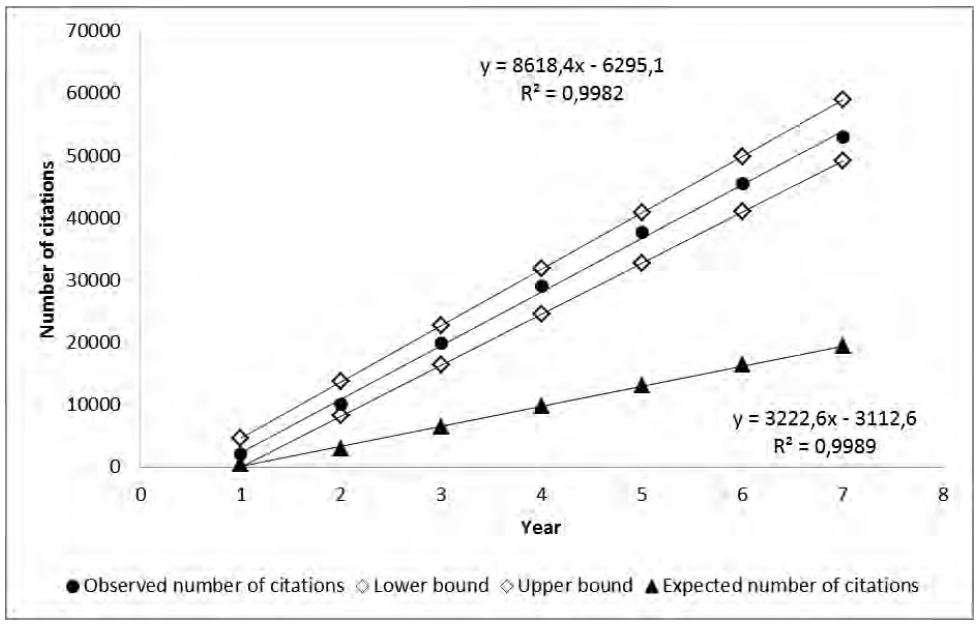


Figure 2. Number of accumulated citations for the sampled core-documents during a six-year citation window and number of accumulated expected citations. Lower and upper bounds at the 95 % confidence level are displayed for the observed citations.

However, for a 6-year citation window, the expected percentage growth was 3849 % and for the point estimate 2436 %, CI [2043 %, 2491 %].

Discussion

Results convincingly showed that the average field normalized citation score ($\bar{C}_f$) for core documents was well above the world standard (the expected). With regard to the two-year citation window, $\bar{C}_f$ was almost three times the expected according to the point estimate. Considering the *lower* bounds of the confidence intervals, over all three citation windows with 95 % confidence, the corresponding population parameter was within the interval 2.34 – 2.71. The corresponding interval for journal normalized citation counts ($\bar{C}_j$), was 1.89 – 2.09. These figures indicate a substantial difference between the expected and the observed. The different results arrived at when applying field normalization respectively journal normalization should reflect that core-documents are often published in high impact journals.

Mapping the impact of core-documents in terms of their percentage distribution over top *n* % categories is complementary to elaborations on averages. Given the

narrowest citation window, a quarter of all sampled core-documents belonged to the top 5 % most cited publications with a margin of error of ± 2 %. With approximately the same margin of error, corresponding figures for top 10 % and top 20 % was 36 % and 52 % respectively. These findings are actually not in line with the claim that core-documents belong to the set of high impact papers of specialties (Glänzel & Czerwon, 1996), at least not in a general sense.
A more elaborated depiction of core-documents' citedness was provided by the histogram in Figure 1. We can appreciate that a majority of the sampled core-documents are cited above the world average (1.0) and a little less than half below. About ten percent is never to be cited during a six-year citation window, while approximately 18 percent have a citation frequency at least four times the world average. The modal group of sampled core-documents are cited above the world average but within the limit of a factor of two. The definition of core-documents as such suggests a close relationship with the research front, that is, with the portion of current papers within a field that is tied to a relatively small and select group of earlier papers by citation (cf. Price, 1965). However, this would not per see imply a high citation rate as other markers of high quality such as originality and immediacy play important roles. In fact, results arrived at here indicate that for every core-document cited more than twice the expected, we would find a core-document cited below the world average. Conclusively, core document attributes are not in themselves sufficient markers of "outstanding research performance" (cf. Glänzel and Czerwon, 1996). However, it would be complementary to explore to what extent high impact papers possess core document attributes.

Notably, the 1992 core-document population showed up with a considerably larger figure for the share of core-documents cited above average. In Glänzel and Czerwon (1996), 62.4 % were cited above average while the corresponding figure in this study was 54 %, ± 2.2 %. One may assume that there is a trend of an increasing number of core-documents of lower quality. Another, assumption is that the difference between the two populations is due to the fact that review papers generally have a higher citation impact than research articles (Glänzel and Moed, 2002).

Considering the accumulation of citations to core-documents, a much faster than expected growth during the 6 year citation-window was observed. This is in line with expectations and other findings. However, it was also observed that indicator values decline notably over time (cf. Table 1 and Table 2). This indicates that core-documents have a higher obsolescence than expected. Consequently, the percentage growth for core-documents was substantially lower than expected, also when considering the upper bound of the confidence interval.

Conclusions

In spite of the obvious limitations of basing inferences on a random sample, it has been feasible to map citation impact of core-documents at the 95 % confidence level. Findings indicate that core-documents are well cited above baselines and that a large share of the 2003 *Science Citation Index* core-document population belongs to the top-cited papers of the world. This is basically in line with previous findings, though considerably more detailed information with regard to relevant impact indicators lay ground for a more nuanced interpretation of core-documents' role in the scientific communication system. Hence, previous claims of core-documents key-position and impact should be moderated on grounds that the citation impact of core-documents is unevenly distributed.

References

Adams, J., Gurney, K. & Marshall, S: (2007). Profiling citation impact: A new methodology. *Scientometrics*, 72(2):325-344.

Glänzel, W. & Czerwon, H. J. (1995). A new methodological approach to bibliographic coupling and its application to research-front and other core documents, *Proceedings of 5th International Conference on scientometrics and Informetrics*, held in River Forest, Illinois, June 7-10: 167-176.

Glänzel, W. & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2):195-221.

Glänzel, W. & Moed, H. (2002). Journal impact measures in bibliometric research. *Scientometrics*, 53(2):171-193.

Glänzel W, Thijs B. (2011). Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1):297-309.

Glänzel W, Thijs B. (2012). Using 'core documents' for detecting and labeling new emerging topics. *Scientometrics*, 91(2):399-416.

Isserlis, L. (1918). On the value of a mean as calculated from a sample. *Journal of the Royal Statistical Society*: 75–81.

Jarneving B. (2007a). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, 1(4):287-307.

Jarneving B. (2007b). Complete graphs and bibliographic coupling: A test of the applicability of bibliographic coupling for the identification of cognitive cores on the field level. *Journal of Informetrics*, 1(4):338-56.

Kessler, M. M. (1960). An experimental communication center for scientific and technical information. Massachusetts Institute for Technology, Lincoln Laboratory.

Kessler, M. M. (1962). An experimental study of bibliographic coupling between technical papers. Massachusetts Institute for Technology, Lincoln Laboratory.

Kessler, M.M. (1963a). Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10-25.

Kessler, M.M. (1963b). Bibliographic coupling extended in time: Ten case histories. *Information Storage and Retrieval*, 1:169-187.

Kessler, M.M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation*, 16(3):223-233.
Lundberg, J. (2007). Lifting the crown—citation z-score. *Journal of Informetrics*, 1(2):145–154.
Peters, H. P. F., Braam, R. R. and van Raan, A. F. J. (1995). Cognitive resemblance and citation relations in chemical engineering publications. *Journal of the American Society for Information Science*. 46(1):9-21.
Price, D. J. de Solla (1965), Networks of scientific papers. *Science*, 149(3683):510-515.
Sen, S. K. & Gan. S. K. (1983). A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library Science and Documentation*, 30(2):78-82.
Vladutz, G. & Cook, J. (1984). Bibliographic coupling and subject relatedness. *Proceedings of the ASIS Annual Meeting*, 47: 204-207.

IMPACT OF META-ANALYTICAL STUDIES, STANDARD ARTICLES AND REVIEWS: SIMILARITIES AND DIFFERENCES

Maite Barrios[1], Georgina Guilera[1,2], and Juana Gomez-Benito[1,2]

mbarrios@ub.edu, gguilera@ub.edu, juanagomez@ub.edu
[1]Department of Behavioral Sciences Methods, University of Barcelona, Spain
[2]Institute for Brain, Cognition and Behavior (IR3C), University of Barcelona, Barcelona, (Spain)

Abstract
Meta-analysis refers to the statistical methods used in research synthesis for combining and integrating results from individual studies. In this regard meta-analytical studies share with narrative reviews the goal of synthesizing the scientific literature on a particular topic, while as in the case of standard articles they present new results. This study aims to identify the potential similarities and differences between meta-analytical studies, reviews and standard articles as regards their impact in the field of psychology. To this end a random sample of 335 examples of each type of document were selected from the Thomson Reuters Web of Science database. The results showed that meta-analytical studies receive more citations than do both reviews and standard articles. The implications of these results for the scientific community are discussed.

Conference Topic
Scientometrics Indicators: Relevance to Science and Technology, Social Sciences and Humanities (Topic 1)

Introduction
For many decades narrative reviews were the preferred way for researchers to combine the results of different articles about a specific topic. The aim of such reviews was to gather together a set of studies on a given subject, summarizing their results and drawing conclusions regarding the question of interest. This approach had a number of limitations, notably the lack of transparency or subjective nature of many of the decisions made when preparing the review (Cooper & Hedges, 1994). For example, the criteria for including studies or the level of confidence assigned to each one of them might vary from one set of reviewers to another, and in some cases this could mean that two reviews reached substantially different conclusions (Borenstein, Hedges, Higgins, & Rothstein, 2009). Furthermore, the number of scientific publications now being produced is so great that any attempt to synthesize research by means of narrative reviews is likely to prove ineffective due to the unmanageable amount of information, unless, that is, the process can be made more systematic. It is in this context that systematic reviews and meta-analyses have emerged as a way of making more

rigorous the process of document localization and the definition of inclusion/exclusion criteria, among other aspects of the review procedure. While these approaches do not completely eliminate the subjectivity that is characteristic of narrative reviews, they at least ensure a more transparent synthesis, since they make explicit all the decisions made during the process. In a systematic review the statistical synthesis of data is based on what is known as meta-analysis, an approach that includes a range of statistical methods and formulas designed to synthesize and compare the results of a set of studies (Littell, Corcoran & Pillai, 2008). Meta-analyses and systematic reviews, however, are not free from criticism (Bailar, 1997); yet, the problems detected in studies of this kind are the same as those that narrative reviews have to face (Borenstein et al., 2009). Meta-analysis has become a highly popular way of synthesizing research literature and it is now widely accepted within the scientific community (Cooper, 2010), to the extent that when a team of scientists plans a new study it is highly likely that they will first seek to locate a meta-analysis in order to design their own investigation. In this regard the field of psychology is no exception, not least because the first study to be regarded as a meta-analysis examined the effectiveness of psychotherapy (Smith & Glass, 1977). The considerable influence of the meta-analytic procedures that were being used in psychology and education led to them being transferred to many other areas of knowledge.

In science a common way to measure the relevance of a study is to count the number of citations it has received (Moed, 2005). When a scientific paper is published in a journal other scientists can use its findings to elaborate, corroborate or contrast their own research. They then indicate the use of that paper by means of a formal citation in their own research. The number of citations that a study receives has therefore been used as an objective quantitative indicator of its usefulness, importance and the interest it arouses in the scientific community. However, as Glänzel and Moed (2002) point out, the citations that a paper receives are themselves influenced by at least five factors: (i) the type of document (e.g. articles, reviews, notes or proceedings papers, among others); (ii) the discipline, since not all scientific fields have the same citation habits; (iii) the paper's age, since older papers have a greater chance of being cited; (iv) the paper's 'social status', for example, the impact factor of the journal in which it was published or the standing of its author(s); and (v) the observation period, due to the influence of aspects such as obsolescence or the citation curve of the literature. Moreover, other authors have shown that a high number of citations are associated with a higher number of co-authors (Bearer, 2004; Glänzel, Rinia & Brocken, 1995; Lawani, 1986; Vieira & Gomes, 2010), a greater number of both pages (Bornmann & Daniel, 2007) and references (Haslam et al., 2008; Bornmann, Mutz, Neuhaus & Daniel, 2008; Peters and van Raan, 1994; Vieira & Gomes, 2010), English language publication (van Raan, 2005) and a greater international collaboration (Askes, 2003; Glänzel et al., 1995).

The fact that reviews receive more citations than do standard articles is widely known (Amin & Mabe, 2000; Braun, Glänzel & Moed, 2002; Dong et al., 2005;

Glänzel & Schubert, 1989; Seglen, 1997; Sigogneau, 2000; Vieira & Gomes, 2011). Although the Thomson Reuters Web of Science does not provide a clear description of how papers are classified into the different document types (e.g. articles, reviews or proceedings papers) it is accepted that in social sciences, and in psychology in particular, that review articles do not normally contain original data but simply collect, review and synthesize earlier research, without including substantial theoretical or conceptual development (Harzing, 2013). In this regard, meta-analytical studies fall halfway between the original articles and reviews. They share with narrative reviews the goal of synthesizing the scientific literature on a particular topic, while as in the case of original articles they present new results, which in the case of meta-analyses is done by combining the results of the set of articles they consider. Thus, meta-analytical studies would be expected to arouse considerable interest in the scientific community, and consequently they receive as many citations as do review articles.

It should be noted that Thomson Reuters Web of Knowledge classifies each document into a particular document category. As regards the 'review' category Thomson Reuters Web of Knowledge uses a wide criterion and a paper may be classified as a review either when it is published in the 'review' section of a journal or when the words 'review' or 'overview' appear in the title of the document (Thomson Reuters, 1994). When it comes to meta-analytical studies, Thomson Reuters Web of Knowledge does not have a consistent way of classifying them. Although most meta-analytical studies are classified as standard articles, some are classified as reviews. In a previous study (Guilera, Barrios & Gómez-Benito, 2012), in which we examined a whole set of meta-analytical studies in the field of psychology, we found that 68.0% were classified as standard articles and just 32.0% as reviews (unpublished data). One of the reasons for this ambiguous classification is likely to be that Thomson Reuters proposed that any article containing more than 100 references should also be coded as a review (Thomson Reuters, 1994). However, as some authors point out (Seglen, 1997; Sigogneau, 2000) this criterion is open to criticism because the number of references in a paper is discipline-dependent, which means that one should be wary of using it as an indicator of the level of originality of a study (Harzing, 2013). Nonetheless, since the number of citations which a paper can receive in a specific research field is directly proportional to the mean number of references per article (Seglen, 1997), and given that some authors (Bornmann et al., 2008; Haslam et al., 2008; Peters & van Raan, 1994; Vieira & Gomes 2010) have found that citation counts are associated with a higher number of references, then meta-analytical studies classified as reviews would be expected to be cited more often than would those classified as standard articles.

In light of the above the aim of the current paper is to conduct a comparative analysis of meta-analytical studies, reviews[82] and standard articles[83] in order to

[82] Throughout the article, the term 'review' is used to refer to documents classified as 'Review' in the Thomson Reuters Web of Knowledge, excluding meta-analyses that have been classified as such in this study.

explore potential differences and similarities as regards their impact. The specific focus is on the field of psychology, where we compare these three types of documents while controlling for the paper's age and journal. We hypothesized that (i) meta-analytical studies would be cited as often as reviews; and (ii) those meta-analytical studies classified as reviews would receive more citations than would those classified as standard articles.

Method

Data collection and sample

The meta-analytical studies included in the present analysis corresponded to a subsample of the articles which Guilera et al. (2012) identified as being empirical meta-analytical studies in the field of psychology (n = 2,605). Three hundred and thirty-five papers were selected from that whole sample so as to work with an accuracy of 5% and a confidence level of 95%. A stratified sampling approach was used to ensure the new sample was proportionately representative of the general data set. Year of publication and Bradford zone were used as stratification variables. The sample was proportionally and randomly selected from among the journals classified in the different Bradford zones because in the general sample (Guilera et al., 2012) the results showed a relationship between Bradford zone and the number of citations per article, such that those articles classified in the core and first zones presented a higher number of citations.

In accordance with the document type classification used by Thomson Reuters Web of Knowledge, 335 standard articles and 335 reviews were selected using the Thomson Reuters Web of Science database. In order to select this set of standard articles and reviews, methodological and empirical meta-analytic studies were excluded. The studies included were randomly selected from among those published in the same journal and year as the meta-analytical studies under study. In the event that no standard article or review was published in a specific journal in the same year, previous years were checked in succession in order to find a matched standard article and/or review. If this procedure failed to identify a standard article or review that had been published relatively close to the date of publication of the meta-analytical paper we then examined, with the same purpose, the years subsequent to the year of publication of the meta-analytic paper.

Thus, the three types of documents (meta-analytical studies, standard articles and reviews) were matched for the following variables: year of publication and journal. The sample selection was conducted between 26 April and 31 May 2012.

[83] Throughout the article, the term 'original article' is used to refer to documents classified as 'Article' in the Thomson Reuters Web of Knowledge, excluding meta-analyses that have been classified as such in this study.

Variables and data analysis

The number of citations for each article, from its year of publication until the date of its downloading, was obtained from the Web of Science database in order to study the impact of the research. As expected, citations were highly positively skewed. Given that many statistical procedures assume that the variables are normally distributed, we applied log transformation to the data in order to improve the normality of this variable. The non-parametric Kolmogorov-Smirnov goodness-of-fit test was used to assess the normality of the data after log transformation.

Analysis of variance (ANOVA) was applied in order to determine whether there were any differences between the meta-analytical studies, reviews and standard articles in terms of the number of citations. In addition, analysis of covariance (ANCOVA) was used to study any differences in the number of citations corresponding to the three types of documents while controlling for the effects of extraneous variables. Thus, the number of authors, pages per document and references were analysed as covariates, as suggested by Bornmann, Mutz, Neuhaus and Daniel (2008). In order to study differences between meta-analytical review studies and meta-analytical standard articles, impact factor and years since publication were also added as covariates.

Results

Of the 335 meta-analytical studies selected the majority were classified as standard articles by the Thomson Reuters database (n = 226, 67.5%), with only 32.5% (n = 109) being classified as reviews. The main characteristics of this sample are shown in Table 1 (i.e. number of journals, number of articles, mean years since publication, and number of citations received by the articles classified in each Bradford zone). Note that the mean number of citations is higher in the areas closer to the core.

After logarithmic transformation the citation data followed a normal distribution (Kolmogorov-Smirnov $Z = 0.655$, $p = .784$). The ANOVA revealed statistically significant differences in the number of citations received depending on the type of document (Table 2). Specifically, meta-analytical studies received a significantly higher number of citations compared to both review and standard articles. As expected, reviews were cited more often than were standard articles. Covariance analysis showed that after controlling for possible extraneous variables (number of co-authors, references and pages) the statistically significant differences between the three document types were maintained ($F(2, 992) = 28.190$, $p < .001$). Table 3 shows the descriptive statistics for the extraneous variables and figure 1 illustrates the mean number of citations and 95% confidence intervals corresponding to the different types of document.

Table 1. Characteristics of the sample of meta-analytical studies according to Bradford zones.

Bradford's area (number of journals)	*Articles* n (%)	*Citations* Mean (SD) CI	*Years* Mean (SD)
Core (n = 1)	26 (7.76)	176.62 (228.09) 84.49 ÷ 268.74	9.23 (7.62)
Zone 1 (n = 2)	29 (8.66)	118.21 (121.05) 72.16 ÷ 164.25	8.86 (7.20)
Zone 2 (n = 4)	29 (8.66)	80.79 (86.84) 47.76 ÷ 113.82	8.41 (6.28)
Zone 3 (n = 6)	31 (9.25)	72.32 (91.38) 38.81 ÷ 105.84	8.42 (7.13)
Zone 4 (n = 12)	36 (10.75)	62.86 (72.51) 38.33 ÷ 87.39	8.86 (7.70)
Zone 5 (n = 25)	46 (13.73)	42.37 (43.45) 29.47 ÷ 55.27	9.13 (6.96)
Zone 6 (n = 36)	52 (15.52)	42.42 (56.40) 26.72 ÷ 58.12	8.73 (7.11)
Zone 7 (n = 38)	48 (14.33)	49.15 (91.46) 22.59 ÷ 75.70	9.33 (7.11)
Zone 8 (n = 37)	38 (11.34)	22.89 (25.10) 14.65 ÷ 31.14	8.55 (6.99)

SD: standard deviation, CI: confidence interval at 95%, Years: years since publication

Table 2. Citation differences between meta-analytical studies, reviews and standard articles.

Document types	*Mean (SD)*	*Median (IQR)*	*F(d.f.)*	*p-value*	*Groups*[a]
Meta-analysis	66.42 (203.30)	29.0 (68)	35.951 (2, 1002)	< .001	MA vs R** MA vs SA** R vs SA **
Reviews	44.77 (74.55)	18.0 (41)			
Standard articles	24.32 (45.42)	11.0 (21)			
Meta-analysis-Review	84.86 (135.186)	39.0 (70)	1.077 (1, 328)	.300	
Meta-analysis-Article	57.53 (82.61)	25.5 (68)			

[a]Only significant group differences are shown. SD: standard deviation, IQR: interquartile range, F: Snedecor's F test, d.f.: degrees of freedom, MA: Meta-analytical studies, SA: Standard Articles, R: Reviews. Meta-analysis-Review: Meta-analytical studies classified as reviews, Meta-analysis-Article: Meta-analytical studies classified as standard articles.
** $p < .001$

The data show that the number of references was not the only criterion used by Thomson Reuters to classify an article as a review, since 32.1% (n = 35) of the meta-analytical studies classified as reviews contained fewer than 100 references, while conversely, 5.8% (n = 13) of the meta-analytical studies classified as standard articles included more than 100 references. Table 4 shows for each type of document the percentage of documents with 100 references or fewer and the percentage with more than 100 references.

Table 3. Descriptive statistics of extraneous variables.

	Mean (SD)	*Median (IQR)*
Number of authors		
Meta-analysis	2.80 (1.56)	3.0 (1)
Reviews	2.81 (2.12)	2.0 (2)
Standard articles	3.19 (2.15)	3.0 (2)
Number of references		
Meta-analysis	80.64 (52.88)	69.0 (57)
Reviews	116.61 (67.76)	109.0 (45)
Standard articles	51.09 (36.48)	44.0 (36)
Number of pages		
Meta-analysis	16.59 (8.82)	15.0 (11)
Reviews	18.67 (9.44)	17.0 (12)
Standard articles	12.77 (7.17)	11.0 (8)

SD: standard deviation, IQR: interquartile range.

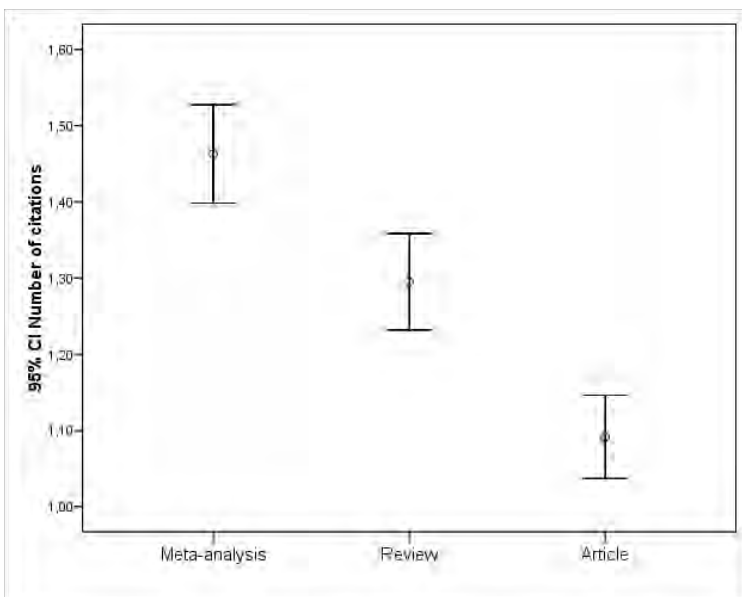


Figure 1. Mean number of citations and 95% confidence intervals for meta-analytical studies, reviews and standard articles.

Covariance analysis was then performed to determine any differences between meta-analytical studies, according to the classification of Thomson Reuters (i. e., standard articles and reviews) and taking as covariates the number of years since publication, the journal impact factor and the number of pages, references and co-authors. The analysis revealed no statistically significant differences between meta-analytical studies classified as reviews and those classified as standard articles (Table 2).

Table 4. Percentage of documents with 100 references or fewer and the percentage with more than 100 references for each type of document.

	Equal or less than 100 references	*More than 100 references*
Document types	Percentage (n)	Percentage (n)
Meta-analysis	74.0 (248)	26.0 (87)
Meta-analysis-Review	32.1 (35)	67.9 (74)
Meta-analysis-Article	94.2 (213)	5.8 (13)
Reviews	29.6 (99)	70.4 (236)
Standard articles	95.2 (319)	4.8 (16)

Meta-analysis-Review: Meta-analytical studies classified as reviews,
Meta-analysis-Article: Meta-analytical studies classified as standard articles.

Discussion

This paper compares the impact and structural features of a randomly selected sample of meta-analytical studies, reviews and standard articles in the field of psychology. To our knowledge this is the first study to compare the impact of these three types of documents. In terms of impact, reviews have been conclusively identified as the type of document which receives more citations in comparison with standard articles, notes, proceedings, etc. (Amin & Mabe, 2000; Braun et al., 1989; Dong et al., 2005; Glänzel & Moed, 2002; Seglen, 1997; Sigogneau, 2000; Vieira & Gomes, 2011). Our first hypothesis, based on the fact that the purpose of meta-analytical studies is to synthesize results from empirical literature, was that they would be cited as often as reviews. However, a notable finding of the present study is that the citation rate for meta-analytical studies was, on average, higher than that of both standard articles and reviews. This result was independent of the number of authors and the number of references and pages in the document, and neither did it depend on whether the meta-analytic study was classified by Thomson Reuters as a review or a standard article. One explanation for the high citation rate of meta-analytical studies is the considerable importance ascribed to them by the scientific community, such that meta-analytical studies may be used both to remain up to date on a particular topic and to guide the design of new studies based on meta-analytical results.

We also hypothesized that meta-analytical studies classified as reviews by Thomson Reuters would receive more citations than those classified as standard articles. However, after controlling for the number of authors, references and pages, as well as the years since publication and the journal impact factor, the data revealed no significant differences between these two types of documents. This means that after controlling for extraneous variables the interest shown by the scientific community in meta-analytical studies is similar, regardless of how Thomson Reuters classifies the type of document. A likely explanation for this result is that meta-analytical studies usually incorporate the term meta-analysis in their title (Guilera et al., 2012) and also as a keyword. Thus, when researchers are

looking for a meta-analytical study they probably use 'meta-analysis' as a search term rather than filtering by type of document.

Another notable result of the present study concerns the criterion used by Thomson Reuters to classify review and standard articles. Although the vast majority of meta-analytical studies classified as reviews contain more than 100 references, whereas those classified as articles have fewer than 100 references, our data show that this cut-off was not always applied, thereby suggesting that Thomson Reuters must apply other parameters when making this classification. In fact, in their discussion of journal impact factor, Thomson Reuters (Thomson Reuters, 1994) state that articles in 'Review' sections of research or clinical journals are also coded as reviews, along with articles whose titles contain the word review or overview.

These results have a number of implications that should be of interest to the scientific community, not just scientists themselves but also research evaluators, journal editors and bibliometricians.

Firstly, meta-analytical studies receive a high number of citations, more than in the case of reviews and standard articles. This finding has important implications for scientists, who aim to select the most relevant articles to read and to complement their research, and who are also aware that scholarly publishing is central to academic success. Article selection, on the one hand, may depend on the article's impact and this might mean that meta-analytical studies are perceived as being more relevant pieces of research, while scientists fail to select other types of documents with less probability of being cited. The importance, on the other hand, of the quantity and impact of a scientist's publications in determining future performance evaluations, funding decisions, promotion and salaries (Borrego, Barrios, Villarroya & Ollé, 2010) means that the possibility of publishing a study with a high probability of receiving a high number of citations may be perceived by scientists as an opportunity of boosting their chances of obtaining funding, promotion or a tenure position.

Secondly, although there is a need for future studies to investigate the citation patterns of meta-analytical studies, it is likely that those journals which are able to accumulate a high number of meta-analytical studies will be able to increase their impact factor. This is supported by a recent study (Guilera et al., 2012) in which we found that the citation of meta-analytic papers makes a strong contribution to a journal's impact factor. Consequently, journal editors may be especially interested in publishing meta-analytical studies, since they know that a paper of this kind is likely to receive a high number of citations, even higher than for other reviews, thereby increasing the impact factor of the journal. However, this can lead to journals having a highly skewed distribution of citation rates for its articles, and therefore, as Seglen (1997) advised, it is important to avoid judging a paper by its wrapping rather than by its contents. As others authors have advised (Bloch & Walter, 2001, Kurmis, 2003, Pendlebury, (2009), research evaluators and scientists in general should avoid taking the journal impact factor as a measure of the quality of a piece of research, that is, using it, for instance, to assess a

candidate's suitability for promotion or to choose the journal to which an article will be submitted or from which a paper will be selected to read.

Thirdly, the ambiguous criterion applied by Thomson Reuters to classify meta-analytical studies into reviews or standard articles can lead to misunderstanding among the research community. For instance, the claim that reviews are, on average, more likely to be cited than are standard articles does not always hold true due to the mix of meta-analytical studies. This finding is also of interest for research evaluators, who may assess the papers of scientists differently according to document type (Gonzalez-Albo & Bordons, 2011). Therefore, if a document is classified as a review it may be interpreted as a piece of research of minimal originality and which does not include significant conceptual development. This result should also be taken into account by bibliometricians who, when selecting the most valuable studies on the basis of their impact or when analysing citation rates according to the type of document, may unwittingly introduce a source of bias. In this regard, a limitation of the present study that results from this criterion is that some of the reviews which were randomly selected for the sample were, in fact, standard articles with more than 100 references.

Finally, it should be noted that this study offers a broad overview of the behaviour of meta-analytical studies in the field of psychology, and therefore the results cannot be generalized to other disciplines. Further analyses focusing on other scientific fields are now needed to confirm the higher impact of meta-analytical studies compared with reviews and standard articles.

Acknowledgments

This study was supported by grants 2009SGR00822 from the 'Departament d'Universitats, Recerca i Societat de la Informació' of the Generalitat de Catalunya and PSI2009-07280 from the 'Ministerio de Ciencia e Innovación' of Spain.

References

Amin, M., & Mabe, M. (2000). Impact factors: Use and abuse. Perspectives in Publishing, 1, 1–6.

Aksnes, D. W. (2003). Characteristics of highly cited papers. Research Evaluation, 12(3), 159–170.

Bailar, J.C. (1997). The promise and problems of meta-analysis. New Engl J Med, 337(8), 559–561.

Beaver, D. B. (2004). Does collaborative research have greater epistemic authority? Scientometrics, 60(3), 399–408.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. United Kingdom: John Wiley and Sons, Ltd.

Bornmann, L., & Daniel, H. D. (2007). Multiple publications on a single research study: Does it pay? The influence of number of research articles on total citation counts in biomedicine. Journal of the American Society for Information Science and Technology, 58(8), 1100–1107.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.D. (2008). Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. Ethics in Science and Environmental Politics, 8, 93-102.
Borrego, A., Barrios, M., Villarroya A., & Ollé, C. (2010). Scientific output and impact of postdoctoral scientists: a gender perspective. Scientometrics, 83(1), 93–101.
Bloch, S. Walter, G. (2001). The Impact Factor: time for change. Australian & New Zealand Journal of Psychiatry, 35(5), 563-568.
Braun, T., Glänzel, W., & Schubert, A. (1989). Some data on the distribution of journal publication types in the Science Citation Index Database. Scientometrics, 15(5-6), 325–330.
Cooper, H. (2010). Research synthesis and meta-analysis: A step by step approach (4th edition). Thousands Oaks, CA: Sage Publications, Inc.
Cooper, H., & Hedges, L. V. (Eds.) (1994). The Handbook of Research Synthesis. New York: Russell Sage Foundation.
Dong, P., Loh, M., & Mondry, A. (2005). The ‘‘impact factor’’ revisited. Biomedical Digital Libraries, 2, 7.
Glänzel, W.; & Moed, H. F. (2002). Journal impact measures in bibliometric research. Scientometrics, 53(2), 171–193.
Glänzel, W., Rinia, E. J. & Brocken, M. G. M. (1995). A bibliometric study of highly cited European physics papers in the 80s. Research Evaluation, 5(2), 113–122.
Gonzalez-Albo, B., & Bordons, M. (2011). Articles vs. proceedings papers: Do they differ in research relevance and impact? A case study in the Library and Information Science field. Journal of Informetrics, 5, 368-381.
Guilera, G., Barrios, M., & Gómez-Benito, J. (2013). Meta-analysis in psychology: A bibliometric study. Scientometrics. Advance online publication. doi: 10.1007/s11192-012-0761-2.
Harzing, A. W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics. 14*(1), 23-34.
Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology. Scientometrics, 76(1), 169–185.
Kurmis, A. P. (2003). Current concepts review - Understanding the limitations of the journal impact factor. Journal of Bone & Joint Surgery-American Volume, 85A(12), 2449-2454.
Lawani, S. M. (1986). Some bibliometric correlates of quality in scientific research, Scientometrics, 9(1–2), 13–25.
Littell, J. H., Corcoran, J., & Pillai, V. (2008). Systematic reviews and meta-analysis. Oxford, UK: Oxford University Press.
McVeigh, M. E., & Mann, S. J. (2009). The journal impact factor denominator. Defining citable (counted) items. JAMA, 302(10), 1107–1109.

Moed, H. F. (2005). Citation analysis in research evaluation. Dordrecht, The Netherlands: Springer.
Pendlebury, D. A. (2009). The use and misuse of journal metrics and other citation indicators. Archivum Immunologiae et Therapiae Experimentalis, 57(1), 1-11.
Peters, H. P. F., & van Raan, A. F. J. (1994). On determinants of citations scores—a case-study in chemical engineering. Journal of the American Society for Information Science, 45(1), 39–49.
Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. British Medical Journal, 314, 498–502.
Sigogneau, A. (2000). An analysis of document types published in journals related to physics: Proceeding papers recorded in the Science Citation Index database. Scientometrics, 47(3), 589-604.
Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.
Thomson Reuters (1994). The Thomson Reuters Impact Factor. Retrieved August, 2012, from http://thomsonreuters.com/products_services/science/free/essays/impact_factor/.
van Raan, A. F. J. (2005) For your citations only? Hot topics in bibliometric analysis. Measurement, 3(1), 50–62.
Vieira, E. S., & Gomes, J. A. N. F. (2010). Citations to scientific articles: Its distribution and dependence on the article features. Journal of Informetrics, 4, 1–13.
Vieira, E. S., & Gomes, J. A. N. F. (2011). The journal relative impact: an indicator for journal assessment. *Scientometrics, 89*(2), 631–651.

THE IMPACT OF R&D ACTIVITIES ON HOSPITAL OUTCOMES (RIP)

Antonio Garcia Romero[1], Josep A. Tribó[2] and Alvaro Escribano[3]

[1] *help@eco.uc3m.es*
Department of Economics. Universidad Carlos III de Madrid, c/Madrid, 126 28903 Getafe Madrid (Spain)

[2] *joatribo@emp.uc3m.es*
Department of Business, Universidad Carlos III de Madrid, c/Madrid, 126 28903 Getafe Madrid (Spain)

[1] *alvaroe@eco.uc3m.es*
Department of Economics. Universidad Carlos III de Madrid, c/Madrid, 126 28903 Getafe Madrid (Spain)

Abstract
In recent years, research policy stakeholders have emphasized their interest in the societal returns of research. The goal of this study is to assess the impact of research activities on Spanish hospitals clinical outcomes. To do so, we use a panel data set of Spanish hospitals, and we consider two fixed effects models, one for medical and the other for surgical specialties respectively. The use of panel data set allows us to explain causality among variables. Preliminary results show that scientific research contributes to reduce the length of stay in, both, medical and surgical specialties. In further research, we plan to enlarge our data set as well as the structure of the estimation models in order to explain other outcome indicators (i.e.: hospital discharges) as well as the temporal lags among the causal relationships. Preliminary evidence suggest that basic research has longer-lasting effects in comparison to more applied research

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3) Research Fronts and Emerging Issues (Topic 4).

Introduction
Traditionally, the evaluation of scientific research has been based only on output measures of performance. This approach has played a pivotal influence in the design on the most efficient research policies. Nevertheless, in recent years, research policy has been leaned towards a more stakeholder-oriented view that emphasizes the societal returns of research. Consistently, new methods and indicators capable of measuring the "real" effects of research on society have been developed (Smith, 2001; Cozzens, 2004).

In the specific case of health-related research, the analysis of its societal impacts is especially difficult due to three main reasons. First, there are few datasets that include the societal returns of biomedical research. Second, the way that research activity is connected to relevant issues for societal impact is very complex (Lewison, 2002). And third, there are many other factors, not related to biomedical research, which can also be associated with the same outcomes (Mushkin, 1979).

Previous works on this topic are scarce compared with the scientific literature relevant to other aspects of research policy. Nevertheless, several approaches have been applied successfully to the analysis of the relationship between research and clinical practice. For instance, the method proposed by Lewison, et al. (1998) and Grant, et al. (2000) to identify the flows of scientific knowledge from biomedical research to clinical practice. Another interesting approach is based on the Payback model to organize the assessment of the outcomes of health research (Hanney et al, 2004). Finally, some econometric approaches using a panel data set (Bonastre et al, 2011) have found no association of scientific production on the length of stay in French public hospitals.

This research is aimed to examine the impact of medical research activities on Spanish hospitals outcomes. To do so, we use a panel data set of Spanish hospitals, and we estimate two fixed effects models, one for medical specialties and the other for surgical ones. The final objective of this piece of research is to investigate an eventual causal relationship from research activities to hospital clinical outcomes.

Data and Methods

Data

The data used in this study were gathered from two different sources. First we used the Spanish Survey of Hospitals (ESCRI) hosted by the Ministry of Health. This survey provides relevant information regarding the human resources, organization, clinical outcomes or financial issues for a number of 1,000 hospitals in Spain. Although this survey was conducted annually during 1994-2011, we just used the period 1996-2004 due to the limitations in the bibliometric data set. Second, the bibliometric data set have been gathered from the data set "Bibliometric map of Spain 1996-2004: biomedicine and health sciences" (Méndez et al., 2005). This data set was built from SCI and SSCI after a harmonization and disambiguation of addresses. In addition, each article was assigned to a clinical specialty or basic research subfield. The definition of each variable used in this research and its descriptive statistics is shown in Table 1 and Table 2 respectively.

Table 1. Definition of the variables used in this analysis.

Variable	*Definition*	*Source*
LOS_MEDIC	Average length of stay for medical specialties	ESCRI
LOS_SURGIC	Average length of stay for surgical specialties	ESCRI
CLINPUB100	Clinical papers /100 (Physicians, surgeons and residents)	Méndez et al. (2005)
BASICPUB100	Basic papers /100 (Physicians, surgeons and residents)	Méndez et al. (2005)
CLINIMPACT	Normalized impact of clinical papers	Méndez et al. (2005)
BASICIMPACT	Normalized impact of basic papers	Méndez et al. (2005)
PHYSICIAN100	Number of physicians x 100 beds	ESCRI
SURGEON100	Number of surgeons x 100 beds	ESCRI
NURSING100	Number of nurses x 100 beds	ESCRI
ASSIST100	Number of nursing assistants x 100 beds	ESCRI
DUE100	NURSING100 + ASSIST100	ESCRI
RESIDENT100	Number of residents x 100 beds	ESCRI
ANALYSIS100	Number of clinical tests items x 100 beds	ESCRI
CT100	Computerized tomography x 100 beds	ESCRI
MRI100	Magnetic resonance imaging x 100 beds	ESCRI
XR100	X-Ray imaging x 100 beds	ESCRI
DRUGS100	Drugs expenditure x 100 beds	ESCRI
SURGINST100	Surgical instrument expenditure x 100 beds	ESCRI
INTASSETS100	Intangible assets expenditure x 100 beds	ESCRI
RDEXPEND100	R&D expenditure (CRO) x 100 beds	ESCRI
EMERGEN	Emergency overload: emer. admissions/total admissions	ESCRI
COMPLEXITY	Complexity index	ESCRI

The set of explanatory variables used in this study can be classified into five groups: (i) bibliometric indicators, (ii) human resources; (iii) diagnosis activity; (iv) hospital investment; and (v) hospital characteristics.

Regarding the bibliometric section we have included two indicators of production and other two of impact. We measured the scientific productivity by the number of documents published in clinical topics (CLINPUB100) and basic research sub-fields (BASICPUB100). We used the number of physicians, surgeons and residents to normalize these indicators. To evaluate the scientific impact, we selected indicators of the normalized impact, as the citation rate an institution receives compared to the world average. As in the former case we used an indicator for the clinical scientific production (CLINIMPACT) and another one for the basic research subfields (BASICIMPACT). When the value of these indicators is equal to one, it means that the observed citation rate of a hospital is similar to the world average in the same disciplines.

For the remaining group of variables we take advantage of a unique dataset composed of those hospitals included in the ESCRI data set. We matched this dataset with the bibliometric one using some key variables as well as an auxiliary data set that included the name of the hospitals. The matching process was initially made using computer assistance and two curators cleaned the result afterwards.

Table 2. Descriptive statistic for the variables used in this analysis.

Variable	*N*	*Mean*	*Std.Dev*	*Min*	*Max*
LOS_MEDIC	6890	10.7537	53.98185	1	2463.75
LOS_SURGIC	6988	4.481821	4.12757	0.235	155.6471
CLINPUB100	935	2.727498	6.138576	0	94.44444
BASICPUB100	926	5.851543	11.55026	0	150
CLINIMPACT	1060	.8774906	.9478966	0	18.92
BASICIMPACT	1077	.7708914	1.535795	0	33.33
PHYSICIAN100	9945	7.725919	11.28879	0	190
SURGEON100	9945	1.293509	3.939874	0	147.4359
NURSING100	9945	54.51605	51.13807	0	1250
ASSIST100	9945	50.40513	47.36602	0	1175
DUE100	9945	104.9212	98.35351	0	2425
RESIDENT100	9945	3.527671	7.962401	0	155.5556
ANALYSIS100	9945	333203.2	684892.4	0	1.37e+07
CT100	9945	1309.458	2878.603	0	108980
MRI100	9945	997.3003	6059.984	0	211320
XR100	9945	20642.46	31342.61	0	758693.1
DRUGS100	9945	733489.1	1658295	0	7.00e+07
SURGINST100	9945	36996.78	134444.3	0	4944600
INTASSETS100	9945	82966.59	1079794	0	7.21e+07
RDEXPEND100	9945	2403.78	31844.19	0	1623600
EMERGEN	9945	0.3523448	0.31093	0	1
COMPLEXITY	10101	1.291456	0.6226701	1	3

Method

Given the time-series nature of our data, we employed panel data techniques to test our hypotheses. We used hospital fixed-effect estimation because the Hausman test revealed a correlation between the hospital-specific error component and the explanatory variables. The persistence of our dependent variable of hospital productivity, led us to cluster standard errors by hospitals and prevent potential bias in the estimations (Petersen, 2009).

The models

We have considered two different models for medical and surgical specialties respectively based on the assumption that these fields could take slightly different patterns. For instance, there are some variables such as drug acquisition strongly associated with medical specialities while surgical instrument is clearly a variable associated to surgery outcomes. Taking into account these circumstances we have estimated two fixed effect models to explain the average length of stay (LOS) for medical and surgical specialties respectively.

$$LOS\_MEDIC = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i$$

$$LOS\_SURGIC = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i$$

where β_i represents the coefficient associated with the x_i variable.

Table 3. Empirical results for the average stays in medical and surgical specialties

Variables	*Y=LOS_MEDIC*	*Y=LOS_SURGIC*
Fixed effects		
Intercept	10.487*** (0.592)	5.947*** (0.594)
Scientific production		
CLINPUB100	0.037 (0.023)	-0.007 (0. 023)
BASICPUB100	-0.012 (0.017)	-0.036* (0.017)
CLINIMPACT	0.025 (0.016)	-0.011 (0.016)
BASICIMPACT	-0.089* (0.046)	0.040 (0.046)
Human resources		
PHYSICIAN100	-0.045*** (0.010)	--
SURGEON100	--	-0.013 (0.011)
NURSING100	0.003 (0.029)	--
ASSIST100	0.001 (0.031)	--
DUE100	--	-0.003*** (0.001)
RESIDENT100	0.012 (0.007)	0.007 (0.007)
Diagnosis activity		
ANALYSIS100	0.000 (0.000)	--
CT100	-0.0003*** (5.73e-05)	-0.0002** (5.42e-05)
MRI100	-0.0003** (7.6e-05)	-0.0002** (7.37e-05)
XR100	-0.00001 (6.43e-06)	4.62e-06 (6.39e-06)
Hospital investment		
DRUGS100	-8.05e-08* (3.18e-08)	--
SURGINST100	--	-8.67e-07 (7.29e-07)
INTASSETS100	2.70e-08* (1.18e-08)	3.45e-10 (1.17e-08)
RDEXPEND100	-3.25e-07 (2.77e-07)	-2.60e-06 (2.17e-06)
Hospital characteristics		
EMERGEN	1.057 (0.677)	4.474*** (0.679)
COMPLEXITY	-.606*** (0.120)	-0.444*** (0.119)
F (p-value)	9.55 (0.000)	7.18 (0.000)
R^2	0.192	0.249

[a] Parameter estimation and standard errors in parentheses
* p< 0.05, ** p<0,01, *** p<0,001

Results

Preliminary results show that scientific research contributes to reduce length of stay in both, medical and surgical specialties (Table 3). First, for patients treated

by physicians, the relative impact of basic research contributes to reduce the length of stay (-0.089 with $p<0.05$). Second, regarding the surgical specialties, we have found a negative impact of the number of papers published per 100 doctors in surgical length of stay (-0.036 with $p<0.05$). These results could be justified by the fact that Medicine has a stronger dependence on basic knowledge than Surgery has. With regard to the significance of BASICPUB100 in the model for the length of stay in surgical specialties, we may argue that for surgery it is more important productivity research (intensive measure) rather than extensive research, which was relevant to explain the stay length in medicine. Besides, such relationship between length of stay in surgery and research productivity may capture the characteristics of the hospitals where better surgeons are attracted to hospital with better researchers. We plan to explore this issue in all of its depth in further research. Regarding the role of the rest of variables for each model, we found that the relative number of physicians, computerized tomography, as well as magnetic resonance imaging, contribute to reduce the stay length of stay.

In the case of the surgery patients, the length of stay seems to be reduced both by the nursing staff and the use of advanced imaging techniques (CT and MRI). Regarding the hospital characteristics, we observed a positive effect of the emergency overload on the length of stay for surgical specialties. Finally, the effect of hospital complexity (based on hospital discharges), suggest that the complex hospitals are also more efficient. However, this effect could be due to the chronic patients who are treated typically in less experienced centers.

Conclusions and further research

The effect of R&D activities on hospitals outcomes is a relevant issue for policy decision makers in health and research activities. The apparently intuitive positive relationship between research and clinical outcomes has not been unambiguously shown probably because of a lack of reliable data sets. We assemble a panel data set by combining two sources. On the one hand, the Spanish Survey of Hospitals hosted by the Spanish Ministry of Health, and on the other hand, the bibliometric map of Biomedical research in Spain elaborated by Mendez et al. (2005) using data from the Web of Science and SCOPUS. This rich dataset have allowed to show the existence of a clear relationship between medical research in a hospital and clinical performance in this hospital

In further research, we aim to advance this research in four ways. Firstly, we will update the bibliometric information to recent years as well as to include additional bibliometric indicators. Secondly, we will investigate the strength of the effects shown by considering different temporal in the explanatory variables like those of research productivity. Thirdly, we will also explore the impact of research on other clinical outcomes such as hospital discharges. Lastly, we will examine whether there are interactions among variables introducing multiplicative terms into the models.

Acknowledgments

We are grateful to financial support from the Ministry of Science and Innovation (ECO2009-10796 and ECO2012-36559) and the Cátedra Telefónica-UC3M "Economía de las Telecomunicaciones".

References

Bonastre, J., Le Vaillant, M. & De Pouvourville, G. (2011). The impact of research on hospital costs. *Health Economics*, 20, 73-84.

Méndez-Vásquez,RI., Suñén-Pinyol, E., Cervelló, R. & Camí J. (2005). Mapa bibliométrico de España 1996-2004: biomedicina y ciencias de la salud. *Medicina Clínica*, 130, 246-253.

Cozzens, S.E. (2004). "Socioeconomic impact indicators: old measures, new models". *8th Science and Technology Indicators Conference*. Leiden (The Netherlands).

Grant J., Cottrell R., Cluzeau F. & Fawcett G. (2000). Evaluating "payback" on biomedical research from papers cited in clinical guidelines: applied bibliometric study, *BMJ*, 320, 1107-1111.

Hanney S., Grant J., Wooding S. & Buxton M. (2004). Proposed methods for reviewing the outcomes of research: the impact of funding by the UK's Arthritis Reserach Campaign, *Health Research Policy and Systems*, 2.

Lewison, G. & Dawson, G. (1998). The effect of funding on the outputs of biomedical research, *Scientometrics* 41: 17-27.

Lewison, G. (2002). From biomedical research to health improvement, *Scientometrics*. 54:179-92.

Mushkin, S. (1979). *Biomedical research: costs and benefits*, Ballinger Publishing Company, Cambridge (MA).

Petersen, M. A. 2009. Estimating standard errors in finance panel data sets: comparing approaches. *Review of Financial Studies*, 22, 435-480.

Smith, R. (2001). Measuring the social impact of research. Difficult but Necessary, *BMJ*, 323: 528.

INDUSTRY RESEARCH PRODUCTION AND LINKAGES WITH ACADEMIA: EVIDENCE FROM UK SCIENCE PARKS

David Minguillo, Mike Thelwall

D.MinguilloBrehaut2@wlv.ac.uk; M.Thelwall@wlv.ac.uk
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, WV1 1SB Wolverhampton (UK)

Abstract
The aim of this study is to identify the research areas, geographic regions, university-industry (U-I) collaborations, quality, and impact of the research associated with the research-intensive organisations based in the UK science parks. An analysis of scholarly publications (1975-2010) revealed three main research domains: food-biotechnology and bio-pharmacology; physics and material engineering; and agro-biotechnology. These three types of research were mainly produced in East England, South East England, and Scotland, respectively. Only a quarter of the research results from inter-institutional cooperation. The high involvement of private sector in the physics and material engineering domain involves the highest rate of U-I collaboration but the lowest citation impact. The research quality, defined in terms of the journals where research is published, is significantly higher than the average across research areas, although its impact is not significantly higher than the national average. In terms of inter-sector differences, the higher the involvement of Higher Education Institutions (HEIs) and Research Institutions (RIs) the greater the impact of the publications produced. The low level of impact of private research suggests that citations may not be the best indicator to assess academic researchers with close and operational linkages with industry.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5); Scientometrics Indicators: Relevance to Science and Technology (Topic 1).

Introduction
The sustainability of socio-economic development among developed countries increasingly depends on the capacity to foster dynamic and strong research-based industries. In this regard, European and national policies highlight the potential role of university as a main source of research, technology and innovation, and actively promotes closer links with industry (Dyson, 2010; Hauser, 2010; Lambert, 2003). However, this university-industry (U-I) collaboration is not always a straightforward process as the academic and private communities belong to systems that differ in their identity and mission, bringing about transaction costs associated with the efforts employed to bridge the gap between both communities (Abramo, et al., 2009; Arvanitis, Kubli, & Woerter, 2008). In fact,

this interaction barrier has led to create an entire constellation of actors oriented to encourage and facilitate the multidimensional and complex process of capitalisation and transference of academic knowledge (Minguillo & Thelwall, 2011; Suvinen, Konttinen, & Nieminen, 2010).

One of the most important and long-standing members of this support constellation are intermediary infrastructures: incubators, science parks, research and technology parks, and innovation parks. These policy tools are widely known as science parks (SPs), and are basically physical infrastructures established in partnerships between research-intensive universities, public authorities and private investors to create favourable conditions to facilitate U-I collaboration and boost technological innovation, and ultimately generate local socio-economic growth (Link & Scott, 2007; UKSPA, 2012; Vedovello, 1997). Yet the pivotal role of SPs in the commercialisation of academic research and technology (R&T) obviously has a significant impact on the goals and functions of universities, and in turn on part of the scientific community. The assessment of SPs mainly focuses on finding out to what extent the links with universities are able to stimulate the growth of cutting-edge industries and a competitive advantage for businesses located on SPs in comparison to their off-park counterparts (Quintas, Wield, & Massey, 1992; Rothaermel & Thursby, 2005; Schwartz & Hornych, 2010; Siegel, Westhead, & Wright, 2003; Westhead & Storey, 1995).

A growing interest in studying factors that may strengthen U-I interaction and encourage a stronger research-orientation in industry has led to suggestions that the use of a scientometric approach may give a fuller understanding of the impact of SPs on the synergy between industry and academia (Bigliardi, et al., 2006; Fukugawa, 2006; Link & Scott, 2003; Siegel et al., 2003). Although, there are two relevant studies regarding the Hsinchu SP in Taiwan, employing bibliographic (Hu, 2011) and patent data (Hung, 2012), and a third one using web-based data to study the SPs in the region of Yorkshire and the Humber in the UK (Minguillo & Thelwall, 2012), it is necessary to conduct further studies that map the research capability and properties of on-park businesses across regions and countries. This could shed new light on the intermediary role of SPs, provide empirical evidence for the literature regarding U-I collaboration in general (Teixeira & Mota, 2012), and most importantly guide and support more effective U-I collaboration processes in developed countries.

With this in mind, this study mainly analyses the capacity of the UK SP movement to encourage and generate R&T. The focus is on providing a better understanding of two specific aspects; (1) the research areas that attract most of the on-park research and the contribution of the geographic regions and U-I collaboration across different areas; and (2) whether the research production associated with SPs has a greater quality and impact than the average research across the different areas. These aspects provide an insight into the R&D activities and U-I links that are expected to be fostered by the different support infrastructures, and to what extent the on-park research is integrated into the wider scientific community.

Data and methodology

Publications associated with UK SPs were retrieved from Elsevier's Scopus database covering a period of 35 years (1975-2010). We used two different approaches to retrieve the records of the research publications produced by any organisation located within a SP in the UK. First, with the help of the SP list provided by the United Kingdom Science Park Association (UKSPA) and the electronic version of the *Atlas of Innovation* created by the *World Alliance for Innovation* (Wainova) we identified the names of 82 full members across the country. This allowed for the creation of queries with the specific names of the different SPs (e.g., *AFFIL ("norwich research park") AND (LIMIT-TO(AFFILCOUNTRY, "United Kingdom"))*). Second, to extend the first search and identify non-members of the UKSPA we used truncated queries with terms that are broadly used to name research-based infrastructures in the country, such as science-, technology-, innovation park, and incubator, as well as terms for commercial-based infrastructures, such as business-, industrial-, enterprise park, and business centre (i.e. *AFFIL("sci* park") AND (LIMIT-TO(AFFILCOUNTRY, "United Kingdom")*). Both specific and truncated queries were restricted to the year 2010 covering journals, book series, and conference proceedings, while excluding editorials, erratum, letters, and notes.[84] The search yielded 10,920 records.

A similar search strategy was used on the *Web of Science* (WoS) database (Thomson Reuters) but approximately two thousand fewer records were retrieved using this method. Note that not all onsite organisations mention the SPs where they are located as part of their affiliation addresses in research publications, so this search approach may not take all the relevant publications into account. Data cleaning and standardisation was used to identify all publications listing at least one author address referring to a UK SP, and the author address was checked for a correct assignment to the organisation stated by the author. The research produced by departments, sub-units, or company groups was assigned to the parent entity, and only research centres associated with HEIs were treated independently in order to get more fine-grained results. In the case of firms, name changes, mergers, or acquisitions were taken into account where possible but in most cases organisations with different physical locations were treated separately to quantify the impact of SPs on the immediate environment. Most hospitals in SPs are teaching hospitals and were classified as HEIs, as recommended in the *Frascati Manual* (2002).The organisations were grouped into six groups (higher education, industry, government, on-park organisation, non-profit organisation, and research institute), and other main attributes (type of organisation, location, type of location, and district). We obtained 9,771 publications produced by at least one onsite-organisation.

[84] This selection of document types is based on their relevance as public communication channels for industry research outputs (Cohen, Nelson, & Walsh, 2002).

The research subject areas were taken from the Scopus journal classification scheme, and publications placed in journals indexed in more than one subject area are counted in each one. These areas are also used to identify the degree of participation of the private and academic sectors, of the regions, and of the U-I collaboration. Reputation, in form of citations given by the research community, was used to determine the popularity and impact of the research. The prestige was determined in two ways. First, quality was approximated by the number of citations received by the journals of the publications. This is quantified by the two citation based indicators; Scimago Journal Ranking (SJR) and Source Normalised Impact per Paper (SNIP), as both are designed to evaluate the prestige and visibility of journals in relation to the particular characteristics of a research area. Second, impact was approximated by the number of citations received by each individual publication. Finally, the Wilcoxon signed-rank test, which is the non-parametric equivalent of the t-test, was applied to assess if there is a significant difference between the observed and expected quality and impact of the research across subject areas.

Results

As background information, the data set extracted from *Scopus* outperforms the *Web of Science* in terms of representing the heterogeneous publication output of a mainly private oriented research community associated to the SP movement (see Figure 1).

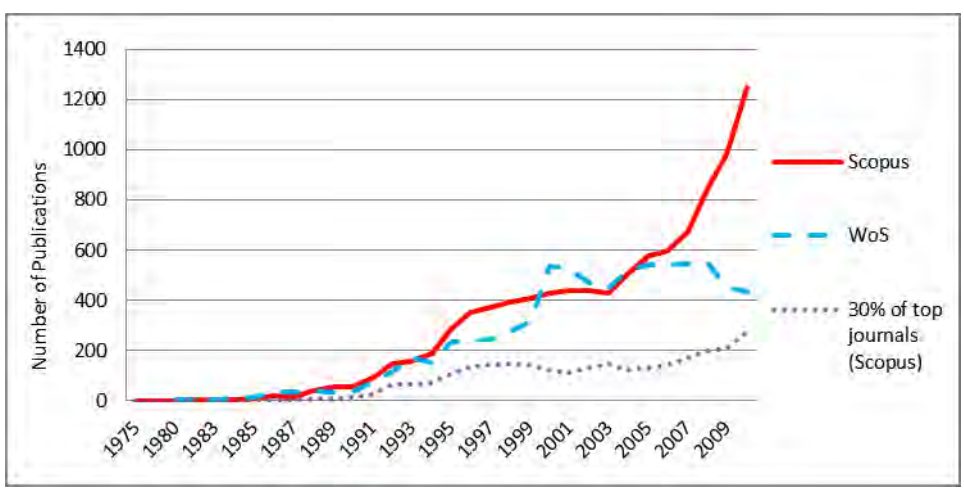


Figure 1. Publications from the UK SP movement from 1975 to 2010.

The coverage of WoS and Scopus seem to be very similar until the mid 90s, after which Scopus exhibits an exponential growth compared to the flat and even decreasing WoS coverage. No bias that would account for the difference could be identified by the publication sources or type of sources indexed by Scopus, as

demonstrated by the normal distribution of the top 30% largest journals in Scopus. The WoS output trend confirms previous findings indicating that WoS-indexed research produced by industry is steadily declining (Tijssen, 2004). These findings strongly suggest that the publication output of the SP movement is underrepresented in WoS.

The chronological development of the SP movement reported in Figure 2 contains the number of infrastructures which have been research-active every year of their existence in terms of research publication output. This shows that the constant growth of the output, shown in Figure 1, coincides with an increase in SPs that are involved in research activities. Before the 1990s there were, on average, 4.5 research-active SPs every year. During one decade this number increased to 24.5, resulting in a more than a two-fold increase by 2010 to a total of 61 SPs. Similarly, the output trend started to become substantial in the beginning of the 1990s, reaching over 400 publications in 2000 with a further three-fold increase by 2010.

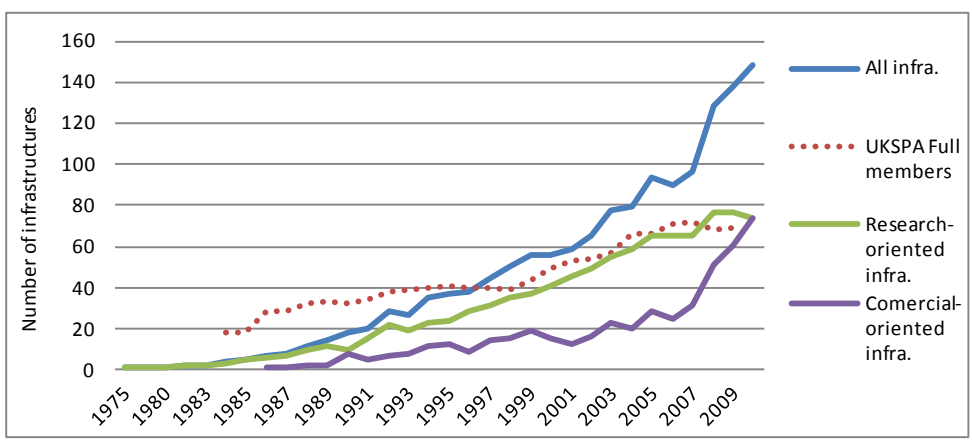


Figure 2. Number of research- and commercial infrastructures producing research publications in each year (Scopus data).

Figure 2 also illustrates that one of the reasons for the remarkable increase of records in Scopus could be an increase in the number of commercial-oriented infrastructures producing research in the last years. The distribution followed by the research-oriented infrastructures publishing every year shows a similar distribution to the records in WoS (see Figure 1).

Research subject areas, collaborative efforts, quality and impact of the SP movement

Scholarly journals are the main venues for formal interaction and communication for different scientific communities, making it possible to identify the intellectual and social aspects shared. These two aspects provide the framework that forms

each knowledge domain, and the distance between domains can be determined by the degree of similarity between their cognitive and reputational systems, which in turn shapes the structure of science as a whole (Minguillo, 2010). Hence, the output of the SP movement helps, among other things, to shed light on their degree of intellectual and social integration into the wider scientific community. To do this, the research areas with the largest number of publications were identified based on the journals where the research is frequently disseminated.

Research subject areas and Collaborative efforts

The most frequent Scopus-indexed type of source for the research generated by SPs is journals (91%), in comparison to conference proceedings (7%), serials (1%), and generic (1%). The low rate of conference proceedings is somewhat surprising because conferences are considered as potential venues of interaction for industry and academia (D'Este & Patel, 2007; Lee & Win, 2004), and indeed, in the last ten years there has been an increasing trend for participating in conferences, as shown by the fact that 83% of all conference publications were published between 2005-2010, representing 12% of all publications over the last five years. This growth is the result of the intensification of R&D activities in technology areas, such as *Engineering*, *Physic and Astronomy*, and *Materials Science*.

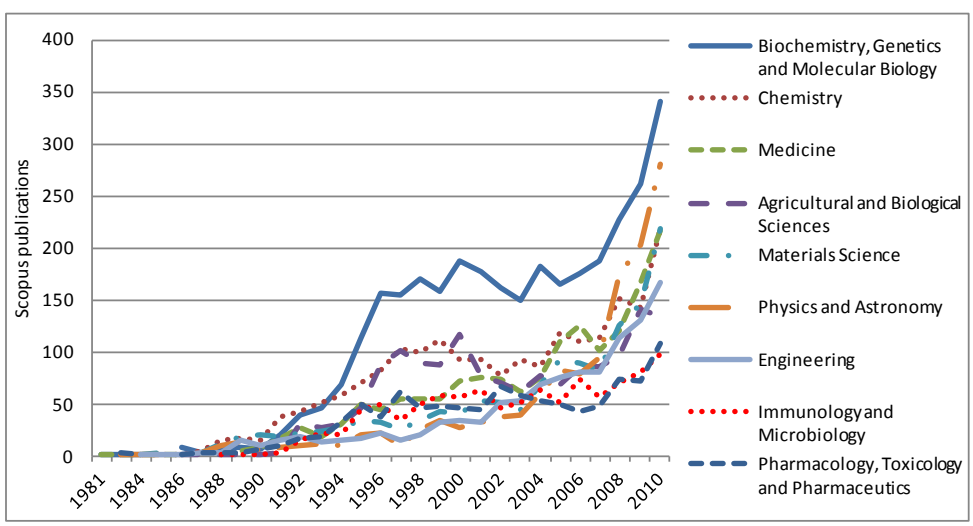


Figure 3. Chronological development of the top nine subject areas for the SP movement.

Regarding the most important research fields, the chronological development of the top nine subject areas, covering 80% of the total output, shows that *Biochemistry, Genetics and Molecular Biology* is the largest research area with 18% of the total output (Figure 3). It started in the mid 80s and has its first

breakthrough in the mid 90s due to the establishment of RIs (e.g. Institute of Food Research, and the John Innes Centre), the parallel relocation of the pharmaceutical industry (e.g. GlaxoSmithKline) and the emergence of new spin-outs. In 2005 it again had exponential growth partially caused by the diversification and maturity of the industry and new emerging RIs (e.g. Babraham Institute). This trend differs from the relative decline suffered by *Chemistry*, and *Agricultural Biological Sciences* during 2000 and 2007. The other top subject areas have followed a constant growth and have similarly achieved a remarkable upward increase since 2005. Three related subject areas have been subject to recent exponential growth, namely *Physics and Astronomy*, *Material Science*, and *Engineering*, and this is partially caused by the RIs *Rutherford Appleton Laboratory* and the private sector (e.g. AkzoNobel R&D, Diamond Light Source, TWI). On one hand, these two sets of fields represent the emerging physics and material engineering industrial sector and, on the other hand, the partially weakening health and life science industrial sector, consisting of three subject areas: *Biochemistry, Genetics and Molecular Biology*, *Chemistry*, and *Agricultural Biological Sciences*. Both groups also differ in terms of research and technology producers as the first is slightly dominated by firms (64%) and the second by public RIs & HEIs (72%) (see Table 1), suggesting the maturity of new research-based industrial sectors, mostly produced by the private sector, that coexists with the well-established and publicly backed bio-tech industry within the SP movement.

The ranking of the top 15 subject areas in output (Table 1) illustrates characteristics of the research associated with the SP movement, the research profile of the three regions with the greatest research-intensive innovation structures, and the collaboration between on-park organisations (firms or HEIs/RIs) with on- or off-park organisations (firms or HEIs/RIs). At the regional level, the most productive is the East of England with the top subject area *Biochemistry, Genetics and Molecular Biology*. This depends upon the high concentration of small and large biotech firms (Birch, 2009), that in turn are highly dependent upon public RIs, as shown by the low share of private research (38%). This region also produces significant research in *Agricultural Biological Sciences* and *Chemistry*, and despite generating considerable research in other research fields, the region seems to be public science-based and specialised in food-biotechnology and bio-pharmacology. The research and technology from the South East is framed within four important areas *Physics and Astronomy*, *Materials Science*, *Engineering* and *Chemistry*, and even though there are public RIs that support the two first research areas, the role of industry as a research producer is significant (63%). Another region with a similar profile is the North East. Hence, the South East region seems to rely on private research to develop an industrial sector around physics and material engineering. Finally, Scotland, with a reduced private research capacity (35%), relies on public research (e.g. *Moredun RI, Roslin Institute, Veterinary Laboratories Agency*) to concentrate research related to *Immunology*, *Medicine*, *Veterinary* and *Biochemistry, Genetics and*

Molecular Biology, which in turn is exploited by the agro-biotech industry, confirming previous findings (Cooke, 2001). On the other hand, the subject areas with the highest rate of private participation are *Pharmacology* (81%), *Materials Science* (67%), and *Engineering* (66%); conversely the highest academic contribution is found in *Agricultural and Biological Sciences* (85%) and *Immunology* (80%).

Table 1. Distribution of the top subject areas according to private and academic output, regions, and inter-institutional collaborative efforts.

		Output						Three main regions' Output						Collaboration			
#	Research area	*n* = 17,341	%	#	Industry *n* (45%)	#	HEIs/Ris *n* (52%)	#	a	#	b	#	c	#	All *n* (25%)	#	U-I *n* (56%)
(1)	Biochemistry, Genetics & Molecular Biology*	3182	18%	(10)	36%	(5)	62%	(1)	26%	(5)	9%	(4)	10%	(11)	18%	(9)	46%
(2)	Chemistry*	2009	12%	(6)	58%	(9)	41%	(3)	12%	(4)	12%	(12)	2%	(6)	34%	(5)	67%
(3)	Medicine***	1572	9%	(9)	39%	(7)	55%	(4)	8%	(7)	5%	(2)	15%	(12)	15%	(12)	44%
(4)	Agricultural and Biological Sciences*	1535	9%	(15)	12%	(1)	85%	(2)	13%	(13)	2%	(5)	8%	(13)	14%	(14)	32%
(5)	Physics and Astronomy**	1334	8%	(7)	58%	(11)	39%	(7)	4%	(2)	13%	(11)	4%	(7)	33%	(2)	73%
(6)	Materials Science**	1300	7%	(2)	67%	(13)	32%	(8)	4%	(1)	20%	(10)	4%	(1)	52%	(1)	74%
(7)	Engineering**	1097	6%	(3)	66%	(14)	31%	(9)	4%	(3)	12%	(9)	5%	(5)	35%	(3)	71%
(8)	Immunology and Microbiology***	1015	6%	(13)	18%	(2)	80%	(6)	7%	(16)	1%	(1)	15%	(14)	14%	(13)	33%
(9)	Pharmacology, Toxicology & Pharmaceutics	1006	6%	(1)	81%	(15)	17%	(5)	7%	(9)	4%	(8)	5%	(8)	20%	(7)	65%
(10)	Chemical Engineering	551	3%	(5)	58%	(10)	41%	(10)	3%	(10)	3%	(14)	1%	(3)	38%	(6)	65%
(11)	Environmental Science	484	3%	(11)	34%	(6)	62%	(11)	2%	(12)	2%	(7)	6%	(9)	20%	(10)	45%
(12)	Computer Science	391	2%	(4)	64%	(12)	33%	(14)	1%	(6)	5%	(13)	1%	(4)	36%	(4)	68%
(13)	Mathematics	294	2%	(8)	55%	(8)	44%	(16)	1%	(8)	4%	(15)	1%	(2)	39%	(8)	63%
(14)	Veterinary***	287	2%	(14)	16%	(3)	80%					(3)	12%	(15)	10%	(15)	25%
(15)	Earth and Planetary Sciences	285	2%	(12)	33%	(4)	63%			(11)	2%	(6)	8%	(10)	18%	(11)	44%

a East of England (n=54%; I=38%); *b* South East (n=14%; I=63%); *c* Scotland (n=12%; I=35%)

* Food-biotechnology and Bio-pharmacology; ** Physics and Material engineering; *** Agro-biotechnology

Regarding inter-institutional collaboration, only 25% of all the research output has been co-authored by two or more different institutions, with *Material Science* being the area with the highest collaborative effort. From these collaborations, more than half (56%) are U-I, and there is a strong relationship (r_s=0.86) between the ranking of private output and U-I collaboration across the research areas. This shows that the research-intensive industries within the SP movement are able, to some extent, to capitalise on academic knowledge. Interestingly, the comparison between research areas in terms of U-I collaboration shows that the three top areas belong to the physics and material engineering industry, implying that the South Eastern agglomeration is the most successful in fostering U-I interaction. On the other hand, the low ranking of the other two main industrial agglomerations, food-biotechnology and bio-pharmacology, and agro-biotechnology – mainly based in East of England and Scotland respectively - is affected by the central role of the public research and especially RIs. Although most RIs are meant to closely support and cooperate with local businesses, they are industry-related and the outcome of the cooperation with private sector may not necessarily lead to the publication of research articles.

Quality and Impact

The quality is basically defined by capacity to place publications in journals that attract a considerable amount of citations from its research area. The quality of the

output was obtained through comparing the expected quality (the average value of the SJR and SNIP given to each subject area in 2010) with the observed quality (the average value of the 2010 SJR and SNIP of the journals where on-park organisations publish). If the observed quality is higher than the expected quality then this is evidence that the research of on-park organisations is good enough to be disseminated among the most prestigious journals in the area. On the other hand, the impact of the output, defined by the number of citations that each publication receives, is obtained through comparing the expected impact (the average number of citations received by the publications in each subject area), with the observed impact (the average number of citations received by on-park organisations' publications). Then, if the observed impact is higher than the expected one it is assumed that the on-park research is relevant and attracts the attention of the research community.

Table 2. Quality and impact of the top subject areas.

	Quality				Impact (1996-2010)		
	SNIP		SJR		Observed		Expected
	Observed	Expected	Observed	Expected	n=18.44	St dev	n=16
Biochemistry, Genetics and Molecular Biology*	1.42	0.78	0.68	0.42	25.12	40.66	28.46
Chemistry*	1.35	0.88	0.23	0.15	16.50	27.45	18.76
Medicine***	1.26	0.77	0.41	0.13	18.75	33.88	17.86
Agricultural and Biological Sciences*	1.33	0.64	0.25	0.10	23.21	36.97	18.51
Materials Science**	1.15	0.91	0.14	0.10	11.06	23.35	11.57
Physics and Astronomy**	1.12	1.14	0.13	0.11	8.01	21.48	15.18
Engineering**	1.34	0.80	0.12	0.06	7.52	21.35	8.12
Immunology and Microbiology***	1.39	1.45	0.63	0.40	21.45	28.98	24.01
Pharmacology, Toxicology and Pharmaceutics	1.03	0.49	0.29	0.15	18.87	30.52	17.72
Chemical Engineering	1.42	0.63	0.28	0.09	15.43	29.65	10.7
Environmental Science	1.37	0.67	0.13	0.08	15.03	27.04	18.55
Computer Science	1.62	1.49	0.70	0.06	6.56	43.30	10.23
Mathematics	1.20	1.01	0.07	0.05	6.36	49.56	9.95
Veterinary***	1.02	0.56	0.10	0.06	13.12	24.71	9.23
Earth and Planetary Sciences	1.40	0.51	1.10	0.07	10.13	17.09	17.96

* Food-biotechnology and bio-pharmacology; ** Physics and material engineering; *** Agro-biotechnology

SNIP Source: www.journalindicators.com

SJR Source: www.scimagojr.com

Table 2 illustrates that the SP movement as a whole is capable of publishing in the most influential journals and these publications have a higher impact than the national average. Based on the SNIP indicator, the difference between the observed and expected quality suggests that the areas with highest quality are *Earth and Planetary Sciences*, *Chemical Engineering*, and *Agricultural and Biological Sciences*, while those with lower quality are *Immunology and Microbiology* and *Physics and Astronomy*. The comparison based on the SJR supports the high quality of on-park research, with the areas of highest quality being *Earth and Planetary Sciences* and *Computer Science*. In terms of impact of the output, between the period 1996 and 2010, 79% of the publications have been cited and the observed impact is higher (18.44) than the expected one (16). However, only five areas seem to have higher impact than expected, the highest being; *Chemical Engineering*, *Agricultural and Biological Sciences*, and

Veterinary. On the other hand, the areas with the lowest relative impact are: *Earth and Planetary Sciences* and *Physics and Astronomy*.

The Wilcoxon signed-rank test compares the expected and observed values, confirming that the quality measured by the SNIP (z=-3.238, p<.05) and SJR (z=-3.409, p<.05) of the journals within the different subject areas is significantly higher than the expected. On the other hand, the level of impact obtained by the publications is only slightly higher than expected with a difference that is not statistically significant (z=-.966, p>.05). This reveals that the organisations associated to the SP movement are able to publish in high-quality journals, although the impact of these publications on the scientific community varies across areas and tends to be only slightly greater than the average.

Different factors may lead areas with high quality to have low impact and vice versa. When the top quality research areas are compared based on the three main regional agglomerations (non shown), the observed quality reveals that research in food-biotechnology and bio-pharmacology industries in the East of England has a much higher value (2.63) than the agro-biotech industry in Scotland (2.40), and the physics and material engineering industry primarily located in the South East (2.0). The citations, however, show that only the agro-biotech sector has a positive impact (0.74), whereas the impact of food-biotechnology and bio-pharmacology (-0.3) and physics and material engineering sectors (-2.75) are below the expected values. The main reason for this could be the nature of the research. As Godin (1996) claims, basic research produced by industry in biotechnology and chemistry is more useful for the research community and thus more cited than the applied research produced by industry in physics. The applied nature of the research generated in physics and material engineering is reflected in the greater dissemination of research in the form of conference proceedings, for example. Another reason could be that the private-oriented sectors have only experienced a strong increase over the last ten or five years, and thus, have had less time to be cited.

Table 3. Citation rates of regions, infrastructures, and organisations.

	Citations per publication								
Region	**IN**		**OUT**		**Infrastructure**	**IN**	**Organisation**	**IN**	**OUT**
	#	*n=19.2*	*#*	*n=22.1*		n=19.7		*n=19.7*	*n=21.2*
East of England	1	26.9	1	30.2	Research Camp	48.6	Research Institutes	25.7	25.6
North West England	2	16.0	2	29.4	Research Pk	27.8	Firms	15.2	17.4
Scotland	3	13.3	3	28.2	Incubator	16.0	HEIs	14.3	19.9
North East England	4	13.3	4	27.1	Science Pk	14.8	Government	6.3	10.2
South West England	5	12.4	5	19.9	Innovation Pk	13.8	Non-profit organisations	5.8	182.0
East Midlands	6	11.7	6	19.1	Science & Innovation Cent	12.6	**% of uncited publications**	**IN**	**OUT**
London	7	10.4	7	17.7	Industrial Pk	8.9	Organisation	*n=0.21*	*n=0.21*
West Midlands	8	9.7	8	16.8	Business Pk	8.6	Research Institutes	0.13	0.15
Yorkshire and the Humber	9	8.5	9	15.5	Technology Pk	8.3	Firms	0.27	0.25
South East England	10	7.9	10	15.3			HEIs	0.29	0.21
Wales	11	7.3	11	12.3			Government	0.40	0.23
Northern Ireland	12	3.4	12	11.0			Non-profit organisations	0.43	0.26

To find the reason for the inconsistency between the quality and impact of the output the characteristics of the impact across regions, infrastructures, and types of organisations were examined. First, Table 3 reports the citation rates of the on- and off-park organisations. Interestingly, at the national level the evidence indicates that on-park research production, chiefly conducted by the private sector, had a slightly lower impact (19.2) than the off-park production (22.1) which is chiefly conducted by HEIs. At the regional level, the low impact of the private research base in the South East, which occupies the tenth position, differs from the top positions of the primarily public research generated in the East of England and Scotland. The impact of the off-park organisations shows that the exchange of research with off-park organisations located in the North East, London, and the East of England attracted the interest of the research community, increasing its impact.

Similarly, the level of impact of the infrastructures and organisations (see Table 3), clearly shows that the closer the research production is to public RIs the greater the research impact. Infrastructures whit a greater part of the output generated by RIs, research- campuses (48.6) and parks (27.8), and, to a lesser extent, incubators (16), and science parks (14.8), have a greater impact than the business-oriented infrastructures, namely industrial- (8.9) and business- parks (8.6). Most of these RIs are recognised centres of excellence and the research produced by RIs, regardless of being on (25.7) or off park (25.6), leads to the highest impact for the on-park research community. On the other hand, it is difficult to argue that the research produced with the participation of either firms or HEIs could receive more citations due to the high level of collaboration between both.

Discussion

The result showed that Scopus provides a wider coverage of the research output of the SP movement in comparison with WoS. Scopus' broad coverage policy, with about 70% more sources than WoS (López-Illescas, Moya-Anegón, & Moed, 2008), offers a more comprehensive representation of the industrial research. This is especially true when conference proceedings are important (Meho & Rogers, 2008). The likely underrepresentation of private research in WoS represents a significant limitation for U-I studies, as any conclusions drawn are related to the properties of the bibliographical database used.

Overall, the SP movement prefers to publish in journals and the expansion of technology fields has recently increased the use of conference proceedings as source of communication. Besides this, the growing interest from commercial-oriented business parks to promote R&D activities as a means to add value to the products and services of their tenants involves new opportunities for further expansion of the SP movement, as it has been able to redefine itself to nurture a greater research production in the last two decades.

Quantitatively speaking, the interdisciplinary field of *Biochemistry, Genetics and Molecular Biology* is the main research field of the movement, and the East of

England possesses the main private and public agglomeration across the country, which in turn is related food-biotechnology and bio-pharmacology, in line with other findings (Birch, 2009). Despite two closely related areas (*Chemistry*, and *Agricultural Biological Sciences*) to the food-biotech and bio-pharma sector suffering a slight decline between 2000 and 2007, the research output of this important sector is underpinned by the convergence of recognised centres of research excellence that form an important public science base, along with a considerable group of international companies and spin-outs. The high visibility of this sector is also partially the result of the heavy publishing activity of bio-related companies (Cockburn & Henderson, 1998). The other two sets of top agglomerations are tightly related with either the South East or Scotland; the first is configured by an emerging private and multidisciplinary research base that is exploited by the physics and material engineering sector, while the latter is characterized by a considerable public research base focused on agro-biotechnology. The characteristics of both agglomerations also have been highlighted by Cooke (2001), while the slight decline in research of areas considered within food-biotechnology and bio-pharmacology may reflect the important weakening of the pharmaceutical industry in the UK and Europe (Rafols et al., 2012). The chronological trend followed by, at least, these three main agglomerations illustrates the potential influence of public strategy in the establishment of research units and partnerships within SPs as a way to support the emergence of new industries. Link and Scott (2003), also show how the historical development of SPs in the United States is influenced by public policies, promoting an early emergence of medical centres and aerospace technology that are then replaced by a biotechnology and biomedical industry. This policy-driven development may also be the reason for the difference between the subject areas distribution of the SP movement with those found among patenting off-park firms where *physics*, *engineering*, *clinical medicine*, *chemistry*, and *biomedical science* are the most popular fields, for example (Godin, 1996).

In terms of collaborative efforts, only a quarter of the output is the result of an inter-institutional collaboration, of which more than half is between HEIs/RIs and industry. This national rate of U-I collaboration is considerable low in comparison with the 34% found on the Hsinchu science park, for example (Hung, 2012). The significant involvement of the private sector in the research production related to physics and material engineering, in turn leads this domain to be the most successful in bridging the U-I gap and represents an attractive market niche for the commercialisation of academic R&T. The explanation for the active participation of industry in R&D activities in this domain is that industry needs to develop their own expertise in physics, while the life science sector relies more on external research (Godin, 1996). However, the central role of the public research infrastructure, mostly RIs, in the high visibility of the other two main domains (Food-biotechnology and bio-pharmacology, and Agro-biotechnology), seems to generate an unexpectedly low rate of U-I collaboration. Most RIs tend to have a lower publication average in comparison with Universities, as factors such as,

human resources, value to publishing, and rewarding system differ between HEIs and RIs (Hayati & Ebrahimy, 2009; Noyons, Moed, & Luwel, 1999). In fact, the top position for the areas related to *Physics and Material engineering*, in terms of U-I collaboration, coincides with the study of Abramo and his colleagues (2009) who found that U-I collaboration in Italy is chiefly established in *Electronic and engineering*, outperforming other domains, such as *Chemistry* and *Agro-biotechnology*. The authors' explanation is the low level of development of the Italian industry, however this finding suggests that this domain is more likely to encourage a closer interaction between both sectors.

In terms of quality and impact, the publications of the SP movement have the quality to appear in leading journals and may have a slightly higher impact than the national average (not significant), being consistent with the higher quality (Cockburn & Henderson, 1998) and impact (Marston, 2011) of private research in biomedicine, for example. Thus, the observed quality and impact on the different fields do not seem to be related to each other, even though a journal's prestige is the most important factor for future impact in some science and technology areas (Bornmann & Daniel, 2007). The evidence suggests that the degree of impact, is determined by the public or private origin of the research. Hence, the regions with a greater public research base, such as the East of England and Scotland, have a higher impact on the research community, while those with a higher rate of private research, such as the South East, have less impact. In support of this, the output related to research oriented infrastructures and organisations (e.g. Research- campuses and Parks, and RIs) draws greater interest from the scientific community. This difference is also apparently linked to the applied nature of the research conducted by the private sector, and which has less scientific impact (Godin, 1996). This finding also reflects the distance between basic and applied research, as it is widely considered as one of the main interaction barriers between the public and private sectors (Bruneel, D'Este, & Salter, 2010). Thus, despite the private sector tending to establish collaborations with research leaders; they tend not to be able to publish their publication in top quality journals (Abramo et al., 2009), however this fact is partially contradicted as the on-park research in general have a significant higher quality. For this reason, the use of citations as a proxy to assess the quality of private research may not be suitable, as the diverse objectives of both communities from research differ in terms of intellectual and reputational goals, undermining to some extent the interest of private research in the actions of the scientific community.

Conclusions

This study draws on bibliographic data from at least one on-park organisation in the UK with the aim of expanding the knowledge of the SP movement as a whole. In particular, the focus has been on; (1) identifying the research areas that attract most of the on-park research and the contribution of the geographic regions and U-I collaboration across the different areas, and (2) finding out whether the

quality and impact of the research production associated with SPs have a greater quality and impact than the average research.

In answer to the first goal, the findings reveal that the R&D activities are frequently generated in four subject areas: *Biochemistry, Genetics and Molecular Biology*, *Chemistry*, *Medicine*, and *Agricultural and Biological Sciences*, and the mass of research accumulated in the three top regions are characterised by; (1) public science-based research specialised in food-biotechnology and bio-pharmacology in the East of England, (2) private science-based research specialised in physics and material engineering in the South East, and (3) public science-based research specialised in the agro-biotech sector in Scotland. *Pharmacology*, *Engineering*, and *Materials Science* are the areas with the highest rate of private participation. The synergy expected within SPs is again questioned here as it is found that inter-institutional collaboration is only limited to a quarter of the output, of which more than half are U-I collaborations. The domain with the highest U-I interaction is private research-oriented physics and material engineering, while the rate of knowledge transference from the other two main domains seems to be punished for their high reliance on on-park RIs and then, their different approach to get involved into the research and dissemination process.

In answer to the second goal, the findings regarding the quality and impact of the output, reveal that in general on-park organisations publish in significantly higher quality journals, and that the research has similar impact to the national average. The relationship between quality and impact varies for the same research area, especially among the set of areas related to the three top domains and regions. A closer look at the impact produced by the regions, infrastructures, and organisations reveals that the closer the output is to HEIs and RIs the greater the impact, while the closer the output is to firms the lower the impact. This is a sign of the interaction barriers between the public and private sectors that are usually caused by the focus on either basic or applied research, which is also illustrated by the limited impact of the private research on the scientific community.

In conclusion, this study provides evidence that research impact is likely to be associated with the nature of the organisation producing the research rather than its relation to a physical intermediary infrastructure. The low level of interest in private research from the scientific community suggests that citation-based indicators may not be the best tools to assess the private research community and especially the academic research organisations, such as, schools, departments and RIs, which have built up strong links with industry. Furthermore, that important aspects, such as geographically high concentrations of on-park research activities, low U-I collaboration rates, and limited integration into the research community, question the idea of SPs as the catalysts behind a knowledge-based development across regions, and policy tools intended to support the transition from declining to innovative industries as a way of reducing the unequal distribution of research-intensive industry across the UK. Thus, this evidence is helpful for policy makers

in assessing the actual impact of policies and in guiding the directions of a more effective and realistic transfer policy for SPs and U-I collaboration in general.
An important limitation is that the results here are only indicative because although the main goal of SPs is to facilitate R&T transfer, formal research dissemination only uncovers part of this transference, and not all U-I interactions result in (co-authored) articles (Katz & Martin, 1997). Another important limitation is that it might not cover all the research generated within the SP movement due to the fact that not all on-park organisations mention the name of the infrastructures where they are based as part of their affiliation address. In addition, the rapid increase of the output over recent years can generate bias against part of the publications as they have less time to be cited. Similarly, the results could also favour the visibility of some research intensive industrial sectors where publications are more important. Finally, the identification of the research community associated with the SP movement allows qualitative studies that should disclose interesting insights into the real impact of support infrastructures on effective knowledge transfer. The central role of most RIs in supporting local industries makes it necessary to map their research performance and links with the private sector. There are also other interesting aspects of on-park research output which suggest that the development of the UK SP movement is characterised by a constant increase in the research production from the 90s with exponential growth since 2000. On the other hand, the coverage gap found in the WoS database suggests that the sources where industry in general is able to publish and interact with the wide scientific community might be less likely to be indexed in the WoS. It is therefore necessary to empirically examine the bias of this database against private research.

Acknowledgments

The authors would like to thank Robert Tijssen for reading part of the manuscript and for his valuable comments and suggestions, and also the three anonymous referees that helped to clarify the arguments of the paper.

References

Abramo, G., D'Angelo, C., Di Costa, F., & Solazzi, M. (2009). University–industry collaboration in Italy: A bibliometric examination. *Technovation*, *29*(6-7), 498–507.

Arvanitis, S., Kubli, U., & Woerter, M. (2008). University-industry knowledge and technology transfer in Switzerland: What university scientists think about co-operation with private enterprises. *Research Policy*, *37*(10), 1865–1883.

Bigliardi, B., Dormio, A., Nosella, A., & Petroni, G. (2006). Assessing science parks' performances: directions from selected Italian case studies. *Technovation*, *26*(4), 489–505.

Birch, K. (2009). The knowledge-space dynamic in the UK bioeconomy. *Area*, *41*(3), 273–284.

Bornmann, L., & Daniel, H. (2007). Multiple Publication on a Single Research Study : Does It Pay ? The Influence of Number of Research Articles on Total Citation Counts in Biomedicine. *Journal of the American Society for Information Science*, *58*(2000), 1100–1107. doi:10.1002/asi

Bruneel, J., D'Este, P., & Salter, A. (2010). Investigating the factors that diminish the barriers to university–industry collaboration. *Research Policy*, *39*(7), 858–868.

Cockburn, I. M., & Henderson, R. M. (1998). Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. *The Journal of Industrial Economics*, *46*(2), 157–182.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and Impacts: The Influence of Public Research on Industrial R&D. *Management Science*, *48*(1), 1–23. doi:10.1287/mnsc.48.1.1.14273

Cooke, P. (2001). Biotechnology Clusters in the UK: Lessons from Localisation in the Commercialisation of Science. *Small Business Economics*, 43–59.

Dyson, J. (2010). *Ingenious Britain: making the UK the leading high tech exporter in Europe*. London.

D'Este, P., & Patel, P. (2007). University–industry linkages in the UK: What are the factors underlying the variety of interactions with industry? *Research Policy*, *36*(9), 1295–1313.

Fukugawa, N. (2006). Science parks in Japan and their value-added contributions to new technology-based firms. *International Journal of Industrial Organization*, *24*(2), 381–400.

Godin, B. (1996). Research and the practice of publication in industries. *Research Policy*, *25*(4), 587–606.

Hauser, H. (2010). *The Current and Future Role of Technology and Innovation Centres in the UK* (p. 29). London.

Hayati, Z., & Ebrahimy, S. (2009). Correlation between quality and quantity in scientific production: A case study of Iranian organizations from 1997 to 2006. *Scientometrics*, *80*(3), 625–636.

Hu, M.-C. (2011). Evolution of knowledge creation and diffusion: the revisit of Taiwan's Hsinchu Science Park. *Scientometrics*, *88*(3), 949–977.

Hung, W. C. (2012). Measuring the use of public research in firm R&D in the Hsinchu Science Park. *Scientometrics*, (106). doi:10.1007/s11192-012-0726-5

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18.

Lambert, R. (2003). *Lambert Review of Business-University Collaboration*. Norwich: H. M. Treasury.

Lee, J., & Win, H. N. (2004). Technology transfer between university research centers and industry in Singapore. *Technovation*, *24*(5), 433–442. doi:10.1016/S0166-4972(02)00101-3

Link, A. N., & Scott, J. T. (2003). U.S. science parks: the diffusion of an innovation and its effects on the academic missions of universities. *International Journal of Industrial Organization*, *21*(9), 1323–1356.

Link, A. N., & Scott, J. T. (2007). The economics of university research parks. *Oxford Review of Economic Policy*, *23*(4), 661–674.
López-Illescas, C., Moya-Anegón, F., & Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of Informetrics*, *2*(4), 304–316.
Marston, L. (2011). *All together now: Improving cross-sector collaboration in the UK biomedical industry*. *NESTA report London*. London.
Meho, L., & Rogers, Y. (2008). Citation counting, citation ranking, and h-• index of human-• computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, *59*(11), 1711–1726.
Minguillo, D. (2010). Toward a new way of mapping scientific fields: Authors' competence for publishing in scholarly journals. *Journal of the American Society for Information Science and Technology*, *61*(4), 772–786. doi:10.1002/asi.21282
Minguillo, D., & Thelwall, M. (2011). The entrepreneurial role of the University: a link analysis of York Science Park. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of the ISSI 2001 Conference - 13th International Conference of the International Society for Scientometrics & Informetrics, Durban, South Africa, July 4-8* (pp. 570–583). South Africa.
Minguillo, D., & Thelwall, M. (2012). Mapping the network structure of science parks: An exploratory study of cross-sectoral interactions reflected on the web. *Aslib Proceedings*, *64*(5).
Noyons, E. C. M., Moed, H. F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science*, *50*(2), 115–131.
OECD. (2002). *Frascati Manual: Proposed Standard Practice for Surveys on Research and Experimental Development*. *Frascati Manual: Proposed Standard Practice for* (p. 252). Paris: OECD.
Quintas, P., Wield, D., & Massey, D. (1992). Academic-industry links and innovation: questioning the science park model. *Technovation*, *12*(3), 161–175.
Rafols, I., Hopkins, M. M., Hoekman, J., Siepel, J., O'Hare, A., Perianes-Rodríguez, A., & Nightingale, P. (2012). Big Pharma, little science? *Technological Forecasting and Social Change*.
Rothaermel, F., & Thursby, M. (2005). Incubator firm failure or graduation?The role of university linkages. *Research Policy*, *34*(7), 1076–1090.
Schwartz, M., & Hornych, C. (2010). Cooperation patterns of incubator firms and the impact of incubator specialization: Empirical evidence from Germany. *Technovation*, *30*(9-10), 485–495.
Siegel, D., Westhead, P., & Wright, M. (2003). Assessing the impact of university science parks on research productivity: exploratory firm-level evidence from the United Kingdom. *International Journal of Industrial Organization*, *21*(9), 1357–1369.

Suvinen, N., Konttinen, J., & Nieminen, M. (2010). How Necessary are Intermediary Organizations in the Commercialization of Research? *European Planning Studies*, *18*(9), 1365–1389.
Teixeira, A. a. C., & Mota, L. (2012). A bibliometric portrait of the evolution, scientific roots and influence of the literature on university–industry links. *Scientometrics*, *93*(3), 719–743.
Tijssen, R. J. W. (2004). Is the commercialisation of scientific research affecting the production of public knowledge? Global trends in the output of corporate research articles. *Research Policy*, *33*(5), 709–733. doi:10.1016/j.respol.2003.11.002
UKSPA. (2012). United Kingdom Science Park Association. Retrieved December 12, 2012, from www.ukspa.org.uk/
Vedovello, C. (1997). Science parks and university-industry interaction: Geographical proximity between the agents as a driving force. *Technovation*, *17*(9), 491–531.
Westhead, P., & Storey, D. J. (1995). Links between higher education institutions and high technology firms. *Omega*, *23*(4), 345–360.

INFLUENCE OF UNIVERSITY MERGERS AND THE NORWEGIAN PERFORMANCE INDICATOR ON OVERALL DANISH CITATION IMPACT 2000-12

Peter Ingwersen[1;2], Birger Larsen[1],

[1] *{pi; blar}@iva.dk*
Royal School of Information and Library Science, University of Copenhagen, Birketinget 6, DK 2300 Copenhagen S (Denmark)

[2] Oslo University College, St. Olavs Plass, 0130 Oslo (Norway)

Abstract
This paper analyses the patterns of Danish research productivity, citation impact and (inter)national collaboration across document types 2000-2012, prior to and after 1) the university mergers in 2006 and 2) the introduction of the Norwegian publication point-based performance indicator 2008/09. Document types analysed are: research articles; conference proceedings papers excluding meeting abstracts; and review articles. The Web of Science citation index (WoS) combined with the Danish Research & Innovation Agency's basic statistics is used for data collection and analyses. Findings demonstrate that the overall productivity and citation impact steadily increases over the entire period, regardless the university fusions and the introduction of the performance indicator. The collaboration ratio between purely Danish and internationally cooperated research articles remains stable during the period while that of proceedings papers decline. The number of countries with which Denmark collaborate increases for all publication types during recent years in line with citation impact of international cooperation. Simultaneously, the citation impact for conference proceedings papers as such remains substantially the same over the period except for a drop from 2010; their productivity declines slightly since 2009. The ratio between proceedings papers and research articles starts declining from 2009 in WoS corresponding to actual developments observed in the point-based performance indicator itself. Since 2009 the WoS coverage of proceedings papers is declining. The positive growth in research articles derives primarily from the Natural Sciences and Technology published in prestigious Level 2 journals. The introduction of the publication performance model, rather than the university mergers, is regarded the accelerator of these processes in recent years.

Conference Topics
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability; Visualisation and Science Mapping: Tools, Methods and Applications

Introduction
The rationality behind mergers of smaller university units and research centres into fewer but larger universities at a national scale is commonly of economic and

management nature. In Denmark the university merger exercise was finalized in 2006. Simultaneously the new universities obtained a quasi-autonomous status with a board appointed by government and became reorganized in larger research units rather than in departmental entities. Democratically elected heads of units, university directors and deans now belong to an earlier age. From a political (governmental) perspective the idea of *New Public Management* applied to the science sector is to benefit from expected scientific synergies, higher research productivity and quality and innovation, increased private-public sector collaboration, faster through-put of students, increased bureaucratic control, and management streamlining. Owing to increased monitoring of research outcomes and administrative regulation of research funding distribution, however, the segment of the administrative staff in universities including university hospitals is growing, not declining (Danmarks Statistik, 2012).

Commonly monitoring of institutional and national productivity and citation impact are based on peer reviewed journal articles (van Raan, 1999; 2005; Moed, 2005). Since the Norwegian performance indicator system also takes into account proceedings papers, albeit hitherto assigning less scoring points to the latter, we have included this document type as well in the present investigation. Earlier studies of possible influence of institutional mergers have not demonstrated substantial effects on research quality or decrease in bureaucratization. For instance Kyvik (2002, p. 53) discussed "[the] merger of 98 vocationally-oriented colleges into 26 state colleges in Norway. The mergers, which took place in 1994, [had] in many ways proved to be a successful reform. The colleges now have more competent administration and professional leadership, and they have become far more visible and acquired a higher status. Still, several of the aims of the reform – to improve teaching and research and to make the colleges more cost-effective – can so far not be said to have been fulfilled. In addition, many academic staff feels that the new colleges have become bureaucratized, that the identity of the individual vocational programs have been weakened, and they blame the reform for a general retrenchment in financial resources." In the Danish case one may argue that the university mergers took place at the top levels of the institutions, whereas the scientific staff carried on as usual continuing their projects and collaboration. Since not many research groups have been split or made redundant after the mergers one might indeed argue that they may have had a positive effect on research productivity and quality.

As part of the research monitoring measures the Norwegian performance model based on assigned publication points was introduced in 2009 into the Danish academic landscape (Schneider, 2009). The starting point was to establish 67 groups of researchers from the Danish universities to list and assign points to peer reviewed journals that published scientific material authored by Danish academics 2008. The performance indicator takes into account published peer reviewed research and review articles, monographs, anthology papers and proceedings

papers. In the publication period 2008-2011 proceedings papers were assigned fewer points (.7) than journal articles (1.0 in Level 1 journals and 3.0 in Level 2 journals, i.e. the leading journals of a field as judged by the relevant researcher group and covering maximum 20 % of the field journal output). From 2012 proceedings papers receives similar points as articles, depending on the level of the conference, as assessed by the relevant group. For each document the points are fractionalized according to the collaborating universities and institutions; then cumulated per institution. Also from 2012 the model encourages collaboration by multiplying the fraction obtained (min. 0.1) by 1.25. Each of the 67 groups represents an academic field or specialty. Since 2009 the past year's research output has been assigned points annually that are used to distribute a portion of public research funding among the universities the following year. Only the cumulated results are publicly available per university and major academic area, such as the Humanities or Health Sciences (Forskningsstyrelsen, 2013); the intermediate or more detailed publication point distributions and document lists per unit and department are not publicly accessible. This is in difference to Norway where no multiplication of fraction takes place and all the documents and their point assignments are transparent as well as publicly accessible through an open access database (Sivertsen, 2010). In Belgium the Flemish BOF-key applies whole counting at the institutional level (Debackere & Glänzel, 2004; Engels, Ossenblok & Spruyt, 2012).

With respect to the publication performance indicator a major underlying idea was to encourage publishing in so-called 'Level 2' journals when implemented in Norway (Aagaard & Schneider, 2012). This has been studied in Norway and results demonstrate a substantial increase of 55 % 2005-09 for publications in Level 2 journals (Sivertsen, 2010; Sivertsen & Schneider, 2012). The Belgian experience for the social sciences and humanities is analysed by Ossenblok, Engels and Sivertsen (2012). The influence of peer reviewed conference papers on citation performance has not been studied extensively (Butler & Visser, 2006) – and then mostly in relation to particular fields like computer science (He & Guan, 2008; Wainer et al., 2011). They have not been studied at all in relation to performance indicator models like the Danish/Norwegian one based on publication points.

The present analysis investigates the patterns of research productivity and citation impact, as indication of research quality across document types, prior to and after 1) the university mergers in 2006 and 2) the introduction of the Norwegian assessment system 2008/09. Due to the change of and adaptation to the novel management and institutional structures within the new university units a certain stand-still in productivity immediately following the fusions might be expected, because not all the involved institutions were fusion-ready or would have preferred other constellations than the ones enforced by the government. One might also expect a decrease in institutional collaboration after 2009 at

international as well national levels owing to the fractionalisation principles in the assessment system, in particular from 2009-2012, prior to the introduction of the multiplication factor 2012. By some (science and engineering) universities fractionalisation was seen to penalize international collaboration by the research communities. From the perspective of Humanities the entire measurement system was regarded as an attack on the freedom of research and many critical opinions have been posted on academic blogs (e.g. http://professorvaelde.blogspot.com; http://www.forskeren.dk). From the government perspective the hopes were to reinforce an increase of the overall Danish research production and citation impact owing to better research quality caused by the mergers and encouraged by the performance system.

Motivated by the aforementioned conjectures the present investigation has the following three research questions:

1. Did the merger of universities 2006 alter the productivity and/or citation impact for Danish academic research, including research and review articles and proceedings papers (but excluding the humanities and monographs) in the following years, compared to the period immediately prior to the merger?
2. Did the university mergers influence the patterns of (inter)national collaboration?
3. Did the introduction of the Norwegian performance indicator for research publications in 2009 alter the Danish productivity patterns, citation impact or (inter)national collaboration in the following years?

It is important to stress that in Denmark the public funding of universities and research has not declined as a result of the economic crisis from 2008. It is fairly constant at a 0.9-1.1 % of the national BNP and its potential influence on productivity and research quality may be regarded as neutral.

From a methodological standpoint the investigation makes use of the Web of Science (WoS) citation indexes SCI, SSCI, CPCI-S and CPCI-SSH (Thomson-Reuters) as basis for the annual analyses and covers a period of 13 years: 2000-2012. Monographic material and the Humanities fields are not explicitly dealt with in the investigation owing to the language bias in WoS. However, some humanistic documents are involved by the application of CPCI-SSH. For comparative reasons the point-based performance indicator statistics 2009-12 are included since they demonstrate the real number of research documents published in Denmark (Forskningsstyrelsen, 2013).

The paper is organized as follows. Data collection procedures and analysis methods including three collaboration indicators are described. This is followed by three sections on findings. One section deals with the overall development of

productivity, citations to and impact of Danish research over the period across research articles, proceedings papers[85] and review articles. This is followed by a section on (inter)national cooperation across document types and citation impact developments. Analyses of the average number of collaborating countries and Danish research institutions across document types provide indications of publication behaviour that might have been influenced by the university mergers and introduction of the Norwegian performance indicator. The third section compares statistics from the development of the system to the WoS-based observations. Discussion and conclusion sections close the paper.

Methodology

The data collection was carried out in WoS on April 21, 2013 on Science Citation Index (SCI), Social Science Citation Index (SSCI), Conference Proceedings Citation Indexes for Science (CPCI-S) and Social Science and Humanities (CPCI-SSH). For each year the Danish share of WoS indexed materials was observed to detect any anomalies in database developments. Nothing particular was detected: the Danish world share remains rather constant at .80 % 2000-08; then it increases to almost 1.0 %. Research quality is measured in terms of citation impact. The citation window is kept at three years. This implies that 2010 is the last year with a workable three-year citation window (2010-2012). Citation and publication analyses are studied for each document type separately: research articles; review articles; proceedings papers. 'Other' types of documents that include meeting abstracts, editorials, book reviews, letters to editors, errata, etc. are taken into account but omitted from further analysis, which solely concerns the former three types. The WoS document category 'proceedings papers' is used to retrieve conference papers or contributions. It derives from the two CPCIs as well as from the original citation indexes (SCI and SSCI). In the latter case they are also commonly tagged by the category 'article'; but in the CPCIs there exists a partial overlap between the two document categories, which changes over time. Also over time, the two conference citation indexes display a great variety in coverage that actually declines since 2008. The discussion section includes an analysis of the WoS coverage of Danish and world proceedings papers in the CPCIs. In order to avoid the said overlap between the categories, foremost between research articles and proceedings papers, all documents indexed by both tags were kept as proceedings papers and thus excluded from the article category. Samples drawn from the overlap showed that such documents are indeed conference papers or contributions but published in a serial or thematic journal issues; thus the exclusion from the research article category.

Further, the ratio of proceedings papers vs. research articles is calculated per annum. These two publication types are regarded the channels that directly

[85] Proceedings papers include this WoS document category and exclude the category 'Meeting Abstracts'.

communicate scientific knowledge; review articles are seen as submissions that summarize already published knowledge. In relation to (inter)national cooperation the investigation operates with the following indicators:

1) *International cooperation ratio*, i.e., the ratio (between 0.0 and 1.0) of documents that are published in collaboration between Denmark and at least one other country. This ratio is calculated annually for research articles and proceedings papers separately. The number of collaborating countries constitute an additional sub-indicator;
2) *Average Number of countries per internationally collaborated document*;
3) *Average number of Danish institutions collaborating per document* within the set of purely national Danish publications for each document type.

In order to divide each annual set of research articles and proceedings papers into a purely national set of publications and a set of internationally authored documents for each type the analytic tools provided by WoS were applied to list, select and retrieve the documents from the collaborating countries to form a separate set of records, named the international cooperative set. The number of individual countries was detected in this set. The total number of documents containing at least one country was calculated by aggregating the number of documents assigned each country in the set. This aggregated number of documents was then divided by the number of documents in the international cooperative set to produce indicator (2).

The set of purely national Danish publications in a document type was retrieved by means of Boolean NOT logic of the international cooperative set on the initial set of that document type. The resulting purely Danish set was then analyzed by the Analyze Result tool of WoS for each document type with respect to the metadata category of 'Organizations Enhanced'. The total number of documents containing at least one institutional name was calculated by aggregating the number of documents assigned each 'Organization Enhanced' in the set. This aggregated number of documents was then divided by the number of documents in the national Danish set to produce indicator (3). It is important to stress that in this calculation name form control of institutions is not necessary. Since only one name form of each affiliated institution is commonly assigned each document, logic dictates that this calculation involving institutional names signifies the average number of *different* institutions collaborating per document. Thus, the analysis does not inform about the real number of different institutions that collaborate. Indicators (2) and (3) were calculated for the seven selected years 2001; 2003; 2006; 2008; 2010; 2011; and 2012. Citation impact for each document type divided into purely national and international collaborative sets was calculated for the five selected years 2001; 2003; 2006; 2008; and 2010.

In case of sets too large for WoS to handle when generating online citation reports, i.e. sets above 10,000 items, the set was logically divided into subsets according to the indicator (2) method above; later the analysis results were aggregated. The Danish research article sets from 2010 to present constitute such large sets (Table 1). In total the analyses deal with almost 171,000 source documents and more than 830,000 citations.

The annual statistics from the assessment system (Forskningsstyrelsen, 2012) was used to form new descriptive statistics dedicated the point assignments to the three document types as well as to the overall academic areas of Science & Technology, Social Sciences and the Health Sciences covering the period 2008-2011. The number of publications per academic area was included in the public agency statistics 2009-11. For 2008 the number of publications was estimated from the assigned points. The statistics cover more publications than indexed by WoS. Nevertheless, the trends can be compared between our findings through WoS and those observed by the agency.

Findings

Table 1 displays the annual number of Danish research publications indexed by WoS 2000-2012 including the three dominant document types, and the corresponding citation volumes. Diagrams 1-2 provide the corresponding citation impact development over the entire period.

The general trend for research articles, Table 1, is a steady increase of productivity over the entire period. For proceedings papers the years 2001, 2004 and 2006 display negative growth. The highest productivity is reached in 2007. From 2008 and onwards the productivity, according to WoS indexing, is declining fast. For review articles three years 2001, 2007 and 2010 demonstrate negative growth. The major type of documents in the document category 'Other types' consists of 'Meeting abstracts' throughout the period.

Table 1. Annual Danish research publications and citations 2000-2012 with three year citation windows in (parenthesis)(WoS, April 2013)

	2000 (2000-02)		*2001 (2001-03)*		*2002 (2002-04)*		*2003 (2003-05)*		*2004 (2004-06)*		*2005 (2005-07)*	
Document types	Publ.	Citations	Publ.	Citations	Publ.	Citations	Publ.	Citations	Publ.	Citations	Publ.	Citations
Research articles	6712	40687	6891	42950	6772	44136	7079	51972	7383	53582	7588	58928
Proc. Papers	1596	3208	1479	3255	1489	3116	1696	3539	1595	3887	1777	5146
Review articles	341	4458	304	4696	357	3838	351	5805	446	6817	472	7291
Other types:	1523	1272	1262	927	1645	953	1620	1022	2042	1093	2412	1507
Total types:	10172	49625	9936	51828	10263	52043	10746	62338	11466	65379	12249	72872
Online:	10172		9936		10263		10746		11466		12249	

	2006 (2006-08)		*2007 (2007-09)*		*2008 (2008-10)*		*2009 (2009-11)*		*2010 (2010-12)*		*2011*	*2012*
Document types	Publ.	Citations	Publ.	Citations	Publ.	Citations	Publ.	Citations	Publ.	Citations	Publ.	Publ.
Research articles	7988	60655	8532	69693	9178	78130	9836	83365	10921	98017	12391	13423
Proc. Papers	1660	4634	1852	4745	1538	4553	1485	4407	1233	3054	915	610
Review articles	558	8656	545	9605	651	12846	735	13648	695	11691	787	827
Other types:	2432	1568	2755	2023	2760	1910	2976	2236	2800	2937	2755	3185
Total types:	12638	75513	13684	86066	14127	97439	15032	103656	15649	115699	16848	18045
Online:	12638		13684		14127		15032		15649		16848	18045

Diagram 1 shows the annual ratio of proceedings papers vs. research articles to the left, for WoS covering the entire period and according to the Danish Research Agency for 2008-2011 (Forskningsstyrelsen, 2012), and the cumulated 2-year citation impact for research and review articles combined (a kind of journal impact) as well as proceedings papers separately to the right. The WoS ratio illustrates the same trend as shown for the productivity, Table 1, with decline from 2008 in WoS. The Agency statistics also demonstrate a similar negative trend from 2009. While the citation impact is fairly constant at 2.0 and growing to 2.96 in 2008-09 for proceedings papers with a significant drop in 2010 to 2.48, the impact for journal-based publications (Res.art.+Rev.art.) is constantly increasing including 2010 reaching an impact score of 9.44.

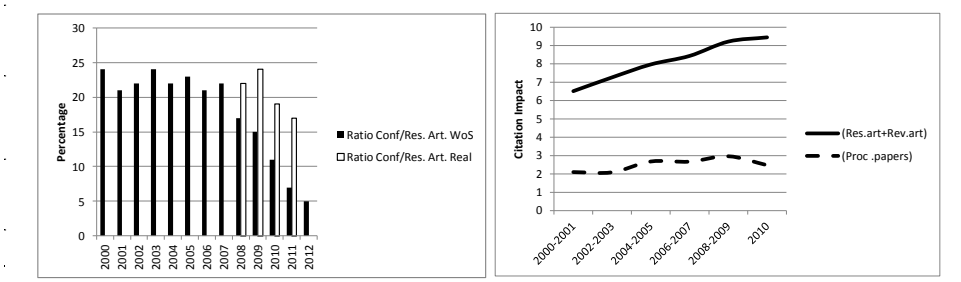


Diagram 1. Annual ratios in percentage of Danish proceedings papers *vs.* research articles, from WoS and Danish Research Agency 2008-11 (left); 2-year citation impact development for research articles and review articles combined vs. proceedings papers (right)(WoS, April 2013)

Diagram 2 (left) demonstrates the detailed annual trends for the different document categories. One observes a drastic drop in impact for review articles in 2010 to the 2003-06 level; however, the Danish *research articles* constantly increase their citation impact score including 2010, thus compensating the Danish average citation impact that is constantly rising during the entire analysis period.

(Inter)national cooperation, document types and citation impact

Diagram 2 (right) demonstrates the citation impact obtained by the research articles and proceedings papers published by Danish institutions only or authored in international collaboration with other nations. The impact of the *research articles* made in international collaboration is continuously substantially higher (almost the double) than that received by purely Danish publications, the latter staying level from 2008. In addition, the international cooperative research articles demonstrate a steady impact growth. Notably, the overall trend for the research articles initiated prior to the university mergers 2006 does not change after the fusion (Diagram 2, left); the increase simply continues regardless the mergers and the introduction of the Norwegian performance indicator system 2008.

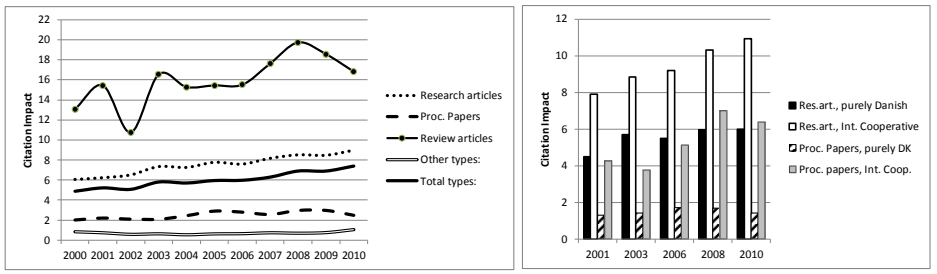


Diagram 2. Annual development of Danish citation impact to publications 2000-2010 with three-year citation window (left). Citation impact in five selected years for research articles and proceedings papers, purely Danish vs. international cooperation (right)(WoS, April 2013).

In contrast, the drop 2010 in citation impact for the Danish *proceedings papers*, Diagram 2 (left), derives from a marked decline in the impact received by the international Proceedings publications that year – as well as from the purely Danish Proceedings papers. The latter set of documents starts losing impact already in 2006 (right), simultaneously with the university mergers.

Table 2. Development of international cooperation, number of cooperating countries and purely Danish authorship across document types during seven selected years (WoS, April 2013)

Research Articles	***2001***	***2003***	***2006***	***2008***	***2010***	***2011***	***2012***
Purely Danish authorship	3375	3396	3453	3821	4342	4969	5192
Int. Coop. Authorship	3516	3683	4535	5357	6579	7422	8231
Total no. of documents	6891	7079	7988	9178	10921	12391	13423
Number of countries	103	120	127	125	137	135	152
Proceedings Papers							
Purely Danish authorship	1031	1201	1131	1168	968	787	521
Int. Coop. Authorship	448	495	529	370	265	128	89
Total no. of documents	1479	1696	1660	1538	1233	915	610
Number of countries	67	65	57	71	85	49	71

For research articles the total *number of unique countries* with which Denmark is collaborating increases steadily over the seven selected years, Table 2: from 103 countries in 2001 to 152 countries in 2012. At the same time the number of countries for *proceedings papers* reaches a peak in 2010; it drops heavily in 2011 but raise again in 2012 – coinciding with the introduction of the multiplication factor for cooperation in the performance indicator system. This drop also coincides with the decline for proceedings paper productivity, shown in Table 1

above. Table 2 demonstrates that already from 2008 a decrease initiates *primarily* among the internationally collaborative papers according to WoS indexing, going from 529 to 370 items. From 2010 also the volume of Danish authorship proceedings papers diminishes.

Diagram 3 displays the *international cooperation ratio* (indicator 1), the *average number of countries* collaborating including Denmark in the Danish/international research publications (indicator 2) and the *average number of Danish institutions* collaborating per document within the set of purely Danish publications (indicator 3) for research articles (left) and proceedings papers (right).

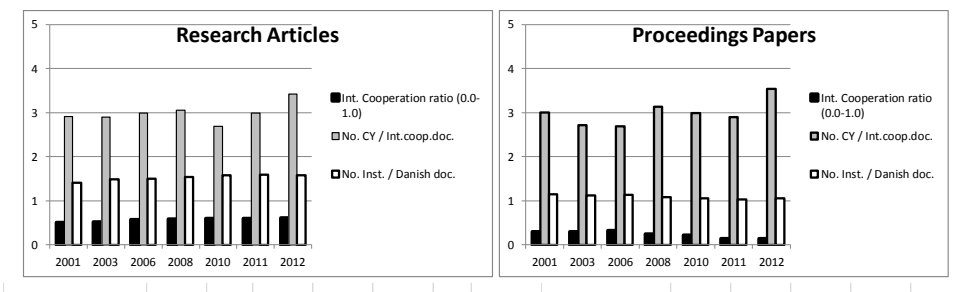


Diagram 3. International cooperation ratio (0.0 – 1.0), average number of countries collaborating in Danish publications and mean number of Danish institutions collaborating per purely Danish publications. Research articles (left); Proceedings papers (right)(WoS, April 2013)

One observes, Diagram 3, that the *international cooperation ratio* according to the WoS indexing is stable for research articles during the period (left) whilst declining for the proceedings papers since 2006 (right). For both document types USA constitutes the dominating partner for Danish research institutions and its share does indeed increase from 14 % in 2001 to almost 18 % in 2012 for the research articles and centres around 4.5 % for proceedings papers (figures not shown in tables/diagrams).

For research articles the *average number of countries* cooperating with Denmark declines in 2010 but increases steadily since then. For proceedings papers a decline starts in 2008, turning into a positive trend from 2012. The performance indicator system may have had a negative effect at its introduction, which has turned positive in recent years, probably affected by the introduced multiplication factor for cooperation: The general trend for both types is a small overall increase in indicator 2 during the entire period.

Indicator 3 (Danish collaborating institutions per document), Diagram 3, demonstrates constant stable scores for research articles (left) but a slight decrease

in the average number of institutions collaborating within the purely Danish proceedings paper space (right).

Comparative statistics of actual publications 2008-11and WoS trends

Diagram 4 demonstrates extracts from the publication statistics published by the Danish Research & Innovation Agency (Forskningsstyrelsen, 2012) for the publication years 2008-2011 associated with the performance indicator scores. The figures for 2008 are estimated since only the indicator scores (in points) are available not the underlying publication volumes. From 2009 the number of publications is provided by the Agency, in addition to the distribution of performance points across document types and the four central academic areas: Natural Sciences & Technology; Social Sciences; Health Sciences; and Humanities. Only the three former areas are dealt with in Diagram 4.

In particular, Denmark is highly productive with respect to *Level 2 articles* (the most leading publication vehicles); their growth is primarily caused by a 36 % increase in articles made in the *Natural Sciences & Technology* area over the three years 2009-11, that is, since the introduction of the performance indicator system. For the *Health Sciences* the growth is only 6.4 % and for the Social Science area 14.1 % during the same time. For *Level 1 articles* the steady growth over the period is equally caused by all three areas, each with an increase of approx. 18 %.

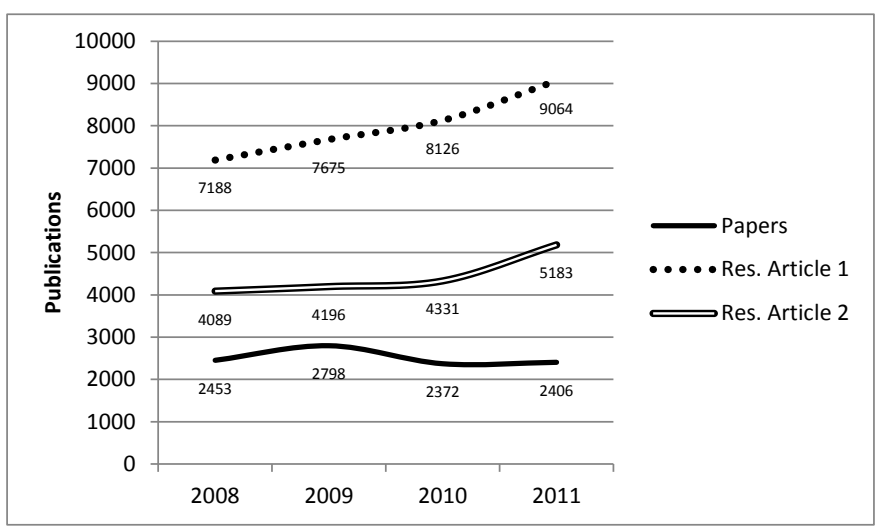


Diagram 4. The development of journal articles, Levels 1 and 2, and papers published in proceedings and anthologies; scores from 2008 are estimated (from Forskningsstyrelsen, 2012).

For actual (anthology and Proceedings) papers the trend, Diagram 4, is slightly negative from 2009 with the Health Sciences as the dominant area in decline (-62 %, although for a small population) and the Social Sciences with -12 %. This decline coincide with the similar decline observed in the WoS indexing space for the same period, Tables 1-2 and Diagram 1.

Discussion

Research questions one and two

In research question 1 we asked if the mergers of universities 2006 did alter the productivity and/or citation impact for Danish academic research (excluding the humanities and monographs) in the following years, compared to the period immediately prior to the mergers?

The answer is probably no. The mergers do not seem to influence the already active and positive developments in *research article production* and *impact*, Table 1 and Diagrams 1-2. They simply continue linearly regardless the events. This is in line with the ealier findings by Kyvik (2002). However, indeed we observe a negative productivity *and* impact development of actual proceedings papers, but first from 2009 and continuing into 2012, Diagrams 1 and 4. Similar trends for the productivity are visible for this document type according to WoS as well as observed by the Research Agency. The productivity decline seems in particular to take place in the Social Sciences (-12 %) and the Health Sciences (-62 %). Findings suggest that the decline in citation impact is caused by both purely Danish and the internationally collaborative proceedings papers (Diagram 2, right), yet mostly by the latter set. Similarly, the ratios of proceedings papers vs. research articles decline from 2009, with respect to WoS indexing *and* according to the research Agency statistics, Diagram 1, left. This negative trend is also observed 2009-12 with respect to the international cooperation ratio and number of Danish institutions collaborating on research in WoS, whereas the number of countries in cooperation with Denmark in the proceedings papers has increased from 2012 after a steady decline, Diagram 3, right. The publication performance indicator seems to function as the central *accelerator* rather than the university mergers for the development of this document type. The university mergers and the new management structures in the larger university units seem not to be influential on the research outcome; the publication and research quality development seems rather unaffected.

With respect to review articles the developments are rather variable across the period; it is hence not definitive to state that the university mergers (or the assessment system) are causing the recent impact drop from 2009 for this document type. The quality of the review articles are simply not recognized at the same high level as in the years 2007-08.

Research question 3

Initially we speculated that the fractionalization in the performance indicator might have a penalizing effect on the collaboration pattern. However, our findings, Diagram 3, left, do *not support* this idea for the research articles. On the contrary, the international cooperation ratio as well as the mean number of Danish institutions in cooperation per article is entirely stable according to WoS indexing; and the average number of countries per article does actually *increase* from 2009. For proceedings papers, as outlined above, the international publication behaviour is more negative. Probably the lower scores assigned this type of publication and the fractionalization method applied 2008-2011 by the publication performance system has discouraged some from publishing in proceedings papers.

In contrast the publication performance system seems positively to have encouraged and thus *affected* researchers positively to publish research articles, in particular through *Level 2 journals*, which are assigned higher scores, Diagram 4. The growth of this particular type of research articles over the three years 2009-11 is 24 %, with the Science & Technology fields showing a growth of 36 %, the Social Sciences 14 % and the vast Health Sciences 6.4 %. This growth is almost in line with that found for the six years 2005-09 in Norway at 55 % (Sivertsen, 2010; Sivertsen & Schneider, 2012), also after the introduction of their version of the performance system.

Methodological issues associated with WoS coverage

The entire set of Danish publications assigned performance points by the Danish Research Agency is logically containing the WoS-defined set analysed in the present study. One may consequently assume that the overall trends observed in the WoS-defined set mirror the trends in the agency-defined Danish set; for a comparison, see for instance Diagram 1, left. Diagram 5 demonstrates the growth patterns in the CPCI-S and SSH combined and the equivalent share of Danish proceedings papers 2000-12 across 7 data points (from Table 1). The diagram shows that 1) the two conference citation indexes decrease dramatically their coverage of that document type since 2006 and 2) similar (negative) growth trends occur for both segments. By knowing the real number of Danish publications, Diagram 4, this implies that the real number of proceedings papers published is far from being indexed in those two indexes, but that certain indicators, such as impact and international collaboration ratios and trends probably are valid. The proceedings paper/research article ratio scores are thus not realistic – although their trend pattern may very well be (Diagram 1, left). By comparing the productivity obtained from WoS with that provided by the Danish Research Agency (Forskningsstyrelsen, 2012), Diagram 4, one observes, for instance, that for 2011 the Agency stipulate the publication of 14,247 journal articles out of which WoS covers 13,178 (Table 1) with a coverage of 92 %. For

proceedings papers in WoS vs. anthology plus conference papers given by the Agency, the coverage is only 38 %.

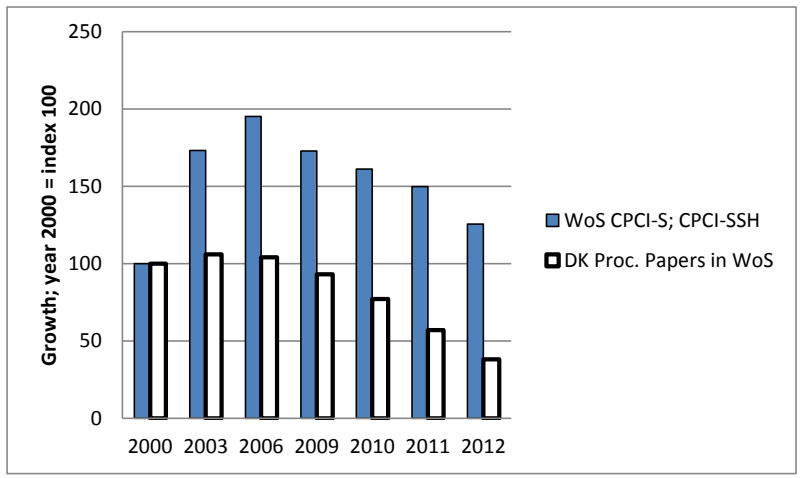


Diagram 5. Growth of CPCI-S, CPCI-SSH 2000-2012 and Danish proceedings papers (WoS, April 2013)

We may thus infer that the proceedings paper decline observed in the WoS set, although less pronounced in reality, with a high probability takes place in the Social Sciences (-12 %) and the Health Sciences (-62 %), and that the substantial growth of research articles, and continued positive impact development detected in WoS, with great certainty primarily is caused by high productivity and growth of *Level 2 articles* (Diagram 4), in particular published by the Natural Sciences & Technology fields..

Conclusions

The publication behaviour regarding *research articles* seems not influenced at all by the university mergers in 2006 but probably positively affected by the introduction of the publication performance indicator in 2009 with respect to publishing in leading (level 2) journals. The overall positive trends of steady publication and citation impact growth already in progress from 2001 have continued linearly, regardless these events. From a research political perspective this is acknowledgeable. So far the resources spend on the re-organizing of the Danish university system, on streamlining the administrative infrastructures and on re-shaping the research foci can only be seen to provide extra trade-offs regarding the growth of Level 2 research articles during recent years, in particular published by the Natural Science & Technology fields. Indeed, this may be positively influenced by the performance indicator rather than by the mergers. The slight drop in the productivity of proceedings papers initiated 2009 according to the Research Agency, and the decrease in the international collaboration ratio

as well as in the number of Danish research institutions cooperating derive from the Social and Health Sciences and is with some probability caused by the performance indicator system's fractionalization mode.

Finally, it is evident that the introduction 2009 of the publication performance indicator, which assigns points to the published peer reviewed publications, thus far has not introduced a 'salami-tactics´ in the production behaviour in the Danish science system and a consequential decline in citation impact, as witnessed in Australia in connection with other but more simplistic point-based assessment systems (Butler, 2003; 2004).

References

Aagaard, K. & Schneider, J.W. (2012). Den danske bibliometriske model; En dårlig kopi af den norske. *Forskningspolitikk*, March, 2012 (1), 28-29.

Butler, L. (2003). Explaining Australia's increased share of ISI publications—The effects of a funding formula based on publication counts. *Research Policy*, 32(1), 143-155.

Butler, L. (2004). What happens when funding is linked to publication counts? In: Moed, H.F., Glänzel, W; Schmoch, U. (Eds.), *Handbook of Quantitative Science and Technology*, Kluwer Academic Publishers, 389-340.

Butler, L. & Visser, M.S. (2006). Extending citation analysis to non-source items. *Scientometrics*, 66(10), 327-343.

Danmarks Statistik (2012). University administrative staff developments 2007-10. (In Danish) Available at: http://www.statistikbanken.dk/statbank5a/selectvarval/saveselections.asp

Debackere, K. & Glänzel, W. (2004). Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key. *Scientometrics*, 59, 253-276.

Elleby, A. & Ingwersen, P. (2010). Publication point indicators: A comparative case study of two publication point systems and citation impact in an interdisciplinary context. *Journal of Informetrics*, 4, 512-523.

Engels, T.C.E., Ossenblok, T.L.B. & Spruyt, E.H. J. (2012). Changing publication patterns in the social sciences and humanities, 2000–2009. *Scientometrics*, 93, 373-390.

Forskningsstyrelsen (2013); the Danish Research & Innovation Agency; (in Danish) available at: http://www.fi.dk/viden-og-politik/tal-og-analyser/den-bibliometriske-forskningsindikator

He, Y. & Guan, J.C. (2008). Contribution of Chinese publications in computer science: A case study on LNCS. *Scientometrics*, 75(28), 519-534.

Katz, J., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, 40(3), 541-554.

Katz, J., & Martin, B. (1997). What is research collaboration? *Research Policy*, 26(1), 1-18.

Kyvik, S. (2002). The merger of non-university colleges in Norway. *Higher Education*, 44(1), 53-72.
Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.
http://www.forskeren.dk. Visited April 22, 2013.
http://professorvaelde.blogspot.com. Visited April 22, 2013.
Ossenblok, T.L.B., Engels, T.C.E. & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science. A comparison of publication patterns and incentive structures in Flanders and Norway (2005-2009). *Research Evaluation,* 21(4), 280-290.
Schneider, J. W. (2009). An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway. *European Political Science*, 8(3), 364–378.
Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. *ISSI Newsletter*, 6(19), 22-28.
Sivertsen, G. & Schneider J.W. (2012). *Evaluering av den bibliometriske forskningsindikator*. Oslo: NIFU. 75 p. (Rapport 17/2012, in Norwegian)
van Raan, A.F.J. (1999). Advanced bibliometric methods for evaluation of universities. *Scientometrics*, 45(3), 417-423.
van Raan, A.F.J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*(1), 133–143.
Wainer, J., de Oliveira, H.P. & Anido, R. (2011). Patterns of bibliographic references in the ACM published papers. *Information Processing & Management*, 47(13), 135-142.

INFORMATION AND LIBRARY SCIENCE, CHANGES THAT INFLUENCED IT'S NEW CHARACTER, DIRECTION AND RESEARCH: A BIBLIOMETRIC STUDY, 1985-2006

Luba Gornstein[1] and Bluma C. Peritz[2]

[1] *lubag@savion.huji.ac.il*
The Hebrew University of Jerusalem, Library Authority (Israel)

[2] *bluer@cc.huji.ac.il*
The Hebrew University of Jerusalem (Israel)

Abstract
The present bibliometric study intended to characterize research literature in the field of Information and Library Science (LIS) during the years 1985-2006. The results revealed, that the field has undergone significant changes, mainly paradigmatically. The process paradigm prevalent until mid-1970's in research, was replaced by the user paradigm that, during the years under study, gained new perspective of the sociological aspect of information technology integration that is developing rapidly. The following methodologies were used for this study: content analysis, reference analysis, faceted classification. Content analysis was performed by using an updated taxonomy designed for this research. In this research we found that the field of LIS is influenced mainly by disciplines belonging to the social sciences. Through the years, the field tends to become more interdisciplinary. Theoretically, we surmise that this research contributes on a conceptual level to the process of self-knowledge of the field. Methodologies used in this study are not based on large research populations, but require intensive examination of the literature and evaluation of the contents of articles for the purpose of in depth subject analysis. Follow up studies are popular with many fields of the sciences like: sociology, education, economics, demography, medicine and others.

Conference Topic
Bibliometrics in Library and Information Science (Topic 14)

Introduction
The main goal of this study is to define by bibliometric methods the conceptual outlines of LIS as reflected by the topics covered by research during the period 1985-2006. The year 1985 appears to be a turning point, followed by a period of the application of technological advancements in libraries and information centers. Technological advancements, including the development of the internet and other innovative communication technologies led to significant changes in the field of Information and Library Science.

Throughout the development of human thought, philosophers, sociologists and historians of science have attempted to define the various fields of the sciences and the relationships between them. These definitions varied greatly from one period in history to another, in accordance with the leading streams of philosophy and their current concept of science. In earlier periods, most sciences, for example medicine, belonged to the humanities. In some periods, only the natural sciences were considered science, while the social sciences and humanities were not.

Library and Information Science (LIS) is a fairly new discipline and was declared an academic field only in the 1920s (Waples, 1931). During it's inception, paradigms from social sciences, mostly from sociology, were used. It developed rapidly as an interdisciplinary field, integrating topics from the humanities, social and behavioral studies and later from computer science. Research in LIS apply both quantitative and qualitative methods and cover a wide range of topics, including information retrieval from online databases and the internet, information needs and uses of various populations, development of online information systems, collection management, indexing and automatic abstracting, bibliometrics, scientometrics and webometrics, as well as historical studies.

Nonetheless, there is no definitive answer yet as to whether LIS is a field of science or a professional field. Most fields of the sciences are characterized by two essential components: a theoretical basis supported by an empirical research. In the LIS there is still difficulty in consolidating a theoretical basis, agreed upon among the researchers of the field (Cornelius, 2002; Hjorland, 2000).

Additionally, another problem relates to the question whether LIS is an interdisciplinary field that combines content and methodologies from other sciences and imports them to other fields and vice versa ("outside looking field"), or it's a self-developing, "inside-looking field". Many researchers have tried to answer this question, but have not yet been able to agree unequivocally (Harris, 1986; Houser, 1988; Peritz, 1977).

The present research is a follow-up of Peritz's (1977) long-term study which describes the research in LIS during the period 1950-1975. Peritz formulated fundamental definitions of library science, discerning between research and theory and enabled further research during the past thirty years.

Following Peritz's, other research was published on the field of LIS covering the period of the years 1975-1985 (Nour, 1985; Feehan et al, 1987; Atkins, 1988; Jarvelin & Vakkari, 1993, Koufogiannakis, Slater, & Crumley, 2004). Therefore, in the present research we refer to the period beginning in 1985. Regardless of the fact that other research defining Information and Library Science implement Peritz's methodologies, they are confined to a short period (some up to one year), to a particular country, a very specific topic, or based on limited sources.

Recently, a number of studies of the LIS as a discipline which aim to define the epistemic boards and interdisciplinary character of the LIS were published (Cronin & Meho, 2008; Astrom, 2010; Prebor, 2010; Milojevic, Sugimoto, Yan & Ding, 2011; Lariviere, Sugimoto and Cronin, 2012; Chang and Huang, 2012). These studies used methodologies which are different from the methodologies

applied in the present study, like content analysis based on taxonomy, which has to be created specifically for each one of the subjects under study, reference analysis and faceted classification, which seem to be the proper methodologies for in depth bibliometric research.
The present study is empirical, bibliometric in nature, incorporating various methodologies from different fields of the sciences, and is meant to be extensive for the period covered, the population studied and the sources used.

Objectives

In order to achieve the main goal of this research we must itemize it into several objectives, or operative questions:

1. Is the percentage of research literature growing in the period under study, relative to previous periods?
2. What are the most researched topics in the field during the period in question (1985-2006)?
3. What methodologies are used by researchers in the field?
4. Are advances in technology reflected in research today?
5. What are the characteristics of the citations (based on reference analysis)?
    a. Are there topics that are restricted to LIS (inside looking), as opposed to subjects that penetrate into other fields of the sciences, or, conversely, imported to LIS from other fields?
    b. What fields of the sciences contribute to LIS research in the period under study as opposed to previous periods (outside looking)?
6. Are the core journals in the field growing or changing with time?
7. Which countries contribute mostly to research in the field?

As mentioned above, the main goal of this research is to discover conceptual changes in the field. Thus, a broad historical and philosophical survey was required in order to present a paradigmatic perspective relative to previous researches, but taking into account the time and space restrictions this will be mentioned only in short.
In the historical survey, we found that the most significant paradigmatic change that occurred in the field was the transition from the paradigm of storage and content processing of books or other formats, to the paradigm of availability and access to content. As a result of this paradigmatic change, the emphasis of research in the field of LIS changed from study of the library objects or libraries as institutions, to the user study, in other words to the study of information needs to provide fast and easy access to information. The leading cause for this paradigmatic change is the enormous technological advancement that began in the late 1970's and continues to this day, as well as new philosophical approaches and research methods in science in general, and in LIS in particular, which started to be implemented.

Methodology

The research population is composed of a sampling of 1803 articles, which represents 30% of the total articles published in the leading journals in the field, between the years 1985-2006, in a 3 year leap. The sampling selection process was gradually checked. In the first phase, a list of the leading journals in the field was compiled. This list is based on journals listed in the Journal Citation Reports (JCR) database in the category Information Science and Library Science. In the period under study the JCR listed 56 journals in LIS. From this list, 29 titles were selected based on their high Impact Factor and their continuity, since the goal of the study was to identify research. In the years researched, in three year leaps, 5936 articles were found. Of these, the first two articles in each issue of each one of the journals were selected. These articles constitute the sampling. Other forms of literature that appear in these journals, such as literature reviews, editorials, book reviews, etc., were not included in the research population.

The following methodologies were applied for the analysis of the data: content analysis for classification of the research articles in the sampling, reference analysis of research articles and faceted classification.

Content analysis was performed by using the following taxonomy (Hawkins, Larson & Caton, 2003) which had to be elaborated and updated along the line, in order to be able to classify the papers including new topics and developments in the field. The validity of the taxonomy was ascertained by consulting an expert in classification. The taxonomy comprised of 13 main categories. During the collection of data, we found that this taxonomy is comprehensive and valid and matches all the articles in the sampling with the relevant topics. Classification of the articles by subjects was meant to describe the research literature and define the conceptual borders in the field of LIS. It was meant to discover the changes in the years researched as opposed to previous periods.

Classification of the articles by the methodologies used by researchers of the field, was meant to lead to conclusions on the characteristics of the study and to determine if they have changed in the years under study.

The purpose of the reference analysis accompanying the articles was to check the interdisciplinary level of the field and to determine which fields of the sciences most influence research in LIS.

The faceted classification method was used to correlate between the different findings during the data collection in order to get an overall picture of the status of research in the years researched.

Results

In this study we found that the field of LIS became more scientific in nature during the years 1985-2006 (see table 1).

In the period under study, there is a significant growth of research literature in the field. If one compares the years 1985-1994 to the years 1997-2006, 60.61% of the articles in the sampling in the first group are categorized as research articles, as opposed to 73.33% in the latter one. In other words, the trend is clear: when we

look at the literature that accompanies the field of Information and Library Science appearing in the core journals, we can conclude that this literature is becoming more scientific in nature. This conclusion is reinforced by other findings. A large concentration of the research articles was found in the journals which are associated with Information Systems (see table 2).

Table 1. Numbers and percentages of research articles in the sample

Period	*No. of research papers*	*No. of non-research papers*	*Total*	*% of research papers*	*% of non-research papers*
1985-2006	1234	569	1803	68.44%	31.56%
1985-1994	420	273	693	60.61%	39.39%
1997-2006	814	296	1110	73.33%	26.67%

Table 2. Numbers and percentages of research articles by group of journals which were used for sampling

Group of journals	*Number of papers in sample*	*No. of research papers 1985-2006*	*% of research papers 1985-2006*	*% of research papers 1985-1994*	*% of research papers 1997-2006*
Information systems journals (9)	276	225	81.52%		81.52%
Other journals (20)	1,525	1,009	66.16%	60.61%	70.79%

There is an ongoing discussion regarding the question if Information Systems journals are really belonging to the field of LIS as it classified by JCR. The in depth content analysis of the research articles in the sample of the present research strengthened the inclusion of nine Information Systems journals among the core list of JCR because they dealt mainly with the social aspect of implementation of information systems and not with the technological ones.
When we look at the characteristics of the authors of the articles in the sampling, we can see that most of them are written by researchers, whether they are research articles or not (see table 3).

Table 3.Distribution of the professional affiliation of most of the authors of the articles which are produced by more than one author

Research	*Most R**	*R*=P***	*Most P***	*Total*
Research papers n=728	80.22%	6.73%	13.05%	100.00%
Non-research papers n=186	55.91%	8.61%	35.48%	100.00%

*R- Researcher (affiliated with academic departments)
**P-Practitioner (affiliated with library positions)

The percent of articles where the first author is a researcher has grown through the years, in both research and non-research articles.

An overview of the empirical findings that we have related to until now enables us to conclude, that in the years researched, the field of Information and Library Science has become more scientific in nature, the level of the articles themselves, the level of the journals that are gaining importance, as well as the level of the authors involved.

As a result of classification of the research articles by subject, we found a paradigm change in LIS in the period researched: from a process-oriented research paradigm prevalent in the field to the mid 1970's, through the "user-oriented research" and "literature oriented research" paradigm (terminology of White and McCain, 1998) to the integration of information technology and electronic information services, that is based on the users' paradigm. Researchers of the field of Information and Library Science in the years researched are less concerned with aspects of the process paradigm. Some subject groups are often used as categories of main subjects, and others as categories of secondary subjects.

Table 4. Most frequent combinations of main and secondary subject categories, without sub-category division

Most frequent relations between main and secondary subjects	*Secondary subject*								
Main subject	**Use and user behavior**	**Subject-specific sources, applications and other research aspects**	**Types of institutions**	**Information technology**	**Information industry**	**Electronic information service**	**Knowledge organization**	**Information policy**	**Grand Total**
Information technology	**113**			15	**46**	11	15	11	211
Electronic information services	**42**	18		**37**		11	**23**		131
Bibliometrics, scientometrics and webometrics		**69**							69
Professional issues	10		**42**						52
Use and user behavior		17	11						28
Knowledge organization				10		17			27
Information industry			10						10
Grand Total	165	104	63	62	46	39	38	11	528

If we look at the relationship between main and secondary subject categories, we could see that the most frequent cross-reference was found between the information technology as a main subject and the use and user behavior as a secondary subject. This combination indicates the significant position of the

"user-oriented paradigm" in the development of Information and Library Science. The combination of the subject of bibliometrics, scientometrics and webometrics as a main subject and subject-specific topics as a secondary subject, supports the strong position of the "literature-oriented paradigm" in the field (see table 4).

Table 5. Most popular relationships between the main subject and the main methodology of the research articles

Relations between the main subject and the main research type	*Main research type*													
Main subject category	**Empirical**	**Theoretical**	**Operations**	**Bibliometric studies**	**Evaluative**	**Information retrieval**	**Content analysis**	**Information systems design**	**Case studies**	**Historical**	**Comparative**	**Network analysis**	**Other**	**Grand Total**
Information technology	105	58	49	3	36	13	12	26	36	7	10	5	10	370
Electronic information services	29	31	53	3	18	59	4	19	6		6	1	1	230
Bibliometrics, scientometrics and webometrics	6	16	23	86	3	1	17				4		1	157
Professional issues	53	8	21	4	7		4		3	6	3		4	113
Knowledge organization	11	37	7		12	3	6	6		5	3	2	1	93
Use and user behavior	59	13	1	1		4	4			1			1	84
Social issues	18	9	5	5	2		4		2	1	1			47
Theoretical aspects of library and information science	1	36	2				1			4	1			45
Information industry	14	8	7	2	2	2	1	2		3	1			42
Types of institutions	8	4	2		2		1		1	4	1		2	25
Historical aspects of library and information science, including history of libraries and history of the book.		1								13				14
Information policy	2	6			1	1			3		1			14
Grand Total	306	227	170	104	83	83	54	53	51	44	31	8	20	1234

Classification of the articles by methodologies points to the fact that the nature of the research in the field of LIS has not changed significantly in the period researched, compared to the preceding periods. The prevailing methodologies in the field are empirical, such as surveys and interviews. Methodologies such as operation research and historical research are losing popularity. In contrast, the methodology of case study is gaining popularity. These findings lead us to

conclude that the field of Information and Library Science, in regard to the nature of the research, is an applied science similar to education, social work, medicine etc. Looking at the most popular relationships between the main subjects of the research articles in the sampling and the main methodologies (using the faceted classification theory) enables us to conclude that, though Information Technology is a favorite subject of research, in LIS it's studied and researched by methodologies which are mostly associated with the social sciences (see table 5).

Table 6. General distribution of references from the research articles

Type of references	*1985-2006 N=38,671*	*1985-1994*	*1997-2006*
Inside references (within LIS)	94%	61%	49%
Outside references (to other fields)	15%	39%	51%
References to reference works	%5	1.1%	0.8%
References to electronic resources	%1	0.1%	8.9%

Table 7. Distribution of outside references related to the main subject categories

Main subject category	*No. of outside references*	*% of LS**	*% of SS***	*% of HUM****	*% of CS*****	*% of GEN******
Information technology	9,116	3.20%	**60.91%**	0.60%	27.05%	8.03%
Electronic information services	2,700	6.70%	12.00%	1.78%	**73.26%**	5.56%
Bibliometrics, sciento-metrics and webometrics	1,728	**22.69%**	**33.22%**	2.14%	6.77%	**35.19%**
Use and user behavior	1,535	11.73%	**61.63%**	4.04%	10.81%	11.79%
Knowledge organization	1,181	6.01%	**42.85%**	8.64%	34.46%	11.18%
Professional issues	780	1.41%	**75.26%**	3.08%	7.05%	13.21%
Theoretical aspects of library and information science	756	9.13%	**46.83%**	12.17%	15.87%	16.01%
Information industry	747	6.02%	**60.37%**	4.55%	15.93%	13.12%
Social issues	658	3.95%	**71.12%**	1.37%	8.66%	14.89%
Types of institutions	257	5.84%	**68.87%**	3.89%	5.45%	15.95%
Information policy	231	0.87%	**69.70%**	0.43%	8.66%	20.35%
Historical aspects of library and information science, including history of libraries and history of book.	202	0.00%	16.34%	**38.61%**	0.50%	**44.55%**
Total:	19,891	6.46%	50.95%	2.78%	27.75%	12.07%

**LS - Life Sciences*
***SS - Social Sciences*
****HUM - Humanities*
*****CS - Computer Science*
******GEN - General*

Following reference analysis, we found that LIS is increasingly becoming interdisciplinary through the years (see table 6), and is influenced mostly by disciplines belonging to the social sciences.
The use of the faceted classification method shows, that in various subject categories the interdisciplinary level varies as well.
Thus, subjects belonging to Information Technology are considerably influenced by the social sciences, e.g. management, sociology, psychology, etc. In contrast, the category of professional issues is characterized by a low interdisciplinary level. Moreover, the distribution of the outside references in relation to the main subjects of the research articles shows, that most of the subjects which stay in focus of research in the field of LIS are influenced by the social sciences (see table 7).
Although the results show that most of research in the field is concentrated in North America and Western Europe, the period under study also revealed that research in the field became more international, including other geographical areas (see table 8).

Table 8. Numbers and percentages of article produced by authors associated with the region

Non-collaborated authorship	1985-2006 n=1708	%	1985-1994 n=682	%	1997-2006 n=1026	%
North America	988	57.85%	411	60.26%	577	56.24%
Western Europe	498	29.16%	206	30.21%	292	28.46%
Asia	97	5.68%	22	3.23%	75	7.31%
Oceania	45	2.63%	12	1.76%	33	3.22%
Eastern Europe	31	1.81%	17	2.49%	14	1.36%
Near East	27	1.58%	8	1.17%	19	1.85%
South Africa	14	0.82%	6	0.88%	8	0.78%
Latin America	8	0.47%	0	0.00%	8	0.78%
Total	1708	100.00%	682	100.00%	1026	100.00%

Some 95% (1708) articles are produced by authors who came from the same geographical region and only 5% (91) articles are produced by authors who came from various geographical regions.

Conclusions

In summary, the present bibliometric research intended to characterize research literature in the field of Information and Library Science during the years 1985-2006 and revealed that the field has undergone significant changes, mainly paradigmatically. The process paradigm prevalent until mid-1970's in research was replaced by the user paradigm that during the years under study gained new perspective of the sociological aspect of information technology integration that is developing rapidly. In this research we found that the field of LIS is influenced

mainly by disciplines belonging to the social sciences. Through the years, this field tends to be more interdisciplinary that in the preceding years, but in certain subjects the research does not cross the boundaries of Information and Library Science. Most of the literature in the field comes from North America and Western Europe, but in recent years there is an increase in literature from other regions such as Asia and the Middle East.

Theoretically, we surmise that this research contributes on a conceptual level to the process of self-knowledge of the field. Methodologies used in such studies do not allow to process large research populations, but require intensive examination of the literature and evaluation of the contents of articles for the purpose of in depth subject analysis. In a recent survey of literature (Alexander, 2012) the importance of taxonomies, for the study of the sociology of science, is raised and was found "highly relevant to the information profession". Taxonomy of the research topics that was created for this research can be used for other research, since it covers most of the topics researchers in the field deal with.

In this research we accomplished our goal and have defined the conceptual and the epistemic borders of the field of Information and Library Science in the period studied and identified the important changes that occurred during those years. Rapid technological developments have already caused changes in the field. Therefore, research similar to this, should be conducted every decade in order to define the field also from the perspective of sociology, history and philosophy of science, as done in many disciplines, such as education, sociology, economics, demography, medicine and others.

References

Alexander, F. (2012). Assessing information taxonomies using epistemology and the sociology of science. *Journal of Documentation, 68*(5), 725-743.

Astrom, F. (2010). The visibility of Information Science and Library Science research in bibliometric mapping of the LIS field. *Library Quarterly, 80*(2), 143-159.

Atkins, S. E. (1988). Subject trends in Library and Information Science research, 1975-1984. *Library Trends, 36*(4), 633-658.

Chang, Y., & Huang, M. (2012). A study of the evolution of interdisciplinarity in Library and Information Science: Using three bibliometric methods. *Journal of the American Society For Information Science and Technology, 63*(1), 22-33.

Cornelius, I. (2002). Theorizing information for Information Science. *Annual Review of Information Science and Technology, 36*, 393-425.

Cronin, B. & Meho, L. I. (2008). The shifting balance of intellectual trade in information studies. *Journal of the American Society for Information Science and Technology, 59(4)*, 551-564.

Feehan, P. E., Gragg, W. L., Havener, W. M., & Kester, D. D. (1987). Library and Information Science research: An analysis of the 1984 journal literature. *Library & Information Science Research, 9*(3), 173-185.

Harris, M. H. (1986). The dialectic of defeat - Antimonies of research in Library and Information Sciences. *Library Trends, 34*(3), 515-531.
Hawkins, D. T., Larson, S. E. & Caton, B. Q. (2003). Information Science Abstracts: tracking the literature of Information Science. Part 2: a new taxonomy for Information Science. *Journal of the American Society for Information Science and Technology, 54 (8)*, 771-781.
Hjorland, B. (2000). Library and Information Science: practice, theory, and philosophical basis. *Information Processing & Management, 36*(3), 501-531.
Houser, L. (1988). A conceptual analysis of Information Science. *Library & Information Science Research, 10*(1), 3-34.
Jarvelin, K., & Vakkari, P. (1993). The evolution of Library and Information Science 1965-1985: A content analysis of journal articles. *Information Processing & Management, 29*(1), 129-144.
Koufogiannakis, D., Slater, L., & Crumley, E. (2004). A content analysis of librarianship research. *Journal of Information Science, 30*(3), 227-239.
Lariviere, V., Sugimoto, C., & Cronin, B. (2012). A bibliometric chronicling of Library and Information Science's first hundred years. *Journal of the American Society For Information Science and Technology, 63*(5), 997-1016.
Milojevic, C. R., Sugimoto, C. R., Yan, E. J. & Ding, Y. (2011). The cognitive structure of Library and Information Science: analysis of article title words. *Journal of the American Society for Information Science and Technology, 62(10)*, 1933-1953.
Nour, M. M. (1985). A quantitative analysis of the research articles published in core journals of 1980. *Library & Information Science Research, 7*(3), 261-273.
Peritz, B. (1977). *Research in Library Science as reflected in the core journals of the profession: a quantitative analysis(1950-1975).* University of California, Berkeley.
Prebor, G. (2010). Analysis of the interdisciplinary nature of Library and Information Science. *Journal of Librarianship and Information Science, 42*(4), 256-267.
Waples, D. (1931). The Graduate Library School at Chicago. *The Library Quarterly, 1*(1), 26-36.
White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of Information Science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327-355.

AN INFORMETRIC STUDY OF KNOWLEDGE FLOW AMONG SCIENTIFIC FIELDS (RIP)

Erjia Yan, Ying Ding[1], and Xiangnan Kong[2]

[1]eyan@indiana.edu, dingying@indiana.edu
School of Library and Information Science, Indiana University, Bloomington, Indiana, 47405 (USA)

[2]xkong4@uic.edu
Department of Computer Science, University of Illinois at Chicago, Chicago, Illinois (60607)

Abstract
The production and creation of knowledge is not dependent on any individual or isolated entity; instead, knowledge is diffused, exchanged, and circulated among various entities. Studying knowledge flow and transfer within and across different research areas can help us better understand science innovation and scientific collaboration. This work-in-progress paper presents a methodological framework to study knowledge flow, including a knowledge hierarchy, the construction of knowledge flow network, and measurements that can be used to study disciplinarities. Data set and preliminary results are also introduced.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6) and Modeling the Science System, Science Dynamics and Complex System Science (Topic 11)

Introduction
The production and creation of knowledge is not dependent on any individual or isolated entity; instead, knowledge is diffused, exchanged, and circulated among various entities. Knowledge flow, in the past twenty years, has become more inter-sectoral, more inter-organizational, more inter-disciplinary, and more international (Lewison, Rippon, & Wooding, 2005; Wagner & Leydesdorff, 2005; Autant-Bernard, Mairesse, & Massard, 2007; Ponds, Van Oort, & Frenken, 2007; Buter, Noyons, & Van Raan, 2010).

Similar to many important concepts in informetrics and scientometrics, the transfer of knowledge is an unobservable phenomenon (Jaffe, Trajtenberg, & Fogarty, 2000). As an alternative, researchers rely on proxies to measure the concepts of interest. The quantitative studies of knowledge flow usually use citations as the research instrument. Citations between scientific articles imply a knowledge flow from the cited entity to the citing entity (Jaffe, Trajtenberg, & Henderson, 1993; Van Leeuwen & Tijssen, 2000; Nomaler & Verspagen, 2008). Using the trading metaphor (Stigler, 1994; Lockett & McWilliams, 2005; Cronin

& Meho, 2008), knowledge flow has been explored as the intellectual trading among different disciplines.

The quantitative studies of interdisciplinarity were made available by researching on citation networks aggregated at the field level. Researchers usually choose a subset of representative journals or the full sets of journals from a field based upon certain classification schemas of journals, and then measure the extent to which the chosen field cites publications of other fields. Network-based indicators have also been proposed to measure how interdisciplinary different research fields are. Examples include entropy (Zhang et al., 2010), integration and specialization (Porter, Roessner, Cohen, & Perreault, 2006; Porter, Roessner, & Heberger, 2008), diversity and coherence (Rafols & Meyer, 2010), percentage of multi-assignation (Morillo, Bordons, & Gomez, 2003), and relative openness (Rinia et al., 2002).

Previous efforts on inter-sectoral, inter-organizational, and interdisciplinary knowledge flows laid sound theoretical and methodological foundations to the inquiry of knowledge flow studies. Nonetheless, most of these studies only involved a few disciplines as the research target, and consequently were not able to provide a holistic view of the developments and interactions of various scientific disciplines. This study is thus motivated to conduct a more comprehensive examination of scientific trading, and to obtain a bird's-eye view for the developments and interactions of various scientific disciplines.

Data and proposed methods

Knowledge hierarchy

The data were awarded by the Elsevier Bibliometrics Research Program[86]. The intermediary data file is a journal-to-journal citation network (matrix) for all indexed journals in Scopus with a two-year citation window; that is citations in year *t* to articles published in year *t*-2. Data on the following five pairs of cited year-citing year are therefore obtained: 1997/1999 (i.e., cited journals in 1997; citing journals in 1999), 2000/2002, 2003/2005, 2006/2008, and 2009/2011. The data statistics is shown in Table 1.

Table 1. Data statistics

Year	*Total number of citations*	*Increase (%)*
1997/1999	4,563,187	-
2000/2002	5,712,008	20.11%
2003/2005	7,418,729	23.01%
2006/2008	8,417,970	11.87%
2009/2011	9,463,845	11.05%

[86] http://ebrp.elsevier.com/index.asp

Scopus has a well-defined journal classification schema called All Science Classification Codes (ASJC). The schema is composed of minor subject areas, major subject areas, and top-level divisions. A journal is usually assigned into one or several minor subject areas. In total, there are around 300 minor subject areas. These subject areas are grouped into 27 major subject areas, and these major subject areas are further grouped into 4 top-level divisions: Life Sciences, Physical Science, Health Sciences, and Social Sciences & Humanities. This schema is referred to as knowledge hierarchy and is visualized it in Figure 1. In the proposed study, we will focus on the analysis of the 27 major subject areas.

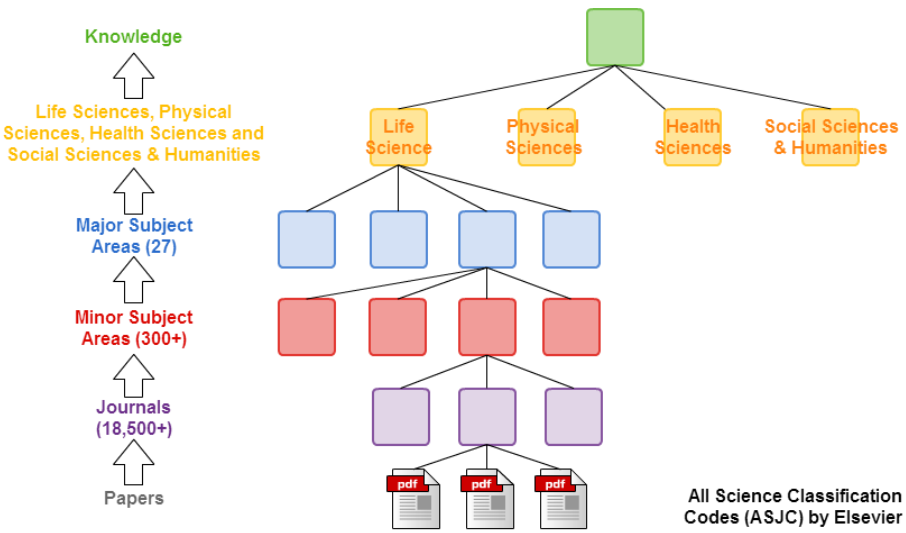


Figure 1. A six-layer knowledge hierarchy

As each journal is associated with one (or several) major subject area, a field-to-field citation matrix can be aggregated based on the journal-to-journal citation matrix (in this case, a field is a major subject area).

Measurements

For an effective evaluation, we propose several measurements, some of them are based on the concept of scientific trading (Yan, Ding, Cronin, & Leydesdorff, in press), such as self-dependence, knowledge exports/imports, trading dynamics, and trading impact. The weighted directed field-to-field citation network can be represented as $G=(V, A)$ where A represents the weighted directed link set and V represents the vertex set of subject areas.

- Self-dependence: $\text{Self\_dependence}_j = \frac{G_{jj}}{\sum_{i=1}^{n} G_{ij}}$, for any subject area j.

- Knowledge exports/imports: $\text{export/import}_{\text{k}} = \frac{\sum_{\text{i}=1}^{\text{n}} \text{G}_{\text{ik}}}{\sum_{\text{j}=1}^{\text{n}} \text{G}_{\text{kj}}}$, for any subject area k.
- Trading impact: $trading\_impact_k = \sum_{i=1}^{n} G_{ik}$, for any subject area k.
- Trading dynamics: $\text{trading\_dynamics}_{\text{k}} = \text{slope}(\text{trading\_impact}_{\text{k,t}}, \text{trading\_impact}_{\text{k,t+1}}, \dots)$, for any subject area k.

Self-dependence measures the extent to which an area depends on its own knowledge. Knowledge exports/imports measures whether a research area is a salient knowledge exporter or importer. Trading impact and dynamics quantify a research area's size of impact and its dynamics.

Other measurements are based on the concept of knowledge path, such as average shortest path length, average shortest path weight, and occurrence in shortest path. The proposed indicators are formally defined as:

- Shortest path (SP) from *i* to *j* ($\text{SP}_{\text{i}\to\text{j}}$) is a path from *i* to *j* in the knowledge flow network such that the sum of the distances of its constituent edges is minimized, where the distance is defined as $reverse\_flow\_width_{i\to j} = \frac{1}{number\ of\ citations\ from\ j\ to\ i}$.
- Shortest path length (SPL) from *i* to *j* ($SPL_{i\to j}$) is defined as the number of nodes traversed in transferring a piece of information in the shortest path from *i* to *j* ($SP_{i\to j}$).
- Average shortest path length (ASPL) for *i* as source of knowledge transfer is defined as: $ASPL_{i:source} = \frac{\sum_{j=1}^{n} SPL_{i\to j}}{n}$, where *n* is the number of subject areas in this study.
- Shortest path weight (SPW) from *i* to *j* ($SPW_{i\to j}$) is defined as the accumulative distances of pairs of nodes in the shortest path from *i* to *j* ($SP_{i\to j}$) where the distance is defined in formula (1).
- Average shortest path weight (ASPW) for *i* as source of knowledge transfer is defined as: $ASPW_{i:source} = \frac{\sum_{j=1}^{n} SPW_{i\to j}}{n}$.
- Occurrence in shortest path (OiSP) for *k* is defined as the number of times *k* occurred in shortest paths between (all potential) pair of nodes: $\sum_{i=1}^{n}\sum_{j=1}^{n}(k\ is\ on\ the\ shortest\ path\ of\ SP_{i\to j}?, 1, 0)$

For each subject area, the average shortest path length (as source of knowledge flow) measures how easily its knowledge can be accessed by subject areas. The average shortest path weight (as source of knowledge flow) measures how remote/different a subject area is from other subject areas. The occurrence in shortest path measures how important a subject area is to others' knowledge flow.

Preliminary results

We report the trading impact and trading dynamics of the 27 major subject areas in Figure 2. Y-axis shows the trading impact, i.e., the amount of incoming citations and the x-axis shows the trading dybamics.

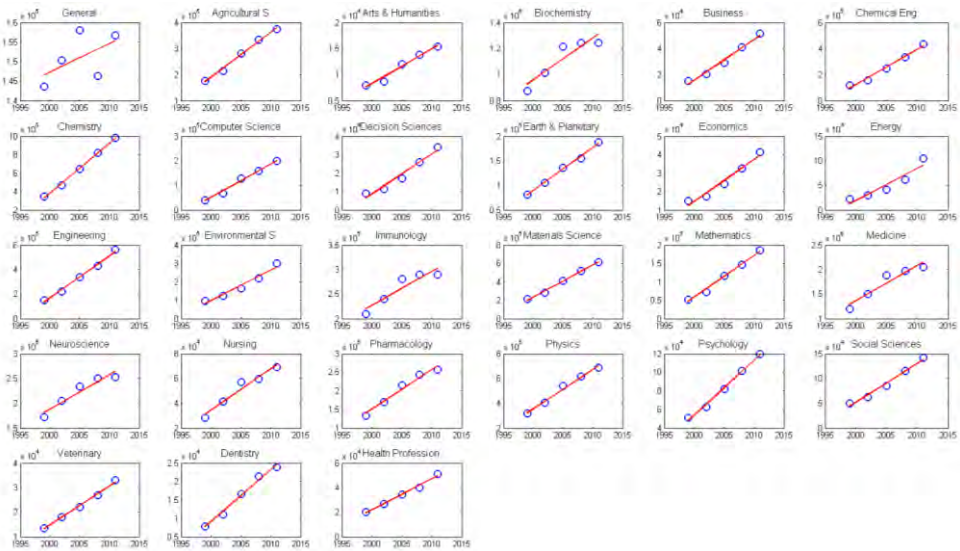


Figure 2. Trading dynamics of 27 major subject areas

Medicine, biochemistry, chemistry, material science, and physics have the highest trading impact, reflecting their dominant relations of economic, social, and political power in society, according to Lenoir (1997)'s theory on disciplinarity. In regards to trading dynamics, all subject areas have received an increased trading impact in the past decade (from 1997 to 2011). Areas such as energy, computer science, decision science, chemical engineering, engineering, and business have received higher increment rates as their slopes are steeper. Since the scientific trading impact of an area tells us whether its domain knowledge is recognized and valued (Merton, 1968; Cronin, 1984), the results suggest that these areas are becoming more visible and valued by other research areas.

The following map shows the critical knowledge paths for the 2011 data. The wider the path, the more frequent they occur in the shortest path. The map layout is based on Map Equation[87].

Different from previous maps of science that mostly are occurrence-based, Figure 3 is a directed map showing critical knowledge paths. Medicine, Chemistry, Social Sciences, Biochemistry, and Physics and astronomy form the backbone of knowledge path facilitating the dissemination of disciplinary knowledge among all other domains.

[87] http://www.mapequation.org/

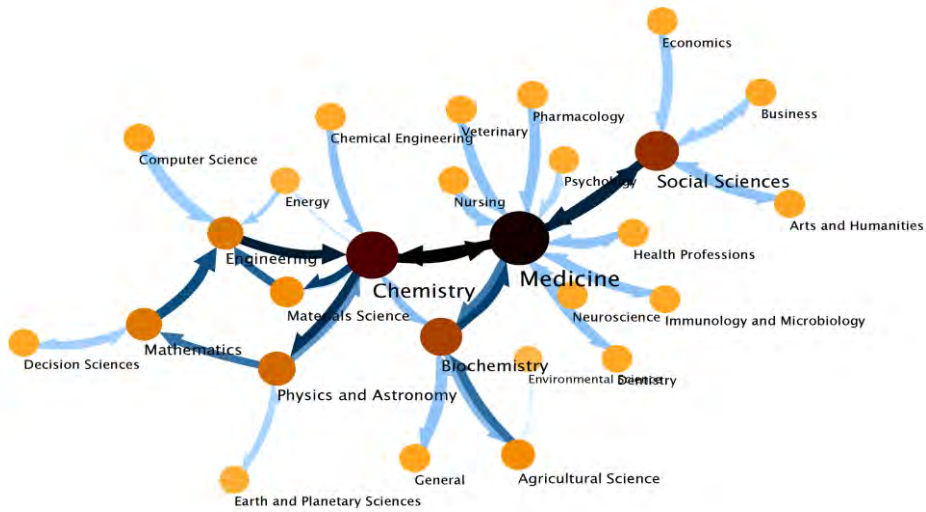


Figure 3. Critical knowledge paths

Future work

In the proposed study, we aim to study the dissemination of scientific knowledge through concepts of scientific trading and knowledge path. The preliminary results show promising findings that inform us on the patterns of knowledge transfer and dissemination. The proposed measurements quantify patterns of knowledge flow and dissemination, providing additional insights into interdisciplinary studies. These measurements are also valuable for scientific evaluation and science policy making. For ongoing studies on this project, we plan to apply these measurements to the data set and further our understanding on disciplinarity and knowledge dissemination.

Acknowledgement

This project is supported by the Elsevier Bibliometric Research Program (EBRP): http://ebrp.elsevier.com/

References

Autant-Bernard, C., Mairesse, J., & Massard, N. (2007). Spatial knowledge diffusion through collaborative networks. *Papers in Regional Science, 86*(3), 341-350.

Buter, R. K., Noyons, E. C. M., & Van Raan, A. F. J. (2010). Identification of converging research areas using publication and citation data. *Research Evaluation, 19*(1), 19-27.

Cronin, B. (1984). The citation process. The role and significance of citations in scientific communication. London: Taylor Graham.

Cronin, B., & Meho, L. I. (2008). The shifting balance of intellectual trade in information studies. *Journal of the American Society for Information Science & Technology, 59*(4), 551-564.
Jaffe, A. B., Trajtenberg, M., & Fogarty, M. S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review, 90*(2), 215-218.
Jaffe, A. B., Trajtenberg, M., & Henderson, A. D. (1993). Geographical localization of knowledge spillovers by patent citations. *Quarterly Journal of Economics, 108*(3), 577-599.
Lenoir, T. (1997). *Instituting science: the cultural production of scientific disciplines*. Stanford, CA: Stanford University Press.
Lewison, G., Rippon, I., & Wooding, S. (2005). Tracking knowledge diffusion through citations. *Research Evaluation, 14*(1), 5-14.
Lockett, A., & McWilliams, A. (2005). The balance of trade between disciplines: do we effectively manage knowledge? *Journal of Management Inquiry, 14*(2), 139-150.
Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56-63.
Ponds, R., Van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science, 86*(3), 423-443.
Porter, A. L., Roessner, J. D., Cohen, A. S., & Perreault, M. (2006). Interdisciplinary research: meaning, metrics and nurture. *Research Evaluation, 15*(3), 187-195.
Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82*(2), 263-287.
Rinia, E. J., Van Leeuwen, T. N., Bruins, E. E. W., Van Vuren, H. G., & Van Raan, A. F. J. (2002). Measuring knowledge transfer between fields of science. *Scientometrics, 54*(3), 347-362.
Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science, 9*(1), 94-108.
Van Leeuwen, T., & Tijssen, R. (2000). Interdisciplinary dynamics of modern science: analysis of cross-disciplinary citation flows. *Research Evaluation, 9*(3), 183-187.
Wagner, C. S., & Leydesdorff, L. (2009). Network structure, self-organization and the growth of international collaboration in science. *Research Policy, 34*(10), 1608-1618.
Yan, E., Ding, Y., Cronin, B., & Leydesdorff, L. (forthcoming). A bird's-eye view of scientific trading: Dependency relations among fields of science. *Journal of Informetrics*.
Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics, 4*(2), 185-193.

INTERACTIVE OVERLAYS OF JOURNALS AND THE MEASUREMENT OF INTERDISCIPLINARITY

Loet Leydesdorff,[1] Ismael Rafols,[2] & Chaomei Chen[3]

[1] *loet@leydesdorff.net*
Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam (The Netherlands)

[2] *i.rafols@sussex.ac.uk*
SPRU (Science and Technology Policy Research), University of Sussex, Falmer, Brighton, East Sussex BN1 9QE, (United Kingdom); *Ingenio* (CSIC-UPV), Universitat Politècnica de València, València, (Spain)

[3] *Chaomei.Chen@cis.drexel.edu*
College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, (USA)

Abstract
Document sets downloaded from the Web of Science can be projected onto global journal maps based on all journals contained in the Journal Citation Reports (JCR) of the Science and Social Science Citation Indices (2011). The disciplinary diversity of a downloaded set is then measured in terms of this map using Rao-Stirling's "quadratic entropy." Since this indicator of interdisciplinarity is normalized between zero and one, the interdisciplinarity of document sets can be compared among one another and across years, both cited and citing. The colors used for the overlays are based on Blondel *et al.*'s (2008) community-finding algorithms operating on the 10,000+ journals included in JCRs. The results can be exported from VOSViewer with different options such as proportional labels, heat maps, or cluster density maps. The maps can also be web-started and/or animated (e.g., using PowerPoint). The "citing" dimension of the aggregated journal-journal citation matrix was found to provide a more comprehensive description than the matrix based on the cited archive.

Conference Topic
Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8); Research Fronts and Emerging Issues (Topic 4).

Introduction
The technique of using overlay maps was introduced into science mapping by Boyack and collaborators in unpublished studies in the mid-2000s and elaborated into interactive overlays at the Internet by Rafols *et al.* (2010). The latter study used Web-of-Science (WoS) Subject Categories that are attributed to journals by professional indexers and semi-automatically by computer programs on the basis of a criteria such as the content, the title, and the citation patterns of journals

(Bensman & Leydesdorff, 2010; Pudovkin & Garfield, 2002: 1113n.). The categories, however, are overlapping and imprecise (Boyack *et al*., 2005; Rafols & Leydesdorff, 2009; cf. Rafols *et al*., 2010: 1887).

Visualization of the entire set 10,000+ journals was previously not possible because of computer capacities and the unsolved problem of the cluttering of the many labels in the layout. The capacity to print or display labels on a single page or screen is limited. Using more than seventy to one hundred labels, the representation of a network as a map can easily become too crowded as the labels begin to overlap and clutter. Both VOSViewer[88] and Gephi[89] have solved this problem by offering the possibility to foreground certain labels (those with a high value of a given node-attribute) more than others. In Gephi, the label size can be set proportionally to the size of the attribute. The downside of this proportional sizing is that labels of specialist journals can become so tiny that they cannot be read without zooming in (Leydesdorff, Hammarfelt, and Salah, 2010). In VOSViewer, the labels of nodes with small values of the network attribute (e.g., degree centrality) are faded for the sake of readability. However, one can zoom in and then these labels again become readable, or the user can move the cursor to a journal with the mouse, and then bring an otherwise suppressed label to the fore. For our purpose, this functionality is optimal: it solves the problem of visualizing large datasets in cases where the labels contain essential information (in our case, the journal names). The labels are available, but hidden when not needed visually.

Unlike network visualization programs such as Pajek and Gephi, VOSViewer uses an MDS-like algorithm (Kruskall & Wish, 1978) to position the nodes instead of a forced-based spring layout (e.g., Kamada & Kawai, 1989; Fruchterman & Reingold, 1991). The latter algorithms operate to minimize the stress in the sum of individual relations in the graph, whereas MDS (and its derivates) minimizes stress in the *system* of relations under study in terms of the dimensions of the latent structure (Leydesdorff, in press). However, in this study we are interested precisely in the structural dimensions of the journal network at the systems level, and therefore the map of the multi-dimensional vector space (i.e., similarities among citation distributions) will be used instead of the network of individual relations (i.e., citations as valued ties between journals). We use the cosine as a non-parametric proximity measure between vectors.

Methods and data

The data was harvested from the Journal Citation Reports (JCR) 2011 in September 2012. First, the JCRs of the Science and Social Science Editions of this database were merged. On the basis of this data an aggregated journal-journal citation matrix of 10,675 journals was constructed.[90] Of the $10{,}675^2$ =

[88] VOSViewer is a program for network visualization freely available at http://www.vosviewer.com .

[89] Gephi is a freeware programs for network analysis and visualization freely available at https://gephi.org/users/download/ .

[90] The Science Edition 2011 contains 8,281 journals, and the Social Science Edition 2011 contains 2,943 journals. Of these journals, 549 are contained in both databases.

113,955,625 cells only 2,207,789 (= 1.94%) are filled with values larger than zero; the grand total of the matrix is 35,295,459 citations, or on average 15.99 per cell with a value larger than zero. The data was gathered from the "citing" side. In the SCI and SSCI, the long tails of low values are sometimes summed up on this (citing) side as "all others". This cutoff at the lower end varies in the JCR with the sizes of the tails. However, since the file contains also 1,226,364 cells (55.54% of the non-zero cells) with a value smaller than five, one can expect the remaining inaccuracy because of the data processing to be small.

The aggregated journal-journal citation matrix was transformed into a cosine-normalized similarity matrix both in the being-cited and the citing directions. Matrices can then be exported in formats that can be read by the various visualization programs. We use SPSS (v.19) for the cosine normalization and Pajek and UCINet for the data manipulation. As noted, VOSViewer is used for the visualizations.[91]

After normalization in terms of the citing patterns, cosine values were larger than zero for 65,349,785 cells (57.34% of N^2). With a threshold of cosine > 0.2, the similarity matrix can significantly be reduced to only 3,151,994 (off-diagonal) values larger than zero (2.77%). Of the 10,675 journals, 10,330 (96.8%) are nevertheless connected into the largest component. This largest component is used for the mapping. Visualization software uses largest components because isolated and non-related components cannot be positioned unambiguously with reference to the largest component.

We worked with a standard laptop with 8 GB internal memory under Windows 7, 64-bits. VOSViewer gave no error message for processing the largest component of 10,330 journals. The computation took approximately two hours, but one needs to generate the basemap only once since the coordinates can thereafter be saved and used again. Mutatis mutandis, the largest component in the cited direction was 10,256 (96.1%). In this case, three more journals were removed because they generated outlier points, distorting the representation in VOSViewer.

The abbreviation "VOS" in VOSViewer stands for "visualization of similarities." The algorithm used for this is akin to that of MDS: VOSViewer minimizes a stress function at the systems level (Van Eck et al., 2010; cf. Kruskall & Wish, 1978). Waltman et al. (2010) have further integrated a clustering algorithm into the program that operates on the basis of the same principles as the positioning of the nodes in the map. The cluster results are automatically colored into the map, but the colors of the clusters can be changed interactively.[92] Additionally, a representation of the map as a density or heat map is provided in VOSViewer.

Eleven clusters were generated in VOSViewer using the citing patterns and the default value for the modularization ($\gamma = 1$; Waltman et al., 2010). Leydesdorff and Rafols (2012) used this default solution, but the new version of the maps will

[91] Chen & Leydesdorff (in press) makes similar functionalities available in CiteSpace.

[92] The clustering algorithm operates with a parameter (γ) that can be changed interactively in order to generate more or fewer clusters in the solution.

be based on modular decomposition using Blondel et al.'s (2008) algorithm for the decomposition (in Pajek). This algorithm is more commonly used. Twelve clusters are then distinguished in the citing dimension, and 40 in the cited. Thus, the cited map is finer-grained than the citing one, whereas the citing one is more clearly structured (Figure 1). In a later section of the paper, we will discuss how the user can replace the classification and coloring with any other one—including the one provided by VOSViewer. For reasons of presentation, we also postpone the discussion about the measurement of interdisciplinarity using the overlay maps.

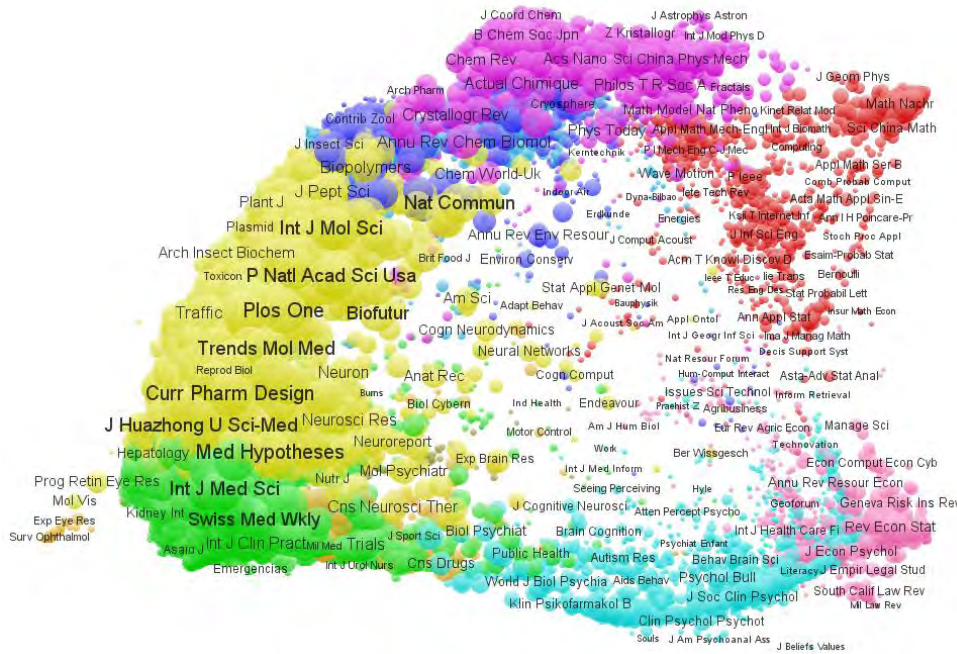


Figure 1: 10,330 journals similar in their citing patterns above cosine > 0.2; 12 colors (clusters; $Q = 0.575$).[93] This map can be viewed directly in VOSViewer via WebStart at http://www.vosviewer.com/vosviewer.php?map=http://www.leydesdorff.net/journals11/citing_all.txt&label_size=1.0&label_size_variation=0.3 .

Construction of the basemaps

Figure 1 provides the map based on the citing patterns (cosine > 0.2) and using Blondel *et al.*'s (2008) algorithm for the coloring of 12 communities. The resemblance to the maps based on WoS Subject Categories is striking (Leydesdorff, Carley, & Rafols, in press; Rafols *et al.*, 2010); and this croissant-like structure also accords with Klavans & Boyack's (2009) conclusion that a

[93] Using single-level refinement (in Pajek): $Q = 0.5747$; multi-level refinement $Q = 0.5750$. The number of clusters is 12 in either case (Blondel *et al.*, 2008).

consensus has increasingly emerged regarding the shape of journal maps based on aggregated citations.

The corresponding figure based on "being-cited" patterns (not shown here) is more compressed because the visibility of relatively isolated groupings in the border regions (that is, peninsulas of the large component) leaves less space for the central grouping. This second figure (in the cited dimension) can be web-started at http://www.vosviewer.com/vosviewer.php?map=http://www.leydesdorff.net/journals11/cited_all.txt&zoom_level=1.2&label_size=1.0&label_size_variation=0.3. As noted, 40 clusters are identified—and therefore differently colored—in this figure (Q = 0.529; Blondel *et al.,* 2008).[94] When one enlarges this picture (interactively or by including, for example, the command "zoom_level=1.2", as above), the borders are removed and the resulting picture is not so different from the one based on citing patterns.

The generation of overlay files

Two programs are made available online for generating overlays, at http://www.leydesdorff.net/journals11/citing.exe and http://www.leydesdorff.net/journals11/cited.exe, respectively. These routines can only process data downloaded from WoS in the so-called "tagged" format (that is, with labels like "AU " for authors, "TI " for titles, etc.). The user also needs the table files citing.dbf and/or cited.dbf, respectively, in the same folder; these table files can also be downloaded from http://www.leydesdorff.net/journals11. In addition to the coordinate information for the maps, the full titles of the journals as provided by JCR are listed in these files. Because there are differences in some cases between the abbreviations in JCR and the *Science Citation Index*, we use the full titles of the journals as keys for the matching. In the case of an unforeseen mismatch—for example, because a journal title was changed—this record will be skipped unless one edits (or duplicates) the title in the corresponding table file.[95]

When the programs and tables are brought into a single folder with the input file, which is downloaded from WoS and renamed "data.txt," an output file can be generated. This file is called either "cited.txt" or "citing.txt" depending on the routine in use. These files can be opened as so-called map-files by VOSViewer. Thereafter, all options commonly available in VOSViewer for the visualization can be used for improving the representations. The resulting figures can be exported as graphic files (.jpg, .png, etc.) or scalable vector graphics (.svg) that can further be edited in InkScape[96] of Adobe Illustrator™. Thus, in addition to the

[94] Using single-level refinement (in Pajek): Q = 0.5297; multi-level refinement Q = 0.5294. The number of clusters in 40 in the latter case (used here) and 41 in the former. In summary, Q = 0.57 in the citing and Q = 0.53 in the cited dimension: the citing matrix has a slightly higher modularity than the cited, which means that the citing behavior is slightly more organized than what is cited. This conclusion is consistent with the small number of 12 citing clusters compared to the larger number of 40 cited clusters.

[95] Table files in the .dbf format can be read into Excel, but are easier to save after making changes using the spreadsheet editors of OpenOffice or SPSS.

[96] InkScape is freeware available for download at http://inkscape.org.

label view (default), one can choose the density view or a heat map; other options make it possible to vary labels in size so that they can be made equally visible, or to change the colors of clusters, etc.

Our routines set the sizes of the nodes equal to the $\log_4(n + 1)$. The value of n is augmented by one in order to prevent the disappearance of a node in the case of a single publication (since $\log(1) = 0$). The base 4 for the logarithmic was chosen for pragmatic and esthetic reasons. Depending on the relative sizes, the user may wish to use a function other than the logarithm. The values of n, for example, can be retrieved from the table file named "overlay.dbf" which is generated at each run; this file can be read into Excel. By replacing the column labeled "normalized weight" in the map-file (cited.txt or citing.txt) with the values in the column *NPubl* in the file overlay.dbf, for example, one can obtain a map which exhibits a linear relation between the sizes of nodes and their respective publication volumes.[97]

In VOSviewer, one can choose between weighted sizes (normalized by dividing all weights by the average weight) or "normalized weight," that is, using the weights as already normalized by the user. Default output of our routines contains the label "normalized weight"; that is, the base-4 logarithm of the number of papers is used for the sizing of the nodes. This constant normalization enables the user to compare across overlays and to animate them for different years. By removing the word "normalized" from the header of the map-files of VOSViewer (i.e., by replacing the header with "weight"), however, the resulting figures can be esthetically optimized for each dataset independently using the normalization of VOSViewer. The user can change these column headings in the first lines of the files "citing.txt" and "cited.txt" after running the programs citing.exe or cited.exe, but before importing these (map) files into VOSViewer.

If one wishes to assume another classification as the *default* for generating overlays, one has to change the cluster indication in the column named "Blondel" in the tables cited.dbf and/or citing.dbf in this respect (in Excel or SPSS) and save these files thereafter again as .dbf tables with the same name.[98] These table files contain the clustering results of VOSViewer (for the default value of $\gamma = 1$; cf. Waltman *et al*., 2010) in the column headed "cluster" that can be copied to the column with the header "Blondel". Our programs use standardly the values provided in this latter column. The files citing.txt and cited.txt can also be edited, and then the last columns with cluster numbers that dictate the coloring and/or the labels can also be changed specifically using a text editor (or Excel); for example, if one wishes to highlight a specific group by using a different color or a marker.

[97] The files cited.txt or citing.txt are "comma-separated variable" files (.csv) that can be read and saved, for example, by using Excel.

[98] Exporting files in the .dbf format may not be easy in newer versions of Excel, but it is possible using the same spreadsheet in Open Office (or using other programs, including SPSS).

Measurement of interdisciplinarity

The maps enable us to propose an indicator (between zero and one) for the interdisciplinarity of any set downloaded from the Web of Science in terms of the set's distribution across the journals in terms of their distances on the map. These distances can be expressed as a percentage of the maximum distance, that is, the diagonal of the base map. The ratios are then weighted with the proportions of publications in each of the categories (that is, journals) using Rao-Stirling diversity (Δ). This measure is defined as follows:

$$\Delta = \sum_{ij} p_i p_j d_{ij} \quad (1)$$

where d_{ij} is a distance measure between two categories i and j, and p_i is the proportion of elements assigned to category i—that is, the relative frequency of each journal.

The Rao-Stirling diversity measure was introduced by Rao (1982a and b) and has also been named "quadratic entropy" (Izsák & Papp, 1995) because it measures not only diversity in terms of the spread of the elements among the categories of the classification, but also takes into account the distances among the categories (that is, in this case, among the journals on the map). Stirling (2007, at p. 712) proposed this measure as a general framework for measuring diversity in science, technology, and innovation. Porter *et al*. (2007) also used this measure in their integration score of interdisciplinarity.

Note that diversity can be considered as a specific—albeit common—operationalization of interdisciplinarity among other possible ones (cf. Barry *et al*., 2008; Klein, 1990; Wagner *et al*., 2011). For example, the concept of "interdisciplinarity" also contains the notion of "intermediation"—which can, for example, be operationalized using betweenness centrality (Leydesdorff, 2007; Leydesdorff & Rafols, 2011a)—and "coherence" (Rafols & Meyer, 2010).[99] Using betweenness centrality (an attribute to the nodes of the network), journals can be ranked in terms of their "interdisciplinarity," but Leydesdorff & Rafols (2011a, at p. 96) found different components between betweenness centrality and other (diversity-based) measures of interdisciplinarity (using factor analysis of the JCR 2008 as data). In summary, our measure in this study does *not* address the (inter)disciplinarity of journals measured in terms of, for example, betweenness centrality, but only the interdisciplinarity of *document sets* measured as Rao-Stirling diversity.

In our opinion, "interdisciplinarity" or its measures (such as diversity) should not be used without specification of the unit of analysis (cf. Wagner *et al*., 2011). In

[99] "Coherence" was operationalized by Leydesdorff & Rafols (2011b, at p. 856) as follows: $C = \sum_{ij(i \neq j)} p_{ij} \cdot d_{ij}$. Coherence C and diversity Δ (Eq. 1) can also be compared as observed versus expected values of interdisciplinarity in the set (cf. Rafols *et al*., 2012, at p. 1286), but has no interpretation in this map since journal names are unique attributes to papers.

this case, the measure applies only to the interdisciplinarity of downloaded document sets. In a next section, we will extend the options for generating overlays using the journal names in the cited references of these documents, and then specify this as the interdisciplinarity of their respective knowledge bases.[100] Analogously, one can ask for the "interdisciplinarity" of the sets that cite these documents ("the audience set"; Zitt & Small, 2008; cf. Carley & Porter, 2012). A publication set can be monodisciplinary (not diverse in the journals where it is published), but it can be cited interdisciplinarily (by diverse journals) or the other way round. Interdisciplinarity measured as Rao-Stirling diversity can be compared across sets and over time because the same basemaps are used for the normalization. We shall specify these possible extensions to cited references and citation patterns in a further section.

In this study, we use the distance on the map $\| x_i - x_j \|$ between each two journals participating in the set as the distance parameter d_{ij} in Eq. 1, as a proportion of the maximally possible distance (that is, the diagonal of the map). This distance measure is an optimization and projection in two dimensions (x and y) of the multi-dimensional distances (1 – *cosine*) among journals. Leydesdorff, Kushnir, & Rafols (in press) used the latter measure straightforwardly for an analogous mapping of (USPTO) patents in terms of International Patent Classifications (IPC). However, the number of IPC classes is currently 637, whereas the number of journals is more than 10,000. The number of distances would therefore be on the order of 10^8. Even after setting the threshold of cosine > 0.2, this number would be on the order of 10^6, and the size of the files would remain on the order of 50 Mbytes both cited and citing. (Furthermore, the threshold would be too coarse, because more distanced journals may often have a smaller cosine value between them than 0.2, and the variation in the distances (1 – *cosine*) would unnecessarily be reduced from zero to 0.8 given this threshold.)

Initial explorations led us also to the empirical conclusion that the results would be confounding because of the relative failing of relatedness in interdisciplinary sets above the level of the threshold. By using the distance on the map $\| x_i - x_j \|$ between two journals, these problems are circumvented. Since the MDS-like algorithm of VOSViewer already optimizes in terms of distances, we can use these distances between points directly for the computation of the Rao-Stirling diversity.[101] By normalizing these distances first against the maximum (diagonal)

[100] This extension assumes that the user has ticked the box within WoS for downloading "cited references" before the downloading, and thus one addresses the interdisciplinarity of another unit of analysis; for example, the knowledge bases of the sets (Bornmann & Marx, 2013; Leydesdorff & Goldstone, in press).

[101] A related program of VOSViewer, VOSmapping.exe at http://www.vosviewer.com/relatedsoftware/ , allows for specification of the dimensionality to more than two (Ludo Waltman, *personal communication*, December 30, 2012). However, one then loses the relation with the visible distances on the map. Furthermore, the extraction of a third dimension is not expected to add a large percentage to the explanation of the variance in the matrix (Schiffman *et al*., 1981). Given today's hardware systems limitations to the number of variables, it is not possible to specify the percentages of variance explained by the two first and/or later factors, using SPSS v. 20 (Leydesdorff, 2006).

value, one defines the diversity indicator between zero and one (since the *p*-values of the proportions are also fractions of one).

As an example, we return to the document sets used by Leydesdorff & Rafols (2012, at p. 328), namely the comparison of the publication portfolios 2006-2010 of the London Business School (LBS) and the Science and Technology Policy Research Unit (SPRU) at the University of Sussex. Using a number of indicators, Rafols *et al*. (2012) showed that the latter unit is far more diverse than the former even though both units are assigned to the same heading of Business & Management in the upcoming UK-wide evaluation, the so-called Research Excellence Framework. Figure 2 provides the two portfolios of 148 SPRU[102] (to the left) and 343 LBS publications[103] (to the right) as overlays on the 2011 "citing" maps, respectively. Table 1 provides the Rao-Stirling diversities for the two schools in both the cited and citing dimensions.

The difference in the values between cited and citing is caused by the larger distances in the citing map when compared with the cited one. The discriminating power of the citing map is therefore larger and the graphs are clearer. Furthermore, "citing" refers to the current knowledge base in 2011—as the running variable—whereas "cited" refers to the structure in the (cited) archive. We therefore recommend using the routine citing.exe unless one has theoretical reasons for focusing on "cited" or when the more fragmented clustering in the latter map is important for the argument.

Table 1: Rao-Stirling diversity for 143 SPRU and 343 publications (2006-2010) in both the citing and cited dimensions.

		Citing	*Cited*
SPRU	(*N* = 148)	0.218	0.136
LBS	(*N* = 343)	0.092	0.082

The routine provides at each run the value of Rao-Stirling diversity measure on the screen, and this value is saved to a file rao.txt. Note that this file is overwritten in each subsequent run; thus, these values have to be noted separately. Although the coordinates of VOSViewer can vary and take values larger than one (or less than minus one), the diversity values are normalized between zero and one, and the cited or citing values in Table 1 may therefore be compared,[104] and one can also compare results using sets of documents for different years. Using

---

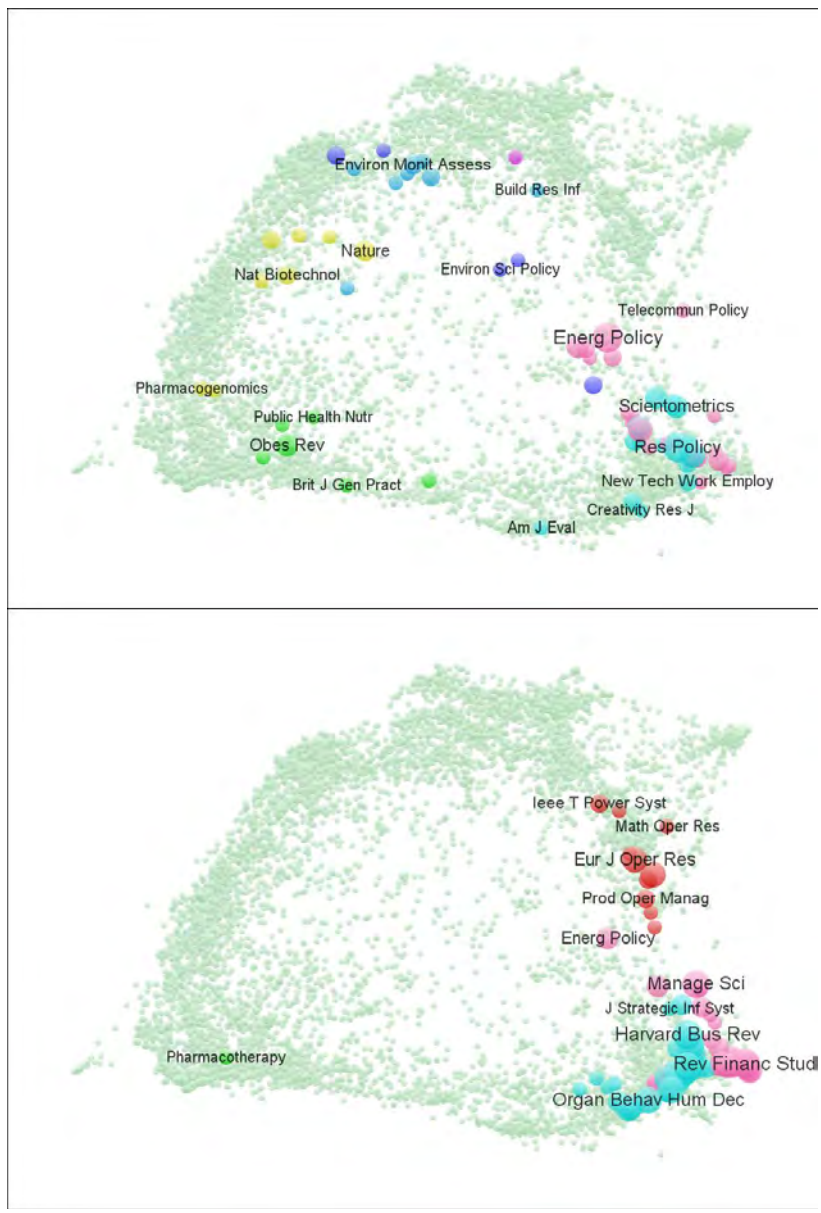
The dimensionality chosen remains therefore a bit arbitrary, unless one were able to use the (1–*cosine*) measure in the *N* = 10,330 dimensions of the full matrix of aggregated journal-journal citations. As noted, this approach would be computationally too intensive given the large value of *N*, but one can pursue such a more precise approach offline.

[102] Of the 155 SPRU papers, 148 were included in the largest component.

[103] Of the 348 LBS papers, 343 were included in the largest component.

[104] Because the map in the "cited" direction is more compressed, however, one cannot directly compare the projections in the cited and citing dimensions in terms of distances.

PowerPoint, the sequences for different years can also be animated on top of the otherwise stable base map.



Figures 2a and b: Overlay maps 2011 comparing journal publication portfolios from 2006 to 2010 between the Science and Technology Policy Research Unit SPRU at the University of Sussex (on the left; *N* = 148; available at http://www.vosviewer.com/vosviewer.php?map=http://www.leydesdorff.net/journals11/fig3a.txt&label_size=1.35) and the London Business School (on the right; *N* = 343; available at http://www.vosviewer.com/vosviewer.php?map=http://www.leydesdorff.net/journals11/fig3b.txt&label_size=1.35).

Further extensions to cited and citing sets of documents

As noted above, one can extend the analysis to the cited and citing sets of documents in terms of other units of analysis—or any unit of analysis that contains full journal names or the conventional abbreviations in the WoS format. For example, all journals titles in documents in WoS (or Scopus, PubMed, etc.) can be matched against the keys contained in the files cited.dbf and citing.dbf that are used for the overlay mapping and the measurement of interdisciplinarity. The two table files contain two keys: the full journal titles and the abbreviated ones using the conventions of WoS for the abbreviations. Our routines automatically correct for variations in upper and lower-case in these titles.

For example, document sets downloaded in WoS contain the abbreviated journal titles in the cited references (field-tag: "CR") in addition to the full journal names of each document which is tagged as "SO" (as an abbreviation of "source"). The journal names in the field "CR" may contain misspellings (Leydesdorff, 2008, Table 4 at p. 285), but Thomson-Reuters has recently invested in v5 of WoS (in 2011) to improve standardization in the CR-field. Two additional (sister) programs are brought online that operate on this field when properly downloaded in the tagged-format and renamed as "data.txt" (as above). These routines ("crciting.exe" at http://www.leydesdorff.net/journals11/crciting.exe and "crcited.exe" at http://www.leydesdorff.net/journals11/crcited.exe) operate on the journal names in the cited references in the document set under study using the standard abbreviations of WoS for the comparison (whereas the original programs citing.exe and cited.exe use the full journal titles).[105]

For example, the above used set of 155 documents (Figure 3a) published by authors with an address at SPRU between 2006 and 2010, contains 7,545 cited references of which 2,552 can be matched with the WoS keys for the journal abbreviations. Figure 3 is not shown here, but can be web-started at http://www.vosviewer.com/vosviewer.php?map=http://www.leydesdorff.net/journals11/figure4.txt&label_size =1.35; it provides the map of the knowledge base of the SPRU authors overlaid on the basemap in the citing direction. The Rao-Stirling diversity is marginally down to 0.214 (from 0.218 in Table 1).

The relatively low rate of matching (2552 of 7545; 33.8%) is perhaps itself indicative of the interdisciplinary nature of these articles, which often result from policy-oriented and externally funded reports. Gibbons *et al*. (1994) called this type of interdisciplinarity the "Mode 2"-type of knowledge production: not only more "interdisciplinarity," but also more engagement with social actors. In the case of the more disciplinarily oriented ("Mode 1") London Business School, 348 documents contain 16,713 cited references, of which 10,034 (60.0%) could be validated in terms of sources at WoS ($\Delta = 0.096$). Thus, the knowledge base of

[105] The abbreviated journal titles are stored in a file cr.dbf that contains a fieldname "journalcr" used by the routines. When one uses other installations of the *Science Citation Index* such as on Dialog or STN, one may have to rename both the file (to cr.dbf) and this fieldname to "journalcr" (Lutz Bornmann, *personal communication*, 5 January 2013).

these authors is also more "disciplined" in the sense of being more oriented toward academic objectives.
WoS offers the possibility to download also the *citing* journals of a set by creating a so-called "citation report." This screen allows only for downloading as comma-separated variables or Excel sheets. The download contains the names of the journals, but not the cited references. After changing the field-name in the Excel sheet into "SO" (as in the tagged format) and saving the file using the name "core.dbf", citing.exe and cited.exe will use this file as source information in the absence of data.txt, and thus produce the overlay files citing.txt or cited.txt, respectively, and Rao-Stirling diversity values. (As noted, saving an Excel sheet in the .dbf format is easier in OpenOffice or SPSS than in more recent versions of Excel.)

Conclusion

The journal map based on a cosine-normalized matrix and using the MDS-like solution of VOSViewer captures journals as positions in a vector space that is reduced to the two dimensions of the plane. The first two main dimensions of the underlying citation matrix can be expected to capture the major part of the variance in the matrix (Schiffman *et al*., 1981; see footnote 22 above). However, a map remains a projection (in two dimensions). Unlike spring-embedded solutions, the projection of MDS is not dependent on a seed, but the *system* of journal-citations is projected deterministically.[106] The journals are positioned in the vector space on the basis of the aggregates of their mutual relations (Leydesdorff, in press).
The journal is a more precise unit of analysis when compared with the journal grouping using WOS Categories (Rafols *et al*., 2010; Leydesdorff & Rafols, 2012; Leydesdorff, Carley, & Rafols, in press). The WOS Categories are both divisive and overlapping, since journals can be attributed to several categories, on the one hand, but the cuts between categories remain sharp, on the other. The consequent error reflects uncertainty in the networks about the delineations (Rafols and Leydesdorff, 2009; Rafols *et al*., 2010). In a lower-level networked system of journals, such decisions are not needed since all cosine-normalized distances among journals can be introduced concurrently into the computation. (The threshold of cosine > 0.2 was set above because of technical limitations.)
A network system at the article level would be even more precise, but dysfunctional in terms of overlay files for studying sets, for example in terms of their interdisciplinarity, because the journals are no longer considered as relevant categories. One would be able to position articles, but one cannot position other articles in terms of a baseline of articles. Another advantage of positioning papers in terms of locations on the map of journals is the availability of network measures of interdisciplinarity. Rao-Stirling diversity measure of

[106] VOSViewer uses a seed, but the algorithm tends to converge to the global maximum of the quality function (Van Eck & Waltman, 2012, at p. 2).

interdisciplinarity, for example, operates directly on the values that are visible on the map, that is, the distances between the nodes and the (logarithmically normalized) sizes of the nodes given the document set(s) under study.

Acknowledgement

We would like to thank Nees Jan van Eck and Ludo Waltman for suggestions and comments, and are grateful to Thomson-Reuters for access to the data. We acknowledge support by the ESRC project 'Mapping the Dynamics of Emergent Technologies' (RES-360-25-0076). A full version of this paper (Leydesdorff *et al*., in press) was in the meantime accepted for publication in the *Journal of the American Society for Information Science and Technology.*

References

Barry, A., Born, G. & Weszkalnys, G. (2008). Logics of interdisciplinarity. *Economy and Society 37*(1), 20–49.

Bensman, S. J., & Leydesdorff, L. (2009). Definition and Identification of Journals as Bibliographic and Subject Entities: Librarianship vs. ISI Journal Citation Reports (JCR) Methods and their Effect on Citation Measures. *Journal of the American Society for Information Science and Technology, 60*(6), 1097-1117.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 8*(10), 10008.

Bornmann, L., & Marx, W. (2013). The proposal of a broadening of perspective in evaluative bibliometrics by complementing the times cited with a cited reference analysis. *Journal of Informetrics, 7*(1), 84-88.

Boyack, K.W. (2009). Using detailed maps of science to identify potential collaborations. *Scientometrics* 79(1), 27-44.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the Backbone of Science. *Scientometrics, 64*(3), 351-374.

Carley, S., & Porter, A. L. (2012). A forward diversity index. *Scientometrics, 90*(2), 407-427.

Chen, C., & Leydesdorff, L. (in press). Patterns of Connections and Movements in Dual-Map Overlays: A New Method of Publication Portfolio Analysis *Journal of the American Society for Information Science and Technology*.

Fruchterman, T., & Reingold, E. (1991). Graph drawing by force-directed replacement. *Software—Practice and Experience, 21*, 1129-1166.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The new production of knowledge: the dynamics of science and research in contemporary societies*. London: Sage.

Izsák, J., & Papp, L. (1995). Application of the quadratic entropy indices for diversity studies of drosophilid assemblages. *Environmental and Ecological Statistics, 2*(3), 213-224.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters, 31*(1), 7-15.
Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.
Klavans, R., & Boyack, K. (2009). Towards a Consensus Map of Science. *Journal of the American Society for Information Science and Technology, 60*(3), 455-476.
Klein, J. T. (1990). *Interdisciplinarity: History, Theory, & Practice.* Detroit: Wayne State University Press.
Leydesdorff, L. (2006). Can Scientific Journals be Classified in Terms of Aggregated Journal-Journal Citation Relations using the Journal Citation Reports? *Journal of the American Society for Information Science & Technology, 57*(5), 601-613.
Leydesdorff, L. (2007). "Betweenness Centrality" as an Indicator of the "Interdisciplinarity" of Scientific Journals. *Journal of the American Society for Information Science and Technology, 58*(9), 1303-1309.
Leydesdorff, L. (in press). Advances in Science Visualization: Social Networks, Semantic Maps, and Discursive Knowledge. In B. Cronin & C. Sugimoto (Eds.), *Next Generation Metrics: Harnessing Multidimensional Indicators of Scholarly Performance*. Cambridge MA: MIT Press.
Leydesdorff, L., & Rafols, I. (2009). A Global Map of Science Based on the ISI Subject Categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348-362.
Leydesdorff, L., Hammarfelt, B., & Salah, A. A. A. (2011). The structure of the Arts & Humanities Citation Index: A mapping on the basis of aggregated citations among 1,157 journals. *Journal of the American Society for Information Science and Technology, 62*(12), 2414-2426.
Leydesdorff, L., & Rafols, I. (2011a). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics, 5*(1), 87-100.
Leydesdorff, L., & Rafols, I. (2011b). Local emergence and global diffusion of research technologies: An exploration of patterns of network formation. *Journal of the American Society for Information Science and Technology, 62*(5), 846-860.
Leydesdorff, L., & Rafols, I. (2012). Interactive Overlays: A New Method for Generating Global Journal Maps from Web-of-Science Data. *Journal of Informetrics, 6*(3), 318-332.
Leydesdorff, L., & Goldstone, R. L. (in press). Interdisciplinarity at the Journal and Specialty Level: The changing knowledge bases of the journal Cognitive Science, *Journal of the American Society of Information Science and Technology;* available at arXiv preprint arXiv:1212.0823.
Leydesdorff, L., Kushnir, D., & Rafols, I. (in press). Interactive Overlay Maps for US Patent (USPTO) Data Based on International Patent Classifications (IPC). *Scientometrics*.

Leydesdorff, L., Rafols, I., & Chen, C. (in press). Interactive Overlays of Journals and the Measurement of Interdisciplinarity on the basis of Aggregated Journal-Journal Citations. *Journal of the American Society for Information Science and Technology*.

Porter, A.L., Cohen, A.S., Roessner, J.D., Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics* 72, 117–147.

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology, 53*(13), 1113-1119.

Rafols, I., & Leydesdorff, L. (2009). Content-based and Algorithmic Classifications of Journals: Perspectives on the Dynamics of Scientific Communication and Indexer Effects. *Journal of the American Society for Information Science and Technology, 60*(9), 1823-1835.

Rafols, I., & Meyer, M. (2010). Diversity and Network Coherence as Indicators of Interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, *82*(2), 263-287.

Rafols, I., Porter, A., & Leydesdorff, L. (2010). Science overlay maps: a new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871-1887.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy, 41*(7), 1262-1282.

Rao, C. R. (1982a). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology, 21*(1), 24-43.

Rao, C. R. (1982b). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhy : The Indian Journal of Statistics, Series A, 44*(1), 1-22.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling: theory, methods, and applications*. New York / London: Academic Press.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707-719.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538.

Van Eck, N. J., & Waltman, L. (2012). The VOS mapping software: A brief introduction. Retrieved from http://www.vosviewer.com/relatedsoftware/ (Dec. 31, 2012).

Van Eck, N. J., Waltman, L., Dekker, R., & van den Berg, J. (2010). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology, 61*(12), 2405-2416.

Waltman, L., van Eck, N. J., & Noyons, E. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629-635.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to Understanding and Measuring Interdisciplinary Scientific Research (IDR): A Review of the Literature. *Journal of Informetrics, 5*(1), 14-26.

INTERDISCIPLINARY RESEARCH AND THE PRODUCTION OF LOCAL KNOWLEDGE: EVIDENCE FROM A DEVELOPING COUNTRY

Diego Chavarro[1], Puay Tang[2] and Ismael Rafols[3]

[1] *diego.chavarro@sussex.ac.uk*
SPRU - Science and Technology Policy Research University of Sussex, Brighton (England)

[2]*p.tang@sussex.ac.uk*
SPRU - Science and Technology Policy Research University of Sussex, Brighton (England)

[2]*i.rafols@sussex.ac.uk*
*Ingenio (CSIC-UPV),*Universitat Politècnica de València, València (Spain) &
SPRU - Science and Technology Policy Research University of Sussex, Brighton (England)

Abstract
This study shows that interdisciplinary research is important for the development of knowledge pertaining to local issues. Using the Colombian publications from 1991 until 2011 in the Web of Science, we investigate the relationship between degree of interdisciplinarity (inferred from references) and local focus of the articles (as shown by the use of the term 'Colomb*' in the title, keywords or abstracts). We find that higher degree of interdisciplinarity in a publication is associated with more focus on local issues. In particular, publications combining disparate disciplines in balanced proportions are shown to be more likely to relate to local issues, other things being equal. These results support the view that policies fostering cognitively disparate disciplines may be useful for strengthening the local relevance of research.

Conference Topic
Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3)

Introduction
It is widely assumed that research addressing social and economic challenges is best conducted through interdisciplinary approaches because they defy disciplinary categorization and solutions. In discursive terms, interdisciplinarity has become a mantra of science policy. The recognition of the benefits of interdisciplinary research has indeed stimulated a steadily growing interest in developing new knowledge through research that integrates the skills and perspectives of multiple disciplines.

This article aims to add to the body of literature on the role of interdisciplinary research (IDR) to address complex social, cultural, economic and political issues by examining the relationship between interdisciplinary research and the production of "local knowledge". By "local knowledge" we mean it to be knowledge of local conditions or issues. In our case these that are pertinent to a whole country, Colombia. For the purposes of this article, we assume that local knowledge has social and economic relevance to this country. This focus on a particular locus is supported by Barry, Born & Weszkalnys (2008) who have asserted that IDR research (more below) and the salience of the importance of the "context of application as a site for research…. at which knowledge is produced" (p. 21) needs to be accounted for.

Scholars in various fields have increasingly recognized the need to link disciplinary fields in order to more fully respond to pressing societal questions or to deal with a particular problem. For instance, health may not be adequately studied through a disciplinary framework. Instead, poor health results from a constellation of factors: malnutrition, bad eating habits, genetics, age, poverty, ignorance, pollution, environmental conditions, and peer pressure (for instance, anorexia). As once pointed out by Kofi Annan, the ex-Secretary General of the United Nations, "we shall not finally defeat AIDS, tuberculosis, malaria or any of the other infectious diseases that plague the developing world until we have also won the battle for safe drinking water, sanitation and basic health care …" (as cited in Dodd and Munck, 2002, p.2).

There now exists a large body of literature on the benefits of interdisciplinary research. Nowotny, Scott and Gibbons (2001) observed that science is undergoing a shift from a Mode-1 production of science, which is mainly disciplinary and initiated by the interests of the researcher, to a Mode-2 which is interdisciplinary, that displaces "a culture of autonomy of science" (p.89) and addresses socially relevant issues. In this context, interdisciplinary research has received direct support in recent years through public policies as a means of fostering the social relevance of research. As Barry, Born & Weszkalnys (2008) note, "what is novel is the contemporary sense that greater interdisciplinarity is a necessary response to intensifying demands that research should be integrated with society and the economy" (p. 23).

One of the most widely used definitions of interdisciplinary research regards it as

> a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or field of research practice (National Academies, 2004: 66).

Initiatives aimed at interdisciplinary research are based on the view that it strengthens, renews and interweaves science, technology, society and innovation. Hence, not surprisingly, the notion of interdisciplinary research has permeated into the formulation of various Science Technology and Innovation (ST&I) policies and such research has ostensibly come to be regarded as an essential

component of these policies as reflected in a number of documents. Examples of such documents are those by, among others, the OECD, UNESCO (Godin, 2009), the UK Royal Society, research funding agencies, such as the U.S. National Science Foundation (Adams and Clemons, 2011: 218), National Institute of Health and UK Research Councils, government agencies and universities (Brint, 2005), among others.

Despite the apparent wide acknowledgement of the benefits of interdisciplinary research, scholars have found that IDR is in practice discouraged in a variety of ways. One way is research assessment practices that many countries have increasingly implemented. These assessment exercises are based on disciplinary perspectives (see special issue edited by Laudel and Origi, 2006; Martin, 2011; also reviewed in Rafols et al., 2012). This disciplinary emphasis has tended to encourage academics to "game" the evaluation system by publishing in disciplinary journals with the potential result of jeopardizing more interdisciplinary "risky research" that may yield greater social and economic impacts (Nightingale and Scott, 2007, pp. 546-547, Smith et al. 2011). In universities, a prevailing 'silo' mentality also tends to discourage interdisciplinarity. Such behaviour may hinder the ability to address future 'grand challenges,' such as smart cities and climate change although many governments consider these as national priorities.

The relationship between interdisciplinarity and research on local issues

Against the extant literature on the contribution of IDR to a range of public and "real-life" issues and abiding with the importance of context (in our case, a developing country Colombia) in such research, we argue that IDR can be expected to play an important role in the development of local S&T capabilities. Already noted above, its importance is further illuminated below:

> Necessity and complexity have also been cited as reasons for interdisciplinary research in and about developing countries. Shinichi Ichimura cautioned that the conceptual frameworks of traditional disciplines are often too narrow and too compartmentalized for the study of problems in other areas. Norman Dinges made a similar observation about cross-cultural research, suggesting interdisciplinary perspective grows as the "indigenization" of research sensitive to local norms takes place; and Lawrence Murphy, using the example of the Social Research Center of the American University of Cairo (Egypt), has traced the movement from narrow, academically oriented research projects to more appropriate long-term interdisciplinary, multifaceted studies that analyzed problems of immediate concern to the host nation. (Klein, 1990, p. 45)

Scholars have also argued that local contexts are enablers of interdisciplinary research because they require different cognitive approaches to understand and address their specific needs:

> Practical contexts also have aspects that combine perspectives from different disciplines and are seldom intelligible without the development of novel inter-, multi- or transdisciplinary modes of knowledge production. (...) Localized science (...) is not just a 'perturbation' of the claims of universally valid paradigms or a denial of the feasibility of generalizing, reducing and deducing anything and everything. Knowledge production in the context of application is itself a fertile seedbed for the emergence of novelty. Localized investigations create genuine new knowledge. They can be full of surprises, especially when they combine knowledge elements from different realms, and mix them with societal expectations. (Nowotny and Ziman, 2002).

The importance of 'localized' research has been also highlighted by Stiglitz, who points out that "local researchers combining the knowledge of local conditions – including knowledge of local political and social structures -- ……provide the best prospects for deriving policies that both engender broad-based support and are effective…" (Stiglitz, p. 24 in Stone, 2000). This is an argument also underscored by Bones et al. (2011) in their study on the importance of a range of "local knowledge providers" and communication channels to improve the public health of a remote area in western Alaska (see also Gahi, 2004). Specifically for developing countries, the production of locally relevant interdisciplinary knowledge is considered key for achieving what has been called the "indigenisation of science", which results from the selection, adaptation, application, localization and combination of theories and methodologies from different sciences (Alatas, 1993: 312).

However, as Jacobs and Frickel (2006) have argued, the assumptions behind policies for interdisciplinary research have yet to be tested, both theoretically and empirically. Although the claim for the relationship between IDR and local knowledge has been argued theoretically and on the basis of anecdotal evidence, there is a need to test this assumption on a greater scale and to further refine our analytical understanding of the ways in which IDR and the local context of application might be intertwined.

This article attempts to examine empirically the relationship between IDR and the production of local knowledge. We investigate interdisciplinary research by drawing on publications data from journal articles, reviews and proceedings papers indexed by the Web of Science (WoS). For this, we use recently developed bibliometric indicators to gauge the degree of interdisciplinarity (Porter and Rafols, 2009) and a multivariate test to find whether there is a significant relationship between degrees of interdisciplinarity in a publication and the production of publications on local issues. The statistical method chosen for this is logistic regression, which allows one to find the probability that an event (publication of an article on local issues) occurs given the presence of a predictor (degree of interdisciplinarity and other variables). Our study focuses on Colombia as an example of an developing country with a growing ST&I system.

Methods

Data and Sample

The dataset is comprised of articles, reviews and proceedings papers included in the ISI Web of Science Database. These articles are authored by at least one researcher who was affiliated to a Colombian institution at the time of publication. We included records since 1991, one year after the official foundation of the Colombian System of Science and Technology and the designation of Colciencias as the institution in charge of ST&I policy in the country. Given that the method for gauging degree of interdisciplinarity relies on references, we only took into account records with more than three bibliographic references successfully categorized into disciplines. Also we only considered publicaitons that had information on the countries of the participating co-authors (this criteria excluded several records). The application of these filters yielded 14,402 records, approx. 75% of the total sample of reviews, articles and proceedings papers published with a Colombian address after 1990.

Variables and Method

This study is focused on the relationship between two main variables: the first one is orientation of research. It is "local" when it directly mentions the word "Colomb*" in the title or abstract and "non-local" when it does not (in regressions, 1 means "local" and 0 "non-local" orientation). The second one is the degree of interdisciplinarity, which we measure with various indicators of diversity (more below) ranging from 0 to 1 (1 indicates totally interdisciplinary and 0 completely disciplinary).

We chose the country name as the criterion to identify locally oriented research because place-names act both as a coordinate system that locates geographically the action being performed and as a characterizing device that sets the action within a specific socio-economic context (for a conceptualization of place-names as indexical and characterizing signs, see Keates, 1996, pp. 81-82). Place-names "are of such vital significance because they act so as to transform the sheer physical and geographical into something that is historically and socially experienced" (Tilley, 1994, p. 18). This approach has also been used by Ordóñez-Matamoros, Cozzens and Garcia (2010).[107]

When operationalizing the measurement of interdisciplinarity, we follow Yegros-Yegros et al. (2010), who use each of the dimensions of diversity (variety, balance and disparity) separately as well as a synthetic measure of diversity (Rao-Stirling's) which combines all three dimensions. The Rao-Stirling diversity (also

[107] In a quick examination of the use of a "place-name" we found that the percentage of publications that mention the country in their title, keywords or abstracts is much higher in Latin American countries, for instance, Colombia, Brazil, Argentina, Chile and Mexico, than in developed countries such as the U.S., the Netherlands, Germany and the UK. For the former group of countries, papers accounted for 15% to 25% of their total production, whereas for developed countries the percentage is below 5%.

known as 'quadratic entropy') was first proposed as a measure of interdisciplinarity by Porter et al. (2007), who called it an 'Integration score', which was then further developed by Rafols and Meyer (2010). The key advantage of this measure is that it not only takes into account the distribution of references across disciplinary categories, but crucially also considers how cognitively distant these categories are. Intuitively, this means that a publication with references from atomic physics and cell biology is weighted as more interdisciplinary than one with references from cell biology and biochemistry.
The equations for each variable of diversity are found below:

$$Variety = v = \text{Number of disciplines}$$

$$Balance = \frac{1}{\ln(v)} \sum_i p_i \ln p_i$$

$Disparity = \frac{1}{v(v-1)} \sum_{i,j} d_{ij}$, sum only for those categories in the reference set.

$$Rao - Stirling\ Diversity = \sum_{i,j} p_i p_j d_{ij}$$

where v_{max} = variety of the article with a greater number of disciplines identified within the dataset, p_i = proportion of elements in category i, d_{ij} = distance between categories i and j (Rafols and Meyer, 2010: 267).
The variables above are part of a conceptualization of IDR as *diversity* (Rafols and Meyer, 2010 based on Stirling, 2007:710). *Variety* corresponds to the number of categories in which elements can be classified, for instance, if a researcher finds five different species of amphibians in an ecosystem, five is the value of variety. *Balance* describes the evenness of the distribution of elements into categories. A sample is completely balanced if all categories share the same number of elements. *Disparity* is used to reflect the degree of the distinctiveness that exists between the elements of the distribution. If classifications are a means to separate elements, disparity is a property that tells the extent of separation (the distance) between the categories used. For example, soprano voices are closer to mezo-soprano than to contralto voices in terms of tone range. For this, a value for distance between elements has to be set.
The cognitive distances $d_{i,j}$ between categories are drawn from the metrics underlying the global maps of science based on the ISI Web of Science Categories (formerly Subject Categories, see annex) (Rafols et al., 2010). Each measure of diversity is calculated for each article by classifying bibliographic references into one or more WoS Categories, using the software Vantage Point[108]. This attribution of references to WoS Categories is very inaccurate –there is up to 50% disagreement between alternative classifications (Rafols and Leydesdorff, 2009, p. 1828). As a result, the diversity measure of a single article has a large noise and is not reliable, but the robustness of global science maps suggests that

[108] www.thevantagepoint.com

the error is not systematic, and with large numbers, one can still obtain good approximations (Rafols and Leydesdorff, 2009, p. 1829). As our sample consists of 14,402 publications, we can be confident that the aggregation will yield reliable results.

After classifying the references, a script in statistical language R was run on a matrix of articles vs. cited disciplines to determine the indicators. The variables Rao-Stirling Diversity, Variety, Disparity and Balance were calculated by this means.

In addition, we have incorporated two control variables that may have effects on the dependent variable, local knowledge: these are (i) Collaboration and (ii) Field to which an article is more likely to belong, for instance Biosciences or Social Sciences. The variable Collaboration shows whether in an article there is more than one country in the affiliation. It is a dummy variable with the categories International collaboration, National collaboration and No collaboration. This variable was identified from the field C1 in the WoS format. The categorical variable for Field ("Macro-discipline") aims to control how the cognitive context may influence the local or non-local nature of the outcomes of research, given that some disciplinary fields can be more prone to produce local studies than others. This variable is constructed based on the results of a study by Rafols et al. (2010). Using factor-analysis, these authors classified WoS Categories into 18 'Macro-disciplines'. Macro-disciplines are aggregations of into large disciplinary groups with similar citation patterns. We performed a match between the most cited subject of an article and the list of the 18 macro-disciplines. Table 1 shows a description of all the variables.

Regression

To test the relationship between interdisciplinarity and research orientation we use a regression technique. Regression techniques try to find the relationship between a set of explanatory variables on one or more dependent variables. The kinds of regressions used will depend on the types of variables. Here, we used logistic regression. While other techniques, such as discriminant analysis, require meeting strict conditions of multivariate normality and equal distribution of variance and covariance matrices, logistic regression is robust when such conditions are not strictly met (Hair et. al., 2005: 276). For these reasons we have selected logistic regression using the statistical packet SPSS.

Logistic regression is similar to normal regression, but it follows a different approach for estimating the coefficients. As the error term for a dichotomous dependent variable does not follow a normal distribution and the variance is not constant (Hair et. al., 2005: 277), logistic regression does not use the method of least squares (OLS) to estimate the regression model. Instead, it uses a maximum likelihood estimation, which consists in fitting an S-like probabilistic curve to best fit the data. This implies that the interpretation of the coefficients is done by looking at the exponential of the beta coefficients, because they are given in logits (the logarithm of the conditional probability of a variable).

Table 6: Description of the variables used in the study

Name	Type	Values	Role	Description
Research orientation	Categorical	1 = local 0 = non-local	Dependent	If an article has the word Colomb* in the title, abstract or keywords, it is considered local (1)
Rao-Stirling Diversity	Numerical	Between 0 and 1	Independent	This variable synthesizes three properties of disciplinary diversity: variety, balance and disparity.
Variety	Numerical	Between 1 and 222	Independent	Number of Web of Science categories cited by each article.
Balance	Numerical	Between 0 and 1	Independent	Balance in terms of proportion of references in each Web of Science Categories cited by an article.
Disparity	Numerical	Between 0 and 1	Independent	Average distance between the Web of Science Categories cited by an article. Distances are given by cross-citations between Web of Science Categories across all science.
International Collaboration	Dummy	0 or 1	Independent	1 if more than one country participates in an article.
National Collaboration	Dummy		Independent	1 if more than one Colombian author
No Collaboration	Dummy	0 or 1	Independent	1 if there is no collaboration
Macro-Discipline	Dummy	Agricultural sciences Biomedical sciences Business and Mngt. Chemistry Clinical medicine Cognitive sciences Computer sciences Ecology Economics & geography Engineering Environmental S&T Geosciences Health services Infectious diseases Materials sciences Physics Psychology Social studies	independent	This is an aggregation of disciplines in terms of cross-citations made by Rafols et al. (2010). This variable groups articles in terms of their belonging to one of these categories. Each article belongs to one category. The assignation of an article to a category was done by the most referenced discipline in each article.

As explained above, the dependent variable is whether an article is local or not (research orientation), and the main predictor is the degree of interdisciplinarity, firstly as a synthetic variable (Rao-Stirling diversity) and secondly as represented

by its different dimensions (variety, balance, disparity). In order to account for the socio-cognitive context in which research takes place, we have explored the influence of Collaboration and Macro-discipline, also noted before.

Thus, we performed the logistic regression in two blocks: In the first we incorporated Rao-Stirling diversity, Collaboration and Macro-disciplines. In the second, we replaced Rao-Stirling diversity by the set of separate dimensions of interdisciplinarity as diversity: Variety, Balance and Disparity. We also tested for a possible inverted U-shape relationships between IDR variables and the dependent. The reduction in the -2 log likelihood (the variance) of each model is used as a criterion to assess the improvement in each block. We use three Pseudo-R^2 measures to assess the adequacy of the models. The first measure is Hosmer and Lemeshow's R^2, the second Cox and Snell's R^2 and the third Nagelkerke's R^2. These measures calculate the variation that is explained by the model based in -2 LL. The first is calculated as -2LL (new model) /-2LL (original model). 0 means no improvement and 1 means total fit of the model. This measure, however, does not take into account the size of the sample. For that, Cox and Snell's R^2 is used. As this measure cannot reach the theoretical maximum of 1, the correction by Nagelkerke is used. These three statistics can help to assess the goodness of fit of the model (Field, 2009: 269).

Results

We first present the general descriptive values for the key variables in variable in this study. Table 2 shows that the dependent variable (local) has a small share of articles (24%) referencing explicitly Colombia in their texts as compared to articles not mentioning it. Regarding collaboration, we can see that articles in the WoS database are more likely to be done in collaboration with authors from abroad. However, in general terms, the number of Colombian publications in journals covered by WoS has been increasing since 1991. It grew from 85 in 1991 to 2,203 in 2010, which represents an approximate 24 fold increase.

Table 2: Descriptive statistics of measures of interdisciplinarity

	Frequency	%
Research orientation		
Non-Local	10930	75.89%
Local	3472	24.11%
Collaboration		
National	4968	34.50%
International	8749	60.75%
No Collaboration	685	4.76%

Figure 1 provides an intuitive insight of the relationship between Rao-Stirling diversity and research orientation. We see that the proportion of locally focused publications is higher in publications with Rao-Stirling diversity higher than 0.5.

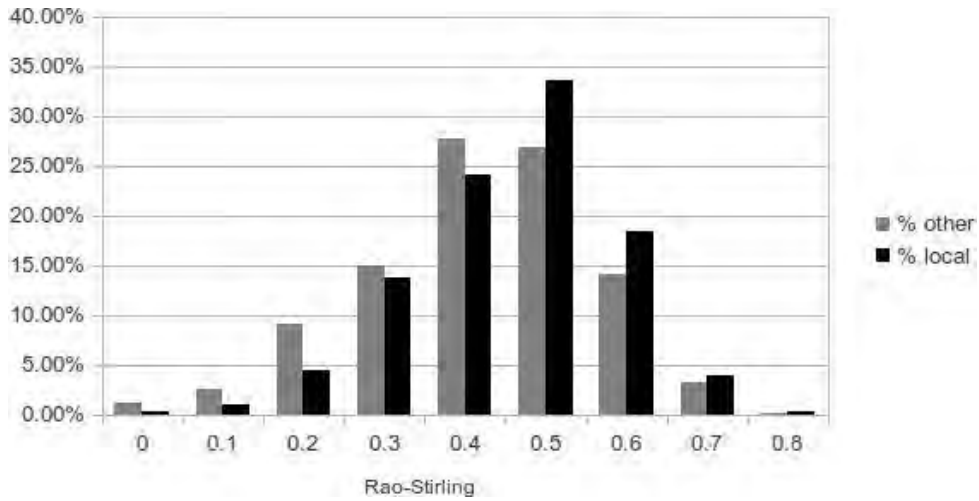


Figure 1. Percentage of local-focus and non-local papers by Rao-Stirling diversity intervals

It is worth noting that most of the publications present an average score in Rao-Stirling diversity, i.e., they are moderately interdisciplinary. The distribution of the variable shows a normal curve, falling between accepted ranges of kurtosis and skewness. Extreme cases like publications with very low (0.1) or very high (0.8) Rao-Stirling diversity are unusual. When exploring Variety, Balance and Disparity in regard to research orientation we found that the share of local papers is slightly greater for higher degrees of disparity and balance, whereas it is lower for variety.

Logistic regression

As explained, the logistic regression was performed in two model. In the first we incorporated Rao-Stirling diversity, adding Collaboration and Field (macro-discipline). In the second, we replaced Rao-Stirling diversity with the set of separate characteristics: Variety, Balance and Disparity. Table 4 presents the results. It is found that IDR variables (Rao-Stirling diversity, variety, balance and disparity) are related with the production of knowledge on local issues. These relationships are statistically significant. The relationships are as follows.

Firstly, Rao-Stirling diversity is positively related to the production of knowledge on local issues. The odds ratio shows that for each unit increase in Rao-Stirling diversity (controlling for Collaboration and Field), it is 1.7 times more likely that an article is related to local issues.

Secondly, when interdisciplinarity is decomposed into its constituent properties, the effects of each property on the probability of finding an article on local issues varies. Disparity and Balance show a positive relationship with the local focus of articles. A unit increase in these variables makes it approximately three times more likely that a paper is on local issues. Variety, on the other hand, contributes

negatively to this relationship. A unit increase in Variety makes it 0.9 times less likely that a paper is local.

Thirdly, it is important to note that the controls used in this analysis have also significant effects on the predicted variable. National collaboration and International collaboration are positively related with the production of knowledge on local issues. As compared to No collaboration, National collaboration increases the probabilities to publish on local issues by about two times, while international collaboration does it by 1.2 times.

Table 3. Coefficients of the logistic regression

Variables	***Model 1***	***Model 2***
Rao-Stirling Diversity	0.539 (1.715) **	
Variety		-0.257 (0.945) ***
Balance		1.051 (2.861) ***
Disparity		1.11 (3.034) ***
Control variables		
National Collaboration	0.743 (2.101) ***	0.77 (2.161) ***
International Collaboration	0.155 (1.168)	0.227 (1.255) *
Fields (macro-disciplines)		
Agricultural_Sciences	0.119 (1.126)	-0.025 (0.976)
Business and Management	0.502 (1.653) *	0.263 (1.301)
Chemistry	-1.925 (0.146) ***	-2.104 (0.122) ***
Clinical_Medicine	-0.181 (0.834) *	-0.32 (0.726) ***
Cognitive Sciences	-0.187 (0.829)	-0.259 (0.771) *
Computer Science	-1.647 (0.193) ***	-1.943 (0.143) ***
Ecology	1.195 (3.305) ***	1.083 (2.955) ***
Economics and Geography	0.212 (1.236)	-0.067 (0.935)
Engineering	-2.291 (0.101) ***	-2.61 (0.074) ***
Environmental ST	-0.29 (0.748) **	-0.504 (0.604) ***
Geoscience	1.805 (6.079) ***	1.619 (5.047) ***
Health Services	1.409 (4.093) ***	1.249 (3.487) ***
Infectious Diseases	0.586 (1.797) ***	0.589 (1.802) ***
Materials Science	-2.891 (0.056) ***	-3.076 (0.046) ***
Physics	-4.406 (0.012) ***	-4.675 (0.009) ***
Psychology	0.397 (1.487) *	0.291 (1.338)
Social Studies	0.956 (2.602) *	0.746 (2.109)
Constant	-1.627	-2.341
Cox and Snell's R2	0.199	0.207
Negelkerke's R2	0.297	0.309

*** p < .001, ** p < 0.01, *** p <0.05

Note: Odds ratios in parentheses. Model 1 includes Rao-Stirling diversity as a single measure for interdisciplinarity. Model 2 replaces Rao-Stirling diversity with variety, evenness and disparity. The reference category for Collaboration is No Collaboration and the reference category for Macro-discipline is Biomedical Sciences.

Different macro-disciplines are related to the production of knowledge on local issues in different ways. As compared to biosciences (used as reference category), some macro-disciplines increase the probability of producing publications on local issues. They are business and management, ecology, geosciences, health services, infectious diseases, psychology, and social studies. Their odds ratios show an increase in odds between two (Social Studies) and five (Geosciences).
Finally, we tested for inverted U-shape relationships in each of the IDR related variables. None of the quadratic variables showed a significant coefficient ($p < 0.05$).

Discussion

Our results support the view that IDR is related to the production of scientific knowledge on local issues. Articles with a local focus tend to be more interdisciplinary. This could be explained by the fact that research related to local issues is often associated with problem-oriented research, which is then associated with interdisciplinary research. An analysis of the top 10 most interdisciplinary articles of the sample supports this view. Six out of them were classified as local and most of them focus on topics directly related to the Colombia: malaria, fruits, management of biotechnology in Colombia, transport. The local paper that appears to be less related to direct application one about history, but even then it is history of engineering education, which is relevant in terms of technological development. The majority of 10 articles appear to involve problem-oriented research, with perhaps the exception of the last article, which appears to be more theoretical.
A finer analysis, unpacking the different dimensions of interdisciplinarity, reveals that articles with a focus on local issues tend to have a more balanced composition of highly disparate bodies of knowledge (more balance and disparity) in their references, but, interestingly, with less categories (less variety). The interpretation of these results is that local knowledge is associated with long range, high risk interdisciplinarity across distant cognitive areas, rather than piecemeal interdisciplinarity across neighbouring fields. Interestingly, these findings are exactly opposite to those by Yegros-Yegros et. al. (2010) about the relation between IDR and scientific performance in terms of numbers of citations, as shown in Table 6. The latter finds a positive influence of variety and a negative influence of disparity and balance on scientific performance, as measured in terms of number of citations per paper. The exact opposite tendency between our findings and Yegros-Yegros et al. (2010) suggests that related mechanisms may be at play: on the one hand, problem-oriented research tends to associated with cognitively disparate IDR, on the other hand, problem-oriented research tends to be less valued in academic terms (less cited) –therefore cognitively disparate IDR gets less citations.
Although this study contributes to testing and to a better understanding of the relationship between IDR and scientific knowledge on local issues, there are some limitations to the conclusions. First, different results might be found in developed

countries, in which the local focus is perhaps not as evident as in a developing country such as Colombia. However, we think that our results could be generalized to other developing countries, in the so called "periphery" of the system, which are trying to participate in the global scientific community and at the same time are making efforts to adapt and develop knowledge relevant to their local contexts with the aim of appropriating the socio-economic returns of S&T.

Table 4. Relation between different dimension of diversity with performance and local focus

	Performance (Yegros-Yegros et al. 2010)	*Local Focus (this paper)*
Variety	+	–
Balance	–	+
Disparity	–	+

Second, the studies relies on the classification of references into WoS categories, which is problematic (Rafols and Leydesdorff, 2009), as has been mentioned. However our sample is big enough to reduce the noise of an inaccurate classification. An article-level classification system might provide a more accurate means of measuring degree of interdisciplinarity (Waldman and van Eck, 2012).

These findings have serious implications for evaluations in developing countries. We conjecture that evaluation exercises that aim for "high impact" in terms of citation counts have the likely perverse consequence of sacrificing IDR that can produce local knowledge, which in turn, could jeopardize the development of local S&T capabilities. This, we suggest is an unintended policy outcome that merits consideration, and calls for a deeper questioning of evaluation methods used in ST&I policies: what kinds of impacts/benefits do policy makers in developing countries expect to obtain from research? What are the objectives of research evaluation exercises for developing countries? What kind of indicators do these objectives require?

Acknowledgments

We acknowledge support from the US National Science Foundation (Award#1064146 - "Revealing Innovation Pathways: Hybrid Science Maps for Technology Assessment and Foresight"). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Adams, J. D. & Clemmons, J. R. (2011). The Role of Search in University Productivity: Inside, Outside, and Interdisciplinary Dimensions. Industrial and Corporate Change, 20(1), 215–251. Retrieved July 31, 2011.

Alatas, S. (1993). On the Indigenization of Academic Discourse. Alternatives: Global, Local, Political, 18(3), 307–338.
Barry, A., Born, G. & Weszkalnys, G. (2008). Logics of Interdisciplinarity. Economy and Society, 37(1), 20–49.
Bones, C., L. Alessa, M. Altaweek, A. Kliskey and R. Lammers (2011). Assessing the Impacts of Local Knowledge and Technology on Climate Change Vulnerability in Remote Communities, International Journal of Environmental Research and Public Health, 8:733-761
Brint, S. (2005). Creating the Future: 'New Directions' in American Research Universities. Minerva, 43(1), 23–50.
Chavarro, D., Orozco, L. & Villaveces, J. (2010). Análisis Del Perfil De Los Grupos De Referencia Del País, in: La investigación en Uniandes. La construcción de una política, (pp. 107–117). Bogotá: Ediciones Uniandes.
Dagnino, R., Thomas, H. & Davyt, A. (1996). El Pensamiento En Ciencia, Tecnología y Sociedad En Latinoamérica: Una Interpretación Política De Su Trayectoria. *Revista Redes*, 3(7), 13–51.
Dodd, R. and Munck, L. (2002). "Dying for change: Poor people's experience of health and ill-health." World Health Organization and World Bank. Geneva and Washington, DC. Available at [http://www.who.int/hdp/publications/dying_change.pdf].
Field, A. (2009). Discovering Statistics Using SPSS: (and Sex and Drugs and Rock 'n' Roll). London: SAGE.
Gahi, R. (2004) Use of local knowledge in impact assessment: Evidence from rural India, *Economic and Political Weekly*, 39(40): 2-8.
Godin, B. (2009). The Making of Statistical Standards: The OECD Frascati Manual and the Accounting Framework, in: The making of science, technology and innovation policy: conceptual frameworks as narratives, 1945-2005, (pp. 67–116). Institut national de la recherche scientifique.
Jacobs, J. A. & Frickel, S. (2009). Interdisciplinarity: A Critical Assessment. Annual Review of Sociology, 35(1), 43–65.
Hair, J., Anderson, R., Tatham, R. & Black, W. (2005). Multivariate Data Analysis. Upper Saddle River N.J.: Prentice Hall.
Keates, J.S. (1996). 'Signs and symbols'. In Understanding maps. Harlow, England : Longman. p.p. 67-83.
Klein, J. T. & Porter, A. L. (1990). Preconditions for interdisciplinary research. In: Birnbaum-more, P., Rossini, F. and Baldwin, D. (eds) (1990). International Research Management. Studies in interdisciplinary methods from business, government, and Academia. Oxford University Press, pp. 11-17.
Larivière, V. & Gingras, Y. (2009). On the Relationship Between Interdisciplinarity and Scientific Impact. Journal of the American Society for Information Science and Technology, n/a–n/a.??
Laudel, G. (2006). Conclave in the Tower of Babel: How Peers Review Interdisciplinary Research Proposals. Research Evaluation, 15(1), 57–68.

Martin, B. (2011) The Research Excellence Framework and the "impact agenda": are we creating a Frankenstein monster? *Research Evaluation*, 20(3): 247-254.
National Academies (Committee on Facilitating Interdisciplinary Research). (2005). Facilitating Interdisciplinary Research. Washington, D.C: The National Academies Press.
Nightingale, P. and A. Scott (2007). Peer review and the relevance gap: ten suggestions for policy makers. Science and Public Policy, 34(8):543-553.
Nowotny, H, Scott, P. & Gibbons, Michael. (2001). Re-thinking Science Knowledge and the Public in an Age of Uncertainty. Cambridge: Polity Press.
Nowotny, H. and Ziman, J. (organizers) (2002). Synopsis of the symposium localized science. Novelty, plurality and narratives. Switzerland, 21st and 22nd January 2002.
Ordóñez-Matamoros, H. G., Cozzens, S. E. & Garcia, M. (2010). International Co-Authorship and Research Team Performance in Colombia. Review of Policy Research, 27(4), 415–431.
Porter, A. L., Cohen, A. S., Roessner, D. & Perreault, M. (2007). Measuring Researcher Interdisciplinarity. Scientometrics, 72(1), 117–147.
Porter, A.L., Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics* 81, 719–745.
Rafols, I. & Leydesdorff, L.(2009). A Global Map of Science Based on the ISI Subject Categories. Journal of the American Society for Information Science and Technology, 60(2), 348–362. Retrieved July 31, 2011,
Rafols, I. & Meyer, M. (2010). Diversity and Network Coherence as Indicators of Interdisciplinarity: Case Studies in Bionanoscience. Scientometrics, 82(2), 263–287.
Rafols, I., Porter, A.L., Leydesdorff, L., 2010. Science Overlay Maps: A New Tool for Research Policy and Library Management. Journal of the American Society for Information Science and Technology 61, 871–1887.
Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P. and Stirling, A.(2012) How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business and Management. *Research Policy*, 41 (7), 1262-1282.
Restrepo, M. & Villegas, J. (2007). Clasificación De Grupos De Investigación Colombianos Aplicando Análisis Envolvente De Datos. Revista Facultad de Ingeniería Universidad de Antioquia, 42, 105–119.
Ruiz, C. F., Bonilla, R., Chavarro, D., Orozco, L. A., Zarama, R. & Polanco, X. (2009). Efficiency Measurement of Research Groups Using Data Envelopment Analysis and Bayesian Networks. Scientometrics, 83(3), 711–721.
Smith, S., V. Ward and A. House (2011). Impact' in the proposals for the UK's Research Excellence Framework: Shifting the boundaries of academic autonomy. Research Policy, 40:1369-1379.

Stirling, A. (2007). A General Framework for Analysing Diversity in Science, Technology and Society. Journal of The Royal Society Interface, 4(15), 707–719.
Stone, D. (2000) Banking on knowledge: the genesis of the Global Development Network. Routledge.
Tilley, C. (1994) A phenomenology of landscape: places, paths, and monuments. Oxford: Berg.
Waldman, L. & van Eck, J (2012). A new methodology for constructing a publication-level classification system of science [online]. CWTS: CWTS Working Paper Series. Available at [http://www.cwts.nl/pdf/CWTS-WP-2012-006.pdf] , accesed 7th of Oct 2012.
Yegros-Yegros, A., Amat, C.B., D'Este, P., Porter, A.L., Rafols, I., 2010. Does interdisciplinary research lead to higher scientific impact? STI Indicators Conference Leiden.

INTERNATIONAL COMPARATIVE STUDY ON NANOFILTRATION MEMBRANE TECHNOLOGY BASED ON RELEVANT PUBLICATIONS AND PATENTS

Lihua Zhai [1], Yuntao Pan [1], Yu Guo [1], Zheng Ma [1], Fei Bi [2]

[1] *zhailh@istic.ac.cn*
Institute of Scientific and Technical Information of China, Beijing (China)

[2] *bifei3446@126.com*
Zhejiang University, Department of Chemical and Biological Engineering, Hangzhou (China)

Abstract
This study adopts a bibliometric approach to quantitatively assessing current research trend on nanofiltration membrane technology, a new type of membrane separation technology widely used in various fields, by using scientific papers published between 1988 and 2011 in journals of all the subject categories of the Science Citation Index and patent data with the same time span from Derwent patent database. Development in basic research and technological innovation on nanofiltration membrane technology is studied. Over the past 24 years, there has been a notable growth trend in publication outputs. Compared with other countries, China has showed a rapid growth, especially in 2000-2011period, and the total number of papers ranks second only after USA in the world. For patents outputs, the rapid growth occurred between 2005-2011. China,USA and Japan ranked top 3 in the world, accounting for 78% of the total number of nanofiltration membrane. But an analysis on the type of patents possessed by the major patentees and their countries shows that, although there are four Chinese institutions in the top 10 patentee list, the main kind of patents from China are application patents, which focus on integrated application of existing nanofiltration membrane, while patents owned by foreign patentees are mostly research patents involving the technology innovation for the nanofiltration membrane itself. Therefore, the research capacity of nanofiltration membrane in China should be further strengthened in order to play a real advantageous role and become international leader in this field.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5)

Introduction
The membrane separation technology, a new separation technology (Zeng YM 2007), emerged in the early 20th century and attracted more and more attention in the 1960s. With the function of separation, concentration, purification and refining, it is widely used in food industry, medicine, water desalination, drinking

water purification, chemical industry, metallurgy, energy, oil processing and other fields, resulting in huge economic and social benefits, and has become one of the most important means of today's separation science (Shi J et al. 2001; Zhu YJ et al. 1997; Keda I et al.1988; Van der Meer WGJ et al. 1997).

In recent years, as one of the common technology to solve the major issues in the field of water resources, energy and environment, membrane separation technology has attached great importance from worldwide, and many countries have had it as a high-tech technology which will be given priority development in the 21st century. In China, the membrane treatment technology is also becoming an important safeguard technology to water quality and safety, energy conservation and clean production.

According to "the 12th Five-Year Special Plan for the high performance membrane materials science and technology development", high performance membrane materials play an important role for a national economic development, industrial technology and the strengthen of the international competitiveness, and to a certain extent, its application level is the reflection of the process industry, energy use and environmental protection for one country.

Among numerous membrane separation technologies, nanofiltration membrane technology has attracted more attention due to its special separation performance. As a new type of separation membrane developed in the early 1980s after the typical reverse osmosis composite membrane, the pore size range of nanofiltration membrane is about lnm(Kong XG 2005), which is between reverse osmosis and ultrafiltration membranes. It has two significant characteristics: Firstly, the molecular weight cutoff (MWCO) is about 200-1000Da, between reverse osmosis and ultrafiltration membranes. Secondly, the surface separation layer of nanofiltration membrane is posed by the polyelectrolyte. Bi F (2011) found that the nanofihration could be used to remove most of the harmful trace organics, and hence to improve the quality of drinking-water and ensure safe drinking-water. Other researchers (Cao M 2011; Han SS 2009; Hou L et al. 2010) also found that nanofiltration was the focus of attention in the field of water treatment and process separation.

Despite the importance and high growth rate of nanofiltration membrane, there have been few attempts to gather data about the worldwide scientific production of nanofiltration membrane. Bibliometric studies provided an accurate and presumably objective method to measure the contribution of a paper to the advancement of knowledge (Huang and Zhao 2008) and had already been widely applied for the scientific production and research trends in many disciplines of science and engineering (M. Zitt and E. Bassecoulard 1994; R. Tang and M. Thelwall 2003; J. Keiser and J. Utzinger 2005). The Science citation index (SCI) from the Web of Science databases is the most widely accepted and frequently used database for analysis of scientific publications (Braun et al. 2000).

The objective of this study is to analyze the status and trends of nanofiltration membrane technology based on relevant publications and patents in the last 24 years in order to help researchers understand the panorama of global

nanofiltration membrane technology research, and provide technical support for science and technology development planning.

Data and methods

Scientific publications related to nanofiltration membrane used in this paper were gathered based on the Scientific Citation Index (SCI) bibliographic database, which was maintained by the Institute of Scientific Information, USA. SCI is the most frequently-used index in scientific output analysis (Kostoff 2000). We performed bibliographic searches and compiled references using an online version of the SCI database. Five search terms, including "nanofiltration membrane、nanofiltration membranes、nanofiltration (NF) membrane、nanofiltration (NF) membranes、Nanometer Filtration membrane" were used to locate publications that contained these words in publication's titles, abstracts, or keyword lists with the time span between 1988 and 2011.

We then retrieved individual document information. As is common in other bibliometric analyses (Ho 2007; Tian et al. 2008; Zhang et al. 2010), research published by authors from England, North Ireland, Scotland, and Wales were labeled as documents originating in the United Kingdom. Although we searched documents published between 1988 and 2011, the earliest publication in the SCI database was published in 1994. Using the above-mentioned searching strategy, a total of 2,195 publications were identified in the SCI database.

The patents were collected based on the Derwent patent database and the same search strategy was used as well as the same time span. A total of 520 patents were identified from the Derwent database.

The analytical methods used in this paper is a data analysis tool software developed by Thomson Reuters (Thomson Data Analyzer), through which statistical and metrological analysis are conducted on the documents and patents data from SCI and Derwent database, and it can help to study the trend of this technology and grasp the distribution of scientific and technological output characteristics deeply from the bibliometric perspective.

Results and discussions

Country analysis of basic research status of nanofiltration membrane technology

Scientific paper is an important output of the basic research in the form of a nation or region (Jiri 2008), through which we can analyze and understand the technical status of basic research in various countries.

According to the statistical results of the SCI database, the numbers of papers related to nanofiltration membrane technology between 1994-2011 were 2195, which were distributed in 68 countries. The top 10 countries and their published papers are shown in table 1.

The productivity ranking of countries was headed by USA, which was responsible for the most number (428). China published the second highest number of papers (279), followed by France (182). The literatures of other countries ranged from 90

to 200. Five of top 10 countries were from G7, while only two were emerging countries. The pattern of domination in publication of the G7 has occurred in most scientific fields (Suk et al. 2011), reflecting the high economy activity and academic level of these countries (Yang et al. 2012).

Table1. Top 10 countries for papers

Rank	*Country*	*The number of papers*	*The number of cooperating countries*
1	USA	428	25
2	China	279	16
3	France	182	18
4	UK	121	23
5	South Korea	117	17
6	Japan	104	14
7	Spain	100	22
8	India	93	13
9	Netherlands	93	14
10	Canada	92	17

The time distribution of papers in the countries which published more than 100 papers was studied and the result was shown in Fig 1. The number of papers in the United States was higher than that in other countries, and maintained a high growth rate. For China, there were three time stages for these papers: before 2000, the number of the papers related to nanofiltration membrane was lower than most of other countries. From 2000 to 2005, the number started to increase, yet remained a low level. After 2005, the growth rate became faster and faster, with a high rate of linear growth, the number had reached a higher level and rank the second in 2005. In 2010, the number had exceeded that in the United States.

Research on nanofiltration membrane began in the 1970s, originally developed as the anti-permeable membrane, early called "loose reverse osmosis membrane (Loose the Reverse Osemosis Membrane). In the 1990s, the concept of nanofiltration membrane was finally formed (Faleshi M. 2001). Therefore papers related to nanofiltration membrane appeared in 1994 for the first time and increased year by year. The total growth rate of papers in the United States and China were ahead of other countries, showing a more prominent in this field. Nanofiltration technology is mainly used in the field of bio-chemical, food, and water treatment and other fields, especially in the field of water treatment. Nanofiltration membrane technology has an important position in membrane separation industry; moreover, it will also have a greater impact on the water treatment industry.

The situation that more and more papers related to nanofiltration membrane appeared year by year in the United States and China make it clear that this technology in the two countries had attracted greater attention, and therefore more

research output were gained, which laid a solid foundation for the development of nanofiltration membrane applications industry.

All in all, although the research on nanofiltration membrane in China had a relatively late start, it had a faster development. Corresponding, the fast development of basic research brought China great potential in the field of nanofiltration membrane applications. The innovation of nanofiltration technology then would be studied by analyzing the related patents to discuss the application of nanofiltration membrane.

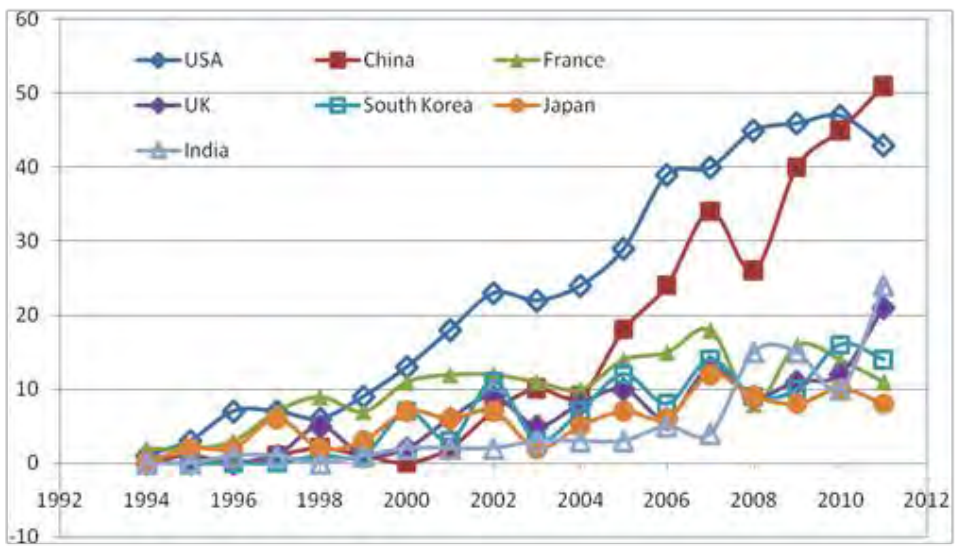


Fig1. Changes of nanofiltration membrane technology over time

Country analysis of nanofiltration membrane technology innovation

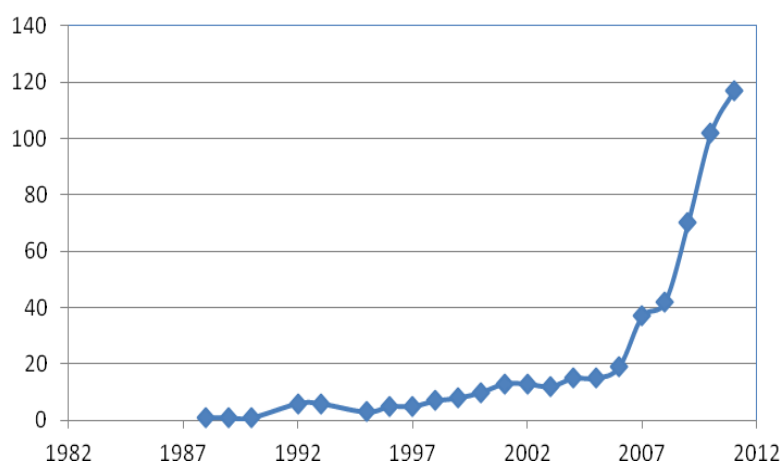


Fig2. The time changes of international nanofiltration membrane technology patents

According to the statistical results of the Derwent database, patents related to nanofiltration membrane technology between 1988-2011 reached the number of 520, which were distributed in 37 countries. The top 10 countries and their patent data are shown in table 2. And the time trends of these patents are shown in Fig.2. For these nanofiltration membrane technology patents from all countries, there were two time stages. From 1988 to 2005, on the international level, the annual number of patents was no more than 20. After 2005, patents increased quickly and reached 120 in 2011, which was six times than that in 2005.

Table2. Top 10 countries for patents

Rank	*Patent priority country*	*The number of patents*
1	China	250
2	USA	93
3	Japan	63
4	South Korea	31
5	France	19
6	Germany	17
7	European Patent Organization	14
8	Canada	11
9	Australia	9
10	UK	7

Requirements for Nanofiltration membrane technology application will become more and more highly as the market continues to expand, which will also promote an increase in the number of patents. Therefore, the rapid growth of Nanofiltration membrane patents was probably accorded with the prediction that membrane technology had become the most widely technology of wastewater treatment (Ortega et al. 2007; Renou et al. 2008).

According to the statistical results of patent priority country, the number of patents in China was up to 250, with the first place, followed by the United States, almost 100 patents. Japan ranked third in the number of patents (63). The total number of patents of the top three in the country rank reached 406, accounting for 78% of the total patents in the field of nanofiltration membrane. Among patent priority countries, China accounted for the largest proportion which showed that China had a strong technical innovation ability.

The patentee is the owner of the patent collectively. That is to say, when the patent application is approved, the one who conduct the applicant and to be granted a patent is the patentee. The patentee can be not only agencies but also an individual who may be the important innovation forces in the market (Bessen J. 2008).

520 patents belonged to more than 600 patentees, 15% of which had more than a patent. Therefore, this study focused on the more important patentee and their countries. The top 10 patentee who have the most number of patents are shown in

table 3. Data shown in table 3 can reflect the concentration situation of the patents, for examples, which agency or which person has the most number of patents.

Table3. Nanofiltration membrane technology patentee and patent

Rank	*Patentee*	*The number of patent*	*The number of research patent*	*The number of application patent*	*Country*
1	Toray industries, Inc	19	9	10	Japan
2	GE, USA	13	6	7	USA
3	Zhe Jiang University	10	6	4	China
4	Nan Jing Zelang Medical Technology Co.Ltd	8	0	8	China
5	Organo Corp	8	7	1	Japan
6	DOW	7	6	1	USA
7	Jin Brand Co.Ltd	6	0	6	China
8	Hangzhou water treatment technology development centre	5	3	2	China
9	Kurita water industries,Inc	5	3	2	Japan
10	Akzo Nobel	4	2	2	Netherlands

From the national level, China and Japan were countries that centralized more nanofiltration membrane technology, and China accounted for 4, Japan accounted for 3 of top 10 countries.

With respect to the type of agency, there were more companies than research institutions and university in top 10 patentees. In addition to the Zhejiang University and Hangzhou Water Treatment Technology Development Center, the others were companies or enterprises, which explained the dominant position in the technology innovation of nanofiltration membrane for companies and enterprises.

Patents owned by the major patentees were studied deeply and divided into two kinds of patents: research patent and application patent. The former was real technology innovation, including the development of nanofiltration membrane having certain characteristics, and improving some of the characteristics of nanofiltration membranes and nanofiltration membrane material technology invention. While application patent was mainly used in the processing equipment of related industry, such as various types of wastewater treatment process, preparation process of some of the compounds and drugs, and processing methods in the relevant industry. Focus of these patents was synthesized through technology to meet the comprehensive needs of the market, but it had less effect on nanofiltration membrane technology innovation itself.

According to the data result of table 3, the main patentees had not only research patents but also application patents. But for the two Chinese companies, ranking

the fourth and seventh respectively, the patents were application patent. Both Organo Corp and Dow had an outstanding performance in research patent. More applications of their products could explain above observation (Uzal N et al. 2010).

The relationship between the basic research and technological innovation activities of nanofiltration membrane

Comparative analysis of international papers and patent of nanofiltration membrane technology is shown in Fig 3, from which the development of nanofiltration membrane technology can be divided into three stages. From 1988 to 1995, patents about nanofiltration membrane started to appear, while not did related papers, which indicated that the nanofiltration membrane technology originated in the application field and the innovation activities of nanofiltration membrane had a strong market-oriented, so this stage was called technical exploratory stage. During 1995-2005, the number of patents didn't change much and maintained at a relatively low level, while the scientific papers showed the obvious linear growth trend, which indicated that basic research was at the leading edge, so called accumulation stage of basic research. From 2005 to 2011, the number of scientific papers and patents showed faster growth, especially in the number of patents, an increase of more than 2 times, so the third stage was called technology leap stage. The explanation for this might be that basic research had significant impact on technology innovation (Szu-chia S. Lo 2010).

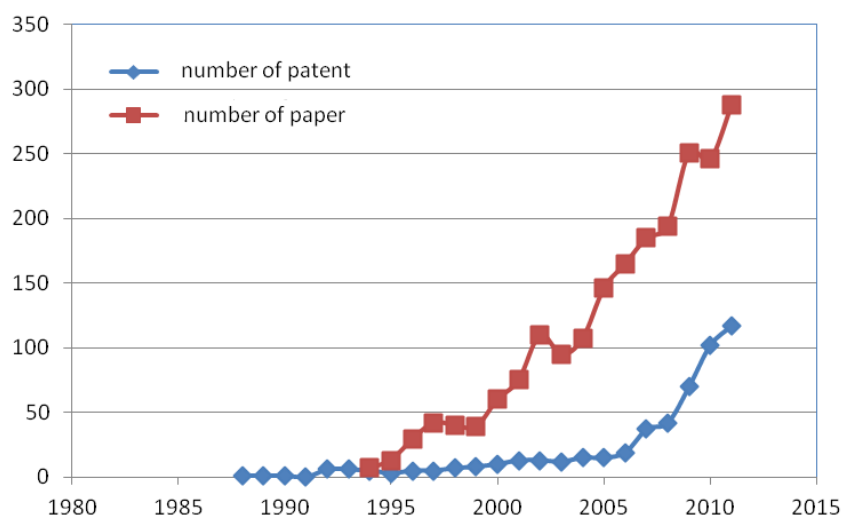


Fig3. Comparison of the number of papers and patents of international nanofiltration membrane technology

Comparison between papers and patents related to nanofiltration membrane technology of China is shown in Fig 4. Papers of nanofiltration membrane

appeared earlier than patents in 1994, while patents appeared first in 2000. After 1995, the number of nanofiltration membranes papers increased over time. The number of patents started to increase fast in 2005 and maintained a rapid growth rate. Until 2009, the number of patents began to exceed that of papers. It can be seen that the law of development in China is very different from that of other countries, taking into account above three stages. In China, Nanofiltration membrane technology started from basic research then arrived at the stage of the accumulation of basic research, with no access to experience technical exploration stage. In 2005, the rapid growth in the number of patents made the development of nanofiltration membrane enter the technological leap stage.

Development of nanofiltration membrane technology is closely related to the development of its applications. In China, nanofiltration membranes are of great concern in the field of water purification and water treatment due to its high efficiency separation characteristics. The development of these industries also has great effect on the development of nanofiltration membranes technology. Starting from 2000, China, while maintaining rapid economic growth, began to explore the economic, social, energy, environmental sustainability coordination mode of development, thus increased environmental governance, and the introduction of a series of guiding policies and safeguards.

In 2007, China formulated and issued a new national standard for drinking water. The number of indicators in the new national standard increased from 35 to 106, essentially flat with the world's most stringent EU water quality standards. These measures are a strong impetus to the research and development of new technologies in the field of water treatment. Among these new technologies, nanofiltration membrane technology is attracting more and more attention because of the quality of the separation efficacy. According to the above analysis, the fact that research on nanofiltration membrane of China has experienced from the accumulation stage of basic research to technology leaps has a closer relationship with national sustainable development strategies and related standards proposed. More and more market demand for nanofiltration membrane made a strong impetus to the nanofiltration membrane science and technology capabilities, and the number of papers and patents showed a rapid growth trend.

It should be noted that, although China was in the leading position not only from the technical scale but from technical concentration, the market of nanofiltration membrane in China showed a poor performance compared that in other countries (Yang YQ 2011). Foreign enterprises occupied a larger market share, in particular, Dow Chemical and Japanese companies.

According to the results obtained from table 3, the patents owned by two Chinese companies were all application patents which focused on integrated application of existing nanofiltration membrane, rather than the film itself technological innovation. Patents of other two institutions in China belonged to both nanofiltration membrane technology invention and application technology. But technological innovation may not be transformated fully into market application for they are just research institutions.

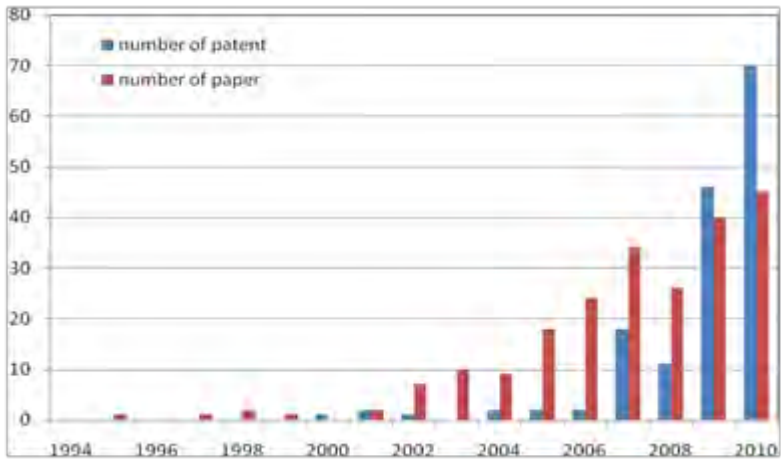


Fig4. Comparison the number of patents and papers of Chinese nanofiltration membrane

Patentees from other countries were all companies according to table 3 which were not only committed to the technical synthesis, but paid more attention to the nanofiltration membrane technology innovation. For example, six of the seven patents owned by Dow Chemical were about the technology innovation for nanofiltration membrane itself. As the main part in the market, these companies have showed an outstanding performance in the industry for they have the innovative technology.

Therefore, although having great potential in innovation technology of nanofiltration membrane according to the patent data, China still needs to strengthen the research capacity to accelerate the process from research into technology.

Conclusion

Nanofiltration membrane technology, a good separation technology, has been widely used in many fields, especially in the field of water treatment. With increasing international attention, the number of research papers and patents is increasing year and year. Overall, the development of nanofiltration membrane technology has the characteristics of the technology-oriented. The earliest scientific and technological output was in the form of patent. Then the scientific papers played a larger role. After a period of basic research Accumulated, the number of patents began to rapidly increase. At present, the technical innovation is still in a rapid rise.

Nanofiltration membrane technology research in China started late, but developed rapidly. With the proportion of energy binding emission reduction targets and the upgrade of the water quality standards, the value of the nanofiltration membrane applications became more apparent and the related research papers increased. At present, the number of Chinese patents related to nanofiltration membrane

technology is in top line in the world, as well as the number of Chinese papers. Whether from basic research or from the technological innovation, the scientific output of China in the nanofiltration membrane technology has strong international advantage, especially in technological innovation.

Nanofiltration membrane technology is eventually a practical technology. Although there is a large potential in the technology innovation in China, the research capacity should be strengthened and the technology transformation process should be accelerated in research institutions in order to play a real advantage role and obtain international leader in this field. By continuing to promote the development of nanofiltration membrane industry, the technology will become innovative technology in many areas, particularly in the water treatment. The leading research will lead to a leading in nanofiltration membrane industry eventually.

Acknowledgments

The research reported here were supported by the Natural Science Foundation of China (No. 70973118) and the National High Technology Research and Development Program （863）of China (No. 2011AA01A206).

References:

Bessen J (2008). The value of US patents by owner and patent characteristics. *Research Policy*, 33(5), 932-945.

Bi Fei, Chen Huanlin, Gao Congjie. (2011). Advances on trace organics removal from drinking-water by nanofiltration. *Modern Chemical Industry*, 31(7), 21-26.

Braun, T., Schubert, A. P., et a1. (2000). Growth and trends of fullerene research as reflected in its journal literature. *Chemical Reviews*, 100(1), 23-37.

Cao Ming. (2011). The Study and Application of Nanofiltration Membrane. *Guangzhou Chemical Industry*, 39(18), 13-14.

Faleshi M. (2001). Progress in membrane science and technology for seawater desalination-a review. *Desalintion*, 134, 47.

Han Shasha, Liu Baoping.(2009). Application of nanofiltration membrane technology in water treatment. *Anhui Chemical Industry*, 3, 7-8.

Hou Li'an, Liu Xiaofang. (2010). Research progress and development prospects of nanofiltration membrane technology to water treatment. *Membrane Science and Technology*, 4, 1-7.

Ho Y.S . (2007). Bibliometric analysis of adsorption technology in environmental science. *J Environ Prot Sc*, 1, 1–11.

Huang, Y., & Zhao, X. (2008). Trends of DDT research during the period of 1991 to 2005. *Scientometrics*, 75(1), 111-122.

Jiri Vanecek (2008). Bibliometric analysis of the Czech research publications from 1994 to 2005. *Scientometrics*, 77(2), 345-360.

J. Keiser and J. Utzinger (2005). Trends in the core literature on tropical medicine: A bibliometric analysis from 1952–2002. *Scientometrics*, 62, 351-365.
Keda I, Nakano T, Ito H, et a1. (1988). New composite changed revise osmosismembrane. *Desalination*, 68, 109-119.
Kong Xiangguo.(2005). Discussion several common methods for the water treatment. *Sci-Tech Information Development & Economy*, 15(20), 295-296.
Kostoff .R.N (2000). The under publishing of science and technology results. *The Scientist*, 14, 6.
M. Zitt and E. Bassecoulard (1994). Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. *Scientometrics*, 30, 333-351.
Ortega, L. M., Lebrun, R., et a1. (2007). Treatment of an acidic leachate containing metal ions by nanofiltration membranes. *Separation and Purification Technology*, 54, 306-314.
Renou, S., Givaudan, J. G., Poulain, S., et a1. (2008). Landfill leachate treatment: Review and opportunity. *Journal of Hazardous Materials*, 150, 468-493.
R. Tang and M. Thelwall (2003). US academic departmental Web-site interlinking in the United States disciplinary differences. *Libr. Infor. Sci. Res.*, 25, 437-458.
Shi Jun, Yuan Quan, Gao Congjie. (2001). *Handbook of membrane technology*. Beijing: Chemical Industry Press.
Suk, F. M., Lien, et a1. (2011). Global trends in Helicobacter pylori research from 1991 to 2008 analyzed with the Science citation index expanded. *European Journal of Gastroenterology and Hepatology*, 23(4), 295-301.
Szu-chia S. Lo (2010). Scientific linkage of science research and technology development: a case of genetic engineering research. *Scientometrics*, 82, 109-120.
Tian YG, Wen C, Hong S (2008). Global scientific production on GIS research by bibliometric analysis from1997–2006. *J Informetr*, 2, 65–74.
Uzal N, Yilmaz L, Yetis U (2010). Nanofiltration and Reverse Osmosis for Reuse of Indigo Dye Rinsing Waters. *Separation Science and Technology*, 45(3), 331-338.
Van der Meer WGJ, van Dijk JC. (1997). Theoretical optimization of spiral-wound and capillary nanofihrationmodel. *Desalination*, 113, 129-146.
Yang, L. Y., Yue, T., et a1. (2012). A comparison of disciplinary structure in science between the G7 and the BRIC countries by bibliometric methods. *Scientometrics*, 93(2), 497-516.
Yang Yuqin.(2011). The Research and Market Progress of Nanofiltration Membrane. *Information Recording Materials*, 12(6), 11-19.
Zeng Yiming. (2007). *Membrane bioreactor technology*. Beijing: National Defense Industry Press.

Zhang L, Wang MH, Hu J (2010). A review of published wetland research, 1991–2008: ecological engineering and ecosystem restoration. *Ecol Eng*, 36,973–980.
Zhu Yujun, Sun Fenghui. (1997). Study of factors affecting separation performance of nanofiltration membrane. *Technology of Water Treatment*, 32(2), 78-82.

IN-TEXT AUTHOR CITATION ANALYSIS: AN INITIAL TEST (RIP)

Dangzhi Zhao[1] and Andreas Strotmann[2]

[1] *dangzhi.zhao@ualberta.ca*
School of Library and Information Studies, University of Alberta, Edmonton T6G 2J4 (Canada)

[2] *andreas.strotmann@gesis.org*
gesis – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne (Germany)

Abstract
In-text author citation analysis refers to author-based citation analysis using in-text citation data from full-text papers rather than reference data from citation databases. In-text author citation analysis has the potential to support more refined author citation and co-citation counting for improved citation analysis results and to help with the application of citation analysis to research fields such as the social sciences that are not covered well by citation databases. This work in progress reports results from an initial test on how well in-text author citation analysis works as compared to traditional author citation analysis.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Visualisation and Science Mapping: Tools, Methods and Applications (Topic 8).

Introduction
Problems with citation databases (Web of Science, Scopus) have been one of the major sources for criticisms of citation analysis as it has relied heavily on these databases. For example, the journal-only coverage has limited the usefulness of citation analysis of research fields where conference proceedings or books are as important as journals (e.g., computer science, social sciences and humanities); a first-author-only practice for indexing cited authors (Web of Science) has limited application of citation analysis in highly collaborative research fields; coverage bias against non-English publications has made cross-country comparisons difficult; and insufficient coverage of journals in research fields like the social sciences and humanities has made citation analysis unreliable in these research fields.

Although some limitations (e.g., in indexing or download) may be worked around by researchers (Strotmann & Zhao, 2010; Zhao & Strotmann, 2011), fixing problems in coverage of these databases is completely at the mercy of the

companies that run these commercial databases. It is therefore very important to find alternative methods and data sources that may alleviate near-complete reliance on these databases in bibliometrics.

Another source for criticisms of citation analysis is that current citation and co-citation counting methods treat all citations equally and do not take into account how heavily a cited work is actually used in the citing work, or where in a work it is cited. It is important to study refined counting methods that weigh citations based on their frequency and location for improved citation analysis results.

The present study attempts to contribute to these areas of research by testing *in-text author citation analysis*, which collects citation data from the full text of articles or books rather than from citation databases. This study is inspired by Strotmann and Bleier's (2013) author co-mention analysis, which combines the basic ideas of co-word analysis and author co-citation analysis as a way to address problems with the application of bibliometrics to the international social sciences, and extends the document co-mention idea of Rosengren (1968) to author-based analysis. By relying exclusively on in-text citations, the present study also differs from methods recently introduced by Boyack, Small, and Klavans (2012), where classic citation data sources are primarily used, augmented by in-text citation distances.

Feasibility, benefits and limitations of in-text author citation analysis

Scholarly writing requires that the author of an article (or a book) cite relevant works in the text where they are referred to and list the details of the cited works in the reference list at the end of the article. There are a number of standard citation styles that specify the details about how in-text citations and reference lists should be done, and each citation style has been adopted by one or more scholarly communities. The APA style specified in the Publication Manual of the American Psychological Association, for example, requires that all citations in the text should be placed in parentheses inside which the last names of up to three authors of each cited work are listed along with its publication year. Alternatively, the authors' last names are listed in the text, followed by a year in parentheses. If more than one work by the same author(s) published in the same year is cited, these works are differentiated by adding lower-case letters to the publication year. All this is done consistently throughout an article.

As scientific communication has moved to electronic publishing, journal articles and books are now available in full text. For in-text citations clearly delimited by parentheses and following a set of prescribed rules, automatic identification and extraction of in-text citations from full text and the parsing of author names and years from these citations do not have the complex problems that co-mention analysis in text mining research has to deal with (Strotmann & Bleier, 2013), and can therefore be quite easy and accurate.

The APA citation style is used not only by the Psychology community but also by a number of other scholarly communities especially in the social sciences, such as Linguistics, Sociology, Economics, Criminology, Business, Education, Nursing, or Library and information studies. Other widely used citation styles such as Chicago are very similar to APA in terms of format for in-text citations. In-text citation analysis may therefore extend citation analysis to these research fields, most of which have long been covered insufficiently by citation databases.

In-text citations in the author-year format, delimited by parentheses, can support not only all of the author citation and co-citation counting methods that have been used in citation analysis, but also some refined counting methods that are at least in theory improvements of traditional methods.

The number of citations an author receives from a set of articles is currently calculated in two ways: (a) as the number of papers in this set of articles that lists one or more of this author's works in their reference lists, or (b) as the total number of this author's works that appear in the reference lists of this set of articles. For example, if two works published by author A are cited by article X and three by article Y, the number of citations A receives from X and Y is two using method (a) and five using method (b).

The co-citation count between two authors is traditionally the number of papers that list at least one article from each author's oeuvre in the same reference list. For example, if articles X and Y above also cite one and two articles written by author B respectively, the co-citation count between A and B is two, i.e., two articles (X and Y) that cite them together, and has nothing to do with how many works by A and B are actually cited how heavily in X and Y.

When calculating citation and co-citation counts using in-text citation data, in-text citation strings in the format of author-year are first identified and extracted from each citing paper (or book). All in-text citation strings of a citing paper can be combined into a single long string. Whenever this long string contains an author's name, this citing paper would contribute one to this author's citation count defined in (a). If this long string contains the last names of both authors A and B, this citing paper contributes one to the co-citation count of authors A and B.
Citation counts as defined in (b) are more difficult to calculate as they require identification of each cited paper rather than just each cited author by looking up all author names listed in each of these strings for each cited work along with the publication year in the full-text reference list.

None of these traditional citation and co-citation counting methods takes into account how many times or how heavily an article is used in the citing article or where in the citing article it is used. In other words, all citations are treated equally in these methods, which has been another source of criticisms of citation

analysis. Some articles are real inspirations for the work being developed and are therefore referred to specifically many times in many of the major sections including methodology and discussion. Other articles are simply mentioned once along with many others in the literature review section of the citing article. Researchers often need to weigh if they should cite an article at all in the latter case. It is clearly problematic not to treat the real inspirations for research "better" when using citation counts to measure research impact.

Using in-text citation data, the location or frequency of each in-text citation can be recorded and counted, and the resulting information can be used to weigh citations or co-citations. Such weighted citation and co-citation counts may lead to better measures for author impact or relatedness. For example, it is relatively easy to calculate a rough citation count weighted by citation frequency using in-text citation data as the number of an author name's total appearances in a citing paper's long in-text citation string summed over all citing papers, as we do below. Weighted co-citation counting is also feasible along the lines proposed for counting author bibliographic coupling frequencies (Zhao & Strotmann, 2008).

Regarding the limitations of the proposed method, (1) in-text citation analysis is limited to author citation and co-citation analysis of research fields where standard APA-like citation styles are used, and would require significantly more sophisticated techniques to be used for document or journal-based citation analysis, for bibliographic coupling analysis, or for the study of research fields that use numbers in superscripts or brackets as in-text citations to link to the numbered references at the end of the articles. (2) In-text citation analysis may not work well for highly collaborative research fields as only up to 3-5 authors of each cited work are available in in-text citation data, although this may be better than using Web of Science data which only indexes first authors. (3) In-text citation analysis is also more sensitive to author name ambiguity problems than citation counting using citation databases because only last names are available in in-text citation data. In-text citation analysis therefore has higher requirements for author name disambiguation.

Initial test of in-text author citation analysis

Research questions

We conducted an initial test of the feasibility, usefulness and limitations of in-text author citation analysis. This test aims to address the following questions.

- How do author rankings by citations compare between in-text citation data and citation data obtained from citation databases?
- How do author co-citation matrices compare between in-text citation data and citation data from citation databases?

- How do citation and co-citation counts weighted by occurrence frequency compare with unweighted counts?

Data collection and analysis

We downloaded two datasets: one consisting of the full text of all articles published in JASIST 2009-2011, and the other of all full records from a search in Scopus for JASIST 2009-2011 restricted to articles and reviews. The two datasets are comparable in size: 564 full text articles and 565 Scopus records.

We used Linux' pdftotext to convert the PDFs to Unicode text files before extracting the in-text references from each file using a relatively large but straightforward Python regular expression. The Scopus records were processed by our own Python scripts to extract the information relevant for author citation and co-citation analysis.

From each of the datasets, we calculated citation counts for all cited authors, selected the top 500 authors ranked by these citation counts, and calculated co-citation counts for these 500 authors. For the in-text citation dataset, we also calculated citation and co-citation counts weighted by their frequencies in the full text citing articles.

First-author-based counting was used here, which means that only the first author of each cited work is counted towards citation counts and two authors are counted as being co-cited whenever they both appear as the first in one of the in-text citation strings of a citing paper. Counting all authors of each cited work provided in the in-text citation data can be tricky because all authors are listed for works that have up to three authors but only the first author is provided if a work has more than three authors. This should not be a serious problem for most of the research fields that use APA style as the collaboration level in those fields is normally low. For a diverse field like information science, however, counting all authors listed in in-text citation data may introduce a systematic bias against areas of research (e.g., those closely related to computer science) where large-group collaborations happen regularly.

We did not perform author name disambiguation on the data for this initial test despite being well aware of author name ambiguity problems in citation analysis (Strotmann & Zhao, 2012). Instead, we manually extracted a subset of 60 highly cited author last names common to all the analyses which (a) did not contain non-ASCII Unicode characters, (b) consisted only of a single word, and (c) were not at first blush extremely common last names. Table 1 lists the names of authors (including their first name initials) that we used in this test; for in-text citation analysis, only their last names were actually used.

Results, discussion and conclusion

Here we report results from comparisons of rankings and mappings of 60 authors. These authors are top ranked authors by first-author-based citation counts from Scopus data after removing authors with highly ambiguous Chinese and Korean names (Strotmann & Zhao, 2012) or names that our current computer programs does not handle well yet (e.g., composite last names, or names with non-ASCII Unicode characters).

Table 1 presents rankings of these 60 authors by three counting methods: simple counting based on Scopus data (Scopus), simple counting based on in-text citation data (InText Simple), and weighted counting based on in-text citation data (InText Weighted).

Table 1. Rankings of 60 authors by three counting methods

Author	*Scopus*	*InText Simple*	*InText Weighted*	*Author*	*Scopus*	*InText Simple*	*InText Weighted*
Leydesdorff L	1	1	1	Brin S	31	34	49
Garfield E	2	11	11	Dumais S	32	57	55
Salton G	3	3	15	Vakkari P	33	28	19
Cronin B	4	4	5	Case D	34	54	52
Egghe L	5	7	3	Fidel R	35	37	46
Moed H	6	20	23	Porter M	36	22	27
Small H	7	10	4	Voorhees E	37	27	34
Hirsch J	8	8	10	Wasserman S	38	36	30
White H	9	2	2	Zitt M	39	52	39
Jansen B	10	13	6	Rieh S	40	38	21
Newman M	11	9	7	Savolainen R	41	44	42
Bornmann L	12	17	22	Aksnes D	42	53	51
Spink A	13	16	13	Bensman S	43	41	29
Marchionini G	14	12	20	Davis F	44	5	18
Wilson T	15	6	9	Hearst M	45	55	53
Ingwersen P	16	19	14	Meho L	46	58	40
Saracevic T	17	14	12	Borlund P	47	47	47
Borgman C	18	24	31	Ellis D	48	39	35
Bates M	19	15	8	Harter S	49	42	56
Rousseau R	20	31	41	Kleinberg J	50	49	54
Merton R	21	23	26	Narin F	51	43	57
Schubert A	22	25	45	Sebastiani F	52	50	36
Manning C	23	30	48	Seglen P	53	45	58
Thelwall M	24	26	17	Stvilia B	54	51	38
Belkin N	25	18	16	Braun T	55	59	59
Kuhlthau C	26	32	32	Golder S	56	48	50
Robertson S	27	21	33	McCain K	57	40	25
Boyack K	28	56	43	Rogers E	58	35	24
Joachims T	29	29	44	Wagner C	59	33	28
Bollen J	30	60	60	Watts D	60	46	37

The Pearson's r value between the two author rankings by unweighted citation counts from Scopus data and from in-text citation data is 0.81. A visual inspection of the rankings suggested that ambiguous author names may have been a major source of discrepancies contributing to the unexpectedly low r value. We therefore recalculated Pearson's r value between these rankings after removing author names that we knew to be fairly (although not extremely) common, such as White, Wilson, Davis, and Newman, and obtained an r value of 0.93, confirming this suspicion.

Surprisingly, Garfield's positions differ significantly between the rankings, even though his last name is close to unique in this field. On closer examination, we find that Garfield's name appears in two almost equally common forms in the full ranking, one spelling his name the obvious way with separate *f* and *i* characters, and the other, unexpected one, spelling it with an *fi* ligature Unicode character. Unicode normalization following text extraction from the downloaded PDFs might be able to fix this problem relatively easily – in previous author co-citation studies, we used this method with some success.

It thus appears that in-text citation counts calculated with our first quick and simple computer programs are already quite comparable with citation counts from Scopus data. As expected, author name ambiguity problems are more serious with in-text citation data than with Scopus data, however. We therefore expect that in-text citation analysis can work well if (and only if) author name disambiguation is performed reasonably well.

Comparable results are also seen from co-citation counts between Scopus data and in-text citation data: the two unweighted co-citation matrices, which ideally should be identical, show a vector cosine similarity measure of 0.91. (This is calculated from the upper triangle matrices of the co-citation matrices - without diagonal values - from the vectors obtained by concatenating the component row vectors of each matrix.)

An initial examination of factor analysis results shows that the groupings resulting from these two co-citation matrices are the same for most of the 60 authors representing five major specialties in IS: IR systems, Users and their interaction with IR systems, Evaluative bibliometrics, Relational bibliometrics, and Web science. Detailed analysis of these results is to be completed, but we have the impression that the few authors that are placed in different specialties by factor analysis of these two co-citation matrices are mostly those with common names, such as Rogers, Wagner, or Davis. These authors are grouped into a separate factor when using in-text citation data, as names that correspond to many different individuals apparently tend to do (Strotmann & Zhao, 2012), but they are placed into their own respective speciality factors based on Scopus data (e.g., Rogers into

Web science and Wagner into Relational bibliometrics), which disambiguates these names to some extent via first initials.

We tentatively conclude that in-text author citation analysis (and especially co-citation analysis) can work well *if* author name disambiguation is performed properly. Our ad-hoc approximation to this by removing "obviously" problematic names from the analysis introduces a bias (a) against non-British European names (e.g., Börner), (b) against authors from cultures with frequent compound last names (e.g., Bar Ilan, van Leeuwen), and (c) against with Chinese or Korean last names (e.g., Chen, Park). While this may be admissible in the present context, where we merely test similarity of performance under the assumption that the author name ambiguity problem can be resolved sufficiently, such biases would likely invalidate actual applications of the method tested here whenever this problem is not addressed appropriately.

References

Boyack, K.W., Small, H., & Klavans, R. (2012). Improving the accuracy of co-citation clustering using full text. Proceedings of the 17th International Conference on Science and Technology Indicators, Montreal, Quebec, Canada, 5-8 September, 2012.

Rosengren, K.E. (1968). *Sociological aspects of the literary system*. Stockholm: Natur och Kultur.

Strotmann, A., & Bleier, A. (submitted). Author Name Co-Mention Analysis: Testing a poor man's author co-citation analysis method. Submitted to *14th International Society for Scientometrics and Informetrics Conference*, 2013.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, 63(9), 1820-1833.

Strotmann, A., Zhao, D., & Bubela, T. (2010). Combining commercial and Open Access citation databases to delimit highly interdisciplinary research fields for citation analysis studies. *Journal of Informetrics, 4*(2), 194-200.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in Information Science 1996-2005: Introducing author bibliographic coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070-2086.

Zhao, D., & Strotmann, A. (2011). Counting first, last, or all authors in citation analysis: A comprehensive comparison in the highly collaborative Stem Cell research field. *Journal of the American Society for Information Science and Technology*, 62(4), 654-676.

KNOWLEDGE CAPTURE MECHANISMS IN BIOVENTURE CORPORATIONS: A CASE STUDY

Thomas Gurney[1], Antoine Schoen[2], Edwin Horlings[3], Koichi Sumikura[4], Patricia Laurens[5], Peter van den Besselaar[6], and Daniel Pardo[7]

[1] *t.gurney@rathenau.nl,* [3] *e.horlings@rathenau.nl*
Rathenau Institute, Science System Assessment, Anna van Saksenlaan 51, 2593 HW, The Hague (The Netherlands)

[2] *a.schoen@esiee.fr,* [5] *laurensp@esiee.fr*
Université Paris-Est, ESIEE – LATTS – IFRIS, 2, bd Blaise Pascal, Noisy le Grand, 93160 (France)

4 sumikura@grips.ac.jp
GRIPS - National Graduate Institute for Policy Studies, 7-22-1 Roppongi, Minato-ku, Tokyo 106-867, (Japan)

6 p.a.a.vanden.besselaar@vu.nl
VU University Amsterdam, Network Institute & Department of Organization Science, De Boelelaan 1105, Amsterdam (The Netherlands)

7 daniel.pardo@wanadoo.fr
CNRS - Aix-Marseille Université, LEST UMR 7317, 35 avenue Jules Ferry, 13626 Aix en Provence Cedex 01 (France)

Abstract
Mechanisms of knowledge transfer from academia to industry have long been debated. The knowledge inputs required may stem from research conducted many years prior to a technology being adopted and adapted by industry, and a supporting base of knowledge is required to facilitate this. In this case study we utilise the publishing and patenting history of an individual scientist, and link their output to the technologies with which the scientist is involved. A detailed description of knowledge sources of these technologies is discussed, including the role absorptive capacity plays in priming their development. This study addresses the contributions of the researcher, particularly in relation to the contributions of their academic and industrial co-authors and co-inventors. We find clear linkages, and varied degrees of knowledge transformation, between the technologies in their present form and long-past outputs of the individual, via the publications of the inventor and the literature cited by the patent applications. We also find that the individual demonstrates a high level of absorptive capacity, incorporating and adapting exogenous knowledge into their own knowledge base.

Conference Topic
Technology and Innovation Including Patent Analysis (Topic 5).
Collaboration Studies and Network Analysis (Topic 6)

Introduction

In innovation research, analyses have encompassed various levels of aggregation and address different aspects. For analyses concerning knowledge transfer mechanisms, when examining the minutiae of mechanisms and mediums (such as those of tacit or codified knowledge, R&D networks, formal or informal collaborations), difficulties arise. These difficulties stem from enormous complexities of the knowledge involved in the science and related technologies. The end technological object is the result of the knowledge input and accretion over time into a coherent, and critical, mass. We elaborate upon a method by Gurney *et al* (2012) to discern the knowledge contributions of a specific inventor/author to a patent corpus and the technologies they represent. We utilise two of the output indicators typically used in this and other studies, those of patents and publications. The concepts and practices embodied and codified in the publications and patents were linked to each other, through the citations to literature found in the patent documents. Through linking the two corpora of knowledge the actual knowledge contributions to the development of an idea from inception to product were demonstrated.

The core of this paper discusses the multiple aspects of absorptive capacity, knowledge transfer and transformation, including how scientific knowledge is incorporated into practices, skill sets and eventually artefacts. We then discuss the context and history of our test case. Following this, we briefly summarise the methodology, along with descriptions of the indicators we use followed by the visualisation and clustering techniques employed in our analysis. Our results and conclusions follow, ended with our discussion and implications for further analyses and policy.

Conceptual Framework

The most common and widely cited knowledge transfer mechanisms and inputs are patents, publications, informal and formal interactions, personnel hiring, licensing, R&D collaborations, contract R&D and consulting (Cohen, W.M. et al., 2002). With each of these mechanisms the medium of knowledge transfer can be either codified (such as, for example, patents and publications) or tacit (such as, for example, R&D collaborations and personnel hiring). Key to the reception and implementation of these mediums is the absorptive capacity of the unit under study.

The organisational infrastructure required for facilitating the development and transfer of knowledge depends heavily on the recipient knowledge platform. The knowledge assets (Nonaka, 1994), sector roles (Baba et al., 2009) and older science-push and demand-pull concepts (Langrish et al., 1972), factor into the knowledge base's receptivity. This receptivity is known as 'absorptive capacity' (Cohen, W. M. & Levinthal, 1990) and can best be described as "[t]he ability of a firm to recognize the value of new, external information, assimilate it, and apply it to commercial ends is critical to its innovative capabilities," (p.128).

On an individual level, select individuals act as gatekeepers, such as star (Zucker, L. G. & Darby, 1996) or core (Furukawa & Goto, 2006) scientists. The concept of absorptive capacity has been expanded on significantly by Zahra & George (2002) to include potential and realised absorptive capacity and address (1) Acquisition – the role of prior knowledge or capabilities and the infrastructure already in place; (2) Assimilation – exogenously generated knowledge needs to be understood prior to incorporation; (3) Transformation – the ability to meld exogenous and endogenous knowledge, to create novel fundamental or applied knowledge and (4) Exploitation – the usage of novel knowledge generated during transformation.

Patents have been used as indicators (Schmookler, 1966) for multiple purposes (e.g. Griliches (1998), Schmoch (1993) and Fleming (2001)) as they are highly detailed evidence of technological progress (Tijssen, 2002). Some drawbacks exist, for example, not all innovations are patented (Arundel, 2001; Arundel & Kabla, 1998) or some innovations are kept secret (Brouwer & Kleinknecht, 1999). Publications serve as the primary indicators for the defining characteristics and development of science. They are the most visible outcome of scientific endeavours, and an extensive range of indicators and methodologies have been developed. Analyses using patents or publications are typically based around the meta-data e.g. Title words, abstract words and keywords (Courtial et al., 1993; Engelsman & van Raan, 1994), patent classifications (Leydesdorff, 2008; Tijssen & Van Raan, 1994), and publication/patent citations (Karki, 1997; Meyer, M. S., 2001).

Citation studies using patent-to-literature citations (Meyer, M., 2000; Meyer, M. S., 2001; Meyer, M., 2002; Narin, 1976, 1994) typically rely on direct citation linkages. Non-patent literature references (NPLRs) exhibit different characteristics based on their source, who includes the reference, the patenting offices and completeness of inclusion (Criscuolo & Verspagen, 2008) and their scientific-ness (Callaert et al., 2006). NPLRs from applicants or examiners have typically been treated as being of differing importance (Karki, 1997) but we choose to utilise both types as the presence of citations to literature in patent documents indicates a cognitive link to, or awareness of, the related scientific concepts (Tijssen, 2001), no matter the source of the NPLRs.
University-based scientists publish primarily to extend their professional and intellectual prowess and regular publishing is considered a requirement. There has been an increase in the rate of university patenting linked to institutional and national level changes (Owen-Smith & Powell, 2003; Zucker, L. G. & Darby, 1996), and the increased interest in academic spin-offs and spin-outs (Owen-Smith & Powell, 2003; Zucker, L. G. & Darby, 1996; Zucker, L.G. et al., 1999).
With firm-based publishing efforts, the firm stands to gain (or lose) more from the publication process than the author, such as – higher rates of approval of patents (McMillan et al., 2003), a window and source into various fields (Schartinger et

al., 2002) and to stronger ties with future progenitors of knowledge (Hicks, 1995; Zucker, L. G. & Darby, 1996).

Case selection

Our case study involves a prominent Japanese biotechnology researcher, Professor Yusuke Nakamura, who is heavily involved in cancer therapeutics at the University of Tokyo, where he was head of the Human Genome Center. Nakamura founded OncoTherapy Science Inc. (OTS) in April of 2001 to research and develop anti-cancer medicine, cancer therapy and cancer diagnosis based on oncogenes and proteins. He maintains direct links between his research at the University of Tokyo and research conducted at OTS allowing us to draw upon his extensive publishing history as well as his numerous patenting activities, both at the University of Tokyo and OTS.

Method

Data collection

The sources and type of data come from (1) Patents – all patent applications with OncoTherapy listed as an applicant were extracted from the EPO PatSTAT database (2000-2008) with all inventors; (2) Publications – all publications with OncoTherapy listed as an institution were downloaded from WoS (all up to 2011); and all publications with Nakamura listed as any of the authors. These base data were parsed using SAINT (2009) and managed in a relational database. Further data were collected from the patents – specifically (where found) (a) In-text non-patent literature references (IT-NPLRs) and (b) Bibliographic NPLRs (B-NPLRs). The patent documents were grouped by INPADOC family and the associated data aggregated to the parent INPADOC family with each collective representing a specific technology (Martinez, 2010). Where possible the NPLR were identified and matched to their ISI WoS twins and added to the extant set. The origins of each document within the combined set were recorded.

The similarities between publications (both NPLR and Nakamura's) were calculated based on their shared cited reference and title word combinations (van den Besselaar & Heimeriks, 2006). A network was constructed using the publications as nodes and the edges representing the degree of similarity as calculated above. The research streams of publications within the network were assigned by utilising a community detection algorithm developed by Blondel et al (2008). Once the initial research stream assignment was completed, the general streams were isolated and the community detection algorithm was run again to produce smaller concept clusters.

The INPADOC families were clustered using the International Patent Classifications (IPC) codes, the use of which for indicators of knowledge-relatedness has been well-developed (Breschi et al., 2003; Jaffe, 1986).

The NPLRs were co-located within the general research streams based on the level of similarity of shared title word and cited reference combinations. By linking the INPADOC families to the general publication communities in which their NPLRs are co-located, we can infer that there is at least a degree of shared knowledge features between the publication community and the citing INPADOC families.
For more specific knowledge features, the second layer of concept clusters provided a finer-grained view into the communities. Within each concept cluster, the source composition of publications varies. In our case study, in which Nakamura is the primary producer of the publications, each concept can potentially contain a mixture of publications authored by Nakamura and either cited or not, and NPLR not authored by Nakamura. Varying proportions of source publications imply differing levels of imparted or similar knowledge features of the publications. Where Nakamura is not cited but his publications are highly similar, we assume similar skillsets and familiarity of topics and processes of the research. With a concept cluster containing both NPLR and non-NPLR publications by Nakamura, this implies direct contributions of the concepts researched and implemented skill sets. Where there is a combination of all three types, we assume there are direct contributions to concepts and skill sets, and a shared knowledge base and minimum required skill sets.

To visualise the publication community structure over time, we employ a method introduced by Horlings & Gurney (2012) where cognitive communities or research trails over time are transformed based on the time ranges of each community to latitude and longitude coordinates to be displayed on an equirectangular map.

Knowledge capture mechanisms

Following on Zahra & George's (2002) dimensions of absorptive capacity (acquisition, assimilation, transformation and exploitation), we are able to examine in detail: (1) the reputational and applicability aspects of the scientific base work (Hullmann & Meyer, 2003) conducted by Nakamura; (2) the markers for what other fields of science are being utilised by the technologies (Karki, 1997; Schmoch, 1993); (3) the degree of shared knowledge features (such as concepts, knowledge bases and, to a certain extent, skill sets); (4) the level of input from co-inventors of Nakamura; (5) and if Nakamura incorporated skill sets acquired during the development of the technologies and applied them to further his fundamental scientific research by knowledge creation feedback (Fischer, 2001; Tijssen, 1998).

Results

Patents and patent families

In total we collected 242 patent application documents via PatSTAT (Oct 2011) with Nakamura listed as inventor and OncoTherapy as assignee. The patent documents came from 90 INPADOC families, and were composed of 115 priority patents. The earliest patent filing date was March 2000, and the latest was November 2008. The maximum, minimum, average and median numbers of patent applications per INPADOC family are, respectively, 23, 2, 5.3 and 4.

Clustering of INPADOC families by IPC

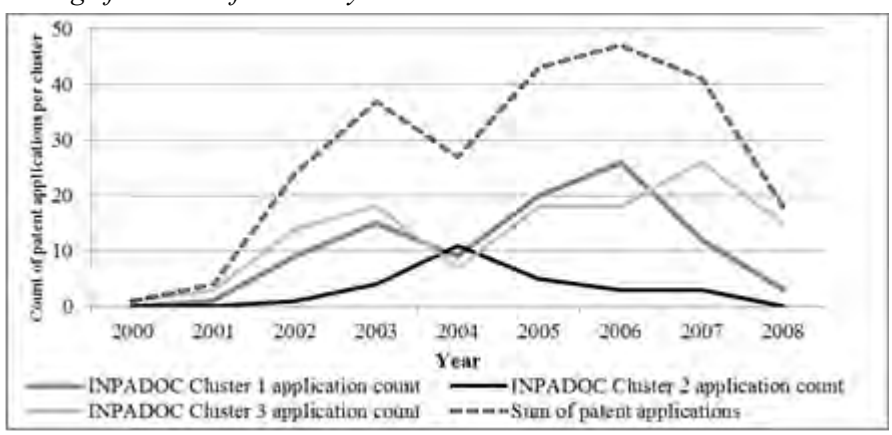


Figure 1(a) INPADOC cluster patent count.

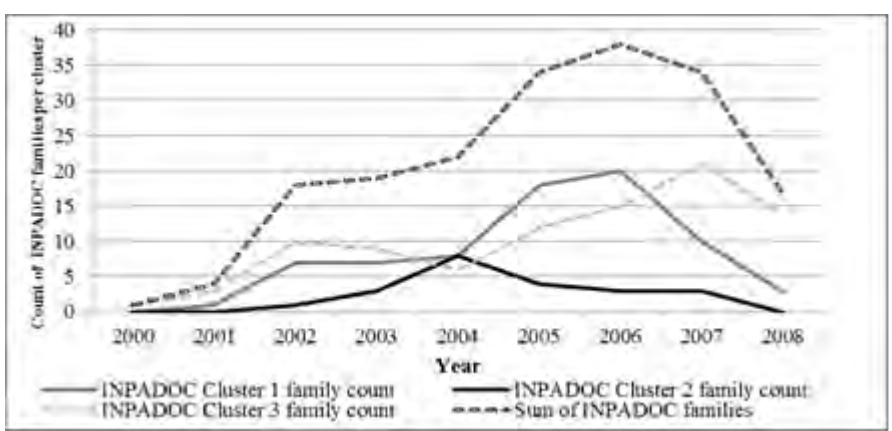


Figure 1(b) INPADOC cluster family count.

Three primary INPADOC clusters were found, using main group IPC data. The growth in the number of patent applications and INPADOC families per cluster are shown in Figures 1(a) and (b). In 2002 and 2004, the number of unique INPADOC

families increased at a slower rate suggesting a period of specialisation within OncoTherapy. From 2004, the increased application rates and increased number of unique families suggest a diversification period. Between 2002 and 2004, Clusters 1 and 3 (dark grey and light grey lines respectively) displayed specialisation whilst Cluster 2 (black border) tended to diversification. In 2004, Cluster 2 peaked and tended to specialisation, whilst 1 and 3 showed overall decreases.

The 2-mode network in Figure 2 demonstrates the specific areas shared by each INPADOC cluster and also serve to highlight which clusters have specialised technological areas that are only applicable to each cluster. As shown in Figure 2, the primary areas at the main group IPC levels addressed by the INPADOC clusters relate primarily to the use of micro-organisms, enzymes, peptides and growth factors, recombinant DNA technologies and medicinal preparations using the peptides and RNA.

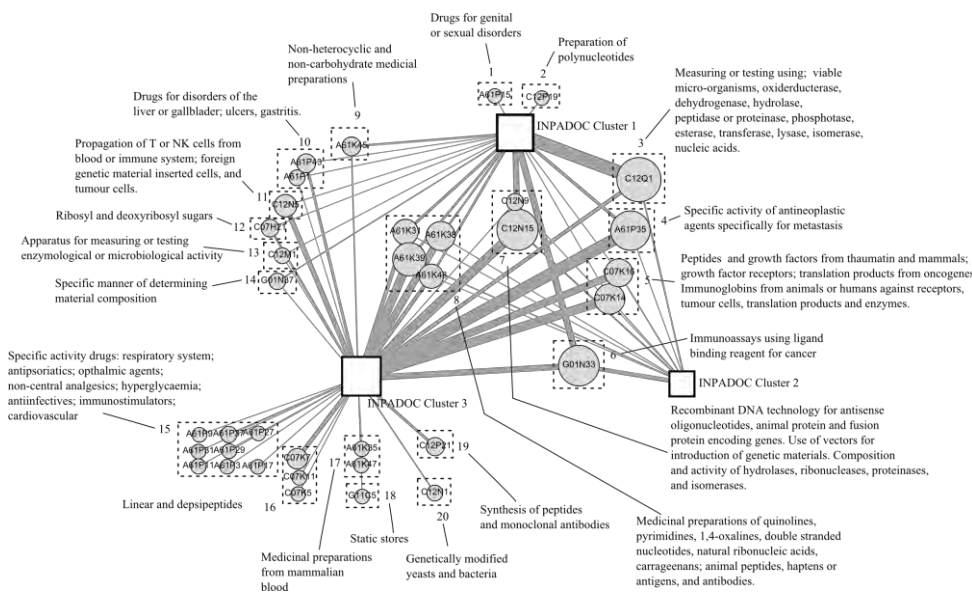


Figure 2 Annotated 2-mode network of main group level IPC and inpadoc family clusters. (Note: Node inpadoc clusters= count of inpadoc families, node size main group IPC nodes=count of patent applications citing main group IPC code. Edge weight=proportional count of number of patent applications utilising the main group IPC code.)

Publications, NPLRs and patents

Nakamura has published a large number of publications with 931 publications over 33 years. His first publication was in 1977 and at a rate under 5 per year until 1987. Between 1988 and 1994, he published between 5 and 10 publications a year and at present, he (co)publishes at a rate of 50 a year.

In total we were able to positively link 525 unique occurrences of B- and IT-NPLRs to the 242 patent applications. Of these NPLRs, 147 were uniquely B-NPLRs, 313 were uniquely IT-NPLRs and 65 NPLR were shared. The most cited NPLR is cited by 41 different patent applications. The most cited publications come from the time period of 1996-2004 with less than 10% of NPLR citations going to publications older than 1996.

Table 1 summarises the distribution of NPLR, the content of each stream, and the links to INPADOC clusters. Figure 3 shows the similarity network of the publications and NPLRs over time. Most Nakamura-authored NPLR are located in streams 7 and 13 and the bulk of patent citations are to Streams 9 and 13.

Table 1 Publication stream summary.

Stream	Total (Nakamura/NPLR /Both)	Start	End	Summary	INPADOC clusters
1	157(84/73/0)	1978	2011	Cell biology, nuc. acids, proteins, polypeptides, factor regul.	1, 2, 3
2	273(182/90/1)	1978	2007	Gene-mapping, novel genes, human genes	1, 2, 3
3	2(0/2/0)	1979	1987	RNA	2, 3
4	85(5/80/0)	1987	2008	Cancer gene expression	1, 2, 3
5	169(133/35/1)	1987	2009	Breast cancer, gene mutation	1, 2, 3
6	2(2/0/0)	1988	1989	Mouse liver	-
7	135(97/18/20)	1988	2011	Gene expr., cancer (prostate, liver, pancreas), therap. targets	1, 2, 3
8	15(0/15/0)	1988	2005	Endocrinology, mouse-human models, porcine spinal-cord	2, 3
9	78(6/72/0)	1989	2007	Lymphocytes, melanomas, peptides, antigens	1, 2, 3
10	8(0/8/0)	1991	2002	Endometriosis, fertility and sterility	3
11	6(5/1/0)	1992	1995	Pharmacology, analogs, glycines	2
12	15(0/15/0)	1993	2005	Methylation (histone and glycine)	2, 3
13	159(110/16/33)	1994	2010	Gene expression, cdna microarrays	1, 2, 3
14	8(6/2/0)	1996	2005	Phospholipase, cell receptors	2
15	15(15/0/0)	1996	2005	OLETF rats, diabetes	-
16	20(20/0/0)	1997	2003	Congenital disorders	-
17	2(0/2/0)	1998	2001	Hepatology	2, 3
18	183(183/0/0)	1999	2011	Japan and population specific cancers	-
19	2(0/2/0)	1999	2000	NFAT mechanisms and inhibition	2

Co-inventors and partner institutes

Figure 4 shows the distribution of Nakamura's co-inventors in the publication corpus. Many publications are authored with Nakamura's co-inventors, with some publications cited as NPLR where Nakamura is not an author. This would seem to

indicate that the knowledge utilised by the patent applications stems not only from Nakamura, but also from his co-inventors. However, the relative scarcity of cited NPLR without Nakamura as author but with one of his co-inventors authoring would suggest that the knowledge comes from within Nakamura's research group.

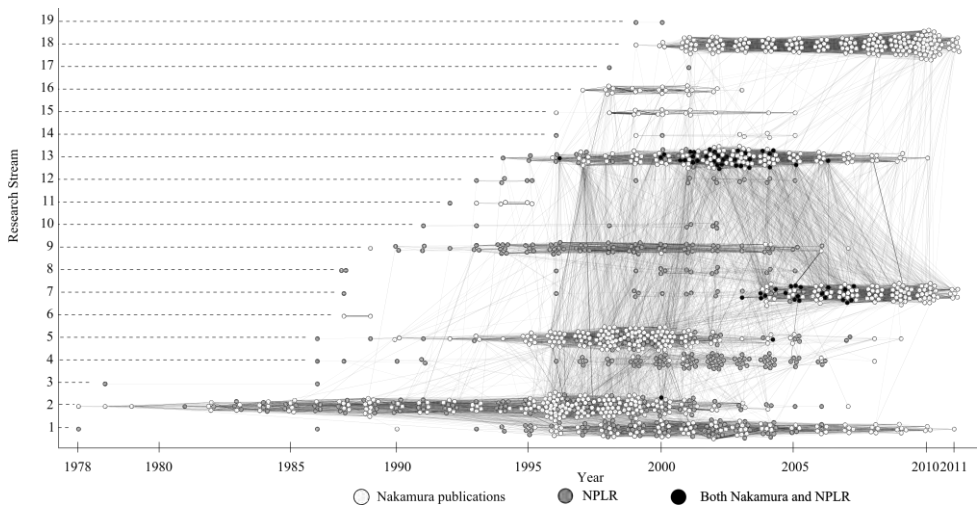


Figure 3 Longitudinal and research stream clustering of Nakamura and NPLR publications. (Note: edges=degree of title word/reference combination similarity. Node colour=source where white=Nakamura publications, Grey=NPLR, Black= Both Nakamura and NPLR

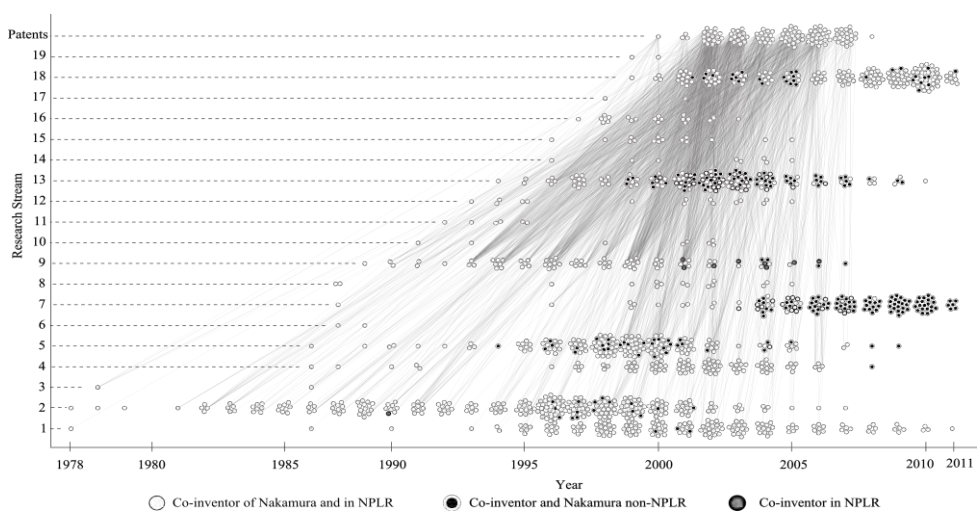


Figure 4 Co-inventor location in research streams. (Note: Only edges between patent applications and cited publications are shown (both IT-NPLR and B-NLPR).

Within the 77 INPADOC families on which OncoTherapy is listed as assignee and Nakamura as inventor, Nakamura has 10 recurring co-inventors, with 4 of these co-inventors also patenting without Nakamura. OncoTherapy has 6 researchers that patent without Nakamura, but the vast majority of INPADOC families primarily stem from patent applications with Nakamura listed as inventor.
OncoTherapy collaborates on patents with only two organisations, the University of Tokyo in 26 different INPADOC families, and Sentan Kagaku Gijutsu Incubation Center in one INPADOC family. The University of Tokyo is present in just under a third of OncoTherapy's INPADOC families, which, considering Nakamura is based at the university, is not particularly high. The fact that, overall, there is only 1 significant patenting organisational partner for OncoTherapy's technologies is interesting.

Concept clusters

From the 19 research streams, we extracted 66 concept clusters (CCs) that contain NPLRs (both non-Nakamura- and Nakamura-authored). We linked these CCs to the citing INPADOC families and the designated INPADOC clusters. Presented in Figures 5 (a)-(c) are citations to CCs from the INPADOC clusters. Due to space constraints, we have chosen to focus on streams 1, 7, 9 and 13 and their CCs.
From Figure 5(a) – containing only NPLR not authored by Nakamura thus outside Nakamura's expertise, the INPADOC clusters rely heavily, and from an early stage, on CC 9/0 and CC 9/1 (research related to the cytotoxic effect of lymphocytes, and human leukocytes and antigens). INPADOC cluster 1 exclusively cites research from CC 7/2 (increasing rates of bile duct cancer) and CC 1/2 (mRNA binding proteins expression and cancer proteins).
Figure 5(b) shows CCs containing both non-Nakamura-NPLRs and non-NPLR-Nakamura publications. This combination of sources indicates that there is some immediate similarity between research performed by Nakamura and the cited publications. In many cases, the research is cited from an early stage (as seen by the grey edges between nodes) but there is a fair degree of research cited later in the technologies' development phases (dashed and solid black edges). CCs 9/2, 9/3 and 9/4 are cited early by all three clusters, and Nakamura only starts to publish much later in these topics (also seen in Figure 3).
All three INPADOC clusters cite research in CCs 1/4 and 1/5, but again Nakamura's publications related to those topics are only published later. For CC 1/1, cited exclusively by INPADOC cluster 1 in the middle phase of its development, Nakamura - whilst having published extensively in that concept cluster – is not cited at all.
Figure 5(c) shows the CCs considered to contain the most specific aspects of research performed by Nakamura. In most cases, the INPADOC clusters cite the CCs from an early stage but in many cases Nakamura only published later in these topics. This is a strong indicator that Nakamura recognized the necessity of the knowledge in these CCs to further develop the technologies, and assimilated and transformed the content for future research purposes.

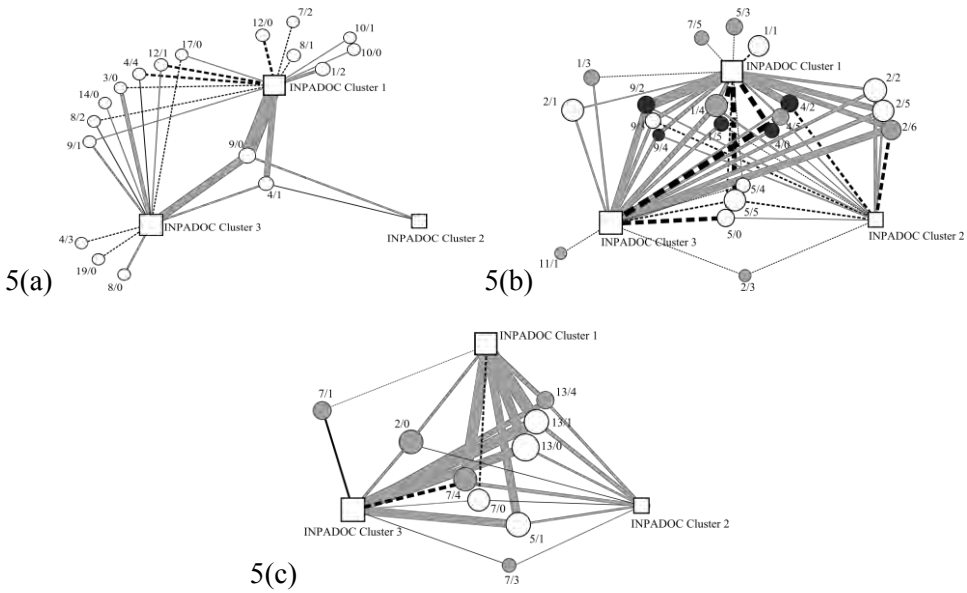


Figure 5 (a)-(c) Concept clusters cited by inpadoc clusters containing (a) only NPLR not authored by Nakamura; (b) NPLR not authored by Nakamura and publications by Nakamura not cited by the patent applications and (c) NPLR authored by Nakamura (Note: For concept labels, a/b, a=parent stream ID, and b=concept ID. Size of nodes=count of publications or count of inpadoc families. Thickness of edges =number of citing inpadoc families. Edge colours: age of the inpadoc cluster the concept is cited, grey=early, dashed=middle, black=late. CC node colours for (b) and (c): White=Nakamura publications present from start, gray= Nakamura publications present from middle time period, black= Nakamura publications present at end of time period))

Summarising, in stream 1, Nakamura publishes extensively but is not cited by the patent applications at all. The degree of exogenously-generated knowledge is high, with no direct contributions by Nakamura. However, the shared knowledge base and shared minimum skill set is significant as only one of the five CCs cited do not contain any Nakamura publications.

With stream 7, initially the INPADOC clusters barely cite the stream at all. Up to 2004 the first cited NPLRs were all non-Nakamura NPLRs but from 2004 onwards Nakamura publishes prolifically and is often cited. The proportionally large number of Nakamura-NPLRs and Nakamura's knowledge base and skill sets are now integral to the technologies.

The technologies cite stream 9 extensively but Nakamura's role is limited. He is not directly cited but does publish at later stages in all of the cited CCs. In short, the necessary scientific aspects derived from stream 9 are exogenously sourced. However some of the topics relate to background information.

Nakamura-authored publications dominate stream 13 with a third of his publications cited by the technologies. In one CC (13/4) Nakamura is not the first

to publish, with some NPLRs coming from others. The role of Nakamura's research in stream 13 and its contributions to the technologies of clusters 1-3 is more obvious as the publications in this stream are authored almost entirely by Nakamura.

Summary and conclusion

Considering the enormous volume of data available with our approach, we chose to focus on four specific streams of publications and their impact (through citation links and topic similarity) on the patent applications. We also reduced the specificity of the technologies by aggregating the patent applications into INPADOC families and then further into INPADOC clusters. At an obvious loss of detail, we feel that the aggregation was necessary to better analyse the knowledge and skillset contributions of Nakamura as an individual.

Nakamura's impact within these four streams on the INPADOC clusters was viewed through the lens of the adoption and adaptation aspects of Zahra and George (2002) and their respective source of knowledge, be they exogenously or endogenously generated.

Acquisition – this dimension primarily details the role of prior knowledge or capabilities and the infrastructure already in place. The first step in this aspect is recognising knowledge that is or would be useful to the development of the technologies. By examining the degree of required knowledge through Figures 4 and 5(a) we gain insight to this aspect. Co-inventors are considered here as they provide necessary expertise and skillsets.

Assimilation – By conducting research in the topic areas required for the technologies, whether through a non-concerted approach or a cumulative directed approach, the codified and tacit skills and insights developed directly impact the development of the technologies. In this sense, the process of 'learning-by-doing' seems to be prevalent. In examining Figures 5(a) and (b) we can see the specific topics and levels of contribution by Nakamura and at what stages of the science his contributions become visible.

Transformation – addresses the ability to meld exogenous and endogenous knowledge, to create novel fundamental or applied knowledge. Streams 7 and 13 from Figure 3 provide examples of this. Taking into consideration the degree of similarity between stream 5 and stream 13, we see a strong link, particularly around 1996 and 2000 coinciding with bursts of publishing one year later in stream 13. The skill sets and knowledge acquired in practising research in topics within stream 5 have had a significant influence on the required knowledge and skills sets for stream 13. The same behaviour can be discerned between streams 13 and 7, where stream 13 provides the required knowledge and skill sets for the topics in stream 7. Translating this to the patent applications: where the technologies previously relied on exogenously generated knowledge from streams 7and 13, the endogenously generated knowledge of stream 5 was successfully acquired, assimilated and transformed for use in streams 13 and 7.

On a methodological level, our approach benefits from its ability to encompass both the macro and micro views. Our approach can isolate and highlight specific aspects of utilised knowledge in relation to the knowledge features already in place. We are able to co-locate the knowledge features of individuals who contribute to the publications and patent applications, not through the direct citations of NPLRs, but through the co-location of NPLRs in the environment.
A disadvantage of our method as outlined above is the complexity of the process. Due to this complexity we chose to aggregate the technologies into clusters of INPADOC families. This limits our attention to detail within the technologies but allows a thorough examination of the contributions of an individual (in our case Nakamura). The possibility exists to aggregate on the publication side and examine in detail the characteristics of the technologies being produced.
We see this method aiding in the evaluation of technologies and the contributions of those involved in the development of the technologies. With the addition of funding information in the meta-data extracted from WoS it would then be possible to trace the results of such funding to its exploitation phase and results. The scaling up of this method would allow research groups, departments or entire research institutes or infrastructures to map their contributions in the early stages of the development of a technology right through to their exploitation or implementation. This would be invaluable to funding agencies and universities for reporting on their research achievements, as in many cases the end-point of fundamental and applied research may be so far removed from the origin as to be unrecognisable.

Acknowledgements

We gratefully acknowledge the support of a JSPS/CNRS Grant (DP) and Aix Marseille University (KS). The authors would like to thank the two anonymous reviewers for their useful comments and input.

References

Arundel. (2001). The relative effectiveness of patents and secrecy for appropriation. *Research policy, 30*(4), 611-624.

Arundel, & Kabla. (1998). What percentage of innovations are patented? Empirical estimates for European firms. *Research policy, 27*(2), 127-141.

Baba, Shichijo, & Sedita. (2009). How do collaborations with universities affect firms' innovative performance? The role of. *Research Policy, 38*(5), 756-764.

Blondel, Guillaume, Lambiotte, & Lefebvre. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*, P10008.

Breschi, Lissoni, & Malerba. (2003). Knowledge-relatedness in firm technological diversification. *Research policy, 32*(1), 69-87.

Brouwer, & Kleinknecht. (1999). Innovative output, and a firm's propensity to patent.:: An exploration of CIS micro data. *Research policy, 28*(6), 615-624.

Callaert, Van Looy, Verbeek, et al. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics, 69*(1), 3-20.
Cohen, & Levinthal. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly, 35*(1, Special Issue: Technology, Organizations, and Innovation), 128-152.
Cohen, Nelson, & Walsh. (2002). Links and impacts: the influence of public research on industrial R&D. *Management Science*, 1-23.
Courtial, Callon, & Sigogneau. (1993). The use of patent titles for identifying the topics of invention and forecasting trends (Vol. 26, pp. 231-242): Springer.
Criscuolo, & Verspagen. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research policy, 37*(10), 1892-1908.
Engelsman, & van Raan. (1994). A patent-based cartography of technology. *Research policy, 23*(1), 1-26.
Fischer. (2001). Innovation, knowledge creation and systems of innovation. *The Annals of Regional Science, 35*(2), 199-216.
Fleming, & Sorenson. (2001). Technology as a complex adaptive system: evidence from patent data. *Research Policy, 30*(7), 1019-1039.
Furukawa, & Goto. (2006). Core scientists and innovation in Japanese electronics companies. *Scientometrics, 68*(2), 227-240.
Griliches. (1998). Patent statistics as economic indicators: a survey: University of Chicago Press.
Gurney, Schoen, Horlings, et al. (2012). *Knowledge Capture Mechanisms in Bioventure Corporations*. Paper presented at the Proceedings of 17th International Conference on Science and Technology Indicators.
Hicks. (1995). Published papers, tacit competencies and corporate management of the public/private character of knowledge. *Industrial and Corporate Change, 4*(2), 401.
Horlings, & Gurney. (2012). Search strategies along the academic lifecycle. *Scientometrics*, 1-24.
Hullmann, & Meyer. (2003). Publications and patents in nanotechnology. *Scientometrics, 58*(3), 507-527.
Jaffe. (1986). Technological Opportunity and Spillovers of R & D: Evidence from Firms' Patents, Profits, and Market Value (Vol. 76, pp. 984-1001): JSTOR.
Karki. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information, 19*(4), 269-272.
Langrish, Gibbons, Evans, & Jevons. (1972). *Wealth from knowledge: a study of innovation in industry*: Halstead Press Division, Wiley.
Leydesdorff. (2008). Patent classifications as indicators of intellectual organization. *Journal of the American Society for Information Science and Technology, 59*(10), 1582-1597.
Martinez. (2010). *Insight into different types of patent families*: OECD.
McMillan, Mauri, & Robert III. (2003). The impact of publishing and patenting activities on new product development and firm performance: the case of the

US pharmaceutical industry. *International Journal of Innovation Management, 7*(2).
Meyer. (2000). Does science push technology? Patents citing scientific literature. *Research Policy, 29*(3), 409-434.
Meyer. (2001). Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology (Vol. 51, pp. 163-183): Springer.
Meyer. (2002). Tracing knowledge flows in innovation systems. *Scientometrics, 54*(2), 193-212.
Narin. (1976). Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity. *Cherry Hill, NJ: Computer Horizons, 338*, 27.
Narin. (1994). Patent bibliometrics. *Scientometrics, 30*(1), 147-155.
Nonaka. (1994). A dynamic theory of organizational knowledge creation. *Organization science, 5*(1), 14-37.
Owen-Smith, & Powell. (2003). The expanding role of university patenting in the life sciences: assessing the importance of experience and connectivity. *Research policy, 32*(9), 1695-1711.
Schartinger, Rammer, Fischer, & Fröhlich. (2002). Knowledge interactions between universities and industry in Austria: sectoral patterns and determinants. *Research Policy, 31*(3), 303-328.
Schmoch. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics, 26*(1), 193-211.
Schmookler. (1966). *Invention and economic growth*: Harvard University Press Cambridge, MA.
Somers, Gurney, Horlings, & Van den Besselaar. (2009). *Science Assessment Integrated Network Toolkit (SAINT): A scientometric toolbox for analyzing knowledge dynamics*. The Hague: Rathenau Insitute.
Tijssen. (1998). Quantitative assessment of large heterogeneous R&D networks: the case of process engineering in the Netherlands1. *Research policy, 26*(7-8), 791-809.
Tijssen. (2001). Global and domestic utilization of industrial relevant science: patent citation analysis of science-technology interactions and knowledge flows. *Research policy, 30*(1), 35-54.
Tijssen. (2002). Science dependence of technologies: evidence from inventions and their inventors. *Research policy, 31*(4), 509-526.
Tijssen, & Van Raan. (1994). Mapping changes in science and technology. *Evaluation Review, 18*(1), 98-115.
van den Besselaar, & Heimeriks. (2006). Mapping research topics using word-reference co-occurrences: a method and an exploratory case study. *Scientometrics, 68*(3).
Zahra, & George. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of management review*, 185-203.

Zucker, & Darby. (1996). Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proceedings of the National Academy of Sciences of the United States of America, 93*(23), 12709.
Zucker, Darby, & Brewer. (1999). *Intellectual capital and the birth of US biotechnology enterprises*: National Bureau of Economic Research.

LEAD-LAG TOPIC EVOLUTION ANALYSIS: PREPRINTS VS. PAPERS (RIP)

Ying Ding, Erjia Yan, Cassidy Sugimoto, Staša Milojević

{dingying, eyan, sugimoto, smiloje}@indiana.edu
School of Library and Information Science, Indiana University

Abstract
Publishing or perishing reflects the fierce competition in scholarly communication. Claiming authorship of innovation is essential in many fields, especially in physics or astrophysics. Over the last decade, Web 2.0 technologies have dramatically changed the style of our scholarship and scholarly communication. People can claim and share their initial and innovative thoughts/ideas via online archiving systems, personal blogs, twitter, and facebook. In this paper, we applied the topic modeling algorithm (the LDA model) and the time series analysis to conduct the lead-lag analysis and identify different topic evolution patterns for preprints from arXiv and papers from Web of Science (WOS) in astrophysics during the last twenty years (1992-2011). We found that both arXiv and WOS share similar topic evolution trends in popular topics and identified some knowledge transfer delay in WOS.

Conference Topic
Collaboration Studies and Network Analysis (Topic 6) and Modeling the Science System, Science Dynamics and Complex System Science (Topic 11)

Introduction
Publishing or perishing reflects the fierce competition in scholarly communication. Being the first person to claim new ideas, new methods, or new discoveries is critical in science. Scholarly communication, especially through the formal channels, plays a critical role in this process. However, over the last decade the Internet and especially the Web 2.0 technologies started having an impact on our scholarly communication by enabling fast and broad dissemination. Researchers thus started sharing their initial and innovative thoughts/ideas via their personal blogs, tweets, facebook comments, and online discussion groups. These ideas can then be downloaded, discussed, tweeted/retweeted, forwarded, commented, and tagged via different online platforms, such as Twitter, Mendely, Citeseer, and CiteULike. This informal scholarly communication significantly speeds up the process of knowledge dissemination.
arXiv is an online repository of e-prints in a number of fields – most notably in physics, mathematics, and computer science. Since its creation by Paul Ginsparg in 1991, arXiv has become central to the diffusion of research in those fields. Today, arXiv is one of the largest open access self-archiving systems which hosts over 0.8M e-prints in science covering physics, mathematics, computer science,

quantitative biology, statistics, and quantitative finance. This online archiving system, with the policy of allowing every author to submit his/her research output, offers the ideal platform to swiftly propagate knowledge. Before manuscripts enter lengthy peer-review processes that can take anytime from 3 months up to 1.5 years (depending on different journals or disciplines), they can already be viewed, criticized, even cited by public audience. arXiv has also become one of the major open access venues for an ever growing number of researchers who want to reach the wider audience, but do not have means to pay extremely high open access fees to journals. Thus, in its mixed role arXiv contains papers in different stages of their life-cycle: from true pre-prints to post-prints.

An aspect of arXiv that has generated quite an interest is its role in changing scholarly communication and accelerating knowledge transfer. Shuai et al. (2012), for example, analyzed the online responses to 4,606 e-prints submitted to arXiv using downloads, mentions on Twitter, and citations in scholarly articles. They studied the delay and time span of article downloads and Twitter mentions to understand the temporal difference. Through the regression and correlation tests, they found that Twitter mentions and arXiv downloads follow distinct temporal patterns with Twitter mentions having shorter delays and narrower time spans. Shi et al. (2011) on the other hand conducted the topical lead-lag analysis on papers and funding proposals to study whether research grants lead publications or vice versa. They proposed a general method for lead-lag estimation based on the LDA (Latent Dirichlet Allocation) model and time series analysis, and applied them on 20,000 grant proposal abstracts and half a million computer science research paper abstracts. They found that the lead-lag of research papers with respect to research grants is topic specific. Thus, on the topic of Security and Cryptography, research papers lead by two years ahead of grant proposals, while on the topic of Neural Network, grants lead by three years ahead of research papers. However, so far there has been no research conducted to analyze the difference of topic evolution in informal scholarly communication (i.e., e-prints) and formal scholarly communication (i.e., publications). In this paper, we applied LDA and time series analysis to conduct the lead-lag analysis and identify different topic evolution patterns for e-prints and papers in astrophysics for the last twenty years (1992-2011). As this is an ongoing effort funded by National Science Funding, we present some preliminary results here and point out our future work.

Data and proposed methods

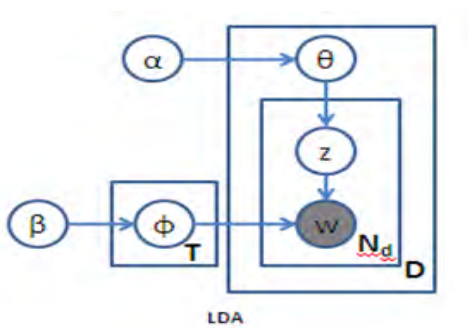
Data

We used two major sources of data: arXiv and Web of Science (WOS). Data were crawled from arXiv through the category called astrophysics. In arXiv, astrophysics was categorized to include the following subfields: cosmology and extragalactic astrophysics, earth and planetary astrophysics, galaxy astrophysics,

high energy astrophysical phenomena, instrumentation and methods for astrophysics, and solar and stellar astrophysics. arXiv started to host astrophysics e-prints in April 1992. We collected 117,913 astrophysics e-prints from 1992 to 2011. In WOS, astrophysics is listed as one of the WOS subject categories. All papers in different document types were collected from this subject category for the period of 1992-2011. Finally, 166,191 research articles were collected from WOS.

Methods

Latent Dirichlet Allocation (LDA) captures the topical features of nodes by postulating a latent structure for a set of topics linking words and documents. The LDA method has been reliable for detecting multi-nominal word distribution of topics (Blei, Ng, & Jordan, 2003). After the success of the LDA model, the basic model has been extended to various levels. The Author-Topic model proposed by Rozen-Zvi and her colleagues (2004) depicts the content of documents and the interests of authors simultaneously. Later, Tang and his colleagues extended LDA to reveal the topic distribution of authors, conferences, and citations concurrently (Tang, et al., 2008). LDA has been applied in scholarly communication to identify the topic distribution in dynamic research communities (Yan et al., 2012), to analyze disciplinary development using domain specific dissertations (Sugimoto et al., 2011), to study scientific collaboration and endorsement at the topic level (Ding, 2011a), and to calculate topic-based PageRank (Ding, 2011b).



Notations	Meaning
w	word
z	topic
N_D	the number of words in the entire collection of documents
α	hyperparameter for generating Θ from Dirichlet Distribution
β	hyperparameter for generating φ from Dirichlet Distribution
Θ	a multinomial distribution over topics
φ	a multinomial distribution over words
D	collection of documents
T	collection of topics

Figure 1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) provides a probabilistic model for the latent topic layer (Blei, Ng, & Jordan, 2003). For each document d, a multinomial distribution θ_d over topics is sampled from a Dirichlet distribution with parameter α. For each word w_{di}, a topic z_{di} is chosen from a topic-specific multinomial

distribution $\phi_{z_{di}}$ sampled from a Dirichlet distribution with parameter β. The probability of generating a word *w* from a document *d* is:

$$P(w|d,\theta,\phi) = \sum_{z\in T} P(w|z,\phi_z)P(z|d,\theta_d)$$

Therefore, the likelihood of a document collection *D* is defined as:

$$P(Z,W|\Theta,\Phi) = \prod_{d\in D}\prod_{z\in T}\theta_{dz}^{n_{dz}} \times \prod_{z\in T}\prod_{v\in V}\phi_{zv}^{n_{zv}}$$

where n_{dz} is the number of times that a topic z has been associated with a document d, and n_{zv} is the number of times that a word w_v has been generated by a topic z. The model can be explained as: to write a paper, an author first decides topics and then uses words that have a high probability of being associated with these topics to write the article. The limitation of LDA is that the number of topics to be extracted should be decided beforehand which is usually based on perplexity. Labeling and judging the quality of topics can only be empirically evaluated.

Preliminary results

Overview

The LDA model was applied to articles in astrophysics collected from WOS (166,191) and arXiv (117,913) for the last 20 years (1992-2011). Fifty topics were extracted using the LDA model. The 50 topics of arXiv were matched to the 50 topics of WOS. Table 1 shows these 50 topics. Each topic was labeled using the top five ranked words (i.e., words with high probability in this topic). The extracted 50 topics demonstrate the major research topics in astrophysics including: astrophysical laws (e.g., blackhole radition, radiative transport, gravity, and star structure), stellar physics (e.g., stellar evolution, chemical dependency, white dwarfs, neutron stars, and black holes), interstellar medium (e.g., heating, gas dynamics, magnetic fields, and shocks), cosmology (e.g., models, dark matter, inflation, and accelerating), and galaxies (e.g., spiral, disk, surface, Milky Way, and density waves).

Figure 2 shows the topic distribution of these 50 topics for arXiv and WOS. In arXiv (the blue line), topic was distributed with topic 12 (Cosmic Microwave Background (CMB)) as the most popular topic throughout the 20-year time span, then followed by topic 25 (dark matter researcha), topic 7 (Gamma-ray burst (GRB)), topic 28 (Lyman-alpha systems and cosmology), and topic 48 (Redshift-Luminosity Distance Relation). In WOS (the red line), the topic distribution is comparably stable with lower amplitude oscillations. Topic 31 (active galactic nucleus and Seyfert galaxy) remains the most popular topic during the later years, followed by topics 40 (accretion-disk-simulation-wind-jet), topic 25, topic 23 (pulsar-X1-PSR-source-transient) and topic 28.

Table 1. Fifty topics of astrophysics in arXiv and WOS

Topic 00	blackhole-massive-accretion-binaries-disk	Topic 25	dark-matter-universe-cosmology-milky
Topic 01	rotating-model-stability-theory-cosmology	Topic 26	how-what-meteor-astronomy-why
Topic 02	impact-meteorite-origin-chondrite-lunar	Topic 27	type-supernovae-neutron-core-nucleosynthesis
Topic 03	region-maser-source-infrared-line	Topic 28	alpha-redshift-field-quasar-absorption
Topic 04	line-excitation-transition-atomic-irons	Topic 29	coronal-region-loop-flux-heating
Topic 05	binaries-spectroscopy-eclipsing-light-photometric	Topic 30	spiral-surface-disk-brightness-gas
Topic 06	variable-period-cataclysmic-majoris-variation	Topic 31	active-nuclei-seyfert-variability-line
Topic 07	gammaRay-burst-GRB-afterglow-blazar	Topic 32	magellanic-cloud-globular-cloud-photometry
Topic 08	motion-observation-photograph-determination-reference	Topic 33	dwarf-group-globular-compact-elliptical
Topic 09	line-profile-polarization-spectrum-absorption	Topic 34	dwarf-lowMass-open-sequence-binaries
Topic 10	motion-orbit-theory-satellite-peridic	Topic 35	supernova-comet-remnant-cygus-coma
Topic 11	planet-system-extrasolar-satellite-jupiter	Topic 36	sky-source-sample-catalog-digital-rosat
Topic 12	background-cosmic-microwave-power-spectrum	Topic 37	dust-circumstence-disk-tauri-envelope
Topic 13	costmic-energy-ray-gammaRay-highEnergy	Topic 38	spectral-distance-distribution-determination
Topic 14	oscillation-model-pulsation-mode-delta	Topic 39	wave-convection-dynamo-flow-rotating
Topic 15	nova-outburst-spectrum-cygni-dwarf	Topic 40	accretion-disk-simulation-wind-jet
Topic 16	photometry-cepheids-open-photoelectric-distance	Topic 41	acceleration-plasma-wave-shock-radiation
Topic 17	telescope-space-hubble-imaging-observation	Topic 42	data-analysis-method-astronomy-application
Topic 18	spectra-ultraviolet-analysis-atmosphere-wolfRayed	Topic 43	measurement-atmosphere-mars-atellite-venus
Topic 19	sunspot-rotation-activity-cycle-variation	Topic 44	flare-hard-burst-observed-coronal
Topic 20	abundance-giant-chemical-red-branch	Topic 45	interstellar-could-dust-grain-diffuse
Topic 21	gas-interstellar-hydrogen-neutral-cloud	Topic 46	molecular-cloud-core-region-dense
Topic 22	nebula-planetary-central-orion-bipolar	Topic 47	gravitation-lensing-microlens-quasar-weak
Topic 23	pulsar-X1-PSR-source-transient	Topic 48	luminositiy-function-relation-distribution-redshift
Topic 24	source-radio-polarization-object-compact	Topic 49	transfer-radiative-method-equation-radiation

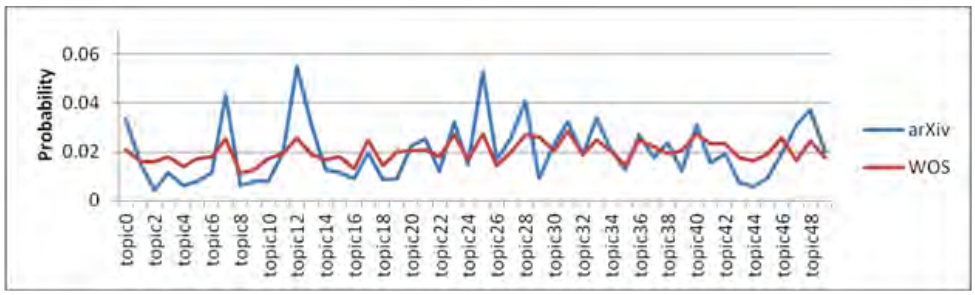


Figure 2. Overview of the topic distribution for astrophysics paper in WOS and arXiv (1992-2011) (Note: horizontal axis represents topics, and vertical axis represents the topic distribution probability)

Lead-Leg Analysis

In Table 2, Topic 7 (Gamma-Ray Burst (GRB)) is the most popular topic in 1992 and 1999 in arXiv, which takes 13 years to climb to the top in WOS. But topic 12 (Cosmic Microwave Background (CMB)) which has stayed as the most popular topic in arXiv during the period of 1993-2001 never managed to reach the top in WOS. Topic 28 (Lyman-alpha systems and cosmology) was very popular in 1998 in arXiv but took an additional 5 years to become the most popular topic in WOS in 2003. Topic 25 started to get popular in arXiv since 2002, and two years later it became the most popular topic in WOS.

Table 2: Most popular topic in each year in arXiv and WOS (1992-2011)

Year	arXiv	WOS	Year	arXiv	WOS
1992	topic 7	topic 39	2002	topic 25	topic 23
1993	topic 12	topic 30	2003	topic 12	topic 28
1994	topic 12	topic 31	2004	topic 25	topic 25
1995	topic 12	topic 40	2005	topic 25	topic 7
1996	topic 12	topic 40	2006	topic 25	topic 25
1997	topic 12	topic 22	2007	topic 25	topic 25
1998	topic 28	topic 23	2008	topic 25	topic 25
1999	topic 7	topic 31	2009	topic 25	topic 25
2000	topic 12	topic 31	2010	topic 11	topic 25
2001	topic 12	topic 23	2011	topic 11	topic 25

Topic Evolution

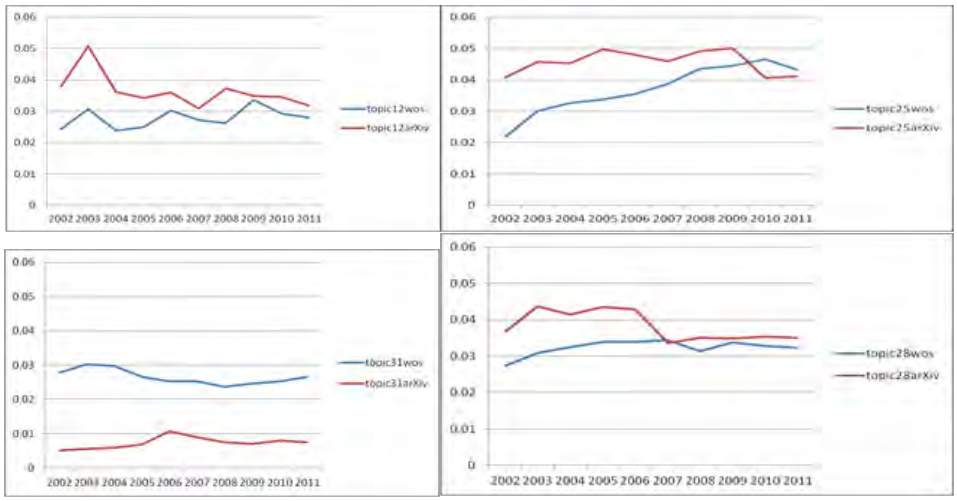


Figure 3: Topic evolution for several individual topics

Figure 3 displays the topic evolution of the latest 10 years (2002-2011) for several individual topics. The general evolving trends of WOS and arXiv are similar. We can see that topic 25 (dark matter research) reaches the highest topic distribution in 2009 in arXiv, while in 2010 in WOS. So there is a one year delay in WOS. The highest peak for topic 12 (Cosmic Microwave Background (CMB)) in arXiv is in 2003, while in WOS it is in 2009, same for topic 28, which becomes very popular in 2003 (arXiv) and in 2005 (WOS). Thus, we can see that there is a delay of one up to six years in popular topics from arXiv and WOS. However, for particular topics (e.g., topic 31), one can observe the reverse trend. Namely, it gains its high popularity in arXiv in 2006, while in WOS in 2003.

Future work

This paper outlines the preliminary results of the ongoing project aiming to study the topic evolution patterns for e-prints and published articles. We found that both arXiv and WOS share similar topic evolution trends, as both have similar most popular topics during 1992-2011. We identified some knowledge transfer delays in WOS, ranging between one and six years. Future work includes: 1) thorough comparison of topic evolution patterns for each topic in arXiv and WOS to categorize the evolution patterns and identify the reasons; 2) application of other topic modeling algorithms to extract topic distributions for authors and journals in order to compare the topic evolution for authors and journals, and 3) comparison of the evolution patterns in astrophysics with other domains in science, social science and humanities.

Acknowledgement

This project is supported by Dig Into Data Project funded by NSF (NSF grant SMA-1208804). We thank Vincen Larivière for data collection.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1033.

Ding, Y. (2011a). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, 5(1), 187-203

Ding, Y. (2011b). Topic-based PageRank on author co-citation networks. *Journal of the American Society for Information Science and Technology*, 62(3), 449-466

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, p487-494, Banff, Canada.

Shi, X., Nallapati, R., Lescovec, J., McFarland, & Jurafsky, D. (2011). Who leads whom: Topical lead-leg analysis across corpora. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media,* July 17-21, 2011, Barcelona, Spain.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PLoS ONE* 7(11): e47523. doi:10.1371/journal.pone.0047523

Sugimoto, C. R., Li, D., Russell, T., Finlay, S., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing Library and Information Science (LIS) dissertations. *Journal of the American Society for Information Science and Technology*, 62(1), 185-204.

Tang, J., Jin, R., & Zhang J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings of 2008 IEEE International Conference on Data Mining (ICDM2008)*, p1055-1060, Dec 15-19, 2008, Pisa, Italy.

Yan, E., Ding, Y., Milojevic, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics*, 6(1), 140-153.

LITERATURE RETRIEVAL BASED ON CITATION CONTEXT

Shengbo Liu[1] Chaomei Chen[2] Kun Ding[1] Bo Wang[1] Kan Xu[3]

[1] *liushengbo1121@gmail.com, dingk@dlut.edu.cn, bowang1121@gmail.com*
WISElab, Dalian University of Technology, No. 2, Linggong Road, Ganjingzi district, Dalian,116024, China

[2] *chaomei.chen@drexel.edu*
College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104-2875, USA

[3]*xukan@dlut.edu.cn*
Information Retrieval Laboratory, Dalian University of Technology, China

Abstract
While the citation context of a reference may provide detailed and direct information about the nature of a citation, few studies have specifically addressed the role of this information in retrieving relevant documents from the literature primarily due to the lack of full text databases. In this paper, we design a retrieval system based on full texts in the PubMed Central database. We constructed two modules in the retrieval system. One is a reference retrieval module based on citation contexts. Another is a citation context retrieval module for searching the citation contexts of a specific paper. The results showed that the retrieval system performed very well on searching highly cited papers and classic papers. The citation contexts of a paper might be related to many topics. Finally, tag cloud is employed to present these topics.

Conference Topic
Bibliometrics in Library and Information Science (Topic 14)

Introduction
Literature retrieval is concerned with searching the most relevant bibliographic information. When writing a paper, researchers have to find some papers as the intellectual base of their work. These papers should be the most relevant papers not only to the subject of the paper in discussion but also to the sub-topics of the paper. Normally, researchers will search the relevant papers on the web. But the great amount of scientific information being published makes it difficult for users to identify the most relevant information. For example, in the biomedical domain alone, around 1,800 new papers are published each day (Hunter, 2006).
With the development of the field of scientometrics, citations are often used in literature retrieval to improve the retrieval efficiency. Four types of citations can be applied to enhance the performance of literature retrieval. The first type is the citation count as an indicator for ranking the retrieval results, and finding the most

cited papers. Bibliographic coupling and co-citation measures are another two types based on citation linkages to find the most relevant papers. Bibliographic coupling refers to a linkage between two documents which have one or more identical references (Kessler, 1963), whereas co-citation is defined as a linkage between two documents concurrently cited by another document (Small, 1973). These two types of citations can be used to reveal the relationships between documents. Some examples have showed that they can improve the performance of information retrieval (Eto, 2012; H. Nanba, Kando, N., Okumura, M, 2000; Pao, 1993; Small, 1973). Many popular literature search engines, such as CiteSeer[109] and Google Scholar[110] also use the links between articles and documents provided by citations to enhance their ranked retrieval results. The fourth type of citations is the citation context. The citation context of a given reference can be defined as the sentences that contain a citation of the reference. For instance, the sentence "This comparison is made using BLASTX [18]" is a citation context of the reference [18]. One may also define a citation context based on more sentences before and/or after the citation sentence. Many researchers have tried to enhance search performance by incorporating citation context into information retrieval systems (Bradshaw, 2003; Mercer, 2004; Nakov, 2004; O' Connor, 1982).

Actually, citation context can provide direct information about an instance of citation. Researchers did not use these citation contexts directly to retrieve literature, but use these citation contexts to improve the traditional retrieval systems. One of the most important reasons is that it is very hard to collect all the citation contexts of the retrieved literatures. In the past, information regarding citation context was not readily accessible due to the lack of full text of citing papers. Researchers often had to extract the necessary information through a manual process. For example, O'Connor (O' Connor, 1982, 1983) extracted single words from citation contexts. Small (Small, 1986) extracted concepts from citation contexts to name a cluster of a co-citation network. In recent years, full text literatures are more accessible. PubMed Central provides full text documents in XML format. In this paper we will introduce the design of a literature retrieval system based on all full text documents in PubMed Central.

We design two modules for the retrieval system. One is the reference retrieval module based on citation contexts. Another is the citation context retrieval module for searching the citation contexts of a specific paper. We expect that this system could help researchers find the needed documents more quickly and accurately.

[109] Scientific Literature Digital Library, http://citeseer.ist.psu.edu

[110] Google search engine, for peer-reviewed scholarly literature, http://scholar.google.com

Related work

Citation context analysis

The citation context analysis includes the application of the citation position and citation content.

Citation positions are considered in co-citation analysis. Elkiss (Elkiss, 2008) and Liu (Liu, 2012) studied co-citations in an article at four levels: the sentence level, the paragraph level, the section level, and the paper level. Elkiss found that papers co-cited at a finer granularity are more similar to each other than papers co-cited at a coarser granularity. For example, papers co-cited at the sentence level have a stronger relationship than papers co-cited at the section level. Liu found that sentence-level co-citations are potentially more efficient candidates for co-citation analysis. Gipp (Gipp, 2009b) classified the co-citation into five categories based on occurrence positions: within the same sentence, the same paragraph, the same chapter, the same journal and the same journal but different editions. In each category, a co-citation is given a different value of 1, 1/2, 1/4, 1/8 or 1/16. The result shows that the weighted co-citation analysis yields much more similar documents than traditional co-citation analysis. Callahan (Callahan, 2010) used a similar method to calculate the co-citation strength; a co-citation can occur at different levels of a paper. A co-citation at the paper level is assigned a weight of 1, and for each level deeper an additional weight of 1 is added. Recently, Boyack (Boyack, 2012) used the co-citation proximity to improve the co-citation clustering performance. He found that taking into account reference proximity from full text can increase the textual coherence of a co-citation cluster solution by up to 30% over the traditional approach based on bibliographic information.

Citation content can be used to identify the nature of a citation. The attributions and functions of a cited paper can be identified from the semantics of the contextual sentences (Siddharthan, 2007). Nanba and Okumura (H. Nanba, Okumura, M, 1999, 2005) collected citation context information from multiple papers cited by the same paper and generated a summary of the paper based on this citation context information. They also extracted citing sentences from citation contexts and generated a review. Mei (Mei, 2008) and Mohammad (Mohammad, 2009) found that the summarization of citation contexts is very different from the abstract of the cited reference. Nakov (Nakov, 2004) referred to citation contexts as citances – a set of sentences that surround a particular citation. Citances can be used in abstract summarization and other Natural Language Processing (NLP) tasks such as corpora comparison, entity recognition, and relation extraction. Small (Small, 1979)studied the context of co-citation and analyzed the context in which the co-citation paper was mentioned. In addition, he analyzed the sentiment of the co-citation context (Small, 2011). Mei (Mei, 2008) defined the length of citing sentences as 5, 2 sentences before the citation and 3 sentences after. In this study, we use the sentence with the citation tag as the citation context.

Anderson (Anderson, 2010) analyzed the citation context of a classic paper in organizational learning which was published by Walsh and Ungson in the Academy of Management Review. The results provided a richer understanding of which knowledge claims made by Walsh and Ungson have been retrieved and have had the greatest impact on later work in the area of organizational memory, and also what criticisms have been leveled against their claims. Our research also designed a module for searching citation contexts of any specific paper. It is very helpful for researchers to understand the citation value of a reference.

Citation context used in citation retrieval

O'Connor (O' Connor, 1982, 1983) assumed that citing statements give some information about the cited document. Cue words were extracted from the citation context and applied as index terms for the cited document. Then these index terms were used to improve the search effectiveness. Bradshaw (Bradshaw, 2003) proposed a Reference Directed Indexing (RDI) method to improve information retrieval of scientific literature. RDI also used a similar method to O' Connor's to create index terms from citation contexts. RDI considered both the relevance of a document to the query terms and the number of papers citing it.

Mercer and Di Marco also described their work on using citances to improve indexing tools for biomedical literature (Mercer, 2004). The first step of their work was using cue phrases in citances to predefine the citation classification. Then they applied these classifications to improve existing citation indexes. Ritchie (Ritchie, 2008) take the explicit, content words from citation contexts and index them as part of the cited document. And the results showed that the citation-enhanced document representation increases retrieval effectiveness across a range of standard retrieval models and evaluation measures.

Our reference retrieval module is similar to RDI, but we directly use the citation contexts as the retrieval field and rank results according to frequency of references which are corresponding to the citation contexts. The advantage of this approach is that the citation contexts could reveal the citation values of a reference.

Data and Method

Our procedure consists of four major components: 1. Data collection, 2. Citation context extraction, 3. Index creation, and 4. Retrieval system design (See Figure 1).

Data collection

All full text papers in PubMed Central were selected in this research. The data was downloaded on July 23 2012. There are 3431 journals with 622801 papers. All of these papers and their references were used to build the database for citation retrieval.

Papers published on December 2012 in BMC Bioinformatrics were chosen as the test dataset. There are 26 papers and 751 citation contexts.

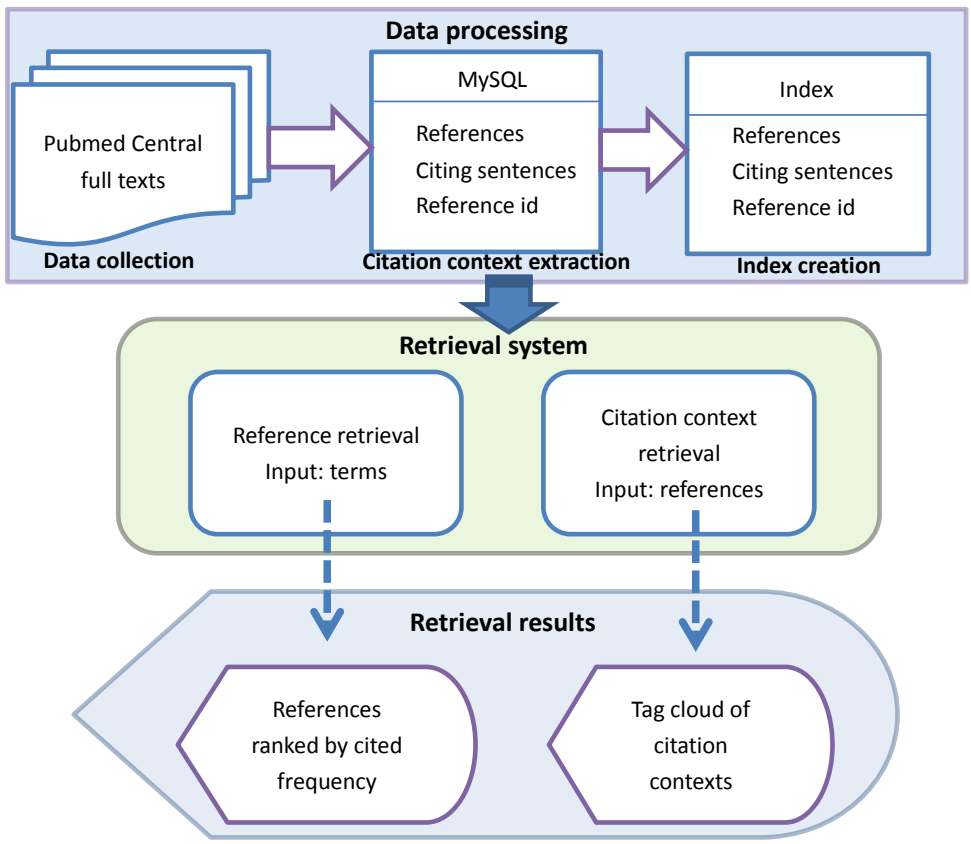


Figure 1. The citation retrieval system design

Citation context extraction

The full text literatures in PubMed Central are XML files. Figure 2 shows an example of a XML file with reference information. The citation context and its corresponding reference information are extracted and saved in MySQL database. In this paper, citation context is defined as one citing sentence with the reference tag. 17551920 citing sentences were extracted from 622801 papers.

Index creation

The aim of creating an index is to speed up the retrieval. Although citing sentences are stored in MySQL, the retrieval speed is very slow due to the large size of the citation context dataset. Therefore, indexing is necessary in this research. Lucene v3.5 is employed to create indexes for the retrieval field of citation context and cited reference.

Retrieval System design

The system includes two modules. One is the reference retrieval module; the other is the citation context retrieval module.

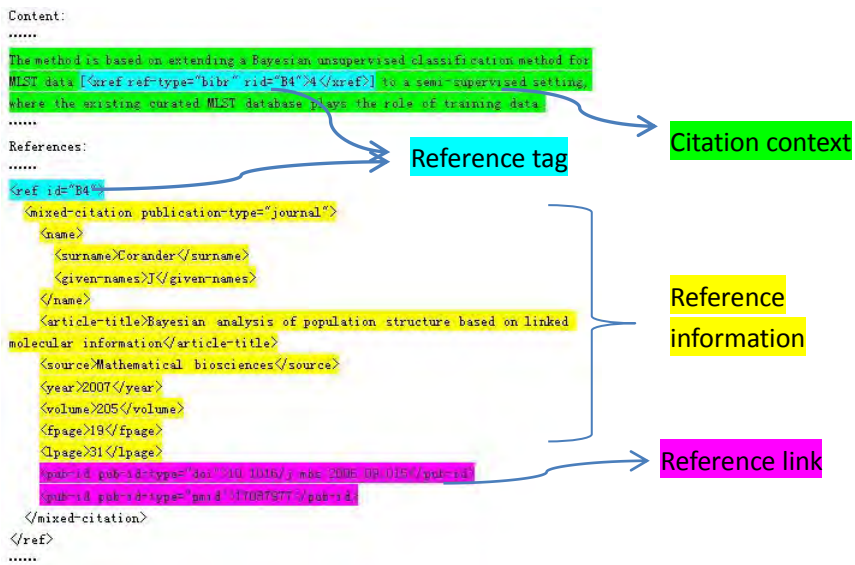


Figure 2. Extracting citation context from XML files

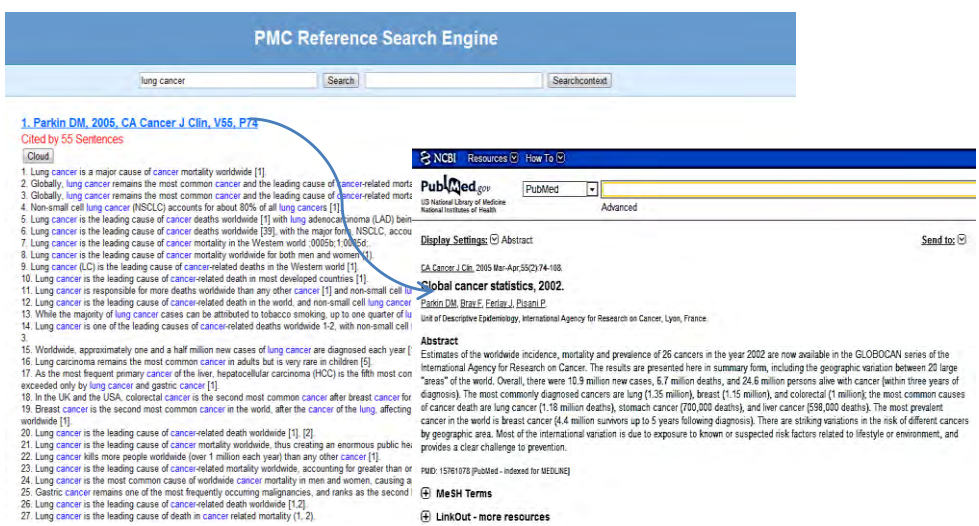


Figure 3. An example of reference retrieval

1）Reference retrieval module

In this module, the retrieval field is the citation context. The indexes of 17551920 citation contexts have been created. Researchers use topic terms to search the relevant citation contexts. But the citation contexts are not the final results. The references corresponding to these citation contexts are the results that researchers want to get. Each citation context corresponds to one or more references. The results will be ranked by corresponding counts of the citation context. The higher corresponding counts are, the more papers cite this reference on the querying topics . Each retrieved reference has a unique reference link to the title and abstract of the reference. Figure 3 shows an example of retrieval references related to "lung cancer". "Parkin DM,2005,CA Cancer J Clin,V55,P74" ranked first in the results. It was cited by 55 sentences, which means that "Parkin DM,2005,CA Cancer J Clin,V55,P74" was cited 55 times on the topic of "lung cancer". The general information about this paper can be found through the linkage. "Parkin DM,2005,CA Cancer J Clin,V55,P74" might also have been cited numerous of times on other topics. The citation context retrieval module which we discussed later provides the total cited times and topics of a chosen reference.

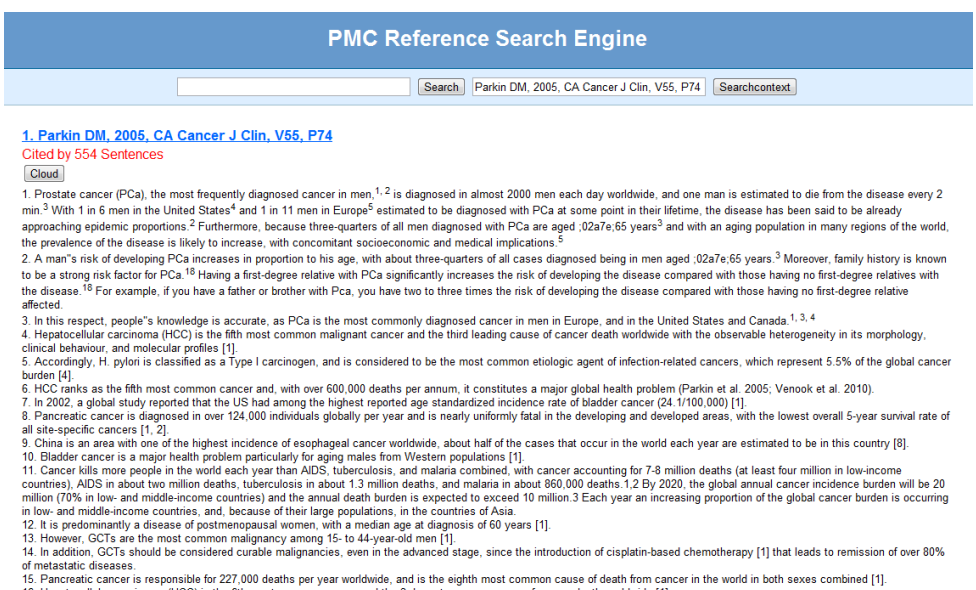


Figure 4. An example of citation context retrieval

2）Citation context retrieval module

The retrieval field of this module is the reference field. Researchers could use author, year, and/or journal information to find target references. The results show the citation frequency and citation contexts of the references. One reference could have one hundred citation contexts or even more. It is time consuming to

distinguish topics in these citation contexts manually. Tag cloud is employed to represent the citation contexts with topic terms in this module. Tag cloud (word cloud or text cloud) is a visual representation for text data, typically used to depict keyword metadata (tags) on websites, or to visualize free-form text. Tags are usually single words, and the importance of each tag is shown with font size or color (Halvey, 2007). An example is showed in Figure 4. The reference "Parkin DM,2005,CA Cancer J Clin,V55,P74" is used in this example which is the one we used in the reference retrieval module. 554 citation contexts have been retrieved. The reference retrieval module has retrieved 55 of 554 citation contexts related to "lung cancer". The other citation topics of this reference were represented in a tag cloud. Figure 5 shows the tags cloud of the citation contexts with single words. The main citation topic of this reference is the common causes of cancer death. The citation subtopics involve different kinds of cancer, different countries and genders that cancer occurs. Lung cancer is just one aspect of the citation topics.

Figure 5. Tag cloud of citation contexts

3) Test

In order to check the performance of the retrieval system, 26 new papers with 751 citation contexts from BMC Bioinformatrics were collected. The topics of each citation context were identified with 1-4 topic words manually. For example, the sentence "As a feature of reaction rules, some techniques focus on physicochemical properties and structures [25]" will be tagged with "physicochemical", "properties", and "structures". These topic words are used as retrieval terms to search references. Not all the citation contexts have topic words: for example, "It evolves the two different populations within the context of each

other [11][13]". The citation topic of this reference might have been expressed in the sentences before or after this citation context. The dataset was divided into four groups by time period, in order to check the influence of time. We chose 50 citation contexts which have explicit topic words for each period. The papers published earlier tend to receive more citations. So we expect that the retrieval system will perform better on the early time period. If the corresponding reference of a citation context appears among the top 10 retrieval results, we mark this retrieval as a successful retrieval. Otherwise, we mark it as an unsuccessful retrieval.

Results

The testing result of the retrieval system is shown in Table 1. The testing data was separated into four time periods based on the number of references in each year. The four periods are 1973-2000, 2001-2005, 2006-2008, and 2009-2011. The results show that the retrieval system performs very well for the early time period with the accuracy rate of 68% which is higher than the CRM-crosscontext method performs (He, 2010). The CRM-crosscontex is a citation recommendation method with the precision 42%. For the period 2001-2005 and 2006-2008, the accuracy rates are the same. They both have reached 60% which is a little lower than 1973-2000. For the most recent time period, the system did not perform very well. The accuracy rate of this period is only 38% which is the lowest in the four time periods.

Table 1. Retrieval performance of the retrieval system

	1973-2000	*2001-2005*	*2006-2008*	*2009-2011*	*total*
Successful	34	30	30	19	113
Unsuccessful	16	20	20	31	87
Accuracy rate	68%	60%	60%	38%	56.5%

Table 2 shows 10 instances of the successfully retrieved topics and references. The topics are extracted from citation contexts and the original references that the citation contexts used are ranked the first in all retrieval results respectively. Most of these successfully retrieved topics are about tools and methods. The highly cited conclusions could also be retrieved successfully. For example, "Han JD, 2004, Nature, V430, P88" is retrieved on topic "data party hubs". This paper was cited 100 times on this topic.

Although some of the citation contexts with explicit topics were not retrieved successfully, it did not mean that the retrieval system does not fit for these topics. Table 3 shows three examples of comparisons of the original references with the recommended references retrieved from our system on the same topics. The testing dataset used "Chang CC, 2011, ACM Trans. Intell. Syst. Technol, V2" as the reference of topic "LIBSVM". But our system recommended another paper of Chang's which was published in 2001 and received 34 citations on topic

"LIBSVM". For topic "BLAST e-value", the original reference was Karlin's paper which had just one citation on this topic. The recommended reference had been cited 66 times on this topic. It is hard to judge which reference is better. It is impossible to read all the related articles while we are conducting our research. Our recommended references are retrieved based on the behavior of all other authors. Our system definitely has some value which cannot be ignored.

Table 2. 10 instances of successful retrieved topics

Topics	*References*	*Freq*
Weblogo	Crooks GE, 2004, Genome research, V14, P1188	376
Date party hubs	Han JD, 2004, Nature, V430, P88	100
BiMax	Prelic A, 2006, Bioinformatics, V22, P1122	40
PredictNLS	Cokol M, 2000, EMBO Rep, V1, P411	20
SVMLight	Joachims T, 1999, Making large-scale SVM learning practical	11
Bron-Kerbosch algorithm	Bron C, 1973, Commun ACM, V16, P575	10
Amino acid compositions	Hua S, 2001, Bioinformatics, V17, P721	7
PMSprune	Davila J, 2007, TCBB, V4, P544	6
APBioNet	Tan TW, 2010, BMC Genomics, V11, PS27	5
ChemicalTagger	Hawizy L, 2011, J Cheminf, V3, P17	4

Table 3. Comparison of original references and retrieved references

Topics	*Sources*	*References*	*Freq*
LIBSVM	Original	Chang CC, 2011, ACM Trans. Intell. Syst. Technol, V2	3
	Retrieved	Chang CC, 2001, LIBSVM: a library for support vector machines	34
Graphviz	Original	Ellson J, 2001, Lecture Notes in Computer Science Springer-Verlag, P483	0
	Retrieved	Ellson J, 2003, Graph Drawing Software, P127	5
BLAST e-value	Original	Karlin S, 1990, Proceedings of the National Academy of Sciences of the United States of America, V87, P2264	1
	Retrieved	Altschul SF, 1990, J Mol Biol, V215, P403	66

Discussion

The retrieval system designed in this paper is based on the large amount of full text papers in PubMed Central. Most of the databases do not provide the full texts. Therefore, the retrieval system in this paper is particularly suitable for the field of biomedicine. With the development of information science, the citation retrieval system will extend to other fields where full text databases are available.

The reference retrieval module shows its effectiveness on searching papers published early and papers with high citation frequencies which is what we expected. It is also very effective in retrieving papers that regarding introduce methods or tools. The reference retrieval module will perform better on retrieving

basic or classic papers in a specified field. But papers with lower citation frequencies will be hard to find in this system, since the retrieval field of this module is citation context.

The citation context retrieval module provides all the citation contexts of a specific reference. These citation contexts may contain many topics. Tag cloud is employed to represent these topics. The topics of the citation contexts greatly enhance the meaning of a reference. The retrieval results enrich our understanding of which knowledge claims by the references have been used and have had the greatest impact on subsequent work, and also what criticisms have been leveled against their claims. They also can be used to evaluate the impact of a reference together with the citation frequency.

A test version of the literature retrieval system is available on the World Wide Web at: http://ir.dlut.edu.cn:8090/PMCSEARCH/.

Conclusion

We designed a literature retrieval system based on citation contexts extracted from full text publications in biomedicine. The reference retrieval module is for searching publications which have been cited on topics related to the querying terms. The citation context retrieval module is for searching the citation contexts of a specific paper and for visualizing the contributions of the specific paper in a tag cloud. The results showed that this retrieval system was particularly accurate in retrieving highly cited papers and classic papers, whereas the accuracy was reduced when searching less cited papers and newly published papers. The citation context retrieval module could identify different citation topics of a reference. In summary, our work demonstrates the potential of using citation contexts in enhancing the retrieval of scientific publications and improving our understanding of the impact of a specific publication on subsequent work.

Acknowledgments

This research is supported by National Natural Science Foundation of China (grant number 71003011, 61272370), Fundamental Research Funds for the Central Universities of China (DUT12RW414), the specialized research fund for doctoral tutor (20110041110034) . Part of the research was conducted during Shengbo Liu's visiting doctoral studentship at the iSchool at Drexel University. Thanks to Howard White for the suggestions.

References

Anderson, M. H.& Sun, P.Y. T. (2010). What have scholars retrieved from Walsh and Ungson (1991)? A citation context study. *Management Learning,* 41, 131-145.

Boyack, K. W., Small, H.& Klavans, R. (2012). Improving the Accuracy of Co-citation Clustering Using Full Text. *Journal of the American Society for Information Science and Technology,* preprint.

Bradshaw, S. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. *Proceedings of the 7th European conference on digital libraries* (pp.499-510), Trondheim: Springer.

Callahan, A., Hockema, S., & Eysenbach, G. (2010). Contextual Cocitation: Augmenting Cocitation Analysis and its Applications. *Journal of the American Society for Information Science and Technology,* 61, 1130-1143.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D. & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology,* 59, 51-62.

Eto, M. (2012). Evaluations of context-based co-citation searching. *Scientometrics, Preprint*, 1-23.

Gipp, B. & Beel, J. (2009b). Identifying related documents for research paper recommender by CPA and COA. *Proceedings of international conference on education and information technology* (pp.636-639), Berkeley: International Association of Engineers.

Halvey, M. & Keane, K. (2007). An Assessment of Tag Presentation Techniques. *The 16th International World Wide Web Conference*. Banff : IW3C2.

He, Q., Pei, J. & Kifer,D. (2010). Context-aware Citation Recommendation. *The 19th International World Wide Web Conference* (pp.421-430). Raleigh: IW3C2.

Hunter, L. & Cohen, K. (2006). Biomedical language processing: What' s beyond pubmed? *Molecular Cell, 21*, 589-594.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*, 10-25.

Liu, S. & Chen, C. (2012). The proximity of co-citation. *Scientometrics, 91*, 495-511.

Mei, Q. & Zhai, C. (2008). Generating impact-based summaries for scientific literature. *The Proceedings of ACL '08* (pp.816-824). Columbus: ACL.

Mercer, R. E. & Marco, CD. (2004). A design methodology for a biomedical literature indexing tool using the rhetoric of science. *The BioLink workshop in conjunction with NAACL/HLT* (pp.77-84). Boston: Association for Computational Linguistics.

Mohammad, S., Dorr, B., Egan, M., Hassan, A.,Muthukrishan, P., Qazvinian, V., Radev, D. & Zajic, D. (2009). Using citations to generate surveys of scientific paradigms. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp.584-592). Boulder: Association for Computational Linguistics.

Nakov, P. I., Schwartz, A.S. & Hearst, M.A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. *SIGIR 2004Workshop on Search and Discovery in Bioinformatics*. Sheffield: SIGIR.

Nanba, H., Kando, N. &Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article

generation. *Proceedings of the American society for information science* (pp.117-134). Chicago: ASIS.
Nanba, H. & Okumura, M. (1999). Towards multi-paper summarization using reference information. *The 16th International Joint Conference on Artificial Intelligence* (pp.926-931). Stockholm: IJCAI.
Nanba, H. & Okumura, M. (2005). Automatic detection of survey articles. *The Research and Advanced Technology for Digital Libraries*. Berlin: IJCAI.
O'Connor, J. (1982). Citing statements: Computer recognition and use to improve retrieval. *Information Processing and Management, 18*, 125-131.
O'Connor, J. (1983). Biomedical citing statements: Computer recognition and use to aid full-text retrieval. *Information Processing and Management, 19*, 361-368.
Pao, M. L. (1993). Term and citation retrieval: A field study. *Information Processing and Management, 29*, 95-112.
Ritchie, A. (2008). *Citation context analysis for information retrieval.* University of Cambridge, New Hall.
Siddharthan, A. & Teufel, S. (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. *Proceedings of NAACL/HLT-07*. Rochester : Association for Computational Linguistics.
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American society for information science and technology, 24*, 265-269.
Small, H. (1979). Co-citation context analysis: The relationship between bibliometric structure and knowledge. *Proceedings of the ASIS Annual Meeting* (pp.270-275), Medford: Information Today.
Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. *Journal of the American Society for Information Science, 37*, 97-110.
Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investgation. *Scientometrics, 87*, 373-388.

AUTHOR INDEX

Partners

ISBN: 978-3-200-03135-7
ISSN: 2175-1935