The Research of Paper Influence Based on Citation Context -- A Case Study of the Nobel Prize Winner's Paper

Shengbo Liu[1], Kun Ding[1], Bo Wang[2], Delong Tang[1], Zhao Qu[1]

[1] *liushengbo1121@gmail.com, dingk@dlut.edu.cn, tangdl@mail.dlut.edu.cn, qz_031@mail.dlut.edu.cn*
WISElab, Dalian University of Technology, No. 2, Linggong Road, Ganjingzi district, Dalian, 116024, China

[2] *bowang1121@gmail.com*
School of Management, Dalian Ploytechnic University, #1 Qinggongyuan, Dalian, 116000,China

Abstract
Citation context was used to measure the influence of highly cited papers. The themes of citation context were analyzed with bibliometrics methods. The citation context was classified into three categories as positive, negative and neutral. And the neutral citations were also classified into three sub categories, related work in background or introduction, theoretical foundation, and experimental foundation. The citation contexts of a highly cited paper of O'Keefe were extracted as the experiment data set. The results showed that the co-occurrence method was very useful for describing the themes of the citation contexts. The citation contexts of the selected paper were divided into five themes. The classification of citation contexts could provide more information about how and why a paper was highly cited. There was no negative citation in this experiment, and more than 10% citation contexts were positive citation. About 50% of the neutral citations were belonging to related work in background or introduction. The detailed influence of the target paper was also illustrated in our research.

Conference Topic
Citation and co-citation analysis

Introduction
Citation frequency is a commonly used indicator to measure the importance of a paper. Recently, *Nature* asked Thomson Reuters, which now owns the SCI, to list the 100 most highly cited papers published from 1900 to 2014. The results revealed some surprises, many of the world's most famous papers do not rank in the top 100 (Van Noorden, Maher, & Nuzzo, 2014). John P. A. Ioannidis and colleagues surveyed the most-cited authors of biomedical research for their views on their own influential published work. The results showed that the most important paper was indeed one of author's most-cited ones. But they described most of their chart-topping work as evolutionary, not revolutionary (Ioannidis, Boyack, Small, Sorensen, & Klavans, 2014). Although the citation frequency is an important indicator to measure the influence of a paper, it is hard to reveal why others always cited this paper and what influence it makes. Citation context refers to the text surrounding the references (Henry Small, 1982). It could provide more detailed information about citation.

In this paper, we take John O'Keefe's (Nobel Prize winner in Physiology or Medicine 2014) most influence paper as instance. The influence of this paper will be analyzed based on citation context. Our analysis will provide a richer understanding of which knowledge claims made by O'Keefe have had the greatest impact on later work.

Related work

Citation context analysis

Citation context can be defined as the sentences that contain the citation of a particular reference. For example, the sentence "This comparison is made using BLASTX [18]" is the citation context of reference [18].

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 (A. Siddharthan, Teufel, S., 2007). Nanba and Okumura (Nanba, 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. Elkiss et al. (Elkiss, 2008) generated the citation summarization based on citation context to describe the topic of cited paper. Mei (Mei, 2008) and Mohammad (Mohammad, 2009) found that the summarization of citation contexts is very different from the abstract of the cited reference. Liu and Chen(Liu & Chen, 2013) studied the differences between latent topics in abstracts and citation contexts. The results showed that topics from citing sentences tend to include more specific terms than topics from abstracts of citing papers. Nakov (Nakov, 2004) referred to citation contexts as citances – a set of sentences that surrounding 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 (H. Small, 1979) studied the context of co-citation and analyzed the context in which the co-citation paper mentioned. In addition, he analyzed the sentiment of the co-citation context (H. Small, 2011).

Anderson (Anderson, 2010) analyzed the citation contexts 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. Chang(Chang, 2013) compared the citing topics of *Little Science*, *Big Science* in natural sciences and humanities and social sciences through citation context. He found that the citing topics in natural sciences and humanities and social sciences were very similar, but the cited motivation had some differences.

The classification and function of citation context

Citation context contains the direct related information between cited paper and citing paper. It could be used to reveal the nature of a citation. The cited motivation of each citation is different, so the value of each citation will be different. For example, some of the citation contexts support the claims in the cited paper, and some of them may take the opposite opinion about the views or methods in the cited paper. Spiegel-Rösing (Spiegel-Rösing, 1977) studied the citation context of Science Study in 1977 and classify the citation context into 13 categories, including use the data of cited paper, use the method of cited paper, compare the work of cited paper and citing paper and so on. In order to provide more information for literature management, Teufel reclassified the above 13 categories into four categories, (1) Explicit statement of weakness, (2) Contrast or comparison with other work, (3) Agreement /usage /compatibility with other work, (4) A

neutral category(Teufel, Siddharthan, & Tidhar, 2006) . Cue phrases were used to identify the category of each citation context. The similar method was also employed in Liu's (Liu et al.) work in which the citation context was classified as positive citation, negative citation, and neutral citation. Other people like Small (Henry Small, 1982), McCain (McCain & Turner, 1989), Siddharthan (A. Siddharthan & Teufel, 2007), Swales (Swales, 1990) also did some work about citation context classification.

Data and Method

Our procedure consists of three major components, 1. Data collection and preprocessing, 2. Theme analysis of citation context, and 3. The classification of citation context. Details are explained in corresponding sections.

Data collection and preprocessing

The 2014 Nobel Prize in Physiology or Medicine is awarded to Dr. John M. O'Keefe, Dr. May-Britt Moser and Dr. Edvard I. Moser for their discoveries of nerve cells in the brain that enable a sense of place and navigation. The scientific background was introduced in the document "The Brain's Navigational Place and Grid Cell System" . The keywords this document were selected manually and used to retrieve the award field in Web of Science. The search query was shown as follows：

TI=(hippocamp* AND (place OR Position* OR spatial)) OR (("grid cell*" OR Position* OR Navigation* OR spatial OR place) And ("entorhinal cortex" OR brain OR cerebral))

The time period was from 1945 to 2014, and 4441 papers were collected.

The citation context collection was built through three steps. First, the paper with the first author O'Keefe and the highest citation frequency was selected. Second, the papers which cited the chosen paper were downloaded with full text. Actually, we could just find less than 20% full text papers. Last, the citation contexts of the chosen papers were extracted from the full text for further analysis. The extraction method has been introduced in our previous work (Liu & Chen, 2013).

The theme analysis of citation context

The theme analysis includes two tasks. One is counting the frequency of noun phrases appeared in citation contexts. Another is mapping the co- occurrence network of noun phrases.

Part-of-speech is needed before extract noun phrases. There are many tools to label part-of-speech, such as PosTagger, CLAWS POS tagger. Stanford Log-linear Part-Of-Speech Tagger (Toutanova & Manning, 2000) was employed in this work, which was developed by NLP group of Stanford University. The noun phrase formation rules was designed with the same method described in Wang's paper (Wang, Liu, Ding, Liu, & Xu, 2014). When counting the frequency of noun phrases. If one citation context contains two same noun phrases, it will count once.

In bibliometrics analysis, co-occurrence method was often used to detect subjects/themes (Hofer, Smejkal, Bilgin, & Wuehrer, 2010; Lee, 2008; Zhang et al., 2012). But few of the related works use this method to detect the theme of citation context. Pajek software was employed to mapping the noun phrases co-occurrence network of citation context. We expect to identify the citing themes through drawing the co-occurrence map.

The classification of citation context

Following the work of Spiegel-Rösing (Spiegel-Rösing, 1977) and Teufel (Teufel et al., 2006), citation contexts will be classified into three categories as positive, negative and neutral. Table 1 shows the description of each category. We divided the positive category into three sub categories and the negative category into two sub categories.

Table 1. The description of each category

Category		Description
Positive	(1)	Affirm or praise the cited work
	(2)	Apply the methods, tools or databases of the cited paper
	(3)	Comparison of methods and results
Negative	(1)	Point out the weakness of the citation
	(2)	Contain negative cue words
Neutral	(1)	Contain no cue words

To our knowledge, the proportion of neutral citations occupy more than others. So we will classify the neutral citation into three sub categories based on the citation motivation.

(1) Related work in background or introduction. Introduce the related work with no comments.
(2) Theoretical foundation. Concepts, principles, methods, or results which will be used in citing paper.
(3) Experimental foundation. Including experimental conditions, processes, environment, and results.

Results and discussion

Target paper detecting

Table 2 shows top ten highly cited papers in Nobel Prize award field. The highest cited paper was "PLACE NAVIGATION IMPAIRED IN RATS WITH HIPPOCAMPAL - LESIONS" which published in *Nature* in 1982. It has been cited 3589 times. Although this paper got highly cited in Nobel Prize award field, it did not appear in "Scientific background" document, which was the instruction of why the winner got this prize. The author Morris R.G.M did not get Nobel Prize. The Nobel Prize was given to the author of the second highest cited paper "HIPPOCAMPUS AS A SPATIAL MAP - PRELIMINARY EVIDENCE FROM UNIT ACTIVITY IN FREELY-MOVING RAT". The result is similar to the work of Van Noorden (Van Noorden et al., 2014) that the Nobel Prize winner's paper did not get the highest citation frequency.

O'Keefe who is the Nobel Prize winner had three papers ranked in top ten high cited papers in Nobel Prize award field. The highest cited paper had been cited 1812 times. This paper was selected as the target paper. The seminal work of this paper was the discovery of "place cell".

It is hard to download all the 1812 citing papers. So 200 citing papers with full text were selected in our experiment. There were 228 citing sentences. The target paper was average cited 1.14 times in each citing paper.

Table 2. Top ten high-cited papers in Nobel Prize award field.

Author	Title	Journal	Year	Cited frequency
Morris, R. G. M., P. Garrud, et al	PLACE NAVIGATION IMPAIRED IN RATS WITH HIPPOCAMPAL-LESIONS	Nature	1982	3589
Okeefe, J. and Dostrovs.J	HIPPOCAMPUS AS A SPATIAL MAP - PRELIMINARY EVIDENCE FROM UNIT ACTIVITY IN FREELY-MOVING RAT	Brain Research	1971	1812
Okeefe, J. and M. L. Recce	PHASE RELATIONSHIP BETWEEN HIPPOCAMPAL PLACE UNITS AND THE EEG THETA-RHYTHM	Hippocampus	1993	1033
Tsien, J. Z., P. T. Huerta, et al	The essential role of hippocampal CA1 NMDA receptor-dependent synaptic plasticity in spatial memory	Cell	1996	919
Grant, S. G. N., T. J. Odell, et al	IMPAIRED LONG-TERM POTENTIATION, SPATIAL-LEARNING, AND HIPPOCAMPAL DEVELOPMENT IN FYN MUTANT MICE	Science	1992	827
Hafting, T., M. Fyhn, et al	Microstructure of a spatial map in the entorhinal cortex	Nature	2005	773
Cohen, L., S. Dehaene, et al	The visual word form area - Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients	Brain	2000	755
Burgess, N., E. A. Maguire, et al	The human hippocampus and spatial and episodic memory	Neuron	2002	669
Packard, M. G. and J. L. McGaugh	Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning	Neurobiology of Learning and Memory	1996	666
Okeefe, J	PLACE UNITS IN HIPPOCAMPUS OF FREELY MOVING RAT	Experimental Neurology	1976	657

The themes of citation context

299 noun phrases were extracted from the citation contexts. Table 3 listed twenty high frequency noun phrases. The term “place cell” got the highest frequency of 76, because the most contributing work of the target paper was the discovery of place cell. Hippocampus, environment, rat, fire, neuron were all the important terms in target paper. Some of the terms were not mentioned in the target paper, such as cognitive map and ca3.

Table 3. Top twenty high cited papers in Nobel Prize award field.

No.	Noun phrase	Frequency	No.	Noun phrase	Frequency
1	place cell	76	11	discovery	17
2	hippocampus	74	12	place field	15
3	environment	55	13	rodent	13
4	rat	44	14	ca3	13
5	animal	40	15	space	12
6	cell	31	16	ca1	12
7	location	29	17	position	11
8	fire	25	18	pyramidal cell	9
9	cognitive map	19	19	region	9
10	neuron	18	20	navigation	9

Figure 1 showed the co-occurrence map of the noun phrases. Each node represents a noun phrase. The size of the node was proportional to the number of terms co-occurred with it. We set the co-occurrence threshold as more than once and got 71 nodes in the map.

The map could divide into five parts manually based on the relationship of terms. Part A was mainly involving navigation, which was not mention too much in cited paper. It was the following research of place cell. Part B was related to neuron region, including CA1 and CA3. CA1 was discussed in the cited paper, but CA3 was found in the later work. Part C was related to experimental process about firing pattern of rat. Part D was the experimental environment. The definition of place field was widely cited. Part E was about the concept of place cell.

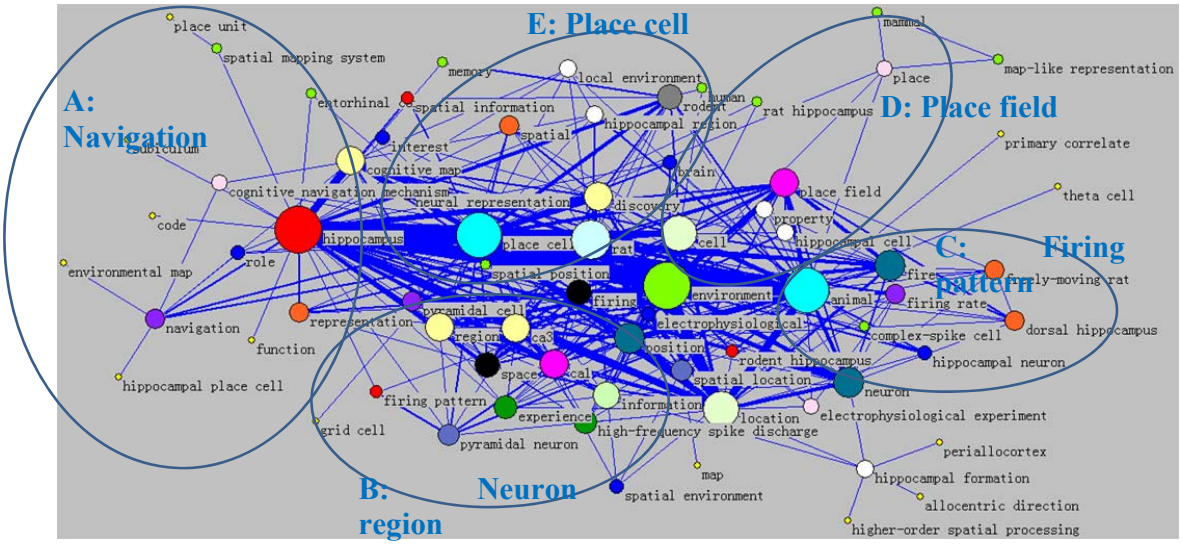


Figure 1. Co-occurrence map of the noun phrases.

Table 4. Example of positive citations

No.	Positive citation
1	The discovery of place cells [1]-[5] in the hippocampal regions of rats **consolidated** the idea that hippocampus probably represents a cognitive map of the local environment of an animal……
2	The concept of cognitive map for navigation, carried out mainly by Tolman [10], was **fuelled** by the discovery of the so-called place cells in the hippocampus of the rat and has widely increased our understanding of cognitive navigation mechanisms [11]
3	The **breakthrough** came in 1971 with the discovery of the rat s cognitive map in the cells of the hippocampus [16]……
4	The idea of the formation of a cognitive map was first proposed by Tolman [45] in the late 40s and was later **supported by the discovery** of place cells by o keefe and dostrovsky [35]
5	The **striking discovery** of place cells in the rat hippocampus [51] has triggered a wave of interest on spatial learning that holds until today

Table 5. Sub categories distribution of neutral citations

Category	Related work	Theoretical foundation	Experimental foundation
Counts	114	49	41

The classification results

The classification results showed that most of the citations were neutral citation. There was no negative citation in our datasets. 24 of 228 citation contexts were positive citations and 204 citations were neutral citations. Table 4 listed some examples of positive citations.

The sub categories distribution of neutral citations was shown in table 5. Nearly half of the citations were cited as related work. Theoretical foundation had 49 citations, and most of them were related to place cell or place field. 41 of 204 neutral citations were classified into experimental foundation, including ca1 neuron fire experiment, rodent studies and so on.

Conclusion and discussion

Citation context was used to measure the influence of paper in this research. The influence was identified from two aspects, the theme of the citation context and the classification of the citation context. The results showed that the traditional bibliometrics methods could be utilized in identify the themes of citation context. The citation contexts were divided into five themes in our experiment. The classification results showed that there were no negative citations of O'Keefe's most influential paper. More than 10% citation contexts were positive citations.

Through the citation context analysis of the influence paper, the detailed influence of the high influence paper could be revealed. The influence themes are more wide than the abstract of the target paper and the proportion of the positive citations takes more account than it appears in some journals (Liu et al., 2014).

There is only one case study in this paper. Although we could get some insightful results from this case study, comparative experiments are still needed in our future work.

Acknowledgments

This research is supported by National Natural Science Foundation of China (grant number 61272370), the specialized research fund for doctoral tutor (20110041110034), and the ISTIC-THOMSON REUTERS Joint Laboratory Open Foundation (IT201002). Part of the research was conducted during Shengbo Liu's visiting doctoral studentship at the iSchool at Drexel University. Thanks to the reviewers for the kindly 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*(2), 131-145.

Chang, Y.-W. (2013). A comparison of citation contexts between natural sciences and social sciences and humanities. *Scientometrics, 96*(2), 535-553.

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*(1), 51-62.

Hofer, K. M., Smejkal, A. E., Bilgin, F. Z., & Wuehrer, G. A. (2010). Conference proceedings as a matter of bibliometric studies: the Academy of International Business 2006–2008. *Scientometrics, 84*(3), 845-862.

Ioannidis, J. P. A., Boyack, K. W., Small, H., Sorensen, A. A., & Klavans, R. (2014). Bibliometrics: Is your most cited work your best? *Nature, 514*(7524), 561-562.

Lee, W. H. (2008). How to identify emerging research fields using scientometrics: An example in the field of Information Security. *Scientometrics 76*(3), 503.

Liu, S., & Chen, C. (2013). The differences between latent topics in abstracts and citation contexts of citing papers. *Journal of the American Society for Information Science and Technology, 64*(3), 627-639.

Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., & Lin, Y. (2014). Literature retrieval based on citation context. *Scientometrics, 101*(2), 1293-1307.

McCain, K. W., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics, 17*(1), 127-163.

Mei, Q. & Zhai, C. (2008). Generating impact-based summaries for scientific literature. *Proceedings of ACL '08*, Columbus.

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*, Boulder.

Nakov, P. I., Schwartz, A.S. & Hearst, M.A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. *SIGIR 2004 Workshop on Search and Discovery in Bioinformatics,* Sheffield.

Nanba, H. & Okumura, M. (1999). Towards multi-paper summarization using reference information. *16th International Joint Conference on Artificial Intelligence*, Stockholm.

Nanba, H., & Okumura, M. (2005). Automatic detection of survey articles. *The Research and Advanced Technology for Digital Libraries*, Berlin.

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.

Small, H. (1979). Co-citation context analysis: The relationship between bibliometric structure and knowledge. *Proceedings of the ASIS Annual Meeting,* Medford.

Small, H. (1982). Citation context analysis. *Progress in communication sciences, 3*, 287-310.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investgation. *Scientometrics, 87*(2), 373-388.

Spiegel-Rösing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science*, 97-113.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press (Cambridge England and New York).

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*.

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature, 514*(7524), 550-553.

Wang, B., Liu, S., Ding, K., Liu, Z., & Xu, J. (2014). Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology. *Scientometrics, 101*(1), 685-704.

Zhang, J., Xie, J., Hou, W., Tu, X., Xu, J., Song, F., Wang, Z. & Lu, Z. (2012). Mapping the knowledge structure of research on patient adherence: Knowledge domain visualization based co-word analysis and social network analysis. *PloS One, 7*(4), e34497.

Time to First Citation Estimation in the Presence of Additional Information

Tina Nane

g.f.nane@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 905, 2300 AX, Leiden (The Netherlands)

Abstract
We are interested in modelling the time to first citation, that is how long does it take for a publication to be cited for the first time after it has been published in a journal. We argue that both cited and uncited publications should contribute to the distribution of the time to first citation. Moreover, our objective is to model the time to first citation nonparametrically, hence under no parametric assumption. Due to the similarities with the observed data in survival analysis, we employ the techniques based on censored data and describe the distribution of the time to first citation in terms of the hazard rate, that is the instantaneous rate of being firstly cited. We find that publications receive their first citation at increasing rates in the first 24 months after their publication date and at decreasing rates afterwards. Moreover, we observe that the hazard rate and hence the time to first citation is influenced by the document type, number of authors and collaboration type and field. We also investigate the difference in the time to first citations for publications grouped by their collaborative status or the assigned field.

Conference Topic
Citation and co-citation analysis

Introduction
The first citation a publication receives is an important event in the bibliometric data, as it is not only a simple citation count, but also marks a change in the status of the publication, i.e. from being uncited the publication becomes cited. Certainly, observing the first citation of a publication depends on the considered time frame. Regardless the period of analysis, certain publications will never receive their first citation, in other words we will not observe the first citation received by some publications for any finite time period we consider.

Another important aspect concerns the time it takes for a publication to receive its first citation. For some publications it takes a small amount of time, such as 1-2 months, while for others it can even take more than 10 years. Due to overlong reviewing and publication procedures, some publications might even have negative times to first citation, meaning that the publication has been cited before it has been published.

The event that a publication received its first citation, as well as the time to the first citation received considerable attention over the years, starting with Schubert and Glänzel (1986), Glänzel (1992), Rousseau (1994), Glänzel and Schoepflin (1995). Since 2000, Egghe (2000), Egghe and Rao (2001), Burrell (2001), and Glänzel et al. (2012) continued to model the first citation data. Additionally, we acknowledge the work of van Dalen and Hekens (2005) and Bornmann and Daniel (2010), that is specifically close to the present research and will be referred to later on. Most of the previous work relies on the parametric modelling of the time to first citation distribution, such as the double exponential model (Rousseau, 1994), mixtures of non-homogeneous Poisson process

(Burrell, 2001), etc. The modelling in the existing literature focuses only on publications in certain journals or fields and the uncited publications do not always contribute to the time to first citation distribution, yet they emerge as a consequence of the model (Burell, 2001). Additionally, in Egghe (2000), the proportion of the uncited documents emerges from the model.

It should be stressed however that the time to first citation distribution derived from a set of publications that contains both uncited and cited documents does not coincide with the time to first citation distribution of the publications that receive a citation. From a probabilistic perspective, the first distribution is the sub-distribution of the latter. Furthermore, not accounting for the uncited publications can lead to biases in the estimation of the distribution of the time to first citation.

Our present study aims to continue and extend the research on the time to first citation analysis. We consider all the publications, regardless the document type and field, that appeared in Web of Science (WoS) in 2000 and their first citations received until the end of 2013. The time to first citation is registered in months. Additional data is recorded for each publication, such as document type, the number of authors, institutions and countries, and information on collaboration.

We propose an approach that aims to model the time to first citation distribution by accounting for all observations (both uncited and cited publications). Our approach assumes that the event of interest is the first citation, which is time dependent and we are interesting in modelling the time to this event of interest, namely the time to first citation. The time to event analysis has been employed in many fields. In sociology, it is known as event history analysis, in economy as duration analysis and in engineering is called reliability theory. Nevertheless, it is best known in biostatistics, where most research has been performed and where it is called survival analysis.

Consequently, the terminology employed in survival analysis is ubiquitous. In biostatistics, a frequent event of interest is death and the time to the event is then expectedly called survival time. Different functionals of the distribution of the time to the event of interest are successively termed survival function, hazard or cumulative hazard function. We will employ this unfortunate terminology in the analysis of the time to first citation.

A typical feature of the data in survival analysis is that not all events of interest are observed within the period of analysis. These observations are referred to as censored observations. The uncited publications are therefore regarded as censored observations. The uncited publications are in fact right censored observations, since their first citation is conditioned to take place after the period of analysis ended, i.e. at the right of the period of analysis. This approach circumvents the issue of not having a time to first citation for the uncited publications.

In survival analysis, the distribution of the time to event data is usually characterized by its survival function, as well as its hazard rate. The hazard rate provides information on the evolution in time of the event rate, in our case first citation rate. An attractive feature of the hazard rate compared to the density function, for example, is that the hazard rate accounts for the aging effect, while the density does not. Based on our data, we provide the time to first citation distribution and investigate its behaviour via the hazard rate.

Another important aspect in survival analysis is how additional information on observations, referred to as covariates or explanatory variables influence the time to the

event of interest. The Cox model (Cox, 1972) is probably the most popular method to model the influence of covariates on the time to the event of interest. In this study, we aim to infer on the effect of different characteristics of publications on the time to first citation. In other words, is the document type, number of authors, collaboration type or the field of a publication influencing the time it takes for that publication to receive the first citation? To our best knowledge, the influence of the explanatory variables document type, collaboration or field have not been accounted so far in the time to first citation analysis.

These methods in survival analysis have been previously used to model the time to first citation distribution by van Dalen and Henkens (2005) and Bornmann and Daniel (2010). Both studies restrict themselves to publications in a specific area of research, i.e. demography and chemistry. van Dalen and Henkens (2005) propose to model the hazard rate of the time to first citation distribution under the parametric assumption of a Gompertz distribution, which, in turn, lead to hazard rate which are decreasing over time. This restriction is unintuitive and in particular, it does not fit the data of the present study. Bornmann and Daniel (2010) are very brief in explaining the methods and, more importantly, the results of the analysis are not consistent in presenting their results, as they first refer to the differences in the survival curves and later on to the differences in the hazard rate. It is not very clear, for example, if the publication characteristics have an effect on the hazard rate.

Time to first citation distribution

We consider all the publications in Web of Science (WoS) that appeared in 2000 and their first citations up until 2013. That accounts for 1,202,371 publications, from which 62.62% received their first citation until the end of 2013. The first citation of publication A is defined as the publication date (month) of a publication B that cites firstly publication A, that is the minimum publication date of all publications that cite publication A. Needless to say that since the study is restricted to WoS, we refer to the first citation covered by WoS. Moreover, we exclude self-citations, hence we condition on publication B having no common authors with publication A.

The time to first citation of publication A is the time period (in months) between the publication date of publication A and the publication date of a publication B that cites firstly publication A. The time to first citation can sometimes be negative, but this is mostly an artefact due to the slow reviewing or publication process in different journals, etc. We exclude such observation from our study.

We chose the publication date to be registered in months given the availability of the data, but also for a better insight in the first citation process. Moreover, this avoids the issue of highly discrete data. Nonetheless, it is noteworthy that the publication date in months is not available for all data. For these cases, the first month of the year (January) or the middle one (July) is usually reported.

The histogram of the time to first citation for the publications in WoS that appeared in 2000 and received their first citation within the period 2000-2013 is presented below.

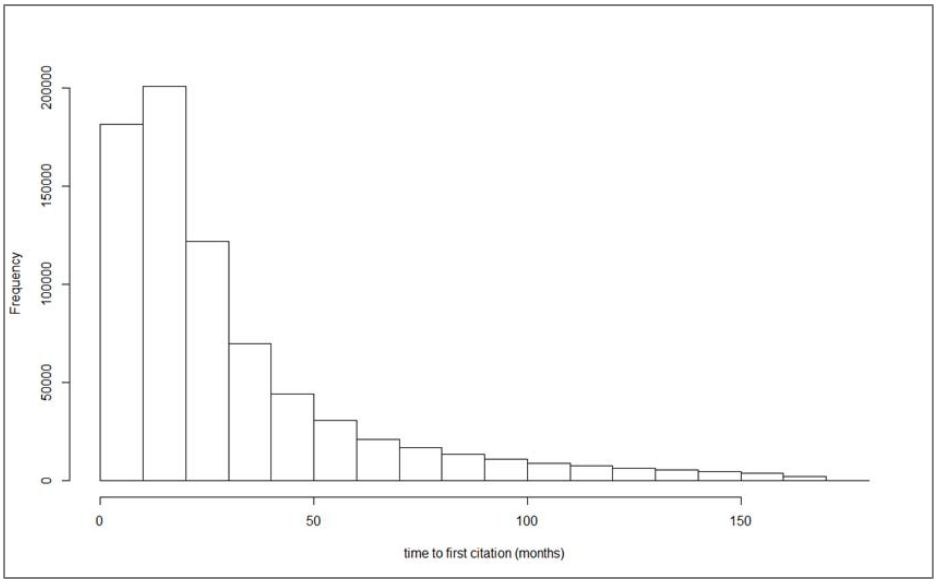


Figure 1. Histogram of the time to first citation for publications in 2010.

Most of the publications received their first citation shortly after publication. As expected, the proportion of publications that receive citations decreases over time. There are however publications that receive their first citation 13 years after their publication.

The histogram provides information on the time to first citation distribution of publications that received at least a citation until 2013. As mentioned beforehand, there is however no information on the publications that have not received any citation, apart from the percentage of the uncited publications.

Censored observations

It would be desirable though that the uncited publications also contribute to the distribution of the time to first citation, as they influence the probability of being firstly cited. Within this framework, the uncited publications did not experience the event of interest (first citation) by the duration of the study. What it is known is that their first citation occurs after the analysis ended.

In survival analysis, these observations are referred to as right censored observation. The publications that received their first citation within the period of analysis are called uncensored observations. Modelling time to event data requires that observations, both censored and uncensored have an observed time of interest, denoted as the follow-up time. For the uncensored observations, the follow-up time is the time to their first citations. For the censored observations, the follow-up time is the time period (in months) between their publication date and the end of analysis, that is December 2013, and it is referred to as the censored time.

For example, the censored time of a publication that appeared in January 2000 is 168 moths, whereas the censored time of a publication from June 2000 is 163 months. It needs to be distinguished between a publication with its time to first citation 163 months, for example it appeared in January 2000 and was firstly cited in December 2013, and a publication with its censored time 163 months. For this, we use an indicator Δ that is 1 if the publication has been cited and 0 if the publication remains uncited for the period of analysis.

The hazard rate

We are now interested in modelling the first citation rate on small units of time and its evolution in time. For this we will make use of the hazard rate, a functional of the time to first citation distribution. The hazard rate is referred to as the force of mortality in sociology, or the failure rate, in reliability. All these terms adhere to the pessimistic tone consistently used in survival analysis.

The hazard rate quantifies the rate at which first citations occur per unit of time relative to the proportion of publications that have not been yet cited. For a continuous random variable X, the hazard function is defined as

$$\lambda(t) = \lim_{\Delta t \searrow 0} \frac{P(t \leq X < t+\Delta t | X \geq t)}{\Delta t}.$$

In our case X denotes the time to first citation. We assume that the underlying time to first citation is continuous, while the observed data is discretized by measurement.

In order to compute the hazard rate at a given time point t, one needs to calculate the conditional probability in the numerator. In the present study, this is the probability of being firstly cited in the time interval [t,t+Δt), given that the publication has not been cited before time t. The conditioning ensures that at each time point t, only the publications that have not been cited up until time t are considered, therefore also the publications that are not cited throughout the entire period of analysis, i.e. the censored observations. Dividing this conditional probability by Δt, that is the width of the interval [t,t+Δt), we obtain the rate of the first citation occurrence per unit of time. By taking the limit Δt↘0 gives the instantaneous rate of occurrence of first citation. Note that, by definition, the hazard rate is not a (conditional) probability, or a density.

The hazard rate is a functional of the time to first citation distribution and can be derived for any parametric distribution and also estimated for a nonparametric distribution. The most straightforward example is the exponential distribution, for which the hazard rate is a constant function.

The hazard rate for the publications in the study is depicted in Figure 2 below.

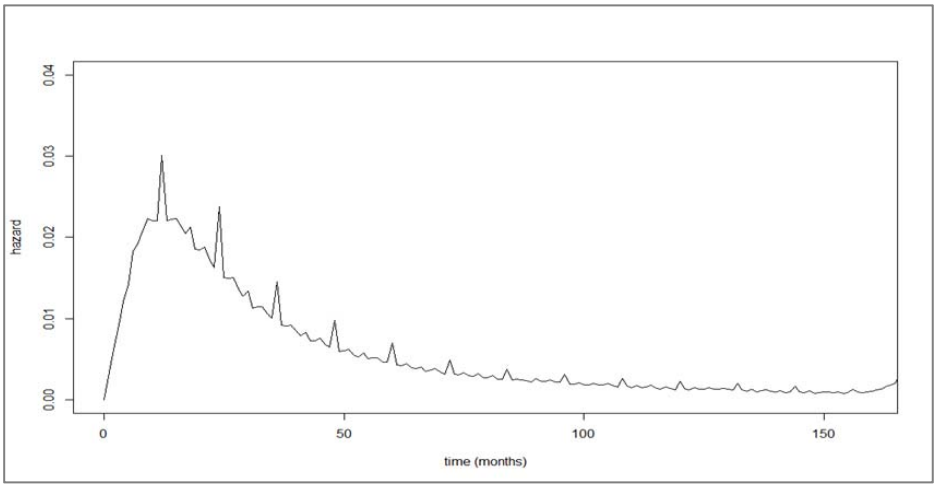


Figure 2. Hazard rate of publications in 2010.

First of all, we notice some spikes in the hazard function, which occur at the beginning and in the middle of each year in the citation window. This is due to the fact that certain journals publish once or twice a year. Moreover, when the publication date of certain

journal issues is unknown, the publication date is typically assigned to the beginning or middle of the year.

It seems that, per unit of time, publications receive their first citation at an increasing instantaneous rate up until a given time, that we refer to as the first citation peak, and despite the spikes, at decreasing instantaneous rates after the first citation peak. This shape suggests an unimodal hazard rate.

The first citation peak is for this dataset 24 months. In terms of conditional probabilities, the results can be interpreted as follows. Given that publications have not been cited before, on small unit intervals, they get cited for the first time with higher probability in the first 2 years after publications and with lower probability afterwards. The conditional probability decreases with time, but flattens after a while. That is, the decrease of the hazard is rather steep until 50 months and flattens afterwards. It can be inferred that first citation instantaneous rate is low and does not change significantly for documents that have not been cited for 4-5 years after publication.

Additional information – covariates

We are now interested in what can possibly influence the time to first citation and its hazard rate. This additional information is recorded as explanatory variables that are typically referred to as covariates in survival analysis, or as control variables in econometrics.

We consider the following covariates: document type, number of authors, collaboration type and field. By field we refer to the 250 subject categories to which journals are assigned in WoS. Surely, other covariates might be included, such as number of institutions or countries, number of pages, journal impact, etc.

Assume that covariates do not change over time, that they have a fixed value at the publication date. There can be however, covariates that change over time (time dependent covariates), such as journal impact, authors' visibility or performance.

The Cox model

The most famous model that incorporates the information on certain covariates in survival analysis is the Cox model (Cox, 1972). Regardless the fact that the model is more than 40 years old, it has been widely used and numerous versions, for particular issues with the data, have been proposed and investigated ever since.

The Cox model specifies the hazard rate at time t of a publication with a given covariate vector z as

$$\lambda(t|z) = \lambda_0(t)\exp(\beta' z),$$

where λ_0 is the underlying baseline hazard and β' is the transpose of the vector of underlying regression coefficients. Notice that if we take all covariates to be zero, we obtain the baseline hazard.

Within the Cox model, the hazard has two components. The first one, the baseline hazard, is the nonparametric part and it indicates how the hazard varies in time. The second term specifies parametrically, via an exponential function, the dependence on the covariates. It is then obvious why the Cox model is considered a semi-parametric model. Moreover, it is worth mentioning that the baseline hazard can be left unspecified when one want to estimate the regression coefficients and this flexibility has been particularly attractive for researchers.

Ever since the model was proposed, there was a great interest in estimating the regression coefficients β, that reflect how changes in the covariates produce a change in the hazard rate. The estimates were obtained via a partial likelihood method that avoided the bothersome issue of estimating the baseline hazard λ_0.
We have fitted the Cox model with the following covariates

- Document type
- Collaboration type
- Number of authors.

We will focus on estimating the (baseline) hazard and not on the regression coefficient estimation. We need to stress that conditioning on the covariates to be at a baseline value, i.e. z=0, is not the same thing as not accounting for covariates. This can be determined from the equation specifying the Cox model, but also from the figure below.

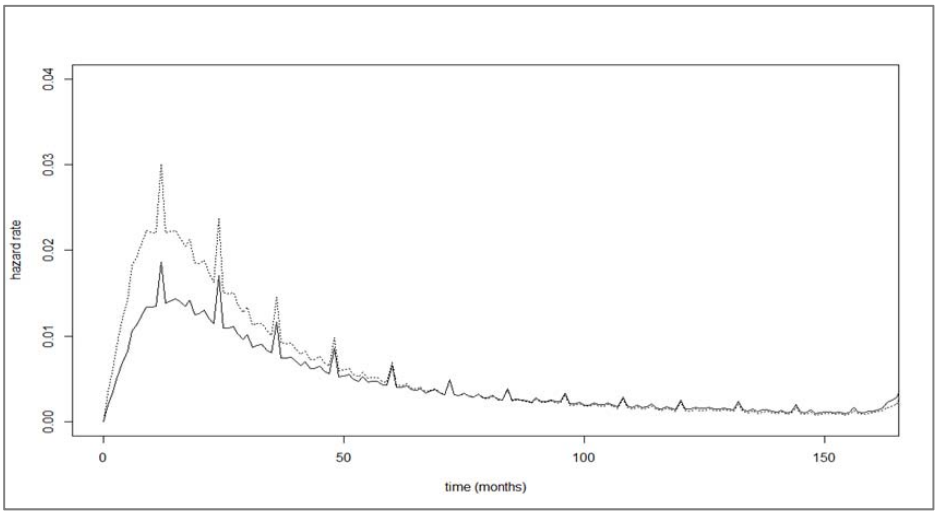


Figure 3. Hazard rate in the presence of no covariates (dotted) and baseline hazard (solid line).

Apparently, accounting for covariates shifts the hazard down in the first 60 months after the publication date and has no effect afterwards. The baseline hazard follows the same trend as the hazard rate in the presence of no covariates that is increasing until 24 months after the publication date and decreasing afterwards. Therefore, we can conclude that the covariates have a scale effect rather than a shape effect on the hazard. Furthermore, it seems that there is a proportional effect of the covariates on the baseline hazard, at least in the first 50 months. This represent a visualization of the goodness of fit of the Cox model and additionally, several tests suggest that the model fits the data well.
We want to investigate now whether certain characteristics of the publication, such as the collaborative status or the field have an impact on the instantaneous first citation rates.

Collaboration

It is commonly thought that publications that have resulted from an international collaboration are more visible to the academic community and hence receive more citations than national collaborative publications or publications that do not result from any inter institutional collaboration. It would be interesting to see if the collaboration type also influences how fast a publication receives its first citation.

As mentioned beforehand, we have fitted a Cox model with document type, collaboration type and number of authors as covariates. All the covariates have a (statistical) significant influence on the time to first citation.

To show the difference in the hazard rates among the different types of collaboration, we compute the hazard rate for publications with international, national and no collaborations. All the other covariates are set to their baseline level. Figure 4 depicts these differences.

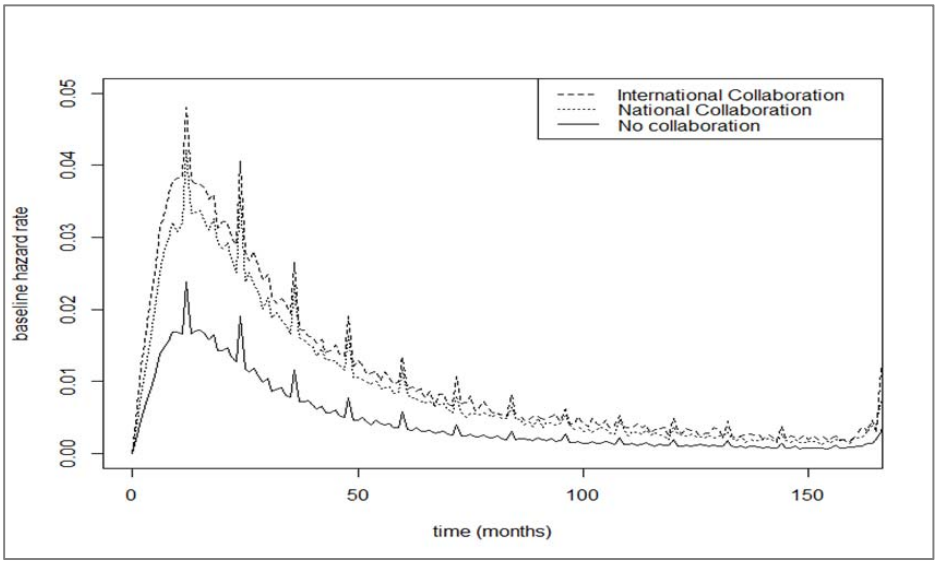


Figure 4. Baseline hazard rates in terms of collaboration type: international collaboration (dashed), national collaboration (dotted) and no collaboration (solid line).

It seems that there is a significant scale difference in the instantaneous first citation rate among publications that represent international and international collaborations and those that do not result from such collaborations. There are however small differences between baseline hazard of the international and national collaborative publications. Nonetheless, the publications that resulted from an international collaboration register higher instantaneous first citation rates than publications that represent national collaborations and these publications have, in turn, higher instantaneous first citation rates than publications whose authors are affiliated to a single institution. Similar to the overall (baseline) hazard rates, there are less and less differences in the hazard rates of different collaboration types 100 months after publication.

Contrary to the popular belief however, it seems that, apart from a scaling factor, publications receive their first citation at similar rates irrespective their collaboration type. The maximum hazard function is attained by publications of all collaboration types at the same time point, which is 24 months after the publication date. This is not different from the overall baseline hazard.

To condition further on specific values of the other covariates, we have considered the document type ‘Article’ and assume the publications has 3 authors, which is close to the overall average of the entire dataset, that is 3.31.

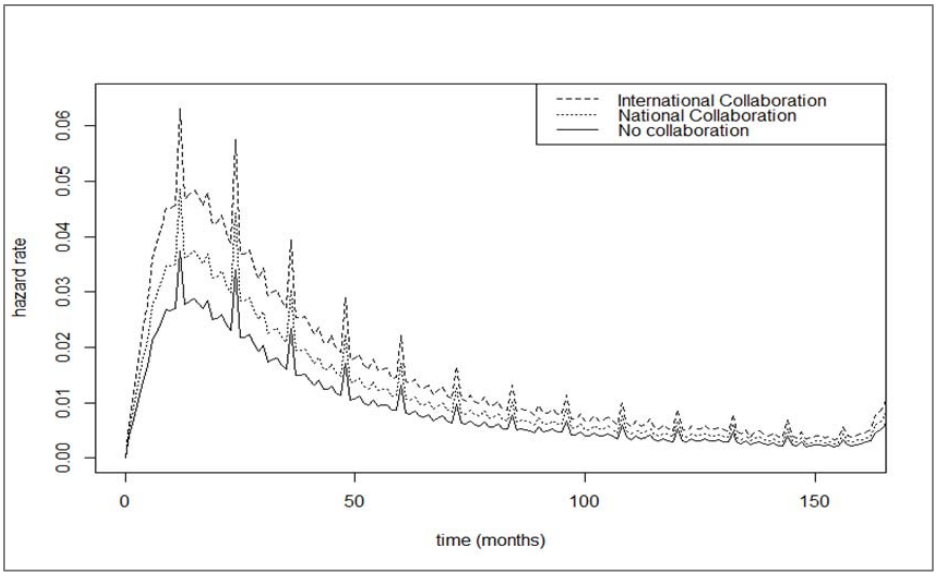


Figure 5. Hazard rates for articles with mean number of authors. International collaboration (dashed), national collaboration (dotted) and no collaboration (solid line).

Figure 5 depicts the hazard rates of articles that result from different collaborations and are written by three researches. We notice that the differences in the hazard rates have decreased. Despite similar behaviour over time, international collaborations still achieve the highest hazard rates over time, followed by national collaborations and articles produced by the same institution (no collaboration).

Field

We are also interested to see whether the field assigned to a certain publication affects the rate of being firstly cited. Nonetheless, more than half of the journals in WoS are assigned to at least two fields and some journals are assigned to six fields. This means that the field covariate cannot be uniquely defined for each publication. This difficulty cannot be overcome by using the WoS subject category assignment and hence the field cannot be included as a covariate in the Cox model. A solution is to adopt the publication-level classification system proposed by Waltman and van Eck (2012). Within this approach each publication is assigned to an unique cluster. Employing the publication-level classification system is deferred to future research.

In order to still assess the influence of the field on the time to first citation distribution, we have limited the data of all publication from 2000 to three fields: Biochemistry & Molecular Biology, Economics and Mathematics. We have now a number of 80,745 publications that have been published in 2000 and are assigned to the three fields.

We have fitted the Cox model with the following covariates

- Document type
- Collaboration type
- Number of authors
- Field

All four covariates have a (statistical) significant effect on the hazard rate. We are interested in the baseline hazard rates for the data grouped by the field. The differences

between the three baseline hazards can be observed in Figure 6. Once again, the other covariates have been set to zero.

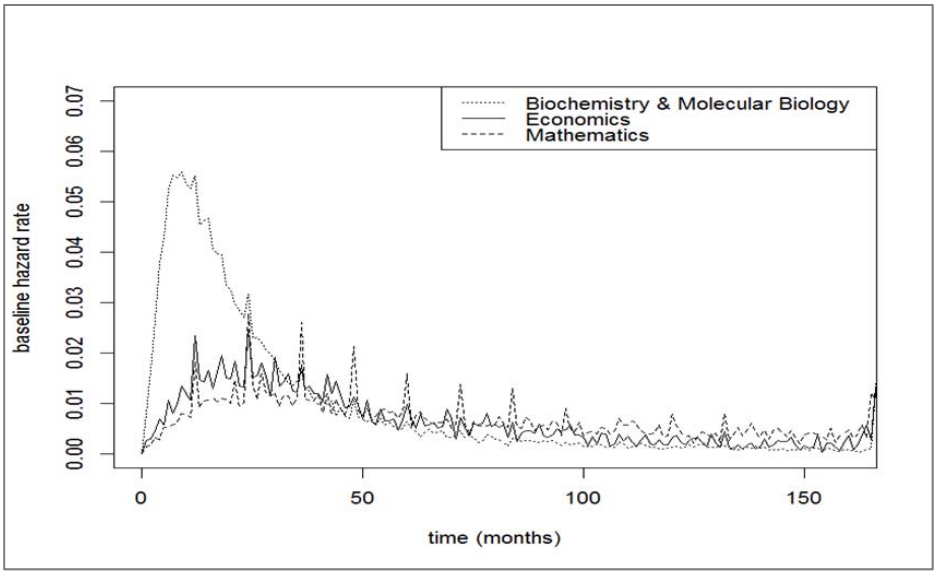


Figure 6. Baseline hazard rates in terms of field: Biochemistry and Molecular Biology (dotted), Mathematics (dashed) and Economics (solid line).

The three baseline hazard rates differ in both shape and scale. Firstly, it seems that the publications that appeared in 2000 in Biochemistry and Molecular Biology achieve their maximum first citation rate earlier than publications in Economics or Mathematics. The peak in Biochemistry and Molecular Biology is registered at 12 months, whereas the publications in Economics and Mathematics have a baseline hazard rate peak around 24 months.

We observe that there are large changes over time in the baseline hazard rate of publication in Biochemistry and Molecular Biology. Moreover, during the first part of the citation window, publications in Biochemistry and Molecular Biology have an instantaneous first citation rate three times as higher than the instantaneous first citation rates in Economics and Mathematics. The publications in Economics and Mathematics exhibit similar hazard rate behaviour.

It is noteworthy and interesting that after 60 months, the order of the three baseline hazard rates completely reverse, that is publications in Mathematics have higher baseline hazard rates than publications in Economics, that have higher baseline hazard rates than the publications in Biochemistry & Molecular Biology.

Discussion and conclusions

The first citation is probably the most important citation a publication receives. It can determine entirely the number or speed of further citations. Besides a simple citation count, it also changes the status of a publication, from being uncited to being cited. In some fields, being cited is even sufficient to become frequently cited.

The time to first citation also contributes to the number or speed of further citations. Apart from the famous sleeping beauties (van Raan, 2004), it is obvious that the more it takes for a publication to receive its first citation, the lower the probability of receiving further citations.

Time to first citation is the first step in modelling how publications accumulate citations in general over time. It is still unknown whether the time to first citation differs significantly from the time to second citation, etc.

We aimed to model the time to first citation and used a set of publications that appeared in 2000 and are included in the WoS database. Probably the most important aspect of our approach is that we employed nonparametric or semi-parametric methods of estimation. In other words, we let the data speak for itself. This ensures a greater flexibility and avoids the bothersome issue that a given model fits a particular data well, say publications that appear in a certain year and within a specific field, but fails to fit another particular data appropriately. While this is not a problem specific only to the first citation analysis, for an example on this matter in the first citation analysis, see Rousseau (1994). Another important drawback of the parametric approach is that certain employed parametric models cannot incorporate specific shapes of the first citation data. Van Dalen and Hekens (2005) for example make use of a Gompertz hazard model that cannot incorporate an unimodal hazard, as we obtained in the present study.

Apart from the nonparametric choice of estimation, we have also incorporated the uncited publications in the distribution of the time to first citation by using methods developed in survival analysis. We stress the fact that the information on uncited publications should be accounted for in modelling the time to first citation distribution, otherwise the results of the estimation can be seriously biased, especially given the high percentage of uncited publications.

We have investigated the time to first citation distribution through its hazard rate, the instantaneous rate of being firstly cited. We observe that the hazard rate increase over the first 24 months and decreases afterwards. This is somehow expected, that publications receive their first citations at higher rates until a maximum and afterwards at lower and lower rates. What is surprising is the relative short period of time over which the hazard rate is increasing. It means that the probability of a publications being cited for the first time is increasing over the first 24 months, and decrease afterwards.

Furthermore, it is of high interest to investigate whether certain characteristics of publications influence their time to first citation. We included the document type, number of authors, collaboration type and the field. We have found that all these explanatory variables (covariates) influence the time to first citation and investigated the differences between the hazard rates of publications grouped by collaboration type. The hazard rates of the three collaboration types differ in scale and not in shape and attain the maximum at the same time point. Hence, it seems that publications receive their first citations at an increasing rate up to the same time point, namely 24 months regardless their collaboration type.

A different dataset has been chosen to investigate the influence of the field on the time to first citation. It seems that, for the three selected fields, the hazard rate of the publications differ not only in scale but also in shape. The publications in Biochemistry and Molecular Biology register higher rates than publications from Economics and Mathematics, but also they have increasing first citation rates over a shorter period of time than the publications from the other two fields. The order of the three hazard rates reverse after 60 months.

As mentioned in the previous section, the problem of the overlapping fields in WoS needs to be addressed in future research and this can be overcome by considering the

publication-level classification system proposed by Waltman and Van Eck (2012). Numerous investigations are further required and desired. For example it would be very interesting to investigate whether the time to first citation distribution, and in particular the hazard rate including self citations differs from the time to first citation excluding self citations. Other covariates can be included in the analysis, such as the impact of the journal, the performance or visibility of authors, etc. Of course, it is very interesting to see whether the shape of the hazard rate changes over the time of publication, not only through the citation window. The author expects that the hazard would have the same unimodal shape, but the maximum point would be attained at different time points that is the first citation peak would be time dependent.

In terms of estimation, it is highly desirable to account for the monotonicity of the (baseline) hazard that is to provide estimates of the baseline hazard rate under the assumption of monotonicity. This is in line with the research of Lopuhaä and Nane (2013), but needs some refinement to incorporate the estimation of a unimodal baseline hazard.

References

Bornmann, L. & Daniel, H.-D. (2010). Citation speed as a measure to predict the attention an article receives: An investigation of the validity of editorial decisions at *Angewandte Chemie International Edition*. *Journal of Informetrics, 4*, 83-88.

Burrel, Q.L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics,* 52, 3-12.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*. *Series B (Methodological)*, *45*, 187-220.

van Dalen, H.P. & Henkens, K. (2005). Signals in science – On the importance of signaling in gaining attention in science. *Scientometrics*. *64*, 209-233.

Egghe, L. (2000). A heuristic study of the first-citation distribution. *Scientometrics,* 48, 343-359.

Egghe, L & Rao, I.K.R. (2001). Theory of first-citation distributions and applications. *Mathematical and Computer Modelling*, *34*, 81-90.

Glänzel, W. (1992). On some stopping times of citation processes. From theory to indicators. *Information Processing & Management, 28*, 53-60.

Glänzel, W., Rousseau, R. & Zhang, L. (2012). A visual representation of relative first-citation times. *Journal of the American Society for Information Science and Technology*, *63*, 1420-1425.

Glänzel, W. & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature, *Journal of Information Science*, *21*, 37-53.

Lopuhaä, H.P. & Nane, G.F. (2013). Shape constrained nonparametric estimators of the baseline distribution in Cox proportional hazards model. *Scandinavian Journal of Statistics, 40*, 619-646.

van Raan, A.F.J. (2004). Sleeping beauties in science (short communication). *Scientometrics*

Rousseau, R. (1994). Double exponential models for first-citation processes. *Scientometrics, 30*, 213-227.

Schubert, A. & Glänzel, W. (1986), Mean response time – a new indicator of journal citation speed with application to physics journals. *Czechoslovak Journal of Physics (B), 36*, 121-125.

Waltman, L. & van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system in science. *Journal of the American Society for Information Science and Technology*, *63*, 2378-2392.

Author Relationship Mining based on Tripartite Citation Analysis

Feifei Wang[1], Junwan Liu[2], Siluo Yang[3]

[1] *feifeiwang@bjut.edu.cn,* [2] *liujunwan@bjut.edu.cn*
School of Economics and Management, Beijing University of Technology, 100024 Beijing (China)

[3] *58605025@qq.com*
School of Information Management, Wuhan University, 430072 Wuhan (China)

Abstract
This study scrutinizes potential author relationships according to the findings based on the tripartite citation analysis. It focuses on Author co-citation analysis (ACA), author bibliographic-coupling analysis (ABCA) and author direct citation analysis (ADCA). By algorithm design and empirical analysis, the deduction from results of ACA, ABCA and ADCA to potential author relationships mining could be probable, and the empirical process would be practicable.

Conference Topic
Citation and co-citation analysis

Introduction
Citation analysis is a mature quantitative research method in Bibliometrics and Scientometrics. It is widely used in scientific evaluation, scholarly communications, academic behavior analysis, and information retrieval. Author citation analysis mainly includes three types: author co-citation, author coupling, and author direct citation.

Author co-citation analysis (ACA) is the most widely used method for the empirical analysis of disciplinary paradigm, and is frequently studied and improved upon. Many ACA studies have been conducted since Small (1973) introduced document co-citation analysis and White and Griffith (1981) introduced ACA. Bibliographic coupling was proposed as early as 1963 (M. M. Kessler, 1963). However, author bibliographic-coupling analysis (ABCA), i.e. author-coupling relationships, did not get much attention until it is formally put forward and empirically studied by Zhao (Zhao & Strotmann, 2008).

Direct citation is sometimes also called inter-citation or cross citation (Zhang et al., 2009). Compared with co-citation and bibliographic coupling, direct citation is a direct citing relationship without a third party paper. Although researchers are aware of direct citation analysis and employed from time to time (Shibata et al., 2008), it was ignored because of the unavailability of data, difficulty of implementation, and long time windows to obtain a sufficient linking signal for clustering. However, scholars are gradually paying attention to this topic (Boyack & Klavans, 2010). A number of studies have focused on journal direct citation or comparative analysis of methods. The author direct citation analysis was more clearly explored by Wang et al. (2012). Wang used this method to reveal the knowledge communication and disciplinary structure in Scientometrics. This process is named "author direct citation analysis" (ADCA) (Yang & Wang, 2015).

All of these three kinds of citation analysis methods can reveal separately the author relationship in a field. Then, how about the similarities or diversity among the tripartite citation relationships at author level? And, how could the tripartite relationships be synthetically presented to the readers or the result users? We have tried to answer these two questions in previous studies (Wang, 2014), even though the effort is still limited. Persson (2010) and Gómez-Núñez et al. (2011, 2014, 2015) tried to combine these citation measures

in a normalized way to weight existing direct citation relationships between articles or journals.

The following question is worthy of investigation as well: Could we discover potential author relationships according to the findings based on the tripartite citation analysis? To give an example: in a field, author A's paper and author B's paper both are cited by the same paper C, then A and B have co-citation relationship, which can be marked as (A, co-citation, B). Author C and author D, when citing the same paper in their respective articles, have bibliographic-coupling relationship, marked as (C, bibliographic-coupling, D). In addition, if C and A cite each other, then C and A have direct-citation or cross-citation relationship, marked as (C, directly citing, A) or (A, directly citing, C) or (A, cross citation, C). According to these primary relationships, could we deduce an integrated relationship between A and D, or B and C, even B and D? And, what will be the association strength in these potential relationships? These are the key problems that we answer in this study.

Data and methodology

Basic Data

Since the journal *Scientometrics* is one of the most representative communication channels in the field of Scientometrics, it reflects the characteristic trends and patterns of the past decades in scientometric research (Schubert A 2002). This study is based on bibliographic data based on all types of documents published in *Scientometrics* from 1978 to 2011, retrieved from the Web of Science. Author names including the cited authors were normalized because some authors may report their names differently in different papers. We identified each author by his or her surname and first initial only; the same applies to cited authors.

Methodologies

In this study, bibliometrics method is applied to identify the core authors (94 first authors who have published 5 or more papers and simultaneously have a cited frequency over 10) in Scientometrics filed. Author co-citation analysis (ACA), author bibliographic-coupling analysis (ABCA) and author direct citation analysis (ADCA) are respectively used to discover author relationships with co-citation, bibliographic-coupling and direct-citation. Co-occurrence analysis and deductive reasoning methods are used to mine potential author relationships on the findings of the tripartite citation analysis. VBA program processes all kinds of citation analysis data. The final results of author relationship are visualized with Pajek.

Results and discussion

According to the tripartite citation analysis, we obtain three original relation matrixes and their corresponding normalized matrixes (Fig. 1). The normalization method is based on Salton's Cosine similarity measures, which returns similarity values ranging between 0 and 1. In order to describe the directivity of citing behaviour and achieve vectorial deducing, the direct citation matrix is unsymmetrical.

Core author co-citaion matrix

	Garfield E	Glanzel W	Braun T	Egghe L
Garfield E	1	0.7535	0.8359	0.5426
Glanzel W	0.7535	1	0.8916	0.7579
Braun T	0.8359	0.8916	1	0.5736
Egghe L	0.5426	0.7579	0.5736	1

Core author bibliographic-coupling matrix

	Garfield E	Glanzel W	Braun T	Egghe L
Garfield E	1	0.4249	0.3612	0.2881
Glanzel W	0.4249	1	0.9171	0.4069
Braun T	0.3612	0.9171	1	0.26
Egghe L	0.2881	0.4069	0.26	1

Core author direct citation matrix

	Garfield E	Glanzel W	Braun T	Egghe L
Garfield E	1	0.0022	0.0312	0.0027
Glanzel W	0.2844	1	0.414	0.1365
Braun T	0.2511	0.173	1	0.0073
Egghe L	0.0974	0.221	0.1058	1

Figure 1. Normalized matrixes of tripartite citation analysis.

The following five steps could help us realize author relationship mining based on tripartite citation analysis, such as "A→C, B→D, B→C". These steps can also be seen as an algorithm in relation mining.

First step: Obtaining the fundamental citation relationship with strength(>0) among core authors from original matrixes

Tripartite adjacency matrixes are transformed into corresponding adjacency lists. ACA list $\{L_{1i},Q_{1i}\}$ versus matrix $\{O_{1i}, P_{1j}\}$, and relational degree X_i (i stands for the ID of author pair) in list can replace X_{ij} (i/j stand for different authors in the matrix) . ABCA list $\{L_{2i},Q_{2i}\}$ versus matrix $\{O_{2i}, P_{2j}\}$, and relational degree Y_i versus Y_{ij}. ADCA list $\{L_{3i},Q_{3i}\}$ and $\{L_{3j},Q_{3j}\}$ versus matrix $\{O_{3i}, P_{3j}\}$, and relational degree Z_i and Z_j versus Z_{ij} (the order between i and j denotes the citing direction). We used the Adjacency list in calculation process.

Second step: Filtering no-explicit-relationship author pairs

The no-relationship author pairs (X_i=0, Y_i=0, Z_i=0, and no cooperation), are filtered as $\{O_{4i},P_{4j}\}$ in the Adjacency matrix, and $\{L_{4i},Q_{4i}\}$ in the Adjacency list, which forms the basic object in subsequent analysis.

Third step: Mining the relationship of A →C from$\{L_{1i},Q_{1i}\}\{L_{3i},Q_{3i}\}\{L_{4i},Q_{4i}\}$

Remark the $\{L_{4i},Q_{4i}\}$ as $\{A_k,C_k\}$ (k stands for the number of author pairs), the goal is finding the Dk with the relations $\{A_k→D_k, C_k\text{-}D_k\}$. We looked for the synchronous relations with strengh between A_k and D_k, C_k and D_k, from $\{L_{1i},Q_{1i}\}\{L_{3i},Q_{3i}\}$, and matched the author pairs in $\{A_k,C_k\}$. The pseudo code is as follows:

If one author in the pair of $\{A_k,C_k\}$= one author in a pair of $\{L_{1i},Q_{1i}\}$, and another one in the pair of $\{A_k,C_k\}$= one author in a pair of $\{L_{3i},Q_{3i}\}$, and another one in the pair of $\{L_{1i},Q_{1i}\}$= another one in the pair of $\{L_{3i},Q_{3i}\}$

Then mark the "one author in the pair of $\{A_k,C_k\}$" (so as the "one author in a pair of $\{L_{1i},Q_{1i}\}$") as Cα, the "one author in a pair of $\{L_{3i},Q_{3i}\}$" (so as the "another one in the pair of $\{A_k,C_k\}$") as Aα, the "another one in the pair of $\{L_{1i},Q_{1i}\}$" (so as the "another one in the pair of $\{L_{3i},Q_{3i}\}$") as Dα

End with the relation between A_α and C_α according to D_α, and their relation strength equaling to the product of X_α and Y_α. If the order of author pair in $\{L_{4\alpha},Q_{4\alpha}\}$(i.e., $\{A_k,C_k\}$) is in reverse of the order of author pair in $\{L_{3\alpha},Q_{3\alpha}\}$(i.e., $\{A_k,D_k\}$), then the relation strength between Aα and Cα will be the negative value.

Finally, choose the top value (Take the absolute value of the negative value) as the final relation strength of A_α and C_α.

Fourth step: Mining the relationship of B →D from$\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}\{L_{4i},Q_{4i}\}$

Remark the $\{L_{4i},Q_{4i}\}$ as $\{B_k,D_k\}$ (k stands for the number of author pairs), the goal is to find the A_k with the relations $\{A_k→D_k, A_k\text{-}B_k\}$. We looked for the synchronous relations with strengh between A_k and D_k, A_k and B_k, from $\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}$, and matched the author pairs in $\{A_k,C_k\}$.This process is similar with the process of A→C, so the pseudo code is omitted.

Fifth step: Mining the relationship of B →C from$\{L_{1i},Q_{1i}\}\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}\{L_{4i},Q_{4i}\}$

Remark the rest (no relationship like A→C and B→D) of $\{L_{4i},Q_{4i}\}$ as $\{B_k, C_k\}$ (k stands for the number of author pairs), the goal is to find the A_k and D_k with the relations $\{A_k→ D_k, A_k\text{-}B_k, C_k\text{-}D_k\}$. We looked for the synchronous relations with strengh between A_k and D_k, A_k and B_k, Ck and D_k, from $\{L_{1i},Q_{1i}\}\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}$, and matched the author pairs in $\{B_k,C_k\}$. The pseudo code as follows:

If one author in the pair of $\{B_k,C_k\}$= one author in a pair of $\{L_{2i},Q_{2i}\}$, and another one in the pair of $\{B_k,C_k\}$= one author in a pair of $\{L_{1i},Q_{1i}\}$, and another one in the pair of $\{L_{2i},Q_{2i}\}$=one author in the pair of $\{L_{3i},Q_{3i}\}$, and another one in the pair of $\{L_{1i},Q_{1i}\}$= another one in the pair of $\{L_{3i},Q_{3i}\}$

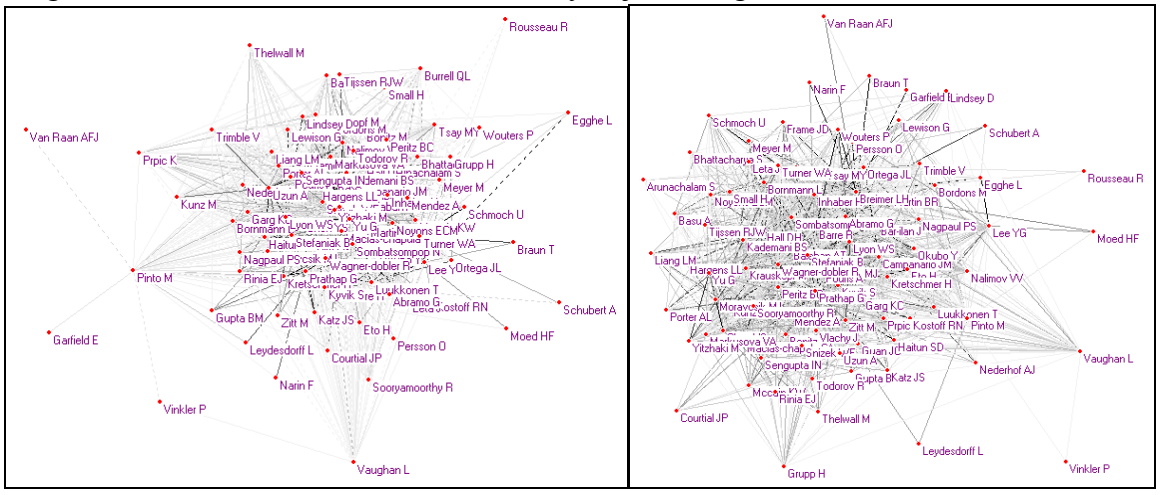
Then mark the "one author in the pair of $\{B_k,C_k\}$" (so as the "one author in a pair of $\{L_{2i},Q_{2i}\}$") as B_χ, "another one in the pair of $\{B_k,C_k\}$" (so as "the one author in a pair of $\{L_{1i},Q_{1i}\}$") as C_χ, one author in the pair of $\{L_{3i},Q_{3i}\}$(so as the "another one in the pair of $\{L_{2i},Q_{2i}\}$") as A_χ, another one in the pair of $\{L_{1i},Q_{1i}\}$(so as the "another one in the pair of $\{L_{3i},Q_{3i}\}$) as D_χ

End with the relation between B_χ and C_χ according to A_χ and D_χ, and their relation strength equaling to the product of X_χ and Y_χ and Z_χ. If the order of author pair in $\{L_{4\chi},Q_{4\chi}\}$(i.e., $\{B_k,C_k\}$) is in reverse of the order of author pair in $\{L_{3\chi},Q_{3\chi}\}$(i.e., $\{A_k,D_k\}$), then the relation strength between B_χ and C_χ will be the negative value.

Finally, choose the top value (take the absolute value of the negative value) as the final relation strength of B_χ and C_χ.

So far, all relationship among author pairs in $\{L_{4i},Q_{4i}\}$ have been built.

According to the above algorithm, potential relationships among not-directly-related core author set could be discovered by VBA programme and Access databases. The final results among A→C, B→D and B→C are visulized by Pajek as Figure 2 and 3.



(a) (b)

Figure 2. (a) Author relationship network of A→C. (b)Author relationship network of B→D.

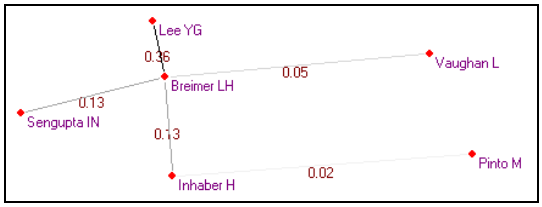


Figure 3. Author relationship network of B→C.

In Figure 3, the labels in the lines denote the value of the relationship similarity for authors in pairs. According to the results, there are different levels of potential relationship between Breimer LH and other authors, such as Inhaber H、Lee YG、Sengupta IN、Vaughan L.

Conclusions

Based on the algorithm design and empirical analysis, the deduction from results of ACA, ABCA and ADCA to potential author relationships mining could be probable, and the

empirical process would be practicable. The findings in Scientometrics field can help scholars discover more research fellows, which can promote scientific research cooperation and broader knowledge communication.

References

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.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers, *Journal of the American Society for Information Science and Technology*, *14*(1), 10-25.

Gómez-Núñez, A. J., Batagelj, V., Vargas-Quesada, B., Moya-Anegón, F., & Chinchilla-Rodríguez, Z. (2014). Optimizing SCImago Journal & Country Rank classification by community detection. *Journal of Informetrics*, *8*(2), 369-383.

Gómez-Núñez, A. J., Vargas-Quesada, B., de Moya-Anegón, F. & Glänzel, W.(2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, *89*(3), 741-758.

Gómez-Núñez, A. J., Vargas-Quesada, B. & Moya-Anegón, F. (2015). Updating the SCImago journal and country rank classification: A new approach using Ward's clustering and alternative combination of citation measures. *Journal of the Association for Information Science and Technology*, published online. http://dx.doi.org/10.1002/asi.23370.

Persson,O.(2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, 4(3), 415-422.

Schubert, A. (2002). The web of Scientometrics: A statistical overview of the first 50 volumes of the journal. Scientometrics, *53*(1), 3-20.

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), 758-775.

Small, H. (1973). Cocitation 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.

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.

Wang, F. (2014). Influence Analysis of Core Authors in Scientometrics from an Integrated Perspective of Publication and Citation. *Science of Science and Management of S.&T.(China)*, *35*(12): 45-55.

Wang, F., Qiu, J. & Yu, H. (2012). Research on the cross-citation relationship of core authors in scientometrics. *Scientometrics*, *91*(3), 1011-1033.

White, H.D. & Griffith, B. (1981). Author cocitation: A literature mesaure of intellectual structures. *Journal of the American Society for Information Science*, *32*(3), 163-171.

Yang, S. & Wang, F. (2015). Visualizing Information Science: Author Direct Citation Analysis in China and around the World. *Journal of Informetrics*, *9*(1), 208-225.

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.

Charles Dotter and the Birth of Interventional Radiology: A "Sleeping-Beauty" with a Restless Sleep

Philippe Gorry[1] and Pascal Ragouet[2]

[1] *philippe.gorry@u-bordeaux.fr*
GREThA UMR CNRS 5113, University of Bordeaux, Av. Leon Duguit, 33608, Pessac (France)

[2] *pascal.ragouet@u-bordeaux.fr*
Centre E. Durkheim CNRS UMR 5116, University of Bordeaux, 3ter Pl. de la Victoire, 33076, Bordeaux (France)

Abstract
Charles Dotter is described as the father of interventional radiology, a medical specialty born at the cross-border of radiology and cardiology. Dotter's landmark paper published in 1964 was poorly cited until 1979 and can be considered from a scientometric point of view as a sleeping beauty. Sleeping-beauties are article that suffer of a delayed recognition. This paper, will explore the bibliometric characteristics of this case study and the accuracy of Van Raan's criteria to define "sleeping beauty" in science will be discussed. "The prince" is identified through citation network analysis, and the sleeping period has been documented as a restless sleep period with science and social controversy that could be documented in publications databases by differentiating bibliographic references. Therefore, a category of "sleeping beauty" –like paper should be introduced. By the end, these observations should open new avenues in identifying "sleeping beauties".

Conference Topic
Citation and co-citation analysis

Introduction

Charles Dotter, father of interventional radiology

Charles Theodore Dotter (1920–1985) was a pioneering US vascular radiologist, credited with developing interventional radiology (IR): he invented the angioplasty and the catheter-delivered stent. On January 16, 1964, he percutaneously dilated a tight, localized stenosis of the superficial femoral artery in an 82-year-old woman with painful leg ischemia and gangrene who refused leg amputation. Percutenous transluminal angioplasty (PTA) was born, and Dotter with his trainee Dr. Melvin P. Judkins, described their technique in a landmark paper published in the medical journal "Circulation" (Dotter, 1964).

Today, Charles Dotter is described as the father of interventional radiology (IR), a sub-specialty of radiology using minimally invasive image-guided procedure to diagnose, as well as to treat diseases in every organ. The Oregon Health Sciences University (OHSU), where he spent his entire medical career, boasts the Dotter Interventional Institute. Furthermore, the Society of Interventional Radiology named a Dr. C.T. Dotter lecture to honor annually extraordinary contributions to the IR field (Rösch, 2003).

However, at first, the relationship between surgeons and radiologists was adversarial because the Dotter technique was a paradigmatic revolution, inviting radiologists to transgress medical specialty boundaries. It can be summed up by Dotter's formula at that time: "The angiographic catheter can be more than a tool for passive means for diagnostic observations; used with imagination, it can become an important surgical instrument". (Payne, 2001).

Therefore, as we found out, Dotter's landmark paper was poorly cited until 1979 and can be considered from a scientometric point of view as a sleeping beauty paper.

Sleeping beauty in scientific literature

In Scientometrics, the phenomenon of delayed recognition has been well described since the pioneering observations of Garfield, and referred to as premature discoveries, resisted discoveries, delayed recognition or sleeping beauties (Burrell, 2005; Braun, 2010). Van Raan (2004) defined "sleeping beauties" as articles that go unnoticed ("sleeps") for a long period of time and then, suddenly, receives a lot of citations by a "prince" (another article). Three variables were defined for such papers: depth of sleep, length of the sleep and awaking intensity. Some publication had heaping before sleeping, and are described as "all-element-sleeping beauties" (Li, 2012).

Objectives

In the present work, we explore the bibliometric characteristics of this case study, question the sleeping-beauty definition, explore the diffusion of Dotter concept during the sleeping period, and document the awaking phase and identify "the prince" through citation network analysis.

Method

A literature search on Dotter C.T. scientific production was conducted both in PubMed and Scopus databases. Citations of Dotter work were extracted from the Web of Science database until 12/31/2013. Then, a descriptive statistics analysis was led on the corpus (219 publications; 7866 citations). Scientific collaborations of C.T. Dotter was explored with Intellixir® to draw co-publications graph. Citations network pattern during time of the landmark paper was drawn using CitNetExplorer software tool (Van Eck, 2014). Complementary queries were run using Dotter or PTA as a keyword in different search fields for different types of documents.

Result

The scientific production of Charles Dotter

Dotter published his first paper in 1948 in a top medical journal, the New England Journal of Medicine (Jan 13; 239(2):51-4). During his 33 years at OHSU, he issued 219 publications; a quarter of his scientific production was disclose in high quality journals, and split between 2 main medical disciplines: radiology and cardiology (Table 1).

Table 1. Journal distribution of C.T. Dotter scientific production.

Source title	*Publications number*	*Impact factor*
Radiology	46	5,561
Am. J. Roentgenol. Radium Ther. Nucl. Med.	27	na
Circulation	19	12,755
New England J Medicine	8	52,589
Am. J. Roetgenol.	6	2,47

Dotter had many relations in the academic community: all along his career he co-published with 140 different authors, mainly with J. Rosch, F. Keller & J. Melvin (340, 215 & 68 co-publications respectively; Fig.1 and Table 2).

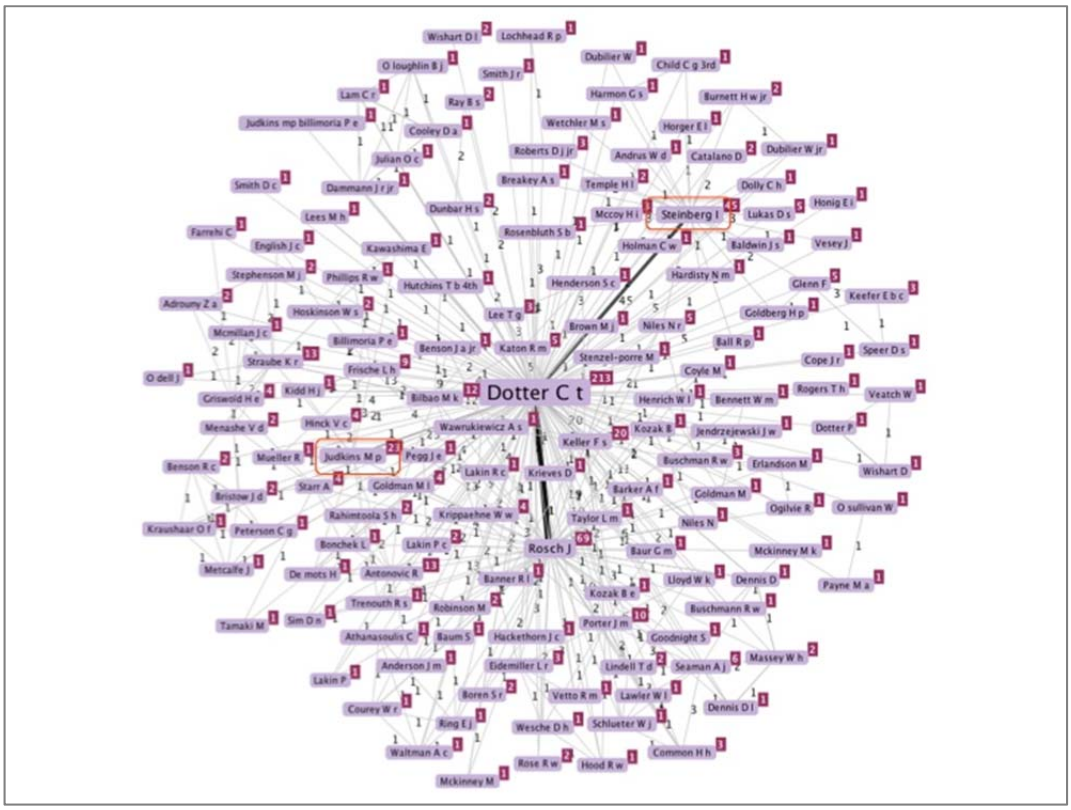


Figure 1. Network of C.T. Dotter co-publications.

Table 2. C.T. Dotter main scientific collaborators.

Author	Lab. / Dpt.	Institution	Publi.
Rösch, Johannes	Center of Cardiac Surgery	Friedrich Alexander University (DE)	340
Keller, Frederick S.	Dotter Interventional Inst.	Orgeon Health & Sciences Medical Center (USA)	215
Steinberg, Israel	Dpt. of Surgery, Medicine & Radiology	New Loma Linda Univ. (USA)	174
Judkins, Melvin P.	Coordinating Center for Collaborative studies in Coronary Artery Surgery	New York Hospital – Cornell Univ. (USA)	68
Bilbao, Marcia K.	Dpt. of Radiology	University of Oregon Mecial School (USA)	22

He published his last paper in 1981, four years before his death. By the end of his career, his scientific work totalized more than 4500 citations and reached 7866 citations at the end of 2013 (Fig. 2).

Dotter successfully diffused his results and obtained recognition from his academic community with an average of 52-251 citations every year.

It is interesting to point out that before his landmark paper was published in 1964, he was already an active researcher with 100 publications, well recognized by his academic community with 1068 citations at that time.

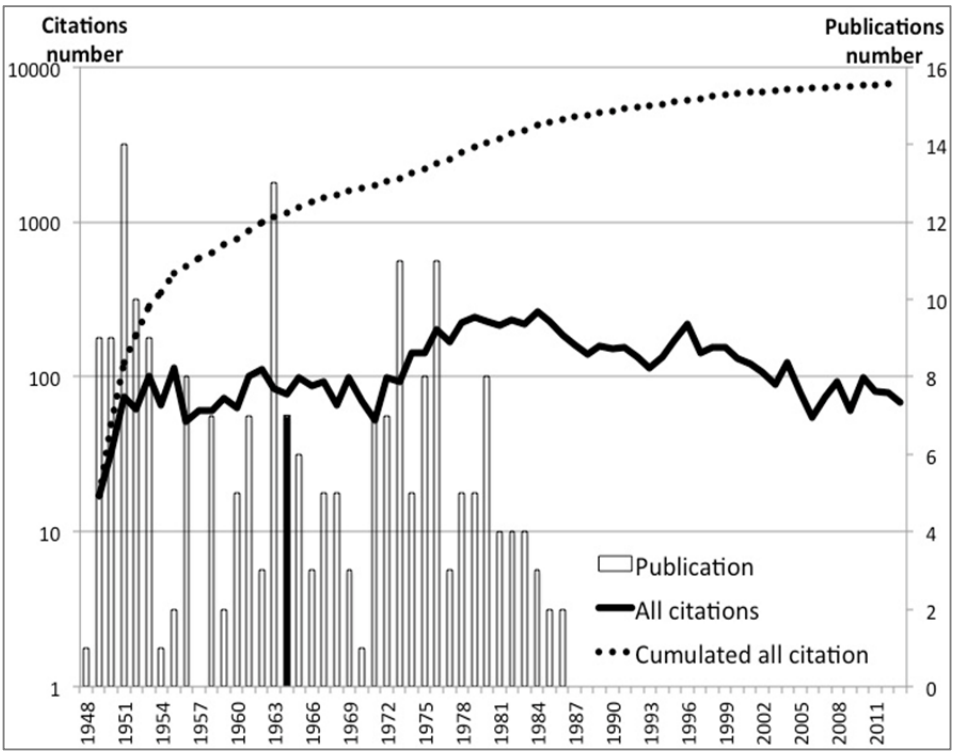


Figure 2. Dotter's publications and citations.

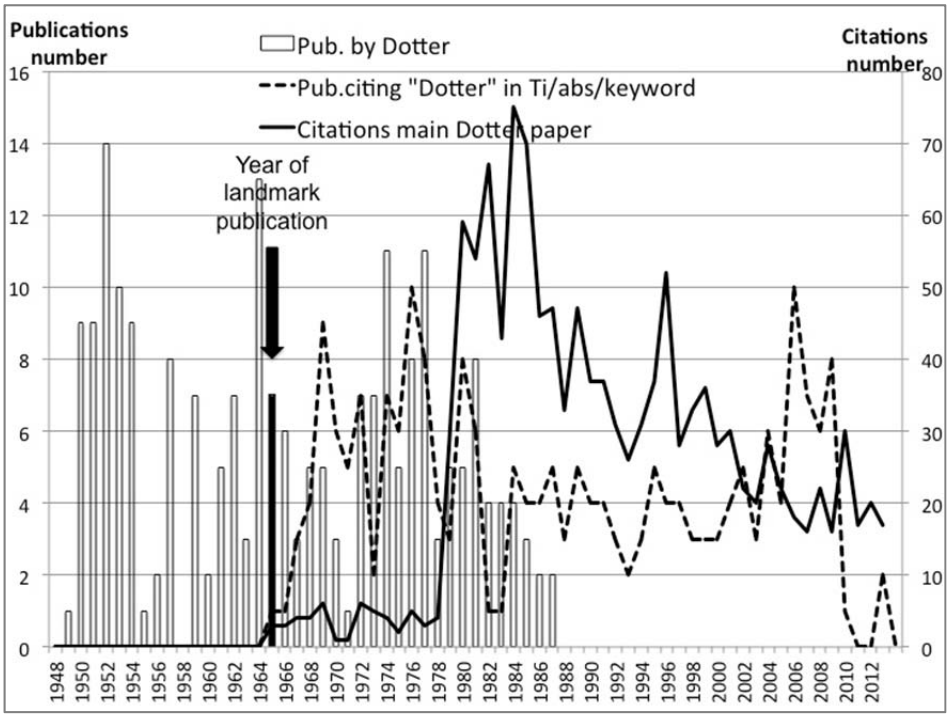


Figure 3. Dotter's main paper citations and Dotter's name apparition in the literature.

Dotter's landmark paper: a sleeping-beauty?

Dotter's landmark paper published in 1964 (Figure 2; black box) was cited with an average of 19.31 citations per year, totalizing 1275 citations today. However, during the first 14 years, his paper was cited only 51 times (Figure 3; full line) before suddenly gaining 29 citations in 1979 and more than 50 citations per year in the latter period.

Therefore, Dotter's main paper has the characteristics of a "sleeping beauty" despite the fact that it does not exactly fit Van Raan's definition (depth of sleep: 3.64 citations/year length of sleep period: 14 years; awake intensity: 52.25 citations/year).

During the delayed recognition period, Dotter was frequently named (n=76) within medical literature (Figure 3: dotted line), as well as his technique, percutaneous transluminal angioplasty (data not show) attesting that the "sleeping period" was traversed by a medical controversy.

The corresponding "Prince" was identified by visualizing the pattern of citations (Fig.4). A German cardiologist, A. Gruntzig, inventor of the coronary balloon angioplasty, was the first to referred to Dotter's previous work. He first did so in a paper published in German in the journal Deutsche Medizinische Wochenschrift in 1974, which had however only very little echo at that time until it was published in English in a well established journal in radiology (American J. of Roentgenology, 132:547-552, April 1979).

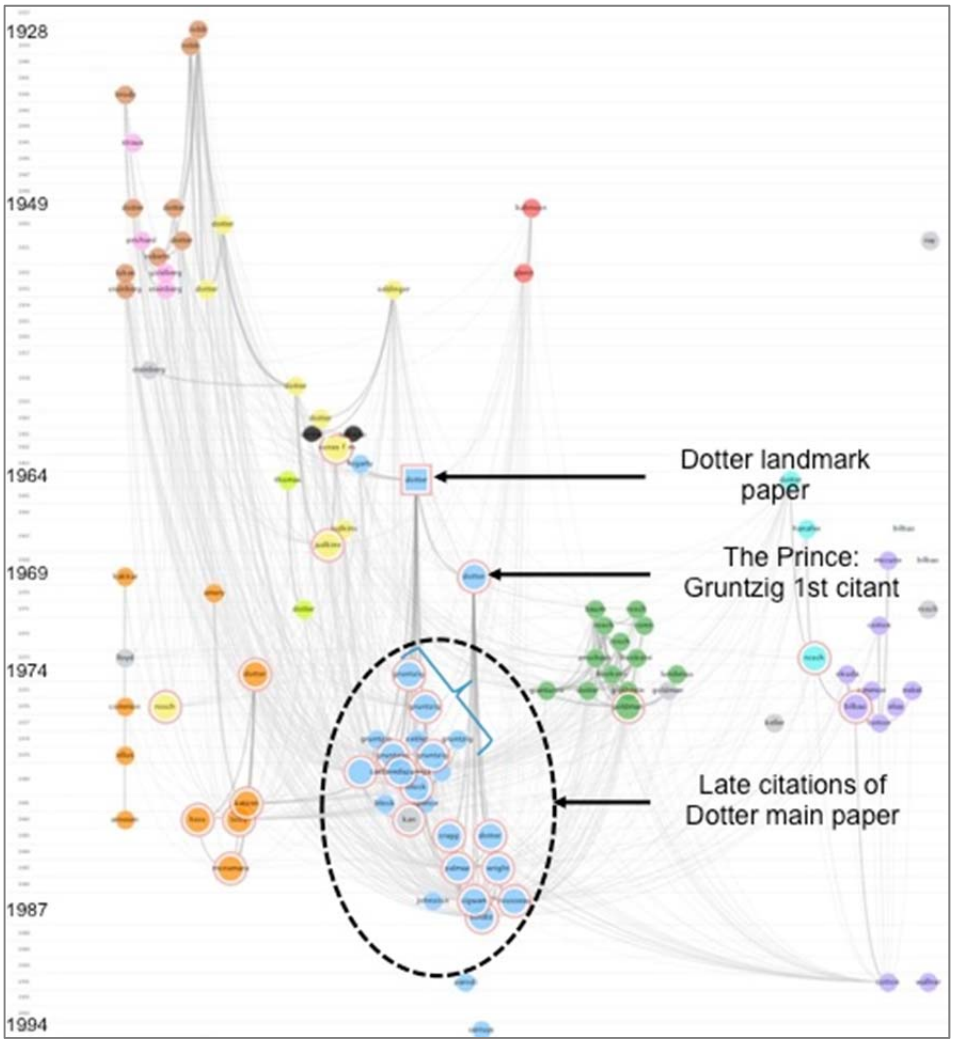


Figure 4. Citation network of CT Dotter paper and its direct and indirect successors.

Later on, Gruntzig's paper, citing Dotter pioneering work, was quickly cited in the medical literature (n=23, year +1) and its peak of citations coincided with the awaking of Dotter landmark paper citations (Figure 5).

Discussions

Dotter landmark paper has the characteristics of a sleeping-beauty but does not fit Van Raan's criteria. Therefore, this case study will discuss the accuracy of Van Raan's criteria to define

"sleeping beauty" in science, and introduce the category of "sleeping beauty" – like as a paper. Beside it is necessary to pinpoint that the sleeping period might indeed be a restless sleep period traversed by scientific controversy that could be traced back in publications databases by differentiating bibliographic references from citations in the text, or by analyzing the nature of the documents, especially article versus editorial, letter or review. These observations should open new avenues in identifying "sleeping beauties" in the literature, and nurture science resistance or controversy study in sociology of science.

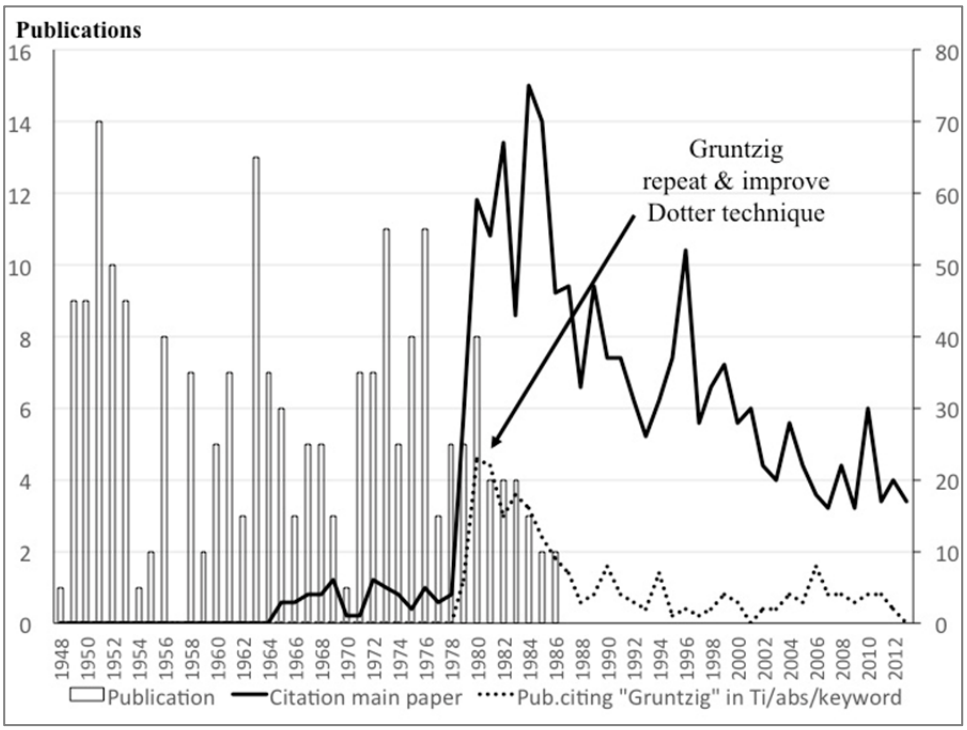


Figure 5. Citations curves of Dotter's paper & Gruntzing refering paper.

Acknowledgments

This work is supported by a grant from the French National Cancer Institute (INCA#6165).

References

Braun, T., Glänzel, W., & Schubert, A. (2010). On Sleeping Beauties, Princes and other tales of citation distribution. *Research Evaluation*, *19*(3), 195-202.

Burrell, Q.L. (2005). Are "sleeping beauties" to be expected?" *Scientometrics* 65, 381-389.

Dotter, C. & Judkins, M. (1964). Transluminal treatment of arteriosclerotic obstruction. Description of a new technic & a preliminary report of its applications. *Circulation 30*, 654-70.

Gruntzig, A. & Hopff. H. (1974), Perkutane Rekanalisation chronischer arterieller Verschlüsse mit einem neuen Dilatationskatheter. *Deutsche Medizinische Wochenschrift, 99*, 2502-2505.

Li, J. & Ye, F.Y. (2012). The phenomenon of all-elements-sleeping-beauties in scientific literature. *Scientometrics,* 92, 795-799.

Payne, M. (2001). C T Dotter : the father of Intervention. *Texas Heart Institute J. 28*, 28-38.

Rösch, J., Keller, F. S., & Kaufman, J. A. (2003). The birth, early years, and future of interventional radiology. *Journal of Vascular and Interventional Radiology*, *14*(7), 841-853.

Van Dalen, H.P. & Henkens, K. (2005). Signals in science - On the importance of signaling in gaining attention in science. *Scientometrics, 64*, 209-233.

Van Eck, N.J. & Waltman L. (2014), CitNeExplorer: a new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 8: 802-

Van Raan, A.F.J. (2004). Sleeping Beauties in science. *Scientometrics, 59*, 467-472.

Citation Distribution of Individual Scientist: Approximations of Stretch Exponential Distribution with Power Law Tails

O.S. Garanina and M.Yu. Romanovsky[1]

[1]*slon@kapella.gpi.ru*
A.M.Prokhorov General Physics Institute of RAS, Vavilov str., 38, 119991 Moscow (Russia)

Abstract
A multi-parametric family of stretch exponential distributions with various power law tails is introduced and is shown to describe adequately the empirical distributions of scientific citation of individual authors. The four-parametric families are characterized by a normalization coefficient in the exponential part, the power exponent in the power-law asymptotic part, and the coefficient for the transition between the above two parts. The distribution of papers of individual scientist over citations of these papers is studied. Scientists are selected via total number of citations in three ranges: 10^2-10^3, 10^3-10^4, and 10^4-10^5 of total citations. We study these intervals for physicists in ISI Web of Knowledge. The scientists who started their scientific publications after 1980 were taken into consideration only. It is detected that the power coefficient in the stretch exponent starts from one for low-cited authors and has to trend to smaller values for scientists with large number of citation. At the same time, the power coefficient in tail drops for large-cited authors.
One possible explanation for the origin of the stretch-exponential distribution for citation of individual author is done.

Conference topic
Citation and co-citation analysis

Introduction
The discussion of how citations of individual authors are distributed has a long history going back even to E. Garfield (1955). In general, there are two points of view on this: the distribution of papers of each scientist is a so-called stretch exponent $W \sim \exp(-x^{\alpha}/T)$, where x is the number of citations, T is some normalization, α is the power exponent coefficient (Redner, 1998; Laherrere & Sornette, 1998). Usually α is considered as 0,3-0,5 (Redner, 1998, Iglesias & Pecharroman, 2006). A slightly more complicated distribution was introduced by (Tsallis & de Albuquerque, 2000).
The second point is that the above distribution has power-law (Pareto, Zipf) character, i.e. $W \sim x^{-\beta}$ where β is the power (Silagadze, 1999; Vazquez 2001; Lehmann et al., 2003). Often, this dependence is treated as the asymptote (tail) of distribution for comparably large x. In this case, the main body is considered as log-normal (Redner, 2005; Stanley, 2010). It should be noted that there are more complicated models of citation distribution.
The idea of our work is to consider the citation distribution of individual scientists taking into account that the distributions for "various-ranking" scientists can be different. Also, it is interesting to join the above stretch-exponential distributions and power-law distributions: observation of tails of citation distributions of individual scientists often demonstrates a presence of small number of extremely-high cited articles, while other articles of considered scientists can be cited much more moderately. From this point of view, the consideration of citation data of a large set of authors (like in (Redner, 1998) etc.) provides rough enough results. Thus, we concentrate on analysis of citation distributions of individual scientists, taking into account some differences in the total number of citations of each. The cumulative distribution of the number of articles with some or larger number of citations will be analyzed.
Of course, the proposed approach is rough enough, since it does not take into account the co-authoring of cited articles. The authors think that it should be considered in further studies in case of wide scientific interest.

The descriptive model is based on our previous works for tailed distributions: Gauss for stock return distributions (Romanovsky & Vidov, 2011), and exponential Boltzmann distribution for new car sells, incomes and weights (Romanovsky & Garanina, 2015). The authors do not know consistently introduced mathematical formulae for distributions with exponential main part and power law asymptote.

Multi-parametric family of curves with stretch exponential main part and power law tail

To define the general form of the desired distribution, one may proceed from the results presented in (Romanovsky & Vidov, 2011) as a starting point. According to (Romanovsky & Vidov, 2011), the sum of a large quantity N of random values similarly distributed with the probability density function (PDF) of the Student's (generally, non-integer) type $\sim z_0^{2\beta}/(z_0^2+f^2)^{2\beta}$ has the distribution of the Gaussian form for comparably small values of fluctuations f:

$$W_G(f) \approx \frac{1}{\sqrt{\pi}} \exp(-f^2)$$

and $\sim 1/f^{2\beta}$ for large f (z_0 being a normalization constant, the sum is treated as random walks in (Romanovsky & Vidov, 2011)). The obvious mathematical generalization to get the exponential part with power-law tail is to perform the transformation $f^2 \to R/T$ (here T can be interpreted as an effective "temperature"). Upon switching from parameters N, z_0, β to parameters θ, T, $\sigma\beta$, the transformation yields the curve with the stretch exponential main part and a transition to power law at the tail in an explicit form of a PDF (Romanovsky & Garanina, 2015):

$$W_{T(\sigma\beta)\theta}(R) = \frac{1}{\sqrt{\pi T}} \int_0^\infty \cos(xR^\sigma) \left\{ \frac{2}{\Gamma(\beta - 1/2)} \left[(\beta - 3/2) \frac{xT}{4\theta} \right]^{\beta/2 - 1/4} K_{\beta - 1/2} \left[\sqrt{(\beta - 3/2) \frac{xT}{4\theta}} \right] \right\}^\theta dx \quad (1)$$

Here R is variable, Γ is the gamma-function, $K_{\beta-1/2}$ is the modified Bessel function of the 2nd kind (also known as ''McDonald function'').

The approximation of Eq. (1) for comparably small R (up to several units of $T^{1/2\sigma}$) is easily reduced to only a dependence on parameter T

$$W_T(R) \cong \frac{1}{T} \exp\left(-\frac{R^{2\sigma}}{T}\right) \quad (2)$$

The general drop off law for $W_{T\beta\theta}$ in the case of large R is $R^{-\beta\sigma}$. The parameter θ describes transition among (stretch) exponential and power-law part of (1). This transition goes under larger R (and smaller values of $W_{T(\sigma\beta)\theta}$) under larger values of θ.

To obtain a general form of W, note that

$$I_\beta(x) = \frac{2}{\Gamma(\beta - 1/2)} \left[(\beta - 3/2) \frac{xT}{4\theta} \right]^{\beta/2 - 1/4} K_{\beta - 1/2} \left[x \sqrt{(\beta - 3/2) \frac{T}{\theta}} \right], \quad (3)$$

It is easy to see that it is a monotonic function of β. Indeed, if $\nu = \mu + 1$, one finds, considering the rule for modified Bessel functions of the 2nd kind, that the ratio $I_\mu(x)/I_\nu(x)$ becomes

$$\frac{I_\mu(y)}{I_\nu(y)} = \frac{K_{\mu+1/2}(y) - K_{\mu-3/2}(y)}{K_{\mu+1/2}(y)} = 1 - \frac{K_{\mu-3/2}(y)}{K_{\mu+1/2}(y)} < 1$$

Furthermore, $\forall \eta : \nu > \eta > \mu$, and one finds that $I_\nu > I_\eta > I_\mu$. Thus, it is not necessary to investigate (1,3) with an arbitrary β. It is enough to consider the integer $\beta = 2, 3, \ldots$, while integrals with intermediate β will be ''locked'' among integrals with neighboring integers β that are expressed by means of elementary functions. Then $n = \beta - 1$,

$$K_{\beta - 1/2} \left[x \sqrt{(\beta - 3/2) \frac{T}{\theta}} \right] = K_{n+1/2} = \sqrt{\frac{\pi}{2x \sqrt{(\beta - 3/2) \frac{T}{\theta}}}} \sum_{k=0}^{n} \frac{(n+k)!}{k!(n-k)! \left[2x \sqrt{(\beta - 3/2) \frac{T}{\theta}} \right]^k} \quad (4)$$

The three functions $W_{T(\sigma\beta)\theta}$ for $\sigma\beta = 2, 1, 0.8$ are:

$$W_{T(\sigma\beta)\theta}(R) = \frac{1}{\sqrt{\pi T}} \int_0^\infty \cos(xR^{\sigma\beta}) \exp\left(-x\sqrt{\frac{\theta T}{2}}\right)\left(1 + x\sqrt{\frac{T}{2\theta}}\right)^{\theta} dx \qquad (5)$$

We used here the simplest form of the function (1) for β=2 for the following approximations of empirical data. The functions $W_{T(\sigma\beta)\theta}$ for σ= 0.5, 0.25, 0.2 are shown in Fig.1. It is seen as a well-coincidence of general functions with corresponding approximation exponents for comparably small values of variable R.

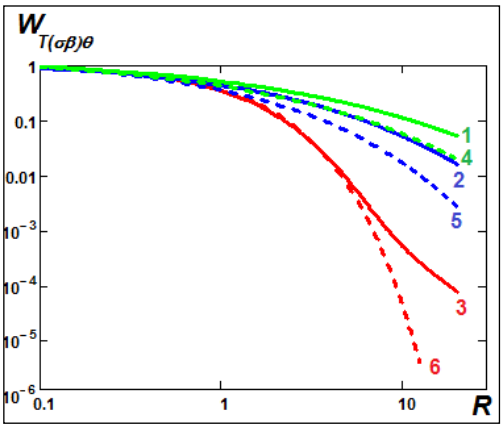


Figure 1. Functions $W_{T(\sigma\beta)\theta}$ for β=2 and σ=0.5 (curve 3), σ=0.25 (curve 2), σ=0.2 (curve 1) for comparably small R. The straight lines (4-6) are exponents $exp(-R^{2\sigma}/T)$ for σ=1,0.5,0.4, respectively. Here T=1, θ=300.

For large R, these functions drop off as R^{-2}, R^{-1}, $R^{-0.8}$, respectively (see Fig.2):

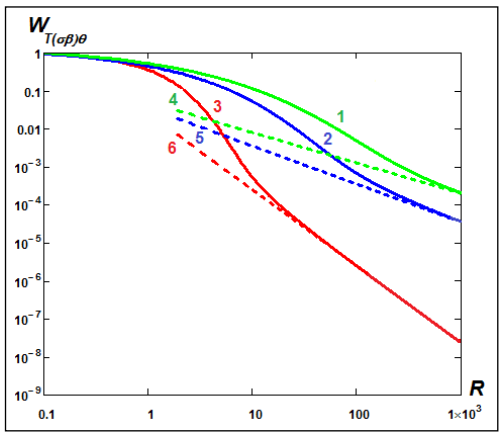


Figure 2. Functions $W_{T(\sigma\beta)\theta}$ for the same β and σ (curves 3-1) as on Fig.1. Hyperboles $R^{-\beta\sigma}$ (straight lines 6-4 on double-logarithmic plot) have σ=0.5, 0.25, and 0.2 (curve 4), respectively. Parameters T, θ are the same as on Figure 1.

Thus the introduced function (1) well-describes the stretch exponent for small (and moderate) values of argument, and provides power-law asymptotes for large R. We used these functions in the next section.

Distribution of citation of individual authors

It was found that the distributions of citations of individual authors are different. It can be expected due to, for example "Matthew effect" (see Bonitz et al., 1997; Bonitz & Scharnhorst, 2001; Stanley, 2010). One may expect that scientists with total number of citation in range 10^2-10^3, 10^3-10^4, and 10^4-10^5 have different distributions of citations. Let us call the scientists with total number of citations in these ranges as the "first-type scientist", etc. We study these intervals for physicists in the ISI Web of Knowledge. The scientists who started their scientific publications

after 1980 were taken into consideration only. We took 20 scientists for the first two ranges, and several scientists for the third. Typical examples of citation distributions are presented below on Figs. 3-5.

On Fig. 3, the cumulative citation distribution (i.e. the number of articles with citations larger than the value R) for experienced scientists with total number of citations in the first range 10^2-10^3 is presented:

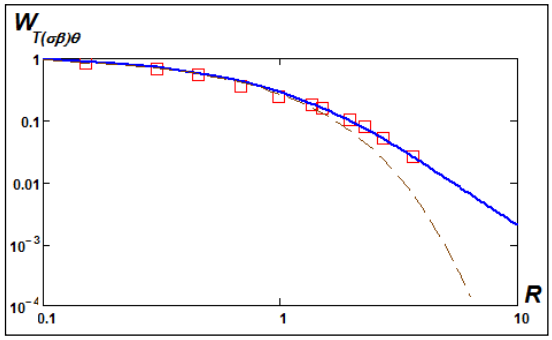


Figure 3. The distribution of articles over citations for the first-type scientist. Open squares are empirical points, the solid curve is $W_{T(\sigma\beta)\theta}$ (5) for β=2, σ=0.5, T=6.5, θ=10, dashed line is an exponent (2) with σ=0.5, T=6.5.

The function $W_{T(\sigma\beta)\theta}$ on Fig.3 is normalized on total number of articles of the first-type scientists in ISI Web of Knowledge. The variable R is the number of citations normalized on T that is the mean citation of this author. It is seen that the function $W_{T(\sigma\beta)\theta}$ (5) well describes the empirical data, the clear difference from the exponent (2) is on-site. At the same time, the total exit on the asymptotic curve $\sim R^{-2}$ does not realize. The last was observed for other-types scientists.

The citation distribution of the second-type scientist (this is a range of world well-known person) is demonstrated on Fig. 4:

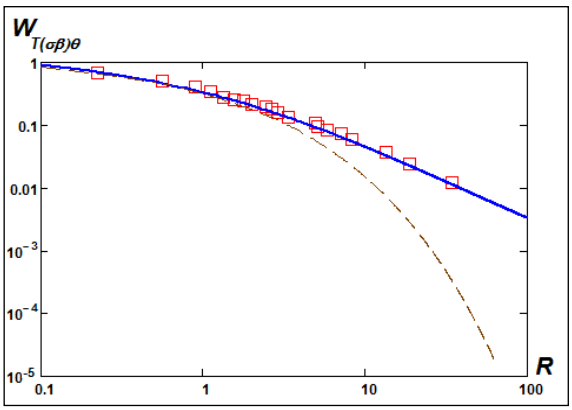


Figure 4. The distribution of articles over citations for the second-type scientist. Open squares are empirical points, the solid curve is $W_{T(\sigma\beta)\theta}$ (5) for β=2, σ=0.25, T=47.4, θ=5, dashed line is an exponent (2) with σ=0.25, T=46.

The normalization of $W_{T(\sigma\beta)\theta}$ on Fig.4 was on total number of articles also. Indeed, the variable R is normalized now on $T^{2\sigma} = (47.4)^{2\sigma} = 6.9$. The "difference" between empirical data as well as function (5) with pure stretch exponent $exp(-R^{1/2}/T)$ is larger than on Fig.3 for the first-type scientist. The total exit on the asymptotic curve $\sim R^{-1}$ is also not realized.

The citation distribution of the third-type scientist (this is a range of Nobel Prize winners) is demonstrated on Fig. 5:

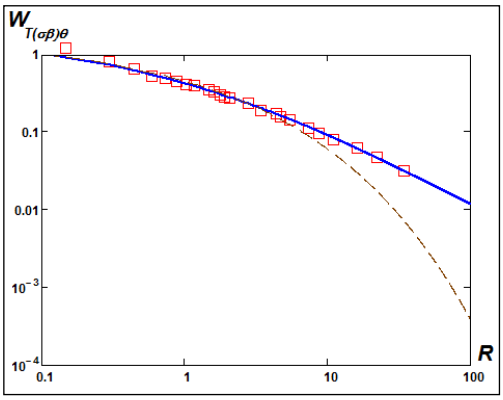


Figure 5. The distribution of articles over citations for the third-type scientist. Open squares are empirical points, the solid curve is $W_{T(\sigma\beta)\theta}$ (5) for β=2, σ=0.2, T=340, θ=5, dashed line is an exponent (2) with σ=0.2, T=340.

The normalization of $W_{T(\sigma\beta)\theta}$ on Fig.5 is the same, the variable R is normalized now on $T^{2\sigma} = 340^{2\sigma} = 10.3$. It is interesting that all values $T^{2\sigma}$ for all three-types scientists are close to each other and may characterize the citation distribution of individual scientists.

Explanation attempt

Let us try to explain the appearance of stretch exponents in cumulative distribution of such random values like citations. We start from the standard exponential distribution

$$W_1 = \exp(-x) \tag{6}$$

where we used normalization T=1 to simplify the following expressions.

How can these calculations be "translated" into the language of citations? The first cause of a citation of some article is the scientific results of this article. Since the author who can potentially cite the above article may find or not find this article, the process of citation due to the scientific significance looks like the two-body exchange (of information in this case) and is provided by distribution (6). Thus it may be that the basic cumulative distribution of citations arises due to the scientific significance of the article and looks like (6).

There are clear additional independent causes for citations. One of them is the name of author (or one of authors in case of co-authoring) of a potentially cited article. It may be the name of scientist in the group that works in the same area of science studied with the author of the cited paper, there arises another causes to cite some scientist. Since this scientist may also be chosen randomly in the process of information exchange, the probability distribution to cite this scientist looks like (6) as well. If now the citation is realized due to two causes: by scientific significance and cited article author, the random value of such citation is the factor of two random values characterized by distribution (6).

Since the causes for citation are independent, they can be considered as some coordinates. For two cases, they are above "scientific significance" and "author's name". The variation of these coordinates here are from small to large scientific significance and from large to small reputation of cited scientist. At the same time, we observe citation as being a principally one-dimensional value: the citation either exists or does not exist. Therefore, all distributions (6) reduce to one dimension. The transformation of coordinates in (7) $x^2 \rightarrow y$ provides than for cumulative distribution function

$$W_2(y) = \exp\left(-\sqrt{y}\right) \tag{7}$$

i.e. the main part of stretch exponent (2) with σ=0.25. These stretch-exponents distributions were observed by us and described in the chapter of this paper "Distribution of citation of individual authors".

The same procedure in case of three clearly existed "coordinates" provides cumulative distribution

$$W_3(y) = \exp\left(-\sqrt[3]{y}\right) \quad (8)$$

The same conduction for power-law tailed stretch exponential distributions should take into consideration the power exponents in tails for original distributions of "scientific significance" etc., and needs the volumetric calculations.

Conclusion

The 4-parametric family of functions representing the stretch exponential distribution for small and medium values of the argument combined with a power-law asymptotic tail, along with various transitions between these two parts, is introduced. These functions are demonstrated as good fits of the available empirical data for the cumulative distribution of citations to individual scientists.

Abstracting from the co-authoring of a cited paper, one may conclude that these cumulative distributions of papers of individual authors versus their citations have character of stretch exponent for small and moderate values of citations, and power-law form for asymptotic part. It looks that the "power of stretch", i.e. the introduced coefficient σ depends on the total number of citations, moreover, this coefficient starts from ½ (i.e. distributions start from normal exponent) and becomes smaller with an increase of the total number of citations. The power-law force becomes smaller in return.

The first attempt to explain the "main body" of distributions (stretch exponents) is provided.

Acknowledgements

The paper is support by RFBR grant 13-07-00672.

References

Bonitz, M., Brukner, E., & Scharnhorst, A. (1997). Characteristics and impact of Matthew effect for countries. *Scientometrics*, *40*, 407-422.

Bonitz, M., & Scharnhorst, A. (2001). Competition in science and the Matthew core journals. *Scientometrics*, *51*, 37-54.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*. 108–111.

Iglesias, J.E., & Pecharroman, C. (2006). Scaling the h-Index for Different Scientific ISI Fields. Online: http://arxiv.org/ftp/physics/papers/0607/0607224.pdf

Laherrere, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B*, *2*, 525-539.

Lehmann, S., Lautrup, B., & Jackson, A.D. (2003). Citation networks in high energy physics. *Physics Review E,* 68. 026113.

Petersen, A.M., Fengzhong Wang, & Stanley, H.E. (2010). Methods for measuring the citations and productivity of scientists across time and discipline. *Physics Review E, 81*, 036114.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B, 4*, 131-134.

Redner, S. (2005). Citation statistics from 110 years of Physical Review. *Physics Today,* 58. 49–54.

Romanovsky, M.Yu., & Vidov, P.V. (2011). Analytical representation of stock and stock-indexes returns: Non-Gaussian random walks with various jump laws. *Physica A*, *390*, 3794–3805.

Romanovsky, M.Yu., & Garanina, O.S. (2015). New multi-parametric analytical approximations of exponential distribution with power law tails for new cars sells and other applications. *Physica A*, 427, 1-9.

Silagadze, Z.K. (1999). Citations and the Zipf-Mandelbrot's law, http://arXiv.org/abs/physics/9901035v2

Tsallis, C., & de Albuquerque, M.P. (2000). Are citations of scientific papers a case of nonextensivity? *The European Physical Journal B*, 13, 777-780.

Vazquez, A. (2001). Statistics of citation networks. *E-prints* arXiv:condmat/0105031.

Influence of International Collaboration on the Research Citation Impact of Young Universities

K. A. Khor and L.-G. Yu

mkakhor@ntu.edu.sg; mlgyu@ntu.edu.sg
Research Support Office and Bibliometrics Analysis, Nanyang Technological University, #B4-01, Block N2.1, 76, Nanyang Drive, Singapore 637331 (Singapore)

Introduction

It is widely presumed that international collaboration benefits the researchers and the organisations involved, and enhances the quality of research (Persson, 2010). However, research also suggests that the effects of international collaboration may vary across disciplines and the authors' countries (Moed, 2005).

In this study, we investigated the effect of international collaboration on the impact of publications of selected young universities, and compared to that of renowned old universities. The 5-year citations per paper (CPP) data, the international collaboration rate, the CPP differential between publications with and without international collaborations, and the difference between the percentages of international collaborated publications falling in the global top 10% highly cited publications and the percentage of overall publications falling in the global top 10% highly cited publications (Δ%Top10%) are used as the impact indications. These data are extracted from the Thomson Reuters Web of Science (WoS) database and Essential Science Indicator (ESI) based on papers published from 2004 to 2013. Young institutions ranked by the 2014 Times Higher Education (THE)'s 100 under 50 Universities are selected in this study, and some renowned universities (> 100 years old) are selected as references for "old universities".

To eliminate the discipline difference effect, the increment of 5-year (2010-2014) field weighted citation impact (FWCI) of internationally collaborated papers over the 5-year overall FWCI of the institutions in SciVal® of Elsevier is used as another indicator. The collaboration among 8 old institutions and 8 young institutions are investigated.

Results and Discussion

Correlation between International Collaboration rate and CPP in 5-year interval

Figure 1 shows the 5-year ESI CPP trends as a function of 5-year international collaborations rate trends for selected young and old universities. While old universities have higher CPP in general, there are strong correlation between international collaboration rate trends and 5-year CPP trends. For example, for old universities, the CPP increased 4.12 for every 10% increase in international collaboration rate for MIT, 3.42 for Univ Oxford, and 3.01 for Stanford Univ. Among young universities, for Nanyang Technol Univ (NTU), it is 2.24 CPP per 10% Intl Collab increment, and that for Plymouth Univ is 3.02, and 0.73 for King Fahd Univ of Petr and Min.

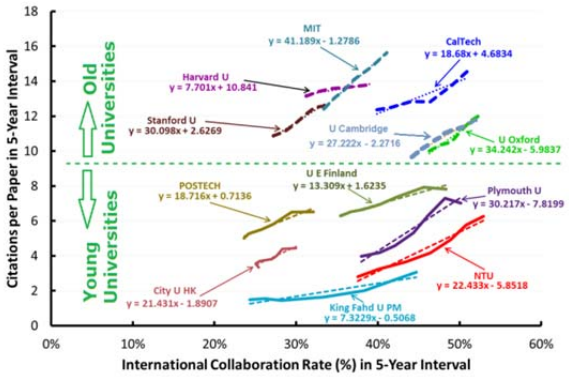


Figure 1. 5-Year CPP Trends vs. 5-Year International Collaborations Rate Trends for Selected Young and Old Universities.

The ΔCPP trends for publications with and without international collaborations for selected institutions are examined, and listed in Table 1.

Table 1. 5-Year Citations per Paper Differential between Publications with and without International Collaborations.

5-Year Period	Citations per Paper Difference between Publications with and without International Collaboration											
	Caltech	U E Finland	Univ Florence	Univ Tsukuba	Univ Melbourne	Univ Waikato	Kyushu Univ	MIT	NUS	HKUST	NTU	USM
2004-2008	5.5	3.04	3.59	5.19	4.5	2.68	2.26	3.24	0.78	1.03	0.2	1.08
2005-2009	5	3.38	3.68	5.65	4.06	1.68	2.62	3.25	0.66	1.44	0.51	0.97
2006-2010	4.2	3.42	3.79	4.87	4.3	2	2.55	3.38	0.63	0.55	0.43	0.65
2007-2011	4.2	4.1	3.91	4.85	4.42	2.1	1.85	2.68	0.82	1.33	0.47	0.11
2008-2012	4.8	4.44	4.38	4.65	4.77	2.86	1.75	2.29	1.28	1.44	0.05	-0.3
2009-2013	6.1	5.28	5.3	5.2	4.87	3.61	2.4	2.16	1.67	0.87	0.02	-0.7
ESI 2009-2013 CPP	15	8.53	7.68	6.48	8.66	5.43	5.28	15.7	7.83	6.7	6.92	3.47

From Table 1, we can find that in the case of Caltech, U Melbourne and U Tsukuba, the CPP difference between their international collaborated

publications and their publications without international collaboration is roughly 4 to 5. This explains the typical 5-year ESI CPP vs. international collaboration rate trends of these institutions: with the increase of international collaboration rate in their publications, the overall CPP of their papers has more weight from their international collaborated publications, and the overall CPP of their publications increased. Yet, for Hong Kong Univ of Sci & Techn (HKUST), Natl Univ Singapore (NUS) and NTU, the CPP gaps between publications with and without international collaboration are relatively small (around 0 to 1 CPP). This is because the fact that these institutions have attracted a lot of researchers with international background to work in these institutions, which makes the difference between their national research and international collaborated research relatively small.

Trends of difference between percentage of international collaborated publications falling in global top 10% highly cited publications and that for all publications (Δ%Top10%)

The study on difference between the percentage of international collaborated publications for an institution falling in the ESI global top 10% highly cited publications and the percentage of all publications of the same institution falling in the ESI global to top10% highly cited publications (Δ%Top10%) shows that, for all the selected young and old institutions, this difference is generally positive, means that internationally collaborated publications generally have a higher rate of high citation publications among all publications. Yet, this difference varies from one institution to another institution. For some renowned top universities like Caltech, Stanford University and University of Cambridge, although their overall CPP for their publications is already very high, the Δ%Top10% is still higher than the percentage of their overall publications falling in the global top 10%. Further investigation is needed to have an adequate explanation for this phenomenon.

Increment of field weighted citation impact (FWCI) of internationally collaborated papers over the FWCI of the involved institutions

Figure 2 shows the increment of FWCI for internationally collaborated papers over the overall FWCI of the two collaborating institutions among the selected 8 old institutions and 8 young institutions. 57 bilateral collaboration couples with 50 and more collaborating publications are identified among these 16 institutions, and the FWCI increment data for these collaboration couples are include in the plot. It can be seen that, international collaboration benefits both the young and the old institution, with the old institution to old institution collaboration provides the highest FWCI increment, followed by the old institution to young institution collaboration. Among the 57 bilateral collaborations, only 3 involved young institution to young institution collaboration, indicating that there are untapped potential for enhancement on bilateral collaboration among young institutions.

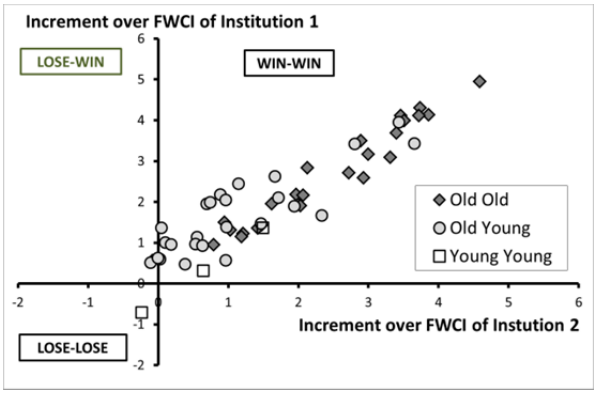


Figure 2. Increment of 5-year FWCI of internationally collaborated papers over the overall FWCI of the involved Institutions.

Conclusions

The investigation on the effect of international collaboration on the impact of publication of selected young universities and well established renowned universities show that, both young and old institutions received benefit from international collaboration using citation impact of their publications as indicator. For example, for old universities, the CPP increased 4.12 for every 10% increase in international collaboration rate for MIT and 3.42 for U Oxford. Among young universities, for NTU, it is 2.24 CPP per 10% Intl Collab increment, and that for Plymouth U is 3.02 CPP per 10% Intl Collab increment.

The percentage of publications fall in the ESI global top 10% highly cited publications for international collaborated publications is generally higher than that for all journal publications of the same institution. Yet, this difference varies from one institution to another institution.

The international collaboration also increases the FWCI of the institution, yet there are untapped potential to enhance the collaboration among young institutions.

References

Persson, O. (2010). Are highly cited papers more international? *Scientometrics*, *83*(2), 397-401.

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

Which collaborating countries give to Turkey the largest amount of citation?

Bárbara S. Lancho Barrantes

b.lancho@csic.es
Spanish National Research Council (CSIC), Agustín Escardino, 7. 46980 Valencia (Spain)

Introduction

In the scientific world it is recognized that high levels of collaboration, but particularly international scientific collaborations, lead to increase in citations, a better quality of the papers published, and a greater productivity of the authors (Leimu & Koricheva, 2005; Hsu & Huang, 2010).
However this citation increment may vary across nations. For various reasons, there might exist differences on the type of collaboration due to countries and their size (Zhao & Guan, 2011).
Therefore in order to study this phenomenon will concentrate on the scientific collaboration between Turkey and the nine most productive countries in the world in 2004 (USA, China, Japan, UK, Germany, France, Canada, Italy, Spain). When considering these countries, the following concerns emerge:

Research questions

Which countries are working more closely with Turkey? From which countries does Turkey receive more citations? How are the averages in terms of references made by Turkey to collaborators? The main idea examined in this work revolves about the increase in citations occurring when Turkey collaborates with a certain country, since the increase in received citations would be higher compared to a scenario in which the cooperation with such nation had not taken place. Particularly, percentage of citation increase is analyzed through the number of citations received by Turkey from collaborating countries and through the number of references given by Turkey to the nine collaborating countries.

Data and Methods

The same data and indicators from the studies Lancho et al. 2013; and Lancho, Guerrero & Moya, 2013 were used for this analysis.
The main indicators used are as follows:
• Citations per paper: Average citations received by the papers published in 2004 within papers from 2005–2007.
• References per paper: Average references given by papers published in 2005– 2007 to papers from 2004.
• Citation Rate Increment from the Collaborator (CRIC): Citation Rate Increment Average when Collaborating (CRIAC), and the Citation Rate Increment obtained from its Collaborators (CRIOC).

Results

The total number of documents belonging to Turkey during this time period was 18170. 3043 papers (16.74% of the total number of papers) were produced from collaboration with one or more countries.

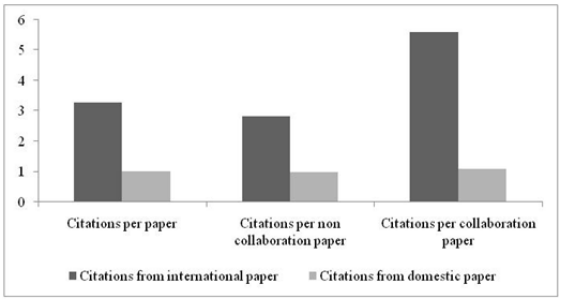


Figure 1. Comparison among the different averages in terms of citations made to Turkey, distinguishing in both cases between domestic and international articles.

The number of citations per collaboration paper is significantly bigger than those of the citations per non-collaboration paper and citations per paper, being international papers the root where this difference is originated.

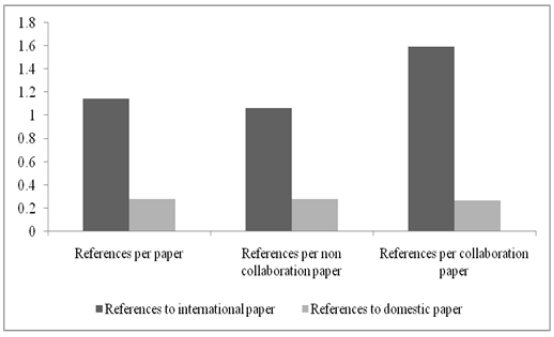


Figure 2. Comparison among the different averages in terms of references made by Turkey, distinguishing in both cases between domestic and international articles.

The number of references per collaboration paper is larger than the one registered by references per non-collaboration paper and references per paper in general. Although these percentages are not much different from each other it notices a slight benefit when collaborating.

Table 1. This chart is referred to the total production in collaboration with Turkey and the total citations made to documents in collaboration with Turkey.

Country	Papers with different countries	Citation to collaboration documents	Citations from collaborators
United States	1368	9206	3978
United Kingdom	411	3082	721
Germany	345	2738	543
France	163	1735	318
Japan	157	869	127
Italy	150	2223	334
Canada	126	963	112
Spain	69	1234	146
China	34	527	53

By observing the above illustration, the United States is the country with which Turkey collaborates more, following this United Kingdom and Germany. And these are the countries that Turkey most benefits from reflected in Citations to collaboration documents and Citations from the Collaborators.

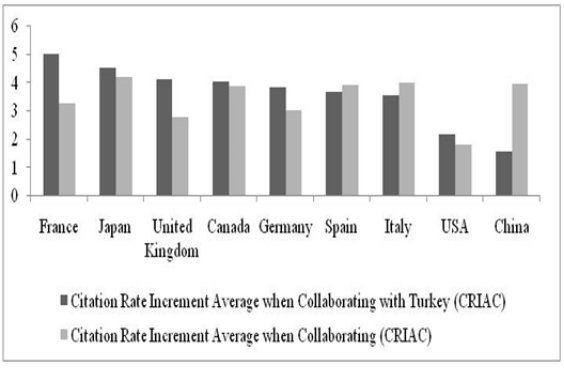


Figure 3. Comparison between CRIAC in general and CRIAC with Turkey.

On a general basis, except in some cases, the increase in citations arising out from collaborating countries is higher in Turkey than in a general study.

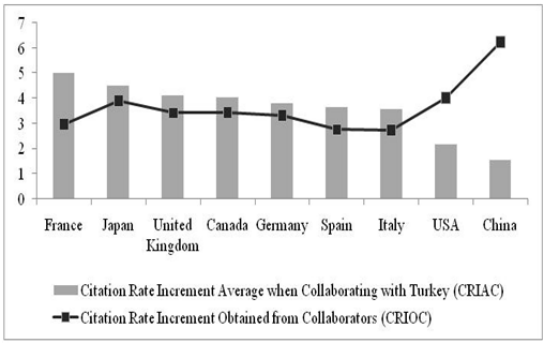


Figure 4. Comparison between the CRIAC with Turkey and the CRIOC among the nine countries with the largest production in 2004.

Values for the CRIAC were higher in some countries than in others in comparison with CRIOC.

Interpretation

Turkey is a country presenting large levels of production, but it has a very low percentage of documents done in collaboration. However, its citation percentage received from its collaborations with countries having larger productions and more collaboration, such as France or Japan it quite high.

If Turkey is involved in collaborations, it receives a positive Citation Rate Increment from the Collaborator (CRIC).

However, Turkey does not receive the same Citation Rate Increment Average when Collaborating (CRIAC) from all the countries. For instance, the largest increases in citations are registered in France, Japan, and the UK.

Finally, this study is only an approximation of how Turkey collaborates and from which it is revealed interesting data that could be developed by a broader study in which more countries and scientific disciplines could take part.

References

Hsu, J.W., & Huang, D.W. (2010). Correlation between impact and collaboration. *Scientometrics, 86*(2), 317–324.

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., & Moya-Anegón, F. (2013). Citation flows in the zones of influence of scientific collaborations. *JASIST, 63*(3), 481–489.

Lancho-Barrantes, B. S., Guerrero-Bote, V. P. & Moya-Anegón, F. (2013). Citation increments between collaborating countries. *Scientometrics, 94*(3), 817 - 831.

Leimu, R., & Koricheva, J. (2005). Does scientific collaboration increase the impact of ecological articles? *Bio Science, 55*, 438–443.

Zhao, Q., & Guan, J. (2011). International collaboration of three 'giants' with the G7 countries in emerging nanobiopharmaceuticals. *Scientometrics*, *87*(1), 159–170.

Do We Need Global and Local Knowledge of the Citation Network?

S.R. Goldberg[1], H. Anthony and T.S. Evans[2]

[1] *s.r.goldberg@qmul.ac.uk*
Queen Mary University of London, School of Physics and Astronomy, London, E1 4NS, (U.K.)

[2] *t.evans@imperial.ac.uk*
Imperial College London, Centre for Complexity Science & Physics Department, London, SW7 2AZ, (U.K.)

Introduction

Models which reproduce key features of the distribution citations to academic papers have a long history (Price, 1965). One aim is to illustrate if certain simple processes can explain important features. In this paper we focus on the fact that the distribution of citations for papers of a similar age scales primarily with the average number of citations (Radicchi, Fortunato, & Castellano, 2008; Evans, Hopkins & Kaube, 2012), with the shape otherwise largely invariant. In particular the width shows no temporal evolution. Simple multiplicative processes or basic models such as the Price model (Price, 1965) give dramatically different results, typically the distributions become narrower over time. The purpose of this study is to find a simple model which can lead to the observed behaviour of citations over time.

Methods

Consider a set of N papers all published in one year with an average number of citations C. We take 'reasonably well cited' papers with $c>0.1C$ and following Evans, Hopkins and Kaube (2012) we fit the number of papers with c citations to a log-normal distribution

$$\frac{n(c)}{N} = \int_{c-0.5}^{c+0.5} \frac{dx}{\sqrt{2\pi}\sigma x} exp\left\{-\frac{(\ln(x/C)+\sigma^2)^2}{2\sigma^2}\right\}$$

The log-normal form is an effective description and our only interest here is that the σ parameter is a reasonable characterisation of the width of the distribution. We want to find a model which has the correct properties for this width, namely it is roughly constant over time and of the right size. We compare outputs from our models against measurements made on data from the citation network of the hep-th section of the arXiv repository (KDD cup 2003).

We tried three models. In model A, with probability p papers are cited in proportion to their current number of citations, Price's cumulative advantage (Price 1965), otherwise the papers cited are chosen uniformly at random. In model B both these probabilities are modified by a factor $\exp((N-t)/\tau)$ for paper number $(N+1)$ where τ is a time scale parameter.

Models A and B are based purely on global information – knowledge of the whole network is required. This reflects authors discovering papers using mechanisms other than the bibliographies of papers.

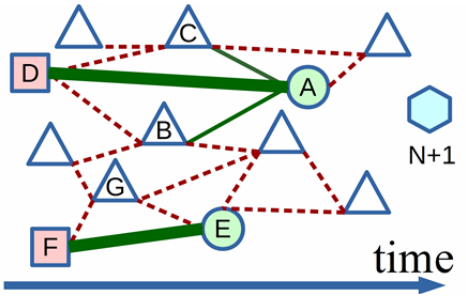


Figure 1. Illustration of Model C. A new paper (hexagon, *N+1*) is set to have four references. The first 'core' paper is chosen, A, using the global process of model B. Then with probability *q*, papers cited by A are also added to the new bibliography. Here B and C are considered (thin solid lines) but only D is added (thick line). The process continues until the required bibliography is complete. Here a second core paper E is chosen and one of its citations, F, is copied. At that point the process stops, paper G is never considered. The new bibliography is A, D, E and F.

For model C we add a second process, which uses only local information, see Figure 1. A set of 'core' papers are chosen as in model B. However each time a core paper is chosen, we examine each of the papers cited by this core paper and with probability q we add each to the new bibliography. This random walk from core papers to subsidiary papers is known to generate an effective cumulative attachment (Evans & Saramäki, 2005). In all cases we choose the length of the bibliography from a normal distribution with the same mean, 12.0, and standard deviation, 3.0, as measured in our hep-th data. The models involve a small number of parameters which have to be chosen. One feature we use is the number of zero cited papers and we match that to the proportion found in our results. We also look at the time it takes a paper in our model to reach half its final citations in order to find an optimal τ value. Finally parameter q in Model C is set by using an approximate form of

transitive reduction (Clough et al., 2014) to estimate the faction of core papers in our data.

Results

Both our Models A and B produced long-tailed citation distributions but in both cases the width parameter σ was significantly smaller than that found in our data. However we were able to find a range of parameters where Model C was consistent with our data, for example see Figure 1. In particular the papers produced in one year had fat tails with a width σ which was roughly constant in time.

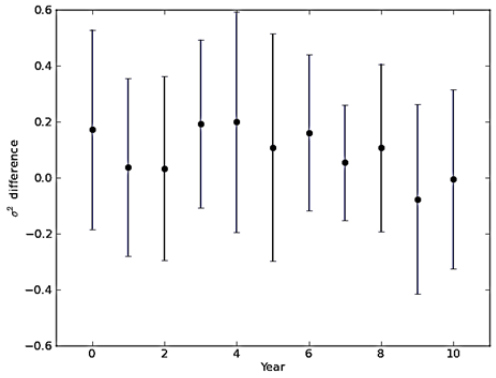


Figure 1. The difference between the width σ of the hep-th data and that found in our Model C for final fitted parameters.

Discussion

We started from the observation that the width of the fat-tailed citations distributions for papers published in one year show some consistent patterns. In particular, in terms of our log-normal width parameter, σ, this width is roughly constant and independent of the age of the papers studied. To keep our work rooted in real citations, we worked with hep-th arXiv data which also shows this characteristic static width.

The difficulty in finding a model which reproduces this key feature was illustrated by results from our first two models: Model A mixed cumulative and uniform random attachment while Model B added a time decay to favour citations to more recent papers. We were unable to find parameter regimes where these models provided good fits to our data.

However our model C with just three parameters was able to produce an accepted fit to the hep-th data over 11 different years, see Figure 1.

The big difference between model C and our earlier attempts is that only in model C was local information as well as global information used to find references for a new paper. We conclude that the citation patterns we see reflect a mixture of local searches of the citation network (reading papers and finding the papers they cite) along with global information providing the recommendation (a chance personal suggestion at a conference perhaps).

Another interesting result is that we find the best fits for our model to our data is when around 70% to 80% of papers cited are 'subsidiary papers', papers found from local searches through the bibliographies of other papers. Interestingly similar results have been found seen by Simkin and Roychowdhury (2005) who arrive at a similar model but for different reasons. Namely they suggested that mistakes in bibliographic entries suggest that around 80% of citations are copied (Simkin & Roychowdhury, 2003). In our terminology these would be citations to subsidiary papers so both sets of results are consistent. Further support for this result comes from the transitive reduction analysis of Clough et al. (2014)

Finally we suggest that more work needs to be done to capture the effect of the variation in the length of bibliographies. We used a normal distribution for this aspect. This encodes some fluctuations in this bibliography length, something usually neglected in other models, but the reference distribution should also be fat-tailed. We failed to get good agreement with data when we modelled bibliography length this way.

Acknowledgments

We would like to thank James Clough, James Gollings, and Tamar Loach for sharing their results on related projects.

References

Clough, J.R., Gollings, J., Loach, T.V. & Evans, T.S. (2014). Transitive reduction of citation networks *J. Complex Networks* (to appear) http://dx.doi.org/10.1093/comnet/cnu039.

Evans, T.S; Hopkins, N. & Kaube, B.S. (2012). Universality of Performance Indicators based on Citation and Reference Counts. *Scientometrics*, *93*, 473-495.

Evans, T.S. & Saramäki, J. (2005). Scale-free networks from self-organization. *Phys.Rev. E*, *72*, 026138.

KDD Cup (2003). Network mining and usage log analysis. Retrieved October 1, 2012 from http://www.cs.cornell.edu/projects/kddcup/datasets.html .

Radicchi, F., Fortunato, S. & Castellano, C. (2008). Universality of citation distributions: Towards an objective measure of scientific impact. *PNAS*, *105*, 17268-17272.

Goldberg, S.R., Anthony, H. & Evans, T.S. (2014). Modelling citation networks, Scientometrics (to appear) [*arXiv:*1408.2970].

Simkin M.V. & Roychowdhury V.P. (2003). Read before you cite! *Complex Systems*, *14*, 269-274.

Simkin M.V. & Roychowdhury V.P. (2005) Stochastic modeling of citation slips. *Scientometrics*, *62*, 367-384.

Citation analysis as an auxiliary decision-making tool in library collection development

Iva Vrkić

ivavrkic@gfz.hr
University of Zagreb, Faculty of Science, Department of Geophysics, Geophysical Library, Horvatovac 95, 10000 Zagreb (Croatia)

Introduction

Academic libraries in Croatia are facing constant budget cuts, making it difficult to obtain access to current scientific and professional journals (Krajna & Markulin, 2011). At the end of 2008 the Croatian economy had plummeted into recession and the *Ministry of Science, Education and Sports* ceased the funding of scientific literature acquisition (Krznar, 2011).

Parallel to budget cuts, the prices of scientific journals increased. The period from 2009 to 2014 showed a threefold increase in prices of the journals acquired by the Geophysical library in Zagreb (Figure 1), making it necessary to review the need for the purchase of each journal.

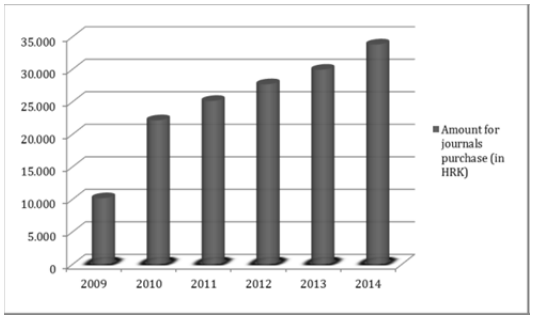


Figure 1. Threefold increase in prices of the journals acquired by the Geophysical library in Zagreb.

Quantitative and qualitative methods can be used to make optimal decisions regarding the purchase of journals (Gomez, 2002). The qualitative method is based upon interviewing lecturers and other competent scientific staff and taking their suggestions on which journals are essential. Their assessment of the journals' relevance is the most important guideline in creating an acquisitions policy. The quantitative method, on the other hand, provides the much-needed objectivity in the acquisitions process, but can only be used as an additional guideline to the qualitative method. This method can come in the form of usage statistics or the assessment of the journal's importance through citation analysis. Such an assessment is described in this paper. Although the quantitative method is objective, its results (list of most used/most relevant journals) cannot replace subject-matter experts' opinion, only inform them.

Methodology

The goal of this study is to determine the importance of certain journals for the geophysical community at the Faculty of Science in Zagreb. This will be done by compiling a list of journals most cited by the scientific staff at the Geophysical department from 2000 to 2014. References from all scientific papers published by the staff at the Department of Geophysics in the last 14 years were collected, and 6120 references were selected from journals cited by our geophysicists. The citation frequency was analysed, and references were listed for each journal.

Results and discussion

Assuming the citation frequency of articles from a certain journal confirms its importance for the scientists, the journals were listed by relevance after the data had been processed. The result is a list of 512 journals ranked by the number of citations. A "Top 15" list has also been created – 15 most cited journals by the members of the Department of Geophysics from 2000 to 2014 (Table 1).

Table 1. *Top 15* – most cited journals by the members of the Department of Geophysics form 2000 to 2014.

	Journal title	∑ citation
1	Journal of Geophysical Research	448
2	Journal of the Atmospheric Sciences	349
3	Quarterly Journal of the Royal Meteorological Society	335
4	Monthly Weather Review	222
5	Boundary-layer meteorology	213
6	Journal of Climate	202
7	Atmospheric Environment	195
8	Journal of Applied Meteorology and Climatology	185
9	Geofizika	162
10	Geophysical Research Letters	115
11	Tellus	114
12	International Journal of Climatology	108
13	The Astrophysical Journal	97
14	Bulletin of the Seismological Society of America	95
15	Annales Geophysicae	93

Data on the age of journal citations (cited by the members of the Department of Geophysics in a 14-

year period) was processed. Citation age is determined as the discrepancy between the publishing years of both the cited and the citing paper.

The citation age median for the whole set is 9 years. The histogram (Figure 2) shows that half of the citations are 0 to 9 years old, and rest of them are 10 to 133 years old. Citation frequency in 1st quartile shows statistically significant greater representation of citations in relation to the 2nd quartile ($\chi^2 = 9{,}86$; $P<0{,}0017$).

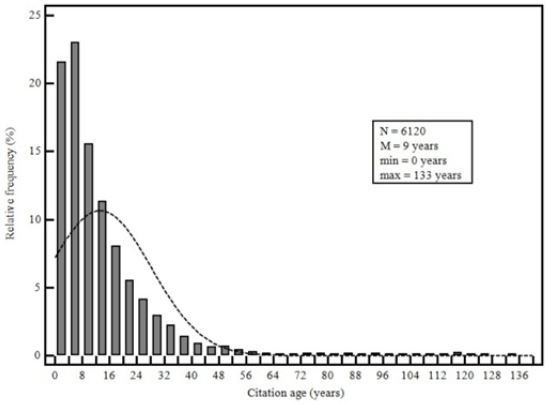


Figure 2. Citation frequency relative to citation age.

Therefore, recent scientific papers are the most cited.

Instead of a conclusion

Why is optimizing the library's acquisitions policy so important? The answer is, of course, because optimization is crucial in creating a list of the most relevant journals to be acquired, which can also be illustrated using the *Pareto principle*.

The Pareto principle is, amongst other thing, used to evaluate periodicals collections. It was named after Vilfredo Pareto, an Italian sociologist, who first used it to explain the distribution of land in Italy, where 80% of the land was owned by 20% of the population.

As previously mentioned, the principle applies to many different areas, so if applied to a periodicals collection, it will show that 20% of the periodicals in the collection will cover 80% of information needs. Also, 80% of the citations will be found in 20% of the periodicals (Dewland & Minihan, 2011).

This analysis further establishes the Pareto principle: 85,87% of the citations were found in the upper 20% of the periodical list. As a relatively low number of periodicals (20%) generates the most citations (85%), it's possible to conclude that, if an academic library strives to acquire the right periodicals and makes an optimal selection, it can provide good coverage of relevant information for its patrons, even if the quantity of said periodicals is low. In other words, a small but optimal selection of periodicals can cover the most of an institution's information needs.

References

Dewland, J. & Minihan, J. (2011). Collective serials analysis : the relevance of a journal in supporting teaching and research. *Technical Services Quarterly*, 28, 265-282.

Gomez, M. (2002). A bibliometric study to manage a journal collection in an astronomical library: some results. *Library and information services in astronomy*, 4, 214-222.

Krajna, T. & Markulin, H. (2011). Nabava knjižnične građe u visokoškolskim knjižnicama. *Vjesnik bibliotekara Hrvatske*, 54, 21-42.

Krznar, I. (2011). Identifikacija razdoblja recesija i ekspanzija u Hrvatskoj. *Istraživanja (Hrvatska narodna banka)*, 32, 1-17.

Is Paper Uncitedness a Function of the Alphabet?

Clément Arsenault[1] and Vincent Larivière[2]

[1] *clement.arsenault@umontreal.ca*
École de bibliothéconomie et des sciences de l'information (EBSI), Université de Montréal, Montréal (Qc) (Canada)

[2] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information (EBSI), Université de Montréal, Montréal (Qc) (Canada), and
Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, Montréal (Qc) (Canada).

Introduction

Citation counts are well-established measures of researchers' scientific impact. One would assume that external factors, such as someone's name, over which an individual has little control over, does not influence such indicators. Yet, reference lists and—to a lesser extent—search results from online databases, are often presented in alphabetical order sorted by first author surname. A large number of scientific journals use parenthetical referencing styles (a.k.a. Harvard referencing style) in which partial parenthetical citations (such as author+date or author+title) are embedded in the text, accompanied by an alphabetized list of complete citations at the end. These lists may be consulted to locate a specific item (known-item search) but are also used in a scanning mode, usually from top (A) to bottom (Z), to identify papers that would potentially provide answers to a question or reinforce an argument.

In marketing and advertising research it is well recognized that product positioning influences choice and selection and that usually "first is best", i.e., that items presented first usually have a better chance of being selected (Carney & Banaji, 2012). Such a phenomenon has also been observed by Haque and Ginsparg (2009, p. 2215) who measured a significant correlation between article position in the arXiv repository and citation impact, due the "visibility" effect that "can drive early readership, with consequent early citation potentially initiating a feedback loop to more readership and citation." Order of presentation (or scanning order) is also central to Cooper's utility theory (1971) since items consulted earlier will find a better chance of being useful to a searcher.

Taking these elements into account, authors with a surname whose initial letter arrives early in the alphabet get more visibility, a situation that is further compounded by the fact that in multi-authored papers, authorship order is sometimes determined by alphabetical rank. This practice is even fairly common in some fields such as economics and finance, mathematics, high-energy physics, marketing, political science, international relations and law (Frandsen & Nicolaisen, 2010, p. 615; Levitt & Thelwall, 2012, p. 725; Waltman, 2012, p. 701). In the field of economics where authorship order is almost always determined alphabetically, research has shown that economists with early surnames (i.e., with initial letters that occur early in the alphabet) publish more articles (van Praag & van Praag, 2008), are more likely to get employment at high standard research departments (Efthyvoulou, 2008) and receive more tenure at top economic departments (Einav & Yariv, 2006), since "the order of authorship, rather than contributorship, is commonly used to assess the prestige that an author incurs from a published research study" (Chambers, Boath, & Chambers, 2001, p. 1461).

Literature Review

Citation likelihood based on author's surname position in the alphabet has also been the subject of some recent studies. McCarl (1993) found that authors receive approximately 0.5% less first author citations per letter the latter their names are in the alphabet. Laband and Tollison (2006) showed that "alphabetized co-authored papers with two authors are more highly cited than non-alphabetized co-authored papers" in both economics and agricultural economics. In a large-scale study Huang (2015, p. 780) revealed that "papers with first authors whose surname initials appear earlier in the alphabet get more citations [and that this effect] is significantly stronger in those fields with longer reference lists."

This later observation reinforces the idea that the browsing effect is to the advantage of papers listed towards the top of alphabetized reference lists since readers are more likely to run out of patience before they get to the end of the list. To corroborate these findings, our study will look at the reverse effect, namely the greater invisibility of papers appearing at the end of reference lists by measuring the uncitedness rates of papers as correlated to the first author's position in the alphabet.

Data and Methodology

The data set used in this study was obtained from the Web of Science databases and consists of all the scientific papers published between the years 2000 and 2013, totalling 15,056,841 source items. Papers are assigned to one of the fourteen disciplines of the National Science Foundation (NSF) classification. Field-normalized citations rates for each paper were calculated, and grouped by the first letter of the surname of the first author, which means that each paper was counted only once in the dataset.

Results and Discussion

Preliminary analysis reveals that, in most of the fourteen NSF disciplines, uncitedness rates tend to increase with the progression of the first author's last name in the alphabet indicating that papers with a first author whose last name starts with a letter that occurs later in the alphabet might be less visible. Correlation coefficients are the strongest in the disciplines of Mathematics and Physics (figure 1) indicating that the practice in these disciplines to list co-authors on the basis of author's position in the alphabet seems to exacerbate this problem.

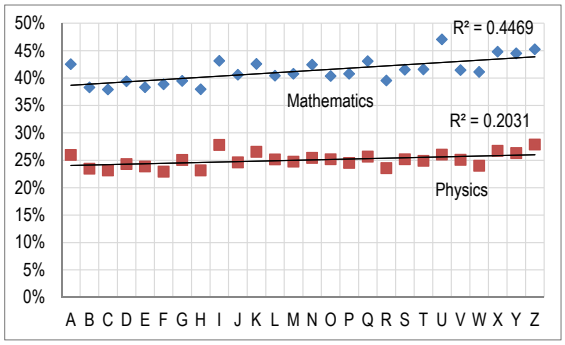


Figure 1. Uncitedness rates of Mathematics and Physics papers by initial letter of first author's surname.

Further analysis at the level of specialty of the NSF classification will validate whether such effects are observable in other fields (such as Economics & Finance) where the tradition of listing co-authors alphabetically is highly prevalent, as well as the potential effect of researchers from specific countries whose surnames are more likely to start with a letter that appear towards the end of the alphabet.

On the whole, these results show that papers whose first author bears a surname that is at the end of the alphabet are at a disadvantage in terms of citation rates, a finding that is likely a consequence of the current structure of reference lists and of search results from online databases.

In a more detailed analysis, confounding factors such as the higher prevalence of names beginning with some letters and the concentration of names from certain regions will be considered.

References

Carney, D. R., & Banaji, M. R. (2012). First is best. *PLoS ONE, 7*(6), e35088. doi:10.1371/journal.pone.0035088

Chambers, R., Boath, E., & Chambers, S. (2001). The A to Z of authorship: Analysis of influence of initial letter of surname on order of authorship. *BMJ,* 323(22–29 Dec.), 1460–1461. doi:10.1136/bmj.323.7327.1460

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval,* 7(1), 19–37. doi:10.1016/0020-0271(71)90024-6

Efthyvoulou, G. (2008). Alphabet economics: The link between names and reputation. *The Journal of Socio-Economics,* 37(3), 1266–1285. doi:10.1016/j.socec.2007.12.005

Einav, L. & Yariv, L. (2006). What's in a surname?: The effects of surname initials on academic success. *Journal of Economic Perspectives, 20*(1), 175–188. doi:10.1257/089533006776526085

Frandsen, T. F. & Nicolaisen, J. (2010). What is in a name?: Credit assignment practices in different disciplines. *Journal of Informetrics,* 4(4), 608–617. doi:10.1016/j.joi.2010.06.010

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. doi:10.1002/asi.21166

Hart, R. L. (2000). Co-authorship in the academic library literature: A survey of attitudes and behaviors. *Journal of Academic Librarianship, 26*(5), 339–345. doi:10.1016/S0099-1333(00)00140-3

Huang, W. (2015). Do ABCs get more citations than XYZs? *Economic Inquiry,* 53(1), 773–789. doi:10.1111/ecin.12125

Levitt, J. M. & Thelwall, M. (2012). Alphabetical co-authorship in the Social Sciences. *Proceedings of the 17th International Conference on Science and Technology Indicators Conference* (Montreal, Canada). http://2012.sticonference.org/Proceedings/vol2/Levitt_Alphabetical_523.pdf

McCarl, B. A. (1993). Citations and individuals: First authorship across the alphabet. *Review of Agricultural Economics,* 15(2), 307–312. doi:10.2307/1349450

van Praag, C. M. & van Praag, B. M. S. (2008). The benefits of being economics professor A (rather than Z). *Economica, 75*(300), 782–796. doi:10.1111/j.1468-0335.2007.00653.x

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics, 6*(4), 700–711. doi:10.1016/j.joi.2012.07.008

Relative productivity drivers of economists: A probit/logit approach for six European countries

Stelios Katranidis[1] and Theodore Panagiotidis[2]

[1]*katranid@uom.gr,* [2] *tpanag@uom.gr*
Department of Economics, University of Macedonia, Greece

Introduction

Economists talk frequently about productivity. They refer to productivity of the economy in most of the cases. This paper examines the productivity of the economists themselves. There has been an increase interest on the drivers of productivity among scientists and economists in particular. Among them the country of the PhD studies, gender, north vs south and inbreeding (at the departmental or national level) has been suggested. Most of the studies employ absolute measures of productivity. We deviate from this tradition and examine relative productivity. Relative is defined in terms of deviations from the countries mean productivity. The latter is measured as papers per faculty (per year) and citations per faculty (per year). We employ a dataset that consists of 1431 economists from six countries. The north is represented by Belgium, Denmark and Germany whereas the south by Greece, Italy and Portugal[1].

Literature Review

The literature on the factors that affect an economists' productivity has expanded in the last decade. Çokgezen (2006) examined the productivity differentials for economists based in Turkey between private and state universities. Ben-David (2010) considered the case of Israel and how high and low rank academic positions vary with productivity. Katranidis et al (2012) examined differences in academic performance taken into account the country where the doctoral studies have been completed for Portugal and Greece respectively. Using survey data, Kalaitzidakis et al. (2004) provided evidence that European economics departments with links with institutions in North-America are more productive in terms of research output. More recently, Bauwens et al. (2011) stressed that English proficiency is an important factor for higher productivity amongst economists.

Data

Our dataset stems from the Scopus database and from the websites of the corresponding Departments. The data were collected for 1431 economists that were employed in Belgium (125 economists), Denmark (82), Germany (543), Greece (82), Italy (504) and Portugal (95). The number of observations (economists) for each country reflects 25% of the RePec registered economists in each country. The characteristics considered for each economist includes number of papers, number of citations, whether their PhD studies took place in the US or they country they work (inbreeding at the national level), gender and the real research age (number of years since obtaining their PhD).

This paper is trying to advance the relative literature in two ways: We use relative measures of productivity on comparing economists' productivity in more than one country instead of absolute measures of productivity, i.e. papers per faculty per year or citations per faculty per year. More specifically, relative productive is calculated as the difference between a researcher's and the country's average productivity. Researchers get a value of 1 if they exhibit a positive difference in productivity compared to the country's average and 0 otherwise. In this sense, the dependent variable is binary and thus probit and logit models are employed to investigate the drivers of relative productivity among economists in six EU countries. This also represents advancement in the literature since OLS regressions were used to model average response to specific characteristics.

The second is the academic inbreeding that refers to the practice where Universities hire its PhD graduates. The evidence demonstrates that this affects negatively the scholarly output (Inanc & Tuncer, 2011). In this study we will consider inbreeding at a higher level i.e. at the national level. Scientific human capital would, in this respect, reflect the quality of human and social capital in the country. Goudard and Lubrano (2013) introduced a model where social capital complements scientific human capital. We will examine whether hiring economists that hold PhD from the same country affects relative productivity. We will refer to this characteristic as national inbreeding.

Methodology

As noted in the previous section, the goal of this study is to investigate the drivers of relative productivity. The dependent variable takes the value of 0 if the productivity of the researcher is below the country's average and 1 otherwise.

[1] This research is implemented through the Operational Program "Education and Lifelong Learning" and is co-financed by the EU (European Social Fund) and Greek national funds.

A linear probability model (LPM) is used in the form of:

$$P_i=p(y_i=1)=\beta_1+\beta_2(Belgium*PhD^{US})+\beta_3(Denmark*PhD^{US})+\beta_4(Germany*PhD^{US})+\beta_5(Greece*PhD^{US})+\beta_6(Italy*PhD^{US})+\beta_7(Portugal*PhD^{US})+\beta_8(Belgium*PhD^{Belgium})+\beta_9(Denmark*PhD^{Denmark})+\beta_{10}(Germany*PhD^{Germany})+\beta_{11}(Greece*PhD^{Greece})+\beta_{12}(Italy*PhD^{Italy})+\beta_{13}(Portugal*PhD^{Portugal})+\beta_{14}(Belgium*Female)+\beta_{15}(Denmark*Female)+\beta_{16}(Germany*Female)+\beta_{17}(Greece*Female)+\beta_{18}(Italy*PhD^{Italy})+\beta_{19}(Portugal*Female) \quad (1)$$

where y_i is 1 if the difference between papers (citations) per faculty per year and the country's average is positive and 0 otherwise, *Belgium* ,…, *Portugal* are dummy variables denoting the country a research is based, PhD^{US} and $PhD^{Belgium}$ are dummy variables taking the value of 1 if the researcher has completed her/his PhD studies in the US and Belgium, while *female* is a gender dummy taking the value of 1 if the research is female.

Results

Equation 1 is estimated for two relative measures of productivity. We consider above country average papers per faculty per year and citations per faculty per year. In the probit model, the factors that affect in a negative and significant way relative productivity (at the 90% significance level) are: (i) having a US PhD and work in Germany, (ii) a German PhD and work in Germany (national level inbreeding), (iii) a Greek PhD and work in Greece, (iv) Italian PhD and work in Italy, (v) Portuguese PhD and work in Portugal and (vi) being female in Germany, Denmark and Italy.

In the logistic model these factors are (negative and significant at the 90%): (i) having a US PhD and work in Germany or in Denmark, (ii) a German PhD and work in Germany (national level inbreeding), (iv) a Danish PhD and work in Denmark, (v) an Italian PhD and work in Italy and (vi) being female in Germany, Greece, Italy and Portugal.

The only variable that affects citations per faculty per year in a positive way is holding a US PhD and working in Italy. Variables that affect in a negative and significant way (90%) are: (i) a German PhD and work in Germany, (ii) a Greek PhD and work in Greece, (iii) an Italian PhD and work in Italy, (iv) a Portuguese PhD and work in Portugal and (vi) being female in Belgium, Germany, Denmark and Italy. The results are similar in the case of the logistic function: (i) a PhD from Belgium and work there, (ii) German PhD and work in Germany, (iii) a Danish PhD and work in Denmark, and (iv) being female in Germany, Greece, Italy and Portugal.

Overall the highest marginal effects are observed for the above average papers per faculty per year: (i) being female in Denmark (-0.502), (ii) holding a Greek PhD in Greece (-0.410) and (iii) holding a Portuguese PhD in Portugal (-0.331) (in the probit model). For the logit: (i) holding a Danish PhD in Denmark (-0.585), (ii) being female in Greece (-0.423) and (iii) holding a US PhD in Denmark. For the citations (probit), the largest marginal effects are identified for being female in Belgium and Denmark (-0.311 and -0.252 respectively). In the logit, inbreeding in Belgium and Denmark (-0.337 and -0.257).

Conclusions

This study examines the drivers of relative productivity among 1431 economists from six European countries. Scopus database was the data source for economists based in three northern EU countries (Belgium, Denmark and Germany) and three southern (Greece, Italy and Portugal). We identify the drivers of relative productivity in terms of deviations from the national average in papers per faculty per year and citations per faculty per year. We employ probit and logit models given that the dependent variable is binary (above the national average 1, below 0). For papers the most important variables that were affecting relative productivity in a negative manner were gender in Denmark and national inbreeding in Greece and Portugal; while for the citations, gender and national inbreeding in Belgium.

References

Bauwens, L., Mion, G. & Thisse, J. F., (2011). The Resistible Decline of European Science, *Recherches économiques de Louvain*, **77**, 4, 5-31.

Ben-David, D. (2010). Ranking Israel's economists, *Scientometrics, 82,* 351-364.

Çokgezen, M. (2006). Publication performance of economists and economics departments in Turkey (1999-2003). *Bulletin of Economic Research, 58,* 253-265.

Inanc, O. & Tuncer, O. (2011). The effect of academic inbreeding on scientific effectiveness, *Scientometrics*, 88, 885-898.

Goudard, M., & Lubrano, M. (2013) Human Capital, Social Capital and Scientific Research in Europe: An Application of Linear Hierarchical Models. The Manchester School. 81(6): 876-903.

Kalaitzidakis, P., Mamuneas T.P. Savvides, A. & Stengos, T. (2004). Research spillovers among European and North-American economics departments, *Economics of Education Review*, 23, 191-202.

Katranidis, S., Panagiotidis, T., & Zontanos, C. (2012). An evaluation of the Greek universities' economics departments. *Bulletin of Economic Research.* Advance online publication.

Do First-Articles in a Journal Issue Get More Cited?

Tian Ruiqiang, Yao Changqing, Pan Yuntao, Wu Yishan, Su Cheng and Yuan Junpeng

trq2011@sina.com

Institute of Scientific and Technical Information of China, 15 Fuxing Road, 100038, Beijing (China)

Introduction

As the advice of peers on the quality of a submitted paper prior to publication, peer review can be regarded as the pre-publication evaluation. Bibliographic citations of scientific papers used as indicators of the visibility, impact, and quality of scientific publications, could be regarded as the post-publication evaluation.

Intentionally or not, journal editors often put the accepted manuscript with nice comments by peer reviewers at the top of all papers in an issue. The First-Articles of journal issues are generally regarded with higher importance, intense creativity or superior quality through peer review process. Judge A, Cable M, Colbert E (2007) deemed that journal editors placed the best paper in the "pole position", and they confirmed this anecdotal evidence further in their study. Specifically, 75% of 16 journals indicated that quality played some primary role in selection of the first articles. Wang (2015) also admitted that journals would choose the very best paper of an issue on the cover, "a paper that in 20 year's time might win a Nobel Prize", according to the opinion of Stang, the EIC of *Journal of the American Chemical society* (Ritter 2006).

Since there are evidences that peer reviewers can successfully discriminate between manuscripts that have a greater chance to be cited in future. Further, in this sense, we made a hypothesis that the best articles selected by peer reviews—usually the First-Articles, will be superior in receiving higher citations after publication. In this paper we will illustrate how peer review and the performance of journal papers measured by bibliometric indicators could concordance with each other. In particular, we examined whether there were obvious citation differences between First-Articles and non-First-Articles published in the same issue of a journal.

Data and Methodology

Twins data, a sampling method used in labour economics, reaches "other things being equal" to a certain extent. Twin studies are often employed to evaluate the inheritance of a trait by dissecting the genetic and environmental contributions to the trait.

In this study, we regard the First-Articles and non-First-Articles in the same issue as twins. They were published in the same time and have similar disciplinary backgrounds.

We select First-Articles from Scopus and Web of Science (WoS). First, we choose journals which publish research articles on their first pages rather than other types of documents, such as editorial, letters et al. And we find that most mathematic journals satisfy this criterion well. Thus we select top100 mathematical journals by their Impact Factors from JCR 2013. Then, we acquire twins data by retrieving articles published in those 100 journals between1995-1999 in Scopus and WoS. As a result, we obtained 19,411 articles in 62 journals in WoS on December 25, 2014 and 18,524 articles in 67 journals in Scopus on January 13, 2015 respectively. The difference of journal numbers is resulted that some journals were not indexed as early as 1995-1999 while included in 2013 JCR. And we identified 2050 out of WoS and 2229 out of Scopus First-Articles, excluding those articles published on supplementary issues, special issues. Table 1 provides an overview of the samples.

Table 1. Descriptive statistics of the samples

	Scopus		WoS	
	Fr	**Non -Fr**	**Fr**	**Non-Fr**
Articles	2229	16295	2050	17361
		67 journals		62 journals

Results

First-Articles receive higher CPP&CTC

The indicator CPP (the average number of citations received per article) and CTC (the contributions to total journal citations) were taken as the criterion to assess the citation position of First-Articles and non-First-Articles in their own disciplinary citation environment. It revealed obvious differences in citations between the First-Articles and non-First-Articles. As shown in Table 2, in WoS, the First-Articles received higher average citation (AC) (16.56) since publishing, while the non-First-Articles got 13.69. In Scopus, the First-Articles accumulated 17.00 of AC, those non-First-Articles of 14.00. In WoS, the First-Articles contribute 12.5% to total citations (TC) of the journal when their proportions in total documents remain only 10.6%. Though the non-First-Articles got 89.4% share of total documents, their contributions of TC remain 87.5%. And the case is almost the same in Scopus: the First-Articles contribute 14.2% to TC when the proportions of articles remain only 12%. Though

the non-First-Articles got 88% of articles, their contributions of TC remain 85.8%.

Based on ANOVA test, we found significant difference between TC of 2050 First-Articles and 17361 non-First-Articles in WoS at the 0.05 significance level. Similarly, in Scopus there is also significantly different between 2229 First-Articles and 16295 non-First-Articles. Specifically, TC of First-Articles is significantly higher than non-First-Articles. From WoS, the non-First-Articles received mean TC of 13.69. While under same circumstance, First-Articles received clearly higher mean TC of 16.56. In terms of Scopus, the non-First-Articles reached at 14.00 of mean TC. And this time, the similar backgrounds, First-Articles performed more excellent, reaching notably higher mean TC of 17.00. Therefore, First-Articles are higher impact than non-First-Articles both in WoS and Scopus.

Table 2. TC difference in ANOVA test

	WoS			Scopus		
	Num	**Mean**	**SD**	**Num**	**Mean**	**SD**
Fr	2050	16.56	30.13	2229	17.00	27.08
N-Fr	17361	13.69	24.03	16295	14.00	24.51
P			0.000			0.000

Nearly 24% First-Articles are most highly cited, while non-cited articles account for only 10%

It shows 22.6% First-Articles in average are also the papers with highest TC among papers published in the same journal issues in WoS. And the proportion keeps stable in the observe window. In Scopus, the percentage of the most highly cited papers in First-Articles goes to almost 25%. In 1997, it even reached a peak of 27%.

Table3. Citation difference of First-Articles and non-First-Articles in WoS& Scopus

	WoS	**Scopus**
CPP-Fr	16.56	17.00
CPP-Non-Fr	13.69	14.00
CTC-Fr	0.125	0.142
CTC-NFr	0.875	0.858
Num highC	463	552
Num zeroC	228	179
highC %	0.226	0.248
ZeroC%	0.111	0.080
ZeroC Total %	0.124	0.107

As shown in Table 3, the percentage of non-cited papers in 62 mathematics journals in WoS is 12.4%. While it is much lower for First-Articles, the uncitedness rate drops to 11.1% in a whole through a period of nearly two decades. As for Scopus database, the share of papers never cited in 67 journals in mathematics decline to10.7%. In addition, the proportion of uncitedness for First-Articles stays to 8.0% on average.

Conclusion

To verify the hypothesis that the best articles selected by peer reviewers, usually the First-Articles, will be superior in receiving higher citations after publication compared with non-First-Articles published in the same journal issue, we first obtained twins data of First-Articles and non-First-Articles by retrieving articles published in top 100 (in terms of JCR 2013 JIF) mathematic journals in Scopus and WoS. Then we employed indicators CPP, CTC and TC, based on which we applied ANOVA to contrast citation bias of First-Articles and non-First-Articles in both Scopus and WoS. Results showed that there existed significant difference between First-Articles and non-First-Articles in receiving citations after publication. On the basis of these empirical grounds, we suggested that the First-Articles are biased in citations compared with non-First-Articles. We also found that it revealed a higher proportion of First-Articles to be most highly cited and comparatively lower proportion to be uncited. Furthermore, it presented a good consistency in conclusion in Scopus and WoS.

The results suggest that the peer reviewer's best recommendation go accordance with highest bibliometric indicator performance. Deliberately or not, papers received best recommendations in pre-publication evaluation process often are arranged as the First-Articles in a journal issue. The First-Articles are generally regarded as ones of high importance intense creativity or superior quality judged by peer reviewers; therefore they are expected to have a greater chance to get highly cited in the future. In fact, such understanding is supported by our analysis in this paper. After publication, those First-Articles are more likely to receive higher citations. Accordingly, peer reviewers' best recommendations and the excellent performance of journal papers measured by bibliometric indicators concordance with each other in the case of First-Articles.

Acknowledgments

We acknowledge the National Natural Science Foundation of China (NSFC Grant No.71373252) for financial support.

References

Judge A, Cable M, Colbert E. (2007). What causes a management article to be cited—article, author, or journal? *Academy of Management Journal*, 50(3), 491-506.

Wang, X., Liu, C., & Mao, W. (2014). Does a paper being featured on the cover of a journal guarantee more attention and greater impact? *Scientometrics*, 102(2), 1815-1821.

Ritter, S. K. (2006). Making The Cover. *Chemical & Engineering News*, 84(45), 24-27.

ProQuest Dissertation Analysis

Kishor Patel,[1] Sergio Govoni,[1] Ashwini Athavale,[1] Robert P. Light,[2] Katy Börner[2]

[1]*Kishor.Patel@proquest.com, Sergio.Govoni@proquest.com, Ashwini.Athavale@proquest.com*
ProQuest LLC, 7500 Old Georgetown Road, Suite 1400, Bethesda, MD 20814 (USA)

[2] *katy@indiana.edu, lightr@indiana.edu*
CNS, SOIC, Indiana University, 1320 E. Tenth Street, Bloomington, IN 47405 (USA)

Introduction

Productivity measurement has become a major issue for university leaders. Federal and state governments support teaching and research with significant investments. When university leaders are seeking new funding, it is not uncommon that they need to justify their request with productivity measurement metrics and equally important research output consumption metrics. However, it is often very difficult for university leaders to generate these metrics as they lack access to relevant data and tools to analyse and visualize large amounts of data.

Interested to address the diverse needs of university leaders, ProQuest and Indiana University analysed the ProQuest Dissertation & Theses Global (PQDT Global) database, an extensive and trusted collection of 3.8 million graduate study dissertations with 1.7 million full text records and editorially assigned metadata created by subject area experts. The database offers comprehensive North American and significant international coverage. Worldwide access to the database is logged at the dissertation level by ProQuest. Usage data mining is important for understanding user behaviour (Srivastava et al., 2000). The ProQuest Dissertations Dashboard released in 2014 provides easy access to dissertations, metadata, and usage data. It is available for free to leaders of any university that shares dissertation data with ProQuest.

ProQuest Data Analysis and Visualization

Analyses were conducted and results visualized to answer questions that seemed of particular interest to university leaders and those seeking to assess the performance of a school as a whole.

Study 1: How much attention are my school's dissertations getting?

A school's ability to generate interest in their students' dissertations may not only reflect the reputation of the school, but have long-term effects on those students' marketability and also in attracting future generations of students to join the school.

Figure 1 plots the production and access data for computer science dissertations for a selected institution given in red and labelled 'Subject University' and two groups of peer institutions rendered in green and blue. Other institutions that have published computer science dissertations are given in grey. The three institutions in the top-right corner of the plot—publishing many theses that attract many views—include both well-regarded private research institutions as well as for-profit colleges with practically open admissions. This implies that while thesis production and usage are important, they should not be used as a sole indicator for the quality of a program.

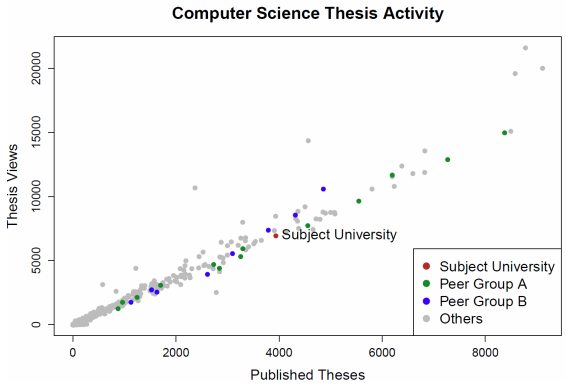


Figure 1. Comparing Subject-Area Specific Thesis Access Activity with Peer Groups.

Study 2: How can I quickly compare the number of dissertations and associated download activity for a large number of universities?

Given all dissertations or dissertations in a certain subject area, university leaders might like to understand the "market share" of an institution within a comparison or peer group.

In Figure 2, two peer groups of institutions are compared. Each institution is represented by a rectangle. Each rectangle is sized based on the total corpus of computer science dissertations available in the ProQuest dataset for that institution.

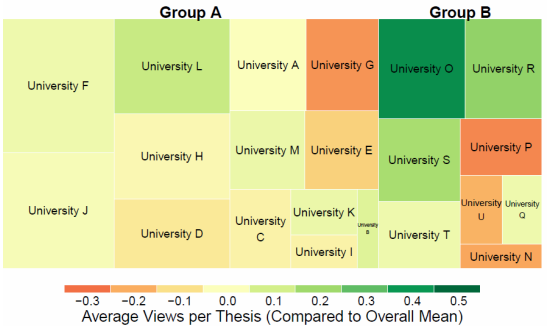


Figure 2. Treemap Comparing Thesis Production and Usage in Computer Science.

Colours tell how frequently the average dissertation at that institution is accessed in comparison to the group average. Computer science dissertations written at Universities L, O, and R are accessed more frequently than the group average, while those published at Universities G or P are accessed less.

Study 3: How is dissertation information flowing in and out of my university?

Universities are both producers and consumers of information (Mazloumian et al., 2013). Administrators are interested to understand which dissertations from which universities are used at their own institution but they also want to know who is accessing their own institution's dissertations. Plus, they might need to compare this in-flow and out-flow of information with the flows calculated for other universities.

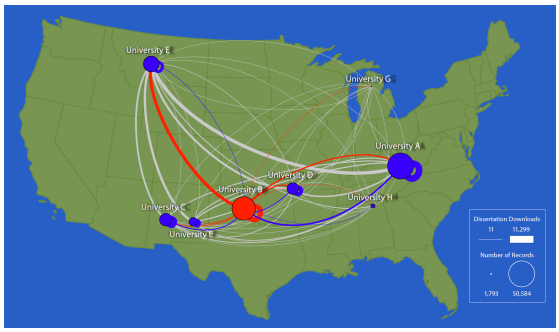


Figure 3. Information Flows within Peer Group

The example in Figure 3 looks at information flow between a group of peer schools. One institution, labelled University B, is highlighted. Red edges depict information flowing out of that institution, while blue flows show information flowing into that institution. The thicker the line, the greater is the number of dissertations. (Information always flows clockwise on the curved lines).

Future Directions

Currently, ProQuest dissertation data is not linked to publication, funding or other data. However, there is much interest in being able to study career trajectories in a more comprehensive manner (Ni & Sugimoto, 2012; Ostriker, Kuh, & Voytuk, 2011) and to examine the reputation and funding of dissertation advisors and the success (in terms of funding and publication records) of their advisees in more detail. Citation counts for dissertations, user ratings and altmetrics data, e.g., social media data, are valuable indicators of impact that we would like to explore. We also think that productivity and usage datasets can be leveraged to study the emergence of new disciplines and cross-disciplinary subject areas (Sugimoto, Li, Russell, Finlay, & Ding, 2011).

Acknowledgments

This work was partially funded by the National Institutes of Health under awards P01AG039347, U01GM098959, and U01CA198934. The authors would like to thank and acknowledge the assistance of Samuel Mills in preparing graphics for this text, Mike Gallant for information technology support as well as the ProQuest dissertations product management, development, and technical teams for their support during this research work.

References

Mazloumian, A., Helbing, D., Lozano, S., Light, R. P., & Börner, K. (2013). Global multi-level analysis of the 'Scientific Food Web'. *Scientific reports*, 3.

Ni, C., & Sugimoto, C.R. (2012). Using doctoral dissertations for a new understanding of disciplinarity and interdisciplinarity. *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*. Baltimore, MD. October 26-30, 2012: ASIST.

Ostriker, J., Kuh, C., & J. Voytuk (Eds.), (2011) A Data-Based Assessment of Research-Doctorate Programs in the United States. Retrieved from: http://www.nap.edu/rdp/

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.* 11, 1 (pp. 92-99). Retrieved from http://doi.acm.org/10.1145/102377.115768

Srivastava, J., Cooley, R., Deshpande, M, & Tan, P., (2000). Web usage mining: discovery and applications of usage patterns from Web data. *ACM SIGKDD Explorations Newsletter*, 1(2), 12-23. http://doi.acm.org/10.1145/846183.846188

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American library and information science dissertations using Latent Dirichlet Allocation. *Journal of the American Society for Information Science and Technology, 62* (1), 185-204. http://onlinelibrary.wiley.com/doi/10.1002/asi.21435/abstract

ISSI

2015

INDICATORS

An Alternative to Field-normalization in the Aggregation of Heterogeneous Scientific Fields

Antonio Perianes-Rodriguez[1] and Javier Ruiz-Castillo[2]

[1] *antonio.perianes@uc3m.es*
Universidad Carlos III, Department of Library and Information Science, SCImago Research Group, C/ Madrid, 128, 28903 Getafe, Madrid (Spain)

[2] *jrc@eco.uc3m.es*
Universidad Carlos III, Departamento de Economía, C/ Madrid, 126, 28903 Getafe, Madrid (Spain)

Abstract
A possible solution to the problem of aggregating heterogeneous fields in the all-sciences case relies on the normalization of the raw citations received by all publications. In this paper, we study an alternative solution that does not require any citation normalization. Provided one uses size- and scale-independent indicators, the citation impact of any research unit can be calculated as the average (weighted by the publication output) of the citation impact that the unit achieves in all fields. The two alternatives are confronted when the research output of the 500 universities in the 2013 edition of the CWTS Leiden Ranking is evaluated using two citation impact indicators with very different properties. We use a large Web of Science dataset consisting of 3.6 million articles published in the 2005-2008 period, and a classification system distinguishing between 5,119 clusters. The main two findings are as follows. Firstly, differences in production and citation practices between the 3,332 clusters with more than 250 publications account for 22.5% of the overall citation inequality. After the standard field-normalization procedure where cluster mean citations are used as normalization factors, this figure is reduced to 4.3%. Secondly, the differences between the university rankings according to the two solutions for the all-sciences aggregation problem are of a small order of magnitude for both citation impact indicators.

Conference Topic
Indicators; Citation and co-citation analysis

Introduction
As is well known, the comparison of the citation impact of research units is plagued with obstacles of all sorts. For our purposes in this paper, it is useful to distinguish between the following three basic difficulties. (i) How can we compare the citation distributions of research units of different sizes even if they work in the same homogeneous scientific field? For example, how can we compare the output of the large Economics department at Harvard University with the output of the relatively small Economics department at Johns Hopkins? The next two difficulties have to do with the heterogeneity of scientific fields: the well-known differences in production and citation practices makes it impossible to directly compare the raw citations received by articles belonging to different fields. Given a classification system, that is, a rule for assigning any set of articles to a number of scientific fields, field heterogeneity presents the following classic hindrances in the evaluation of research units' performance. (ii) How can we compare the citation impact of two research units working in different fields? For example, how can we compare the citation impact of MIT in Organic Chemistry with the citation impact of Oxford University in Statistics and Probability? Finally, (iii) how can we compare the citation impact of two research units taking into account their

output in all fields? For example, how can we compare the citation impact of MIT and Oxford University in what we call the *all-sciences* case?

As is well known, the solution to the first two problems requires size- and scale-independent citation impact indicators. We will refer to indicators with these two properties as *admissible* indicators. Given an admissible indicator, in this paper we are concerned with the two types of solutions that the third problem admits. Firstly, the problem can be solved in two steps. One first uses some sort of normalization procedure to make the citations of articles in all fields at least approximately comparable. Then, one applies the citation indicator to each unit's normalized citation distribution. Secondly, consider the Top 10% indicator used in the construction of the influential Leiden and *SCImago* rankings. In the Leiden Ranking this indicator is defined as "*The proportion of publications of a university that, compared with other similar publications, belong to the top 10% most frequently cited…Publications are considered similar if they were published in the same field and the same publication and if they have the same document type*" (Waltman et al., 2012a). A similar definition is applied in the SCImago ranking (Bornmann et al., 2012) Note that this way of computing this particular indicator in the all-sciences case does not require any kind of prior citation normalization. For our purposes, it is useful to view this procedure as the average (weighted by the publication output) of the unit's Top 10% performance in each field. We note that this important precedent can be extended to *any* admissible indicator. Thus, given a classification system and an admissible citation indicator, we can compute the citation impact of a research unit in the all-sciences case as the appropriate weighted average of the unit's citation impact in each field. Independently of the conceptual interest of this proposal, we must compare the consequences of adopting it versus the possibility of following a normalization procedure.

Intuitively, the better the performance of the normalization procedure in eliminating the comparability difficulties across fields, the smaller will be the differences between the two approaches. Consider, for example, what we call the standard field-normalization procedure in which the normalized citations of articles in any field are equal to the articles' original raw citations divided by the field mean citation. Under the universality condition, that is, if field citation distributions were identical except for a scale factor, then the standard field-normalization procedure would completely eliminate all comparability difficulties. However, the universality condition, once claimed to be the case (Radicchi et al., 2008), is not usually satisfied in practice: even appropriately normalized, field citation distributions are seen to be significantly different from a statistical point of view (Albarrán et al., 2011a; and Waltman et al., 2012a). Therefore, at best, normalization procedures provide an approximate solution to the original comparability problem.

Using a measuring framework introduced in Crespo et al. (2013), recent research has established that different normalization procedures perform quite well in eliminating most of the effect in overall citation inequality that can be attributed to differences in production and citation practices between fields. This is the case for large Web of Science (WoS hereafter) datasets, classification systems at different aggregation levels, and different citation windows (Crespo et al., 2013, 2014; Li et al., 2013; Waltman & Van Eck, 2013; Ruiz-Castillo, 2014). The reason for the good performance of target (or cited-side) normalization procedures is that field citation distributions, although not universal, are extremely similar (Glänzel, 2007; Radicchi et al., 2008; Albarrán & Ruiz-Castillo, 2011; Albarrán et al., 2012; Waltman et al., 2012a; Radicci & Castellano, 2012; Li et al., 2013). It should be noted that this research on target normalization procedures uses WoS classification systems distinguishing at most between 235 sub-fields.

In principle, given the good performance of normalization procedures, we expect that the differences between the two approaches would be of a small order of magnitude.

Nevertheless, this is an empirical question that has never been investigated before. To confront this question, in this paper we conduct the following exercise.

- Ruiz-Castillo & Waltman (2015) apply the publication-level algorithmic methodology introduced by Waltman and Van Eck (2012) to a WoS hereafter dataset consisting of 9.4 million publications from the 2003-2012 period. This is done along a sequence of twelve independent classification systems in each of which the same set of publications is assigned to an increasing number of clusters. In this paper, we use the classification system recommended in Ruiz-Castillo and Waltman (2015), consisting of 5,119 clusters, of which 4,161 are referred to as significant clusters because they have more than 100 publications over this period. For the evaluation of research units' citation impact, we focus on the 3.6 million publications in the 2005-2008 period, and the citations they receive during a five-year citation window for each year in that period. It should be noted that, using the size- and scale-independent technique known as Characteristic Scores and Scales, Ruiz-Castillo and Waltman (2015) show that, as in previous research, significant clusters are highly skewed and similarly distributed.
- Our research units are the 500 universities in the 2013 edition of the CWTS Leiden Ranking (Waltman et al., 2012b). We analyze the approximately 2.4 million articles – about 67% of the total– for which at least one author belongs to one of these universities. We use a fractional counting approach to solve the problem –present in all classification systems– of the assignment of responsibility for publications with several co-authors working in different institutions. The total number of articles corresponding to the 500 universities is approximately 1.9 million articles –about 50% of the total.
- We evaluate the citation impact of each university using two size- and scale-independent indicators. Firstly, we use the Top 10% indicator, already mentioned. Secondly, one characteristic of this indicator is that it is not monotonic in the sense that it is invariant to any additional citation that a high-impact article might receive. Consequently, we believe that it is interesting to use a second indicator possessing this property. In particular, we select a member of the Foster, Greer, and Thorbecke (FGT hereafter) family, introduced in Albarrán et al. (2011b). We apply this indicator to the set of high-impact articles mentioned before. As will be seen below, the fact that both of our indicators are additively decomposable facilitates the comparability of the two solutions to the all-sciences aggregation problem.
- Using Crespo et al.'s (2013) measurement framework, Li et al. (2013) indicate that the best alternative among a wide set of target normalization procedures is the two-parameter system developed in Radicci and Castellano (2012). However, recent results indicate that the standard, one-parameter field-normalization procedure exhibits a good performance in reducing the effects on overall citation inequality attributed to differences in production and citation practices between fields (Radicchi et al., 2008; Crespo et al., 2013, 2014; Li et al., 2013; and Ruiz-Castillo, 2014). Consequently, in this paper we adopt this procedure in the usual solution to the all-sciences aggregation problem.
- We present two types of results. Firstly, we assess the performance of the standard normalization procedure in facilitating the comparability of the citations received by articles belonging to different clusters. Secondly, we assess the consequences of adopting the two solutions to the all-sciences aggregation problem by comparing the corresponding university rankings according to the two citation impact indicators.

The rest of the paper is organized into three sections. Section II presents the citation impact indicators, as well as the two solutions to the all-sciences aggregation problem. Section III describes the data, and includes the empirical results, while Section IV concludes.

The aggregation of heterogeneous scientific fields in the all-sciences case

Notation and citation indicators

It is convenient to introduce some notation. Given a set of articles S, and J scientific fields indexed by $j = 1,\ldots, J$, a *classification system* is an assignment of articles in S to the J fields. Let I be the number of research units, indexed by $i = 1,\ldots, I$. In this Section, the assignment of articles in S to the I research units is taken as given. Let $\boldsymbol{c_{ij}} = \{c_{ijk}\}$ be the *citation distribution of unit i in field j*, where c_{ijk} is the number of citations received by the k-th article, and let $\boldsymbol{c_j}$ be the *citation distribution of field j*, that is, the union of all research units' citation distributions in that field: $\boldsymbol{c_j} = \cup_i \{\boldsymbol{c_{ij}}\}$. Finally, let $\boldsymbol{C} = \cup_i \cup_j \{\boldsymbol{c_{ij}}\}$ be the *overall citation distribution*, or the citation distribution in the all-sciences case. For later reference, let N_{ij} be the number of articles in distribution $\boldsymbol{c_{ij}}$, let $N_i = S_j N_{ij}$ be the total number of articles published by unit i, let $N_j = S_i N_{ij}$ be the total number of articles in field j, and let $N = S_i S_j N_{ij}$ be the total number of articles in the all-sciences case.

A *citation impact indicator* is a function F defined in the set of all citation distributions, where $F(\boldsymbol{c})$ is the citation impact of distribution $\boldsymbol{c}$. Let $\boldsymbol{c^r}$ be the r-th replica of distribution $\boldsymbol{c}$. An indicator F is said to be *size-independent* if, for any citation distribution $\boldsymbol{c}$, $F(\boldsymbol{c^r}) = F(\boldsymbol{c})$ for all r. An indicator F is said to be *scale-independent* if for any $\lambda > 0$, and any citation distribution $\boldsymbol{c}$, $F(\lambda\boldsymbol{c}) = F(\boldsymbol{c})$. An indicator F is said to be *additively decomposable* if for any partition of a citation distribution $\boldsymbol{c}$ into G sub-groups, indexed by $g = 1,..., G$, the citation impact of distribution $\boldsymbol{c}$ can be expressed as follows:

$$F(\boldsymbol{c}) = S_g\, (M_g/M) F(\boldsymbol{c_g}),$$

where M_g is the number of publications in sub-group g, and $M = \Sigma_g\, M$ is the number of publications in distribution $\boldsymbol{c}$.

Consider the following two difficulties for comparing the citation impact of any pair of research units: the two units may be of different sizes, and if they work in different fields, then their raw citations are not directly comparable. As it is well known, these two difficulties can be overcome using a size- and scale-independent indicator. The following two indicators are good examples of size- and scale-independent indicators that, in addition, are additively decomposable.

1. Let $\boldsymbol{X_j}$ be the set of the 10% most cited articles in citation distribution $\boldsymbol{c_j}$, and let $\boldsymbol{x_{ij}}$ be the sub-set of articles in $\boldsymbol{X_j}$ corresponding to unit i, so that $\boldsymbol{X_j} = \cup_i\{\boldsymbol{x_{ij}}\}$ with $\boldsymbol{x_{ij}}$ non-empty for some i. If n_{ij} is the number of articles in $\boldsymbol{x_{ij}}$, then the *Top 10% indicator for unit i in field j*, T_{ij}, is defined as

$$T_{ij} = n_{ij}/N_{ij}. \qquad (1)$$

Of course, for field j as a whole, if $n_j = \Sigma_i\, n_{ij}$ is the number of articles in $\boldsymbol{X_j}$, then $T_j = n_j/N_j = 0.10$.

2. Let z_j be the *Critical Citation Line* –CCL hereafter– for citation distribution $\boldsymbol{c_j}$, and denote the articles in $\boldsymbol{c_j}$ with citations equal to or greater than z_j as *high-impact articles*. For any high impact article with citations c_{il}, define the *CCL normalized high-impact gap* as $(c_{il} - z_j)/z_j$.

Consider the family of FGT indicators introduced in Albarrán et al. (2011b) as functions of normalized high-impact gaps. The second member of this family, A_{ij}, is defined as

$$A_{ij} = (1/N_{ij})[S_l(c_{il} - z_j)/z_j], \quad (2)$$

where the sum is over the high-impact articles in citation distribution $\boldsymbol{c_j}$ that belong to unit i. We refer to this indicator as the *Average of high-impact gaps for unit i in field j*. For the entire field j as a whole, the average of high-impact gaps is defined as

$$A_j = (1/N_j)[S_k (c_k - z_j)/z_j],$$

where the sum is over the high-impact articles in citation distribution $\boldsymbol{c_j}$.

To facilitate the comparison with T_{ij}, in the sequel we will always fix z_j as the number of citations of the article in the 90th percentile of citation distribution $\boldsymbol{c_j}$. In that case, the set of high-impact articles coincides with the set of the 10% most cited articles in citation distribution $\boldsymbol{c_j}$. The two main differences between the two indicators are the following. Firstly, one or more citations received by a high-impact article increases A_{ij} but does not change T_{ij}. In other words, A_{ij} is monotonic but T_{ij} is not. Secondly, T_{ij} is more robust to extreme observations than A_{ij}.

The solution to the all-sciences aggregation problem using the standard field-normalization procedure

For any i, let $\boldsymbol{c_i} = (\boldsymbol{c_{i1}},\ldots, \boldsymbol{c_{ij}},\ldots, \boldsymbol{c_{iJ}})$ be the *raw citation distribution of unit i in the all-sciences case*. Differences in production and citation practices across fields make impossible the direct comparison of the raw citations received by articles in different fields. In order to achieve some comparability, one possibility is to use some normalization procedure. For any article k in citation distribution $\boldsymbol{c_{ij}}$, the normalized number of citations c^*_{ijk} according to the *standard field-normalization procedure* is defined as

$$c^*_{ijk} = c_{ijk}/\mu_j.$$

For any i, let $\boldsymbol{c^*_i} = \cup_j \cup_k \{c^*_{ijk}\} = (\boldsymbol{c^*_{i1}},\ldots, \boldsymbol{c^*_{ij}},\ldots, \boldsymbol{c^*_{iJ}})$ be the *normalized citation distribution of unit i in the all-sciences case*. Since normalized citations are now comparable, it makes sense to apply any indicator to citation distribution $\boldsymbol{c^*_i}$. For any i, let $\boldsymbol{F^*_i} = F(\boldsymbol{c^*_i})$ be the citation impact of distribution $\boldsymbol{c^*_i}$ according to the indicator F. For any pair of research units u and v in the all-sciences case, the citation impact values $\boldsymbol{F^*_u}$ and $\boldsymbol{F^*_v}$ are now comparable, and can be used to rank the two units in question.

Note that, since $\boldsymbol{c^*_i}$ for $i = 1,\ldots, I$ forms a partition of $\boldsymbol{C^*}$ and F is assumed to be additively decomposable, we can write

$$F^* = F(\boldsymbol{C^*}) = S_i (N_i/N)\boldsymbol{F^*_i}.$$

Thus, if we rank universities by the ratio $\boldsymbol{F^*_i}/F^*$, $i = 1,\ldots, I$, then the value one can serve as a benchmark for evaluating the research units in the usual way. For later reference, since $\boldsymbol{c^*_{ij}}$ for $j = 1,\ldots, J$ forms a partition of $\boldsymbol{c^*_i}$, for each i we can write

$$\boldsymbol{F^*_i} = F(\boldsymbol{c^*_i}) = S_j (N_{ij}/N_i)\boldsymbol{F^*_{ij}}, \quad (3)$$

where $F^*_{ij} = F(\boldsymbol{c}^*_{ij})$ for all j, that is, F^*_{ij} is simply the citation impact of citation distribution $\boldsymbol{c}^*_{ij}$ according to F.

A solution to the all-sciences aggregation problem without field-normalization

For any i and any j, denote by $F_{ij} = F(\boldsymbol{c}_{ij})$ the citation impact of distribution $\boldsymbol{c}_{ij}$ according to F. A convenient measure of citation impact for unit i in the all-sciences case, F_i, can be defined as the weighted average of the values F_{ij} achieved in all fields, with weights equal to the relative importance of each field in the total production of unit i:

$$F_i = S_j\,(N_{ij}/N_i)F_{ij} \qquad (4)$$

The comparison of expressions (4) and (5) illustrate the differences between the two solutions to the all-sciences aggregation problem when the evaluation of the units' citation impact is made with additively decomposable indicators. Finally, it is convenient to compute the weighted average of these quantities as follows:

$$F = S_i\,(N_i/N)F_i.$$

Thus, as before, if we rank universities by the ratio F_i/F, $i = 1,\ldots, I$, then the value one can serve as a benchmark for evaluating the research units in the usual way. In practice, we have information concerning some but not all research units. Therefore, we compute F as the following weighted average: $F = \mathrm{S}_j\,(N_j/N)F_j$, where $F_j = \mathrm{F}(\boldsymbol{c}_j)$.

The aim of the paper

The main aim of this paper is the comparison between the rankings of research units obtained with and without the standard field-normalization procedure, $(F^*_1, \ldots, F^*_I)$ and $(F_1, \ldots, F_I)$, respectively.

To understand the way the results will be presented, we need to review the connection between the performance of the normalization procedure and the relationship between the solutions to the all-sciences aggregation problem. For that purpose, we need to introduce some more notation. For any j, let $\boldsymbol{x}_j$ be the set of high-impact articles in distribution $\boldsymbol{c}_j$, that is, the set of articles in $\boldsymbol{c}_j$ with citations equal to or greater than z_j, or the set of the 10% most cited articles in $\boldsymbol{c}_j$. Let us denote by $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots, \boldsymbol{x}_j,\ldots, \boldsymbol{x}_J)$ the set of high-impact articles in the all-sciences case. On the other hand, let $\boldsymbol{Y}$ be the set of the 10% most cited articles in the overall normalized citation distribution $\boldsymbol{C}^* = \cup_j \{\boldsymbol{c}^*_j\}$. Let $\boldsymbol{y}_j$ be the sub-set of articles in $\boldsymbol{Y}$ belonging to field j, so that $\boldsymbol{Y} = (\boldsymbol{y}_1,\ldots, \boldsymbol{y}_j,\ldots, \boldsymbol{y}_J)$. Note that, in practice, the sets $\boldsymbol{y}_j$ might be empty for some j.

Under the universality condition, that is, if all fields are equally distributed except for a scale factor then, at every percentile of field citation distributions, normalized citations will be the same for all fields. In other words, the normalization procedure will work perfectly. In particular, in this situation we would have $z_j/\mu_j = z^*$ for all j. Consequently, we would have $\boldsymbol{y}_j = \boldsymbol{x}_j$ for all j, and $\boldsymbol{Y} = \boldsymbol{X}$. Since citation distributions $\boldsymbol{c}^*_{ij}$ and $\boldsymbol{c}_{ij}$ have the same number of articles and our indicators are a function solely of high-impact articles, we would have $F^*_{ij} = F(\boldsymbol{c}^*_{ij}) = F_{ij} = F(\boldsymbol{c}_{ij})$ for all i and j. In view of equations (4) and (5), we would have $F^*_i = F_i$ for all i. In other words, the rankings $(F^*_1, \ldots, F^*_I)$ and $(F_1, \ldots, F_I)$ will be identical.

As we know, in practice the universality condition is not satisfied. However, the better the performance of the normalization procedure, that is, the closer is the set $\boldsymbol{Y}$ to set $\boldsymbol{X}$, the more similar the rankings $(F^*_1, \ldots, F^*_I)$ and $(F_1, \ldots, F_I)$ are expected to be for any F. Note that this

conjecture has to be verified in practice. In any case, the empirical section begins by assessing the performance of the normalization procedure.

On the other hand, independently of the normalization procedure's performance, we should measure the consequences of adopting the two solutions to the all-sciences aggregation problem using indicators with different properties. The reason, of course is that whenever $\boldsymbol{Y}$ and $\boldsymbol{X}$ differ, that is, when the set of high-impact articles under the two solutions differ, the consequences for the university rankings might be of a different order of magnitude depending on the citation impact indicator we use. This is the reason why we will study the situation using the Top 10% and the Average of high-impact gaps.

Empirical results

The data and descriptive statistics

As indicated in the Introduction, our dataset results from the application of a publication-level methodology to 9,446,622 distinct articles published in 2003-2012 (see Ruiz-Castillo & Waltman, 2015). Publications in local journals, as well as popular magazines and trade journals have been excluded (see Ruiz-Castillo & Waltman, 2015 for details). We work with journals in the sciences, the social sciences, and the arts and humanities, although many arts and humanities journals are excluded because they are of a local nature. The classification system consists of 5,119 clusters, and citation distributions refer to the citations received by these articles during a five-year citation window for each year in that period. In this paper, we focus on the set of 3,614,447 distinct articles published in 2005-2008. In terms of the notation introduced in Section II.1, we have $\boldsymbol{C} = \cup_j \{\boldsymbol{c_j}\} = (c_1,\ldots, c_N)$ with J = 5,119, and N = 3,614,447.

The research units are universities. Publications are assigned to universities using the fractional counting method that takes into account the address lines appearing in each publication. An article is fully assigned to a university only if all addresses mentioned in the publication belong to the university in question. If a publication is co-authored by two or more universities, then it is assigned fractionally to all of them in proportion to the number of address lines. For example, if the address list of an article contains five addresses and two of them belong to a particular university, then 0.4 of the article is assigned to this university, and only 0.2 of the article is assigned to each of the other three universities.

We know the total number of address lines of every publication, but we have information about the number of address lines of specific institutions only for the 500 LR universities. This number is well below I, the total number of research units in the notation introduced in Section II.1. There are 2,420,054 distinct articles, or 67% of the total, with at least one address line belonging to a LR university. The total number of articles in the LR universities according to the fractional counting method is 1,886,106.1, or 52.2% of the total. The distribution of this total among the 500 universities is available in Perianes-Rodriguez & Ruiz-Castillo, 2014a.

The performance of the normalization procedure

We assess the performance of the normalization procedure using the measurement framework introduced in Crespo et al. (2013), we first estimate the effect on overall citation inequality attributable to differences in production and citation practices between clusters, and then the reduction in this effect after applying the standard field-normalization procedure. Given the many clusters with very few publications (see Ruiz-Castillo & Waltman, 2015), we apply this method to the 3,332 clusters with more than 250 publications. These clusters include 3,441,666 million publications, or 95.2% of the total.

We begin with the partition of, say, each cluster citation distribution into P quantiles, indexed by $p = 1,.., P$. In practice, in this paper we use the partition into percentiles, that is, we choose $P = 100$. Assume for a moment that, in any cluster i, we disregard the citation inequality within every percentile by assigning to every article in that percentile the mean citation of the percentile itself, μ_i^p. The interpretation of the fact that, for example, $\mu_i^p = 2\ \mu_j^p$ is that, on average, the citation impact of cluster i is twice as large as the citation impact of cluster j in spite of the fact that both quantities represent a common underlying phenomenon, namely, the same *degree of citation impact* in both clusters. In other words, for any π, the distance between μ^p and μ_j^p is entirely attributable to the difference in the production and citation practices that prevail in the two clusters for publications with the same degree of excellence in each of them. Thus, the citation inequality between clusters at each percentile, denoted by $I(p)$, is entirely attributable to the differences in citation practices between the 3,332 clusters holding constant the degree of excellence in all clusters at quantile π. Hence, any weighted average of these quantities, denoted by *IDCC* (*I*nequality due to *D*ifferences in *C*itation impact between *C*lusters), provides a good measure of the total impact on overall citation inequality that can be attributed to such differences. Let $\boldsymbol{C'}$ be the union of the clusters citation distributions, $\boldsymbol{C'} = \cup \{\boldsymbol{c_j}\}$ for $j = 1,\ldots, 3{,}332$. We use the ratio

$$IDCC/I(\boldsymbol{C'}) \qquad (6)$$

to assess the relative effect on overall citation inequality, $I(\boldsymbol{C'})$, attributed to the differences in citation practices between clusters (for details, see Crespo et al., 2013).

Finally, we are interested in estimating how important scale differences between cluster citation distributions are in accounting for the effect measured by expression (6). For that purpose, we use the relative change in the *IDPC* term, that is, the ratio

$$[IDCC - IDCC^*]/IDCC, \qquad (7)$$

where *IDCC** is the term that measures the effect on overall citation inequality attributed to the differences in cluster distributions after applying the standard field-normalization procedure (for details, see again Crespo et al., 2013). The estimates of expressions (6) and (7) are as follows:

Table 1. The effect on overall citation inequality, I('C), of the differences in citation impact between clusters before and after standard field-normalization, and the impact of normalization on this effect.

		Normalization impact =100 [*IDCC – IDCC*/IDCC*]
Before MNCS normalization, 100 [*IDCC/I(C')*]	22.5 %	-
After MNCS normalization, 100 [*IDCC*/I(C')*]	4.3 %	84.3 %

It can be observed that the effect of the differences in citation practices between such a large number of clusters represents 22.5% of overall citation inequality, a figure much larger than what has been found in the previous literature for at most 235 sub-fields. Nevertheless, the standard field-normalization procedure reduces this effect down to 4.3%, quite an achievement.

Differences in university rankings under the two solutions to all-sciences aggregation problem

The university rankings without and with normalization according to the Top 10% indicator, T_i and T^*_i, and according to the Average of high-impact gaps, A_i and A^*_i can be found in Perianes-Rodriguez & Ruiz-Castillo (2014a). We begin with the comparison of university rankings according to T_i and T^*_i. The Pearson correlation coefficient between university values is 0.995, while the Spearman correlation coefficient between ranks is 0.992. However, high correlations between university values and ranks do not preclude important differences for individual universities. In analyzing the consequences of going from T_i to T^*_i, we must take two aspects into account. Firstly, we should analyze the re-rankings that take place in such a move. Secondly, we should compare the differences between the university values themselves. Fortunately, we have a relevant instance with which to compare our results: the differences found in Ruiz-Castillo and Waltman (2015) in going from the university rankings according to T_i using the Web of Science classification system with 236 journal subject categories, or sub-fields, and the classification system we are using in this paper with 5,119 clusters.

As much as 38.4% of universities experience very small re-rankings of less than or equal to five positions, while 67 universities, or 13.4% of the total, experience re-rankings greater than 25 positions. These figures are 20.2% and 39.0% when going from the WoS classification system to our dataset. Among the first 100 universities, 61 experience small re-rankings in going from T_i to T^*_i, while only 44 are in this situation in the change between classification systems. As far as the cardinal changes is concerned, 78.4% of universities have changes in top 10% indicator values smaller than or equal to 0.05 when going from T_i to T^*_i. This percentage is 71% among the first 100 universities. These figures are 50.1% and 60.0% in the change between classification systems. For most universities, the differences are more or less negligible. Although for some universities more significant differences can be observed, the conclusion is clear. The differences observed in university rankings according to the top 10% indicator when we adopt the two solutions for solving the all-sciences aggregation problem are considerably less than according to the same indicator when we move from the WoS classification system to our dataset (Perianes-Rodriguez & Ruiz-Castillo, 2014a).

The Pearson correlation coefficient between the university rankings according to the average of high-impact gaps, A_i and A^*_i, is 0.596, while the Spearman correlation coefficient between ranks is 0.984. However, the low Pearson correlation coefficient is due to the presence of the well-known extreme observation of the University of Göttingen (Waltman et al., 2012b; Ruiz-Castillo & Waltman, 2015). Without this university, this correlation coefficient becomes 0.986. In any case, as before, high correlations between university values and ranks do not preclude important differences for individual universities. The ordinal differences in university rankings according to this indicator with and without field-normalization are of a similar order of magnitude as those obtained with the top 10% indicator. For example, 33.0% of universities experience very small re-rankings of less than or equal to five positions, while 80 universities, or 16.0% of the total, experience re-rankings greater than 25 positions. Among the first 100 universities, only 44 experience small re-rankings in going from A_i to A^*_i (in comparison with 61 when going from T_i to T^*_i). As far as the cardinal changes is concerned, 64.2% of universities have changes in indicator values smaller than or equal to 0.05 when going from A_i to A^*_i –a comparable figure with 78.4% when going from T_i to T^*_I (Perianes-Rodriguez & Ruiz-Castillo, 2014a).

The conclusion is inescapable. In spite of the fact of the limitations of the standard normalization procedure in the presence of so many clusters, the differences observed in university rankings when we adopt the two solutions for solving the all-sciences aggregation problem are of a relatively small order of magnitude regardless of which of then two rather different citation impact indicators is used in obtaining the university rankings.

Conclusions

The heterogeneity of the fields distinguished in any classification system poses a severe aggregation problem when one is interested in evaluating the citation impact of a set of research units in the all-sciences case. In this paper, we have analyzed two possible solutions to this problem. The first solution relies on prior normalization of the raw citations received by all publications. In particular, we focus on the standard field-normalization procedure in which field mean citations are used as normalization factors. The second solution extends the approach adopted in the Leiden and SCImago rankings for computing the Top 10% indicator in the all-sciences case to any admissible indicator. This solution does not require any prior field-normalization: the citation impact of any research unit in the all-sciences case is calculated as the appropriately weighted sum of the citation impact that the unit achieves in each field.

Using a large WoS dataset consisting of 3.6 million publications in the 2005-2008 period and an algorithmically constructed publication-level classification system that distinguishes between 5,119 clusters, this simple alternative has been confronted with the usual one when the citation impact of the 500 LR universities are evaluated using two indicators with very different properties: the top 10% indicator, and the average of high-impact gaps.

The shape of the citation distributions of 4,161 significant clusters with more than 100 publications in our dataset has been previously shown to be highly skewed and reasonable similar (Ruiz-Castillo & Waltman, 2015). Previous results with WoS classification systems that distinguishes at most between 235 sub-fields indicate that, when this is the case, the standard field-normalization procedure performs well in reducing the overall citation inequality attributed to the differences in production and citation practices between fields. In this paper we have shown that this is not exactly the case, even when we restrict the attention to 3,332 clusters with more than 250 publications. Therefore, a priori it was not obvious what to expect when confronting the solutions to the all-sciences aggregation problem with and without prior field-normalization.

Interestingly enough, the differences between the university rankings obtained with both solutions is of a relatively small order of magnitude independently of the citation impact indicator used in the construction of the university rankings. In particular, these differences are considerably smaller than the ones obtained in Ruiz-Castillo and Waltman (2015) for the move from the WoS classification system with 236 sub-fields to the one used in this paper with 5,119 clusters.

In principle, it seems preferable to evaluate the citation impact of research units in the all-sciences case avoiding any kind of prior normalization operation. However, the empirical evidence presented in this paper indicates that that the use of the traditional methodology does not lead to very different results. This is a convenient conclusion, since there are instances when normalization is strongly advisable. For example, when one is interested in studying the research units citation distributions in the all-sciences case –as we do in the companion paper Perianes-Rodriguez and Ruiz-Castillo (2014b).

It should be noted that, before being accepted, it would be convenient to replicate the results of this paper for other datasets, other classification systems, other types of research units, and other ways of assigning responsibility between research units in the case of co-authored publications.

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*, 40–49.

Albarrán, P., Crespo, J., Ortuño, I., & Ruiz-Castillo, J. (2011a). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*, 385–397.

Albarrán, P., Ortuño, I., & Ruiz-Castillo, J. (2011b). The measurement of low- and high-impact in citation distributions: technical results. *Journal of Informetrics*, *5*, 48–63.

Bornmann, L., De Moya Anegón, F., & Leydesdorff, L. (2012) The new excellence indicator in the world report of the SCImago Institutions Rankings 2011. *Journal of Informetrics*, *6*, 333-335.

Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLoS ONE*, *8*, e58727.

Crespo, J. A., Herranz, N., Li, Y., & Ruiz-Castillo, J. (2014). The effect on citation inequality of differences in citation practices at the Web of Science subject category level. *Journal of the Association for Information Science and Technology*, *65*, 1244–1256.

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.

Li, Y., Castellano, C., Radicchi, F., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics*, *7*, 746–755.

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014a). *An alternative to field-normalization in the aggregation of heterogeneous scientific fields*. Working Paper, Economic Series 14-25, Universidad Carlos III (http://hdl.handle.net/10016/19812).

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014b). *University citation distributions*. Working Paper, Economic Series 14-26, Universidad Carlos III (http://hdl.handle.net/10016/19811).

Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE*, *7*, e33833.

Radicchi, F., Fortunato, S., & Castellano, C. (2008), "Universality of citation distributions: Toward an objective measure of scientific impact", *PNAS*, *105*, 17268-17272.

Ruiz-Castillo, J. (2014). The comparison of classification-system-based normalization procedures with source normalization alternatives in Waltman and Van Eck. *Journal of Informetrics*, *8*, 25–28.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*, 102-117. (DOI: 10.1016/j.joi.2014.11.010).

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*, 2378–2392.

Waltman, L., Van Eck, N. J., & Van Raan, A. F. J. (2012a). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*, 72–77.

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. (2012b). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*, 2419–2432.

Waltman, L., & Van Eck, N. J. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*, 833–849.

Correlating Libcitations and Citations in the Humanities with WorldCat and Scopus Data

Alesia Zuccala[1] and Howard D. White[2]

[1] *spl465@iva.ku.dk*
Royal School of Library and Information Science, University of Copenhagen
Birketinget 6, DK-2300 Copenhagen S (Denmark)

[2] *whitehd@drexel.edu*
College of Computing and Informatics, Drexel University
32nd and Chestnut Streets, Philadelphia, PA, 19104 (USA)

Abstract
The term *libcitations* was introduced by White et al. (2009) as a name for counts of libraries that have acquired a given book. Somewhat like citations, these library holdings counts, which vary greatly, can be taken as indicators of the book's cultural impact. Torres-Salinas and Moed (2009) independently proposed the same measure under the name *catalog inclusions*. Both articles sought an altmetric for authors of books in, e.g., the humanities, since the major citation indexes, oriented toward scientific papers, have not served them well. Here, using very large samples, we explore the libcitation-citation relationship for the same books by correlating their holdings counts from OCLC's WorldCat with their citation counts from Elsevier's Scopus. For books cited in two broad fields of the humanities during 1996-2000 and 2007-2011, we obtain positive, weak, but highly significant correlations. These largely persist when books are divided by main Dewey class. The overall results are inconclusive, however, because the Scopus citation counts for the books tend to be very low. Further correlational research should probably use the much higher book citation counts from Google Scholar. Nevertheless, a qualitative analysis of widely held and widely cited books clarifies the libcitation measure and helps to justify it.

Conference Topic
Indicators

Introduction
Journal-oriented scientists have long had citation counts as an indicator of the impact of their articles, and journal-based citation indexes cater to them. But the same indexes cover citations to books less well, and book-oriented scholars in the humanities and softer social sciences feel themselves at a disadvantage, especially if citation measures are going to be used in performance evaluations and funding decisions (see Kousha, Thelwall, & Rezaie 2011 for a review). White et al. (2009) responded to this lack by proposing that one measure of a book's cultural impact could be the number of libraries that hold it. The idea behind this altmetric was that librarians who acquire a book are somewhat like scholars who cite it, in that both acts involve assessment and choice on behalf of communities of readers. To bring out the parallel, White et al. called the librarians' formal act of acquisition a *libcitation* (first syllable as in "library"). They wrote that the libcitation count (also known as a library holdings count) for a particular book "increases by 1 every time a different library reports acquiring that book in a national or an international union catalog. Readers are invited to think of union catalogs in a new way: as 'librarians' citation indexes'" (p. 1084). OCLC's WorldCat was mentioned as a prime example of a union catalog—that is, one that pools the cataloging records of OCLC member libraries and reports how many of them hold each cataloged item.

At the same time and wholly independently, Torres-Salinas and Moed (2009) made an identical proposal. Their name for libcitations (our term here) was *catalog inclusions*, and they, too, stressed the parallel between such inclusions and citations to journal articles (p. 11). They, too, named WorldCat as a potential source of library holdings data. Moreover, both

they and White et al. raised the possibility of empirically testing the relationship between libcitation counts and citation counts for the same set of books: are the two correlated?

The question is important because citation counts, when scrupulously used, have become a standard performance indicator in many disciplines, and, given the inadequacies of citation data for books, it would be very interesting if libcitations could serve a similar purpose. Torres-Salinas and Moed (2009, p. 24) saw correlation research of this sort in terms of validating the holdings-count idea:

> One way of doing this is to examine...the degree of correlation between the number of times book titles are cited in the serial literature on the one hand, and the number of library catalogs in which they are included on the other.

That is just what the present paper does for books (aka titles) in two broad fields in the humanities: *History* and *Literature & Literary Theory*. It draws on a special database of book citation data from Elsevier's Scopus and libcitation data for the same books from WorldCat, as described in Zuccala and Guns (2013), a research-in-progress paper. White et al. (2009, p. 1094) had anticipated what would be found:

> It is an open question whether libcitation counts for books and book chapters will correlate significantly with citation counts for the same works. Indeed, they may not. Our exploratory trials have shown some books to be high in both citation and libcitation counts. Occasionally, a book turns up that is well cited despite being held by relatively few libraries. More common are books that are meagerly cited, but relatively widely held. This overall mix produces low correlations.

These remarks were occasioned by spot-checking citation counts in the Web of Science. Using Scopus instead, Zuccala and Guns (2013) provided the first empirical answer to the open question: they found low but significant correlations.

The present paper continues this line of analysis (also described in Sieber and Gradmann, 2011). We do not hypothesize that libcitations *cause* citations (or the reverse)—merely that the two variables may positively co-vary.

Our database covers more than 100,000 books, and it now allows correlations to be obtained in the 10 main Dewey subject classes. As before, it has a total libcitation count for each book, but also disaggregates that total into counts for members of the Association of Research Libraries (ARL) and counts for non-members. The non-members include thousands of academic and public libraries whose collections are not primarily intended to support advanced research. In contrast, the 125 ARL institutions own very large subject collections that support graduate degree programs and specialized faculty research in many disciplines. (When multiple libraries in ARL institutions buy the same book, its count can go well beyond 125.) The books with the greatest cultural impact achieve libcitation counts in the thousands by appealing to ARL members and non-members alike. Plum Analytics, a commercial firm specializing in altmetrics, now includes a book's holding count in WorldCat as one of its indicators of "usage."

The results of our analyses, while interesting and suggestive, return us to a common criticism of both the Web of Science and Scopus: within the time frame of our study, they pick up too few citations to books to correlate those citations with libcitations on a firm basis. Both WoS and Scopus have recently expanded their efforts to capture citations to books, but it is too early to assess the full effect of these new data on bibliometrics. Kousha, Thelwall and Rezaie (2011) demonstrate that Google Books and Google Scholar give considerably higher citation counts for books than Scopus does. Our findings point to the same conclusion.

Overview of the database

Here we re-present several details about our database from Zuccala and Guns (2013) and add some new ones. The Scopus database from Elsevier supplied our citation data, which was

granted through the Elsevier Bibliometrics Research Program. Having requested separate datasets in *History* and *Literature & Literary Theory*, we further limited them to citations that appeared in journal articles during two periods, 1996-2000 and 2007-2011. We examined the Scopus data to determine the overall frequency with which various types of publications were cited: books, research articles, conference proceedings, review papers, notes, and other materials. Cited materials that were "non-sourced"—that is, that did not have a Scopus identification number linking them to a source journal—were classified as books, the unit of analysis in which we were interested.

Table 1 shows the number of journals in each field (as classified by Scopus) from which we drew citing articles. The lower part of Table 1 gives the numbers (N's) of books cited in the journal articles in each field and period. It will be seen that, in both fields, the N's of books cited in the earlier period are much smaller than those in the later, because Scopus covered fewer humanities articles in the 1990s.

Table 1. Journals and journal citation data in Scopus (April 2011).

Journal counts and classification codes	
History (N=494 source journals)	ASJC 1202 (Scopus Classification Code)
Literature & Literary Theory (N=419 source journals)	ASJC 1208 (Scopus Classification Code)
Both History and Literature (N=110 source journals)	Both ASJC 1202 and ASJC 1208
Counts of books cited during 1996-2000	*Counts of books cited during 2007-2011*
History (N=20,996)	History (N=50,466)
Literature & Literary Theory (N=7,541)	Literature & Literary Theory (N=35,929)

We searched the apparent books in WorldCat, using an API developer key granted to us by the Online Computer Library Center (OCLC). The key allowed us to match titles cited *at least once* in Scopus with titles held by *at least one ARL and one non-ARL library* covered by WorldCat. (These libraries, while mostly North American, include participants worldwide.) For every matched title (confirming that it was a book), we retrieved the OCLC accession number, ISBN number, publisher's name, publisher's location, and library count data. These were added to the book's citation data from Scopus to create a unique Scopus-WorldCat relational database.

Once a book has been published, it takes time for it to be acquired and cataloged by a library. A book published in a given year could have been acquired by a library no earlier than that year, but might have been acquired up to and including November 2012. Our holdings counts were current as of that cut-off date.

To improve publication-date accuracy, we analyzed only books published in the six years immediately preceding our two five-year citation windows. Thus, the books cited in 1996-2000 were limited (by filtering their Scopus records) to those published during 1990-1995. The books cited in 2007-2011 were likewise limited to those published during 2001-2006.

Converted to the four files at the bottom of Table 1, our book data come to 114,932 cases in all, 81 percent of which are unique titles. The remaining 19 percent are titles that appear more than once. Some were cited in both our earlier and later periods. Others were cited in both the History and the Literature journals, or in the journals that Scopus has assigned to both fields jointly, as shown in Table 1. We did not attempt to re-assign these latter titles to a single field, but allowed them to enter into the counts for both fields. There seems no easy way to avoid double counting, because that is the way in which Scopus has structured the data. Even so, a trial analysis with duplicates removed does not greatly affect the correlations.

Data analyses and results

Our data analyses were conducted with SPSS, the Statistical Package for the Social Sciences. Table 2 gives summary statistics for the titles in History and Literature. Means and standard deviations have been rounded to whole numbers. As noted in Zuccala and Guns (2013, p. 357), both citations and libcitations exhibit the highly skewed distributions typical of bibliometrics. However, the subsets of ARL libcitations for both History and Literature have bimodal distributions, with peaks at 1-4 and 100-104 holding libraries, and a low point at 45-54 libraries. In other words, the ARL libraries tend to acquire large numbers of rarely held titles, large numbers of widely held titles, and markedly fewer titles with holdings counts in between. This saddle-shaped distribution may reflect the opposing needs of specialized researchers: on their behalf, ARL libraries acquire many books held by few other members, but also many books that almost every member *must* have. The titles with the maximum counts in Table 2 (e.g., 92 citations; 4,725 libcitations) will be named in Tables 6 through 9.

Table 2. Summary statistics for two fields in combined time periods.

History combined periods N=71462					
	Minimum	Maximum	Mean	Std. Dev.	Median
Citations	1	92	2	3	1
ARL libcitations	1	212	59	40	63
Non-ARL libcitations	1	4603	278	351	178
Total libcitations	2	4725	338	372	250
Literature combined periods N=43470					
	Minimum	Maximum	Mean	Std. Dev.	Median
Citations	1	91	2	3	1
ARL libcitations	1	215	62	38	67
Non-ARL libcitations	1	4603	305	395	189
Total libcitations	2	4725	367	412	267

In Table 3, *citation* counts for every book are correlated with total *libcitation* counts for every book in major subsets of the database. Citation counts are also separately correlated with the libcitation counts for ARL members and non-members. (Only the libcitation variables are labeled, but the unlabeled citation variable is present in all the cells.) These are Spearman rho correlations, calculated with ranks of the count values rather than the counts themselves. Unlike Pearson r's, rho's do not require the assumption of normally distributed populations and so accommodate bibliometric skew (Zuccala & Guns, 2013: 357).

Table 3. Total, ARL, and non-ARL libcitations to books correlated with citations to the same books in two fields, two periods, and combined periods.

History 1996-2000			**History 2007-2011**			**History combined**		
Total	ARL	Non-ARL	Total	ARL	Non-ARL	Total	ARL	Non-ARL
0.26	0.29	0.25	0.25	0.28	0.24	0.24	0.26	0.23
N=20996			N=50466			N=71462		
Literature 1996-2000			**Literature 2007-2011**			**Literature combined**		
Total	ARL	Non-ARL	Total	ARL	Non-ARL	Total	ARL	Non-ARL
0.23	0.28	0.22	0.18	0.24	0.17	0.20	0.24	0.19
N=7541			N=35929			N=43470		

The rho's are all positive and weak, with values much like those in Zuccala and Guns (2013, p. 357). Because of the large numbers of books involved, all are significant at $p < .001$ by

one-tailed test. The hypothesis of no relationship can thus be safely rejected: citations and libcitations do capture a certain amount of scholarly impact in common. A sign of this in Table 3 is that citations, which are essentially a researchers' practice, always correlate a bit more highly with libcitations from research libraries—that is, ARL members. However, none of the rho's are strong enough to indicate that libcitations can substitute for citations as a measure. Libraries, especially ARL members, do buy many books that turn out to be well cited, but they buy even more books that are not highly cited in the journals covered by Scopus. This raises questions about the citation-libcitation relationship that we will return to later with specific examples.

Table 4 may clarify the situation in our two subject fields. The total libcitation counts for books have been divided at their medians. Citation counts for the same books have been collapsed into three groups, as shown in the column labels. In both History and Literature, the two variables are directly related: as citation counts rise, the percentage of books with above-median libcitation counts also rises sharply. For example, in History, only 43% of books cited once have libcitation counts in the top half, whereas for books cited two to four times the comparable figure is 59%, and for books with five or more citations, 79%. The percentages in the Literature table are almost identical.

Table 4. Libcitations and citations cross-tabulated in two fields for combined periods.

History combined periods				
	Citations			
Libcitations	1	2-4	5 or more	
GT Median	43%	59%	79%	50%
LE Median	57%	41%	21%	50%
	100%	100%	100%	100%
N =	46578	19165	5719	71462
Literature combined periods				
	Citations			
Libcitations	1	2-4	5 or more	
GT Median	44%	59%	78%	50%
LE Median	56%	41%	22%	50%
	100%	100%	100%	100%
N =	29876	10668	2926	43470

However, this effect must be viewed in light of the extreme skew of the citation counts seen in the column marginals. Roughly two-thirds of all books in our samples have only one citation each, and roughly another quarter have only two to four citations. The fraction of titles with five or more citations is relatively small. Thus, the Spearman rho's for these grouped variables, though highly significant ($p < .001$), are even lower than when the variables are ungrouped in Table 3—only 0.22 for History and 0.19 for Literature.

We turn to a finer breakdown of the data. As mentioned in Zuccala and Guns (2013, p. 358), historians who publish in History journals do not exclusively cite works of history, nor do literary scholars who publish in Literature journals exclusively cite works of literature or literary theory. Instead, both groups cite books across the full range of subjects covered by the Dewey Decimal Classification. We were able to get the Dewey class numbers for most of our book titles from WorldCat. (Some books do not receive Dewey classifications.) In Table 5 we subdivide the books cited in History and Literature journals in our two time periods by their main Dewey classes.

Class 000 in Dewey is formally "Computer science, information, general works." This class is traditionally used for general reference books and books in trans-disciplinary fields such as librarianship, journalism, publishing, and reading. Historians and literary scholars mainly cite

books in areas like these, rather than in computer science. Hence, we have shortened the long label here to "General works."

The Table 5 cells contain 120 replications of our correlational study in subsets of the data. We are again correlating each book's total citations with its total libcitations, as well as the libcitation counts from ARL members and ARL non-members. In making comparisons, be aware that non-ARL libcitations make up by far the larger share of total libcitations. The two categories thus tend to produce correlations that are similar or identical, and so the non-ARL results will not be separately discussed here.

Table 5. Libcitations correlated with citations to books by field, period, and main Dewey classes.

	History 1996-2000				**History 2007-2011**			
Main Dewey Classes	Libcites	ARL	Non-ARL	N =	Libcites	ARL	Non-ARL	N =
000 General works	0.20	0.21	0.20	350	0.23	0.28	0.22	794
100 Philosophy and psychology	0.20	0.21	0.19	1055	0.18	0.20	0.17	2041
200 Religion	0.27	0.27	0.26	1766	0.27	0.29	0.25	4186
300 Social sciences	0.26	0.28	0.26	8067	0.23	0.25	0.21	16585
400 Language	0.11	0.11	0.12	247	0.17	0.16	0.17	672
500 Science	0.20	0.27	0.19	914	0.13	0.23	0.11	1543
600 Technology	0.25	0.35	0.23	824	0.12	0.24	0.09	1990
700 Arts and recreation	0.21	0.24	0.20	1056	0.19	0.26	0.18	3788
800 Literature	0.17	0.26	0.15	1620	0.20	0.26	0.19	4725
900 History and geography	0.28	0.31	0.27	4388	0.27	0.29	0.25	10439
	Literature 1996-2000				**Literature 2007-2011**			
Main Dewey Classes	Libcites	ARL	Non-ARL	N =	Libcites	ARL	Non-ARL	N =
000 General works	0.09	0.08	0.09	155	0.17	0.36	0.14	548
100 Philosophy and psychology	0.19	0.22	0.18	585	0.23	0.27	0.22	1919
200 Religion	0.13	0.19	0.12	398	0.25	0.29	0.23	2221
300 Social sciences	0.14	0.16	0.14	1344	0.19	0.22	0.18	6322
400 Language	0.22	0.24	0.21	505	0.22	0.24	0.20	1218
500 Science	0.04	0.09	0.04	115	0.06	0.12	0.06	516
600 Technology	0.13	0.28	0.11	130	0.09	0.24	0.07	703
700 Arts and recreation	0.18	0.21	0.17	591	0.22	0.26	0.20	3268
800 Literature	0.23	0.31	0.21	2616	0.26	0.31	0.25	11171
900 History and geography	0.14	0.25	0.12	742	0.21	0.26	0.20	3963

Even with Table 5's extensive partitioning, the N's underlying the correlations are large enough that most of the rho's remain highly significant ($p < .001$ by one-tail test). Of the correlations between citations and total libcitations, 21 out of 40 remain at or above 0.20. Large N's can cause correlations that are statistically but not substantively significant (Babbie 2015, p. 469). Nevertheless, certain patterns do lend substance to the overall analysis:

- Some 33 of the 40 ARL correlations remain in the 0.20s or higher.
- Some 37 of the 40 ARL correlations are higher than those for the non-ARL libraries in their row. This reinforces the supposed connection between citations and libcitations in research environments.
- As examples of subject accord, the ARL correlation for books classed in *900 History and geography* is second-highest (0.31) in History 1996-2000, and tied-highest (0.29) in History 2007-2011.
- As further examples of subject accord, the ARL correlation for books classed in *800*

Literature is highest (0.31) in Literature 1996-2000, and second-highest (0.31) in Literature 2007-2011.

- In both our History periods, the lowest correlations occur for books classed in *400 Language*. The N's for books in this class, which is historically Dewey's smallest, are likewise small. While historians make use of research from all fields, it is unsurprising that books on language are not their chief resource.
- In both our Literature periods, the lowest correlations occur for books classed in *500 Science,* and the N's for books in this class are small as well. One would not expect literary scholars to cite numerous science books. However, one might expect them to cite more books in *400 Language* than historians, and that is what the data show.
- Table 5 in fact shows wide variation in the number of books that Scopus authors have cited in each class. In both History periods, books classed in *300 Social Sciences* are most numerous. This makes sense because of the close interplay between historical and social scientific topics. Books classed in *900 History and geography* are the second-most numerous, and books in *800 Literature* are third. In both Literature periods, the same three classes dominate but in another order: *800 Literature* first, as seems fitting, then *300 Social Sciences* and *900 History and geography*. For our two broad fields in the humanities, these are reassuringly reasonable outcomes.

Since libcitations are a new altmetric, we think it informative to display the titles that have top-ranked libcitation counts in particular contexts (as do both Torres-Salinas and Moed, 2009 and White et al., 2009). This allows a qualitative as well as a quantitative analysis. White (2005) proposed the label *bibliograms* for bibliometric distributions in which not only the ranked counts but also the terms associated with them are analyzed as communications. "Bibliograms," he wrote (p. 443), "consist of (1) at least one seed term that sets a context, (2) terms that co-occur with the seed across some set of records, and (3) counts of how frequently terms co-occur with the seed by which they can be ordered high to low." Here, we use main Dewey class names as seed terms. We then rank the books that co-occur with them (as OCLC accession numbers) by their libcitation or citation counts. Lastly, the OCLC numbers are used to retrieve full bibliographic data from WorldCat so that we can comment on the authors, titles, and nature of the top-ranked books.

Table 6 comprises extracts from 40 bibliograms. We display, for our two fields and two time periods, the titles with the highest *total* libcitation counts in each of the 10 main Dewey classes. Many of these books have subtitles, but they have been omitted in favor of authors' surnames (or those of first authors in collaborations). We also display their ARL libcitation counts and their citation counts in Scopus.

The books in Table 6 do not resemble typical scientific articles. They are the sort of titles that present readers, like everyone else, may have purchased for reasons having nothing to do with bibliometrics. They exemplify the broad cultural impact of the humanities—for example, standard reference works on language, music, religion; biographies of famous men (Peter Gay's *Freud*, David McCullough's *Truman* and *John Adams*); novels (Toni Morrison's *Paradise*, Dan Brown's *The Da Vinci Code*); popularizations of science (Dava Sobel's *Longitude*, Malcolm Gladwell's *Blink*, Carl Sagan's *Cosmos*); best-selling social critiques (Susan Faludi's *Backlash*, Robert Hughes's *Culture of Complaint*); advice for business executives (James Collins's *Good to Great*, Thomas Peters and Robert Waterman's *In Search of Excellence*). While some exemplify high scholarship, others are not scholarly at all (Ernest Hemingway's *A Moveable Feast*); some are even children's books (David Wiesner's *Flotsam*, Peter Spier's *Noah's Ark*, both Caldecott Medal winners). They come to the fore here because they were bought by thousands of libraries, and they had citation counts of at least one in Scopus. Persons at research universities who specialize in manifestations of popular culture are legion.

Table 6. Books with highest libcitation counts by field, period, and main Dewey class.

History 1996-2000					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
1	160	2592	General works	The Oxford dictionary of modern quotations	Augarde
1	143	2936	Philosophy and psychology	Freud	Gay
1	101	2789	Religion	Crossing the threshold of hope	John Paul II
1	124	4233	Social sciences	My American journey	Powell
1	105	3433	Language	The story of English	McCrum
2	108	2572	Science	Longitude	Sobel
1	112	3204	Technology	Healing and the mind	Moyers
1	130	2133	Arts and recreation	Culture of complaint	Hughes
1	122	4132	Literature	Paradise	Morrison
4	137	4724	History and geography	Truman	McCullough
History 2007-2011					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
1	160	2592	General works	The Oxford dictionary of modern quotations	Augarde
2	145	4059	Philosophy and psychology	Blink	Gladwell
1	93	2931	Religion	Under the banner of heaven	Krakauer
4	152	3967	Social sciences	Freakonomics	Levitt
5	182	2760	Language	The Oxford English dictionary	Simpson
4	104	3284	Science	A short history of nearly everything	Bryson
2	148	4496	Technology	In search of excellence	Peters
4	123	2596	Arts and recreation	New Grove dictionary of music	Grove
6	122	4725	Literature	The Da Vinci code	Brown
5	140	4655	History and geography	John Adams	McCullough
Literature 1996-2000					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
2	155	2076	General works	Double fold	Baker
3	145	4059	Philosophy and psychology	Blink	Gladwell
1	87	3511	Religion	Noah's ark	Spier
3	152	3967	Social sciences	Freakonomics	Levitt
1	105	3433	Language	The story of English	McCrum
1	125	3884	Science	Cosmos	Sagan
1	141	4195	Technology	Good to great	Collins
1	86	4133	Arts and recreation	Flotsam	Wiesner
13	122	4725	Literature	The Da Vinci code	Brown
1	140	4655	History and geography	John Adams	McCullough
Literature 2007-2011					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
1	115	3342	General works	The road ahead	Gates
1	75	2455	Philosophy and psychology	Care of the soul	Moore
1	128	3083	Religion	The Oxford companion to the Bible	Metzger
2	154	3169	Social sciences	Backlash	Faludi
2	148	3119	Language	The Oxford companion to the English language	McArthur
1	112	2068	Science	Black holes and time warps	Thorne
1	93	4314	Technology	Men are from Mars, women are from Venus	Gray
1	130	2133	Arts and recreation	Culture of complaint	Hughes
1	125	3455	Literature	A moveable feast	Hemingway
1	121	3600	History and geography	The fifties	Halberstam

Thus, even the most pop-cultural books in Table 6 are widely held by ARL members. It is a misconception that these libraries acquire only works of rarified scientific or scholarly status. In fact, they buy innumerable works that would also be found in public and school libraries. The best example is the single most widely held item in our database—*The Da Vinci Code*, owned by 122 (of 125) ARL members. Whatever one may think of this novel, it had a huge impact for several years, and scholars in the humanities will want copies on hand, if only to attack Dan Brown's transgressions. Nevertheless, the citation counts for these books in Table 6's leftmost column are very low. Brown's novel has the most, and these may include book reviews.

By contrast, Table 7 displays the titles that are *most* highly cited in our categories. As implied earlier, relatively high citation counts tend to signal a research orientation, and these 40 books, which have the top counts in their respective Dewey classes, are almost all distinctly more academic than those in Table 6. Their *total* libcitation counts tend to be lower than those in Table 6, suggesting more specialized readerships. (The exception is *The Guardian*, a Nicholas Sparks novel.) A fair number of them address themes prominent in the humanities (race, class, gender, imperialism), and their authors include names famous to postmodern scholars, if not to the general public (e.g., Edward Said, Gilles Deleuze, Judith Butler, Donna Haraway, Gayatri Spivak, and, with two books, Giorgio Agamben).

Three-fourths of these books are held by a hundred or more ARL libraries. Of those that are not, some may reflect genuinely narrower acquisition by ARL members. Others (if not errors) may reflect delayed or incomplete reporting of an acquired book that makes its libcitation count deceptively small. That may have happened, for instance, with Spivak's *Death of a Discipline*, whose ARL count in Table 7 is only 22, but whose count as an e-book in WorldCat is 1,246 at this writing.

In any event, ARL libcitation counts range unbrokenly over values from 1 to 215. Given this variation, why are the correlations of ARL counts with citations not higher? We have already noted that they tend to be higher than correlations of *total* libcitations with citations, but only slightly. In both cases the problem is the same: the great majority of books in our database have only one citation (or at most a few). Thus, a key variable in our study has little variability. As one illustration, Table 8 lists the five books with the highest ARL libcitation counts in our two fields (time periods combined, and omitting the *Oxford English Dictionary*, already shown). These books are best-sellers not only among ARL members but in libraries of all kinds. Yet their citation counts in Scopus are minuscule and much the same, just as they were for the books in Table 6. To anyone familiar with these titles, it is incredible that Table 8 reflects their full citation records. Rather, their true counts are not being captured.

Not too long ago, this assumption could only have been checked with data from the Web of Science, but now we can spot-check citations to books in Google Scholar. When that is done, the results are very different from what Scopus shows, whether the Scopus figures are as low as one or as high as 92. Table 9 suggests the nature of the problem. The counts there reflect our judgment calls, such as to include only those for the 2000 edition of *DSM-IV-TR* or the 2007 edition of *The Elements of Style*. Google Scholar itself does not break down by edition the many citations to the feminist classic *In a Different Voice*. Nor does it allow us to extract citations to books in our two periods of study. Nevertheless, the Google Scholar counts indicate where further correlational research should be directed (see also Prins et al., 2014).

Table 7. Books with highest citation counts by field, period, and main Dewey class.

History 1996-2000					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
14	117	573	General works	The letters of the Republic	Warner
30	115	798	Philosophy and psychology	The production of space	Lefebvre
19	111	689	Religion	Ritual theory, ritual practice	Bell
75	129	1195	Social sciences	Imagined communities	Anderson
11	76	509	Language	Biblical Hebrew syntax	Waltke
29	107	450	Science	Bayes or bust?	Earman
25	84	364	Technology	Curing their ills	Vaughan
13	108	650	Arts and recreation	Orientalism	MacKenzie
56	119	1381	Literature	Culture and imperialism	Said
71	119	1406	History and geography	Britons	Colley
History 2007-2011					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
24	114	546	General works	"The tyranny of printers"	Pasley
39	26	413	Philosophy and psychology	The navigation of feeling	Reddy
37	109	478	Religion	Formations of the secular	Asad
92	114	602	Social sciences	Carnal knowledge and imperial power	Stoler
22	12	481	Language	Bilingualism and the Latin language	Adams
31	115	556	Science	The body of the artisan	Smith
32	100	342	Technology	Contagious divides	Shah
17	92	412	Arts and recreation	The reformation of the image	Koerner
26	32	2802	Literature	The guardian	Sparks
83	116	813	History and geography	The birth of the modern world, 1780-1914	Bayly
Literature 1996-2000					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
71	110	415	General works	The reading nation in the Romantic period	St. Clair
79	102	391	Philosophy and psychology	The open	Agamben
36	87	404	Religion	Saint Paul	Badiou
91	117	545	Social sciences	State of exception	Agamben
37	101	377	Language	The translation zone	Apter
12	95	294	Science	The spacious word	Padron
37	71	259	Technology	The companion species manifesto	Haraway
27	104	348	Arts and recreation	In the break	Moten
85	22	559	Literature	Death of a discipline	Spivak
87	106	462	History and geography	Writing history, writing trauma	LaCapra
Literature 2007-2011					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
6	117	573	General works	The letters of the Republic	Warner
17	108	632	Philosophy and psychology	Difference and repetition	Deleuze
6	114	771	Religion	Fragmentation and redemption	Bynum
41	131	1049	Social sciences	Gender trouble	Butler
19	84	301	Language	Discourse and social change	Fairclough
9	117	1034	Science	The origins of order	Kauffman
5	112	475	Technology	The commodity culture of Victorian England	Richards
11	122	983	Arts and recreation	Gone primitive	Torgovnick
38	120	843	Literature	The location of culture	Bhabha
23	125	891	History and geography	Imperial eyes	Pratt

Table 8. Books with the top five ARL libcitation counts in two fields.

History combined				
Cites	**ARL**	**Libcites**	**Title**	**Author**
2	212	4101	Diagnostic and statistical manual of mental disorders: DSM-IV-TR	
2	194	2478	In a different voice	Gilligan
3	180	1282	The alchemy of race and rights	Williams
2	176	1348	On the law of nations	Moynihan
1	176	1136	Theoretical perspectives on sexual difference	Rhode
Literature combined				
Cites	**ARL**	**Libcites**	**Title**	**Author**
1	215	3792	Publication manual of the American Psychological Association	
1	204	3436	The elements of style	Strunk, White
1	203	2046	A theory of justice	Rawls
3	178	1466	There's no such thing as free speech, and it's a good thing, too	Fish
1	175	995	Sex and reason	Posner

Table 9. Same data, but with citations in Scopus replaced by citations in Google Scholar.

History combined				
Cites	**ARL**	**Libcites**	**Title**	**Author**
5364	212	4101	Diagnostic and statistical manual of mental disorders: DSM-IV-TR	
30044	194	2478	In a different voice	Gilligan
2431	180	1282	The alchemy of race and rights	Williams
146	176	1348	On the law of nations	Moynihan
102	176	1136	Theoretical perspectives on sexual difference	Rhode
Literature combined				
Cites	**ARL**	**Libcites**	**Title**	**Author**
1393	215	3792	Publication manual of the American Psychological Association	
2988	204	3436	The elements of style	Strunk, White
782	203	2046	A theory of justice	Rawls
616	178	1466	There's no such thing as free speech, and it's a good thing, too	Fish
1546	175	995	Sex and reason	Posner

Discussion

The correlations in this paper suggest that libcitations and citations are not entirely different measures of impact. However, we are left wanting citation counts for books that do not have so many low, tied values. It is possible that better data would again produce low or even negligible correlations. It is also possible that the correlations would be much higher than those seen here. The libcitation measure draws on a varied mix of assessments, and they are not necessarily the same as those that go into scholars' acts of citation. But, as our data make plain, they indicate major intellectual achievements no less forcefully than citations. In fact, one can argue that many of the humanities titles in Table 6 are *truly* major achievements, in that they have reached large publics beyond academe.

What, then, do libcitations measure? Briefly, they estimate the potential readerships, or users, of a given book. Citations, in contrast, measure actual uses to which the book has been put within research-oriented communities. It is therefore not surprising that citations and libcitations are associated, especially if the latter come from libraries that serve researchers,

such as those in ARL. But libcitations also measure broad cultural impacts that citations may miss, because libcitations rest on chains of judgments within the world of publishing, and this world, which subsumes the scholarly one, extends into every part of life. The chains include authors, agents, past editors who have built publishers' reputations, present-day editors of various kinds, referee-readers, marketers, and wholesalers. Librarians are only the last link.

This speaks to the common objection that librarians do not evaluate individual titles, but put their acquisitions on automatic pilot through approval plans and the like; how, then, can libcitations reflect genuine worth? On the contrary, librarians are highly attuned to potential demand in their communities, and it is they who approve the approval plans and buy into the pre-formed collections. It is quite true that such moves favor some publishers over others, but that is because librarians trust the chains of judgment those publishers represent. And so do their communities, who routinely expect librarians to have acquired certain books they learn about and are displeased if they have not.

Libcitations are sales figures—a market measure. They reflect virtual unanimity on the worth of some titles, but they vary enormously. In our database, although the counts run to the high values seen in our tables, many titles are held by only one ARL and one non-ARL library, just as many papers have only a citation or two. Research on libcitation-citation correlations should continue, but even if they remain low, that does not invalidate the libcitation measure. It is better thought of as a free-standing gauge of authors' cultural impact. Having published a book, what author would not prefer a thousand libraries to hold it rather than 10?

Acknowledgments

The authors are grateful to the Elsevier Bibliometrics Research Programme (http://ebrp.elsevier.com/) and OCLC WorldCat for granting access to the data used to build the unique database for this study. We also thank Dr. Roberto Cornacchia for helping to develop the database, as well as Maurits van Bellen and Robert Iepsma for their data cleaning and standardisation work. Dr. Stefanie Haustein of the École de bibliothéconomie et des sciences de l'information (EBSI), Université de Montréal, kindly provided information on the library holdings-count measure at Plum Analytics.

References

Babbie, E. R. (2013). *The practice of social research*. 14th edition. Boston, MA: Cengage Learning.

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.

Prins, A., Costas, R., van Leeuwen, T., & Wouters, P. (2014). Using Google Scholar in research evaluation of social science programs, with a comparison with Web of Science data. *Proceedings of the Science and Technology Indicators Conference 2014 Leiden*, 434-443.

Sieber, J., & Gradmann, S. (2011). How to best assess monographs? An attempt to assess the impact of monographs using library infrastructure and Web 2.0 tools. European Educational Research Quality Indicators. Retrieved December 17, 2014 from

http://edoc.hu-berlin.de/docviews/abstract.php?id=38002

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. D. (2005). On extending informetrics: An opinion paper. Proceedings of the 10th International Society for Scientometrics and Informetrics Conference, 2, 442-449.

White, H. D., 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.

Zuccala, A., & Guns, R. (2013). Comparing book citations in humanities journals to library holdings: Scholarly use versus 'perceived cultural benefit' (RIP). *Proceedings of the 14th International Society for Scientometrics and Informetrics Conference, 1*, 353-360.

A Vector for Measuring Obsolescence of Scientific Articles

Jianjun Sun[1], Chao Min[1] and Jiang Li[2]

[1] *sjj@nju.edu.cn*
School of Information Management, Nanjing University, Nanjing (China)

[1] *marlonmassine@yeah.net*
School of Information Management, Nanjing University, Nanjing (China)

[2] *li-jiang@zju.edu.cn*
Department of Information Resource Management, Zhejiang University, Hangzhou (China)

Abstract
Diachronous studies of obsolescence categorized articles into three general types: "flashes in the pan", "sleeping beauties" and "normal articles", by using quartiles to identify first 25% and last 75% articles reaching 50% of their total citations, or by using averages to define threshold values of sleeping and awakening periods. However, the average-based and quartile-based criteria, sometimes, less effectively distinguished "flashes in the pan" and "sleeping beauties" from normal articles. In this research, we proposed a vector for measuring obsolescence of scientific articles, as an alternative to these criteria. The obsolescence vector is designed as $O = (G_s, A^-, n)$, where n is the age of an article, G_s and A^- are parameters for revealing the shape of citation curves. Among Nobel laureates' 28,340 articles, each of which received over 20 citations, we identified 265 flashes in the pan (approximately 1%) and 40 sleeping beauties (approximately 0.1%) by the obsolescence vector. By a few case studies, it is verified that obsolescence vector yielded more reasonable classifications than did the average-based and quartile-based criteria.

Conference Topic:
Indicators

Introduction
In a previous study (Li et al., 2014), we introduced G_s index, an adjustment of Gini coefficient, for measuring the inequality of "heartbeat spectrum" of "sleeping beauties". "Sleeping beauty" in science was first proposed by van Raan (2004), in order to describe a phenomenon where papers did not achieve recognition in citations until many years after their original publication. As in the fairy tale, a princess (an article) sleeps (goes unnoticed) for a long time and then, almost suddenly, is awakened (receives a lot of citations) by a prince (another article). "Heartbeat spectrum" was defined as a vector of a sleeping beauty's annual citation(s) received in the sleeping period.

How to categorize recognition to a paper as "early", "delayed" or "normal"? Diachronous studies of obsolescence answered this question, by using quartiles to identify first 25% and last 75% articles reaching 50% of their total citations (Costas et al., 2010), or by using averages to define threshold values of sleeping and awakening periods (van Raan, 2004; van Dalen & Henkens, 2005). In this research, we propose an obsolescence vector based on the G_s index, as an alternative to both approaches.

Literature review
"Obsolescence" (or "ageing") studies, in the field of bibliometrics, attempt to answer the question how long does the information in a research paper remain current, by measuring the number of citations the paper received since publication (Cunningham & Bocock, 1995). There are two approaches to measure obsolescence: "synchronous" and "diachronous" distribution (Nakamoto, 1988). They are also referred to as "citations from" and "citations to" approaches (Redner, 2005), or "retrospective citation" and "prospective citation" approaches

(Burrell, 2002; Glänzel, 2004). The former considers the age distribution of references of a paper in a particular year, while the latter analyzes the distribution of citations over time.

A number of metrics has been proposed, from a synchronous perspective, to measure obsolescence of scientific literature. "Half-life" was described (Burton & Kebler, 1960) as "half the active life", which means the time during which one-half of the currently active literature was published. Price (1970) suggested the percentage of references (from all articles) up to five years old as an index to reveal obsolescence of scientific documents, which is also named "Price Index".

From a diachronous perspective, a citation curve (Garfield, 1989; Avramescu, 1979; Li et al., 2014) is the time distribution of citations a paper received. It is also referred to as "life-cycle" (Cunningham & Bocock, 1995), "citation patterns" (Li & Ye, 2014; Wang, Song, & Barabási, 2013; Guo & Suo, 2014; Redner, 2005), or "citation history" (Redner, 2005; ABT, 1981; Persson, 2005; Vlachý, 1985; Costas et al., 2010). A "typical citation curve" describes the history of an article which received a few citations in the first following years after publication, then rose to a citation peak, but afterwards was gradually less cited with time. It is identified that lognormal function best fits typical citation curves (Egghe & Rao, 1992). For most scientific papers, death (no longer being cited by other papers) comes within ten years after publication (Price, 1976). Nevertheless, the minority appears exponential increase in citations in a long time, whose citation curves fit exponential function (Li & Ye, 2014).

The peaking time of citations features the shape of citation curves, reflecting the immediacy of publications. Some articles were noticed immediately after publication but ignored very soon, and hence were named as "flashes in the pan" (van Dalen & Henkens 2005; Costas et al., 2010). Their citations peaked much earlier than typical citation curves. Some went unnoticed for a long time and then, almost suddenly, received a lot of citations, and hence were referred to as "sleeping beauties" (van Raan, 2004), "premature discoveries" (Stent, 1972; Wyatt, 1975), "resisted discoveries" (Barher, 1961) or "delayed recognition" (Cole, 1970). Their citations peaked much later than typical citation curves. Van Raan (2004) suggested three criteria for distinguishing sleeping beauties: (1) they deeply slept (receive at most 1 citation per year on average), or less deeply slept (between 1 and 2 citations per year on average) for a few years after publication; (2) they slept at least five years; and (3) they were awakened by over 20 citations during the four years following the sleeping period. However, the criteria are not always applicable to answer Garfield (1980)'s question how abrupt a citation boost must be to suggest delayed recognition. Moreover, the criteria ignored the citations received after the awakening period (Li, 2014; Li & Ye, 2012).

Different from van Raan's average-based criteria, Costas et al. (2010) used quartiles. They identified the year after publication in which the document received for the first time at least 50% of its citations ("Year 50%"), then calculated, for all documents of the same year of publication in the same field, the percentiles 25 and 75 of the distribution function of the value of "Year 50%", and recorded them as "P25" and "P75". As a result, the articles were categorized into "flashes in the pan" ("Year 50%" <"P25"), "delayed recognition" ("Year 50%" >"P75") and the rest as "normal publications" ("P25"≤"Year 50%"≤"P75"). These criteria considered the whole citation history of articles rather than only sleeping and awakening periods, and avoided the deficiency of van Raan's definitions. However, the excessive percentages of early and delayed recognition identified by these criteria caused the originally rare phenomena normal.

Methodology

Design of the obsolescence vector

Suppose there are seven ten-year old articles whose citation curves are drawn in Figure 1. P_1 is a sleeping beauty who deeply slept for six years (received no citations) but was suddenly awakened by 40 citations in the following four years. P_2 is a flash in the pan, which immediately received 32 citations within the first two years after publication, but was ignored afterwards and rarely received citations. P_3 is a typical citation curve, which reached citation-peak in the fourth year. It was successfully fitted by the lognormal function in the program OriginPro 8 ($R^2 = 0.972$). P_4 is an article whose citations increase exponentially. Exponential function successfully fits the curve with $R^2 = 0.983$. Both P_5 and P_6 are waveform curves, but they have different initial values, hence have distinct normalized curves in Figure 1. P_7 is a horizontal line, and coincides with the 45 degree diagonal in the right side of Figure 1, which is called "the line of equality" and indicates absolutely even distribution.

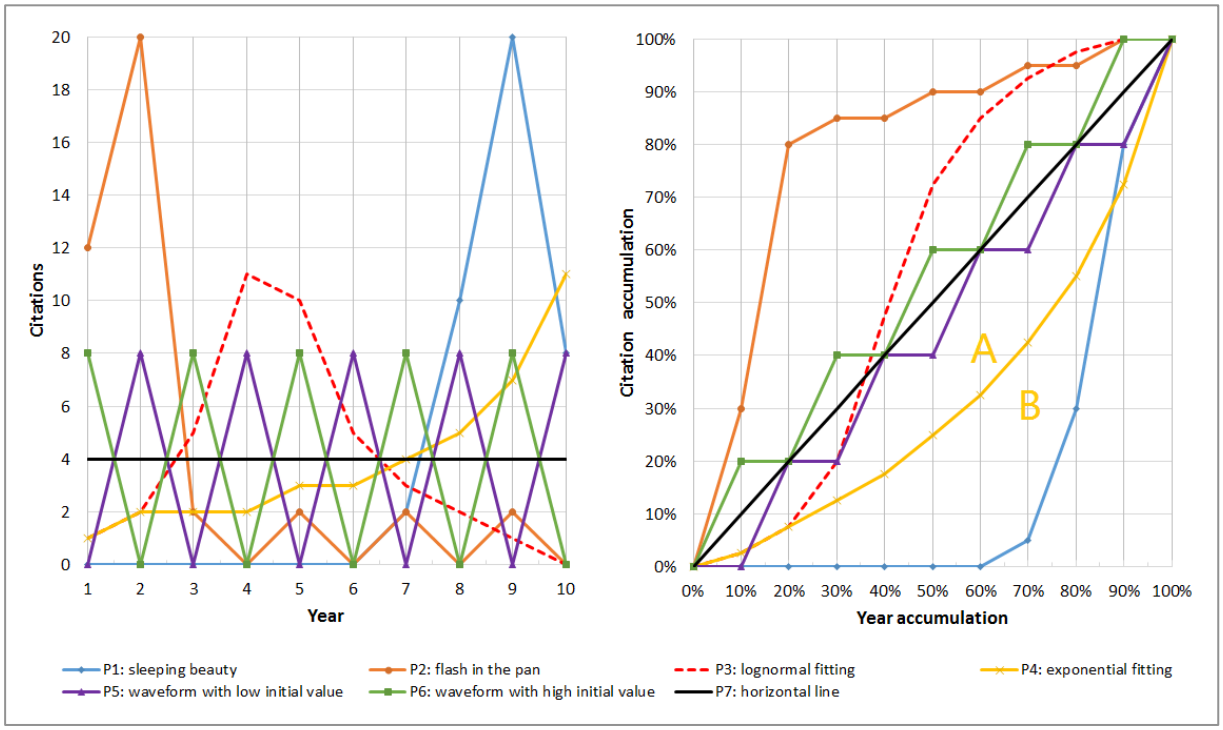


Figure 1. From citation curves to normalized cumulative citation curves of P1-P7 (left: citation curves; right: normalized cumulative citation curves).

The value of G_s, taking P4 as an example, equals to the ratio of the area that lies between the line of equality and the normalized cumulative citation curve (marked A in Figure 1) over the total area under the line of equality (sum of A and B), i.e.,

$$G_s = \frac{A}{A+B}. \quad (1)$$

The normalized cumulative citation curve (hereafter "normalized curve") of P4 is a "Lorenz curve", because the sequence of citations is in an ascending order. Since the areas A and B form an isosceles right triangle, we have

$$A + B = \frac{1}{2}. \quad (2)$$

Thus, putting Eq. (2) into Eq. (1), we have

$$G_s = 2A. \quad (3)$$

The calculation of G_s is determined by the calculation of the area B which can be divided into several trapeziums and a triangle. In this study, we remain the expression of the segment function of G_s in our previous study (Li et al., 2014),

$$G_s = \begin{cases} 1 - \frac{2\times[n\times c_1+(n-1)\times c_2+\cdots+c_n]-C}{C\times n}, & C > 0 \\ 1, & C = 0 \end{cases} \quad (4)$$

but redefine the parameters. In the new definition, n is the age of a paper, C is the total number of citations the paper received during the n years, and $c_i (i \in \{1,2,\cdots,n\})$ is the number of citations the paper received in the i^{th} year after publication. Here, $Gs \in (-1,1]$ and depends on the age (n) of articles. The value of G_s gradually approaches to -1, if the article no longer receives citations.

The value of G_s, to certain extent, characterizes the shape of citation curves:

(1) large G_s indicates delayed recognition, while small G_s denotes early recognition, as P_1 and P_2 shown in Table 1;

(2) $G_s < 0$ implies that there exists leaping early in citation curves, for example, both P_2 and P_6 received a large number of citations immediately after publication, while P_3 has a fast rising period although it does not have immediacy; and

(3) $G_s = 0$ suggests a horizontal citation curve (as P_7), or a citation curve including at least one high-citation period (to guarantee $A^- < 0$) which is offset by at least one low-citation period.

The value of A is not always positive. For P_2, $A<0$, since its normalized curve in Figure 1 is above the line of equality. Since

$$A = A^+ + A^- , \quad (5)$$

putting Eq. (5) into Eq. (3), we have

$$A^- = \frac{1}{2}G_s - A^+. \quad (6)$$

A^+ is the area between the line of equality and the normalized curve under the line of equality. Similar to the calculation of G_s, we calculate A^+, and accordingly have the value of A^-. In case of P_3, the intersection of the normalized curve and the line of equality in Figure 1 exists in between the accumulation year 30% and 40%. Therefore, there is a minor error (a difference) between the output and target of A^+ values of P_3. In cases of P_1, P_4 and P_5, there is no error in the calculation of A^+.

The fast rising period of a citation curve is hidden from the value of G_s if $A^- < 0 < A^+$. In case of $A^+ = 0$, we have

$$A^- = A = \frac{1}{2}G_s. \quad (7)$$

Hence, the value of A^- provides complementary explanation to the shape of citation curves:

(1) recognition to the article is normal or delayed rather than early if $A^-=0$;

(2) there exists leaping in the citation curve of the article if $A^-<0$; and

(3) citation leaping appears early if $A^- = \frac{1}{2}G_s$.

We propose a vector for measuring obsolescence of scientific articles: $O=(G_s, A^-, n)$, where G_s is an index revealing the history of citations, A^- is a parameter uncovering citation leaping and age n is an adjusting parameter. We calculated the obsolescence vectors for P_1-P_7 as shown in Table 1.

Table 1. Obsolescence vectors for P1-P7.

Article	*Citation curve*	*Citations*	*A*	A^+	*Obsolescence vector*		
					G_s	A^-	*n*
P1	Sleeping beauty	40	0.335	0.335	0.670	0.000	10
P2	Flash in the pan	40	-0.300	0.000	-0.600	-0.300	10
P3	Lognormal fitting	40	-0.075	0.028	-0.150	-0.103	10
P4	Exponential fitting	40	0.183	0.183	0.365	0.000	10
P5	Waveform with low initial value	40	0.050	0.050	0.100	0.000	10
P6	Waveform with high initial value	40	-0.050	0.000	-0.100	-0.050	10
P7	Horizontal line	40	0.000	0.000	0.000	0.000	10

Criteria for categorizing the patterns of obsolescence

In this research, we use the terms "flashes in the pan", "sleeping beauties" and "normal articles" as the patterns of obsolescence, but provide three different approaches for measurement, in order to characterize obsolescence vector. We remain van Raan's average-based criteria in the first approach. By following the criteria, we define variables for "flashes in the pan": "noticed" (van Dalen and Henkens, 2005) as receiving over 10 citations, "ignored" as receiving less than two citations per year on average and "immediately" as within two years since publication. We also define the duration of light disappearing for at least five years, since a flash is likely to reappear. Then, we suggest average-based criteria as follows:

flashes in the pan (F_1): articles which received more than 10 citations in the first two years since publication, and then in the next five years received no more than 2 citations per year on average;

sleeping beauties (S_1): articles which received no more than 2 citations per year on average in the first five years since publication, and then in the next four years received more than 20 citations; and

normal articles (N_1): which neither satisfy the criteria for F_1 nor for S_1.

The second approach uses quartiles. We adjust "relative ranking in a field" in Costas et al. (2010) to "relative age", since the former requires the population of articles in a filed which involves a huge dataset. Thus, for a single article, we record the percentiles 25 and 75 of its age as "A25" and "A75". Then, we define quartile-based criteria for the patterns of obsolescence as follows:

flashes in the pan (F_2): articles that reached "Year 50%" within 25% of its age, i.e., "Year 50%" <"A25";

sleeping beauties (S_2): articles that reached "Year 50%" with the time exceeding 75% of its age, i.e., "Year 50%" >"A75"; and

normal articles (N_2): which neither satisfy the criteria for F_1 nor for S_1, i.e., "A25"≤"Year 50%"≤ "A75".

Based on the obsolescence vectors of the seven cases in Table 1, we propose new criteria for categorizing the patterns of obsolescence as follows,

flashes in the pan (F_3): $G_s \leq$ -0.6 and $A^- = \frac{1}{2} G_s$;

sleeping beauties (S_3): $G_s \geq 0.6$ and $A^- = 0$; and

normal articles (N_3): which neither satisfy the criteria for F_3 nor for S_3.

Data

A dataset was prepared to make comparisons of the above three sets of criteria, and to verify the efficiency of the proposed obsolescence vector. From the Web of Science, we collected 58,963 articles of 629 Nobel Prize winners during the period of 1901-2012, in the fields of Chemistry, Physics, Physiology or Medicine, and Economic Sciences. The definition S_2 requires that a sleeping beauty should have more than 20 citations. For the purpose of comparisons, we eliminated articles, which received no more than 20 citations, and remained a collection of 28,340 articles published between 1900 and 2000. Then, we searched the number of annual citations to these articles up to 2011 in the Web of Science. Thus, every article in this collection aged at least eleven, which is sufficient for a sleeping beauty with the shortest sleeping period to be awakened.

Results

Obsolescence vector as an alternative to average-based and quartile-based criteria

The life-cycles of most articles in the dataset have already drawn to their close. As shown in Table 2, the peak of G_s distribution appears in the interval (-0.4,-0.2] and the values of G_s for 84.3% articles are negative. Moreover, 95.0% of the articles have $A^-<0$. Small G_s values (minus) indicate the end of cife-cycles, as shown by article P_2 in Figure 1. It is calculated that 68.4% of the articles with $G_s > 0$ have $A^- < 0$. Thus, there are only a small fraction of citation curves having the shape of P_1, P_4 and P_5 in Figure 1. What they have in common is that there is no citation rise and fall in the initial stage of citation curves. The rise and fall of citations must be a citation leaping or like a lognormal shape. Articles with the largest and smallest G_s values are categorized into sleeping beauties (S_3) and flashes in the pan (F_3), respectively. The obsolescence vector for the former (Rayleigh, 1914) is O = (0.892, 0, 98). Although published as early as in 1914, it received no citations until 1992. It does not satisfy S1, since it was not awakened by more than 20 citations within four years after sleeping period. However, it satisfies S_2, since recognition to it was delayed to the last four years of its age. This example reveals the deficiency of S_1. The latter (Ryle & Bailey, 1968) has an obsolescence vector O = (-0.960, -0.480, 44). The article received 26 citations immediately in the publication year, but the number rapidly fell to zero four years later and it was never cited till the end. It satisfies both F_1 and F_2.

Table 2. Comparisons of the three approaches to measuring obsolescence.

G_s	N	$N(A^-<0)$	F_1	S_1	F_2	S_2	F_3	S_3	$F_1\&F_3$	$F_2\&F_3$	$S_1\&S_3$	$S_2\&S_3$
(-1,-0.8]	494	494	41	0	489	0	265	0	34	262	0	0
(-0.8,-0.6]	3,897	3,897	62	6	3,856	0	1,734	0	57	1,704	0	0
(-0.6,-0.4]	6,808	6,808	30	16	5,250	0	0	0	0	0	0	0
(-0.4,-0.2]	7,213	7,213	21	22	985	0	0	0	0	0	0	0
(-0.2,0]	5,477	5,477	7	25	25	0	0	0	0	0	0	0
(0,0.2]	2,894	2,344	7	27	0	15	0	0	0	0	0	0
(0.2,0.4]	1,140	543	5	26	0	228	0	0	0	0	0	0
(0.4,0.6]	348	141	2	7	0	304	0	0	0	0	0	0
(0.6,0.8]	65	17	1	1	0	65	0	37	0	0	1	37
(0.8, 1)	4	0	0	0	0	4	0	3	0	0	0	3
Total	**28,340**	**26,934**	**176**	**130**	**10,605**	**616**	**1,999**	**40**	**91**	**1,966**	**1**	**40**

It seems that the condition $G_s \leq$ -0.6 and $A^- = \frac{1}{2} G_s$ for flashes in the pan is a loose condition, since it yields 1,999 flashes in the pan in the dataset. If it is intensified to be $G_s \leq$ -0.8 and $A^- = \frac{1}{2} G_s$, the number of flashes in the pan shrinks to 262, closer to the result of criterion F_1. Considering that 81.6% of the articles aged over 20, we suggest the criterion for flashes in the pan be $G_s \leq$ -0.8 and $A^- = \frac{1}{2} G_s$ on condition that n≥20.

The criterion S_3 for sleeping beauties is more stringent than S_1 and S_2, and selected only 40 qualified articles from the dataset. The 40 articles is a subset of the collection by S_2, but covers 39 articles out of the collection by S_1. In Table 2, there are six articles satisfying S_1 whose G_s values exist in the interval (-0.8, -0.6]. For example, the article in Figure 2 received only nine citations within the first five years after publication, but suddenly received 25 citation in the following four years. It also satisfies S_2, since it reached "Year 50%" within ten years (13.9% of its age) after publication. Nevertheless, this article is more like a "typical citation curve" which spent seven years to gradually reach citation-peak and slowly declined to death afterwards. The obsolescence vector of this article is O = (-0.648, -0.324, 72) which does not satisfy S_3. Moreover, we identified 3,897 articles of its kind, which have $G_s \in$ (-0.8, -0.6]. Therefore, it is more reasonable to categorize it as a "normal article" rather than a "sleeping beauty".

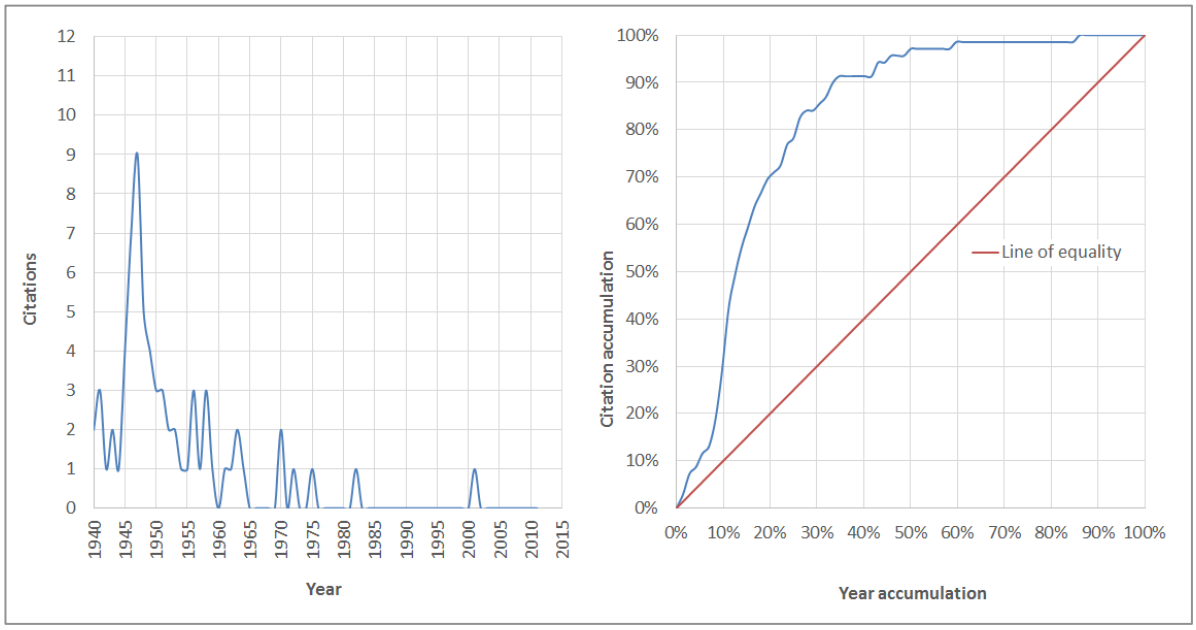


Figure 2. A sleeping beauty by average-based and quartile-based criteria, but a normal article by obsolescence vector (Landsteiner, 1940).

Citation-curve differences of obsolescence

The calculation of G_s values, sometimes, remains citation leaping under cover. As shown in Figures 3, Zewail's and Corey's articles were published in the same year of 2000, and have the same G_s values 0.083. However, they received different citations and have different citation curves. The obsolescence vector of the two articles are O=(0.083, 0, 12) and O=(0.083, -0.004, 12), respectively. Due to the citation leaping since 2007, the normalized curve of Corey's article in Figure 3 surpassed the line of equality in 2010 and yielded $A^- < 0$ which does not appear in the normalized curve of Zewail's article. Therefore, it is a sign of citation leaping to have A^-<0.

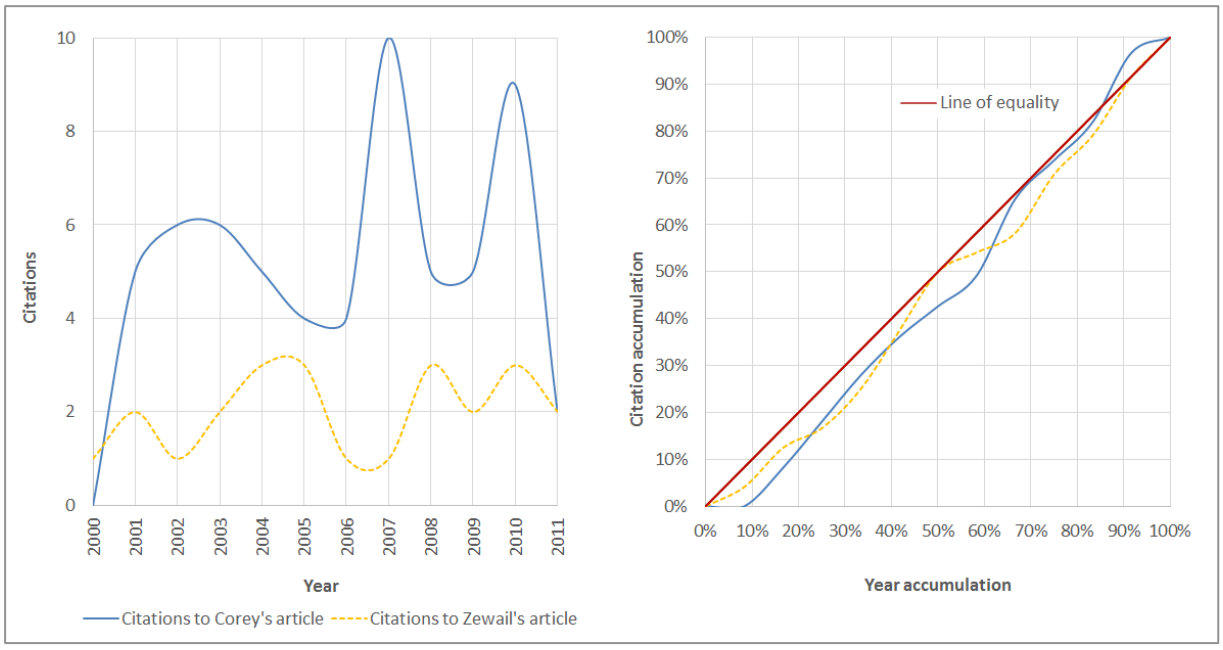


Figure 3. Zewail's article with O = (12, 0.083, 0) and Corey's article with O = (12, 0.083, -0.004).

Age differences of obsolescence

The years of 1950, 1990 and 2000 were selected for the publication years for sampling articles, in order to explore age differences of obsolescence. They were aged 62, 22 and 12, respectively. It appears that older articles have smaller G_s values while younger ones have larger G_s values. It is clear in Table 3 that the peak of G_s distribution among the intervals shifted from (-0.6, -0.4] in 1950, to (-0.4,-0.2] in 1990, even to (-0.2, 0] in 2000. Most of the old articles have been ignored and receive rare or no citations after recognition, similar to the example in Figure 2. Therefore, their G_s values gradually decline. It is hence identified that age exerts significant influence on the values of G_s.

Table 3. Age differences of obsolescence.

G_s	*Year* 1950		*Year* 1990		*Year* 2000	
	N	*N*(A^-<0)	*N*	*N*(A^-<0)	*N*	*N*(A^-<0)
[-1,-0.8]	11	11	12	12	0	0
(-0.8,-0.6]	65	65	45	45	8	8
(-0.6,-0.4]	66	66	190	190	31	31
(-0.4,-0.2]	42	42	250	250	81	81
(-0.2,0]	28	28	148	148	216	216
(0,0.2]	22	16	80	68	173	117
(0.2,0.4]	8	3	27	9	46	10
(0.4,0.6]	6	0	5	2	8	1
(0.6,0.8]	0	0	0	0	0	0
(0.8, 1]	0	0	0	0	0	0
Total	**248**	**231**	**757**	**724**	**563**	**464**

Disciplinary differences of obsolescence

The obsolescence of economic sciences is slower than that of fundamental sciences, including chemistry, physics and physiology & medicine. It is a sign of slow obsolescence to have more positive G_s values and less $A^- < 0$. In Table 4, the distribution of G_s values of economic sciences peaked in the interval (0, 0.2], while in other disciplines, it peaked in the interval (-0.4,-0.2] or (-0.6,-0.4]. The percentage of $A^- < 0$ in positive G_s values is only 50.4%, far less

than 69.8-75.8% in fundamental sciences. Moreover, older articles tend to have higher absolute G_s values, in each of the four disciplines.

Table 4. Disciplinary differences of obsolescence

G_s	*Chemistry*			*Physics*			*Physiology & Medicine*			*Economic sciences*		
	N	*N(A-<0)*	*Age*	*N*	*N(A-<0)*	*Age*	*N*	*N(A-<0)*	*Age*	*N*	*N(A-<0)*	*Age*
[-1,-0.8]	34	34	56.1	124	124	36.4	336	336	51.0	0	0	0.0
(-0.8,-0.6]	625	625	49.8	653	653	35.1	2,615	2,615	45.9	4	4	38.3
(-0.6,-0.4]	1,727	1,727	41.4	1,185	1,185	33.2	3,850	3,850	41.0	44	44	36.2
(-0.4,-0.2]	2,690	2,690	37.5	1,212	1,212	35.0	3,193	3,193	36.2	118	118	36.8
(-0.2,0]	2,236	2,236	35.3	1,008	1,008	34.6	1,972	1,972	30.7	263	263	35.6
(0,0.2]	1,099	926	39.3	576	483	42.2	730	594	34.5	489	341	30.0
(0.2,0.4]	307	161	53.9	289	180	58.9	155	78	49.8	389	124	28.2
(0.4,0.6]	67	34	71.1	147	63	71.9	33	13	60.4	101	31	37.2
(0.6,0.8]	10	3	90.5	38	10	86.9	5	0	47.2	12	4	52.3
(0.8, 1]	0	0	0.0	4	0	90.0	0	0	0.0	0	0	0.0
Total	**8,795**	**8,436**		**5,236**	**4,918**		**12,889**	**12,651**		**1,420**	**929**	

Discussion

Further discussion on $A^- < 0$

Significant citation leaping is likely to result in recurring appearance of $A^-<0$ area. For example of Hsu et al.'s article (1997), citation leaping appeared twice in the citation curve. The first citation peak appeared in 1998, the second year after publication, which led the normalized curve to reach the line of equality. In 1999, the article received six citations. The normalized curve hence surpassed the line of equality. However, the citation leaping disappeared afterwards, and the normalized curve dropped under the line of equality. Nevertheless, the second citation peak, higher than the first one, appeared in 2005 and boosted the normalized curve above the line of equality again. Comparing this example with the supposed waveform citation curves, i.e., P_5 and P_6 in Figure 1, it is identified that the appearance of $A^-<0$ area is originated by citation leaping. Furthermore, double appearance of $A^-<0$ area indicates double citation leaping in which the first one happened immediately after publication and the second one is higher. However, the characteristics of double or multiple appearance of $A^-<0$ area are not in consideration of the new designed obsolescence vector, since the number of this kind is rare.

Limitations

The obsolescence vector cannot differentiate two citation curves if there is multiplier relationship between their annual citations. For example, both (0, 8, 0, 8, 0, 8, 0, 8, 0, 8) and (0, 4, 0, 4, 0, 4, 0, 4, 0, 4) have the same obsolescence vector O=(0.1, 0, 10). The obsolescence vector is applicable to categorize articles into "flashes in the pan", "sleeping beauties" or "normal articles", by distinguishing citation leaping in citation curves. It does not characterize citation history of "normal" articles, which account for a large percent. As normal articles, P_3-P_6 in Figure 1 have entirely different obsolescence patterns. However, they cannot be uncovered by obsolescence vector.

It is controversial whether someone who won a major prize has received increased citations on all his/her work (Hugget, 2013; Mazloumian et al., 2011). However, the results are generalized from articles of Nobel laureates rather than randomly sampled authors, and hence are potentially biased. In addition, "recognition" is referred to as a large number of citations,

e.g., 20. Thus, whether the obsolescence vector is applicable to articles receiving less than 20 citations requires further research.

Conclusions

We proposed a vector for measuring obsolescence of scientific articles, $O = (G_s, A^-, n)$, where n is the age of an article, G_s and A^- are parameters for the shape of the article's citation curves. By distinguishing inequality of citation distribution, obsolescence vector is applicable to categorize articles into three general types:

flashes in the pan: $G_s \leq$ -0.8 and $A^- = \frac{1}{2} G_s$ for n≥20 or $G_s \leq$ -0.6 and $A^- = \frac{1}{2} G_s$ for n<20;
sleeping beauties: $G_s \geq 0.6$ and $A^- = 0$; and
normal articles: which neither satisfy the criteria for F_3 nor for S_3.

The age, subject category and citation curve of articles exert significant influence on G_s values. Older articles tend to have higher absolute G_s values. The criterion for "flashes in the pan" is adjustable in terms of the age of articles. In case of articles younger than, e.g., ten years old, as shown in Figure 1, it is feasible to mildly adjust the criterion as $G_s \leq$ -0.6. Disciplinary differences exist in the proposed obsolescence vector. Articles in economic sciences appear higher G_s values than those in fundamental sciences, including chemistry, physics and physiology & medicine. In case of articles receiving no more citations, their G_s values tend to decline, till to -1.

As an alternative to average-based and quartile-based criteria, the obsolescence vector avoided overlooking the period after sleeping beauties being awakened, and tightened the loose conditions by using quartiles. By obsolescence vectors, we identified 265 flashes in the pan (approximately 1%) and 40 sleeping beauties (approximately 0.1%), among 28,340 articles of Nobel laureates, which receive more than 20 citations by the year of 2011. The low percentages of flashes in the pan and sleeping beauties remained them rare phenomena.

Acknowledgement

This research was financially supported by the National Natural Science Foundation of China (NSFC No. 71203193 and 71273125).

References

ABT, H. A. (1981). Long-term citation histories of astronomical papers. *Publications of the Astronomical Society of the Pacific*, *93*, 207-210.

Avramescu, A. (1979). Actuality and obsolescence of scientific literature. *Journal of the American Society for Information Science, 30*(5), 296-303.

Burrell, Q. L. (2002). The nth-citation distribution and obsolescence. *Scientometrics*, *53*(3), 309-323.

Burton, R. E., & Kebler, R. W. (1960). The "half-life" of some scientific and technical literatures. *American Documentation*, *11*(1), 18-22.

Cole, S. (1970). Professional standing and the reception of scientific discoveries. *American Journal of Sociology*, *76*, 286–306.

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 and Technology*, *61*(2), 329–339.

Cunningham, S. J., & Bocock, D. (1995). Obsolescence of computing literature. *Scientometrics*, *34*(2), 255-262.

Egghe, L.,, & Rao, I. K. R. (1992). Citation age data and the obsolescence function: Fits and explanations. *Information and Processing Management, 28*(2), 201-217.

Garfield, E. (1980). Premature discovery or delayed recognition-why? *Current Contents*, *4*, 488-493.

Garfield, E. (1989). More delayed recognition. Part 1. Examples from the genetics of color blindness, the entropy of short-term memory, phosphoinositides, and polymer rheology. *Current Contents*, *38*, 3-8.

Guo, J. L., & Suo, Q. (2014). Comment on" Quantifying Long-term Scientific Impact". *Science*, *345*(6193), 149.

Hugget, S. (2010). Does a Nobel Prize lead to more citations. *Research Trends*, Retrieved April 7, 2015 from http: //www.researchtrends.com/issue20-november-2010/does-a-nobel-prize-lead-to-more-citations.

Hsu, C. P., Song, X.,, & Marcus, R. A. (1997). Time-dependent Stokes shift and its calculation from solvent dielectric dispersion data. T*he Journal of Physical Chemistry B*, *101*(14), 2546-2551.

Li, J. (2014). Citation Curves of "All-elements-sleeping-beauties": "Flash in the Pan" first and then "Delayed Recognition". *Scientometrics*, *100*(2), 595-601.

Li, J., & Ye, F. Y. (2012). The phenomenon of all-elements-sleeping-beauties in scientific literature. *Scientometrics*, *92*(3), 795–799.

Li, J., & Ye, F. Y. (2014). A Probe into the Citation Patterns of High-quality and High-impact Publications. *Malaysian Journal of Library and Information Science*, *19*(2), 31-47.

Li, J., Shi, D., Zhao, S. X., & Ye, F. Y. (2014). A study of the "heartbeat spectra" for "sleeping beauties". *Journal of Informetrics*, *8*(3), 493-502.

Mazloumian, A., Eom, Y., Helbing, D., Lozano, S., & Fortunato, S. (2011). How citation boosts promote scientific paradigm shifts and nobel prizes. *Plos One*, *6*(5), e18975.

Nakamoto, H. (1988). Synchronous and dyachronous citation distributions. In L. Egghe, & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 157–163). Amsterdam: Elsevier Science Publishers.

Persson, O. (2005). "Citation Indexes for Science"-A 50 year citation history. *Current Science*, *89*(9), 1503-1504.

Price, D. (1970). Citation measures of hard science, soft science, technology, and non-science. In C. E. Nelson & D. K. Pollock (Eds.), *Communication among Scientists and Engineers* (pp. 3-22). Lexington, MA: Heath.

Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306.

Rayleigh, L. (1914). On the theory of long waves and bores. Proceedings of the royal society of London series A- Containing papers of a mathematical and physical character, 90(619), 324-328.

Redner, S. (2005). Citation statistics from more than a century of physical review. *Physics Today*, *58*(1), 49–54.

Stent, G. S. (1972). Prematurity and uniqueness in scientific discovery. *Scientific American*, *227*(6), 84–93.

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.

van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–472.

Vlachý, J. (1985). Citation histories of scientific publications. The data sources. *Scientometrics*, *7*(3), 505-528.

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127-132.

Wyatt, H. V. (1961). Knowledge and prematurity-journey from transformation to DNA. *Perspectives in Biology and Medicine, 18*(2), 149-156.

Field-Normalized Citation Impact Indicators and the Choice of an Appropriate Counting Method

Ludo Waltman and Nees Jan van Eck

{waltmanlr, ecknjpvan}@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

Abstract
Bibliometric studies often rely on field-normalized citation impact indicators in order to make comparisons between scientific fields. We discuss the connection between field normalization and the choice of a counting method for handling publications with multiple co-authors. Our focus is on the choice between full counting and fractional counting. Based on an extensive theoretical and empirical analysis, we argue that properly field-normalized results cannot be obtained when full counting is used. Fractional counting does provide results that are properly field normalized. We therefore recommend the use of fractional counting in bibliometric studies that require field normalization, especially in studies at the level of countries and research organizations.

Conference Topic
Citation and co-citation analysis; Indicators

Introduction
In discussions on bibliometric indicators, two topics that receive a considerable amount of attention are field normalization and counting methods. Field normalization is about the problem of correcting for differences in citation practices between scientific fields. The challenge is to develop citation-based indicators that allow for valid between-field comparisons. Counting methods are about the way in which co-authored publications are handled. For instance, if a publication is co-authored by two countries, should the publication be counted as a full publication for each country or should it be counted as half a publication for each country?

The topics of field normalization and counting methods are usually discussed separately from each other. However, we argue that there is a close connection between the two topics. Our argument is that proper field normalization is possible only if a suitable counting method is used. In particular, we claim that properly field-normalized results cannot be obtained when one uses the popular full counting method, in which co-authored publications are fully assigned to each co-author. The fractional counting method, which assigns co-authored publications fractionally to each co-author, does provide properly field-normalized results. The problem of full counting basically is that co-authored publications are counted multiple times, once for each co-author, which creates a bias in favor of fields in which there is a lot of co-authorship and in which co-authorship correlates with additional citations. This is the essence of the argument that we present in this paper. Our argument builds on an earlier paper (Waltman et al., 2012), but in the present paper we elaborate the argument in more detail and we also present an extensive empirical analysis.

This paper is a shortened version of a more extensive working paper (Waltman & Van Eck, 2015). The working paper includes additional empirical analyses comparing different counting methods at the level of institutions and countries. Furthermore, the working paper considers different variants of fractional counting and also studies first author and corresponding author counting methods.

Counting methods

Our focus is on the comparison between full counting and fractional counting. In the case of full counting, a publication is fully assigned to each co-author. For instance, a publication co-authored by four countries counts as a full publication for each of the four countries. In the fractional counting case, a publication is fractionally assigned to each co-author. The weight with which a publication is assigned to a co-author indicates the share of the publication allocated to that co-author. The sum of the weights of all co-authors of a publication equals one. An example of fractional counting is the situation in which a publication co-authored by four countries is assigned to each country with a weight of 1 / 4 = 0.25.

There is a quite extensive literature on counting methods. Because of space limitations, we mention only a few selected studies. A systematic terminology for counting methods is proposed by Gauffriau, Larsen, Maye, Roulin-Perriard, and Von Ins (2007). They refer to full counting as whole counting and to fractional counting as normalized counting. Gauffriau, Larsen, Maye, Roulin-Perriard, and Von Ins (2008) present a comparison of counting methods at the country level. They also provide an overview of earlier literature on counting methods. Another country-level comparison is reported by Aksnes, Schneider, and Gunnarsson (2012). At the institution level, Waltman et al. (2012) present a comparison between full and fractional counting. Interesting work on counting methods can also be found in various papers by Ruiz-Castillo and colleagues, who propose the idea of a so-called multiplicative counting method (e.g. Albarrán, Crespo, Ortuño, & Ruiz-Castillo, 2010).

Relation between counting methods and field normalization

Our aim in this section is to demonstrate the close connection between counting methods and field normalization. In particular, we aim to make clear that full counting is fundamentally inconsistent with the idea of field normalization. We argue that full counting yields results that suffer from a bias in favor of fields in which there is a lot of co-authorship and in which co-authorship correlates with additional citations. This bias is caused by the fact that co-authored publications are counted multiple times in the case of full counting, once for each co-author.

We present our argument by providing two simple examples. Both examples take countries as the unit of analysis and focus on the mean normalized citation score (MNCS) indicator (Waltman, Van Eck, Van Leeuwen, Visser, & Van Raan, 2011). However, the underlying ideas of the two examples are more general, and similar examples can be given with authors or organizations as the unit of analysis and with other field-normalized indicators.

Table 1. Example involving a single field.

	Authors	No. of cit.	Norm. cit. score
Publication 1	Country A	3	0.6
Publication 2	Country A	6	1.2
Publication 3	Country B	1	0.2
Publication 4	Country A; Country B	10	2.0

Example involving a single field

We consider a world in which there are just four publications. These publications have been produced by two countries, labeled as country A and country B. Table 1 shows for each publication the countries by which the publication is authored and the number of citations the publication has received. The table also shows the normalized citation score of each publication. For simplicity, it is assumed that all four publications are in the same field. The normalized citation score of a publication is therefore obtained simply by dividing the number of citations of the publication by the average number of citations of all four publications. The

average number of citations of the four publications equals (3 + 6 + 1 + 10) / 4 = 5, and therefore the normalized citation score of for instance publication 1 equals 3 / 5 = 0.6. Of course, the average of the normalized citation scores of the four publications equals one.
We now calculate both for country A and for country B the MNCS. Using full counting, we obtain

$$\mathrm{MNCS}_\mathrm{A} = \frac{0.6 + 1.2 + 2.0}{3} = 1.27 \text{ and } \mathrm{MNCS}_\mathrm{B} = \frac{0.2 + 2.0}{2} = 1.10.$$

On the other hand, using fractional counting, we get

$$\mathrm{MNCS}_\mathrm{A} = \frac{1.0 \times 0.6 + 1.0 \times 1.2 + 0.5 \times 2.0}{1.0 + 1.0 + 0.5} = 1.12 \text{ and } \mathrm{MNCS}_\mathrm{B} = \frac{1.0 \times 0.2 + 0.5 \times 2.0}{1.0 + 0.5} = 0.80,$$

where publication 4 has been assigned with a weight of 0.5 to country A and with a weight of 0.5 to country B.
The important thing to observe in this example is that in the case of full counting country A and country B both have an MNCS above one. One of the main ideas of field-normalized indicators such as the MNCS indicator is that the value of one can be interpreted as the world average. Under this interpretation, country A and country B both perform above the world average. Since there are no other countries in our example, the conclusion would be that all countries in the world perform above the world average. There are no countries with a below-average performance. In our opinion, the conclusion that everyone is above average does not make much sense. Moreover, this conclusion is fundamentally different from the conclusion that is reached in the case of fractional counting. Using fractional counting, country A has a performance above the world average while the performance of country B is below the world average.
Looking a bit more in detail at our example, we observe that in the fractional counting case we have

$$\frac{2.5 \times \mathrm{MNCS}_\mathrm{A} + 1.5 \times \mathrm{MNCS}_\mathrm{B}}{2.5 + 1.5} = \frac{2.5 \times 1.12 + 1.5 \times 0.80}{2.5 + 1.5} = 1.$$

Hence, the weighted average of the MNCS of country A and the MNCS of country B, with weights given by each country's fractional number of publications, equals exactly one. This is a general property of fractional counting. The weighted average of the MNCSs of all countries in the world will always be equal to exactly one.
In the full counting case, the weighted average of the MNCS of country A and the MNCS of country B equals

$$\frac{3 \times \mathrm{MNCS}_\mathrm{A} + 2 \times \mathrm{MNCS}_\mathrm{B}}{3 + 2} = \frac{3 \times 1.27 + 2 \times 1.10}{3 + 2} = 1.20,$$

where the weight of each country is given by the number of publications of the country obtained using full counting. So in the full counting case the world average at the country level does not equal one but instead equals 1.20. Taking 1.20 as the world average, we conclude that country A, with an MNCS of 1.27, has an above-average performance while

country B, with an MNCS of 1.10, performs below average. This is in agreement with the conclusion reached using fractional counting.

So in our example there is a difference of 1.20 – 1 = 0.20 between the world average obtained using full counting and the world average obtained using fractional counting. We refer to this difference as the full counting bonus. In principle, the full counting bonus can be either positive or negative, but we will see that in practice the bonus is usually positive. The full counting bonus is caused by the fact that publications co-authored by multiple countries are counted multiple times in the case of full counting, and therefore the citation impact of multi-country publications relative to single-country publications determines whether the full counting bonus is positive or negative. The bonus will be positive if publications co-authored by multiple countries receive more citations than publications authored by a single country. Conversely, a negative bonus will be obtained if multi-country publications are cited less frequently than single-country publications. As can be seen in Table 1, in our example the only publication co-authored by multiple countries is publication 4, and this is also the most highly cited publication. In the full counting case, publication 4 is fully assigned both to country A and to country B. Hence, the most highly cited publication in our example is counted two times, once for country A and once for country B. This double counting of publication 4 explains why both countries have an MNCS above one and why the full counting bonus is positive.

Example involving multiple fields

In the example discussed above, all publications are in the same field. We now consider an example that involves more than one field. This example is presented in Table 2. There are six publications, three in field X and three in field Y, and there are four countries. Countries A and B are active only in field X, while countries C and D are active only in field Y. The three publications in field X have all received the same number of citations, and therefore these publications all have a normalized citation score of one. This is not the case in field Y, in which publication 6, co-authored by countries C and D, has received more citations than publications 4 and 5, which are single-country publications. Of course, the average normalized citation score of the publications in field Y equals one, just like in field X.

Table 2. Example involving multiple fields.

	Field	Authors	No. of cit.	Norm. cit. score
Publication 1	Field X	Country A	10	1.0
Publication 2	Field X	Country B	10	1.0
Publication 3	Field X	Country A; Country B	10	1.0
Publication 4	Field Y	Country C	4	0.8
Publication 5	Field Y	Country D	4	0.8
Publication 6	Field Y	Country C; Country D	7	1.4

Using fractional counting, the four countries all have an MNCS of exactly one. For countries A and B this is immediately clear. In the case of countries C and D, the MNCS is calculated as $(1.0 \times 0.8 + 0.5 \times 1.4) / (1.0 + 0.5) = 1$. So fractional counting tells us that all four countries perform at the world average. This is indeed the outcome that we would expect to obtain. The publications of countries A and B have all been cited equally frequently as the average of their field, so countries A and B obviously perform at the world average. In the case of countries C and D, we observe that these countries have exactly the same performance and that they are the only countries active in field Y. Based on these two observations, it is natural to conclude that the performance of countries C and D is at the world average.

We now consider the full counting case. Using full counting, countries A and B have an MNCS of one, while countries C and D have an MNCS of $(0.8 + 1.4) / 2 = 1.10$. The full

counting results seem to suggest that countries C and D have a better performance than countries A and B. However, a more careful analysis shows that this is not a correct interpretation of the results. To see this, we calculate both for field X and for field Y the average of the MNCSs of the countries active in the field. The average MNCS of the countries active in field X equals one, while the average MNCS of the countries active in field Y equals 1.10. Hence, both countries A and B active in field X and countries C and D active in field Y perform at the world average of their field. Like in the fractional counting case, we conclude that all four countries have an average performance. Countries C and D have a higher MNCS than countries A and B only because they are active in a field with a higher full counting bonus. Field Y has a full counting bonus of $1.10 - 1 = 0.10$, while the full counting bonus in field X equals zero.

Conclusions based on the examples

Based on the above examples, two important conclusions can be drawn. The first conclusion is that there is a need to carefully distinguish between two field normalization concepts. We refer to these concepts as weak field normalization and strong field normalization. Weak field normalization requires the average of the normalized citation scores of all publications in a field to be equal to one. Strong field normalization is more demanding. It requires the weighted average of the MNCSs of all countries active in a field to be equal to one, where the weight of a country is given by its number of publications in the field.

As shown in the above examples, full counting yields results that are in agreement with the idea of weak field normalization, but these results may violate the idea of strong field normalization. For instance, in the first example discussed above, the average normalized citation score of the four publications equals one (weak field normalization), but the average MNCS of the two countries does not equal one (no strong field normalization). Fractional counting results, on the other hand, satisfy not only the idea of weak field normalization but also the idea of strong field normalization. Using fractional counting, the weighted average of the MNCSs of all countries active in a field will always be equal to one.

When citation-based indicators are calculated using full counting, there is a risk of misinterpretation. People may confuse the concepts of weak and strong field normalization, and they may fail to understand that the idea of strong field normalization does not apply in the case of full counting. In the second example presented above, they may for instance draw the incorrect conclusion that countries C and D perform above the world average. In the fractional counting case, people will not draw such an incorrect conclusion, because fractional counting results are in agreement with the idea of strong field normalization.

We now turn to the second conclusion that follows from our examples. The fact that full counting yields results that are incompatible with the idea of strong field normalization may in itself be regarded as just a minor issue. Instead of having a world average of one, the average of all countries in the world may for instance be equal to 1.10 or 1.20. Although a world average of one might be somewhat more convenient, the exact value of the world average may in the end seem to be of limited importance.

However, our second conclusion is that deviations of the world average from one actually do have serious consequences, at least when making comparisons between fields. This is what is shown in the second example given above. Using full counting, the average MNCS of the countries active in field X equals one, while the average MNCS of the countries active in field Y equals 1.10. So in field X the world average equals one, while in field Y we have a world average of 1.10. Direct comparisons of the MNCSs of the countries active in field X and the countries active in field Y therefore do not yield valid conclusions. Based on their MNCSs, the countries active in field Y seem to perform better than the countries active in field X, but

taking into account the fact that field Y has a higher world average than field X, it actually should be concluded that all countries perform at the same level.

Essentially, the second conclusion that we draw based on our examples is that full counting is fundamentally inconsistent with the idea of field normalization. Citation-based indicators calculated using full counting yield results that do not allow for valid comparisons between fields, and this is the case even when field-normalized indicators, such as the MNCS indicator, are used. When full counting is used in the calculation of field-normalized indicators, countries that focus their activity on fields with a high full counting bonus have an advantage over countries that are active mainly in fields with a low full counting bonus. Fractional counting does not suffer from this problem. Fractional counting results are compatible with the idea of strong field normalization, and these results therefore do allow for proper between-field comparisons.

Empirical analysis of the full counting bonus

In the previous section, we have introduced the idea of the full counting bonus and we have illustrated this idea using theoretical examples. In this section, we present a large-scale empirical analysis of the full counting bonus. This analysis for instance makes clear which fields benefit most from the full counting bonus, and the analysis shows the differences between fields caused by the bonus.

Calculation of the full counting bonus

We first explain in more detail the way in which we calculate the full counting bonus. For simplicity, we assume that our interest is in the full counting bonus at the level of countries. However, the full counting bonus can be calculated in a similar way at the level of for instance authors or organizations.

Suppose we have a set of n publications. This could be for instance the set of all publications in a specific field and in a specific year. For each publication i, we have a citation score c_i. The citation score of a publication can be defined in different ways. It may be simply the number of times a publication has been cited, but it may also be something more advanced, for instance a field-normalized citation score. We also know for each publication the countries by which the publication has been co-authored. We use m_i to denote the number of countries that have co-authored publication i.

In order to obtain the full counting bonus, we first calculate for each country the average citation score of its publications. We perform this calculation both using full counting and using fractional counting. Next, we calculate a weighted average of the average citation scores of all countries. In the case of full counting, we use the number of publications of a country obtained using full counting as the weight of the country. In the case of fractional counting, we use a country's number of publications obtained using fractional counting as the country's weight. Finally, we calculate the full counting bonus as the difference between the weighted average in the full counting case and the weighted average in the fractional counting case.

The above approach to calculating the full counting bonus is somewhat complicated. However, a mathematically equivalent but much simpler approach is available. In this approach, the full counting bonus is calculated as

$$\mathrm{FCB} = \frac{\sum_{i=1}^{n} m_i c_i}{\sum_{i=1}^{n} m_i} - \frac{\sum_{i=1}^{n} c_i}{n},$$

where the first term equals the above-mentioned weighted average in the full counting case while the second term equals the weighted average in the fractional counting case. In the first term, the citation score c_i of publication i co-authored by m_i countries is counted m_i times. This is because in the full counting case publication i is fully assigned to each of the m_i countries. In the second term, the citation score c_i of publication i is counted only once, regardless of the number of countries m_i by which publication i has been co-authored. This is because in the fractional counting case the total weight with which publication i is assigned to the m_i countries equals one.

In our empirical analysis, we consider two definitions of the citation score of a publication. Both definitions include a normalization for field. In the first definition, the citation score of a publication is obtained by dividing the number of citations of the publication by the average number of citations of all publications in the same field and in the same year. Averaging the citation scores of multiple publications then gives us the MNCS indicator. This indicator was also used in the theoretical examples presented in the previous section. In the second definition of the citation score of a publication, we determine whether a publication belongs to the top 10% most frequently cited publications of its field and publication year. A publication belonging to the top 10% has a citation score of one, while a publication belonging to the bottom 90% has a citation score of zero. When this second definition is used, averaging the citation scores of multiple publications yields the $PP_{top\ 10\%}$ indicator, where $PP_{top\ 10\%}$ stands for the proportion of top 10% publications (Waltman et al., 2012; Waltman & Schreiber, 2013). When the full counting bonus is calculated for the set of all publications in a specific field and in a specific year, the second term in the above equation for the full counting bonus will be equal to one in the case of our first definition of the citation score of a publication. This term will be equal to 0.1 (or 10%) in the case of our second definition.

Empirical results

We perform our analysis using the Web of Science (WoS) database. The analysis is based on publications in the period 2009–2010. Only publications of the WoS document types 'article' and 'review' are taken into account. A four-year citation window is used, including the year in which a publication appeared. For the purpose of the calculation of the field-normalized citation scores of publications, fields are defined by the WoS journal subject categories.

We consider three units of analysis: Authors, organizations, and countries. To determine the number of organizations and the number of countries by which a publication has been co-authored, we take into account both the regular addresses of the publication and the reprint address. The number of organizations and the number of countries of a publication is obtained by counting the number of distinct organization names and the number of distinct country names mentioned in the addresses of the publication.

The full counting bonus depends on two factors. On the one hand, it depends on the variation among publications in the number of authors, organizations, or countries. For instance, if all publications have the same number of authors, there can be no full counting bonus at the level of authors. On the other hand, the full counting bonus also depends on the relation between the number of authors, organizations, or countries of a publication and the citation score of the publication. There can for instance be no author-level full counting bonus if publications with different numbers of authors on average all have the same citation score.

Figure 1 presents the distribution of publications based on their number of authors, organizations, and countries. Not surprisingly, the figure shows that the variation among publications in the number of authors is largest while the variation among publications in the number of countries is smallest. Figure 2 presents the relation between the number of authors, organizations, and countries of a publication and the average citation score given by the MNCS indicator. In general, an increasing relation can be observed between the number of

authors, organizations, and countries of a publication and the average citation score. The relation is strongest for countries and weakest for authors. In fact, when the number of authors is between two and five, there is hardly any dependence of the average citation score of a publication on the number of authors. Publications with three or four authors on average even have a slightly lower citation score than publications with two authors. Results for the $PP_{top\ 10\%}$ are not shown, but are similar to the results for the MNCS indicator.

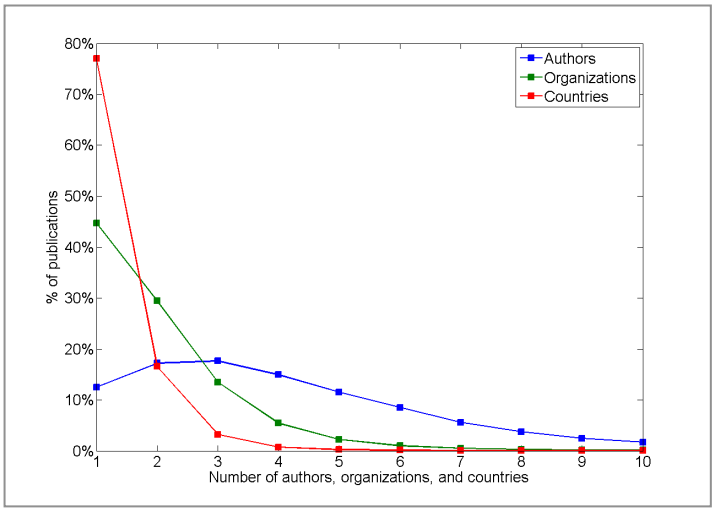


Figure 1. Distribution of publications based on their number of authors, organizations, and countries.

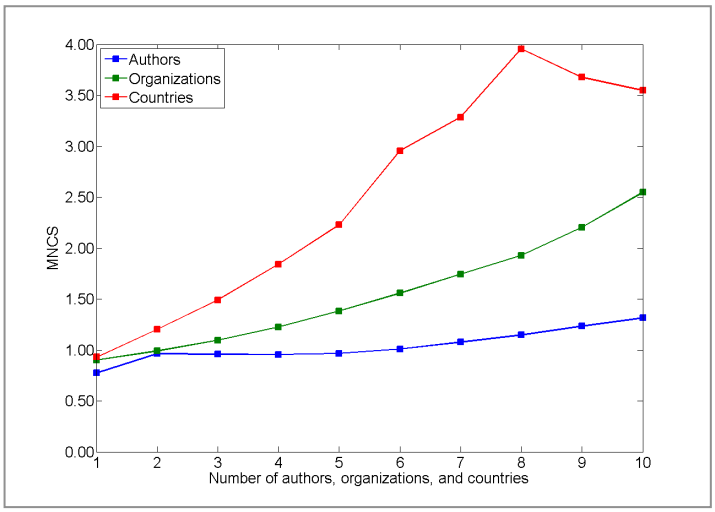


Figure 2. Relation between the number of authors, organizations, and countries of a publication and the MNCS indicator.

Figures 1 and 2 make clear that publications often have multiple co-authors and that the citation impact of a publication tends to increase with the number of co-authors. Co-authored publications are counted multiple times in the case of full counting, and our expectation based on Figures 1 and 2 therefore is to observe full counting bonuses that are positive and of significant size. This is indeed what is reported in Tables 3 and 4. The tables show the full counting bonus at the level of authors, organizations, and countries for five broad fields of science and also for all fields of science taken together. Table 3 relates to the MNCS indicator, while Table 4 relates to the $PP_{top\ 10\%}$ indicator. In order to facilitate comparison between the results obtained for the two indicators, the full counting bonus is presented as a percentage of the average value of the indicator. For instance, in the case of the MNCS

indicator, we obtain a full counting bonus of 0.248 at the level of authors for all fields of science. The average value of the MNCS indicator equals one, and therefore the full counting bonus is reported as 0.248 / 1 = 24.8% in Table 3. Likewise, the $PP_{top\ 10\%}$ indicator has an average value of 0.1 (or 10%), and therefore a full counting bonus of 0.0304 (or 3.04%) is reported as 0.0304 / 0.1 = 30.4% in Table 4.

Table 3. Full counting bonus for the MNCS indicator at the level of authors, organizations, and countries, including a breakdown into five broad fields of science.

	Authors	Organizations	Countries
All fields	24.8%	21.1%	12.6%
Biomedical and health sciences	20.9%	26.8%	16.7%
Life and earth sciences	14.7%	16.2%	12.7%
Mathematics and computer science	8.2%	8.0%	6.9%
Natural sciences and engineering	35.2%	19.3%	10.8%
Social sciences and humanities	14.7%	11.2%	5.6%

Table 4. Full counting bonus for the PPtop 10% indicator at the level of authors, organizations, and countries, including a breakdown into five broad fields of science.

	Authors	Organizations	Countries
All fields	30.4%	26.5%	17.1%
Biomedical and health sciences	24.9%	34.5%	22.6%
Life and earth sciences	22.8%	24.3%	19.7%
Mathematics and computer science	11.3%	11.3%	9.7%
Natural sciences and engineering	43.3%	20.6%	13.0%
Social sciences and humanities	21.3%	17.2%	8.3%

Based on the results for the MNCS indicator presented in Table 3, a number of conclusions can be drawn. At all three analysis levels (i.e., authors, organizations, and countries), there turns out to be a full counting bonus that is positive and of significant size. In general, the bonus is highest at the level of authors and lowest at the level of countries. We have seen in Figure 2 that the number of countries of a publication has a much stronger effect on a publication's citation score than the number of authors, but apparently this is offset by the fact that publications with a large number of countries occur much less frequently than publications with a large number of authors, as shown in Figure 1. The full counting bonus at the level of organizations is generally in between the country-level and author-level bonuses, although there are two main fields (i.e., 'Biomedical and health sciences' and 'Life and earth sciences') in which the organization-level bonus is higher than the author-level one.

The results reported in Table 3 also indicate that at the levels of authors and organizations the full counting bonus is lowest in the 'Mathematics and computer science' main field. At the country level, 'Social sciences and humanities' is the main field with the lowest bonus. The 'Natural sciences and engineering' main field has the highest bonus at the level of authors, while the highest bonus at the organization and country level can be found in the 'Biomedical and health sciences' main field.

The results for the $PP_{top\ 10\%}$ indicator reported in Table 4 are quite similar to the MNCS results presented in Table 3. However, full counting bonuses turn out to be consistently higher for the $PP_{top\ 10\%}$ indicator than for the MNCS indicator.

More detailed results at the level of 250 WoS journal subject categories can be found in an Excel file that is available at www.ludowaltman.nl/counting_methods/. The Excel file also indicates how the five main fields listed in Tables 3 and 4 are defined in terms of the WoS journal subject categories. There turn out to be rather large differences between subject categories in the full counting bonus. For instance, the subject categories with the highest MNCS full counting bonus at the level of organizations and countries are 'Medicine, general

& internal' and 'Physics, nuclear'. The subject categories have bonuses of, respectively, 148% and 176% at the organization level and 89% and 70% at the country level. Other subject categories have bonuses that are close to zero or even negative. Examples of such subject categories include 'Chemistry, organic' and 'Ergonomics'.
It is important to be aware of the consequences of the large differences between subject categories in the full counting bonus. Consider a university that has a full counting MNCS of 2.50 in the 'Medicine, general & internal' subject category and a full counting MNCS of 1.00 in the 'Chemistry, organic' subject category. What should we conclude based on these values? The obvious conclusion may seem to be that in terms of citation impact our university is performing much better in the 'Medicine, general & internal' subject category than in the 'Chemistry, organic' subject category. However, this conclusion does not take into account the effect of the full counting bonus. As mentioned above, the 'Medicine, general & internal' subject category has an organization-level full counting bonus of almost 150%, while the full counting bonus for the 'Chemistry, organic' subject category is close to zero. Taking into account the effect of the full counting bonus, we need to conclude that in both subject categories our university performs around the average level of all organizations worldwide.

Commonly used arguments in favor of full counting

In practice, most bibliometric analyses use full counting instead of fractional counting. Below we list three arguments that are often given to argue against the use of fractional counting and to justify the use of full counting. We also provide a response to each argument.

Argument 1: The different co-authors of a publication usually have not contributed equally. By giving equal weight to each co-author, fractional counting fails to properly represent the contributions made by the different co-authors. Hence, giving equal weight to each co-author is arbitrary and lacks a sound justification.

It is true that there can be large differences between co-authors in the contribution they have made to a publication. At the level of an individual publication, fractional counting may therefore significantly misrepresent the contributions made by individual co-authors. However, at the level of a large set of publications, for instance all publications of an organization or a country, we believe that it is reasonable to assume that the error will be within an acceptable margin. This is because errors at the level of individual publications are likely to cancel out. The contribution of an organization or a country to certain publications may be overestimated, but most probably there will then be other publications for which the contribution of this organization or this country is underestimated.
Furthermore, the argument that giving equal weight to each co-author of a publication is arbitrary may equally well be used as an argument against full counting. Like fractional counting, full counting gives the same weight to each co-author of a publication.

Argument 2: Fractional counting provides an incentive against collaboration, which is often considered undesirable.

We believe that citation impact and collaboration represent different dimensions of scientific performance and that in general these dimensions can best be measured separately from each other. Citation-based indicators should be assessed based on the degree to which they measure citation impact in an accurate way. In this respect, we believe that for many purposes fractional counting performs better than full counting. If in addition to citation impact one also considers collaboration to be a relevant dimension of scientific performance, then additional indicators should be used to measure this dimension. If one desires to do so, these indicators can then be used to provide an incentive to collaboration. By assessing citation-

based indicators based on the effect they may have on collaboration, one fails to make a proper distinction between the citation impact dimension of scientific performance and the collaboration dimension.

Argument 3: Fractional counting is more difficult to understand and less intuitive than full counting.

To a certain degree, we agree with this argument. Fractional counting yields non-integer publication and citation counts. These non-integer counts are more difficult to understand and require more explanation than the integer publication and citation counts provided by full counting. Fractional counting may also be less intuitive than full counting. For instance, consider a researcher who has produced some of his publications on his own while he has produced other publications with one or two co-authors. The researcher may feel that his co-authored publications are of similar importance to his oeuvre as his single-author publications. However, fractional counting gives less weight to the co-authored publications of the researcher than to his single-author publications. This is not in agreement with the feelings the researcher has about the importance of the different publications in his oeuvre, and therefore from the point of view of the researcher fractional counting can be regarded as less intuitive than full counting.

On the other hand, from a different point of view, it can also be argued that fractional counting is actually more intuitive than full counting. Earlier in this paper, we have given two examples showing that field-normalized citation impact indicators calculated using full counting can easily be misinterpreted. Field-normalized indicators calculated using fractional counting are much more easy to interpret in a correct way. As we have explained, this is because indicators based on fractional counting yield results that are compatible with the idea of strong field normalization. Unlike full counting indicators, fractional counting indicators therefore allow comparisons between fields to be performed in an easy and intuitive way. So from this point of view indicators based on fractional counting can be considered more intuitive than their full counting counterparts.

Conclusions

In this paper, we have presented a new perspective on the choice between different counting methods, leading to an important new argument in favor of fractional counting. Building on our earlier work (Waltman et al., 2012), this argument is based on the observation that the problem of choosing an appropriate counting method is closely connected to the problem of field normalization of citation-based indicators.

We have argued that from a field normalization point of view fractional counting is preferable over full counting. As we have shown, properly field-normalized results cannot be obtained using full counting, and field-normalized indicators calculated using full counting can easily be misinterpreted. Fractional counting does provide properly field-normalized results, and these results can be interpreted in a much more straightforward way than results obtained using full counting. Essentially, the problem of full counting is that co-authored publications are counted multiple times, once for each co-author, which creates an unfair advantage to fields with a lot of co-authorship and with a strong correlation between co-authorship and citations. For instance, the average full counting MNCS of all organizations or all countries active in these fields is significantly higher than one. On the other hand, fields in which co-authorship is less common or in which co-authorship does not correlate with citations are disadvantaged. Full counting yields results that are biased against organizations and countries whose activity is focused on these fields. Fractional counting does not suffer from this problem. In the case of fractional counting, each publication is counted only once, regardless

of its number of co-authors, and this ensures that comparisons between fields can be made in an unbiased way.

What are the practical implications of the analysis presented in this paper? In our view, this depends on the level of aggregation at which a bibliometric study is performed. In the case of a study at a high aggregation level, such as the level of countries or organizations (e.g., university rankings), we consider it absolutely essential to use fractional counting instead of full counting. At this level, there is a serious risk of misinterpretation of full counting results. Moreover, we believe that arguments in favor of full counting, such as the ones discussed in the previous section, are of limited relevance at a high aggregation level.

The situation is more difficult at a low level of aggregation, for instance at the level of researchers or research groups. At this level, we believe that reasonable arguments can be given in favor of both full and fractional counting. Especially the third argument discussed in the previous section plays an important role at this level. As pointed out in this argument, full counting is in agreement with the intuitive idea that all publications of a researcher or a research group should be considered of equal importance.

However, there is a more fundamental reason why the argument presented in this paper in favor of fractional counting is less relevant at a low level of aggregation. The argument depends on the connection between counting methods and field normalization, but the entire idea of field normalization may be seen as problematic at a low aggregation level. Field-normalized indicators have a limited accuracy (e.g., Van Eck, Waltman, Van Raan, Klautz, & Peul, 2013), and it is questionable whether these indicators are sufficiently accurate for applications at a low aggregation level. If the accuracy of field-normalized indicators at a low aggregation level is considered insufficient, the argument presented in this paper in favor of fractional counting has no relevance at this level.

In this paper, we have not shown how results obtained using full and fractional counting differ in practice. We refer to our working paper (Waltman & Van Eck, 2015) for an extensive comparison of full and fractional counting in bibliometric studies at the level of institutions and countries. The working paper also considers different variants of fractional counting, and it studies first author and corresponding author counting methods.

References

Aksnes, D.W., Schneider, J.W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *Journal of Informetrics*, *6*(1), 36-43.

Albarrán, P., Crespo, J.A., Ortuño, I., & Ruiz-Castillo, J. (2010). A comparison of the scientific performance of the U.S. and the European Union at the turn of the 21st century. *Scientometrics*, *85*(1), 329-344.

Gauffriau, M., Larsen, P.O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2007). Publication, cooperation and productivity measures in scientific research. *Scientometrics*, *73*(2), 175-214.

Gauffriau, M., Larsen, P.O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics*, *77*(1), 147-176.

Van Eck, N.J., Waltman, L., Van Raan, A.F.J., Klautz, R.J.M., & Peul, W.C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE*, *8*(4), e62395.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., 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., & 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. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. arXiv:1501.04431.

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

Forecasting Technology Emergence from Metadata and Language of Scientific Publications and Patents[1]

Olga Babko-Malaya, Andy Seidel, Daniel Hunter, Jason HandUber, Michelle Torrelli and Fotis Barlos

{olga.babko-malaya, andy.seidel, daniel.hunter, jason.handuber, michelle.torrelli, fotis.barlos}@baesystems.com
BAE Systems, Burlington, MA 01803

Abstract
This paper describes a multidisciplinary study and development effort to analyze full text and metadata of scientific articles and patents for indicators of new disruptive and game-changing technical breakthroughs. The system we are developing can scan millions of documents in two languages, English and Chinese, and extract meaningful trends and predictions. Whereas traditional approaches to innovation analytics rely on citation analysis to analyze impact or identify the most influential patents or researchers in the field, our system takes a step further and combines these methods with an analysis of text in order to identify and characterize emerging technologies. The paper describes the indicators and forecasting models, as well as presents the results of applying these indicators to forecast levels of interest in a particular technology based on the analysis of English and Chinese patents. It further shows how the indicators we developed can provide insights into the nature and the lifecycle of emerging technologies.

Conference Topic
Indicators

Introduction
This paper describes Abductive Reasoning Based on Indicators and Topics of EmeRgence, or ARBITER, an automated system whose purpose is to identify and characterize emerging technologies and emerging fields in science. It does so by processing very large collections of scientific publications and patents in multiple languages and identifies trends, associations, and predictions more rapidly than with current methods. Unlike previous approaches to detecting emergence, which are based on the citation analysis of papers and patents (e.g. Bettencourt et al., 2008; Shiebel et al., 2010; Roche et al., 2010), we are extracting information from the text of publications and patents, identifying authors, their affiliations, addresses, as well as classifying types of organizations and publications. Moreover, we apply natural language processing technologies to extract scientific terminology from the full text of the documents, to identify different types of relationships between citations, authors, terms, and organizations, including contrast, opinion, and related work, and to characterize maturity and other properties of terms based on their contextual patterns. This diverse set of features enables us to efficiently process multiple collections and various types of data without dependency on the presence of a specific feature in a collection. For example, our approach is not hampered by the lack of prior art references in Chinese patents, which is a problem for a standard, citation-based analysis of innovative technologies.

To define indicators of emergent technologies and scientific fields, we have developed a pragmatic theory of technoscientific emergence, described in Brock et al. (2012), which builds on Actant Network Theory (Latour, 2005). An Actant Network is a heterogeneous network of human and non-human elements, including people, institutions, funders, meetings, documents, and scientific terminology, interconnected by disparate relationships. The membership of elements within such a network, and the nature and extent of the relationships

[1] Approved for public release; unlimited distribution.

between these elements, is dynamic and constantly changing. To model emergence, we have developed indicators that measure the character and evolution of Actant Networks, including

- Extent of different types of elements in a network, including prolific and prominent entities
- Number of relationships and the volume of traffic in a network
- Growth of entities and relationships, including average growth rate and slope measures
- Novelty of elements and relationships
- Prevalence of the marketplace actant
- Extent of patenting activities
- Amount of disagreements and uncertainties.

In our previous work, we have shown how these indicators can be applied to characterize communities of practice (Babko-Malaya et al., 2013a), identify the presence of the debate in the community (Babko-Malaya et al., 2013b), as well as determine whether practical applications exist for research fields (Thomas et al., 2013). This paper presents the results of applying these indicators to forecast prominence of technology terms, as measured by a significant increase in term frequency. Whereas ARBITER processes both scientific articles and patents, the results presented in this paper are limited to the analysis of patents.

This paper contains three further sections. First, we give an overview of metadata and full text features, describe different categories of indicators designed to identify emerging technologies, as well as demonstrate how the indicators are combined via Bayesian networks into a forecasting model. The next section presents the results of the correlation analysis of indicators with future term prominence for English and Chinese patents, which measures the ability of our indicators to forecast a significant increase in term usage. The final section outlines how the system can be applied to characterize the nature and the lifecycle of the technology.

System Description

Feature Extraction

ARBITER extracts features from the metadata and full text of scientific papers and patents, including Lexis-Nexis Patent data, which includes granted patents and published patent applications from United States and Chinese national patent offices, and Thomson Reuters Web of ScienceTM (abstracts of journals and conference proceedings for the same time period, ~40M records). The features we extract from these sources include metadata features (such as title, author, author affiliation, patent assignees, etc.), as well as features that are based on the analysis of text. All feature extraction capabilities, including language features, are developed for two languages: English and Chinese. A summary of our features is shown in Figure 1. The entities we extract include people, organizations, documents, and scientific terminology, interconnected by different types of relationships.

To analyze persons, we extract authors from scientific articles and inventors from patents. In order to be able to count unique mentions of researchers, we developed a disambiguation component, which groups them into equivalence classes. Our analysis of researchers builds on features such as researcher impact, including Hirsch index and prolificness (measured by patent/paper productivity), as well as co-authorship and citation graphs.

To identify organizations, we extract author affiliations and patent assignees from metadata, as well as funding organizations from the text of acknowledgements and footnotes of scientific papers. All organizations are classified into three classes: Commercial, Academic, and Government/Nonprofit. The organization classification component allows us to evaluate

the extent and changes in the Academic vs. Commercial involvement in a certain field, as well as the diversity of researchers and organizations.

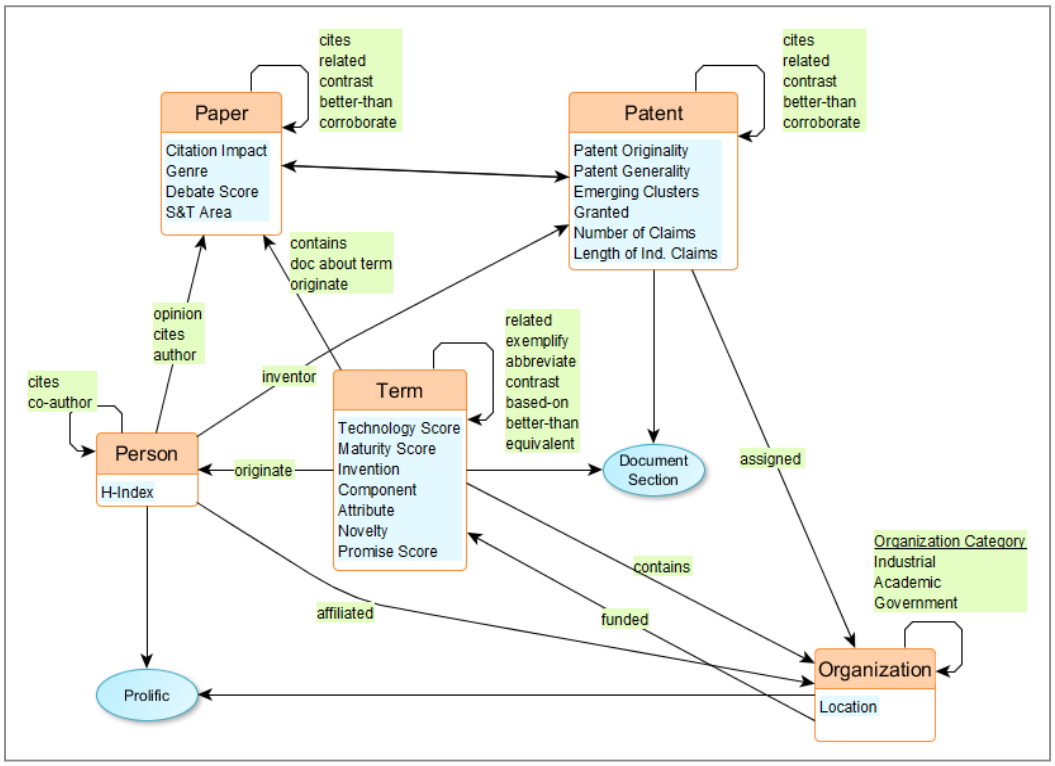


Figure 1. Actant Network extracted from metadata and text.

Our analysis of documents uses citation-based metrics developed by one of our team partners to measure generality, originality, and membership in "emerging clusters" (Breitzman & Thomas, 2015). We further measure mean citation impact of papers and patents, and analyze the structure and length of patent claims.

Our other partners have developed several modules for linguistic processing of text in English and Chinese. For example, to identify scientific terminology, we apply a technology described in Meyers et al. (2010) that extracts scientific noun phrases from the text of papers and patents. The extracted terms are noun phrases that tend to occur frequently in a set of articles from a specific field, but rarely occur in more general or popular articles.

In order to characterize these terms, we score terms based on the extent to which the term behaves like a technology (Anick et al., 2014), as well as assign a maturity score based on how often the term is mentioned in text as being used.

To analyse documents, we apply a genre classifier to evaluate the types of documents that are being published in a certain field, such as review articles or product reviews, as well as to classify documents based on the extent of the debate in the community (Babko-Malaya et al., 2013b). Using the document structure parser, we further identify different sections of documents and categorize claims in patents. To support Chinese extraction, we have adapted a tool to support word segmentation and part of speech tagging to scientific literature and patents (Li & Xue, 2014).

All entities we extract are linked by various types of relations. Whereas some relations are extracted from metadata (e.g. affiliated, invented, assigned, cites, co-author), many relations are extracted from text using information extraction techniques. These relations include opinion relations as well as relations like abbreviate, exemplify, and related work (based on,

better than, contrast, etc), which are described in more detail in Meyers (2013) and Meyers et al. (2014) and are illustrated below.

All entities and relations extracted from full text were evaluated against manually created gold standard corpora. Performance of extraction components is generally comparable across English and Chinese with the f-score above 70-75% in both languages.[2]

Indicators

Using this network, we have developed over 200 indicators that measure different characteristics and changes in the network associated with particular technologies and concepts. The indicators we developed are driven by our pragmatic theory, which defines emergence as the growth in the robustness of actant networks (Brock et al., 2012). The indicators we apply to identify potential disruptive technologies are therefore designed to analyze the relationships between the target entity and other elements in the actant network, including the extent and nature of these relationships, their novelty, dynamic changes, as well as impact, prominence and diversity. Other indicators we explore relate technology emergence to their practicality, as well as the presence of the debate in a community.[3]

Term Momentum Indicators. Our first set of indicators measures momentum in the usage of a particular term. These indicators are time series of annual counts, such as counts of term usage by inventors and organizations, with a further focus on prolific inventors and organizations. In addition, our 'section-based' indicators analyze term usage in independent claims, summary of invention, and abstract sections of patents. The rationale behind an analysis of term usage in specific sections is that these indicators can better measure the extent of the acceptance of the term by the community. For example, if a term occurs in independent claims of patents, it means that it has been legally accepted.

Term Characterization. Beyond indicators based on the momentum associated with individual terms, we also developed indicators that examine different characteristics of these terms. These characteristics include (1) the likelihood that the term describes a technology, (2) the maturity of the technology described by the term, (3) the degree to which the term functions as a description of an invention, and (4) the degree to which a term refers to a component of another technology.

Term characterization scores are calculated by collecting and aggregating evidence from the term's context. For example, to compute maturity scores, we define a set of 'usage' patterns, i.e. patterns that indicate that a term was used or applied: *We used [term] for …, [term] was used for …, employ [term], …* The maturity score is then derived from the number of times these 'usage' patterns are applied to the term. Likewise, the degree to which the term is used as a component is computed based on term usage in 'component'-specific contexts, as illustrated by the sentence *"A typical RFID tag consists of/contains an RFID antenna and RFID chip"*. The terms *RFID antenna* and *RFID chip* are tagged as components in this context, given that they occur as the objects of verbs *consist of* or *contains*. Our expectation is that a time series analysis of maturity of technologies, including their usage as an invention or a component, might be indicative of a change in the lifecycle of a technology, and therefore can be used to identify potentially disruptive technologies (Arthur, 2009).

Semantic Relations. Another class of language-based indicators is based on semantic relations we extract from text. These relations include Opinion, Abbreviate, Exemplify,

[2] Although performance is comparable, there is some variation in the frequency and the type of relations that we extract in the two languages. Some relations are very sparse in Chinese (such as Abbreviations, Contrast, Exemplify (Term1 is an example of Term2). Another difference is that text processing in Chinese is significantly slower than in English due to word segmentation.

[3] The indicators described in this section are focused on the analysis of patents. Similar indicators have also been developed for the analysis of scientific articles, but their analysis is beyond the scope of this paper.

Originate, and different types of Related Work, including Contrast, Based On, and Better Than (Meyers et, 2014). For example, Practical relations represent the author's view that the technology is either being used specially or is useful in some way. Therefore, the indicator that measures the number of Practical relations attached to a term may identify an increase in interest to using a given technology, or its new application. Meanwhile, the relation Abbreviate, which links scientific terms to their abbreviations, can be used to detect the timeline of the acceptance of the term by the community. Finally, relations like Contrast may help to identify the early stages of technology development, given that scientists developing innovative concepts tend to contrast their work with existing research, whereas as the technology becomes more accepted, the number of contrast relations declines.

Document and Inventor Characteristic indicators. This class of indicators measures characteristics of the papers or patents that are using the term. Some of these indicators measure citations to papers containing a given term, or the impact factor of the journals in which the term appears. Others compute dispersion of term usage across technologies or countries, or the number of prior art references in patents.

Inventor Characteristic indicators. In addition to characteristics of documents, we also analyse the inventors and patent assignees who use the term in patents. Examples include the Hirsch index of an inventor or the impact of prior patents granted to inventors or patent assignees.

Novelty. Term Novelty indicators measure the first appearance of a term anywhere in a patent document, as well as the first appearance of a term in specific sections of a patent, such as in the independent claims. Another Novelty indicator computes the first time a term appears with an abbreviation attached. These indicators are thus designed to analyse the timeline of the acceptance of the term by the community.

Most of the indicators described above are time series of annual counts or scores, such as a "number of prominent inventors per year using term in patents." To simplify the modelling process, we reduced each time series to a single value by applying three different methods:

(1) Find the slope of the regression line of indicator values against time (a measure of how fast the indicator is increasing over time);

(2) Calculate the average growth rate for the indicator value over the period selected for the time series;

(3) Compute the sum of indicator values for three years prior to the reference period.

We also experimented with (a) the x2 coefficient of the best-fitting, second-order polynomial for indicator value as a function of year (a measure of curvature, or rate of acceleration), and (b) the two-year prediction of this best-fitting polynomial. These indicators, while sometimes informative, were usually redundant with slope.

Forecasting Models

Our models are tree-augmented Naive Bayes networks (Friedman et al., 1997). Such networks have a structure like that of the network shown in Figure 2. For clarity, we display only a fragment of the model; a complete model may contain 30 to 50 indicator variables.

Bayesian networks provide a factorized representation of a joint probability distribution over a set of variables, and efficiently update the distribution, given evidence in the form of values for variables. In our models, there is a unique root node that represents the unobserved future prominence of an entity. In the above model, this is the node labeled "Prominence3." Prominence is normalized to be between 0 and 1, with a special value of -1 for cases in which the usage of the term decreases. As evidence is entered into the net, the probability distribution over the possible values of prominence is updated.

Bayesian Networks have shown good performance as classifiers (Friedman et al., 1997). We use a version of a Bayesian classifier in which links between indicator variables capture

synergistic effects among those variables – i.e. information about two or more variables tells us more about prominence than the sum of the information value of the individual variables. Capturing synergistic effects has been shown to improve classifier performance (Friedman et al., 1997).

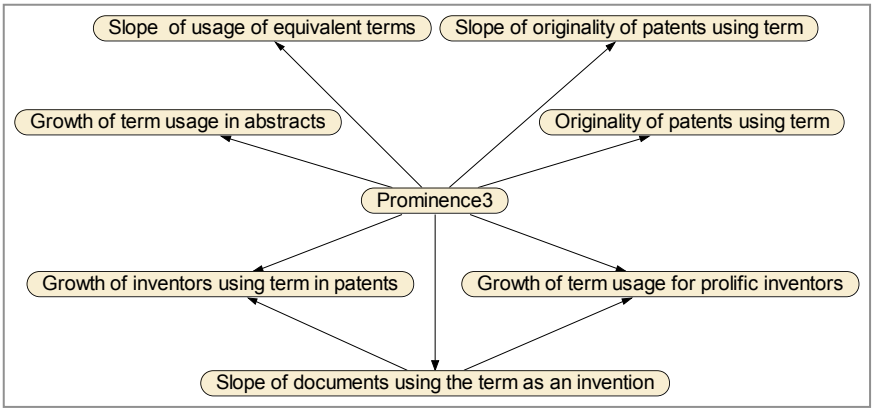


Figure 2. Fragment of model for predicting term prominence.

We chose to use Bayesian networks for several reasons. First, we executed a performance comparison between Bayesian networks (looking at common confusion matrix measurements such as the true and false positive rate, F1 score, etc.) and other classifiers such as JRip, J48, SVM, and meta-classifiers wrapping these, including Bagging and AdaBoostM1. Second, we chose Bayesian networks due to their flexibility and ease of interpretation. Finally, Bayesian networks provide insight into the contribution of indicator variables by supporting the computation of information-theoretic quantities such as mutual information and conditional mutual information.

We use a fine-grained discretization of prominence values instead of a binary prominent/not-prominent variable. This allows more precise computation of information-theoretic relations between indicator variables and prominence than does a binary variable. For example, some variables may be good at predicting very high prominence, while others merely discriminate prominent from non-prominent entities.

Although the prominence variable has a fine-grained discretization, it can be used as a binary classifier by choosing a threshold for prominence. The threshold is chosen through the multi-objective optimization process, described below.

Model Generation and Optimization

Automated model generation must answer the following questions in order to create the desired Bayes net:

- Which indicator variables should be included?
- Which indicator variables should be linked?
- How should continuous variables be discretized?
- How much weight should the training algorithm give to the training data relative to the untrained prior distribution so as to avoid over fitting?
- What threshold for predicting prominence provides the best trade off between recall, precision, and other performance goals?

All of these questions are answered by an optimization loop. This optimization loop uses a multi-objective elitist genetic algorithm (NSGA-II) to search the model parameter space (i.e. answers to the above questions) and rewards solutions that score well relative to specified recall and precision goals. The optimizer uses stratified 10-fold cross validation to compute metrics (e.g. recall and precision) for various combinations of system and ground truth

prominence thresholds. This process leverages the recall ↔ precision trade-off parameter. Finally, the optimizer promotes and further explores solutions that perform relatively well via: (1) uniform crossover, (2) Gaussian mutation for continuous variables, and (3) random flip mutation for discrete variables. The end result is an answer to the above questions that is optimized to the specified objectives.

Indicator Analysis

The analysis described below measures how well the indicators and models can forecast future term prominence, where a term is considered prominent if it has achieved a significant increase in usage.[4] To perform this analysis, we computed indicator values and generated models by processing all documents up to a given year (called the reference period), and then compared system outputs against a ground truth variable measuring an increase in term usage three years after the reference period. This analysis measures the ability of our models to forecast a significant increase in term frequency three years into the future.

By using automated model generation process described above, we generated domain-specific models for different technology areas in English and Chinese patents, including Computer Science, Communications, Biotechnology, and Semiconductors. The performance was higher for Chinese than for English, with the average recall of 0.49 and 0.52 for English patents and recall of 0.47 and precision of 0.61 for Chinese patents. The higher precision for Chinese patents is most likely due to Chinese patents containing a higher percentage of prominent terms than English patents.

To analyze individual indicators, we computed rank correlations between indicators and term prominence. Table 1 illustrates the performance of our indicators for English patents for the domain of Computer Science using Spearman's rank correlation coefficient (Rho) and three approaches to summarizing time series: slope, growth, and sum. For example, in Table 1, Rho slope for the indicator "Number of organizations per year using term in patents" shows the rank correlation for the indicator "the slope of the regression line fitted to the number of organizations using a selected term each year leading up to the reference period."

Table 1 reveals that indicators are significantly correlated with prominence for at least one computation (slope, growth, or sum), with the exception of one — the number of significant opinion relations. This is not unexpected, since opinion relations rarely occur in patents.[5] It also shows that term momentum indicators have the strongest rank correlations with prominence, i.e. measuring past momentum is particularly useful for predicting future prominence. Given that the other classes of indicators are conceptually very different from term momentum indicators, we expect that their effect on the forecasting model is additive to the momentum indicators, rather than duplicative. To test this hypothesis, we computed the partial correlations of non-momentum indicators with prominence, after the most basic term momentum has been accounted for (prior term usage in patents).

[4] One of the limitations of our system is that our analysis applies to individual terms, rather than sets of terms that are representative of technologies or research areas. This limitation is due to the problem of generation of ground truth data for training of our statistical models. In the future, we plan to extend this approach to analyse clusters of related terms, which are representative of technologies and scientific fields.

[5] Our analysis of scientific articles has shown that opinion-type relations (such as positive, standard, and negative opinion) are very infrequent in scientific literature as well, which suggests that opinion-based indicators are not particularly useful for the analysis of scientific literature and patents.

Table 1. Spearman rank correlations with future increase in term usage in English patents.

	Time Series indicators	Rho-Slope	Rho-Growth	Rho-Sum
Term Momentum Indicators	Number of unique organizations per year using term in patents	0.48	0.26	0.47
	Number of prolific organizations per year using term in patents	0.47	0.25	0.46
	Number of unique inventors per year using term in patents	0.50	0.13	0.47
	Number of prolific patenting inventors per year using term in patents	0.45	0.30	0.50
	Number of times per year term is used in patents	0.50	0.26	0.47
	Number of times per year equivalent terms are used in patents	0.48	0.25	0.45
	Number of times per year term is used in summary of invention section	0.52	0.26	0.51
	Number of times per year term is used in Independent claims	0.46	0.38	0.51
	Number of times per year term is used in Abstract section	0.47	0.33	0.52
	Number of industrial assignees using term per year	0.49	0.19	0.46
	Number of academic patent assignees using term per year	0.21	0.26	0.30
Term Characteristics	Annual technology score	N/S	N/S	0.19
	Annual maturity score	0.11	0.13	0.33
	Term usage as an invention	0.12	0.18	0.19
	Term usage as a component	0.23	0.25	0.27
Semantic relations	Annual counts of Exemplify relations	0.33	0.35	0.37
	Annual counts of Practical relations	0.33	0.33	0.37
	Annual counts of Opinion Significant relations	N/S	N/S	N/S
	Term usage with an abbreviation	0.19	0.23	0.24
	Annual counts of Contrast relations	0.20	0.26	0.26
	Annual counts of Based on relations	0.23	0.18	0.24
	Annual counts of Better than relations	0.17	0.13	0.18
Document Characteristic	Originality of patents using the term	N/S	N/S	0.19
	Average citation impact of documents about the term	N/S	N/S	0.31
	Term frequency in an emerging cluster	0.18	0.12	0.42
	Number of prior art references	0.02	-0.12	0.22
	Citations to high impact patents	N/S	N/S	0.31
	Dispersion of term usage across technologies	0.12	N/S	0.46
Inventor Char.	Number of patent inventors using the term as invention	0.12	0.17	0.19
	Hirsch index of the inventor	N/S	N/S	0.19
	Citation impact of prior patents granted to inventor(s)	N/S	N/S	0.29

Table 2 lists the indicators in the descending order of their partial correlations with prominence. An interesting finding is that the indicators that provide information over and above term momentum indicators include the ones that are based on language features, such as Practical and Exemplify relations, as well as term characterization. The indicators that have low or even negative correlations include document- and inventor-based indicators, such as the Hirsch index of the inventor, or the average citation index of document using the term. Having said that, it is important to note that document and inventor indicators are consistently selected by our forecasting models, which indicates that they are not really replaceable by other indicators.

Table 2. Partial correlation of indicators with prominence, controlling for momentum indicator.

Indicator	Partial Correlations
Annual counts of Practical relations	0.199
Term usage as an invention	0.170
Annual counts of Exemplify relations	0.169
Term usage as a component	0.159
Citations to high-impact patents	0.149
Annual maturity score	0.134
Annual technology score	0.129
Annual counts of Based_on relations	0.120
Annual counts of Contrast relations	0.114
Originality of patents using the term	0.101
Term usage with an abbreviation	0.098
Annual counts of Better_than relations	0.080
Citation impact of prior patents granted to inventor(s)	0.019
Average citation impact of documents about the term	-0.023
Number of prior art references	-0.042
Term frequency in an emerging cluster	-0.057
Hirsch index of the inventor	-0.074

Comparing indicators with different rationale, such as practicality versus discursive interest, one interesting finding is that the indicators focusing on the practicality of a field have the strongest correlations with prominence. These indicators include maturity scoring, usage as a component, Practical relations, and term usage by industrial patent assignees. Indicators focused on discursive interest in the term, such as Contrast relations, Better Than relations, and term usage by academic researchers in the field, have weaker (although still significant) correlations with prominence (as shown in Table 1 above). This suggests that, while both practicality and discursive interest are useful characteristics for the analysis of patents, the former is of particular value in forecasting the future prominence of terms.

Our further analysis of indicators focused on trying to identify indicators with complementary strengths. For example, we discovered that many of our indicators are good at predicting whether term usage will increase or decline/remain stable, but there are only a few indicators that are good at predicting different degrees of positive changes in term usage. This is illustrated by Table 3, which shows rank correlations between indicators and future changes in term usage coded as positive versus non-positive (Rho+/), as well as rank correlations considering positive values only (Rho-Pos).

As Table 3 shows, the correlations for the classification problem (Rho+/-) are generally higher, which suggests that it is more straightforward for an indicator to forecast whether or not a term will have a positive prominence, versus forecasting different degrees of positive prominence. It also reveals that some indicators might have particular strengths. For example, while momentum indicators and some document characteristic indicators perform best for delineating between positive and non-positive cases, the best indicator for distinguishing between different levels of positive prominence is “the proportion of granted patents using term relative to published documents”.

Table 3. Spearman correlations for indicators based on different conditions.

	Time Series indicators	Rho+/-	Rho-Pos
Term Momentum Indicators	Number of unique organizations per year using term in patents - Slope	0.50	0.21
	Number of prolific patenting organizations per year using term in patents - Slope	0.49	0.19
	Number of unique inventors per year using term in patents - Slope	0.52	0.22
	Number of prolific patenting inventors per year using term in patents - Slope	0.52	0.22
	Number of times per year term is used in patents - Slope	0.53	0.22
	Number of times per year equivalent terms are used in patents - Slope	0.51	0.20
	Number of times per year term is used in summary of invention section - Sum	0.54	0.24
	Number of times per year term is used in Independent claims section - Sum	0.53	0.25
	Number of times per year term is used in Abstract section - Sum	0.55	0.26
	Number of industrial assignees using term per year - Slope	0.51	0.21
	Number of academic patent assignees using term per year - Sum	0.33	0.09
Term Character ization	Annual technology score - Sum	0.21	0.05
	Annual maturity score - Sum	0.33	0.14
	Term usage as an invention - Sum	0.17	0.12
	Term usage as a component - Sum	0.27	0.13
Semantic relations	Annual counts of Exemplify relations - Sum	0.36	0.19
	Annual counts of Practical relations - Sum	0.37	0.18
	Term usage with an abbreviation - Sum	0.22	0.15
	Annual counts of Contrast relations - Sum	0.24	0.15
	Annual counts of Based_on relations - Sum	0.21	0.15
	Annual counts of Better_than relations - Sum	0.14	0.14
Document Characteristic	Originality of patents using the term - Sum	0.21	0.07
	Average citation impact of documents about the term- Sum	0.30	0.03
	Term frequency in an emerging cluster - Sum	0.46	0.15
	Number of prior art references - Sum	0.27	0.05
	Citations to high-impact patents - Sum	0.33	0.16
	Dispersion of term usage across technologies - Sum	0.50	0.18
Inv- entor Char.	Number of patent inventors using term as invention-Sum	0.18	0.10
	Hirsch index of the inventor - Sum	0.30	-0.02
	Citation impact of prior patents granted to inventor(s) - Sum	0.36	0.07
Single value	Proportion of granted documents using term relative to published documents	0.39	0.29
	The year the term first appeared in a patent	-0.15	0.01
	The year the term first appeared with an abbreviation	0.25	0.17

We further evaluated performance of indicators across one-, two- and three-year gap periods and observed a significant difference. All indicators tend to perform better in predicting longer forecasts (such as three-year gap) than shorter periods (such as one- or two-year gap). This may be because a three-year forecast smoothed out some of the year-by-year volatility in term usage.

Table 4. Spearman correlations for term prominence indicators in Chinese patents.

Time Series indicators	Rho-Slope	Rho-Growth	Rho-Sum
Number of unique inventors per year using term in patents	0.50	N/S	0.46
Number of prolific patenting inventors per year using term in patents	0.50	N/S	0.46
Number of times per year term is used in patents	0.50	0.06	0.46
Number of times per year term is used in Independent claims section	0.50	0.16	0.44
Number of unique organizations per year using term in patents	0.48	N/S	0.43
Number of prolific patenting organizations per year using term	0.48	N/S	0.44
Number of times term is used in summary of invention section	0.18	N/S	0.11
Annual maturity score	0.08	0.08	0.28

Finally, Table 4 shows correlation analysis for some of the indicators that were applied to Chinese Computer Science patents. It is important to note that citations rarely occur in Chinese patents, so indicators that are based on citation metrics cannot be used for the analysis of term prominence in Chinese. A comparison of correlations for English and Chinese (Tables 1 and 4) reveals that the general patterns across two collections are very similar, with Slope and Sum term momentum indicators performing particularly well, along with the Sum version of the Maturity Score.

Future Plans: Term Characterization

In addition to predicting future levels of interest to a technology, we expect that the indicators we developed can also provide some insights into the nature of the technology, its lifecycle, and other term characteristics. An example of this type of analysis is illustrated by 10 computer science terms, shown in Table 5.

Table 5. An analysis of 10 computer science terms.

Term	Pe	Term Characterization Analysis
RFID antenna	0.60	a device, becoming widely used in diff applications in 2007
Instant messaging	0.47	a technology or method, innovative, not a component
Robotics	0.31	a branch of technology, not a specific device, mature
XML	0.31	technology name, active area of research
Speech recognition	0.31	widely accepted technology, but best practice is being debated
Cellular telephone	0.31	a widely used standalone device, still of interest
RDF	0.31	technology name, becoming more widely used
Linux operating system	0.31	a widely accepted mature technology
GPS	0.30	a technology, widely used, mature, active area of research
Quantum computing	0	a principle or concept, innovative, no practical applications

The Pe column shows our predictions for the future changes in term usage, as described above, where zero value indicates that term usage will remain stable or decline in the future, whereas positive values predict that there will be an increased community interest in the term. The terms were analysed using 2007 as the reference period, forecasting term usage in 2010. The most interesting terms in this list include *RFID antenna* and *instant messaging*, the other terms, except for *quantum computing*, have slightly lower positive Pe values, indicating that there will be some growth in their usage between 2007 and 2010. The fact that quantum computing has zero value is not unexpected, considering that the data processed for this analysis included patent literature only, and this term has rarely been used in patents until 2007.

In addition to identifying terms with high prominence, we expect that the indicators described in the paper can also be used to characterize technologies, as illustrated in Table 5. For example, by using individual indicators or groups of indicators, we can potentially identify

widely accepted and mature technologies, terms that function as components of other technologies, active areas of research, as well as areas where best practice is being debated.

For example, Figure 3 reveals the values for the indicator that computes the average growth rate of term usage by academic institutions. This indicator can be used to identify innovative technologies that attract a growing attention from academia. Out of the 10 terms, technologies with the highest growth of academic assignees include *RFID antenna, instant messaging*, and *RDF*.

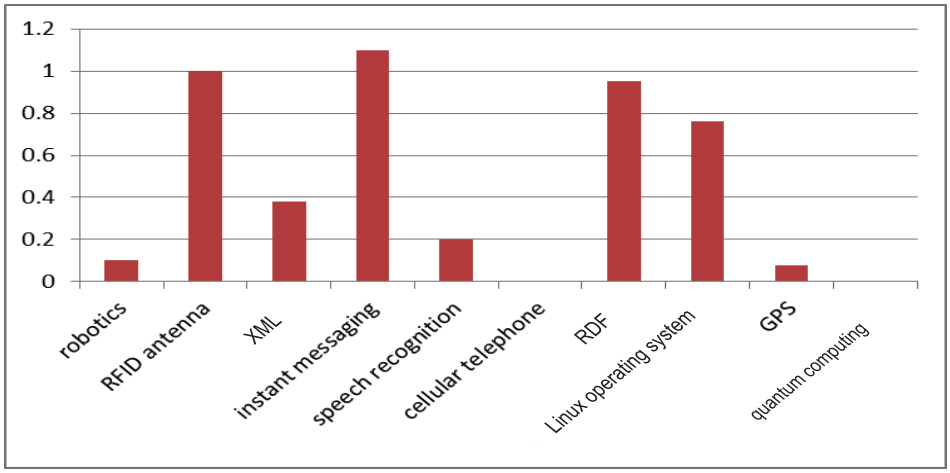


Figure 3. The average growth rate of academic assignees using term from 2002 to 2007.

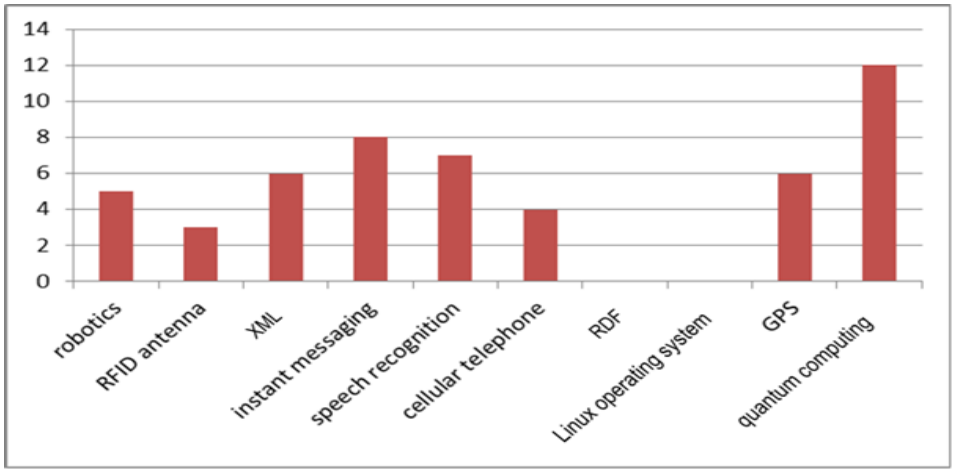


Figure 4. The number of inventors using term as an invention from 2005 to 2007.

Figure 4, on the other hand, illustrates the indicator values for "the number of inventors that were using the term as a description of an invention". Interestingly, the term that has the highest indicator value in this case is *quantum computing*. The terms with the higher values in Figure 3, *RDF and RFID antenna* have the lowest indicator values in Figure 4. This example suggests that individual indicators or groups of indicators may be used to detect different types of emerging technologies and that these differences might be related to their nature or lifecycle. It further illustrates that individual indicators can help to identify newer terms like *quantum computing*, and that high values of specific indicators may be indicative of the future potential of the term.

Conclusion

The system presented is capable of scanning millions of technical documents, extracting key indicators from both text and metadata, and forecasting meaningful trends and predictions from the extracted metrics. In particular, the extracted indicators are useful in predicting levels of interest in particular technologies. We also showed how the indicators provide insight into the nature and the lifecycle of emerging technologies, including their maturity, practicality, stages of development, and acceptance by the community.

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

Anick P, Verhagen M., & Pustejovsky J. (2014). Identification of technology terms in patents. In *Proceedings of LREC 2014*.

Arthur, B. (2009). The Nature of Technology: What It Is and How It Evolves. Free Press.

Babko-Malaya O., Thomas P., Hunter D., Meyers A., Pustejovsky P., Verhagen M., & Amis G. (2013a). Characterizing communities of practice in emerging science and technology fields, In *Proceedings of the International Conference on Social Intelligence and Technology 2013*.

Babko-Malaya O., Meyers A., Pustejovsky J., & Verhagen M. (2013b). Modeling debate within a scientific community. In *Proceedings of the International Conference on Social Intelligence and Technology 2013*.

Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chávez, C., & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, *75*(3), 495–518.

Breitzman, A., & Thomas, P. (2015). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy*, 44(4), 195-205.

Brock, D.C, Babko-Malaya O., Pustejovsky, J., Thomas, P., Stromsten, S., & Barlos, F. (2012). Applied actant-network theory: Toward the automated detection of technoscientific emergence from full-text publications and patents. In *Proceedings of the AAAI Fall Symposium on Social Networks and Social Contagion 2013*.

Friedman N, Geiger, D., & Goldszmidt, M. (1997). Bayesian networks classifiers. *Machine Learning*, 29, 131-163.

Latour B. (2005). *Reassembling the Social: An Introduction to Actor-Network Theory*. Oxford University Press.

Li, S., & Xue, N. (2014). Effective document-level features for Chinese patent word segmentation, In *Proceedings of ACL 2014*.

Meyers, A., Zachary, G., Grieve-Smith, A., He, Y., Liao, S., & Grishman, R. (2014). Jargon-Term Extraction by Chunking. In *Proceedings of SADAATL 2014*.

Meyers, A. (2013). Contrasting and corroborating citations in journal articles, In *Proceedings of Recent Advances in Natural Language Processing 2013*.

Meyers, A., Lee G., Grieve-Smith A., He, Y., & Taber, H. (2014). Annotating relations in scientific articles. In *Proceedings of LREC 2014*.

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.

Roche, I., Besagni, D., François, C., Hörlesberger, M., & Schiebel, E. (2010). Identification and characterization of technological topics in the field of molecular biology. *Scientometrics*, *82*(3), 663-676.

Thomas P., Babko-Malaya O., Hunter D., Meyers A., & Verhagen M. (2013). Identifying emerging research fields with practical applications via analysis of scientific and technical documents. In *Proceedings of ISSI 2013*.

Understanding Relationship between Scholars' Breadth of Research and Scientific Impact

Shiyan Yan[1] and Carl Lagoze[2]

[1] *shiyansi@umich.edu*
School of Information, University of Michigan, Ann Arbor

[2] *clagoze@umich.edu*
School of Information, University of Michigan, Ann Arbor

Abstract
Many existing metrics to evaluate scholars consider their scientific impact without considering the importance of breadth of research. In this paper, we define a new metric for breadth of research based on the generalized Stirling metric that considers multiple aspects of breadth of research. We extract research topics in computer science using concept extraction and clustering from the literature in the ACM dataset. We then assign authors a distribution over these research topics, from which we calculate scores of breadth of research for each author. We design five simulation experiments that evaluate the ability of a metric to measure breadth of research and use these experiments to compare our new metric to traditional metrics. The results show how these metrics perform in different experiments, concluding that no metric consistently outperforms the others. We test the relationship between our new metric and scientific impact and find a weak correlation between them. Finally, we find that the variation of the metric over time illustrates a possible publication pattern for scholars.

Conference Topic
Indicators

Introduction
An increasing number of scholars are engaged in interdisciplinary research (Porter, Cohen, David Roessner, & Perreault, 2007; Wagner et al., 2011). Some of this is due to the emergence of new scholarly "disciplines" that are inherently multi-disciplinary such as information science, while some arises from scientific problems such as climate change that require expertise from multiple fields. Meanwhile, scholarly impact and influence continues, by and large, to be measured by indices that ignore breadth of research and may even penalize scholars who diversify their research portfolio. For example, H-index, which is used extensively to measure scholarly impact, and which has been criticized for its limited focus (Weingart, 2005), may be unfair when comparing scholars with different degrees of breadth of research. Ultimately, a metric or a set of metrics is needed that accounts for breadth of research, so that breadth of research can be measured and be included in an evaluation system of scholars' scientific influence.

In this paper we describe research that explores the area of scholarly impact metrics and breadth of research. The contributions of our work are as follows. We design a new metric to measure scholars' breadth of research that builds on traditional metrics. We develop a multi-stage method for extracting topics from a corpus (in our case computer science papers) and calculate the scores of breadth of research for authors who have published papers in computer science conferences. We design five simulation experiments that compare the relative performance of existing metrics and our new metric for measuring breadth of research. We measure the relationship of breadth of research and H-index for scholars who are authors in our corpus. Finally, we explore the variation of breadth of research for scholars over time to observe their paper publication behavior over their careers.

The structure of this paper is as follows. The next section describes related work in the areas relevant to our work. Following that, we report on the dataset we used in our research. We

then describe our process of dictionary extraction, topic extraction, paper assignment and author assignment to topics. In the subsequent section we illustrate our new metric and compare it to traditional metrics. The penultimate section describes simulation experiments to show the performance of the new metric, the relationship between the new metric and metrics of research impact, and the variation over time of breadth of research for scholars. Our conclusions and possible future work are listed in the final section.

Related Work

There is a variety of existing literature relevant to the area of breadth of research. The areas covered by this literature include topic extraction, topic relationship extraction, metrics design and the relationship between different aspects of research evaluation systems.

There are many methods to associate topics to publication. The simplest one is to use the classification codes in a dataset, such as ISI subject categories in Web of Science, as the set of topics. But these categories are too coarse-grained and hide intra-disciplinary variability. Another method is to use unsupervised learning algorithms to extract some topics according to the content of papers or the citation network of papers. Topic modelling (Blei, Ng, & Jordan, 2003) is one of the popular unsupervised learning algorithms based on content of papers. This model has been used to identify the disciplines that comprise interdisciplinary work funded by NSF (Nichols, 2014). The ACT model (author-conference-topic) (Li et al., 2010) is an adaptation of Blei's model. Another approach is to use community detection in networks as a basis for finding topics. One example is the use of two-round clustering (Rosvall & Bergstrom, 2008) over the citation network to extract topic-associated communities (Velden & Lagoze, 2013). Another method using both the citation network and the word distribution of abstracts (Jo, Hopcroft, & Lagoze, 2011) finds temporally-ordered topics from a corpus of scientific literature, such as the ACM dataset.

Understanding the relationship between topics is also an important step after topic extraction, because the calculation of the similarity of topics is necessary for understanding the breadth of research. Some researchers have extracted the relationships and used information visualization techniques to represent the relationship between different topics. For example, Yan (2013) detects the path between different disciplines to find the evolution of some areas. Another paper describes a new method to find the diversity subgraph in a multidisciplinary scientific collaboration network (He, Ding, Tang, Reguramalingam, & Bollen, 2013). An interesting visualization method leverages the circle of science to visualize the relationship between disciplines in one dimension (Boyack & Klavans, 2009).

Many metrics have been designed to measure factors related to scientific influence. The most common metrics are impact factor and H-index, which measure the number of citations of scholars' papers. Although these metrics have many problems such as lack of universality between different disciplines (Kaur, Radicchi, & Menczer, 2013), they are still widely used in systems like Google Scholar. Some alternative metrics also use the number of citations to measure the scientific influence of scholars (Ruscio, Seaman, D'Oriano, Stremlo, & Mahalchik, 2012). They offer advantages over simple metrics such as H-index, but they also focus solely on the citation count of papers. Other metrics based on the centrality of scholars in a network (e.g., co-authorship) like PageRank and betweeness centrality (Bollen, Van de Sompel, Hagberg, & Chute, 2009) are also widely used. However, the correspondence of centrality to actual influence is unknown.

As mentioned earlier, commonly used metrics of scholarly influence fail to consider breadth of scholars' research. In response a number of researchers have created some metrics for the degree of interdisplinarity and more generally breadth of research. The report of quantitative metrics and context in interdisciplinary scientific research (Wagner et al., 2011) is a good survey for metrics for interdisciplinarity. Specialization and integration (Porter et al., 2007)

are good metrics of interdisciplinarity because they consider similarity between disciplines when measuring interdisciplinarity. They can be modified easily in the context of a diversity of research topics. Some papers discuss different dimensions of interdisplinarity (Rafols & Meyer, 2010; Rafols, Leydesdorff, O'Hare, Nightingale, & Stirling, 2012): diversity, coherence and intermediation. They define diversity as a combination of variety, balance and disparity. Coherence means link strength between different disciplines. Intermediation is based on the network structure and is measured by betweenness centrality, clustering coefficient and average similarity. Other papers describe metrics based on these dimensions. Cassi, Mescheba, and de Turckheim (2014) divides the Stirling metric into "within component" and "between component" to measure the diversity of articles. Jensen & Lutkouskaya (2013) defines six indicators based on the dimensions and measure the breadth of research at two levels (article and laboratory). Karlovčec and Mladenić (2014) defines a new diversity metric based on Generalized Stirling. The metric incorporates connectedness of the citation graph into the original metric and applies it in exploratory analysis of the research community in Slovenia. Roessner, Porter, Nersessian, and Carley (2012) validates the interdisciplinarity metrics with ethnographic materials (field observations and unstructured interviews).

Finally, some research has focused on the relationship between breadth of research and other factors considered in scientometrics (not just scientific influence). One interesting paper finds that the papers with an average degree of interdisciplinarity will get higher impact than papers with too high or too low degree of interdisciplinarity (Sternitzke & Bergmann, 2008). The results are convincing but metrics used in this paper are quite simple (Jaccard similarity and cosine similarity). Two papers find that interdisciplinary papers have potentially lower impact than more focused papers. One of them finds that multidisciplinary papers are not frequently cited in contrast to the disciplinary papers (Levitt & Thelwall, 2008). The other explains how high-ranked journals suppress interdisciplinary research (I Rafols & Meyer, 2010). Other papers describe some factors that can encourage researchers to be involved in interdisciplinary research work (Carayol & Thi, 2005; van Rijnsoever & Hessels, 2011). They provide some theories to explain why scholars choose interdisciplinary projects. Some findings support that there are no correlations between citation ranks and ranked interdisciplinarity indices (Ponomarev, Lawton, Williams, & Schnell, 2014). In contrast, other researchers confirm that the degree of interdisciplinarity is strongly correlated with the impact factor (Silva, Rodrigues, Oliveira, & da F. Costa, 2013).

Dataset

We extract abstracts, full text and other metadata from the ACM digital library for proceedings of major conferences in computer science. From these proceedings we select authors whose names are unambiguous and who have published at least five papers. The standard for unambiguity is whether using the full name as the query sent to Google Scholar returns only one researcher profile with the same name. We extract the citation numbers and H-indexes by crawling over Google Scholar. Overall we crawled H-indexes and citation numbers for 8911 authors from Google Scholar in August 2014. We also used the Wikipedia dataset to extract important terms in computer science.

Topic Extraction and Assignment

Both traditional metrics and the new metric designed in this paper require a distribution over different topics or areas for authors. In order to generate topic distributions, we leverage the text data in the papers of ACM digital library and implement three steps to form distributions: dictionary extraction, topic extraction and author assignment.

Dictionary Extraction

How to define topics is the first problem to be solved in the topic extraction and assignment. In our work, we extract a dictionary of n-grams in computer science and cluster them into topics using the Affinity Propagation algorithm (Frey & Dueck, 2007). Three different sources of dictionaries are used in this paper: grams that are frequently used in papers, grams that can be matched to their abbreviations in the papers, and entries in Wikipedia.

Dictionary extraction follows these steps:

1. Extract bigrams and trigrams that occur frequently in papers using a threshold of more than 10 times for bigrams and more than 5 times for trigrams. The threshold helps to eliminate noisy grams with low frequency.
2. Extract grams from papers that conform to the pattern "n-grams (abbreviation)", e.g. machine learning (ML).
3. Intersect the results of step 1 and step 2 (3816 terms in total).
4. Build a network of entries in Wikipedia according to hyperlinks between them in the website.
5. Make use of grams in step 3 and search their neighbours in the network of Wikipedia terms. If their neighbours also occur frequently in papers (with frequency higher than the thresholds mentioned above), add the terms into the final dictionary (6100 terms)

The top 5 bigrams and top 5 trigrams in the final dictionary are shown in Table 1:

Table 1. Grams with top frequency

Grams	*Frequency*
User Interface	2372
Software development	2102
Programming language	2042
Software engineering	1988
Operating system	1761
Wireless sensor network	586
World wide web	467
Graphical user interface	305
Support vector machine	300
Discrete event simulation	287

Topic Extraction and Assignment

After extracting the dictionary, we count the co-occurrence measure for every pair of terms. We then calculate the similarity between different terms by:

$$Sim_{ij} = \log \frac{Cooccur_{ij} + 1}{Max(Cooccur_{ij}) + 2}$$

The logarithm calculation makes the distribution of similarity more uniform and avoids the influence of outliers of co-occurrence numbers. We weight co-occurrences of terms in abstracts of papers more than those in full text based on the intuition that abstracts generally have a stronger “topic signal”. Using the computed similarity matrix of terms, we then run Affinity Propagation to cluster together similar terms and choose an exemplar for every cluster. The benefits of Affinity Propagation are that there isn’t a need to parameterize the number of clusters and that the exemplars for every cluster provide a straightforward explanation of what these clusters are about. More than two hundred clusters, or topics, are generated. Here are two examples of the clustering results:

Exemplar: digital library

Terms:
citation analysis, citation index, community building, digital earth, digital library, digital library software, digital preservation, digital reference, discourse analysis, dublin core.
Exemplar: machine learning
Terms:
active learning, adaptive control, bayes classifier, belief propagation, clinical trial, computational learning theory , concept learning, conditional random field.
We then assign every paper a probabilistic assignment to the different topics according to their respective frequency of n-grams associated with the particular topic. Therefore, every paper will have a distribution over topics.

Author Assignment

Using the clusters of grams in computer science and the topic distributions for every paper, we assign authors into different topics according to their papers. Every author is represented by a distribution over topics, which are used to calculate scores of metrics. There does not exist a "gold standard" list of researchers that ranks breadth of research that we can use to evaluate how reasonable our topic assignments are. We list below some topic distributions for well-known computer scientists to demonstrate our assignment.

John Koza	
1 genetic programming	0.567
2 programming language	0.083
3 knowledge base	0.063
Peter Denning	
1 memory management	0.107
2 computer systems	0.093
3 information systems	0.050
Eric Horvitz	
1 user interface	0.082
2 information retrieval	0.067
3 machine learning	0.051
4 speech recognition	0.047

Breadth of Research Measurement

With the author distribution of topics established, the key question is how to translate this into a measure of breadth of research for authors. As mentioned in the section describing related work, many metrics have been used to measure the "degree of interdisciplinarity". Compared to previous metrics to measure breadth of research, we design a new metric that considers the topic distribution, similarity distribution and coherence within research topics.

Summary of Old Measurements

There are many measurements of diversity or interdisciplinary, like entropy (Weaver, 1949), Simpson's index (Simpsons, 1949) and generalized Stirling (Stirling, 2007). Each of these is computed as follows. Denote p_i as the probability of topic distribution for an author over topic i, d_{ij} as the distance between topic i and topic j.

$$Entropy = \sum_{i=1}^{n} -p_i \times \log_n (p_i)$$

$$Simpson = 1 - \sum_{i=1}^{n} {p_i}^2$$

$$Generalized\ Stirling = \sum_{i,j} d_{ij}^{\alpha}\ (p_i \times p_j)^{\beta}$$

Comparing them, only generalized Stirling considers not only the distribution of topics but also the similarity between topics. The further the distance between topics in which an author publishes papers, the more diverse will the author's research interest be. However, the traditional metrics do not consider the notion of differing *coherence* between different research topics. And the degrees of influence of topics with small proportions are very limited. The new measurement is a modified version of the generalized Stirling metric and it incorporates the coherence of topics and value of *minor topics* (topics with small proportions).

New Measurement

The new metric for breadth of research is defined as follows.

Denote d_{ij} as the distance between two topics, which are defined as the average distance (inverse of similarity defined above) between terms in the two topics, p_i as the probability of an author's paper belong to topic i, coh_i as the *coherence* of topic i. Coherence of each topic is the proportion of authors for whom the respective topic is their major research topic, which is an important signal to illustrate whether a research topic concentrate on some core research questions. Parameters α, β, γ are used to control the relative weights of different components.

$$Breadth\ of\ Research = \sum_{i,j} d_{ij}^{\alpha}\ (p_i + p_j)^{\beta} (Coh_i \times Coh_j)^{\gamma}$$

We modify the product of p_i and p_j in generalized Stirling to summation of p_i and p_j because the summation will give minor topics more chances to be counted into the measurement of breadth of research. We add the coherence term into the metric because different topics have different "density" within themselves. For example, some topics like digital library are less coherent topics because there are many diverse subtopics in these topics. But for topics like operation systems, researchers concentrate on several narrow subtopics. A researcher focusing on digital library should have larger breadth of research than operating systems researchers if other variables are controlled (so the gamma should have a negative value).

The new metric leverages properties of papers (topic distribution), properties of topics (coherence) and properties of relationship (topic similarity). The tunable parameters give the metric more flexibility to balance between different aspects of breadth of research.

Experiments

Simulation Experiment

There is no established standard for determining the quality of metrics of breadth of research. Furthermore, there is no ground truth to show the rankings of scholars' breadth of research with which to validate the various metrics. We propose an alternative evaluation method based on a set of axioms concerning breadth of research and then test how the metrics perform according to these axioms.

In addition to the definition of d_{ij} and coh_i defined in the previous section, the following definitions relate to the axioms.

- Denote A_i as the article i, $C=\{ A_1, A_2 ... \}$ as a *collection* of articles, and N_C as the number of articles in collection C.
- Denote t_i as the topic i, $D_A(t)$ as the topic distribution of article A over topic t. $(\sum_t D_A(t) = 1)$

- Denote $D_C(t)$ as the topic distribution of collection C over topic t. $D_C(t) = \frac{1}{N_C}\sum_{A_i \in C} D_{A_i}(t)$. $(\sum_t D_C(t) = 1)$
- Denote *score(C)* as the score of a metric over the collection of articles C

Axiom1: Publish in Old Topics

If an author publishes a paper in a topic in which she has published many papers before, her breadth of research should decrease.

Choose t, s.t. $t = arg\,max_t\, D_C(t)$, construct a new article A_{new}, s.t. $D_{A_{new}}(t) = 1$. $C' = C \cup \{A_{new}\}$. Then *score(C') < score(C).*

Axiom2: Publish in New Topics

If an author publishes a paper in a new topic in which she has never published, her breadth of research should increase.

Choose t, s.t. $D_C(t)=0$, construct a new article A_{new}, s.t. $D_{A_{new}}(t) = 1, C' = C \cup \{A_{new}\}$. Then *score(C') > score(C).*

Axiom3: Publish in New Topics Twice

If an author publishes papers in two new topics in a sequence, the increase of breadth of research in the second time should be smaller than the increase of that in the first time.

Choose t_1 and t_2, s.t. $D_C(t_1)=0$, $D_C(t_2)=0$, $t_1 \neq t_2$,construct two new articles A_{new1} and A_{new2}, s.t. $D_{A_{new1}}(t) = 1$ and $D_{A_{new2}}(t) = 1$. $C' = C \cup \{A_{new1}\}$, $C'' = C' \cup \{A_{new2}\}$. Then *score(C')-score(C) > score(C'')-score(C').*

Axiom4: Publish in Close Topics

If an author publishes a paper in a new topic close to the author's research interest, the improvement of her breadth of research should be less than that of publishing a new paper in a randomly chosen topic.

Randomly Choose t_1 s.t. $D_C(t_1)=0$, construct a new article A_{new1}, s.t. $D_{A_{new1}}(t_1) = 1$. $C' = C \cup \{A_{new1}\}$. Choose t_2 s.t. $D_C(t_2)=0$ and $arg\,min_t(inf_{t_0 \in \{t|D_C(t)>0\}} d_{t_0 t_1})$. Construct a new article A_{new2}, s.t. $D_{A_{new2}}(t_2) = 1, C'' = C' \cup \{A_{new2}\}$. Then *score(C'') < score(C')*

Axiom5: Publish in Coherent Topics

If an author publishes a paper in a new topic with high coherence, the improvement of her breadth of research should be less than that of publishing a new paper in a randomly chosen topic.

Randomly Choose t_1 s.t. $D_C(t_1)=0$, construct a new article A_{new1}, s.t. $D_{A_{new1}}(t_1) = 1$. $C' = C \cup \{A_{new1}\}$. Choose t_2 s.t. $D_C(t_2)=0$ and $t_2 = arg\,max_t\,(Cohe_t)$. Construct a new article A_{new2}, s.t. $D_{A_{new2}}(t_2) = 1, C'' = C' \cup \{A_{new2}\}$. Then *score(C'') < score(C').*

We implemented five simulation experiments based on the original dataset with 8911 authors to test how the traditional metrics and our new metric conform to the axioms. The results are shown in Table 2.

Table 2. Probability that metrics satisfy of the axioms

	Entropy	*Simpson's*	*GL Stirling* $(\alpha = 2; \beta = 0.3)$	*New Metric* $(\alpha = 1, \beta = 0.5, \gamma = -0.5)$
Axiom1	0.99	0.99	0.97	0.88
Axiom2	0.89	0.97	0.86	0.86
Axiom3	0.97	0.94	0.50	0.50
Axiom4	0	0	0.76	0.70
Axiom5	0	0	0.54	0.62

The results show that entropy and Simpson's perform well in the first three axioms because they don't consider distances between topics and introduce less noise. Because every new topic will be regarded equally for these metrics, they cannot follow Axiom4 and Axiom5. Generalized Stirling and our metric perform reasonably well in Axiom1 and Axiom2, but worse than entropy and Simpson's. They perform relatively badly in Axiom3 because relatively bad performance on publishing a paper in new topic (Axiom2) will aggregate when testing the performance of publishing two papers in two new topics. But they perform well in Axiom4 because of the consideration of distances. Also we find our metric performs better than generalized Stirling in Axiom5, which means coherences of topics and greater weights on minor topics are beneficial when we consider variation of metrics when people publish in topics with different coherence levels.

Parameter Sensitivity

The performance of new metric is influenced by the value of parameters α, β and γ. We tested the performance of the new metric with different settings. The results are shown in Table 3, Table 4 and Table 5.

Table 3. Average Prob of satisfying the axioms with different α.

	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$
Axiom1	0.40	0.42	0.48	0.62
Axiom2	0.33	0.38	0.44	0.55
Axiom3	0.34	0.32	0.24	0.22
Axiom4	0.38	0.57	0.66	0.64
Axiom5	0.63	0.61	0.57	0.52

Table 4. Average Prob of satisfying the axioms with different β.

	$\beta = 0.1$	$\beta = 1$	$\beta = 10$	$\beta = 100$
Axiom1	0.86	0.67	0.30	0.08
Axiom2	0.69	0.57	0.24	0.16
Axiom3	0.40	0.40	0.29	0.05
Axiom4	0.57	0.57	0.59	0.53
Axiom5	0.61	0.61	0.59	0.52

Table 5. Average Prob of satisfying the axioms with different γ.

	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 10$	$\gamma = 100$
Axiom1	0.58	0.47	0.45	0.45
Axiom2	0.24	0.39	0.47	0.48
Axiom3	0.09	0.26	0.34	0.38
Axiom4	0.49	0.57	0.59	0.59
Axiom5	0.62	0.66	0.58	0.53

The tables show that the metric is very sensitive to the α, β and γ. In order to find the best parameter setting, we calculated the average performance over five different simulation experiments for every parameter settings. We selected the settings with highest average performance and a minimum threshold of at least 0.5 in every experiment. The best setting for Generalized Stirling is $\alpha = 2, \beta = 0.3$. The best setting for the new metric is $\alpha = 1, \beta = 0.5$ and $\gamma = -0.5$. They are used in the comparison of metrics in Table 2.

Summation Modification

One of important modifications of our metric is the replacement of product with summation in the second term of metric. We test the effect of this. If we control the distance term and coherence term in the metric to be the same for every topic and set $\beta = 1$. The metric using summation will definitely follow Axiom2 but not follow Axiom1 and Axiom3.

Let *n* represents the number of topic.

Axiom1: Publish in Old Topics

$$score(C) = \sum_{i,j} d^{\alpha}\ (p_i + p_j)(\text{coh} \times \text{coh})^{\gamma} = (n-1)d^{\alpha}(coh)^{2\gamma}$$

$$= \sum_{i,j} d^{\alpha}\ (p_i{}' + p_j{}')(\text{coh} \times \text{coh})^{\gamma} = score(C')$$

Axiom2: Publish in New Topics

$$score(C) = \sum_{i,j} d^{\alpha}\ (p_i + p_j)(\text{coh} \times \text{coh})^{\gamma} = (n-1)d^{\alpha}(coh)^{2\gamma}$$

$$< \sum_{i,j} d^{\alpha}\ (p_i{}' + p_j{}')(\text{coh} \times \text{coh})^{\gamma} = (n)d^{\alpha}(coh)^{2\gamma} = score(C')$$

Axiom3: Publish in New Topics Twice

$$score(C) = (n-1)d^{\alpha}(coh)^{2\gamma}$$

$$score(C') = (n)d^{\alpha}(coh)^{2\gamma}$$

$$score(C'') = (n+1)d^{\alpha}(coh)^{2\gamma}$$

$$score(C'') - score(C') = score(C') - score(C)$$

From the derivation above, the performance of new metric in Axiom 1 and Axiom 3 should be worse than the metric with product. The performance of Axiom 2 should be better than the metric with product. So we construct a metric using product in the second term and compare the performance of it with the new metric in different parameter settings.

$$Breadth\ of\ Research = \sum_{i,j} d_{ij}^{\alpha}\ (p_i \times p_j)^{\beta}(Coh_i \times Coh_j)^{\gamma}$$

The results in Table 6 shows that the metric using summation outperforms product in Axiom 2, and metric using product outperforms summation in Axiom1, which is consistent with the results of derivation. But the results for the other three axioms are close between the two metrics, which means the interaction between different terms in the metric (distance term, distribution term and coherence term) will influence the results of simulation.

Table 6. Comparison between metric with summation and production.

Metric	*Parameter setting*	*Axiom1*	*Axiom2*	*Axiom 3*	*Axiom4*	*Axiom5*
Production	$\alpha = 0.1\ \beta = 0.1 \gamma = -0.1$	0.99	0.85	0.45	0.22	0.59
	$\alpha = 100\ \beta = 1 \gamma = -1$	0.82	0.62	0.47	0.69	0.53
	$\alpha = 1\ \beta = 1 \gamma = -10$	0.83	0.40	0.39	0.55	0.76
Summation	$\alpha = 0.1\ \beta = 0.1 \gamma = -0.1$	0.97	0.89	0.45	0.22	0.59
	$\alpha = 100\ \beta = 1\ \gamma = -1$	0.69	0.69	0.50	0.69	0.55
	$\alpha = 1\ \beta = 1 \gamma = -1$	0.69	0.47	0.41	0.54	0.77

Relationship between breadth of research and scientific impact

We tested the Pearson correlation between metrics of breadth of research and H-indexes of scholars. Our results (Table 7) show that some metrics have a positive relationship with H-index. Others have weak negative relationship. Because publication numbers may influence

the correlation between breadth of research and scientific impact i.e. the increase of numbers of publications may bring increase of breadth of research and increase of H-index simultaneously to make them positively correlated to each other, we test the partial correlation between metrics of breadth of research to H-index controlling publication numbers (Table 7). They are weaker than Pearson correlations. And all the weak partial correlation scores don't illustrate strong correlation between metrics for breadth of research and H-index for scholars.

Table 7. Correlation between breadth of research and H-index.

	Pearson Corr.	*Partial Corr.*
Entropy v.s. H-index	-0.1722	-0.0769
Simpson's v.s. H-index	0.2102	0.0922
GL Stirling v.s. H-index	0.4283	0.1820
New Metric v.s. H-index	0.4337	0.1832

The Variation of metrics over publication years

We illustrate in Figure 1 the variation of average scores of metrics for all the scholars over publication years. Simpson's, generalized Stirling and our new metric initially increase and then level off, which explains a possible publication pattern of scholars: scholars' breadth of research may increase with the increase of publications in the early stage of their career. But because of accumulation of publications, their accumulative breadth of research will not change dramatically in the late years. For the entropy metric with base n, it is normalized by topic number. So it keeps in a stable level over year, which shows a different pattern compared to other metrics.

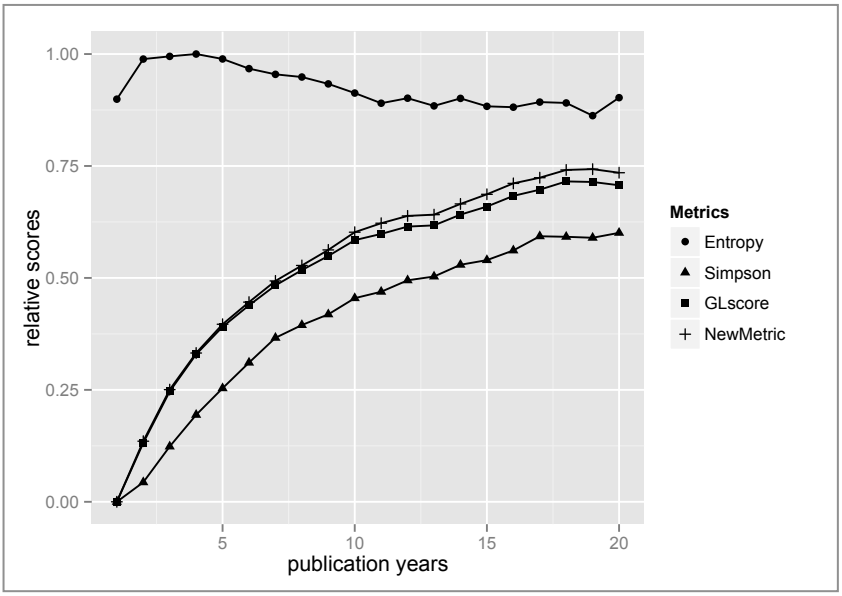


Figure 1. Variation of metrics over publication years.

Conclusion and Future Work

In this paper, we describe a new metric based on generalized Stirling to evaluate breadth of research for scholars in computer science. The metric makes use of topic distribution, similarity between topics, and coherence of topics and it can capture the diversity aspects of breadth of research. The simulation experiments show that traditional metrics can perform well in some axiom, but they don't perform well when coherence within topics and similarity between topics are considered. In contrast, generalized Stirling metric and the new metric for

breadth of research work better in the simulation related to similarity between topics and coherences but perform worse in the experiments of adding new topics. It is a trade-off between the simplicity of metrics and the concern of topic similarity and coherence.

With the new metric for breadth of research, we find the correlation between breadth of research and scientific metrics are weak, especially when we control publication numbers. From our study, there's no evidence to show whether the increase of breadth of research will influence the impact of scholars' publication. Also, after testing the variation of the new metric over years, we find a possible publication pattern of scholars: Breadth of research increases in the beginning with the increase of publications. But they increase slowly when publications have been accumulated.

There are a number of research questions that arise from the work described in this paper. The first one is finding alternative methods to generate research topics. Unsupervised learning models based on both text contents and citation information may be helpful to extract topics and show topic variation for authors. The second question is how to improve the simulation results for the new metric. The new metric performs better than general Stirling and other traditional metrics in some aspects. But if more information from co-author and citation network can be incorporated into the metric, the performance may be better and interpretable.

Acknowledgments

The research is funded by NSF 1258891 EAGER: Collaborative Research: Scientific Collaboration in Time.

References

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS One*, *4*(6), e6022. doi:10.1371/journal.pone.0006022

Boyack, K., & Klavans, R. (2009). Measuring multidisciplinarity using the circle of science. *From WRK1: Tracking and Evaluating Interdisciplinary Research, Workshop at ISSI*, *87122*.

Carayol, N., & Thi, T. (2005). Why do academic scientists engage in interdisciplinary research? *Vasa*.

Cassi, L., Mescheba, W., & de Turckheim, É. (2014). How to evaluate the degree of interdisciplinarity of an institution? *Scientometrics*. doi:10.1007/s11192-014-1280-0

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–6. doi:10.1126/science.1136800

He, B., Ding, Y., Tang, J., Reguramalingam, V., & Bollen, J. (2013). Mining diversity subgraph in multidisciplinary scientific collaboration networks: A meso perspective. *Journal of Informetrics*, 1–18.

Jensen, P., & Lutkouskaya, K. (2013). The many dimensions of laboratories' interdisciplinarity. *Scientometrics*, *98*(1), 619–631. doi:10.1007/s11192-013-1129-y

Jo, Y., Hopcroft, J., & Lagoze, C. (2011). The web of topics: discovering the topology of topic evolution in a corpus. *Conference on World Wide Web*, 257–266.

Karlovčec, M., & Mladenić, D. (2014). Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*. doi:10.1007/s11192-014-1355-y

Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics*, *7*(4), 924–932. doi:10.1016/j.joi.2013.09.002

Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science*, *59*, 1973–1984. doi:10.1002/asi.20914

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Dong, T. (2010). Community-based topic modeling for social tagging. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10*, 1565. doi:10.1145/1871437.1871673

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, 741–754. doi:10.1007/s11192-014-1319-2

Ponomarev, I. V., Lawton, B. K., Williams, D. E., & Schnell, J. D. (2014). Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? *Scientometrics*, 755–765. doi:10.1007/s11192-014-1320-9

Porter, A. L., Cohen, A. S., David Roessner, J., & Perreault, M. (2007). *Measuring researcher interdisciplinarity*. *Scientometrics* (Vol. 72, pp. 117–147). doi:10.1007/s11192-007-1700-5

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. doi:10.1016/j.respol.2012.03.015

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 1–28.

Roessner, D., Porter, A. L., Nersessian, N. J., & Carley, S. (2012). Validating indicators of interdisciplinarity: linking bibliometric measures to studies of engineering research labs. *Scientometrics*, *94*(2), 439–468. doi:10.1007/s11192-012-0872-9

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1118–23. doi:10.1073/pnas.0706851105

Ruscio, J., Seaman, F., D'Oriano, C., Stremlo, E., & Mahalchik, K. (2012). Measuring scholarly impact using modern citation-based indices. *Measurement: Interdisciplinary Research & Perspective*, *10*(3), 123–146. doi:10.1080/15366367.2012.711147

Silva, F. N., Rodrigues, F. a., Oliveira, O. N., & da F. Costa, L. (2013). Quantifying the interdisciplinarity of scientific journals and fields. *Journal of Informetrics*, *7*(2), 469–477. doi:10.1016/j.joi.2013.01.007

Simpsons, E. H. (1949). Measurement of Diversity. Retrieved October 9, 2014, from http://www.nature.com/nature/journal/v163/n4148/abs/163688a0.html

Sternitzke, C., & Bergmann, I. (2008). Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, *78*(1), 113–130. doi:10.1007/s11192-007-1961-z

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society, Interface / the Royal Society*, *4*(15), 707–19. doi:10.1098/rsif.2007.0213

Van Rijnsoever, F. J., & Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy*, *40*(3), 463–472. doi:10.1016/j.respol.2010.11.001

Velden, T., & Lagoze, C. (2013). The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology*, *64*(12), 2405–2427.

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. doi:10.1016/j.joi.2010.06.004

Weaver, W. (1949). Recent Contributions to the Mathematical Theory of Communication 1 Introductory Note on the General Setting of the Analytical Communication Studies.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, *62*(1), 117–131.

Yan, E. (2013). Finding knowledge paths among scientific disciplines. *arXiv Preprint arXiv:1309.2546*, (812), 1–31.

Transforming the Heterogeneity of Subject Categories into a Stability Interval of the MNCS

Marion Schmidt[1] and Daniel Sirtes[1*]

[1] *schmidt@forschungsinfo.de, sirtes@forschungsinfo.de*
iFQ Institute for Research Information and Quality Assurance, Schützenstr. 6a, D-10117 Berlin (Germany)

Abstract
The internal homogeneity of research disciplines in subject categories (SC) of the Web of Science database (WoS) regarding their publication and citation practices is an essential precondition for the field-normalization of citation indicators. This imperative of underlying homogeneity seems not to be met throughout all categories, as has been shown in former research. A keyword-based clustering method displays both the diversity of research areas included in an SC and that the clusters' mean citation rate differ substantially. This proof-of-concept paper on the basis of one country set and two SCs presents a bootstrapping method, which allows quantifying the degree of heterogeneity within subject categories as a stability interval. The MNCS 95% stability interval of our set has a range of 6.7% and 7.3% compared to its score. This kind of robustness measure could be implemented for future evaluative citation analysis in order to convey the coarseness of bibliometric point estimates.

Conference Topics
Methods and techniques; Citation and co-citation analysis; Indicators

Introduction
Field-normalized citation indicators such as the MNCS (Waltman, Eck, Leeuwen, Visser, & Raan, 2011) normalize the citation rate of a given publication corpus based on expectancy values of subject categories which correspond to the respective average citation rates within a research field (Vinkler, 1986; Mcallister, Narin, & Corrigan, 1983). Field normalization has been developed in order to neutralize the obvious diversity of publication and citation practices between field and subfields, as a corrective to otherwise unfair comparisons between the citation impact results of corpora with varying subject distributions.
Various methods for field delineation have been proposed (Glänzel & Schubert, 2003; Glänzel, Thijs, Schubert, & Debackere, 2009; Ruiz-Castillo & Waltman, 2014) including many proposals for clustering methods and arguments to determine the correct levels of aggregation. So far, however, no classification systems other than those provided by the database vendors could be established as standard throughout the bibliometrics community.
However, it is easily observable that the classification of the WoS subject categories diverges in size and specificity. Van Eck et al. (2013) provide furthermore strong evidence of heterogeneity within the medical subject categories along the characteristics of clinical and experimental research: After terms have been extracted from titles and abstracts, substructures are made visible by a term cloud procedure. These substructures can be assigned intellectually to clinical or experimental research and differ significantly in their citation rates along these dimensions. An intuitive explanation for this phenomenon would be the assumption that clinical researchers cite experimental studies, but that experimental researchers cite clinical studies only to a lesser extent.
Van Eck et al. (2013) draw the conclusion that the impact of clinical research is structurally underestimated by classical normalized citation indicators. The substructures made visible correspond to a facet that can be seen as transverse to a valid and comprehensible classification according to medical fields such as Clinical Neurology, Cardiac &

[*] The order of authorship is merely alphabetical.

Cardiovascular Fields, etc. Further theoretical issues beyond classification or clustering criteria seem to be not yet solved: If, for example, publications in so called hot topic areas are compared only with similar publications, even only with those who share not only the same topic, but also the same instruments, etc.? This could be seen as an over-normalization (Sirtes, 2012b; Sirtes, 2012a). Or is it legitimate to aggregate hot topics with less active research areas and thereby highlight the former as particularly successful? With the latter attitude the strategic decision of a researcher for a high impact research fields would be gratified while at the same time an implicit premise would be set that not all delineable areas in a functionally differentiated research landscape would be of equal value, insofar impact differences, which are effects of the functional differentiation, would not be neutralized.

By introducing finer classification systems these issues are addressed, although not answered based on theoretical reasons, as only further normalization options are created, whereas the resulting differences are not directly interpretable. Besides, in-house classifications systems are not easily compatible with a desirable trend towards greater standardization and reproducibility in the bibliometric community.

In the present paper we introduce a concept for quantifying heterogeneity differences within subject categories and thus maintain the WoS subject categories as basis for the field normalization, as they provide community-wide comparability and mutual reproducibility. Heterogeneity differences between subject categories are quantified and used to construct error or stability intervals, which can be integrated into the calculation of the total impacts of an institution or a country as before. The approach thus combines two advantages: on the one hand, we continue to work at the level of a standard classification system and on the other hand, underlying structures on a secondary level are made transparent.

Methods and Data

Keyword terms of all articles, reviews and letters published in journals of two medical subject categories (Parasitology (P), Otorhinolaryngology (O)) of the publication year 2008 have been extracted.[1] WoS keywords are not a controlled vocabulary like, e.g., Medical Subject Headings in PubMed/Medline and are therefore not per se complete and normalized. Table 1, however, shows that the amount of publications that have not assessed with keywords is relatively small. Keywords have, on the other hand, the advantage of simple accessibility; it is not necessary to exclude i.e. filler words. In order to accomplish a basic normalization, a stemming procedure is carried out which neutralizes different inflexions.

All distinct keyword terms are normalized with an Oracle Text stemming function and coupled by the *contains* function, again as provided in Oracle Text. Stemmed terms must therefore not be necessarily identical, but one term can contain the other, respectively. This also applies to keywords, which are phrases and may contain single keywords and be thus coupled with them. These keyword pairs are used for a coupling procedure of the corresponding publications; Salton's Cosine is used to neutralize differing amounts of keywords.

With the aim to reproduce the visual substructures of Van Eck et al. (2013) in a first step with our cluster procedure, these two subject categories have been chosen as they display different types of sub-structures in the discussed work. Parasitology displays quite distinct structures with three visible clusters seemingly characterized by significant differences in citation levels whereas Otorhinolaryngology displays a more fuzzy structure.[2]

[1] All calculations are processed in an Oracle database of WoS raw data (SCI, SSCI, A&HCI, CPCI-S, CPCI-SS) frozen in the 17th calender week 2013.

[2] http://www.neesjanvaneck.nl/basic_vs_clinical/

Table 1: Share of publications with keywords.

	Parasitology	*Otorhinolaryngology*
JARL 2008 (all)	3727	5122
JARL 2008 (percentage of publications with keywords)	98.0%	90.6%

The ratio of realized to theoretical possible relations between all items gives an impression about the broadness of the empirical basis of the coupling results. Table Table 2 gives the percentage of realized to theoretically possible relations of all publications (JARL = Articles, Letters and Reviews with publication type Journal Article) in 2008.

Table 2: Ratio realized relations to possible relations.

	Parasitology	*Otorhinolaryngology*
JARL 2008 (all)	18.2%	11.3%
JARL 2008 (only with keywords)	19.0%	13.8%

The resulting distance measures for publication pairs are imported into the statistical program R, converted into dissimilarity values and the clustering method Ward is used. Ward as a standard hierarchical-agglomerative clustering procedure was chosen, because it is crucial for our approach to have a clustering procedure which does not require a fixed number of clusters as parameter. Furthermore, single linkage with its well-known tendency to dilated cluster structures seems to impose to weak requirements on the clusters' homogeneity and complete linkage too strong requirements.
The usual cut-off-value of 5 was determined manually; however in future iterations of the procedure the optimal cut off value will be estimated.
As shown in Table 2 not all publications in the respective sets are actually assigned with keywords, thus we have added a non-keyword cluster with its mean citation rate in order to represent all publications in our dataset. This appears as a legitimate solution given that fact that non-keyword items have considerably smaller mean citation rates compared to the whole subject category and have to be taken into account in order to appropriately represent the SC.

Results

The visualization for the subject category parasitology as resulting from (Van Eck et al.., 2013) indicates a distribution of three discernable substructures which are clearly different in citation level. With our method, we arrive at eleven clusters. Table 3 shows four of the top keywords[3] and the respective mean citation rates, whereas Figure 1 gives the frequency distribution of the clusters (as the width of the bars) and the mean citation rates in a histogram. The topics of the clusters can only partially confirm Van Eck et al.'s conclusion. The keywords of cluster 5, 6, and 7 have all clear connection to experimental laboratory research, however only 5 (with the most distinctly molecular biology focus) has a very high citation rate compared to the rest. It is possible, that parasitology is rather a special case compared to other medical SCs, as it also encompasses topics such as classical biology (cluster 1), epidemiology (clusters 2 and the more clinical 4), a veterinary cluster (8), and clusters that are joined by common parasites (3, 9,10, and 11).

[3] All keywords were in the top 10 most frequent ones. Redundant keywords (like 'plasmodium' and 'plasmodium falciparum') and keywords that were not informative in understanding the topic of the cluster (like 'parasites') were excluded.

Table 3 - Top keywords and mean citation rate of keyword clusters in parasitology (ordered by cluster size).

Cluster	Top Keywords				Mean Citation Rate
1	Phylogeny	Evolution	Ecology	Morphology	3.91
2	Infection	epidemiology	Seroprevalence	Antibodies	5.76
3	Malaria	plasmodium falciparum	infected erythrocytes	cerebral malaria	6.25
4	Transmission	Children	Resistance	Efficacy	7.02
5	Expression	in-vitro	Protein	gene-expression	7.57
6	Mice	in-vivo	dendritic cells	immune-response	6.69
7	Identification	PCR	linked-immunosorbent-assay	Antibodies	5.50
8	Sheep	Cattle	haemonchus-contortus	Ivermectin	4.11
9	Disease	trypanosoma cruzi	chagas disease	risk-factors	6.09
10	Diptera	Culicidae	aedes-aegypti	anopheles-gambiae	5.32
11	Cryptosporidium	Parvum	Giardia	Genotypes	7.88

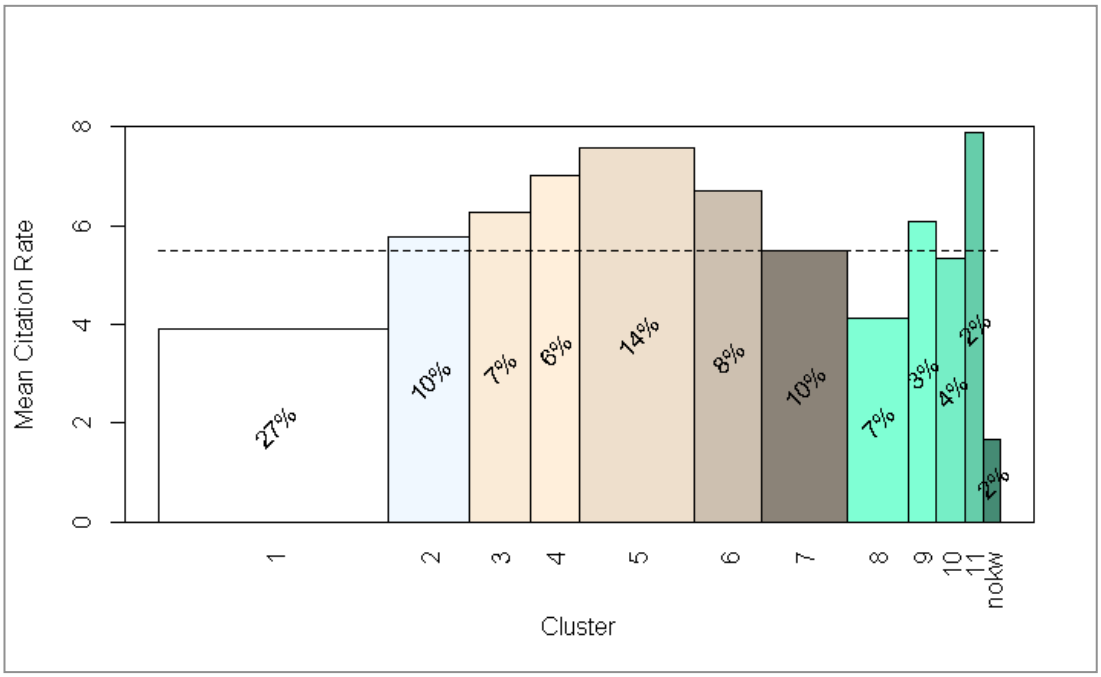


Figure 1: Share and Mean Citation Rate of Parasitology Clusters. The dotted line represents the MCR of the whole SC.

In the second case of otorhinolaryngology, the structure shown by (Van Eck u. a., 2013) is quite fuzzy and less-structured, which is mirrored by our cluster distribution. It consists of one larger and a considerable amount of very small cluster. There are also significant variations between mean citation levels ranging from around 2 to larger than 4, it is however more difficult to interpret the cluster's respective keyword frequencies.

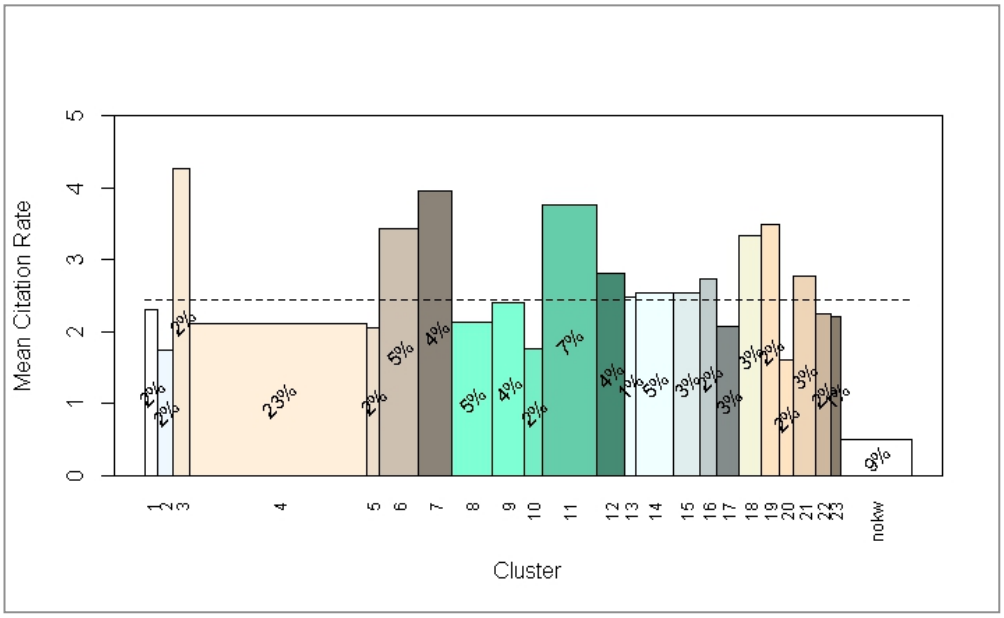


Figure 2: Share and Mean Citation Rate of Otorhinolaryngology Clusters. The dotted line represents the MCR of the whole SC.

In order to calculate the MNCS and its stability, sets of publications with an affiliation in Germany have been selected. The size of the sets were 208 (P) and 486 (O) publications respectively.

On the basis of the resulting cluster distributions, a bootstrapping approach has been utilized.

A set of MCR clusters equal to the size of the German set has been drawn with replacement from the clusters' MCRs with the probabilities equal to the clusters' share. The arithmetic mean of this combination has been calculated and served as the Expected Citation Score (ECS_i). Each raw citation score of the German papers was then divided by the ECS_i and the arithmetic mean of the results delivered the $MNCS_i$. 10'000 iterations of this procedure have been executed. The distribution of the scores are depicted in Figure 3.

Finally, the 2.5% and 97.5% quantiles of this distribution have been calculated.

The resulting MNCS 95% stability interval of the German set for parasitology ranges from 1.35 to 1.46 with an MNCS of 1.40 and for otorhinolaryngology from 0.87 to 0.93 with an MNCS of 0.9. Thus, although parasitology displays a much wider distribution, as can also be seen in Figure 3, the relative deviance of the MNCS ([95% range of $MNCS_i$]/MNCS) is quite similar with 7.3% and 6.7%, respectively.

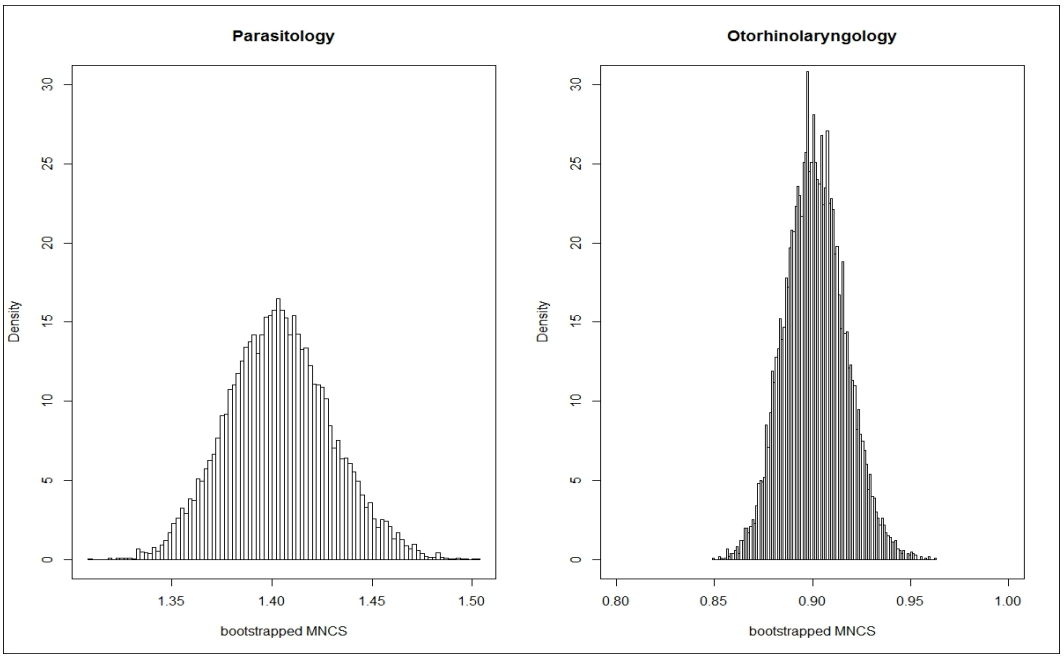


Figure 3: Distribution of $MNCS_i$ for German publications.

Discussion

These preliminary results show in the case of parasitology that clusters can be delineated and differing topical foci can be identified as well. While a dimension clinical versus experimental research is perceivable, other facets also occur: It may be the case that parasitology is a special SC as the clusters have also rather unusual topics compared to other medical disciplines such as classical biology, veterinary sciences and epidemiology. The Mean Citation Rates vary massively with a total range of MCRs of 3.97 citations per publication In the second case of otorhinolaryngology, the cluster distribution is less harmonic, more frayed out and not easily interpretable (confirming here the results of (Van Eck et al., 2013)). The coupling procedure succeeded on a relatively smaller amount of publications and many more clusters have been created. Furthermore, the citation levels are all much lower and the range of MCRs, the publications without keywords notwithstanding, have only a total range of 2.6 citations per publication.

The hitherto work was intended as a proof of concept: We were able to show that subject category substructures with different citation levels exist. Differences in citation homogeneity are however not in both cases concordantly attributable to topical structures. For the current state of this work, some simplifications have been applied: Citation rates should be processed and normalized document type-specific as articles, letters and reviews are cited differently. However, citation level differences in our results are so clear and dominant that they couldn't possibly only be caused by different document type patterns in the clusters. For a final implementation of this method, the calculations will be processed document type-specific and the expansion of the method to sets of multiple SCs, including an SC fractionalization will be developed. An exclusion of letters might be contemplated as for example about half of the publications without keywords in otorhinolaryngology are letters (about three quarters of all letters in this SC). Furthermore, parameters of the study like the clustering method and definition of cut off-values will be systematically varied and analyzed. It is even conceivable to calculate such stability intervals on the basis of percentile based indicators, which are less sensitive to outliers than the MNCS. However, already as it stands this method shows promise in circumventing to problem of calculating normalized citation scores on non-standard classification schemes while taking into account the heterogeneity of research areas in the

classical WoS SC classification. This method could also be combined with already existing bootstrapping methods of the publications sets themselves as implemented for example in the Leiden Ranking (www.leidenranking.com). Together they could account for both the robustness of the citation scores given the size and distribution of the publication sets themselves, as well as the underlying uncertainty of the expected citation rates. We believe that such methods that display the coarseness of bibliometric point estimates, which especially clients of evaluative bibliometric analyses are prone to disregard and thus revel or despair at minute changes of their scores and ranks, are an important step to the correct interpretation of bibliometric indicators and crucial for the development of bibliometrics into a mature science.

References

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367. http://doi.org/10.1023/A:1022378804087

Glänzel, W., Thijs, B., Schubert, A., & Debackere, K. (2009). 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), 165–188. http://doi.org/10.1007/s11192-008-2109-5

Mcallister, P. R., Narin, F., & Corrigan, J. G. (1983). Programmatic evaluation and comparison based on standardized citation scores. *IEEE Transactions on Engineering Management EM*, *30*(4), 205–211. http://doi.org/10.1109/TEM.1983.6448622

Ruiz-Castillo, J., & Waltman, L. (2014). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Working Paper*. Abgerufen von http://e-archivo.uc3m.es/handle/10016/18385

Sirtes, D. (2012a). Finding the Easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. *Journal of Informetrics*, *6*(3), 448–450.

Sirtes, D. (2012b). How (dis-)similar are different citation normalizations and the fractional citation indicator? (And How it can be Improved). In *Proceedings of 17th International Conference on Science and Technology Indicators* (S. 894–896). Montréal: Éric Archambault, Yves Gingras, and Vincent Larivière.

Van Eck, N. J., Waltman, L., Van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE*, *8*(4), e62395. http://doi.org/10.1371/journal.pone.0062395

Vinkler, P. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics*, *10*(3-4), 157–177. http://doi.org/10.1007/BF02026039

Waltman, L., Eck, N. J. van, Leeuwen, T. N. van, Visser, M.S., & Raan, A. F. J. van. (2011). Towards a new crown indicator: an empirical analysis. *Scientometrics*, *87*(3), 467–481. http://doi.org/10.1007/s11192-011-0354-5

Measuring Interdisciplinarity of a Given Body of Research

Qi Wang

qi.wang@indek.kth.se
KTH Royal Institute of Technology, Lindstedsvägen 30, SE-100 44 Stockholm (Sweden)

Abstract
Identifying interdisciplinary research topics is an essential subject, not only for research policy but also research funding agencies. Previous research was constructed on measuring interdisciplinarity mainly at the macro level of research, such as Web of Science subject category and journal. However, these studies lack analysis at the micro level of the current science system. It means few studies have analyzed interdisciplinarity at the level of publications. To cover this gap, we introduce an approach for measuring interdisciplinarity at the level of micro research topics. The research topics are clustered by direct citation relations in a large scale database. According to the characteristics of boundary-crossing research, we provide an alternative approach to measure interdisciplinarity. Comparing with the widely used Rao-Stirring indicator (Integration score), we found that the results obtained by two indicators of interdisciplinarity have a strong correlation, thus we believe that this approach could effectively identify boundary-crossing research topics.

Conference Topic
Indicators

Introduction
In bibliometric and scientometric research, measuring interdisciplinarity is a difficult yet important topic. However, although it has been widely recognized that interdisciplinary research solves complex problems, promotes scientific developments and innovations, there is still no consensus on how to define and measure this type of research. Specifically, a variety of definitions on boundary-crossing research have been proposed, such as interdisciplinary multidisciplinary, transdisciplinary and cross-disciplinary; however the definitions of each term as well as discriminations among them are quite ambiguous (for more details see Huutoniemi K. et al., 2010; Wagner C.S. et al., 2011). In a broad sense, these concepts all refer to the research that cross boundaries between disciplines. We do not intend to explore the nuances among the concepts in this study. Thus, at the very beginning of this article we need to emphasis that, for the purpose of this research, the term *interdisciplinary research topics* used to refer to all type of boundary-crossing research, in other words, it covers all type of research with interdisciplinarity.

Furthermore, due to the controversy in defining research with interdisciplinarity at the conceptual level, there is no consensus on how to measure interdisciplinarity in practices. Various approaches are utilized to analyze interdisciplinarity, including both quantitative methods such as bibliometric indicators, text-mining and qualitative methods such as interviews and surveys. In particular, bibliometric approaches have been widely applied to measure and identify interdisciplinarity, such as citation-based indicators (Porter & Chubin, 1985; Leydestorff, 2007; Porter, Roessner & Heberger, 2008; Porter & Rafols, 2009; Rafols & Meyer, 2010; Leydestorff & Rafols, 2011; Rafols et al., 2012; Lariviere & Gingras, 2014), author-based indicators (Qin et al., 1997; Schummer, 2004; Abramo et al., 2012), as well as similar indicators but relying on a variety of classification systems of science (Tijssen, 1992; Morillo, Bordons, & Gomez, 2001; 2003; Braun & Schubert, 2003; Sugimoto, 2011;

Sugimoto et al., 2011). Additionally, a few studies have applied text-mining approaches, LDA for example, to explore interdisciplinarity of a given issue (Wang et al., 2013; Nichols, 2014).

In this article, we explore a citation-based measurement for identifying interdisciplinary research topics at the level of publications. We also use the Web of Science (WoS) classification system, but with a different approach. More specifically, we first construct micro research topics based on the direct citation relations among individual publications. Meanwhile, the publications are assigned into one or several subject categories on the basis of the journal where the publication has appeared and of WoS classification system. It implies that a research topic constructed might belong to one or several WoS subject categories according to publications within the cluster. In other words, WoS subject categories that attached to publications are regarded as traditional boundaries of scientific disciplines, whereas micro research topics constructed on the relatedness among publications might break the existing knowledge boundaries. We assume, then, that a cluster can be regarded as an interdisciplinary research topic if there is a considerable number of within-cluster citations spanning distant WoS subject categories. The indicator proposed in this article combines knowledge diversity with knowledge integration, in which heterogeneity and connectedness of subject categories within research topics are taken into account. It provides an alternative approach to measure interdisciplinarity and simplifies the previous citation-based approaches.

Data and Methodology

This study was based on data from the in-house WoS database of the Centre for Science and Technology Studies (CWTS) of Leiden University. The database used in this study covers the period from 2002 to 2013, a 10-year period. The total number of publications in our database is about 9 million. The methodology that we introduce for measuring interdisciplinarity of micro research topics can be divided into three steps.

Step 1 Clustering publications into micro research topics

The clustering method is mainly based on the previous studies by Waltman & van Eck (2012; 2013). First, the relatedness of publications was measured by the normalized direct citation relation among individual publications (for details see Waltman & van Eck, 2012). Furthermore, based on the relatedness matrix, an improved Louvain algorithm (Blondel et al., 2008), namely a 'Smart Local Moving algorithm' (SLM) was applied to cluster individual publications (for details see Waltman & van Eck, 2013). Labels of each cluster were selected from titles and abstracts of publications within cluster (for details see Waltman & van Eck, 2012).

Measuring interdisciplinarity on the level of micro research topics, constructed based on the citation relations, is one of the most important distinctions between this study and previous research. There are two reasons for measuring the degree of interdisciplinarity in this approach. First, WoS subject categories attached to journals cannot properly describe publication itself. For instance, although *Journal of the Association for Information Science and Technology* belongs to two categories, INFORMATION SCIENCE & LIBRARY SCIENCE and COMPUTER SCIENCE, it does not necessarily mean that all publications appeared in this journal span the two categories. More generally, some publications associated with the category of INFORMATION SCIENCE & LIBRARY SCIENCE and others related to the category of COMPUTER SCIENCE. The second reason is that WoS assigned journals such as *Nature*, *Science*, and *Plos One* as MULTIDISCIPLINARY SCIENCE. Instead of focusing on a specific scientific field, this sort of journals covers almost the full range of scientific disciplines. When measuring interdisciplinarity on the level of journals, this sort of journals may have high interdisciplinarity scores. However, although the journals are composed of publications

spanning over different scientific disciplines, it does not necessarily mean the integration of knowledge from various sources exists.

In order to avoid the problems mentioned above, we constructed micro research topics based on the relatedness of individual publications, which are expected to provide a more accurate body of research topics within the current science system.

Step 2 Calculating a similarity matrix of ISI subject categories

Porter and Rafols (2009) analyzed a sample of more than 30,000 WoS publications and their cited references, in which publications were assigned to subject categories on the basis of the WOS classification of journals the publications appeared. They constructed a matrix of subject categories using the relations of articles and their cited references, and then applied Salton's cosine (Salton & McGill, 1983) to obtain the similarity matrix of subject categories. The similarity value s_{ij} is high if subject category i and j are cited a lot by the same publications.

However, in this study, two subject categories are considered to be strongly related if they both cite a lot to the same subject categories. Specifically, the construction of a similarity matrix of subject categories is done in two steps.

In the first step, for each pair of a citing subject category i and a cited subject category j, the number of citations from publications in subject category i to publications in subject category j is counted. We use c_{ij} to denote the number of citations from publications in subject category i to publications in subject category j. Note that according to the WoS classification system, one journal might be attributed into multiple subject categories. Therefore a fractional counting strategy is adopted to handle publications belonging to more than one subject category.

The second step is to construct a similarity matrix of subject categories based on the citation matrix created in the first step. The cosine similarity measure is used for this purpose. Hence, the similarity of two subject categories i and j is given by

$$s_{ij} = \frac{\sum_k c_{ik} c_{jk}}{\sqrt{\left(\sum_k c_{ik}^2\right)\left(\sum_k c_{jk}^2\right)}}$$

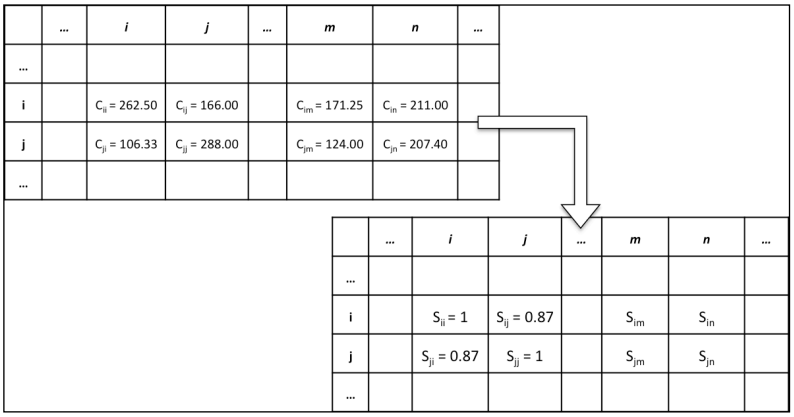


Figure 1. An example of the formula for calculating similarity.

Figure 1 can be used as an example to illustrate how the formula of similarity applied. The top left table is the matrix of citation relations among subject categories, which is not symmetric. Since a fractional counting strategy is used in this study, the numbers of citations are not always integers. As we mentioned above, c_{ij} means the number of citations from subject category i to j. Moreover, according to the above formula, we obtained the symmetric

similarity matrix of subject categories, which is shown in lower right of figure 1. In this case, subject category *i* and *j* are all cite a lot to the categories *i*, *j, m* and *n.* Therefore, the similarity between *i* and *j* is quite high, that is 0.87.

In short, using the cosine similarity measure, s_{ij} is high if publications in the two categories tend to cite the same categories. If publications in two subject categories tend to cite completely different categories, the similarity between the categories is low.

Step 3 Determining the degree of interdisciplinarity

As mentioned above, we suppose that a research topic could be regarded as an interdisciplinary research topic should satisfy two criteria; one is that it contains distant subject categories, the other is there are citation relations among different subject categories within this topic. In short, a cluster that is consisted with citation relations spanning different subject categories might be an interdisciplinary research topic.

Following the criterion discussed above, we explore the indicator to measure interdisciplinarity, whose formula is as follows:

$$Interdisciplinarity = \frac{1}{n\_cit}\sum_i^k \sum_j^k n\_cit_{ij} d_{ij},$$

where $d_{ij} = 1 - s_{ij}$. Within a cluster, n_cit_{ij} is the number of citations between subject categories i and j, and n_cit is the sum of citations obtained by $n\_cit = \sum_i^k \sum_i^k n\_cit_{ij}$. The indicator includes three attributes: *variety*, the number of subject categories within a cluster (denoted as *k*), *connectedness*, the number of cross-citations (denoted as n_cit_{ij}) and *distance*, the degree of distinctiveness between subject categories (denoted as d_{ij}). In short, a research topic can be considered to be more interdisciplinary if the citation relations within that cluster cross various WoS subject categories.

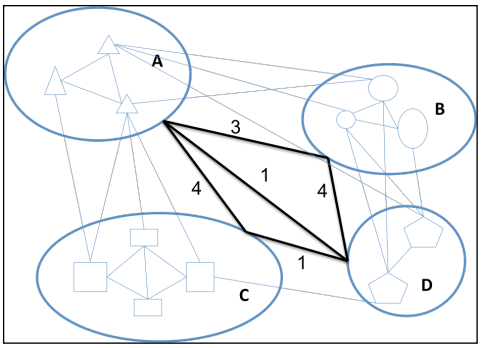


Figure 2. An example of the citation relations within a research topic.

Figure 2 shows a research topic including 12 publications that belong to 4 subject categories. The black lines represent the citation relations among different subject categories, and the blue lines are the links within the same category. In our measurement, the citations crossing subject categories (black lines in the Figure) and distances of subject categories are taken into account.

Results

Clustering analysis

Table 1 provides the basic statistic results of original and restricted database. The restricted database was constructed based on two criteria. First, we expect to analyze research topics with a relatively large number of publications only. Therefore, we set a restriction on the number of publications of each cluster so that clusters with more than 100 publications could

be advanced in the next step. Second since the accuracy of measurement is highly related to the quality of clustering results, we reviewed the clusters with the indicator, *mean citation score*. It obtained by using the total number of citations divided by the total number of publications within a cluster. If the number of citations is less than the number of publications of a cluster, publications belong to the cluster are connected loosely, resulting in the emergence of clusters with poor qualities. In this case, we found 667 clusters with low mean citation scores (defined as less than 2), which accounted for 7% of the total. Thus, it turns out that most of clusters have relatively strong interconnections. The analysis in the following sections is performed base on the restricted database.

Table 1. Basic statistic results of original and restricted database.

	# of pubs	# of topics	Average pubs	Max pubs	Min pubs	St.d pubs
Original	9,146,302	9,565	956	10744	1	1026
Restricted	8,930,360	7,864	1,135	10744	100	1040

Similarity matrix

Using Salton's cosine (Salton & McGill, 1983), we obtained a similarity matrix of WoS subject category, the range of similarity values is between 0 and 1. It implies that the similarity s_{ij} is zero if subject category i and j never cite to the same categories, whereas s_{ij} approaches one if they both cite a lot to the same categories. To test the accuracy and reliability of our similarity matrix, we have compared it with the one obtained by Porter & Rafols (2009), whose method have been introduced above. As expected, the result shows there is positive correlation between the two matrices ($r = 0.7405$). In general, we believe that the results obtained from the two approaches with slight differences are consistent.

Interdisciplinarity of research topics

The average interdisciplinarity score of each research topic is about 0.42 with a standard deviation of 0.11. The largest score is 0.72 associated with the research on respiratory system, while the lowest is close to 0.0086. The distribution of research topics over the interdisciplinarity score is shown in figure2. As can be seen, the majority of research topics have interdisciplinarity scores between 0.35 and 0.55.

In order to better interpret the results, we aggregated the WoS subject category into five main fields according to the Leiden Ranking 2013. Table 2 lists the five main fields. Specifically, a publication appearing in one or several main fields is based on the journal where it has been published. When a publication has appeared in a journal of multi-assignation and these subject categories are assigned into different main fields, the publication is expected to appear in more than one field (more details see CWTS Leiden Rank 2013, pp4). Thus, a research topic might be assigned into several main fields if the publications within this topic belong to more than one field.

Before turning to the interdisciplinarity score, we emphasize that it is quite difficult and almost impossible to define a clear cutting-off point between interdisciplinary and non-interdisciplinary research topics. Considering the difficulty, we selected the research topics with an interdisciplinarity score greater than 0.6143, which account for around 1% of the total. For the purpose of understanding the knowledge integration across main fields in the macro level, we applied following strategy. Regarding a research topic, if the number of publications in one main field is larger than 50% of the total, then the topic is assigned into this main field. Otherwise, the research topic would be assigned into its two dominant main fields. In doing so, the select topics (top 1% of the total) are tabulated in Table 3, in which each row is the main field with the most number of publications and each column is the main field holding the second number of publications. For instance, in the first row, 1 means there

is one research topic whose publications mostly appear in main fields 1 and 2, as well as main field 1 has the most number of publications.

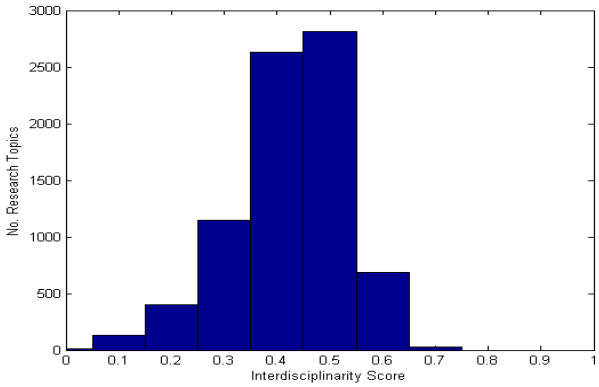


Figure 3. Distribution of research topics over interdisciplinarity score.

Table 2. Labels of main fields.

ID	*Labels of Main Fields*
Main Field -1	Social sciences & humanities
Main Field -2	Biomedical &health sciences
Main Field -3	Natural sciences & engineering
Main Field -4	Life & earth sciences
Main Field -5	Mathematics & computer science

Table 3. Distribution of research topics over the main fields.

	Main field-1	Main field-2	Main field-3	Main field-4	Main field-5	Total
Main field-1	11	1	0	0	0	12
Main field-2	1	33	6	1	2	43
Main field-3	0	2	25	1	0	28
Main field-4	0	0	1	8	0	9
Main field-5	0	2	1	0	5	8

As can be seen, most research topics in the top 1% of the total belong to the main field 2, that is BIOMEDICAL & HEALTH SCIENCES. Meanwhile, among the research topics that across two main fields, the topics whose publications mainly appear in the main field 2 contribute the largest proportion. Primarily, this is because the most number of research topics fall into this main field. In addition, the research conducted by Porter & Rafols (2009) have demonstrated that subject categories MEDICINE- RESEARCH & EXPERIMENTAL and NEUROSCIENCES have high degrees of interdisciplinary according to the Integration score (aka, Rao-Stirling's diversity) (more details see Porter & Rafols, 2009, pp723). In our classification system, the two subject categories both belong to main field 2, which is partially verified that the main field of BIOMEDICAL &HEALTH SCIENCES has relatively high interdisciplinarity. Main field 5, that is MATHEMATICS & COMPUTER SCIENCE, holds the smallest number of research topics with high interdisciplinarity, as shown in table 3. This result is also consist with the research by Porter & Rafols (2009), in which they showed subject category MATHEMATICS that is assigned into main field 5 in our study has the lowest integration score between 1975 and 2005.

For the purpose of examining the quality of the indicator, we now take a more derailed look at research topics. In doing so, we randomly select 5 research topics from the top 1%, one from

each main field. For each research topic, Table 4 gives the three most important subject categories and the two most cited publications.

Table 4. Selected research topics with high interdisciplinarity.

Cluster ID	Information of Publication	
4323	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -1 (53%); Main Field -4 (27%) 705 56 VETERINARY SCIENCES (244); SOCIOLOGY (225); PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH (47) Rijken M et al. (2005). Comorbidity of chronic diseases - Effects of disease pairs on physical and mental functioning (88) Odendaal J.S.J. & Meintjes R.A. (2003). Neurophysiological correlates of affiliative behaviour between humans and dogs (82)
3644	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -2 (54%); Main Field -3 (25%) 875 36 RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING (715); NUCLEAR SCIENCE & TECHNOLOGY (533); ENVIRONMENTAL SCIENCES (464) Stabin M.G. et al. (2005). OLINDA/EXM: The second-generation personal computer software for internal dose assessment in nuclear medicine (370) Gorden A.E.V. et al. (2003). Rational design of sequestering agents for plutonium and other actinides. (227)
4083	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -3 (74%); Main Field -2 (13%) 760 63 NUCLEAR SCIENCE & TECHNOLOGY(282); INSTRUMENTS & INSTRUMENTATION (259); PHYSICS, NUCLEAR (255) Spalding K.L. et al. (2005). Retrospective birth dating of cells in humans (182) Lappin G. & Garner R.C. (2003). Big physics, small doses: the use of AMS and PET in human microdosing of development drugs (137)
7577	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -4 (50%); Main Field -3 (46%) 190 26 ASTRONOMY & ASTROPHYSICS(100); GEOSCIENCES, MULTIDISCIPLINARY (81); METEOROLOGY & ATMOSPHERIC SCIENCES (67) Rietveld M.T. et al. (2003). Ionospheric electron heating, optical emissions, and striations induced by powerful HF radio waves at high latitudes: Aspect angle dependence (91) Pedersen T.R. et al. (2003). Magnetic zenith enhancement of HF radio-induced airglow production at HAARP (45)
8434	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -5 (55%); Main Field -3 (34%) 108 99 ROBOTICS (49); COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (34); INSTRUMENTS & INSTRUMENTATION (22) Vergassola M. et al. (2007) 'Infotaxis' as a strategy for searching without gradients (103) Yoerger D.R. et al. (2007). Techniques for deep sea near bottom survey using an autonomous underwater vehicle (38)

Take two clusters as examples, cluster 3644 and cluster 4083 are randomly selected from BIOMEDICAL & HEALTH SCIENCES and NATURAL SCIENCES & ENGINEERING respectively; however, the two most frequent main fields of both clusters are the same. Apart from that, as can be concluded from table 4, most publications of both clusters belong to the

subject category of NUCLEAR SCIENCE & TECHNOLOGY. Hence we infer that the two research topics are similar at a certain degree. Observing the detailed information of publications in each cluster, we found that both clusters are related to the research on nuclear medicine, that is "a medical specialty involving the application of radioactive substances in the diagnosis and treatment of disease"[1]. However, there is a considerable difference in terms of the degree of interdisciplinary score. Cluster 3644 is much more interdisciplinary than cluster 4083 as shown from table 4. To understand the differences, we visualized the two clusters using the map of subject categories.

The map of subject categories can represent the position of a cluster in the global map of science, as well as show whether the cluster has the characteristics of interdisciplinary research. For instance, we can observe from the map of subject categories whether clusters are dispersed over many distant subject categories. The software VOSviewer (van Eck & Waltman, 2010) was used to construct the map of subject categories. In this study, the baseline map was generated by the citations between WoS subject categories using publications from 2002 to 2013. Figure 4 and 5 were generated by overlaying on the baseline map with circles, in which size of circles represents the number of publications in each WoS subject category, nodes represent subject categories, as well as links shows citations among them.

Comparing the two figures, we found that cluster 3644 are more diverse that it contains citations spanning various subject categories with larger distances (i.e. COMPUTER SCIENCE THEORY AND METHOD, ENGINEERING ELECTRICAL AND ELECTRONIC), as well as its number of publications in various subject categories are quite even. Thus, it is reasonable that cluster 3644 has a higher interdisciplinary score than cluster 4083, although they have a similar research topic. Meanwhile, it can be inferred that the two clusters have different research focuses since the subject categories with the most number of publications of the two clusters are quite different. That also explains why publications with a similar research topic were classified into two clusters.

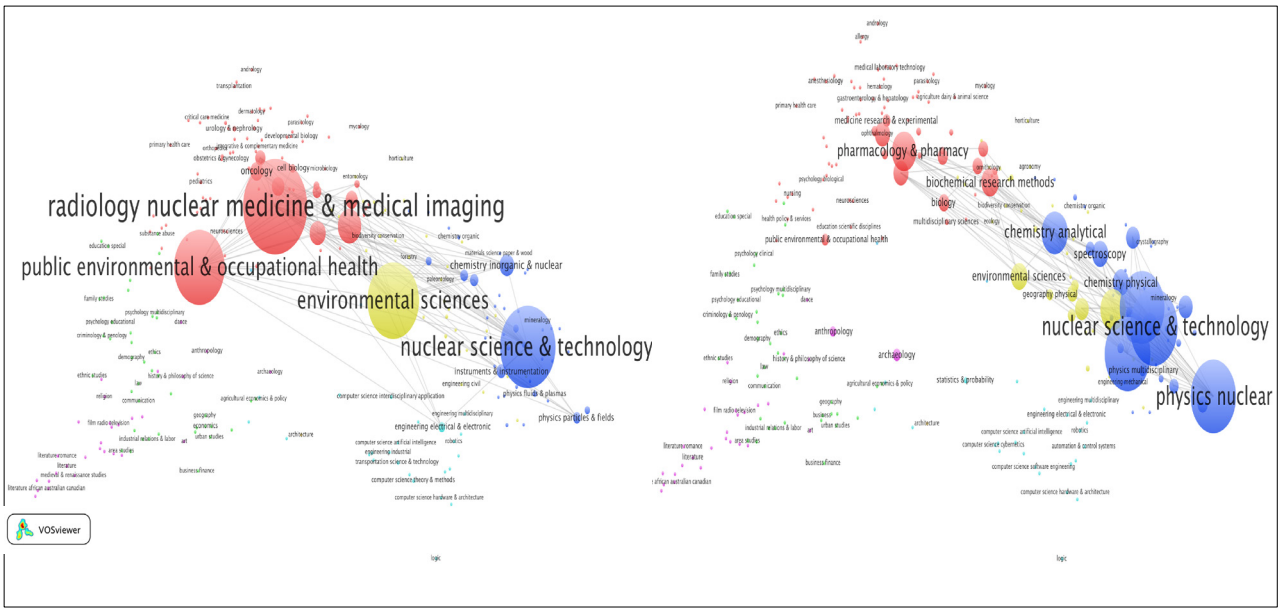


Figure. 4. A map of subject categories (note: the left panel is cluster 3644; the right panel is cluster 4083).

[1] http://en.wikipedia.org/wiki/Nuclear_medicine.

An example of Information Science and Library Science. Readers of this paper might be familiar with research in the field of information and library science; therefore, we now take a specific look at a cluster in this subject category. To give an example, we select the cluster that holds the highest interdisciplinarity value among all the clusters whose most publications belong to this subject category. In doing so, we obtained cluster 4982, which ranks 72 among the top 1% most interdisciplinary clusters. The detailed information of this cluster is shown in table 6.

As can be seen, the cluster includes 565 publications, and most of them belong to main fields of SOCIAL SCIENCES AND HUMANITIES and MATHEMATICS AND COMPUTER SCIENCE, that fit what figure 10 shows. Moreover, it also can be seen that this research topic covers various subject categories, such as computer science research, ergonomics, business, laws, and psychology. Furthermore, based on the most cited publications and the figure of citation network of this cluster, we can estimate that this research topic is rated to the research on information privacy. This is probably in line with what our cognition, that research on information privacy involves studies on either information or computer technology, or social science research such as law and psychology, or studies which overlap the two types of research.

To find more evidence, we searched the courses related to information privacy in MIT OpenCourseWare, using "information privacy" as the key words. Then, 1150 results have been obtained. The courses include from *The Economics of Information*, *Communications and Information Policy* to *Biomedical Computing*, *Information and Entropy*. That proves the research topic of information privacy is interdisciplinary in character.

Table 5. Publication information of cluster 4982.

Cluster ID	Information of Publication	
4982	Main field (R_pubs)	Main Field -1 (52%); Main Field -5 (44%)
	T_pubs	565
	Rank	72
	Subject Categories (N_pubs)	COMPUTER SCIENCE, INFORMATION SYSTEMS (141); BUSINESS (108); INFORMATION SCIENCE & LIBRARY SCIENCE (107)
	Title (Times cited)	Malhotra N.K., Kim S.S. & Agarwal J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model (169)
		Nissenbaum H. (2004). Privacy as contextual integrity (110)

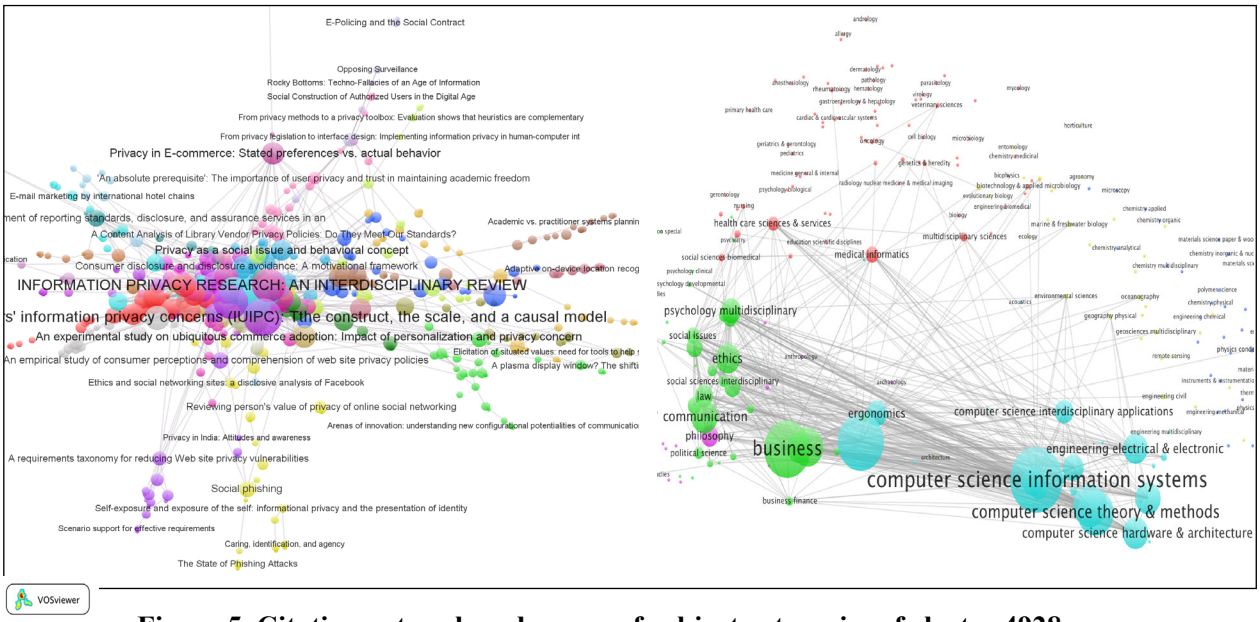


Figure 5. Citation network and a map of subject categories of cluster 4928.

Discussion and Conclusion

In this article, we proposed an alternative approach to investigate interdisciplinarity. The measurement is based on a publication-level and direct citation relations based classification system. Hence, several interdisciplinarity research topics were identified with the new interdisciplinarity score in the current science system.

The interdisciplinarity score proposed not only takes citation relations among various WoS subject categories within a cluster into consideration, but it incorporates a measure of how distant the subject categories. As mentioned above, the indicator proposed in this article is similar, to some extent, with the widely used indicator of interdisciplinarity, that is Rao-Stirling index or Integration score (Porter & Rafols, 2009). The most crucial distinction between the two indicators of interdisciplinarity is that, for each research topic, we use the number of citations among subject categories instead of the number of publications in different subject categories. We consider that the number of citations among subject categories can reflect both how diverse as well as how compact a cluster is. Furthermore, to test the robust of this approach, we estimated Pearson's correlation between the two indicators. The correlation coefficient is 0.9552, which high correlation suggests that there is no difference between the original Rao-Stirling index and the variant proposed in this article.

Another distinction with previous research is that our study is based on a publication-level and direct citation relations based classification system, in which publications were assigned into different research topics according to their citation relations. It implies the research topics constructed can more closely match the current structure of scientific research and provide more detailed information of the research content per se (Waltman & van Eck, 2012). There are 250 WoS subject categories in total, providing a coarse description of science. On the contrary, we worked on a classification with around 10,000 research topics, deriving from large-scale clustering. While the clusters in this study are small compared with WoS classification, it is important and necessary to explore interdisciplinary research topics at different level of classification system of science.

Moreover, we need to emphasis the concept of 'interdisciplinary research topic' that we used in this article again. Here, this term is related to all types of crossing boundary research topics, which can be considered as a loose standard. Since there is a gradual transition from

mono-disciplinary to interdisciplinary research, it is somewhat impossible to define a clear line to distinguish mono-disciplinary and interdisciplinary related research.

In summary, we have introduced an alternative approach for identifying interdisciplinary research topics. By in-depth analysis of some randomly selected topics, especially based on citation networks and overlay maps, we believe that they are boundary-crossing research topics. Since most research on the measurement of interdisciplinarity have conducted based on an existing classification system of science, such as journal and WoS subject category, we expect this study could provide another perspective on the current science system. The identified research topics could more accurately reveal interdisciplinary research within the current structure of scientific research.

Acknowledgments

This paper was written during a research stay at the Centre for Science and Technology Studies (CWTS) of Leiden University. I acknowledge the support of CWTS. I would like to thank Ludo Waltman for the extremely helpful discussions and suggestions on this study. I also appreciate Ulf Sandström for his comments on the early version.

References

Abramo G., D'Angelo, C.A., & Costa, F.D. (2012). Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications. *Journal of American Society for Information Science and Technology*, *63*(11), 2206-2222.

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*, P10008.

Braun, T. & Schubert, A. (2003). A quantitative view on the coming of age of interdisciplinarity in the sciences 1980-1999. *Scientometrics*, 58(1), 183-189.

Huutoniemi, K., Klein, J.T., Bruun, H., & Hukkinen, J. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy*, *39*, 79-88.

Leiden Ranking 2013. Retrieved June 2, 2015 from: http://www.leidenranking.com/Content/CWTS%20Leiden%20Ranking%202013.pdf

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of American Society for Information Science and Technology*, *58*(9), 1303-1319.

Leydesdorff, L. & Rafols, I. (2011). Indicators of the interdisciplinarity of journal: diversity, centrality, and citations. *Journal of Informetrics*, *5*(1), 87-100.

Lariviere, V., & Gingras, Y. (2014). Measuring interdisciplinarity. In Cronin B., & Sugimoto C.R. (Eds.) *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, Cambridge: MIT Press.

Morillo, F., Bordons, M., & Gomez I. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics*, *51*(1), 203-222.

Morillo, F., Bordons, M., & Gomez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. *Journal of American Society for Information Science and Technology*, *54*(13), 1237-1249.

Nichols, L.G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, *100*(3), 741-754.

Porter, A.L., & Chubin, D.E. (1985). An indicator of cross-disciplinary research. Scientometrics, 8(3-4), 166-176.

Porter, A., Roessner, J.D. & Heberger, A.E. (2008). How interdisciplinary is a given body of research? *Research Evaluation*, *17*(4), 273-282.

Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719-745.

Qin, J., Lancaster, F.W., & Allen, B. (1997). Types and levels of collaboration in interdisciplinary research in the science. *Journal of American Society for Information Science and Technology*, *48*(10), 893-916.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicator of interdisciplinarity: case studies in bionanosciene. *Scientometrics*, *82*(2), 263-287.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal ranking can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, *41*, 1262-1282.

Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval.* Auckland, New Zealand: McGraw-Hill.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanonscience and nanotechnology. *Scientometrics*, *59*(3), 425-465.

Sugimoto, C.R., Ni, C., Russell, T.G., & Bychowski, B. (2011). Academic genealogy as an indicator of interdisciplinarity: An examination of dissertation networks in library and information science. *Journal of the American Society for Information Science and Technology*, *62*(9), 1808-1828.

Sugimoto, C.R. (2011). Looking across communicative genres: a call for inclusive indicators of interdisciplinary. *Scientometrics*, *86*(2), 449-461.

Tijssen, R.J.W. (1992). A quantitative assessment of interdisciplinary structures in science and technology: co-classification analysis of energy research. *Research Policy*, 21, 27-44.

Wagner, C.S., Roessner, J.D., Bobb, K., Klein, J.T., Boyack, K.W., Keyton, J., Rafols, I., & Borner, K. (2011) Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of literature. *Journal of Informetrics*, *165*, 14-26.

van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *847*, 523-538.

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.

Waltman, L., & van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, *86*, 471.

Wang, L., Notten, A., & Surpatean, A. (2013). Interdisciplinarity of nano research fields: a keyword mining approach. *Scientometrics*, *94*, 877-892.

How often are Patients Interviewed in Health Research? An Informetric Approach

Jonathan M. Levitt[1] and Mike Thelwall[2]

[1]J.M.Levitt@wlv.ac.uk, [2]m.thelwall@wlv.ac.uk
Statistical Cybermetrics Research Group, University of Wolverhampton, Wolverhampton WV1 1LY (UK)

Abstract
In recent years research funding bodies have increased their emphasis on the engagement between researchers and the public. As part of this increased emphasis, the UK's National Institute for Health Research aims to promote a research-active population. A way in which patients can be research-active is by participating in research interviews. In order to assess the past levels of this type of contribution of patients to research, this paper investigates the extent to which health research refers to patient interviews. Co-word indicators for the interviewing and qualitative interviewing of patients are used to gauge how the levels of interviewing and qualitative interviewing in Web of Science (WoS) articles have varied over time, between science and social science and between WoS categories. The results indicate that the level of interviewing of patients, referred to in WoS articles, rose steadily between 1991 and 2013. Moreover, the amount of interviewing and qualitative interviewing varied substantially between health-related fields, with a marked tendency for more interviews in social science research and fewer in science research.

Conference Topic
Indicators

Introduction
Over the past few years research funding bodies have increased their emphasis on public involvement in health research. For example, the UK's National Institute for Health Research, in a recent strategic plan, listed as a key objective, "Citizens helping to identify and deliver research of the highest quality" (NIHR, 2014), adding that citizen participation health research "is contributing to a 'research active' nation focused on best health for all." In particular, those who are ill seem to be particularly important because they can provide first-hand understanding of the specific illness being researched. In order to understand the potential contribution of ill people to health research, it helps to understand their past contribution to health research. This paper addresses two aspects of past contribution: the extent to which this contribution has varied over time and the extent to which this contribution has varied between subjects. This paper also introduces and demonstrates a novel technique: the use of co-word metrics to gauge the levels of both interviewing and qualitative interviewing of patients, and applies it to Web of Science (WoS) articles.

Background
Informetric techniques Although the individual words in abstracts can be irrelevant to the content of the articles, analyses of the words in academic publications have been used extensively. Collections of articles have been mapped, based on the words in their titles (Leydesdorff & Zaal, 1988; Milojević et al., 2011), their titles and keywords (Whittaker, 1989), their titles and abstracts (Peters & van Raan, 1993), their titles with references used for context (van den Besselaar & Heimeriks, 2006), or their full text (Glenisson et al., 2005). However, other research with similar goals has ignored the text in articles and used subject headings instead (An & Wu, 2011). Automatic analyses of the text of articles have also been used to identify, or differentiate between, different types of methods used. For instance, this approach has been used to track the evolution, over time, of computing technologies within library and information science research and to identify articles that used specific statistical

techniques (Thelwall & Wilson, in press). One particularly relevant study searched for a set of methods-related keywords (e.g., cohort study) in the titles of health-related articles in the Web of Science, and then compared the citation impacts of the articles found for each method (Patsopoulos, Analatos, & Ioannidis, 2005).

Patient involvement in research

In addition to often being involved in decisions about their own care (Charles, Gafni, & Whelan, 1997), patients are routinely the subjects of medical research to investigate the causes of, or cures for, their maladies. Patients can also be more actively involved in research by giving their opinions in open-ended questionnaires, or in interviews, or focus groups and by participating in steering groups for the co-ordination of research. Patients may also be involved in developing or promoting informational material to fellow sufferers (Greenfield, Kaplan, & Ware, 1985) or even in developing research policies (Nilsen et al., 2006). Gaining the patient's perspective can be helpful for research, for example, to get insights into the extent to which symptoms, in practice, vary from the norm (Cotrell & Schulz, 1993) and to understand and prioritise the problems that sufferers believe to be the most important to address (Serrano-Aguilar et al., 2009). Seeking the views of patients is sufficiently widespread for systematic reviews of this practice to be published for specific ailments (Morton et al., 2010). Nevertheless, the apparently widespread knowledge of the importance of patient involvement does not ensure that it occurs for all conditions.

Research questions

This paper investigates a contribution that ill people have made to health research, namely the extent to which health research has interviewed patients. The research questions are:

1. To what extent has the level of the research interviewing (and in particular the qualitative interviewing) of patients varied over time?
2. To what extent has the level of the research of interviewing (and in particular the qualitative interviewing) of patients varied between subject categories?

Method

The main data used to address the research questions is the approximate number of articles that refer to patient interviews and approximate number of articles that refer to qualitative patient interviews. This data, obtained for different WoS databases and subject categories, must be normalised to allow comparisons between findings for different years and subjects.

A simple way of normalising is to calculate the rate of interviewing and qualitative interviewing in each subject category would be to divide by the number of articles in the dataset investigated. For some subject categories only a small proportion of articles are closely related to patients, however, and so this ratio would be flawed. For instance, less than one fifth of Pharmacology Pharmacy articles refer to ‘patient’ in the topic.

In order to normalise the interview metric, this paper divides instead by the number of articles that refer to patients. This interview metric indicates the extent to which articles that refer to also refer to interviews. This choice is based on the reasonable assumption that studies on patient interviews will in generally refer to patient in their abstracts. In order to normalise the qualitative interview metric, this paper divides by the number of articles that refer to patients and interviews. This qualitative interview metric indicates the extent to which articles that refer to patient interviews also refer to the interviews being qualitative. This metric was chosen in order to limit the metric to research that plausibly could qualitatively interview patients (i.e., where patients and interviews are mentioned).

In order to calculate the interview metric and qualitative interview metric the following data was extracted from WoS: (a) the number of articles that contain ‘patient*’ in the topic (patient frequency), (b) the number of articles that contain ‘patient*’ and ‘interview*’ in the topic

(patient interview frequency), and (c) the number of articles that contain 'patient*', 'interview*' and at least one of 'qualitative*', 'open-ended', 'in-depth', ''semi structured' and 'semistructured' in the topic (patient interview qualitative frequency). The interview metric was defined as 1000*patient interview frequency/patient frequency; the qualitative interview metric was defines as 100*patient interview qualitative frequency/patient interview frequency. The multipliers of 1000 and 100 were chosen in order for most of the findings to be expressed between 10 and 100. The definition of the qualitative interview metric was preferred to the alternative definition of 10000*patient interview qualitative frequency/patient frequency as it indicates how the proportion of interviews that are qualitative varied over time and between subjects.

A possible source of inaccuracy in the interview metric is that articles with patient and interview in the topic do not necessarily refer to patient interviews. The accuracy of the interview metric was gauged through content analysis of a random sample of 50 WoS articles containing 'patient*' and 'interview*' in the topic; 90% of the records referred to interviews of patients or people associated with their illness. A possible source of inaccuracy in the qualitative interview metric is that articles with patient, interview and an indicator of qualitative in the topic do not necessarily refer to qualitative patient interviews. The accuracy of the qualitative interview metric was gauged through a content analysis of a random sample of 50 WoS records containing 'interview*' and at least one of ''qualitative*', 'open-ended', 'in-depth', ''semi structured' and 'semistructured'; 96% of the records indicate that the interviews were qualitative. Other possible sources of inaccuracy in these metrics are false positives (e.g., 'patient' can be used in sense not related to health, i.e., not impatient) and omissions (e.g., the list of terms for qualitative research is unlikely to be exhaustive).

As a high proportion of the search terms are in the article abstracts, it is important to confine the study to periods in which a high proportion of WoS records contain abstracts. A total of 84% of the records, of a random sample of 50 WoS articles published in 1991, contain abstracts, whereas the figure for WoS articles published in 1990 is only 8% (for 2013 the figure is 100%). Consequently, this study does not investigate years prior to 1991.

Results

In this paper, ''Patient incidence' denotes the number of articles with 'patient*' in the topic, 'Interview incidence' denotes the number of articles with 'interview*' in the topic per 1,000 articles with 'patient*' in the topic, and 'Qualitative interview incidence' denotes the number of articles with the indicators of qualitative in the topic per 100 articles with 'interview*' in the topic, 'SCI only' denotes articles in the Science Citation Index (SCI) and not in the Social Sciences Citation Index (SSCI), 'SCSI only' denotes articles in the SSCI and not in the SCI, 'SCI & SSCI' denotes articles in both the SCI and SSCI, and 'A&HCI' denotes articles in the Arts & Humanities Citation Index.

Table 1: Patient, interview and qualitative interview incidences for five WoS datasets.

Datasets	*Articles containing patient* in the topic*	*Interview articles per 1000 patient articles*	*Qualitative interview articles per 100 interview articles*
WoS	2,570,556	23.7	26.0
SCI only	2,309,924	11.0	16.5
SSCI only	67,088	134.5	35.1
SCI & SSCI	192,749	137.1	32.1
A&HCI	2,810	74.4	35.9

As can be seen in Table 1, for both SSCI only and SCI & SSCI the incidences of interviews are over 12 times the incidence for SCI only and the incidence of qualitative interviews is 90% higher than the incidence for SCI only. These differences are likely to be partly due to the different sizes of the databases and partly due to differences in the proportion of articles that mention patients. The table also indicates that interviews are relatively prevalent in social science research relating to patients and rare in science research relating to patients. Because of the small number of A&HCI articles that contain 'patient*' in the topic, this paper does not further investigate this dataset.

In response to Question 1 (variation over time) the incidence of interviews for WoS rose by 175% between 1991 and 2013 (Figure 1, left). The incidence for SCI only undulated between 1998 and 2013, (10.2 in 1998, 11.1 in 2013), whereas, during the same period, the levels of SSCI only and SCI & SSCI rose steadily (the 2013 levels are respectively 48% and 36% higher than the 1998 levels). Thus, the use of interviews in patient-related research seems to have risen more rapidly in the social sciences than in science, despite the lower initial prevalence of interviews in science research. The use of qualitative methods in interviews appears to have risen substantially in all the areas investigated. However, the increase is more rapid in social sciences research than in science research (Figure 1, right).

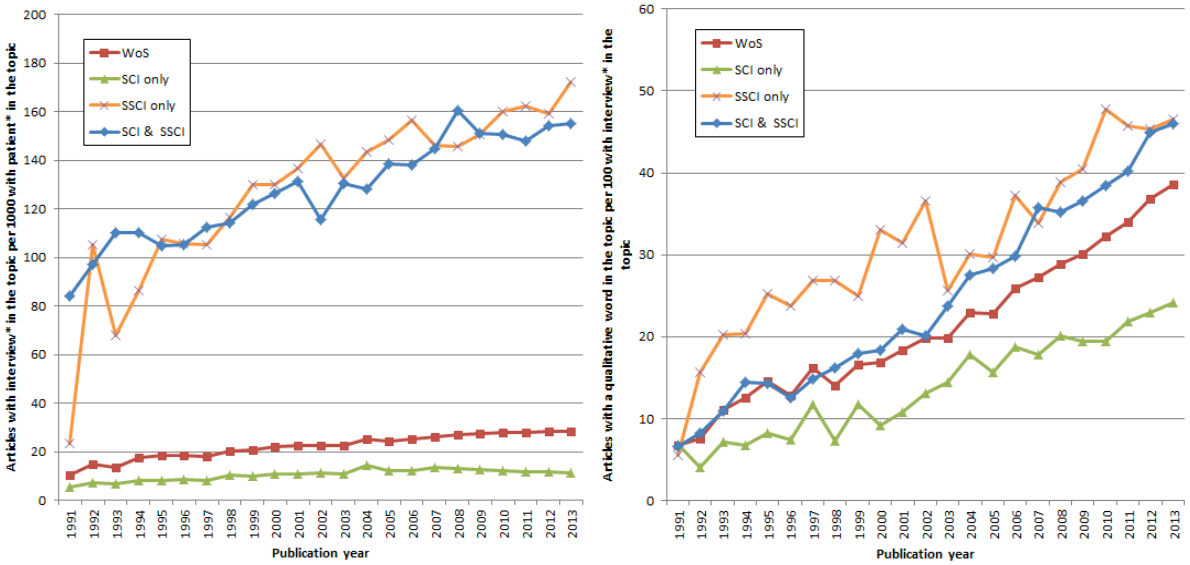


Figure 1. Annual incidence of interviews (left) and qualitative interviews (right).

In order to analyse disciplinary differences in more detail (Question 2), WoS categories were identified for each of the datasets SCI only, SSCI only and SCI & SSCI with at least 50 articles containing patient* and interview* in the topic. The ten categories identified were Clinical neurology, Health care sciences services, Health policy services, Nursing, Oncology, Pharmacology pharmacy, Psychiatry, Psychology, Public environmental occupational health and Rehabilitation. The incidence of interviews varies greatly between the ten categories, in addition to between science and social science research in the same category. The most extreme case is oncology, for which interviews are rare in science, but common in social science research (Table 2).

The incidence of qualitative interviews differs between science and social science in each individual category; qualitative interviews are more prevalent in social science research in 8 out of 10 categories (Table 2). For SCI only, the incidence of interviews is substantially lower for Clinical neurology, Oncology and Pharmacology pharmacy (average 12.0) than for the other seven categories (average 59.6). The incidence of qualitative interviews is also much lower for Clinical neurology, Oncology and Pharmacology pharmacy (average 14.0)

compared with the other seven categories (30.7). Hence, there are substantial disciplinary differences in the incidences of interviews and qualitative interviews within science.

Table 2: Incidence of interviews for ten WoS categories.

	Interviews			*Qualitative interviews*		
WoS category	*SCI*	*SSCI*	*Both*	*SCI*	*SSCI*	*Both*
Clinical neurology	16.5	65.1	107.3	11.9	25.0	17.6
Health care sciences services	92.2	99.9	157.5	41.4	30.3	46.5
Health policy services	76.0	182.7	125.4	31.6	47.6	39.0
Nursing	81.5	199.9	196.4	51.8	53.5	61.0
Oncology	7.2	226.3	195.2	15.0	47.7	45.7
Pharmacology pharmacy	12.3	199.2	67.6	15.2	58.0	17.8
Psychiatry	36.0	136.6	139.7	12.9	21.8	14.4
Psychology	46.0	102.2	115.5	25.9	17.7	19.4
Public environmental occupational health	53.3	219.8	170.6	20.0	44.3	37.0
Rehabilitation	32.5	86.7	137.9	31.0	34.5	52.4
Mean	**45.3**	**151.8**	**141.3**	**25.7**	**38.0**	**35.1**

For Clinical neurology, Oncology and Pharmacology, the percentage of articles in SCI only with patient* in the topic is particularly high: the percentage (in terms of articles in SCI or SSCI with patient* in the topic) for Clinical neurology is 89.3%, for Oncology is 96.4% and for Pharmacology pharmacy is 93.8%, whereas the average percentage for the other seven categories is 30.7%. There is a statistically significant Spearman correlation of -.81 between the interview incidence of SCI only and the percentage of articles with patient* in the topic that are in SCI only. This correlation reflects science categories having few interviews.

Limitations and conclusions

A limitation is that some studies with 'patient*' and 'interview*' in the topic do not interview patients (e.g., they interview physicians or carers of patients) and some studies with 'interview*' or indicators of qualitative in the topic do not conduct qualitative interviews (e.g., they combine quantitative interviews with qualitative analysis of patient records). But, as this research is comparative and the variations over time and between subjects are substantial, it seems likely that this limitation would not greatly affect the overall findings. Another limitation is that the results rely on the WoS journal subject classifications for journals. This may have a significant impact on the results for individual subject categories, as individual journals may have a substantial minority of the articles in a category. It would be useful to apply the techniques here to the full text of papers to help assess how often patient are involved in research but this is not discussed in the abstract of a paper.

After adjusting for the increase in the number of articles with 'patient*' in the topic, the number of WoS articles with 'interview*' in the topic increased by 175% from 1991 to 2013, suggesting that the use of patient interviews has increased substantially over the past 23 years. This may reflect a general trend towards involving patients more frequently in research, or an increase in the amount of research published, or indexed in WoS in research areas that typically involve patient interviews, such as nursing. In addition, after adjusting for the increase in the number of articles with 'patient*' and 'interview' in the topic, the number of articles that also had an indicator of qualitative in the topic increased by 511% from 1991 to 2013. This suggests that qualitative approaches are increasingly prevalent in health interviews, or that the qualitative nature of the research is more frequently specified. An

alternative explanation is that the amount of research published, or covered in WoS, has expanded in areas in which qualitative interviews are particularly common.

The incidences of interviews were particularly low amongst articles that were in SCI only; for 1991-2013 the incidence is less than one twelfth of the incidence for SSCI articles. When confining the study to categories present in both the SCI and the SSCI, there was a very marked difference between the datasets; however, the difference was substantially lower when excluding categories in which over 85% of the articles are in the SCI.

In the context of the NIHR aim of promoting a research-active population, the increased prevalence of patient interviews and qualitative interviews is encouraging, but categories with low percentages of interviews (e.g., Clinical neurology, Oncology and Pharmacology pharmacy) need to be further investigated to check whether individual subject areas are giving too little credence to patient interviews. Finally, this paper indicates that the technique of using simple co-word metrics based on the presence of words in the topic of WoS records can be applied usefully to informetric tasks. However, when investigating articles published prior to 1991, it is important to take into account that only a low percentage of WoS records for articles published in 1990 have abstracts.

References

An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics, 88*(1), 133-144.

Charles, C., Gafni, A., & Whelan, T. (1997). Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). *Social Science & Medicine, 44*(5), 681-692.

Cotrell, V., & Schulz, R. (1993). The perspective of the patient with Alzheimer's disease: a neglected dimension of dementia research. *The Gerontologist, 33*(2), 205-211.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *IP&M, 41*(6), 1548-1572.

Greenfield, S., Kaplan, S., & Ware, J. E. (1985). Expanding patient involvement in care: Effects on patient outcomes. *Annals of Internal Medicine, 102*(4), 520-528.

Leydesdorff, L., & Zaal, R. (1988). Co-words and citations relations between document sets and environments. In: Rousseau, R., & Egghe, L. *Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval* (pp. 105-119).

Milojević, S., Sugimoto, C. R., Yan, E., & 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.

Morton, R. L., Tong, A., Howard, K., Snelling, P., & Webster, A. C. (2010). The views of patients and carers in treatment decision making for chronic kidney disease: systematic review and thematic synthesis of qualitative studies. *BMJ, 340*, c112. 10.1136/bmj.c112

NIHR. (2014). Promoting a research active nation, http://www.nihr.ac.uk/.

Nilsen, E. S., Myrhaug, H. T., Johansen, M., Oliver, S., & Oxman, A. D. (2006). Methods of consumer involvement in developing healthcare policy and research, clinical practice guidelines and patient information material. *Cochrane Database Syst Rev, 3*.

Patsopoulos, N. A., Analatos, A. A., & Ioannidis, J. P. (2005). Relative citation impact of various study designs in the health sciences. *JAMA, 293*(19), 2362-2366.

Peters, H. P. F., & van Raan, A. F. (1993). Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy, 22*(1), 23-45.

Serrano-Aguilar, P., Trujillo-Martin, M. M., Ramos-Goñi, J. M., Mahtani-Chugani, V., Perestelo-Pérez, L., & Posada-de la Paz, M. (2009). Patient involvement in health research: a contribution to a systematic review on the effectiveness of treatments for degenerative ataxias. *Social science & medicine, 69*(6), 920-925.

Thelwall, M. & Wilson, P. (in press). Does research with statistics have more impact? The citation rank advantage of structural equation modelling. *JASIST*.

van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics, 68*(3), 377-393.

Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science, 19*(3), 473-496.

Normalized International Collaboration Score: A Novel Indicator for Measuring International Co-Authorship

Adam Finch[1], Kumara Henadeera[2] and Marcus Nicol[3]

[1] *adam.finch@csiro.au*
CSIRO, Waite Campus, Science Excellence Team, Waite Road, Urrbrae, SA 5064 (Australia)

[2] *kumara.henadeera@nhmrc.gov.au*
National Health & Medical Research Council, Strategic Policy Group, 16 Marcus Clarke Street, Canberra, ACT 2601 (Australia)

[3] *marcus.nicol@arc.gov.au*
Australian Research Council, Research Excellence Branch, 11 Lancaster Place, Canberra ACT 2609 (Australia)

Abstract
International collaboration on research publications is increasingly evaluated as part of a raft of performance measures. Levels of international co-authorship have increased substantially over the last few decades and vary substantially by research field and publication type; however, these variations are not typically accounted for by international collaboration indicators. In this research-in-progress paper, we introduce a novel metric, the Normalised International Collaboration Score, which adjusts the number of countries appearing on publication records using baselines relevant to the subject, age and type of the publication. A pilot analysis shows that these baselines vary substantially and that the application of this metric yields very different results to a more common measure of international collaboration. The limitations of the metric are discussed, along planned extensions for the full version of the study, as well as the relationship between normalised collaboration and citation.

Conference Topic
Indicators

Background and Purpose

Measuring international co-authorship

The availability of author address metadata on publication indices such as Web of Science and Scopus allows the analysis of patterns in co-authorship, including the collaboration by authors from different countries on research outputs. This approach has been used in many studies for decades (such as Glänzel & De Lange, 1997; Narin, Stevens, & Whitlow, 1991; Nederhof & Moed, 1993) and metrics describing international collaboration now appear regularly in bibliometric handbooks (Colledge, 2014; Rehn, Kronman, & Wadskog, 2007) and in reporting tools such as Thomson Reuters' InCites, Elsevier's SciVal and SCImago's Journal & Country Ranking. Such publications tend to receive higher levels of citation, an effect that is not due to the increased propensity for self citation arising from additional authors (Van Raan, 1998), but likely rather shared experience, knowledge and equipment.

Analysis of international co-authorship metadata has highlighted other important aspects of collaboration. Firstly, levels of international collaboration have increased substantially over the last quarter century (Leydesdorff & Wagner, 2008); and secondly, levels of international collaboration vary by field of research (Frame & Carpenter, 1979). A report on Thomson Reuters' InCites (retrieved 7 January 2015) indicates that 2013 articles, reviews and proceedings papers in Tropical Medicine involved international collaboration 46.7% of the time, while for History, this was only 4.3% of the time. Even within Medicine, Emergency Medicine saw only 9.9% foreign collaboration, far lower than Tropical Medicine. Variation is significant over time, with Astronomy & Astrophysics international collaboration rising from 19.4% in 1993 to 45.0% in 2013. To these two aspects, we must add publication type; 2013

Astronomy & Astrophysics articles saw 51.4% international collaboration, but its Proceedings Papers only 0.2%. Such variations exist across the full gamut of subject, years and publication types but most metrics used to evaluate collaboration do not take account of them.

Existing metrics

Frequently, analyses use either the number or proportion of collaborative publications (see for example Boekholt et al., 2009; Colledge, 2014; Luukkonen et al., 1993). Glänzel and De Lange (2002) use a Multilateral Collaboration Index to measure the number of collaborative links compared to the number of collaborative papers, establishing the intensity of collaboration.

Beaudet, Campbell, Côté, Haustein, Lefebvre and Roberge (2014) use a regression model based on power law relationships to establish the expected level of collaboration for a country and an Affinity Index to identify key partners. Degelsegger et al. (2013) propose thematic assessment, normalized either by relating it to the output of the country in the subject, or by comparing it to co-authored output in the same subject but with a different partner. Ding, Yang and Liu (2013) propose using network metrics to evaluate collaboration impact, which is a sound approach within a subject and time frame. Pohl, Warnan and Baas (2014) go the greatest distance to normalizing for the three aforementioned influences, by adjusting the proportion of publications with international collaboration by the number of collaborating countries in each subject. This study only considered a single year, however, did not adjust for publication type and was based on adjusting the share of research with a binary attribute (either internationally collaborative or non-internationally collaborative). The properties and results of this alternative will be compared to our metric in the full version of our study.

The Normalised International Collaboration Score (NICS)

The Normalised International Collaboration Score uses fundamentally the same calculation as the "new" Crown Indicator by which it was inspired (Waltmann, van Eck, van Leeuwen, Visser, & van Raan, 2011). For each publication, a global baseline is constructed, representing the average number of countries contributing to publications of the same type, from the same year and appearing in the same subject area(s). The number of countries contributing to the publication in question is then divided by the relevant baseline to yield a ratio. This ratio is then averaged for all publications in a set (for an institution, country, journal, etc). Our exploratory analysis uses both the mean (as in the Crown Indicator) and the statistically preferable median (Bornmann, Mutz, Neuhaus, & Daniel, 2008), for the purposes of comparison. While the present study only includes a selection of publication types, years and subjects, our full study will include all subjects and publication types back to 1996.

Methodology

The Advanced Search function on Web of Science was used to isolate publications of the Article, Review and Proceedings Paper types with issue cover dates in 1993, 2003 and 2013, and allocated to the subject categories Dance, Engineering (Manufacturing), Evolutionary Biology, Gastroenterology & Hepatology, Political Science, Psychology (Educational), Soil Science and Tropical Medicine. These publication types were selected as those most likely to contain address data; these years as spaced such to demonstrate evolution in collaboration trends and aspects of the data; and those subjects as representing a broad spectrum across science, social science, and the arts and humanities. The selection of a single discipline of period would not have illustrated any variation over time or theme. Record metadata were downloaded, tagged with the relevant subject name and recombined into a single dataset. Individual addresses were broken out, the non-country information in each field deleted and duplicate country entries deleted. A count of unique country contributions per publication was

made. The baselines were constructed by averaging all unique country contribution counts for each combination of year, publication type and subject, using the arithmetic mean and the median. These represented the denominator of the metric's article-level ratio.

The institutional data came from a database of Australian publication records from 2001 to 2014. A query extracted the unique identifier, selected subject areas, year, type and Crown Indicator of each publication, along with the author addresses. The addresses were subjected to a unique contributing country count, yielding the numerator of the metric's article-level ratio. The subject, publication type and publication year data were used to look up the mean and median baseline data (our ratio's denominator). Dividing the latter by the former yielded the article-level NICS, which was then averaged for each Australian institution – using the arithmetic mean and then the median, as appropriate for the baseline.

This gives the following notation for the mean form of NICS:

$$\frac{1}{p}\sum_{i=1}^{p}\frac{n_i}{g_i}$$

And the following notation for the median form of NICS:

$$\widetilde{\left(\frac{n_i}{m_i}\right)}$$

Where p denotes the number of publication produced by a unit of analysis, n_i denotes the number of countries contributing to the unit's publication i, g_i denotes the global mean number of countries contributing to publications of the same type, year and subject(s) as publication i and m_i denotes the global median number of countries contributing to publications of the same type, year and subject(s) as publication i. A third, "hybrid", version of was calculated, finding the median of article level ratios based on a mean:

$$\widetilde{\left(\frac{n_i}{g_i}\right)}$$

Results & Discussion

Table 1 shows the mean baselines for each year in each subject, combining publication types into a single entry. Several points are clear. Some subjects see a substantial increase in average country contributions over time – such as the increase from 1.15 to 1.71 for Evolutionary Biology – indicating a need to normalise for this change if fair comparisons are to be made among publication sets from different year ranges. There are also significant disparities between subjects, with the Engineering subject baseline 1.09 in 2013, compared to 1.78 for Tropical Medicine. It is also notable that, unlike citation counts, there does not seem to be a pattern of lower country contributions for social sciences as opposed to sciences, at least in this very limited dataset; Political Science has one of the higher baseline sets and Engineering, Manufacturing one of the lower. Lastly, some subjects, most likely those in the Arts & Humanities, may be difficult to assess using this metric, due to a paucity of address and a low publication count; the baselines would be based on too low a sample size and very prone to skew from outliers. It is also worth noting that, while country contributions are strongly positively skewed, the variance of the natural log of country contribution counts is lower than that of citation counts for publications of the same year, type and subject, in a all of a selection of the below instances that were considered.

Table 2 shows the number of publications missing address data in each of the three years for each subject. Coverage is a problem in Dance for all years and is more of a problem in the social science subjects than the sciences, but is an issue for all subjects in 1993. In the full analysis, work will be conducted to establish the point at which coverage is sufficient for robust analysis, but the institutional analysis in this pilot study exclude the 1993 publications.

Table 1. Mean Subject Country Contribution Baselines by Year.

	1993		*2003*		*2013*	
Table	*# Pubs*	*# Countries*	*# Pubs*	*# Countries*	*# Pubs*	*# Countries*
Dance	2	1.50	25	1.00	46	1.13
Engineering, Manuf.	1242	1.03	7935	1.06	14513	1.09
Evolutionary Biology	987	1.15	3900	1.38	5543	1.71
Gastroent. & Hepat.	3567	1.09	8595	1.15	11300	1.29
Political Science	987	1.15	3172	1.26	5549	1.76
Psychology, Educ.	483	1.05	1167	1.10	2253	1.20
Soil Science	1724	1.09	3890	1.23	4721	1.36
Tropical Medicine	798	1.45	1381	1.68	3128	1.78

Table 2. Instances of Publication Entries Missing Address Data by Year.

	1993			*2003*			*2013*		
Table	*No Address*	*Total Pubs*	*%*	*No Address*	*Total Pubs*	*%*	*No Address*	*Total Pubs*	*%*
Dance	245	247	99.2%	386	411	93.9%	184	230	80.0%
Eng., Manufact.	936	2178	43.0%	587	8522	6.9%	219	14732	1.5%
Evolutionary Biology	698	1685	41.4%	15	3915	0.4%	10	5553	0.2%
Gastro. & Hepat.	2080	5647	36.8%	158	8753	1.8%	68	11368	0.6%
Political Science	3421	4408	77.6%	965	4137	23.3%	636	6185	10.3%
Psych., Education.	573	1056	54.3%	20	1187	1.7%	47	2300	2.0%
Soil Science	1702	3426	49.7%	132	4022	3.3%	16	4737	0.3%
Tropical Medicine	475	1273	37.3%	11	1392	0.8%	24	3152	0.8%

Table 3 shows the mean baselines for each publication type in each subject, combining years into a single entry. It is clear that publication type is also a major factor for the baselines, with the Proceedings Papers consistently seeing fewer country contributions than other types. However, there is further variation; Political Science, for example, sees higher country counts for Articles than Reviews, while the reverse is true for Soil Science.

Table 3. Mean Subject Country Contribution Baselines by Publication Type.

	Articles		*Proceedings*		*Reviews*	
Table	*# Pubs*	*# Countries*	*# Pubs*	*# Countries*	*# Pubs*	*# Countries*
Dance	73	1.10	-	-	-	-
Engineering, Manuf.	9192	1.20	14398	1.00	100	1.23
Evolutionary Biology	9665	1.54	101	1.00	664	1.59
Gastroent. & Hepat.	20482	1.21	811	1.00	2169	1.24
Political Science	8650	1.57	790	1.18	268	1.28
Psychology, Educ.	3542	1.16	263	1.00	98	1.09
Soil Science	8547	1.31	1641	1.01	147	1.69
Tropical Medicine	5113	1.71	23	1.00	171	1.73

Table 4 shows a comparison of institutional collaboration analysis using the proportion of publications with international collaboration and each of the three variants of the NICS metric. The Median calculation appears the least useful; every baseline in each year, subject and document was 1, so this version essentially reports the median country contribution per article and cannot strongly differentiate among institutions. The version using mean baselines are more useful for ranking but, like the Crown Indicator, remains sensitive to outliers (as in the example of Flinders University, where performance was inflated by a single article with 35 contributing countries). Even though the full study will involve far larger sample sizes, which should be less susceptible to such outliers, it appears that the "hybrid" (median of ratios based on mean baselines) is the strongest option. This would preclude statistical analysis based on parametric data, but it is impossible to tell from the pilot study whether the article level results of the mean calculation would be normally distributed on a global scale either.

Table 4. Selected Australian Institution Ranking.

		% Collaboration		*NICS Mean*		*NICS Median*		*NICS 'Hybrid'*	
Table	*Pubs*	*Value*	*Rank*	*Value*	*Rank*	*Score*	*Rank*	*Score*	*Rank*
Queensland Inst Med Res	55	70.9%	1	1.61	3	2	1	1.19	3
James Cook Univ	98	65.3%	2	1.72	1	2	1	1.44	2
Charles Darwin Univ	40	62.5%	3	1.60	4	2	1	1.12	5
Univ Western Sydney	52	57.7%	4	1.55	6	2	1	1.45	1
Univ Western Australia	219	50.7%	6	1.26	15	2	1	1.12	5
Univ Melbourne	233	50.2%	7	1.48	7	2	1	1.12	5
Univ Adelaide	174	49.4%	9	1.24	22	1	9	0.86	19
Univ Sydney	286	48.6%	10	1.38	10	1	9	0.99	13
CSIRO	210	48.6%	11	1.24	21	1	9	0.99	12
Univ Queensland	271	48.3%	12	1.25	18	1	9	0.91	15
Queensland Univ Technol	58	48.3%	13	1.25	16	1	9	1.00	9
Murdoch Univ	45	46.7%	15	1.23	23	1	9	0.79	22
Univ Newcastle	48	45.8%	16	1.29	14	1	9	1.00	10
Australian Natl Univ	203	44.8%	17	1.08	27	1	9	0.79	22
Univ New S Wales	195	44.1%	18	1.24	20	1	9	0.88	16
Monash Univ	155	43.2%	19	1.35	11	1	9	0.88	16
Curtin Univ Technol	44	43.2%	20	1.20	24	1	9	1.00	11
Howard Florey Inst	48	35.4%	26	1.56	5	1	9	0.78	25
Flinders Univ S Australia	70	30.0%	27	1.65	2	1	9	0.78	25

Discussion

While only a few institutions see a large difference in ranking when applying NICS rather than proportion of international publications, the difference in results and the variations in baselines on which they are based suggest the metric has informational content. It is also worth noting that, at an article level, the Crown Indicator correlates positively and fairly strongly with NICS (Spearman's Rank r=0.384) and that at an institutional level, the two versions of NICS derived from mean baselines correlate more closely with NCI performance (r=0.289 and 0.148) than does share of publications with international collaboration (r=0.09).

There are clearly limitations to this approach. It does not account for collaboration intensity; eight co-authoring institutions in a specific foreign country count the same as one. The NICS

baselines could be rescaled to count not only contributing foreign countries but also the numbers of institutions in those countries, and even potentially types of institutions. As it would require a set of baselines for each country, this would be computationally intensive but will be explored in the full study. This approach would also normalise for the propensity of a country to collaborate, which many of the above-mentioned metrics are aimed at doing. Lower collaboration levels can arise from several causes, including a lower advantage yielded and having a large share of global output (therefore limiting the avenues available for external collaboration); normalising for national collaboration levels may obscure these differences and render accurate national comparisons challenging. In its present form, NICS serves best as a metric to compare the collaboration of countries and institutions, variations in which may then be considered in the context of national motivation and propensity to collaborate.

Other criticisms leveled at the Crown Indicator apply to NICS, most notably a limited representation of global output in some subjects and of some publication types, and the reliance on a subject taxonomy designed for information retrieval rather than bibliometric analysis. In the pilot study, moreover, many articles analysed here appeared in more than one subject area, and yet were normalised only with the baselines for one of those subject areas.

The full study will apply a wide range of statistical tests to the properties of the baselines, the country contribution counts and the resultant ratios; for now, however, and even with the aforementioned caveats, this metric shows potential for robust and meaningful analysis of institutional and national research collaboration abroad.

References

Beaudet, A., Campbell, D., Côté, G., Haustein, S., Lefebvre, C., & Roberge, G. (2014) *Bibliometric Study in Support of Norway's Strategy for International Research Collaboration.* Report to the Science Council of Norway, Science-Metrix. March 2014.

Boekholt, P., Edler, J., Cunningham, P., & Flanagan, K. (2009). *Drivers of international collaboration in research*. Report to EC, DG Research. Technopolis BV,Netherlands, April 2009.

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.

Colledge, L. (2014). Snowball Metrics Recipe Book, Edition 2. Retrieved December 19, 2014 from: http://www.snowballmetrics.com/wp-content/uploads/snowball-recipe-book_HR.pdf

Degelsegger, A., Lampert, D., Büsel, K., Simon, J., Tschant, J., & Wagner, I. (2013) *Assessing international cooperation in S&T through bibliometric methods. Proc. ISSI*, 175-184.

Ding, J., Yang, L. & Liu, Q. (2013). Measuring the academic impact of researchers by combined citation and collaboration impact. *Proc. ISSI*, 1177-1187.

Frame, J.D., & Carpenter, M.P. (1979). International research collaboration. *Social Studies of Science,* 9, 481-497.

Glänzel, W., & De Lange, C. (1997). Modelling and measuring multilateral co-authorship in international scientific collaboration. Part II. A Comparative study on the extent and change of international scientific collaboration links. *Scientometrics*, *40*, 605-626.

Glänzel, W., & De Lange, C. (2002). A distributional approach to multinationality measures of international scientific collaboration. *Scientometrics*, *54*, 75-89.

Leydesdorff, L. & Wagner, C.S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics, 2*, 317-325.

Luukkonen, T., Tijssen, R.J.W., Persson, O., & Sivertson, G. (1993). The measurement of international scientific collaboration. *Scientometrics, 28*, 15-36.

Narin, F., Stevens, K., & Whitlow, E.S. (1991). Scientific cooperation in europe and the citation of multinationally authored papers. *Scientometrics, 21*, 313-323.

Nederhof, A.J., & Moed, H.F. (1993). Modeling multinational publication - development of an online fractionation approach to measure national scientific output. *Scientometrics*, *27*, 39-52.

Pohl, H., Warnan, G., & Baas, J. (2014). Level the playing field in scientific international collaboration with the use of a new indicator: Field-Weighted Internationalization Score. *Research Trends, 9*, 3-8.

Rehn, C., Kronman, U., & Wadskog, D. (2007). *Bibliometric indicators – definitions and usage at Karolinska Institutet*. Retrieved December 19, 2014 from: http://kib.ki.se/sites/kib.ki.se/files/Bibliometric_indicators_definitions_1.0.pdf

van Raan, A.F.J. (1998). The influence of international collaboration on the impact of research results. *Scientometrics*, *43*, 423-428.

Waltmann, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & van Raan, A.F.J. (2011). Towards a new crown indicator: an empirical analysis. *Scientometrics, 87*, 467-481.

Bibliometric Indicators of Interdisciplinarity Exploring New Class of Diversity Measures

Alexis-Michel Mugabushaka[1], Anthi Kyriakou & Theo Papazoglou

[1] *Alexis-Michel.MUGABUSHAKA@ec.europa.eu*
European Research Council Executive Agency[1], Brussels, (Belgium)

Abstract
In bibliometrics, interdsicsiplinatity is often measured in terms of the "diversity" of research areas in the references that an article cites. The standard indicators used are borrowed mostly from other research areas, notably from ecology (biodiversity measures) and economics (concentration measures). This paper discusses a new class of measures, which are used in the study of biodiversity and especially the Leinster-Cobbold diversity measure (Leinster Cobbold 2010). We present a case study based on previously published dataset of 12 journal articles from a group of five researchers from the bio-nano science described and published by Rafols and Meyer (2010). We replicate the findings of this study to show that the various interdisciplinarity measures are in fact special cases of the Cobbold-Leinster diversity measure. The paper discusses some interesting properties of the Cobbold-Leinster diversity measure, which makes it appealing in the study of disciplinary diversity than the standards diversity indicators used as proxy for interdisciplinarity.

Conference Topic
Indicators

Introduction
Considerable efforts have been made to operationalize and measure the concept of interdisciplinarity in bibliometrics (Porter et al., 2007; Rafols & Meyer, 2010). The most commonly used indicators of interdisciplinarity are mostly borrowed from other research areas, notably from ecology (biodiversity measures) and economics (concentration measures).
The purpose of this paper is to bring to discussion a relatively new class of diversity indicators which are used in ecology but so far not been used to investigate disciplinary diversity. Drawing from the literature in ecology, the paper highlights important properties of those measures and discusses how they can help the bibliometric study of interdisciplinarity.
The paper is divided in three parts. The next section briefly presents indicators of interdisciplinarity in bibliometrics. The second section discusses the development of new class of diversity measures used in ecology and presents the Leinster-Cobbold diversity measure, highlighting its properties and why they are relevant for bibliometric usage. The third section presents a case study to illustrate the potential of Leinster-Cobbold diversity indicators as a measure of disciplinary diversity.

Currently used Bibliometric indicators of interdisciplinarity
Bibliometric analyses of interdisciplinarity take as unit of analysis a scientific paper and assume that the extent to which it integrates elements of different disciplines is reflected in the references it cites. References in scientific papers are expected to reflect various aspects of interdisciplinary because researchers will credit what they are indebted to other disciplines: conceptually (concepts, ideas and approaches from other disciplines); analytically (methods for defining, collecting and analyze data) and technically (tools developed in other fields).

[1] The views expressed in this paper are the authors'. They do not necessarily reflect the views or official positions of the European Commission, the European Research Council Executive Agency or the ERC Scientific Council.

Porter et al. (2007) developed the integration score as measure of interdisciplinary which takes into account not only the distribution of the cited references in different subject categories but also how closely related those subject categories are (see also Porter et al., 2006; Porter et al., 2008). In line with Porter's conceptualization, Rafols and Meyer (2006, 2010) introduced a new set of bibliometric indicators to quantify the disciplinary diversity of references as a proxy measure of interdisciplinarity. They are mostly based on the general framework for analyzing diversity developed by Stirling (2007). The most commonly used indicators are summarized in table 1. We note that there are also efforts to use network based measures (Rafols & Meyer, 2010; Karlovčec & Mladenić, 2015) but here we focus on diversity measures.

Table 1. Most common indicators of interdisciplinarity in bibliometric studies .

Indicators	Definition/description	
Variety	The number of different disciplines that a given paper cites**	N
Shannon entropy	As measure of diversity the Shannon Entropy quantifies how diverse the subject categories in the references are.	$H_{SH} = -\sum_{i=1}^{S} p_i \log p_i$ Where p_i is the proportion of elements in a system and S the number of elements in the system.
Simpson diversity	It measures how references are distributed (or concentrated) in subject categories.	$H_{GS} = 1 - \sum_{i=1}^{S} p_i^2$ Where p_i is the proportion of elements in a system and S the number of elements in the system
Rao-Stirling index	Can be understood as the Simpson diversity which takes into account distance/similarity (between disciplines).	$= \sum_{i,j} d_{i,j}\, p_i p_j$ Where di,j is the distance between the ith and jth element in the distance matrix and pi is the proportion of element i

Source: Rafols & Meyer 2010, p. 267 **Its variants includes normalization by the total numbers of subject categories or the shares of references outside a given subject category

New classes of diversity measures in ecology

Effective numbers

The diversity measures listed in table are also among the commonly used indicators of biodiversity in ecology. However, they have recently faced strong criticisms (Jost, 2006; Chao & Lou, 2012).

The main criticism is that those measures fail to satisfy the most basic property that ecologist would expect from a meaningful measure of diversity, namely the replication principle. In simple term, the "replication principle" states that if you have two completely distinct communities (i.e. without any overlap in the species) with each community having a diversity measure X, one would expect that combining those two communities would result in a community with a diversity measure 2X.

One category of diversity measures, which satisfy this replication principle is the so called "Hill-numbers" (also called "effective numbers of species"). They can be interpreted as the

"number of equally abundant specifies that are needed to give the same value of the diversity measure (Chao & Lou, 2012, p. 204).

The Hill numbers have some properties that other measures of diversity based on entropy lack:

- They satisfy the replication principles. i.e. two communities with each 4 effective numbers of species will – if pooled together – result in a community whose effective number equal 8. They therefore give logically consistent answers.
- Their linear scale makes it easier to interpret the magnitude of their change.
- In addition to this this advantage of intuitive consistency, they have another interesting property that we call "unifying framework status". Jost (2006) has shown that practically all traditional measures of diversity can be easily converted to "Hill numbers/ "effective numbers" and vice-versa.

Leinster-Cobbold Diversity Measure

Leinster and Cobbold (2012) developed a measure, which extends the Hill numbers to include the similarities/differences between species. Their measure – called here the Leinster-Cobbold Diversity Measure - can be used with any similarity coefficient between each pair of the species. This extends the scope of its usage to other contexts such disciplinary diversity in bibliometrics. In the following, we first provide its formal definition and discuss its properties as well as its relation to other diversity measures. In the next section we provide a case study of its use in the study of disciplinary diversity.

Consider a system with S elements with relative frequencies translating in estimated probabilities p = (p1, …, pS) so that $\sum_{i=1}^{S} p_i = 1$

The similarity between the elements is encoded in an S x S Matrix Z.

Z = (Z_(i,j)), with Z_(i,j) measuring the similarity between the ith and jth elements.

Whereby 0 ≤Z_(i,j) ≤1, with 0 indicating total dissimilarity and 1 indicating identical elements.

The Leinster-Cobbold diversity measure is defined as

$$ {}^{q}D^{Z}(\boldsymbol{p}) = \begin{cases} \left(\sum\limits_{i:p_i>0} p_i (Z\boldsymbol{p})_i^{q-1} \right)^{\frac{1}{1-q}} & q \neq 1, \\ \prod\limits_{i:p_i>0} (Z\boldsymbol{p})_i^{-p_i} & q = 1, \\ \min\limits_{i:p_i>0} \dfrac{1}{(Z\boldsymbol{p})_i} & q = \infty. \end{cases} $$

where

$$ (Z_p)i = \sum_{j=1}^{S} Z_{i,j} p_i $$

q is in number in range 0 ≤ q ≤ Infinity. It is called a sensitivity parameter and control the relative emphasize that the user wishes to place on common and rare species.

Case Study: Using the Leinster-Cobbold Diversity as a measure of disciplinary diversity

In our view, there are three main advantages in adopting the Leinster and Cobbold diversity measure in the study of disciplinary diversity as well:

- First, Leinster and Cobbold (2012) have discussed the relation between this measure and other diversity measures and showed that they can be seen as its special cases. The advantage here would be to have a single formula which would replace the Shannon entropy, the Simpson Diversity and the Rao-Stirling Index used in bibliometrics.
- Second, because the Leinster and Cobbold measure quantifies diversity on a spectrum which depends on how much emphasis should be given to relatively rare elements (sensitivity parameter q), it provides potentially more information than measures which consider only one value of this sensitivity parameter.
- The third advantage is the intuitive consistency of the Leinster and Cobbold measure. Because it directly produces "effective numbers" which obey the replication principle, the values can be easily interpreted and compared. Consider two publications: one with references from 2 (unrelated) categories and the other with reference from 4 (unrelated) categories. With the Leinster and Cobbold measure, they can be compared to say that the second has a twice as large diversity in references as the first one.

In the following, we present a case study to illustrate the potential of Leinster-Cobbold diversity profiles in quantifying disciplinary diversity.

Disciplinary diversity of selected papers in bio-nanoscience (Rafols & Meyer 2010)

The case study is based on a dataset of 12 journal articles from a group of five researchers from the bio-nano science described and published by Rafols and Meyer (2010). For those 12 papers, Rafols and Meyers published the distribution of their references in Web of Science Categories (Rafols & Meyers, 2010; p. 276, Table 3) as well as the scores on various indicators of diversity (ibid. p. 277, Table 4). The similarity/distance measures between the Web of Science subject categories are taken from the supplementary materials to the paper[2] by Chavarro et al. (2014).

Table 2. Diversity measures for the 12 papers in Rafols and Meyer (2010).

	not considering distance/similarity						considering distance/similarity					
sensitivity parameter q	**0**	**1**	**2**	**3**	**4**	**Inf**	**0**	**1**	**2**	**3**	**4**	**Inf**
Column no.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Papers												
Fun95	16	6,452	4,553	3,989	3,740	3,106	1,656	1,422	1,329	1,288	1,266	1,188
Koj97	17	5,526	4,232	3,848	3,652	2,880	1,479	1,284	1,225	1,203	1,192	1,143
Ish98	15	5,003	3,499	2,990	2,741	2,156	1,342	1,229	1,192	1,176	1,167	1,108
Noj97	16	4,532	3,120	2,665	2,447	1,967	1,280	1,172	1,141	1,128	1,122	1,077
Yas98	16	4,466	3,003	2,537	2,327	1,890	1,231	1,158	1,133	1,122	1,115	1,072
Oka99	16	4,857	3,814	3,557	3,439	3,062	1,253	1,190	1,165	1,154	1,148	1,108
Kik01	14	4,944	3,857	3,534	3,364	2,673	1,251	1,195	1,169	1,155	1,148	1,102
Sak99	14	5,103	4,040	3,764	3,641	3,184	1,245	1,181	1,159	1,149	1,143	1,098
Bur03	14	4,697	3,536	3,230	3,086	2,571	1,178	1,142	1,127	1,120	1,115	1,082
Tom00	15	4,841	3,846	3,625	3,530	3,028	1,227	1,165	1,145	1,136	1,132	1,095
Tom02	14	4,849	3,864	3,630	3,531	3,192	1,242	1,180	1,159	1,149	1,143	1,103

This case study illustrates that the various diversity measures are in fact special cases of the Leinster-Cobbold diversity profiles. We do this by replicating the diversity measures computed by Rafols and Meyer 2010 using the Leinster-Cobbold diversity profiles. We first compute the values of the Leinster Cobbold measure using different values for the sensitivity parameters (0, 1, 2, 3, 4 and infinity) and in two variants: without taking into account the

[2] http://www.interdisciplinaryscience.net/topics/interdisciplinarity-and-local-knowledge

distance/similarity between the subject categories (i.e. the matrix Z is an identity matrix) and by taking into account the distance/similarity between the subject categories (using the similarity data provided in supplementary materials of Chavarro et al. (2014). Using the conversion formulas in the first row of Table 3, we use those Leinster Cobbold values to derive the diversity measures provided in Rafols and Meyer 2010 (table 4 on page 277). The Table 3 below replicates the diversity values reported in Rafols and Meyer 2010. There are some differences, which are due to rounding but also to the fact that some indicators in Rafols and Meyer (2010) were given in normalized form.

Table 3. Deriving diversity measures commonly used in bibliometrics from the Leinster-Cobbold values.

	Variety	**Gini-Simpson**	**Shannon**	**Rao**
computation	**Col 1**	**1- (1/Col 3)**	**ln(Col 2)**	**1- (1/Col 9)**
Papers				
Fun95	16	0,78	1,86	0,25
Koj97	17	0,76	1,71	0,18
Ish98	15	0,71	1,61	0,16
Noj97	16	0,68	1,51	0,12
Yas98	16	0,67	1,5	0,12
Oka99	16	0,74	1,58	0,14
Kik01	14	0,74	1,6	0,14
Sak99	14	0,75	1,63	0,14
Bur03	14	0,72	1,55	0,11
Tom00	15	0,74	1,58	0,13
Tom02	14	0,74	1,58	0,14
Yil04	16	0,76	1,68	0,16

Concluding remarks

In bibliometrics, the interdisciplinarity is operationalized in terms of the diversity of the references in a scholarly article. The most commonly used indicators are derived from the fields of ecology (biodiversity measures) and from the fields of economics (concentration measures). We discuss a new class of biodiversity measures – the "effective numbers" - which not only generalize most of other diversity measures but also have some proprieties which make their interpretation intuitively consistent with the concept of diversity Jost (2006). They were further developed by Leinster-Cobbold (2012) to take into account the similarity/distance of elements (species) in a system (community). We provide an example on how the bibliometric indicators of interdisciplinarity are in fact special cases of this more general Leinster Cobbold indicator.

Future work should not only take a closer look at their statistical properties (distribution, parameters etc.) but also test their reliability and validity. In particular, it would be of interest to analyze how sensitive the indicators are to various degree of granularity of different classifications of research disciplines and to assess extent to which they depend on measures of distances used.

Acknowledgments

We thank Ismael Rafols for helpful comments on an earlier draft of the paper and Diego Chavarro for making the similarity matrix freely available.

References

Chao, A., & Jost, L. (2012). *Diversity measures. In Encyclopedia of Theoretical Ecology* (Eds. A. Hastings and L. Gross), pp. 203-207, Berkeley: University of California Press.

Chavarro, D., Tang, P., & Rafols, I. (2014). Interdisciplinarity and research on local issues: evidence from a developing country. *Research Evaluation*, *23*(3), 195-209.

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*, 363-375.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, *88*, 2427–2439.

Jost, L. (2009). Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008). *Ecological Economics*, *68*,925–928.

Karlovčec, M., & Mladenić, D (2015) Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*, *102*(1), 433-454.

Leinster T., & Cobbold CA. (2012). Measuring diversity: the importance of species similarity. *Ecology*, *93*(3):477-489.

Porter et al., (2006). Interdisciplinary research: meaning, metrics and nurture. *Research Evaluation*, *15*(3), 187-195.

Porter, A.L. & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719-745.

Porter, A.L., Cohen, A.S., Roessner, J.D., & Perreault, M. (2007), Measuring researcher interdisciplinarity, *Scientometrics*, *72*(1), 117-147.

Rafols, I. & Meyer, M. (2006). Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience. *SPRU working paper*.

Rafols, I. & Meyer, M. (2010). Diversity and Network Coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, *82*(2), 263-287.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707-719.

Modeling Time-dependent and -independent Indicators to Facilitate Identification of Breakthrough Research Papers

Holly N. Wolcott[1], Matthew J. Fouch[1], Elizabeth Hsu[2], Catherine Bernaciak[1], James Corrigan[2], and Duane Williams[1]

holly.wolcott@thomsonreuters.com
[1]Intellectual Property & Science, Thomson Reuters, Rockville, MD 20850 (USA)

corrigan@mail.nih.gov
[2]Office of Science Planning and Assessment, National Cancer Institute, Bethesda, MD 20892 (USA)

Abstract
Research funding organizations invest substantial resources to stay current with important research findings within their mission areas to identify and support promising new lines of inquiry. To that end, we continue to pursue the development of tools to identify research publications that have a strong likelihood of driving new avenues of research. This research-in- progress paper describes our work incorporating multiple time-dependent and -independent features of publications into a model that aims to identify candidate breakthrough papers as early as possible following publication. We used multiple Random Forest models to assess the ability of indicators to reliably distinguish a gold standard set of breakthrough publications as identified by subject matter experts from among a comparison group of similar *Thomson Reuters Web of Science*™ publications. These indicators will be selected for inclusion in a multi-variate model to test their predictive value. Prospective use of these indicators and models is planned to further establish their reliability.

Conference Topic
Indicators

Introduction
The National Cancer Institute (NCI) of the US National Institutes of Health (NIH) continues to show a commitment to encouraging transformative research, which the NIH recognizes on its Transformative Research Award website as "unconventional research projects that have the potential to create or overturn fundamental paradigms." Key requirements for identifying and nurturing these potential scientific breakthroughs are an enhanced understanding of the research landscape and awareness of novel approaches with great potential.

Defining Breakthrough Publications
The term "breakthroughs" has been used in prior work by Thomson Reuters (Ponomarev et al., 2014) and operationally, breakthrough publications have previously been defined as those that are highly cited and result in a change in research direction. The body of literature addressing breakthrough publications also uses the term "transformative research." Here, we define a breakthrough publication as an article that results from transformative research. In 2007, the National Science Board (NSB) defined transformative research as "research driven by ideas that have the potential to radically change our understanding of an important existing scientific or engineering concept or leading to the creation of a new paradigm or field of science or engineering. Such research also is characterized by its challenge to current understanding or its pathway to new frontiers" (NSB, 2007).

Prior Work Identifying Breakthrough Publications
Much of the research literature on breakthroughs focuses on retrospective identification of breakthroughs or pivotal points within a specific topic or field (Chen, 2006; Compañó & Hullmann, 2002; Fujita et al., 2012; Huang et al., 2013; Klavans et al., 2013; Ponomarev et

al., 2014). In addition, many of the current approaches require manual selection or curation of all data analysed (Chen, 2006; Klavans et al., 2012). Ponomarev et al. (2014) used variations of a single indicator, citation velocity, to predict highly cited papers while other groups made use of multiple indicators, full-text data and/or co-citation analysis to identify and characterize breakthrough publications in retrospective analyses (Chen, 2006, 2012; Klavans et al., 2012; Klavans et al., 2013). Other efforts focused on the development of analysis and visualization tools for quick visualization and assessment of potential turning points and breakthroughs (Boyack & Börner, 2003; Dunne et al., 2012).

Here, we aim to establish automated and semi-automated approaches to provide early indicators of published research with great potential. The goal is to provide program staff with a robust methodology that highlights pockets of breakthrough research, thereby enabling more informed program management. The methodology leverages an array of indicators to identify work that may contribute significantly to progress in its field. Here we describe work done to identify time-dependent and -independent publication indicators for differentiating breakthrough papers.

Data and Methods

Creating a Gold Standard Data Set

The first challenge in testing the importance of various publication features in predicting research breakthroughs is defining a core set of publications to be used as a gold standard. For our gold standard set of breakthroughs, we selected research articles from the following sources that highlight advances in cancer research:

1. The American Association of Cancer Research (AACR) publishes the AACR Cancer Progress Report annually (176 articles from the 2011-2014 reports).
2. The American Society of Clinical Oncology (ASCO) reports on key research in their annual Report, ASCO Clinical Cancer Advances. (58 articles from the 2009-2013 reports).
3. *Nature Medicine* 2011 special edition focused on advances in cancer research (74 articles spanning publication years 2008-2010).

Using these three sources we identified 287 distinct breakthrough publications that were indexed in the *Web of Science.* Table 1 shows the frequency by *Web of Science* Journal Subject Category. The inclusion of older publications (e.g., publication years of 2008 and 2009) enabled the curation of a dataset that included papers mature enough to have a range of breakthrough characteristics.

Table 1. Top 10 Web of Science Journal Subject Categories by Frequency for the Breakthrough Gold Standard Set (N=287).

Journal Subject Category	Count
Oncology	118
Medicine, General & Internal	109
Multidisciplinary Sciences	31
Cell Biology	17
Biochemistry & Molecular Biology	11
Public, Environmental & Occupational Health	7
Hematology	7
Genetics & Heredity	6
Immunology	6
Medicine, Research & Experimental	5

227 of the 287 breakthrough publications (81.7%) were published in journals in either the Oncology or Medicine, General & Internal *Web of Science* Journal Subject Categories.

Comparison Group Publication Set

We chose a comparison group of publications from a similar set of *Web of Science* Journal Subject Categories. We retrieved 647,879 publications from the 1) Oncology and 2) Medicine, General and Internal categories published between 2008 and 2014. We selected 2,500 publications at random from this dataset for use as the comparison group. We chose to select our control group by matching on the distribution of journal subject categories between the gold standard and comparison sets. However, we did not match the control group on publication year distribution due to the uneven publication year distribution resulting from the gold standard selection criteria.

Publication Indicators- bibliographic, citations, and altmetrics

We collected data from *Web of Science* to generate indicators for inclusion in our assessment. The majority of indicators were derived from the individual *Web of Science* citation records. These indicators were at the publication level (Table 2) and were collected in January 2015. While using a field-normalized Journal Impact Factor (JIF) would have been preferable, some publications in the gold standard set do not have JIFs determined for the publication journal, so we chose to use JIF best quartile as the best available alternative. Npayoffs reflects the inclusion of altmetrics gathered from *Web of Science* usage.

Table 2. Publication-level Indicators Considered For Inclusion in Random Forest Models.

Indicator level	Variable	Description
publication	TimesCitedTotal	total cites
	TimesNSCitedTotal	total cites (non-self)
	TimesCited2y	total cites in past 2 years
	TimeNSCited2y	total non-self cites in past 2 years
	NPages	total number of pages in an article
	NCitedRefs	number of references
	NAuthors	number of authors
	PubYear	publication year
	NCitedJSC	number of JSCs present in cited references
	NCountries	number of countries associated with publication authors
	NOrgs	number of institutions associated with publication authors
	CitVel6m	Citation velocity of specified time period (or maximum number of days since the article was published)
	CitVel1y	
	CitVel2y	
	CitVel5y	
	Bestquartile	Journal's best quartile from the 2013 Journal Citation Report
	DocumentTypeID	Describes publication type (article, review, etc.)
	Npayoffs	Total number of payoff events in Web of Science since January 2013 • A payoff event is when a WoS user downloaded the full-text article, added EndNote library, or saved for future use • Robot data filtered using multiple algorithms

Author-level indicators, person disambiguation

Some of the indicators in the study at the publication-level require a time lag after publication so we sought to increase the number of indicators that could identify potential breakthroughs immediately upon publication. Currently, these additional indicators are based on author publication history characteristics (Table 3). A critical aspect of author-based indicators is ensuring that each author's characteristics are correctly attributed. Therefore, we used a proprietary semi-automated algorithm to disambiguate authors and assign publications to each unique author.

Author-level indicators were assigned to each publication and computed in one of two ways: by averaging the indicator for all authors on a publication or by averaging the indicator for the top three authors on the paper as ranked by the indicator values.

Table 3. Author-level Indicators Considered for Inclusion in Random Forest Models.

Indicator level	Variable	Description
author	AvgNCoAuth	Number of distinct co-authors on all publications in the journal subject categories of oncology or general and internal medicine from 2008-2014
	AvgNCoAuth_Top3	
	AvgHindex	H-index based on all publications in the journal subject categories of oncology or general and internal medicine from 2008-2014
	AvgHindex_Top3	
	AvgPubHist	Total number of publications in the journal subject categories of oncology or general and internal medicine from 2008-2014 divided by six years
	AvgPubHist_Top3	
	NHighCitPubs	Highly cited publications defined by top 10% of publications in a particular year and journal subject category
	AvgNHighCitPubs	
	AvgNHighCitPubs_Top3	

Random Forest™ Model

We used the Random Forest™ machine learning algorithm (Brieman, 2001) as implemented by Liaw and Wiener (Liaw & Wiener, 2002) to assess the relative importance of each of the indicators listed above for differentiating breakthroughs from our comparison group. As Random Forest™ cannot handle null values; we were required to exclude all publications without citations and all publications where authors could not be disambiguated. This resulted in a final dataset of 223 breakthrough publications and 1,170 comparison publications.

The Random Forest™ algorithm is an example of a bagged decision tree algorithm (Breiman, 1996) that combines the classification results of some number N of individual decision trees. This set of N trees comprises the forest and is one of two input parameters that can be specified by the user. The other input parameter is an integer m which specifies the number of variables to consider when deciding how many variables to use for each node in the tree. Details on implementing this algorithm can be found in Liaw 2002 and references therein. As the random forest is built, a random subset of 2/3 of the data is used in the construction of each tree. The remaining 1/3 of the data is referred to as 'out-of-bag' (oob). For the analyses shown, the values $N = 500$ and $m = 4$ were found to minimize the out-of-bag error rate, which is a measure of the misclassification of the oob data by the random forest.

Results

We first examined the correlation among our publication indicators and removed the following indicators that were highly correlated: CitVel6m; CitVel2y; CitVel5y; TimesCitedTotal; TimesCited2y; AvgHindex_Top3; NHighCitedPubs_Top3. With the remaining set of indicators, we then ran the first Random Forest models using both the Mean

Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) to determine the relative importance of the indicators, as shown in Figure 1. The indicators with the highest relative importance are time-dependent (left of the dotted line). However, in order to best inform program management, it would be preferable to predict breakthroughs soon after publication, requiring indicators that can be calculated at, or near, the time of publication.

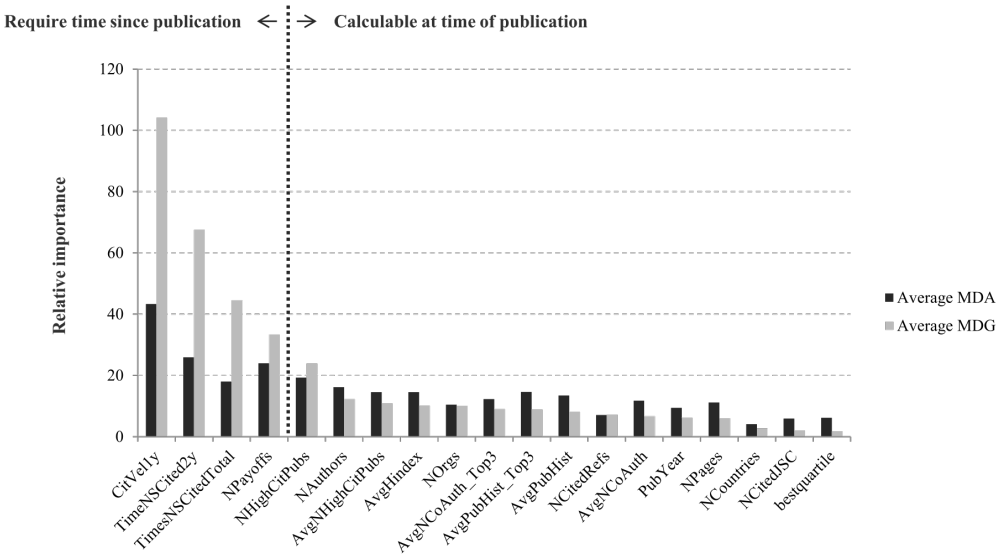


Figure 1. Relative Importance Ranking of Time-Independent and –Dependent Indicators based on Random Forest models (MDG and MDA). Out-of-bag error rate is 4.67%.

Because this work focuses on identification of publications with strong breakthrough potential near time of publication, we then considered only the time-independent indicators and produced new Random Forest models using these data. The relative importance ranking of the time-independent indicators are shown in Figure 2.

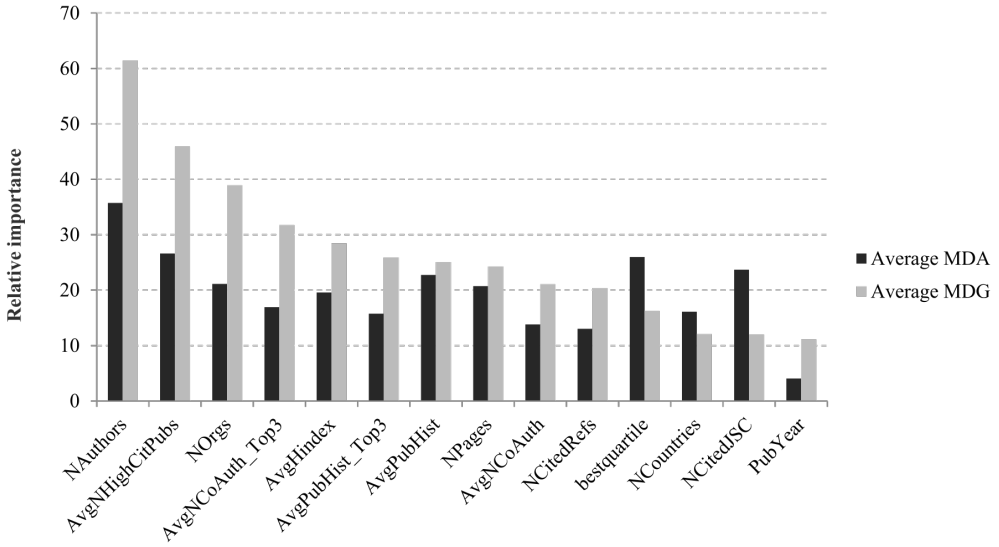


Figure 2. Relative Importance Ranking of Time-Independent Indicators based on Random Forest models (MDG and MDA). Out of bag error rate is 9.48%.

The highest ranked time-independent indicators, sorted by Average MDG, were: NAuthors, AvgNHighCitPubs, NOrgs, AvgNCoAuth_Top3, and AvgHindex. Sorting by Average MDA gives a slightly different set of top five variables: NAuthors, AvgNHighCitPubs, bestquartile,

NCited Journal Subject Category (JSC), and AvgPubHist. While the first two variables are the same for either type of ranking, it would be interesting to explore the divergence of the other variables between the two rankings. The relative importance of these time-independent indicators is consistent with breakthrough work being associated with teams and researchers with a history of strong performance.

Conclusions and Next Steps

We have identified and ranked a set of time-dependent and -independent indicators for their importance in differentiating a set of breakthrough publications from a comparison group. Our results are early steps in developing tools for potentially identify promising emerging research in a timely manner. Our next steps include using a subset of these indicators to establish a multivariate model where the outcome is the estimated probability of being a breakthrough paper based on the existing training set. Using this model, we will prospectively identify candidate breakthroughs and share the results with program officers within NCI to assess the practical value of the model. Future work could include efforts to determine which indicators gain or lose predictive value over time through iterative evaluation of the relative strength and importance of each indicator.

Acknowledgments

This study was improved by contributions from Danielle Daee (NCI); Di Cross, Leo DiJoseph and Joshua Schnell (Thomson Reuters); and extends work by Ilya Ponomarev (formerly Thomson Reuters) and Charles Hackett (National Institutes of Allergy and Infectious Diseases). This work was supported in part by NIH contract #HHS263201000058B.

References

Boyack, K.W., & Börner, K. (2003). Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers, *Journal of the American Society for Information Science and Technology*, *54*, 447-461.

Breiman, L. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

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*, 359-377.

Compañó, R., & Hullmann, A. (2002). Forecasting the development of nanotechnology with the help of science and technology indicators, *Nanotechnology*, *13*, 243.

Dunne, C., Shneiderman, B., Gove, R., Klavans, J., & Dorr, B. (2012). Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization, *JASIST*, *63*, 2351-2369.

Fujita, K., Kajikawa, Y., Mori, J., & I. Sakata. (2012). Detecting Research Fronts Using Different Types of Combinational Citation, *Detecting Research Fronts Using Different Types of Combinational Citation*.

Huang, Y.H., Hsu, C.N., & Lerman, K. (2013). Identifying Transformative Scientific Research, *IEEE 13th International Conference on Data Mining* (ICDM), (pp. 291-300).

Klavans, R., Boyack, K.W., & Small, H. (2012). Indicators and precursors of "hot science", *17th International Conference on Science and Technology Indicators*, (pp. 475-487).

Klavans, R., Boyack, K.W., & Small, H. (2013). Identifying Emergent Opportunities in Science. Retrieved June 2, 2015 from: http://www.mapofscience.com/pdfs/EAGER_Final_v1.pdf

Liaw, A. & Wiener, M. (2002). Classification and Regression by Random Forest. *R News*, 2/3, (pp. 18-22).

National Science Board. (2007). Enhancing Support of Transformative Research at the National Science Foundation, *National Science Foundation*, (p. 14).

ODNI, (2011). IARPA Launches New Program to Enable the Rapid Discovery of Emerging Technical Capabilities.

Ponomarev, I.V., Lawton, B.K., Williams, D.E., & Schnell, J.D. (2014). Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction?, *Scientometrics*, *100*, 755-765.

Reardon, S. (2014). Text-mining offers clues to success: *Nature*, *509*, 410.

Dimensions of the Author Citation Potential

Pablo Dorta-González[1], María-Isabel Dorta-González[2] and Rafael Suárez-Vega[3]

[1] *pablo.dorta@ulpgc.es,* [3] *rafael.suarez@ulpgc.es*
Universidad de Las Palmas de Gran Canaria (Spain)

[2] *isadorta@ull.es*
Universidad de La Laguna (Spain)

Introduction

It is well known that in some fields the average number of citations per publication is much higher than in others (Moed, 2005).

For decades, the number of publications and the number of citations have been the two accepted indicators in ranking authors. Recently, alternative indicators which consider both production and impact have been proposed (Dorta-González & Dorta-González, 2011; Egghe, 2013). However, these indicators based on the h-index do not solve the problem when comparing authors from different fields of science. Given the large differences in citation practices, the development of bibliometric indicators that allow for between-field comparisons is clearly a critical issue (Waltman & Van Eck, 2013).

Traditionally, normalization of field differences has usually been based on a field classification system. In said approach, each publication belongs to one or more categories and the citation impact of a publication is calculated relative to the other publications in the same field.

In our topic normalization we use the aggregate impact factor of three different sets of journals as a measure of the different dimensions in the citation potential of an author.

Dimensions of the author citation potential

Even within the same field, each researcher is working on one or several research lines that have specific characteristics, in most cases very distant from those of other researchers.

Generally, the citation potential in a field is determined within a predefined group of journals. This approach requires a classification scheme for assigning publications to fields. Given the fuzziness of disciplinary boundaries and the multidisciplinary character of many research topics, such a scheme will always involve some arbitrariness and will never be completely satisfactory. Therefore, we propose measuring the citation potential in the specific topic of each author and using this measure as an indicator of the probability of being cited in that topic.

The problem underlying the characterization of the author citation potential is as follows. Given a set of publications from an author in different journals and years, we will try to obtain a measure of the author topic defined by some dimensions of these publications so it can be compared with that of a different author (with publications in different journals and years).

Let us consider a 5-year time window Y. In this paper, we propose characterizing the topic of an author in period Y using three different dimensions (see Figure 1): the weighted average of the impacts in the journals containing the author's papers in Y (production dimension P), the weighted average of the impacts in the journals citing the author's papers in Y (impact dimension I), and the weighted average of the impacts in the journals included as references in the author's papers in Y (reference dimension R).

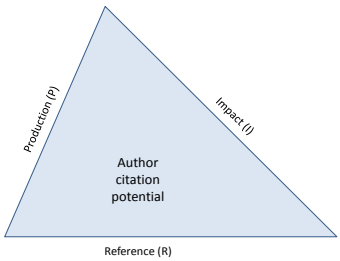


Figure 1. The three dimensions of the author citation potential.

In this characterization we propose the use of journal impact indicators instead of number of citations received by a particular paper. This is because it is necessary that several years pass after the publication of a document, so that the number of citations can be a consistent indicator in comparing similar documents of the same type published in the same year with that of other researchers in the same field. In some fields (e.g., Economics) more than 5 years are needed to obtain a consistent measure of impact (Dorta-González & Dorta-González, 2013). In many fields of the Humanities it is necessary to wait even longer (Dorta-González & Ramírez-Sánchez, 2014).

Materials and Methods

The bibliometric data was obtained from the online version of the Scopus database. Only journal papers in the period 2009-2013 were included, considering for each journal the Scimago Journal Ranking –SJR–. Four subject areas were considered: Chemistry, Computer Science, Medicine, and Physics & Astronomy. This was motivated in order to obtain authors with systematic differences in publication and citation behavior. We designed a random sample with a total of 120 authors (30 in each subject area). They were selected from the highly productive authors of the Consejo Superior de Investigaciones Científicas –CSIC– (Spain).

Results and discussion

The subject areas considered are very different in relation to the citation behavior. For this reason, in the sample there are important differences among the dimensions of the citation potential from one author to another. However, the proportion between production and impact dimensions is very close in all the subject areas considered (Figure 2).

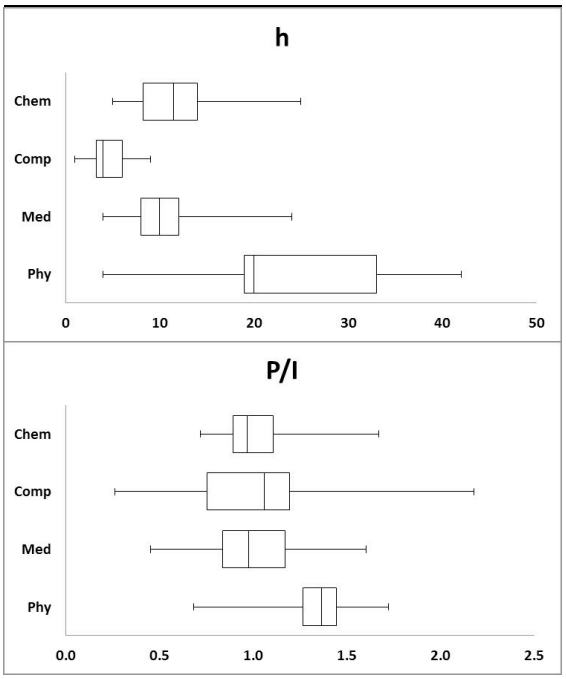


Figure 2. Box-plots comparing the subject areas.

Within- and between-group variability are both components of the total variability in the combined distributions. So: within variability + between variability = total variability.

Note in Table 1 that the proportion between production and impact dimensions produces the greatest percentage reduction of the variance. A more detailed analysis of the results can be found in Dorta-González et al. (2015).

Table 1. Central-tendency and variability.

	P	I	R	P/I
Median	1.521	1.526	2.564	1.065
Mean	1.719	1.546	2.759	1.093
Range of variation	3.692	3.776	7.527	1.915
Within-group variance	46.360	25.089	192.557	9.972
Between-group variance	39.434	17.325	54.463	2.358
Reduction in the variance	14.9%	30.9%	71.7%	76.3%

Conclusions

We have developed a measure of scientific performance whose distributional characteristics are invariant across scientific fields. Such a measure could be employed in the normalization of the impact at the author level in order to allow direct comparisons of scientists in different fields and permit a ranking of researchers that is not affected by differential publication and citation practices across fields.

References

Dorta-González, P., & Dorta-González, M. I. (2011). Central indexes to the citation distribution: A complement to the h-index. *Scientometrics*, *88*(3), 729-745.

Dorta-González, P., & Dorta-González, M. I. (2013). Comparing journals from different fields of science and social science through a JCR subject categories normalized impact factor. *Scientometrics*, *95*(2), 645-672.

Dorta-González, P., Dorta-González, M.I., & Suárez-Vega, R. (2015). An approach to the author citation potential: Measures of scientific performance which are invariant across scientific fields. *Scientometrics*, *102*(2), 1467-1496.

Dorta-González, P., & Ramírez-Sánchez, M. (2014). Producción e impacto de las instituciones españolas de investigación en Arts & Humanities Citation Index (2003-2012). *Arbor*, *190*(770), a191.

Egghe, L. (2013). Theoretical justification of the central area indices and the central interval indices. *Scientometrics*, *95*(1), 25-34.

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

Waltman, L., & Van Eck, N. J. (2013). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. *Scientometrics*, *96*(3), 699-716.

Scholarly Book Publishers in Spain: Relationship between Size, Price, Specialization and Prestige

Jorge Mañana-Rodríguez[1] and Elea Giménez Toledo[2]

[2] *jorge.mannana@cchs.csic.es*
Centre for Human and Social Sciences, Spanish National Research Council. C/ Albazanz, 26-28. Madrid. Spain.

Introduction

The prestige of book publishers is an important element for the assessment of SSH scholars in Spain. Until 2012, that 'prestige' remained based upon subjective, individual judgements from assessment committees' members. In order to provide a more objective reference for the prestige of book publishers, ÍLIA research group developed a ranking of book publishers (so called SPI) based on the opinion of almost three thousand experts from all SSH fields (Gimenez et al., 2013). Nevertheless, the factors underlying the perceived prestige are unknown. Some authors worked on the influence of marketing on the perception of books. Squires (2007) point out that 'we should not underestimate the value or efficiency that the association with a specific publisher provides to its contents'. It is hypothesized that three factors (among others) might be related to the perceived prestige: size of the book publisher (number of titles published), specialization (share of titles in each discipline) and price of the books. This research present the results of a correlational study on prestige, size, specialization and price of SSH book publishers in Spain.

The perception of 'prestige' strongly differs among different subjects to which the term can be applied. When the object is a product or a brand (with book publisher names as equivalent) the quantifiable variables related to the perception by different subjects of the different levels of prestige is relevant for explaining or defining the construct. The overall number of titles published by a book publisher could act as a reinforcement of the perception of prestige since the frequency with which the reader or consumer will be exposed to the brand is statistically more probable and this could lead to a perception of the publisher as able to publish more and better than others. In many goods, the perception of the prestige of competitors, in a similar way to how multi-branding strategies operate (Rahnamaee, A., & Berger, 2013). A brand prestige might also affected by the price (Yeoh & Paladino, 2013), and so the price of book might partially contribute, in a linear fashion, to the perceived prestige of book publishers.

Finally, specialization, as a factor, which might create a link between a specialized scholar with an specialized publisher, might contribute to influence the perception of the publisher as more prestigious in absolute terms. Since Scholarly Publishers Indicators (SPI) is being currently used as a source of information for assessment procedures in Spain (in some SSH fields), it is important to know whether the perceived prestige can be attributed to factors unrelated to the essential issues in research evaluation or if, by the opposite, the perceived prestige is not strongly (linearly) associated to these external factors.

Objectives

The objective of this research is to test the hypothesis stating that there is a linear relationship between prestige, size, specialization and price of books of book publishers in the case of Spain.

The information sources are the following:

-Prestige values: Scholarly Publishers Indicators (SPI, 2012).

-Size, price and specialization: DILVE (DILVE, 2013).

Variable definition:

-Prestige: ICEE (Prestige measure based on extensive survey to researchers and lecturers)

-Size: Raw number of different titles in DILVE for each discipline

-Mean price: the average price of all the titles published by the book publisher in the period analyzed.

-Max. Price: the maximum price of a single title in the whole set of titles published by each publisher.

-Specialization: Share of titles of publisher according to DILVE.

Methodology

For a total number of 119 book publishers (this number was fixed so that the number of lost cases is minimized), their ICEE was retrieved from SPI (2014, and the size, mean price and specialization degree obtained from the extensive database DILVE, for the years 2004 onwards up to 2012. The reason for including data from 2004 onwards is the fact that prestige, as other consumer perceptions, are developed over time so a smaller time span would not provide suitable. Data prior to 2004 is not fully consistent in DILVE database when compared with the publishers resulting from the questionnaire on publishers prestige due to the several changes (splits and merges) which took

place sin that date among book publishers, often involving the disappearance of book publishers names as they were and therefore requiring a much more complex codification of the previous names in order to keep the reliability of the data set. After a verification of the non-normality of the distribution of all the variables, using Kolmogorov-Smirnov nonparametric tests, Spearmans' Rho was selected as the appropriate technique contrasting the linear association hypothesis. The correlation matrix for all the variables was calculated using IBM SPSS (v. 19).

Results

Only significant results (p-value = .05) have been considered, since there is no reason for supposing any bias effect of *n* on the significance of the results (119, in all cases). The following table resumes these statistically significant correlations.

Table 1. Statistically significant correlations (Spearman's Rho).

ρ Publisher Prestige, Raw Size	.269; p < .05
ρ Publisher Prestige, Max Price	.217; p < .05
ρ Raw Size, Max Price	.198; p=.019
ρ Raw Size, Average price	-.232; p < .05
ρ Raw Size, Max Share	.433; p < .05
ρ Max Price, Average price	.593 p < .05

Conclusions

The main conclusion which can be drawn from the results is the seemingly (at least linear) independence of the construct 'prestige' from all the variables hypothesized as potentially influential in the values given to book publishers by the experts. The correlations of publishers' prestige with Raw Size (Number of Titles) and Max. Price, although statistically significant, are small enough as to suppose that the influence of these two variables in the perception of a publisher's prestige is not strong enough as to make necessary normalization measures. These results also suggest (at least from the perspective of a linear relationship) that the rankings in use are not biased by the possible influence of the great number of books, multiple branding and specialization or prices which sometimes can be displayed by some of the publishers belonging to big publishing houses which occupy the highest positions in the rankings.

Discussion

The fact that none of the variables analyzed is linearly related to the perceived prestige of book publishers is consistent with the multi-component structure generally involved in the composition of a concept such as 'prestige'. Also, since it is hardly possible to quantify the 'quality' (an also multi-faceted concept, particularly in the framework of research evaluation) of the contents of the books which, escalated to book publisher level of aggregation could contribute to the perceived prestige, the plausible influence of this factor remains unknown, although further research might offer new insight into this particular relationship. The existence of such relationship between the intrinsic quality of the contents and the prestige of a publisher is also plausible given that the use of books by those who have provided the prestige values presumably use the books as a source of information and as a form of scholarly communication where the quality of the contents might be the core of the perceived prestige, leaving behind other subjectively perceived variables. Also, given the relevance of peer review for assessment processes (Verleysen & Engels, 2013) as well as for the quality of the contents, the use of these filters might be related to the perceived prestige of book publishers.

References

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation*, 22(1), 64–77. doi:10.1093/reseval/rvs036

Rahnamaee, A., & Berger, P. D. (2013). Investigating consumers' online purchasing behavior: Single-brand e-retailers versus multi-brand e-retailers. *Journal of Marketing Analytic*s, 1(3), 138-148.

Squires, C. (2007). *Marketing Literature: The Making of Contemporary Literature.* Basingstoke: Palgrave Macmillan.

Verleysen, F. T. & Engels, T. C. E. (2013). A Label for Peer-Reviewed Books. *Journal of the American Society for Information Science and Technology*, 64, 428-430

Yeoh, M., & Paladino, A. (2013). Prestige and environmental behaviors: Does branding matter*? Journal of Brand Management*, *20*(4), 333-349.

Bootstrapping to Evaluate Accuracy of Citation-based Journal Indicators

Jens Peter Andersen[1] and Stefanie Haustein[2]

[1] *jepea@rn.dk*
Medical Library, Aalborg University Hospital, Sdr. Skovvej 15, 9000 Aalborg (Denmark)

[2] *stefanie.haustein@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal (Canada)

Introduction

Bibliometric indicators ranking aggregate units have a long tradition, including criticisms of methodology, interpretation and application. Despite the criticism, there is a demand for these indicators, and recent developments have led to improvements of methodology and interpretation. An essential element of these interpretations is to provide estimates of the accuracy, robustness, stability and confidence of bibliometric indicators, thereby providing the reader with data required to interpret results. This has, for example, been demonstrated for the set of indicators in the Leiden ranking (Waltman et al., 2012), the Journal Impact Factor (Chen, Jen, & Wu, 2014) and other journal indicators (Andersen, Christensen, & Schneider, 2012) as well as author metrics (Lehmann, Jackson, & Lautrup, 2008). The present study applies the same type of bootstrapping technique to estimate stability, as is used in the Leiden ranking (Waltman et al., 2012), on an array of citation-based journal indicators. The purpose of this analysis is to compare recent methodological advances, as well as traditional approaches. The study is based on clinical medicine journals in the Web of Science (WoS).

Methods

Data acquisition

The dataset contains all articles and reviews in the WoS, published in 2012 in journals classified as clinical medicine according to the National Science Foundation (NSF) classification system. This amounts to 362,556 papers and 2,699 journals from 34 different specialties within the discipline of clinical medicine. Each journal and paper is assigned to exactly one specialty. Citations are observed for a two-year window. In order to account for field differences in citation patterns, relative citations, $\hat{c}$, are computed by normalising observed against expected citations per specialty and year.

Journal indicators

The journal citation indicators selected for this study represent both traditional (means and medians of observed and relative) and novel (percentile) approaches. For a given journal *j*, we calculate the mean citations, μ_c, median citations, M_c, mean relative citations, $\mu_{\hat{c}}$, median relative citations, $M_{\hat{c}}$, top decile ratio of citations, N_{D10}, and relative citations. The top decile ratio for a journal is the percentage of papers present in the overall set of papers with citations in the highest decile range.

Indicator evaluation

Each indicator is evaluated for every journal by performing bootstrapping (Efron & Tibshirani, 1993). The technique involves resampling with replacement, i.e. for a given sample, all observed values are resampled so that a new sample of the same size is drawn randomly, but with the possibility that the same observation can be drawn multiple times. When repeating this resampling numerous times, we can calculate stability intervals to estimate how accurately the observed indicator value describes the underlying observations or whether it is influenced by outliers and thus less robust. To make our results comparable to those reported in the Leiden ranking, we have chosen to iterate each bootstrap 1,000 times and calculate 95% confidence intervals. In addition to this confidence interval we also calculate the standard deviation for each distribution. As the values of the different indicators are observed in very different ranges, we provide an additional mean-standardized version of every indicator. All calculations are performed using the *boot* package (Canty & Ripley, 2015) for *R* version 3.0.3 x64 (R Development Core Team, 2010).

Results and Discussion

We find that bootstrapping can identify outlying indicator scores within a specialty, by showing stability intervals (95% confidence intervals) for every indicator. As exemplified in Figure 1 for the subset of dentistry journals, the stability intervals demonstrate the robustness of rankings based on particular indicators. While, for example, the

stability intervals indicate that the citation impact of the 1st journal in Figure 1 is higher than that of the 5th, the first four journals cannot be clearly distinguished in terms of mean citation impact. Their mean citation rates are heavily influenced by a few highly cited papers.

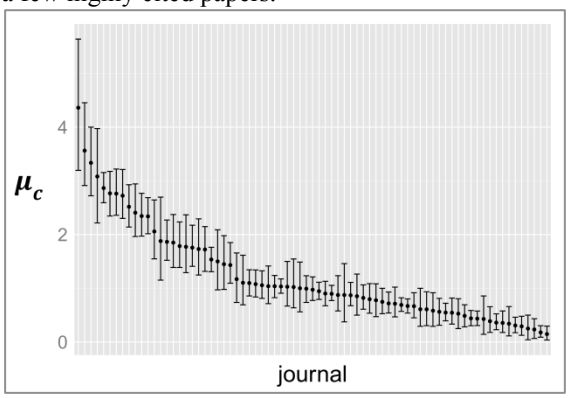


Figure 1. μ_c with stability intervals for all journals in the dentistry specialty.

The study also shows that the percentile-based indicators perform considerably better regarding stability than both mean- and median-based indicators (Figure 2 and Table 1). It is particularly interesting that the medians indicators do not seem to be more stable than the means.

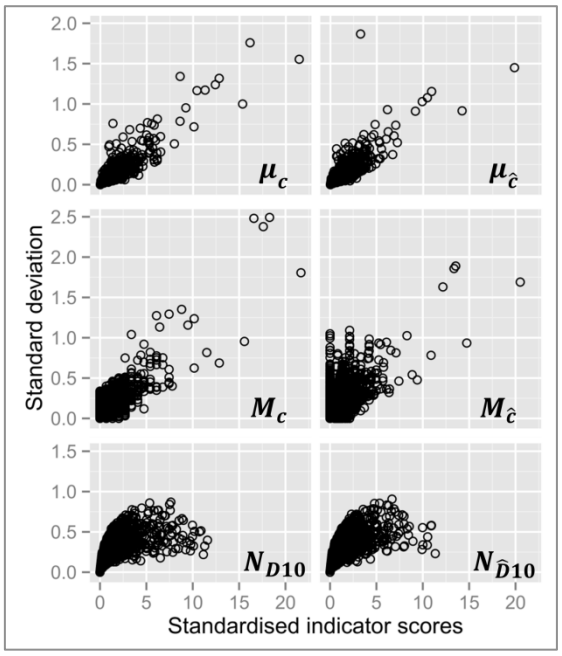


Figure 2. Standard deviation of bootstrapped scores as a function of standardised indicator scores, limited to journals with at least 50 papers.

Finally, we show that indicators are extremely sensitive to sample sizes. Journals with less than 50 papers published in the observation period show significantly larger variance than those publishing at least 50 papers (Table 1). Our results reiterate the importance of testing indicators and providing stability intervals to improve their interpretability. This would identify the limitations of rankings and avoid cases like the 24-fold increase of *Acta Crystallographica A*'s impact factor in 2009 (Haustein, 2012).

Table 1. Mean indicator values and standard deviations for all journals ("All") and journals publishing 50 or more papers ("≥50").

	All				≥50	
	Raw		Standardised			
Indicator	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*
μ_c	2.321	3.897	1.000	1.679	1.052	1.261
M_c	1.477	2.278	1.000	1.543	1.079	1.471
$\mu_{\hat{c}}$	0.835	1.107	1.000	1.326	1.053	1.076
$M_{\hat{c}}$	0.520	0.717	1.000	1.381	1.075	1.297
N_{D10}	0.081	0.131	1.000	1.625	1.107	1.640
$N_{\hat{D}10}$	0.078	0.119	1.000	1.536	1.090	1.513

Further research will include in-depth analyses of multiple indicators and differences of stability intervals across specialties.

References

Andersen, J. P., Christensen, A. L., & Schneider, J. W. (2012). An approach for empirical validation of citation-based journal indicators. In E. Archambault, Y. Gingras, & V. Lariviére (Eds.), *Proc. of STI 2012* (pp. 71–81). Montréal, Canada: 17th International Conference on Science and Technology Indicators.

Canty, A., & Ripley, B. (2015). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-15.

Chen, K. M., Jen, T. H., & Wu, M. (2014). Estimating the accuracies of journal impact factor through bootstrap. *Journal of Informetrics*, *8*(1), 181–196.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the Bootstrap* (p. 456). New York: Chapman & Hall.

Haustein, S. (2012). *Multidimensional Journal Evaluation. Analyzing Scientific Periodicals beyond the Impact Factor*. Berlin / Boston: De Gruyter Saur.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2008). A quantitative analysis of indicators of scientific performance. *Scientometrics*, *76*(2), 369–390.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., Wouters, P. F. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *JASIST*, *63*(12), 2419–2432.

The Lack of Stability of the Impact Factor of the Mathematical Journals

Antonia Ferrer-Sapena[1], Enrique A. Sánchez-Pérez[1], Fernanda Peset[1], Luis-Millán González[2] and Rafael Aleixandre-Benavent[3]

[1]*anfersa@upv.es, easancpe@mat.upv.es, mpesetm@upv.es*
Instituto de Diseño y Fabricación, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia (Spain)

[2]*luis.m.gonzalez@uv.es*
Departamento de Educación Física y Deporte, Universitat de València. Gascó Oliag, 3. 46010 Valencia (Spain)

[3]*aleixand@uv.es*
INGENIO (CSIC-Universitat Politècnica de València). UISYS-Universitat de València. Plaça Cisneros, 4. 46003-València (Spain)

Introduction

Although the 2-year Thomson-Reuters Impact Factor (IF) has become a usual tool for measuring the scientific productivity of all fields of the natural sciences (see Aleixandre-Benavent, Valderrama Zurián, & González Alcaide, 2007), its behavior in the particular case of the journals of pure mathematics (the area MATHEMATICS in the thematic directory of Thomson-Reuters) is far from being stable when its values in consecutive years are considered. If we consider the changes of the values of the IF of a given journal in the last decade, it can be easily seen that the variation of the values is surprisingly high if we compare with other disciplines. Mathematical journals seem to have the worst behavior regarding the time stability both of the IF and the position in the IF list.

A series analysis of a set of journals uniformly distributed in the IF list shows that the variations of the values of the IFs are very big when compared with other scientific disciplines, e.g., APPLIED PHYSICS and MICROBIOLOGY. The reader can see a representation of this behavior for three mathematical journals together with three journals of physics that have been chosen as representatives of these groups in the following graph (Fig. 1).

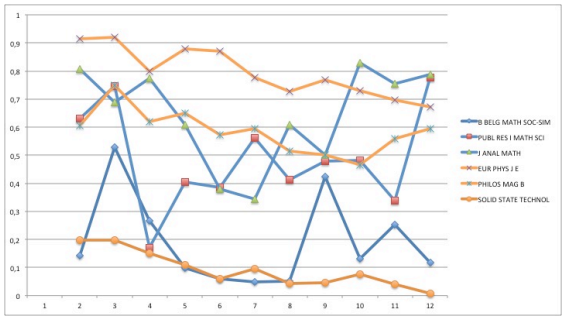


Figure 1. Variations of three journals of mathematics and three journals of physics.

In our study, we analyze the possible reasons for this fact, explaining some typical characteristics of the mathematical journals and of the research in mathematics, that make this science to have unusual properties from the point of view of the bibliometrics.

The research in pure mathematics

In general, mathematicians work in small groups of researchers from different parts of the world that are specialized in some topics, which have a long development period. For instance, it is usual that a group of mathematicians continue with some problems that appeared 50 years ago, or even before (see Behrens & Luksch, 2011). Although some of these topics were intensively studied some years ago, sometimes the research was left at that moment without having complete answers for some central questions, due to the fragility and the small size of the specialized group of researchers working on it. In this context, it is natural that after some years, a new group can recover the research and fruitfully continue with the investigation. The group of interested mathematicians is, almost in all cases, small. Even in new open topics, the size of the interested community of mathematicians is sparse and small. This of course changes when some particular theory becomes important due to the applications. But in these cases, the publication of the mathematical contents is redirected to more applied journals, or to journals of the fields where the theory finds applications.

This research dynamics is not usual at all, if we compare it with the pattern that can be observed in other fields. The main consequence is that the obsolescence of the scientific documents is faster in other sciences than in mathematics.

Mathematical journals

Classical journals that publish papers on pure mathematics follow also a different pattern that the

usual one in other scientific fields that are in some sense similar with respect to some descriptive parameters, as physics or other natural sciences. Although there are a lot of journals that are supported by big publishers—for example, Elsevier and Springer—, some of them preserve the editorial policy and the publication format that they used to have before. Another important group of journals is still published by national societies, universities and research institutes. Very often, these publications are small—in the sense that they publish a small number of papers per year—, but they are prestigious and serious papers are published in them.

This implies that the impact factor of these journals has a strong statistical variability, depending on the number of citations that a small number of papers can receive.

On the other hand, the publication of the papers is slow when compared with journals in other disciplines. Sometimes it takes more than two years for a paper from submission to publication. In general, this does not produce any problem for the dissemination and exchange of information, since the contents are often previously published by the authors in popular open access repositories as arXiv. Moreover, again the small size of the group of specialists interested in the topic reduces the pressure on the authors for a fast publication.

Conclusions: IF-based evaluation of the scientific productivity

The main direct consequence of the properties of the journals of mathematics together with the slow long-term activity in the research of the topics is the small rate of papers that are cited two years after their publication, when compared with other fields. This causes that the value of the IF of the journals is small even if they are prestigious and well-known in the field. For example, an IF of 0.5 is a reasonable impact factor for a journal, and enough to let it to be considered as a serious publication. This value is very small if we compare with other areas (see Bensman, Smolinsky & Pudovkin, 2010; Smolinsky & Lercher, 2012).

However, the 2-year IF is still the main tool in many countries—for example, Spain—to measure the production of a single mathematician or a research institute. This produces some fails in the evaluation systems, and lead the researchers to publish in journals that are considered by the community as less prestigious than others, as a consequence for example of the fact that these journals publish much more papers, and then have a better IF. Therefore, pure mathematics provides an example of a group of disciplines for which the IF-based evaluation clearly distorts the image of the scientific production.

Acknowledgments

This work has benefited from assistance by the National R+D+I of the Ministry of Economy and Competitiveness of the Spanish Government (CSO2012-39632-C02-01) and Prometeo Program for excellent research groups of Generalitat Valenciana (GVPROMETEO2013-041).

References

Aleixandre-Benavent, R., Valderrama Zurián, J. C., & González Alcaide, G. (2007). Scientific journals impact factor: limitations and alternative indicators. *El Profesional de la Informacion*, *16*(1): 4–11.

Behrens, H., & Luksch, P. (2011). Mathematics 1868–2008: a bibliometric analysis. *Scientometrics*, *86*, 179–194.

Bensman, S. J., Smolinsky, L. J., & Pudovkin, A. I. (2010). Mean citation rate per article in mathematics journals: Differences from the scientific model. *Journal of the American Society for Information Science and Technology*, *61*, 1440–1463.

Smolinsky, L, & Lercher, A. (2012). Citation rates in mathematics: a study of variation by subdiscipline. *Scientometrics*, *91*, 911–924.

Using Bibliometrics to Measure the Impact of Cancer Research on Health Service and Patient Care: Selecting and Testing Four Indicators

Frédérique Thonon[1,2], M. Saghatchian[1], R. Boulkedid[2] and C. Alberti[2]
[1]Gustave Roussy, European and International Affairs, Villejuif (France)
[2]Hôpital Robert Debré, unité d'épidémiologie clinique, Paris (France)

Introduction

Traditionally, biomedical research is measured by bibliometric indicators of scientific production and impact (such as number of publications and h-index) and indicators linked to clinical trial activities (Pozen & Kline, 2011). However, there has been an increasing demand in the last few years to measure the impact of medical research in terms of how it improves patients' well-being and public health (Wells & Whitworth, 2007; Ovseiko, Oancea, & Buchan, 2012). Measuring the final impact of research on patients' outcomes is difficult because of attribution problems and time lag between research and outcomes (Ovseiko, Oancea & Buchan, 2012). The aim of our research project is to select and test indicators measuring the impact of cancer research on health service and patient care

First step: indicators selection

See Figure 1 below for details of this process.

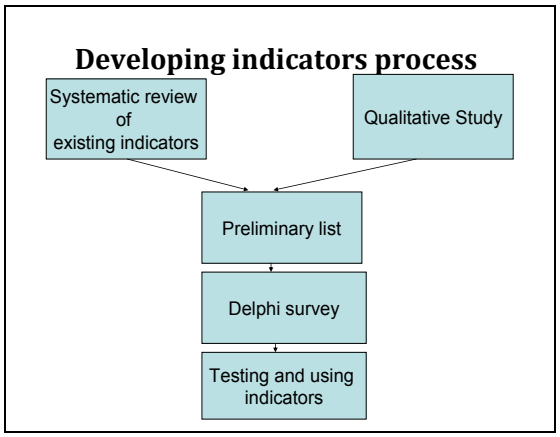


Figure 1: Indicators development process.

Systematic review of indicators

We firstly undertook a systematic review of existing indicators measuring the output and outcome of medical research in order to (1) enlist all the indicators that could potentially be used and (2) to describe their methodology, use, advantages and disadvantages. We took care of designing a study as comprehensive as possible, in order to include indicators ranging from those measuring research activity to those measuring the long-term impact of research. As a result we drew a detailed list of 57 indicators (Thonon et al., 2015).

Qualitative study of researchers

We wanted to develop indicators that would be accepted by those concerned by this evaluation system. Therefore, we undertook a qualitative study to explore the views of actors in translational research on the definitions, issues and evaluation modes of translational research. This study was done to complete the results of the systematic review with an input from the stakeholders directly involved. We interviewed 23 researchers, engineers, administrators and clinicians from diverse backgrounds and engaged in diverse fields of oncological translational research.

Delphi survey

Those two exploratory studies led us to the drawing of an initial list of 61 indicators. We submitted this list to all members of the platform for a modified Delphi survey (N=267). Participants were presented indicators, as well as their methodologies, advantages and disadvantages, and were asked to rate their feasibility and validity on a scale from 1 to 9, and to comment on them. Comments from participants were particularly useful to adjust the methodology of the indicators. In addition, a physical meeting was held where 26 participants discussed the inclusion and methodology of some indicators.

Results

As a result we were able to draw a list of 12 indicators, including 4 indicators that focused on measuring the impact of research on health service and patient care but not used in evaluation systems very often:

- Citation of research in clinical guidelines;
- Citation of research in public health guidelines;
- Number of clinical guidelines authored; and
- Number of validated biomarkers identified in publications.

Second step: indicators testing

We constructed the following methodology to measure those indicators: 17 European cancer centres have been selected in this study. We used

the Scopus database to extract all original articles published between 2000 and 2014 and analysed the data.

Citation of research in clinical guidelines

We selected clinical oncology guidelines published by the European Society of Medical Oncology (ESMO), the American Society for Clinical Oncology (ASCO), and the National Comprehensive Cancer Network. Those guidelines are published in, respectively, *Annals of Oncology, Journal of Clinical Oncology* and the *Journal of the National Comprehensive Cancer Network.* We analysed the number of publications cited in the 'clinical practice guidelines' issues of those journals. We searched the literature for data on the AGREE score of those guidelines to measure the validity of this indicator.

Authorship of clinical guidelines

We extracted and analysed data relative to the clinical oncology guidelines mentioned above.

Citation of research in public health guidelines

From the database of European publications (https://bookshop.europa.eu/en/home/) we searched for public health guidelines related to cancer. Then we extracted the references of the selected guidelines in Scopus and carried out a citation analysis.

Number of validated biomarkers identified in publications

We firstly performed a literature review to identify and list all validated biomarkers used in clinical practice for oncology patients. We then performed a search for all publications related to those biomarkers in the corpus of original articles.

Discussion

This study is still ongoing and the results will be available shortly. We believe those four indicators can provide an additional tool to measure the impact of cancer research on health service and patient care. Citation of research in clinical guidelines is the most investigated indicator (Lewison, 2003; Mostert et al., 2010). There is little literature on indicators linked to the citation of research in public health guidelines (Lewison, 2003) but none linked to indicators measuring the identification of biomarkers, despite the importance of their use for cancer patients' outcomes.

References

Lewison, G. (2003). Beyond outputs: New measures of biomedical research impact. *Aslib Proceedings, 55*, 32-42.

Mostert, S.P., Ellenbroek, S.P., Meijer, I., van Ark, G., & Klasen, E.C. (2010). Societal output and use of research performed by health research groups. *Health Research Policy and Systems, 8*, 30.

Ovseiko, P.V., Oancea, A., & Buchan, A.M. (2012). Assessing research impact in academic clinical medicine: a study using Research Excellence Framework pilot impact indicators. *BMC Health Services Research, 12*, 478.

Pozen, R., & Kline, H. (2011). Defining Success for Translational Research Organizations. *Science Translational Medicine*, *3*(94), 94cm20.

Thonon, F., Boulkedid, R., Delory, T., Rousseau, S., Saghatchian, M., van Harten, W., & Alberti, C. (2015, April 2). Measuring the outcome of biomedical research: a systematic literature review. *PLoS One*, 10(4):e0122239 doi: 10.1371/journal.pone.0122239.

Wells, R., & Whitworth, J.A. (2007). Assessing outcomes of health and medical research: do we measure what counts or count what we can measure? *Australia and New Zealand Health Policy*, *4*, 14.

A New Scale for Rating Scientific Publications

Răzvan Valentin Florian[1]

[1] *florian@epistemio.com*
Epistemio, str. Saturn nr. 26, 400504 Cluj-Napoca (Romania); 20-22 Wenlock Road, London N1 7GU (UK); and
Romanian Institute of Science and Technology, str. Cireşilor nr. 29, 400487 Cluj-Napoca (Romania)

Introduction

Citation-based bibliometric indicators are increasingly being used for evaluating research. This reflects the need of decision-makers to increase the efficiency of allocating resources to research institutions and scientists, while also keeping manageable and cost-effective the evaluation process that grounds the allocation of resources. There often is much room of improvement in how bibliometric indicators are being used in practice. But even state-of-the art bibliometric indicators suffer of a fundamental problem when used for evaluating research: the citations they are based upon are influenced by many factors beyond the quality of cited publications (Bornmann & Daniel, 2008) and these indicators need to be tested and validated against what it is that they purport to measure and predict, which is expert evaluation by peers (Harnad, 2008).

A solution to this problem is aggregating online ratings provided post-publication by the scientists who read the rated papers anyhow, for the purpose of their own research. Online-aggregated ratings are now a major factor in the decisions taken by consumers when choosing hotels, restaurants, movies and many other types of services or products. It is paradoxical that in science, a field for which peer review is a cornerstone, rating publications on dedicated online platforms is not yet a common behavior. For example, if each scientist would provide one rating weekly, it can be estimated that 52% of publications would get 10 ratings or more (Florian, 2012). This would be a significant enhancement for the evaluative information needed by decision makers that allocate resources to scientists and by other users of scientific publications.

For collecting this kind of ratings, a rating scale should be defined. Here I present the choices made during the development of the scale used at Epistemio, an online platform for aggregating ratings and reviews of scientific publications (www.epistemio.com).

Purpose

The expected usage of these ratings is: first, in steering of science by decision-makers, i.e. choosing to whom to allocate resources (typically contributed publicly), such as institutional funding, grants, jobs, positions, tenure, among the institutions, scientists, fields of science, etc. that compete for them; and second, in helping scientists to prioritize and filter the publications that they choose to read or use. For the first purpose, it is important to be possible to aggregate ratings across the set of publications of an individual, of a group of scientists or of an institution; and to be able to use the individual or aggregated ratings to rank the assessed entities. This implies that ratings should be unidimensional. While publications may be assessed across a number of characteristics, such as quality of research, quality of presentation, novelty, and interest, collecting individual ratings across all these dimensions reduces the response rates, and it is not clear how these multidimensional ratings may be aggregated into a scalar one. Therefore, it is desirable that an overall rating that reflects the overall properties of a publication is collected independently of ratings regarding individual characteristics of the publication. Collecting the latter may be left optional. This paper focuses on the overall rating.

What should be rated, exactly?

When experts are asked to rate a publication, the property that should be rated must be named. What is exactly this property? A proper discussion of this issue should analyze the foundations of scientific research, being outside the scope of the present paper. A different way of posing the problem is starting with the needs of expected users of the ratings, which were mentioned above. Typical desired properties of publications (and, therefore, of the results presented in these publications) that are mentioned in the context of steering of science is quality, importance, relevance, and impact. For usability purposes, the text of the question to raters should be kept brief; therefore, a choice must be made among the various wordings that may be used. Importance, long-term societal and scientific relevance, and long-term societal and scholarly impact seem to have similar semantics. Quality seems to be a complementary property: a publication may present potentially important results, but methodology and/or presentation may lack quality, therefore raising uncertainties about the real value of the publication; and a publication may be of high quality while the potential importance is low. We have thus chosen to use the wording “scientific quality and importance” for defining the variable that the ratings are supposed to estimate.

Scale type and range

Online ratings typically take the form of a five-star or ten-star discrete scale: this standard has been adopted by major players such as Amazon, Yelp, TripAdvisor and IMDb. However, these types of scales are likely not being able to measure well the quality and importance of scientific publications, because of the likely high skewness of the distribution of values of this target variable.

Let us consider the number of citations of scientific publications as a relevant proxy for the quality and importance of publications. About 44% of publications in Web of Science have zero citations, and the median number of citations is about 1, yet there is one paper having more than 305,000 citations and 148 papers having more than 10,000 citations (Van Noorden, Maher, & Nuzzo, 2014). In the case of patents, where the monetary value is defined by markets, the top 0.8% were valued at more than 1,000 times the median (Giuri et al., 2007). Let us assume that the main properties of these distributions generalize to the variable we want to measure, i.e. the maximum value can be of about 3 to 5 orders of magnitude larger than the median value. Therefore, a scale of 5, 10 or even 100 discrete categories cannot represent well this variability if the values that the scale represents vary linearly across categories. A logarithmic scale would be suitable, but it is psychologically difficult for most people to estimate values across so many orders of magnitude and to place them on a logarithmic scale.

A solution to this conundrum is asking experts to assess not the absolute value of the target variable, but its percentile rank. Then, the maximum value (100%) is represented by a number just 2 times larger than the median (50%), rather than several orders of magnitude larger. For usability and computational reasons, we limited the precision of the scale to 1%. Theoretically, this limits the capacity of indicating differences between top papers; in the case of the number of citations, in the top 1% the value varies from several hundreds to hundreds of thousands. In practice, test-retest reliability tends to decrease for scales with more than 10 response categories; users consider that a scale with 101 response categories allow them to best express their feelings adequately, but its ease and speed of use is slightly lower than of scales with 11 categories or less (Preston & Colman, 2000).

Because of the skewness of the distribution of absolute values, it is likely that experts are able to discriminate the percentile ranking of high quality papers better than the one of low quality papers. The confidence in rating papers also depends on how close the topic of the publication overlaps the expertise of the rater. For these reasons, raters should be able to express their uncertainty. Therefore, we allowed experts to give the rating as an interval of percentile rankings, rather than a single value. The rating is collected through a graphical interface representing the interval with sliding ends (Fig. 1). For ease of use on mobile devices, the interval can also be expressed using numerical selectors. A review may be associated to the rating, for explaining and supporting the rating.

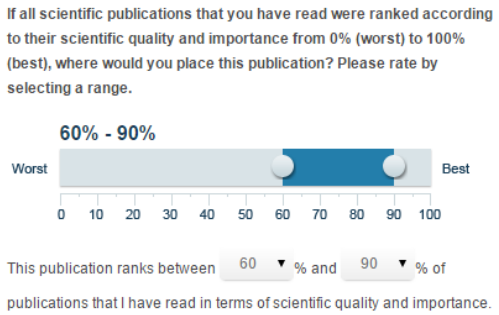


Figure 1. The Epistemio® rating scale for scientific publications.

Acknowledgments

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNDI–UEFISCDI, project number PN-II-PT-PCCA-2011-3.2-0895.

References

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? *Journal of Documentation*, *64*(1), 45-80.

Florian, R. V. (2012). Aggregating post-publication peer reviews and ratings. *Frontiers in Computational Neuroscience*, 6(31).

Giuri, P., Mariani, M., Brusoni, S., Crespi, G., Francoz, D., Gambardella, A., et al. (2007). Inventors and invention processes in Europe: Results from the PatVal-EU survey. *Research Policy*, *36*(8), 1107–1127.

Harnad, S. (2008). Validating research performance metrics against peer rankings. *Ethics in Science and Environmental Politics*, *8*, 103–107.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(2000), 1-15.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, *514*(7524), 550–553.

Analysis of the Factors Affecting Interdisciplinarity of Research in Library and Information Science

Chizuko Takei[1], Fuyuki Yoshikane[2] and Hiroshi Itsumura[3]

naoe.chizuko@ynu.ac.jp, fuyuki@slis.tsukuba.ac.jp, hits@slis.tsukuba.ac.jp

[1]University of Tsukuba, Graduate School of Library, Information and Media Studies, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

[2]University of Tsukuba, Faculty of Library, Information and Media Science, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

Introduction

In recent years, there has been a growing recognition of the necessity for interdisciplinary research that crosses disciplinary boundaries to deal with increasingly complex social issues (Rafols & Meyer, 2010). The relationship between the changes in interdisciplinarity of research over the years and researchers' attributions has rarely been investigated. Understanding the relationship between them will make it possible to gain useful information to foster interdisciplinary research, career-development of researchers, and development of research institutions. Thus, considering different periods, this study examines interdisciplinarity of research and the transdisciplinarity of researchers (targeted researchers themselves and their co-authors).

Methodology

This study targeted full-time faculty members of 2 iSchools, University of Pittsburgh (Pitt) and Syracuse University (SU), as of August 2014. The following data were employed: (1) information about targeted researchers and their co-authors, such as academic degrees or biographies, extracted from web pages; (2) bibliographic data of articles published by targeted researchers, which were extracted from Web of Science (WoS); (3) the title lists of WoS by subject categories acquired from the web site of Thomson Reuters; and (4) a matrix of the distance between categories of WoS, which was computed by Leydesdorff using Stirling's distance (http://www.leydesdorff.net/overlaytoolkit/stirling.htm). The procedure of this study was as follows: First, we examined transdisciplinarity of targeted researchers on the basis of the numbers of different disciplines where they had been engaged. We estimated their disciplines by several points of view such as belonging departments and academic degrees. As for their co-authors, though disciplines were estimated in the same way, we counted only disciplines that were different from those of the targeted researchers who had published the co-authored articles. Next, for each article of (2), by relating its reference list to (3) and (4), we computed indexes regarding interdisciplinarity that were used in later studies. This study applied the following indexes to the distribution of WoS categories assigned to the articles and their citing literature:

a. Total number of categories;
b. Simpson's Index (I);
c. Shannon's Index (entropy, H);
d. Distance between categories; and
e. The proportion of literature cited from different disciplines.

Indexes b and c evaluate the degree of diversity, taking into account both variety and equality in the frequency distribution. Index d indicates the distance between the categories of the articles and their citing literature. It ranges from −1 to 0, multiplying Stirling's distance by −1. As interdisciplinarity grows, their values become higher. Index e indicates the ratio of literature cited from different disciplines. Here, a different discipline is defined as a category with a distance over −0.7. Then, we performed a principal component analysis using these indexes and observed the correlation between the transdisciplinarity of targeted researchers or their co-authors and the interdisciplinarity of their articles along with its time-series variation. We discussed factors affecting the interdisciplinarity of research.

Results

Tendencies of indexes

Table 1 shows the basic statistics regarding transdisciplinarity of researchers and interdisciplinarity of their articles. We targeted 57 researchers, out of 73 faculty members, whose disciplines could be identified on the basis of information from university web sites and WoS.

The result of a principal component analysis for 5 indexes (C to G) revealed that the cumulative contribution rate of the first 2 principal components (PC1 and PC2) is 0.873. The characteristics of the 5 indexes can largely be explained by the first and second principal components. In Table 2, the principal component loading of PC1 suggests strong relationships between all 5 indexes. On the other hand, PC2 is characterized by large negative values of indexes F and G. Figure 1 is a plot of the first and second principal components and indicates that the 5 indexes can be divided into two groups (C, D, and E) and (F and G). It also implies that

highly interdisciplinary articles are remarkably diverse and rarely have common tendencies. In addition, we separated articles into two groups that were roughly equal in size (from 1981 to 2005 and from 2006 to 2014) to investigate the time-series variation related to the transdisciplinarity of researchers and the interdisciplinarity of research. The values of indexes concerning the interdisciplinarity of research (C to G) increased, while there were almost no changes in indexes concerning the transdisciplinarity of targeted researchers and their co-authors (A and B).

Table 1. Basic statistics regarding interdisciplinarity and transdisciplinarity.

		Pitt	SU	ALL
Targeted researchers/all faculties		23 / 30	34 / 43	57 / 73
Number of articles		267	259	526
Number of articles/targeted researchers	median	8	5	6
	range	1-33	1-31	1-33
A: Transdisciplinarity of targeted researchers	median	2	1	2
	range	1-2	1-3	1-3
B: Transdisciplinarity of co-authors	median	1	1	1
	range	0-6	0-4	0-6
C: Total number of categories	median	13	15	14
	range	1-79	1-59	1-79
D: Simpson's Index	median	0.781	0.767	0.777
	range	0-0.949	0-0.934	0-0.949
E: Shannon's Index	median	2.383	2.383	2.383
	range	0-4.385	0-4.061	0-4.385
F: Distance between categories	median	−0.438	−0.413	−0.424
	range	−1–−0.005	−1–−0.013	−1–−0.005
G: Proportion of literature cited from different disciplines	median	79%	79%	79%
	range	0%-100%	0%-100%	0%-100%

Table 2. Principal component loading for 5 indexes.

	PC1	PC2	PC3	PC4	PC5
C	−0.648	0.536	−0.540	0.002	−0.032
D	−0.876	0.301	0.345	0.037	−0.148
E	−0.898	0.350	0.202	−0.051	0.168
F	−0.717	−0.652	−0.089	−0.229	−0.031
G	−0.750	−0.610	−0.093	0.236	0.029

The relationship between transdisciplinarity of researchers and interdisciplinarity of their research

We computed Spearman's rank correlation coefficient for indexes A to G to survey the relationship between transdisciplinarity of researchers (A and B) and interdisciplinarity of their research (C to G) (Table 3). No strong correlation was found between them. However, comparing index A with B, we observed stronger and significant correlation between index B and the indexes concerning interdisciplinarity of research (C to G). In addition, we compared the articles before 2005 with those after 2006 to examine the time-series variation of correlation between indexes. Although there was no distinguished distinction between them, the degree of correlation tended to become stronger and the number of significant coefficients was increased for indexes A and B.

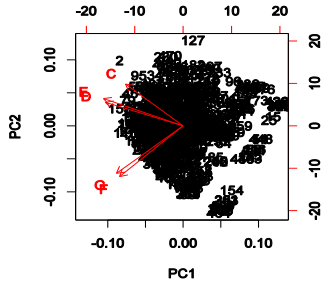


Figure 1. Plot of the first and second principal components.

Table 3. Rank correlation ρ among 7 indexes for all articles.

	A	B	C	D	E	F	G
A	1	0.23*	0.12*	0.17*	0.18*	0.05	0.06
B		1	0.21*	0.20*	0.21*	0.07	0.14*
C			1	0.69*	0.76*	0.17*	0.16*
D				1	0.99*	0.37*	0.30*
E					1	0.37*	0.30*
F						1	0.88*
G							1

*Significant ($p < 0.05$)

Discussion and Conclusions

This study computed indexes for interdisciplinarity of research in library and information science and performed principal component analysis to clarify the relationship among the indexes. The results indicate that the indexes considering the distance between subject categories of WoS have characteristics very different from the indexes considering only the number of categories and their frequency distributions. This suggests that we should consider a more multidimensional approach. Furthermore, we investigated changes over time in the indexes of interdisciplinarity, and observed the progress for interdisciplinarity of research in library and information science. As the results of the correlation analysis between interdisciplinarity of research and transdisciplinarity of researchers, stronger and significant correlations were seen with the transdisciplinarity of co-authors than with that of the targeted researchers themselves. This suggests that interdisciplinarity of research might be more affected by the transdisciplinarity of co-authors than by that of the researchers themselves. We will conduct further investigations with more samples.

Reference

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, *82*, 263-287.

An Analysis of Scientific Publications from Serbia: The Case of Computer Science

Miloš Pavković[1] and Jelica Protić[2]

[1]*milos_pakovic@yahoo.com* [2]*jelica.protic@etf.bg.ac.rs*
University of Belgrade, School of Electrical Engineering, Department of Computer Engineering and Informatics
Bulevar kralja Aleksandra 73, 11000 Belgrade (Serbia)

Introduction

In Serbia, like in other countries all over the world, career opportunities in computing are growing faster than most of the other professions. This trend should be in accordance with the growth of the number of study programs and consequently the number of teaching staff. The most important researchers' and university teaching staff's promotion criteria, according to the regulations in Serbia, are the papers published in journals from the JCR list, which is, for the area of computing, reduced to the SCIe list. The number of such papers is also relevant for projects financed by the Ministry of Education, Science and Technological Deve- lopment of the Republic of Serbia.

In this paper, we present an analysis of the references of Serbian researchers retrieved from the Web of Science. Using the bibliometric indicators from the Web of Science, we also examine the distribution of such references across WoS categories that belong to the broader area of computing. We show the distribution of such publications over the years, cities and universities and identify the relations with global trends in Serbian science.

Data Set

Data used in this paper were taken from Thompson Reuters Web of Science on 29 September 2014, selecting Science Citation Index Expanded (SCIe) journal articles. A basic search was conducted using the keyword ‘‘Serbia’’ in the field address and the retrieved results were limited to articles published during the period 2006–2013. All document information, including names of authors, titles, years of publications, source journals, contact addresses, and number of citations for each article, for every year, were downloaded into Microsoft Excel worksheets. The custom program in C# programming language was developed in order to perform data analysis.

The same data extraction was performed for WoS categories, that we considered the subcategories of the broader scientific area of Computer Science. The distribution of the number of papers from the year 2006 till 2013 (since results for 2014 were incomplete) is presented on Figure 1, and the number of papers over years and WoS categories is presented on Figure 2.

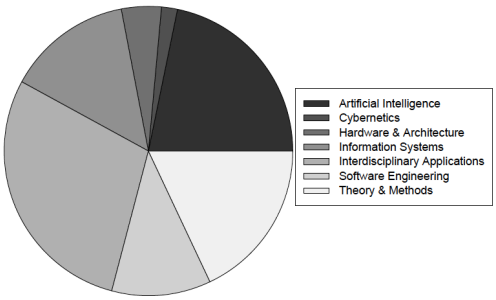


Figure 1. The number of papers in subcategories.

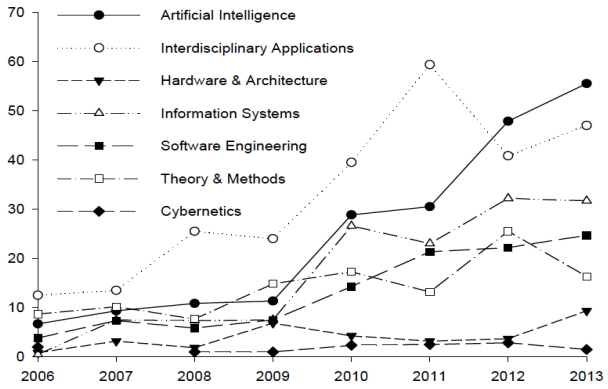


Figure 2. The number of papers in subcategories for each year.

To get numbers presented in Figure 2, disciplinary affiliation is computed fractionally, by assigning 1/N to each category, for a journal paper published in a journal indexed in N different categories.

The name of the country was not always correct for papers submitted before 2006, since our country changed its name to Serbia in 2006, and some papers had the former name Serbia and Montenegro, or even Yugoslavia in their affiliation. Therefore, the additional search was performed using only the names of significant Serbian cities and university centres. It was noticed that our dataset did not hold absolutely correct

information, because of unintentional mistakes in the authors' signatures or other elements of the affiliation. Incorrectly entered data propagate errors to later identification and grouping, as stated in Mitrovic (2014). This issue can be solved partially using text similarity matching algorithms. Our program uses Jaro-Winkler algorithm as proposed in Winkler (1995), also known as JWSF, "Jarod-Winkler similarity function" to overcome this problem.

Distribution of papers over major cities and institutions show interesting results. For the Serbian capital city of Belgrade, only 65.4% of all papers have affiliation of the state University of Belgrade, the biggest and oldest Serbian university, ranked between positions 300 and 400 on the ARWU list. For other university centres in Serbia, the share of publications of state universities is: 93.3% for Novi Sad, 87.4% for Niš and 97.9% for Kragujevac. We conclude that bigger cities have greater potential for scientific productivity outside the university, but this ratio also reflects some problems identified in the past, that institutes belonging to the University of Belgrade did not include the name of the University in affiliation before the initiative to do so, started during the procedure and efforts to qualify for ARWU ranking. The significant growth in the number of papers started in 2008, probably as the result of accreditation procedure regulated by national accreditation body CAQA (www.kapk.org).

Table 1. Journals with more than 20 papers published in the period from 2006 till 2013.

Journal Name	No.	5 years IF
MATCH-communications in mathematical and in computer chemistry	101	1.829
ComSIS - Computer science and information systems	67	0.575
Mathematical and computer modelling	47	2.020
Expert systems with applications	42	1.965
Advances in electrical and computer engineering	28	0.642
Fuzzy sets and systems	27	1.880
International journal of computers communications & control	23	0.694
Information sciences	21	3.893
Journal of multiple-valued logic and soft computing	21	0.667

The list of journals with more than 20 papers in the Table 1 shows that journals in multiple WoS categories are predominant. The journal MATCH publishes the mathematical results and applications in solving chemical problems, without significant content in computing research. The second journal on the list, ComSIS (Computer Science and Information Systems), is an international journal published in Serbia, dedicated to computing, that appeared for the first time on the SCIe list in 2010. In fractional counting, it has been shown that some other disciplines are represented in the comparable quantity to the basic computer science disciplines: Engineering, Electrical & Electronic (71.65), Mathematics, Applied (62.75) and Chemistry, Multidisciplinary (41.67) are in-between Computer Science, Theory & Methods (76.98) and Computer Science, Hardware and Architecture (22.83). Since the leading category is Computer Science, Interdisciplinary Applications (153.00), it is obvious that computer science in Serbia can be viewed predominantly as applied science, blended with electrical engineering, applied mathematics and multidisciplinary chemistry. The leading scientists are I. Gutman with 74 papers in total and 26 in fractional counting, and M. Ivanovic with 23 papers in total and 6.33 in fractional counting.

Conclusions

Considerable growth of publications from Serbia since 2006 was identified in Ivanovic (2014). Serbian national system that transfers data from WoS on weekly bases kobson.nb.rs shows that there were 1746 publications of Serbian authors during 2006 and the yearly production tripled in 2013. At the same time, the number of all publications in Computer Science categories in WoS core collection increased from 123 to 286, while articles only increased from 60 to 204, which was about 3.9% of total Serbian production and 0.47% of the world production in aforementioned categories in the year 2013. The ratio of total world production and total Serbian production is 0.39%, so the results of computer science disciplines are better than average, mostly due to the interdisciplinary approach.

Acknowledgments

We are grateful to Ms. Biljana Kosanović and Ms. Darija Dašić for their assistance in data retrieval, and for valuable advice.

References

Ivanovic, D., Ho, J.-S., (2014). Independent publications from Serbia in the Science Citation Index Expanded: a bibliometric analysis. *Scientometrics*, *101*(1), 603-622.

Winkler, W.E. (1995). Matching and record linkage. In Cox, B. (Ed.), *Business Survey Methods*, Wiley, London, pp. 355-84.

Mitrovic, I., Protic, J., (2014) Problems with affiliations, names and personal identity in the process of evaluating higher education institutions, *EDULEARN14 Proceedings,* Barcelona, 2524-2533.

SCIENCE POLICY AND RESEARCH ASSESSMENT

UNIVERSITY POLICY AND INSTITUTIONAL RANKINGS

SCIENTIFIC FRAUD AND DISHONESTY

A Computer System for Automatic Evaluation of Researchers' Performance

Ashkan Ebadi[1] and Andrea Schiffauerova[2]

[1] *a_ebad@encs.concordia.ca,* [2] *andrea@ciise.concordia.ca*
Concordia Institute for Information Systems Engineering (CIISE), Concordia University,
1515 Ste-Catherine Street West, Montreal, Quebec H3G 2W1 (Canada)

Abstract

The increasing number of researchers and the limited financial resources has caused a tight competition among scientists to secure research funding. On the other side, it has become even harder for funding allocation organizations to evaluate the performance of researchers and select the best candidates. However, it seems that the current evaluation methods are highly correlated with subjective criteria. In addition, the subjective nature of peer-review as one the most common methods in scientific evaluation calls itself for an accurate complementary quantitative method to help the decision makers. This paper proposes an automatic computer system, which is based on machine learning techniques for predicting the performance of researchers. The proposed system uses various features of different types as the input to a complex machine learning module to predict the performance of a researcher in a given year. The method provides the decision makers with fair comparative results regardless of any subjective criteria. Our results show the high accuracy of the proposed system in predicting the performance of researchers.

Conference Topic

Methods and techniques, Science policy and research assessment

Introduction

Research grants is known as one of the crucial drivers of scientific activities that can influence the size and efficiency of R&D sector and its productivity (Jacob & Lefgren, 2011). It can also affect the performance of researchers through providing them with a better access to the research resources (Lee & Bozeman, 2005). In the meantime, policies on R&D activities have evolved over the past fifty years (Elzinga & Jamison, 1995; Sanz-Menendez & Borras, 2000). Funding agencies put a lot of efforts on selecting the best candidates for allocating grants as well as on evaluating the performance of researchers in regards to the amount of funding that they have been receiving. On the other hand, the growing number of researchers worldwide has made the competition for securing the limited financial resources even harder. For example, according to Polster (2007) the contest for receiving research funding is on the rise in Canada especially among the academic researchers mainly due to the changes in federal funding policies, lack of university operating budgets, and increasing research costs. The researchers' demand for funding cannot be fully satisfied by the finite financial capacity of the funding agencies. However, the case could be even worse for the young researchers since the senior researchers are more known within their scientific community that might help them in getting money for research.

Peer review is the oldest measure that has been being used for evaluating researchers' performance and their proposals. Most of the funding agencies use a committee of independent researchers to review the researchers' proposals for funding and select the most appropriate researcher(s) through a competitive process. However, the peer review process has been widely criticized in the literature due to the potential biases since the accuracy of the procedure is highly dependent on the selected experts. For example, preferences of peers can affect the final decision or it can act as a gatekeeper for new research interests since peers may not come into an integrated conclusion (King, 1987). Despite the aforesaid drawbacks, the great advantage of peer review process is that the impact of the proposed research could

be assessed quite easily and accurately (Allen et al., 2009). For this important reason it has still remained as one of the most popular techniques in scientific evaluation. Though, the current trend is to combine the expert review with quantitative performance indicators (Butler, 2005; Hicks et al., 2004) in order to achieve a more balanced evaluation since it cannot be reliable enough as a single indicator. For this purpose, citation and publication counts based indicators are commonly used as the quantitative indicators of researchers' performance.

One of the reasons that scientists publish their work in the form of scientific papers is that in this way they can secure their priority in discoveries (De Bellis, 2009). According to the review of literature done by Tan (1986), performance evaluation of individual researchers and research departments are in most cases based on publication counts measures (at least partially). For the quality of publications, citation counts based indicators, first introduced by Gross and Gross in 1927, are commonly accepted as a proxy for the impact of a scientific publication (Gingras, 1996). In general, they count the number of citations received by an article after the date it is published; hence, papers with higher number of citations are assumed to have higher impact.

Invention of the Internet and availability of the digital data have made it feasible to extract and collect data in a very large scale. In addition, the rapid advancement in the field of computer science has made new ideas and algorithms available to the data scientists. Therefore, large scale digital data and complex algorithms provide researchers with novel opportunities to explore new directions of the information science as well as scientific evaluation. This paper presents an integrated highly accurate automatic productivity prediction system that can assist decision makers (and peers) to detect the most appropriate researchers for funding allocation. The remainder of the paper proceeds as follows: *Data and Methodology* section describes the data gathering procedure in detail while explaining the methods and methodologies that were used; the *Results* section presents the performance evaluation results and interpretations for the proposed system; the paper concludes in *Discussion* section; and limitations and future research directions are stated in the last section of the paper.

Data and Methodology

Data

We decided to focus on performance of the researchers who have been funded by the Natural Sciences and Engineering Research Council (NSERC)[1] of Canada. The main reasons for choosing NSERC was its role as the main federal funding organization in Canada, and the fact that almost all the Canadian researchers in natural sciences and engineering receive at least a basic research grant from NSERC (Godin, 2003). Therefore, as the first stage information about the funded researchers was collected from NSERC[2]. In the next phase, Elsevier's Scopus[3] was used to gather all the information about the funded researchers. The data spans from information about the authors themselves (*e.g.* Scopus ID, their affiliation, number of publications in a given year, *etc.*) to their articles (*e.g.* year of publication, authors of the paper, keywords, *etc.*).

The time interval of the research was set to the period of 1996 to 2010 since the data coverage of Scopus was better after 1996. Moreover, to have a proxy of the quality of the papers we

[1] For more information, see: http://www.nserc-crsng.gc.ca/index_eng.asp

[2] Students were excluded from the data as the goal of the paper is evaluating the performance of researchers.

[3] Scopus is a commercial database of scientific articles that has been launched by Elsevier in 2004. It is now one of the main competitors of Thomson Reuter's Web of Science.

used SCImago[4] to collect the impact factor information of the journals in which the articles were published. SCImago was chosen for two main reasons. Firstly, it provides annual data of the journal impact factors that enables us to perform a more accurate analysis since we are considering the impact factor of the journal in the year that an article was published not its impact in the current year. Secondly, SCImago is powered by Scopus that makes it more compatible with our publications database.

In the next phase of data preparation, we calculated several bibliometric features such as amount of funding received by a researcher in a given year, his/her career age, average number of co-authors, average number of publications, average number of citations, *etc.* In addition, using Pajek[5] software social network analysis techniques were employed to construct the collaboration networks of the researchers within the examined time interval. The created networks were used to calculate various network structure properties (*e.g.* betweenness centrality, eigenvector centrality, and clustering coefficient) of the researchers at the individual level. All the calculated features were integrated in a MySQL[6] dataset. The final database contains 117,942 records of researchers. In the next section, methodologies are discussed in more detail.

Methodology

Several features of various types and from different sources were selected for this study. Funding is acknowledged in the literature as one of the main drivers of scientific activities where a three-year (*e.g.* Payne & Siow, 2003) or a five-year (*e.g.* Jacob & Lefgren, 2007) time window is mostly considered for the funding to take effect. In this paper a three-year time window was considered for all the bibliometric variables, *e.g.* for assessing the productivity of a given researcher in year 1999 his/her amount of funding was summed up for the period of 1996 to 1998 (*sumFund3*). Intuitively, productive researchers are expected to at least maintain their performance level. Various past productivity features were hence included in the model reflecting the quality and quantity of the publications. As a proxy for the rate of publications, number of publications in a three-year time window (*noArt3*) was considered. Two indicators were used as proxies for the quality of publications, *i.e.* average number of citations in a three year time window (*avgCit3*) and the average impact factor of the journals in which the articles were published in a three year time interval (*avgIf3*). Both of the mentioned features can serve as a proxy for quality, but with a slightly different meaning. Impact factor indicates the respectability of the journal, *i.e.* the quality and the level of contribution perceived by the authors and the reviewers of the paper, whereas citation counts show the impact of the article on the scientific community and on the subsequent research.

A multi-level feature representing the scientific field of the researcher (*discip*) was also used in the model since publication and citation habits can be different in various scientific fields. For example, citing habits and the rate of citations may vary across different scientific fields in a way that in some scientific fields authors publish articles more frequently or the published papers contain more references (MacRoberts & MacRoberts, 1996; Phelan, 1999). It is argued in the literature that older researchers in general can be more productive (Merton, 1973; Kyvik & Olsen, 2008) due to several reasons (*e.g.* better access to the funding and expertise sources, more established collaboration network, better access to modern equipments). Hence, the career age of the researcher (*careerAge*) was included in the model representing the time difference between the date of his/her first article in the database and the given year. As a common indicator of the scientific collaboration, the average number of co-authors per paper was also included in the prediction model (*teamSize*). It is expected that

[4] For more information, see: http://www.scimagojr.com

[5] Social network analysis software, for more information see: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[6] Open source relational database management system, for more information see: http://www.mysql.com/

researchers who have on average higher number of co-authors have more connections that might result in relatively higher number of projects or future publications, hence this feature was also considered as one of the influencing factors.

As discussed in the previous section, social network analysis was used to construct the collaboration networks and to measure the structural network properties of researchers. In particular, four network structure indicators were calculated namely betweenness centrality (*bc*), clustering coefficient (*cc*), eigenvector centrality (*ec*), and degree centrality (*dc*). Betweenness Centrality (*bc*) is an indicator of the important players (researchers) in a network who have a control over the flow of knowledge and resources. These players, who are also called as *gatekeepers*, are able to bridge different communities. Theoretically, betweenness centrality of the node *k* is measured based on the share of times that a node *i* reaches a node *j* via the shortest path passing from node *k* (Borgatti, 2005) and is calculated as follows (σ_{ij} is the total number of shortest paths from node *i* to j and $\sigma_{ij}(k)$ is the number of shortest paths from node *i* to node *j* that contains node *k*):

$$bc_k = \sum_{i \neq k \neq j} \frac{\sigma_{ij}(k)}{\sigma_{ij}} \qquad (1)$$

Clustering Coefficient (*cc*), also called *cliquishness*, indicates the tendency of researchers to cluster with other researchers in the network. Hence, researchers with high clustering coefficient may have a relatively high number of connections with the other team members who are collaborating in a tightly knit group. Therefore, this indicator was selected to represent the tight collaboration impact on the overall performance of the team. Theoretically, clustering coefficient of node *i* (cc_i) is defined based on the number of triangles (inter-connected sub-network of three nodes) that contains the node *i* (t_i) normalized by the maximum number of triangles in the given network (Watts & Strogatz, 1998). Let n_i denotes number of neighbors of the node *i*, hence:

$$cc_i = \frac{2t_i}{n_i(n_i - 1)} \qquad (2)$$

Degree Centrality (*dc*) that was also considered as one of the network variables is defined based on the number of ties that a node has (degree) in an undirected graph. Hence, researchers with high degree centrality should be more active since they have higher number of ties (links) to other researchers (Wasserman, 1994). Moreover, in co-authorship networks it can be regarded as the number of direct partners or team members of a given researcher. Hence, it is expected to have an influence on the scientific activities. Degree centrality for node *i* (dc_i) is thus defined based on the node's degree (deg_i) and then the values are normalized between 0 and 1 (dividing by the highest degree in the network) to be able to compare the centralities:

$$dc_i = \frac{deg_i}{deg_{high}} \qquad (3)$$

Eigenvector Centrality (*ec*) takes the importance of a node and its connections into the account. Hence, a researcher has high eigenvector centrality if he/she is connected with other important actors who are themselves occupying central positions in the network. These researchers can be identified as *leaders* in the scientific networks since they are connected

with too many other influential and highly central researchers, and it is hence expected that they shape the collaborations and play an important role in setting priorities in scientific projects that might affect the performance of researchers. A complete list of the selected features is shown in Table 1.

Table 1. List of attributes for the prediction models.[7]

No	*Attribute*
1	Scientific area in which the researcher is working (*discip*)
2	Total amount of funding received by each researcher in a 3 year time window (*sumFund3*)
3	Total number of publications of each researcher in a 3 year time window (*noArt3*)
4	Average number of citations received by researcher's articles in a 3 year time window (*avgCit3*)
5	Average impact factor of the journals in which researcher's articles were published in a 3 year time window (*avgIf3*)
6	Average betweenness centrality of each researcher in a 3 year time window (*btwn3*)
7	Average degree centrality for each researcher in a 3 year time window (*deg3*)
8	Average clustering coefficient of each researcher in a 3 year time window (*clust3*)
9	Average eigenvector centrality of each researcher in a 3 year time window (*eigen3*)
10	Average number of authors per paper for each researcher (*teamSize*)
11	Career age of the researcher (*careerAge*)

The mentioned features were used as an input to the prediction model. Figure 1 shows the whole process of the researchers' performance prediction. Number of publications was considered as the target variable for the performance prediction task. As it can be seen, data is first preprocessed and cleaned. For this purpose, several JAVA programs were coded to check the data for redundancy, out of range values, impossible combinations, errors, and missing values and then data was filtered based on the records that contained all the required data. The resulted data containing all the mentioned features was fed into the data preparation block where at first all the features were normalized to a value between 0 and 1. This was a crucial step since the features were of different units and scales. Local Outlier Factor (LOF) algorithm was then implemented to detect the outliers. LOF that was proposed by Breunig et al. (2000) is based on the local density concept in which the local deviation of a given data is measured with respect to its k nearest neighbors. A given data is outlier if it has a substantial different density from its k neighbors. The final step of the data preparation step was optimizing the attributes' weights. For this purpose we used an evolutionary attributes weights optimizer that employed genetic algorithm to calculate the weights of the attributes. The weighting procedure improved the accuracy of the system by giving more value to the most influential attributes. The resulted data was integrated into a single data repository named as the target data.

[7] The initial list of the selected features was prepared as a result of an intensive statistical analyses performed on the target data. The list was then refined and weighted within the proposed system.

After making the data ready for the analysis, a stratified 10-fold cross validation design was used for the model validation. Cross validation is an analytics tool that is used to design and develop fine tune models. In other words, the data is split into two disjoint sets where one part is used for training and fitting a model (training set) while the other part is employed for estimating the error of the model (test set) (Weiss & Kulikowski, 1991). We used a nested 10-fold cross validation in which the data is split into 10 disjoint subsets in a way that union of the 10 folds results the original data. The method runs 10 times and in each time one fold is considered as the test data while the rest are regarded as the training data.

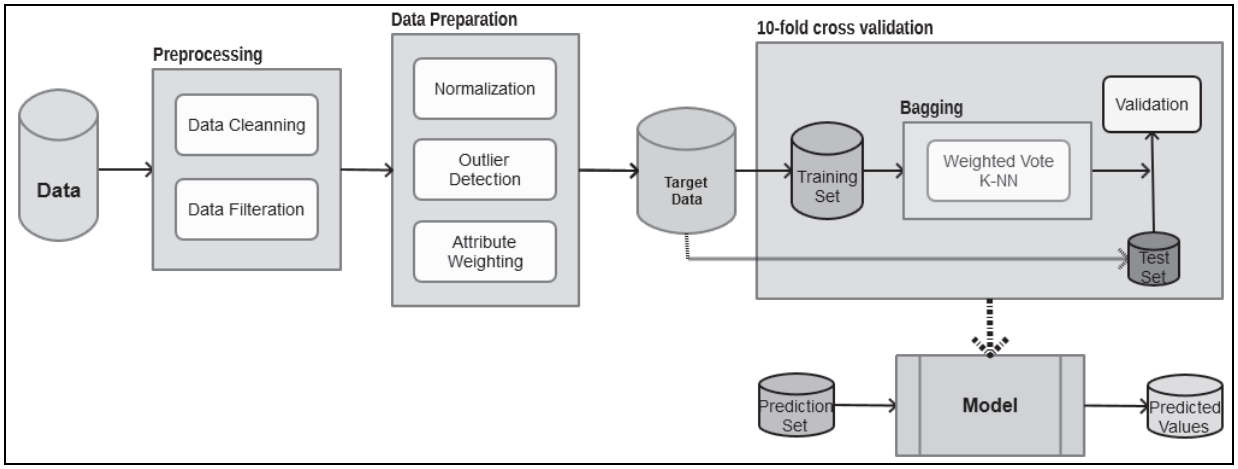


Figure 1. Proposed model for automatic evaluation of researchers' performance.

As mentioned earlier, number of publications was considered as the target variable. To further improve the accuracy of the prediction the ensemble meta-algorithm was employed. For this purpose, bootstrap aggregating (bagging) approach was used. Bagging is an ensemble method that makes random subsets of the data and trains them separately where the final result is obtained by averaging over the results of the separated models (Breiman, 1996). Bagging is a nested module in which we used weighted vote 10-Nearest Neighbor (10-NN) algorithm to train the data and to create the model. In weighted vote 10-NN the distance of the neighbors to the given data is considered as a weight in the prediction in a way that neighbors that are closer to the given data get higher weights. This particularly helped to increase the accuracy of the prediction. Data in the range of 1996 to 2009 was used to train and build the model while a separate disjoint data for 2010 (prediction set) was used for testing the accuracy of the prediction model. The final output of the proposed automatic computer system was the predicted number of publications for the researchers in the prediction set.

Results

In this section the results of the performance evaluation of the proposed automatic computer system (PACS) is presented. As discussed earlier, the model was trained on the data from 1996 to 2009 and a disjoint dataset for 2010 was used for the prediction and the accuracy tests. The accuracy of the proposed model was compared with several well-known machine learning algorithms, however, in this paper the results are presented and compared for the PACS model as well as two other algorithms that showed the highest accuracy in predicting the target variable.

Figure 2 shows the prediction errors of PACS, linear regression, and polynomial regression of degree three[8]. We considered three error measures for comparing the performance of the

[8] Other algorithms (*e.g.* decision trees) were also tested but these listed algorithms were the top two ones with the highest accuracy.

mentioned algorithms. Root mean squared error is one of the main measures for comparing the accuracy of the prediction models and is defined as the square root of the average of the squares of errors. According to Figure 2, PACS is predicating the number of publications of researchers with 1.451 average deviation between the predicted value and the real number of publications. Normalized absolute error is the absolute error (difference between the predicted value and the real value) divided by the error made if the average would have been predicted. The root relative squared error takes the average of the actual values as a simple predictor to calculate the total squared error. The result is then normalized by dividing it by the total squared error of the simple predictor and square root is taken to transform it to the same dimension as the predicted value. As it can be seen PACS is performing better in all the three measures where the degree 3 polynomial fit is the worst.

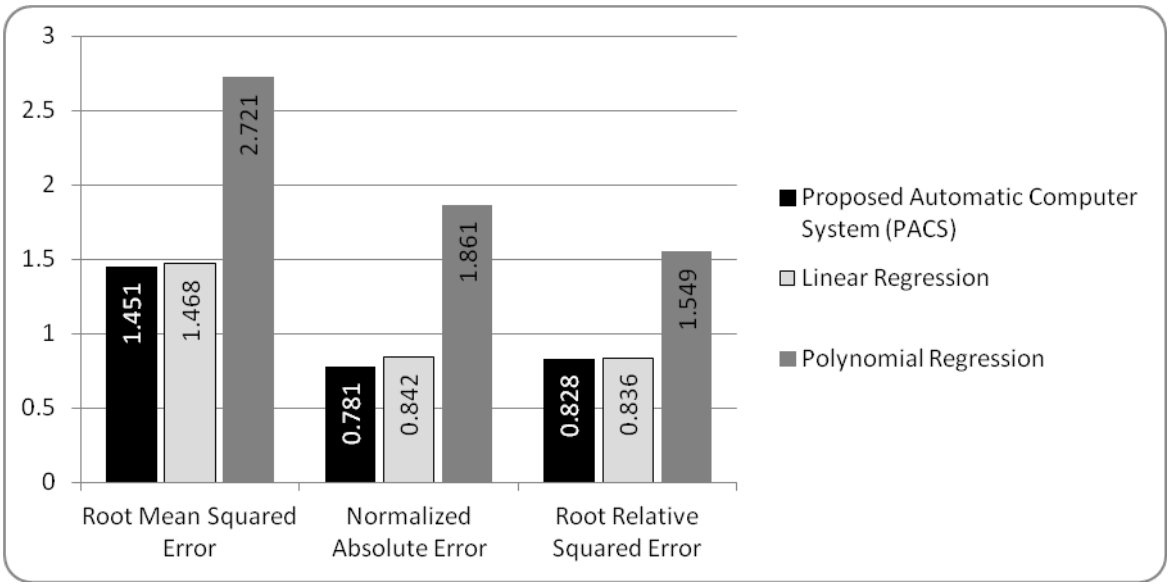


Figure 2. Accuracy test, PACS vs. other two top performing algorithms.

Table 2. Prediction results.

No	*Predicted no of articles*	*noArt*	*sum Fund3*	*avg If3*	*avg Cit3*	*teamSize*	*btwn3*	*clust3*	*deg3*	*eigen3*	*careerAge*	*discip*	*noArt3*
1	0.361	0	0.041	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.737	2	0
2	1.102	0	0.013	0.279	0.028	0.000	0.000	1.000	0.005	0.000	0.632	3	1
3	3.865	7	0.044	0.054	0.005	0.001	0.059	0.125	0.027	0.000	0.737	1	13
4	1.103	0	0.010	0.068	0.083	0.000	0.000	1.000	0.007	0.000	0.737	3	1
5	1.206	1	0.072	0.132	0.020	0.002	0.016	0.409	0.020	0.000	0.526	0	6
6	6.703	4	0.167	0.246	0.080	0.002	0.055	0.158	0.039	0.000	0.737	1	26
7	1.030	4	0.032	0.115	0.017	0.001	0.018	0.455	0.018	0.000	0.737	0	6
8	4.120	3	0.061	0.136	0.041	0.002	0.185	0.109	0.134	0.000	0.737	1	15
9	0.000	0	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.263	0	0
10	5.047	3	0.137	0.141	0.041	0.001	0.133	0.163	0.050	0.000	0.684	0	15
11	1.128	1	0.010	0.091	0.062	0.003	0.003	0.333	0.007	0.000	0.526	1	1
12	1.964	1	0.010	0.113	0.009	0.004	0.053	0.192	0.022	0.018	0.737	1	7
13	12.228	7	0.095	0.399	0.028	0.010	0.197	0.042	0.075	0.000	0.684	0	31
14	2.112	2	0.190	0.228	0.091	0.001	0.011	0.182	0.020	0.000	0.737	1	6
15	2.233	3	0.299	0.230	0.051	0.002	0.013	0.457	0.035	0.000	0.737	0	7
16	3.577	4	0.198	0.259	0.055	0.002	0.042	0.145	0.059	0.000	0.579	4	12
17	11.308	9	0.329	0.309	0.116	0.002	1.000	0.062	0.148	0.000	0.737	1	40
18	4.841	4	0.093	0.458	0.051	0.001	0.027	0.117	0.037	0.000	0.737	0	19
19	5.752	4	0.116	0.253	0.055	0.123	0.003	0.823	0.940	1.000	0.737	1	20
20	7.421	8	0.193	0.270	0.077	0.002	0.153	0.079	0.082	0.000	0.737	1	26

A randomly selected sample of the predictions is presented in Table 2. Each row represents a distinct researcher's profile in 2010 for whom several indicators have been calculated and used in the PACS model as the input features. The real number of articles is shown in *noArt* column that was not fed into the prediction model. Based on the other attributes the proposed system automatically predicted the number of publications of a researcher in 2010, *i.e.* column named *Predicted no of articles* in Table 2 and is highlighted in dark grey. As it can be seen using several features of different types and employing various techniques for data gathering (*e.g.* bibliometrics, social network analysis) and preparation provides the system with highly accurate high-dimensional input data that led to a low error rate and good predictions. Interestingly, it seems that the system successfully considered the differences between various scientific fields in performing scientific activities. According to the results, although the profile of the researchers numbered 1 and 9 in Table 2 are relatively similar, the predicted performance differs as they do not belong to the same scientific field. Hence, the results confirm the importance of the scientific disciplines in predicting the performance of researchers. In addition, comparison of the researchers numbered 6 and 7 highlights the importance of the past productivity as well as the quality of publications in predicting the number of publications.

Discussion

In this paper we used various bibliometric as well as network structural property features to build a model to predict the performance of researchers. Machine learning techniques and availability of the digital data has made it possible to use complex algorithms on high dimensional large scale data. This provides scientometrists with an opportunity to go beyond the current border of using common indicators or simple statistical analyses. Although some researchers recently worked on citation prediction using machine learning algorithms (*e.g.* Fu & Aliferis, 2010; Lokker et al., 2008) to our knowledge this is the first study that focused on the prediction of researchers' productivity using input features of different types and at the individual level of the researchers.

The attribute weighting method to rank features based on their importance that was implemented in the proposed model as well as the outlier detection module for data filtration increased the accuracy of the predictions significantly. Results of the attribute weighting module can also shed light on the most influential attributes in predicting the scientific activities of the target researchers. Another unique approach that was employed in designing the proposed system was using several features of similar nature in building the model that reinforced the prediction power of the system. For example, average number of citations and average impact factor of the journals were used to represent the quality of the paper. Another example is the degree centrality and scientific team size as the former represents the number of direct connections of a researcher while the latter indicates the average number of his/her co-authors. These attributes of similar nature surely empowered the accuracy of the model by providing it with more dimension and flexibility.

To conclude, as it was observed complex computer algorithms can be used to design automatic evaluation systems and prediction tools to evaluate different aspects of scientific activities of researchers. It is obvious that peer reviewing cannot be completely replaced by such tools. However, such systems can help decision makers in setting both long-run and short-term strategies in regard to the funding allocation and/or analyzing researchers' productivity. In addition, the availability of high-dimensional large scale data (in our case, a large dataset spanning from 1996 to 2010) that is intensively cleaned and preprocessed for learning the model will surely contribute to highly accurate predictions that are not based on a limited criteria or a limited feature set. Therefore, this can also help to establish a fairer funding allocation or scientific evaluation system.

Limitations and Future Work

We were exposed to some limitations in this paper. Firstly, Scopus was selected for gathering information about the funded researchers' articles. Since Scopus and other similar databases are English biased, hence, non-English articles are underrepresented (Okubo, 1997). Secondly, due to the better coverage of Scopus before 1996, the time interval of 1996 to 2010 was selected for the analysis. Although Scopus is confirmed in the literature to have a good coverage of articles, as a future work it would be recommended to focus on other similar databases to compare the results.

Furthermore, we were exposed to some limitations in measuring scientific collaboration among the researchers where we used the network structure properties. In particular, we were unable to capture other links that might exist among the researchers like informal relationships since these types of connections are never recorded and thus cannot be quantified. In addition, there are also some drawbacks in using co-authorship as an indicator of scientific collaboration since collaboration does not necessarily result in a joint article (Tijssen, 2004). An example could be the case when two scientists cooperate together on a research project and then decide to publish their results separately (Katz & Martin, 1997). For assessing the quality of the papers based on citation counts we did not account for self citations, negative citations, or special inter-citation patterns among a number of researchers. Although we also used another proxy (average impact factor of journals) to overcome this limitation, it can be addressed in the future works.

References

Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: The role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLoS One*, *4*(6), e5910.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55-71.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140.

Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM Sigmod Record*, *29*(2), pp. 93-104.

Butler, L. (2005). What happens when funding is linked to publication counts? Handbook of quantitative science and technology research (pp. 389-405), Springer.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Scarecrow Press.

Elzinga, A. & Jamison, A. (1995). *Changing policy agendas in science and technology. Handbook of Science and Technology Studies* Ed.by Sheila Jasanoff et al. London: Sage.

Fu, L. D. & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*(1), 257-270.

Gingras, Y. (1996). Bibliometric analysis of funded research. A feasibility study. *Report to the Program Evaluation Committee of NSERC.*

Godin, B. (2003). The impact of research grants on the productivity and quality of scientific research. No. 2003. *INRS Working Paper.*

Gross, P., & Gross, E. (1927). College libraries and chemical education *Science, 66*(1713), 385-389.

Hicks, D., Tomizawa, H., Saitoh, Y., & Kobayashi, S. (2004). Bibliometric techniques in the evaluation of federally funded research in the United States. *Research Evaluation*, *13*(2), 76-86.

Jacob, B. A. & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics*, *95*(9), 1168-1177.

Katz, J. S. & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1-18.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, *13*(5), 261-276.

Kyvik, S. & Olsen, T. B. (2008). Does the aging of tenured academic staff affect the research performance of universities? *Scientometrics*, *76*(3), 439-455.

Lee, S. & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, *35*(5), 673-702.

Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: Retrospective cohort study. *BMJ* (Clinical Research Ed.), *336*(7645), 655-657.

MacRoberts, M. H. & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*(3), 435-444.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Okubo, Y. (1997). Bibliometric indicators and analysis of research systems: Methods and examples. No. 1997/1. *OECD Publishing*.

Payne, A. A. & Siow, A. (2003). Does federal research funding increase university research output? *Advances in Economic Analysis & Policy*, *3*(1).

Phelan, T. (1999). A compendium of issues for citation analysis. *Scientometrics*, *45*(1), 117-136.

Polster, C. (2007). The nature and implications of the growing importance of research grants to Canadian universities and academics. *Higher Education*, *53*(5), 599-622.

Sanz Menéndez, L., & Borrás, S. (2000). Explaining changes and continuity in EU technology policy: The politics of ideas.

Tan, D. L. (1986). The assessment of quality in higher education: A critical review of the literature and research. *Research in Higher Education*, *24*(3), 223-265.

Tijssen, R. J. (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.

Wasserman, S. (1994). Social network analysis: Methods and applications. *Cambridge university press*.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440-442.

Weiss, S., & Kulikowski, C. (1991). *Computer Systems that Learn*: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. San Francisco: Morgan Kaufmann.

Grading Countries/Territories Using DEA Frontiers

Guo-liang Yang[1], Per Ahlgren[2], Li-ying Yang[3], Ronald Rousseau[4], Jie-lan Ding[3]

[1]*glyang@casipm.ac.cn*
Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190 (China)

[2]*perahl@kth.se*
School of Education and Communication in Engineering Sciences (ECE), KTH Royal Institute of Technology, 100 44 Stockholm (Sweden)

[3]*yangly@mail.las.ac.cn, dingjielan@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing 100190 (China)

[4]*Ronald.Rousseau@kuleuven.be*
Institute for Education and Information Sciences, IBW, University of Antwerp (UA), Antwerp B-2000 (Belgium)
KU Leuven, Department of Mathematics, Leuven B-3000 (Belgium)

Abstract

Several approaches exist related to categorizing academic journals/institutions/countries into different levels. Most existing grading methods use either a weighted sum of quantitative indicators (including the case of one properly defined quantitative indicator) or quantified peer review results. An important issue of concern for science and technology management is the efficiency of resource utilization. In this paper we deal with this issue and use multi-level frontiers of data envelopment analysis (DEA) models to grade countries/territories. Research funding and numbers of researchers as used as inputs, while papers and citations are output variables. The research results show that using DEA frontiers we can grade countries/territories on six levels. These levels reflect the corresponding countries' level of efficiency in S&T resource utilization. Furthermore, we use papers and citations as single outputs (with research funding and researchers as inputs) to show changes in country/territory level.

Conference Topic

Science Policy and Research Assessment

Introduction

The efficiency of science and technology (S&T) resource utilization is one of the important issues for S&T management (Yang et al., 2013a; Yang et al., 2014a). Johnes and Johnes (1992) evaluated the efficiency of S&T organizations using data envelopment analysis (DEA) as a performance analysis tool. Rousseau and Rousseau (1997, 1998) assessed the efficiency of countries using gross domestic product, active population and research and development (R&D) expenditure as inputs, and publications and patents as outputs. They showed that DEA can be used in scientometrics as a tool to measure the efficiency of decision making units (DMUs, e.g., countries) by gauging closeness to the efficiency frontier. Similar techniques have been used by other researchers (Kao & Lin, 1999; Roy & Nagpaul, 2001; Shim & Kantor, 1998). Yang and Chang (2009) used DEA under constant and variable returns to scale (RTS) to measure firms' efficiency. Worthington (2001) conducted an empirical survey of frontier efficiency measurement techniques in education. Other researchers have analyzed the efficiency or productivity in the education sector, (e.g., Abbott & Doucouliagos, 2003, Avkiran, 2001, Carrington et al., 2005, Worthington & Lee, 2008, Flegg et al., 2004, Johnes & Johnes, 1995, Johnes, 2006a,b, Kempkes & Pohl, 2010, Wolszczak-Derlacz & Parteka, 2011, and Aristovnik, 2012). When studying the standard university model, Brandt and Schubert (2013) observed that universities are large agglomerations of many (often loosely

affiliated) small research groups. They explained this observation by typical features of the scientific production process. In particular, they argued that there are decreasing RTS on the level of the individual research groups. RTS is a concept with strong relation to scale efficiency. Somewhat similar observations (decreasing RTS) were published earlier by Bonaccorsi and Daraio (2005). Schubert (2014) used non-parametric techniques of multidimensional efficiency measurement, such as DEA, to analyse the RTS in scientific production based on survey data for German research groups from three scientific fields. Based on DEA models, Yang et al. (2013a, 2014a) analyzed the directional RTS of a couple of biological institutes in the Chinese Academy of Sciences (CAS).

Some fairly recent studies have examined the efficiency of countries or regions in utilizing R&D expenditures or other resources. Lee and Park (2005) evaluated R&D efficiency across nations using patents, technology balance of receipts and journal articles as outputs. Wang and Huang (2007) analyzed R&D efficiency of nations by considering patents and papers as outputs. Lee et al. (2009) used DEA to measure and compare the performance of national R&D programs in South Korea. Sharma and Thomas (2008) investigated the R&D efficiency of developing countries in relation to developed countries, taking into account time lags. Other, and similar, studies include Chen et al. (2011), Sueyoshi and Goto (2013), and Zhong et al. (2011).

The literature referred to hitherto focuses on the quantitative measurement of efficiency of resource utilization. In this context, DEA is one of the most popular mathematical tools for estimating the relative efficiency of DMUs. However, Banker (1993) pointed out that DEA efficiency scores usually overestimate efficiency and are biased. Smith (1997) argued that the extent of the overestimation is highly dependent on sample size and the complexity of the production process (as indicated by the numbers of inputs and outputs). However, in many cases we only need to know the general level (grade) of DMUs in terms of efficiency instead of their exact scores or complete ranking.

Several efforts have been made regarding categorization of academic journals, institutions and countries into different levels of standing or quality. Since 2007, the Association of Business School (ABS) has issued the Academic Journal Quality Guide, which classifies journals in business and management into four categories (grade 1 to 4) recognizing the quality of those journals based on a survey of hundreds of experts in the field (Harvey et al., 2007a,b; 2008). From 2010, a new category, termed 4*, was added to the four existing categories to recognize the quality of the top journals (Harvey et al., 2010). Bandyopadhyay (2013) categorized business and management journals into four categories (Excellent, Very Good, Standard, Satisfactory) based on multiple inputs, including Thomson Reuters' Social Science Citation lists of ranked journals and WoS impact factor analyses. In 2005, CAS evaluated its dependent institutes and classified them into three grades (Excellent, Good, and Satisfactory) (CAS, 2006). Glänzel (2011) used characteristic scores and scales as parameter-free tools to identify top journals. Yang et al. (2013b) analyzed the overall development and the balance of the disciplinary structure of China's science based on papers covered by Science Citation Index and with the use of bibliometric methods. These authors further categorized selected countries to reflect their developmental status.

The grading methods in the research reported above use either a weighted sum of quantitative indicators (including the case of one properly defined quantitative indicator) or quantified peer review results. In general, the weighted sum approach normally needs indicator weights and corresponding threshold values as a priori information, while the peer review process usually costs a lot of time and expenditures (Smith, 1996). In the light of these downsides, this paper presents an alternative approach, involving multiple DEA frontiers, to divide various countries/territories into different levels with respect to the efficiency of their S&T resource utilization.

The rest of the paper is organized as follows. The next section introduces the input and output indicators, and the corresponding dataset used in the analysis. The used methods are described in the third section, in which we treat multi-level efficient frontiers and show how to divide the countries/territories into grades using these frontiers. In the fourth section, the results of the study are given, whereas conclusions appear in the final section.

Indicators and data

In this work, research funding and researchers are used as input indicators. Research funding here means Gross Domestic Expenditure on R&D (million current PPP$). The total number of researchers (full time equivalents, FTEs) in one country is used as indicator for researchers. For the output indicators, we used the number of papers covered by the Science Citation Index (SCI) and Social Science Citation Index (SSCI) from the Web of Science (WoS), and the number of citations to these papers in the year 2011. We use OECD statistics and Thomson Reuters' research evaluation tool InCites as sources for input and output data, respectively. All 34 OECD member countries and seven non-OECD member countries/territories were selected for the study. The other non-OECD member countries, covered by OECD statistics, were excluded due to lack of input data. This also holds for the two OECD members Australia and Switzerland (the Gross Domestic Expenditure in 2011 on R&D of these two countries is missing), and thereby the number of OECD member countries included in the study is 32. See Table 1 for details.

Methods

DEA models and their frontiers

DEA is an approach based on linear programming for analyzing performance of organizations and operational processes. This approach was first proposed by Charnes et al. (1978). All DEA models use input and output data to evaluate the relative efficiency of DMUs without prior knowledge of input/output functions and the weights for indicators. Nowadays, numerous theoretical and empirical works on this method have been published, extending the original approach in different ways, and applying them to many areas, including the private and the public sector (e.g., Cooper et al., 2007).

Let $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_s)$ be input and output vectors of n DMUs, respectively of m and s dimensions. Then the Production Possibility Set (*PPS*) is defined by

$$PPS = \{(X, Y): X \; can \; produce \; Y\} \quad (1)$$

There can be different forms of PPS based on different assumptions. Banker (1984) defined the PPS under the assumption of variable RTS to obtain the BCC-DEA model:

$$PPS(X,Y) = \left\{(X,Y) | X \geq \sum_{j=1}^{n} \lambda_j X_j \, , Y \leq \sum_{j=1}^{n} \lambda_j Y_j \, , \sum_{j=1}^{n} \lambda_j = 1 \, , \lambda_j \geq 0, j = 1, \dots, n\right\} \quad (2)$$

where λ_j is a coefficient.

The *PPS* implied in the CCR-DEA model, which was proposed by Charnes et al. (1978) under the assumption of constant RTS, is defined as follows:

$$PPS(X,Y) = \left\{(X,Y) | X \geq \sum_{j=1}^{n} \lambda_j X_j \, , Y \leq \sum_{j=1}^{n} \lambda_j Y_j \, , \lambda_j \geq 0, j = 1, \dots, n\right\} \quad (3)$$

The boundary of the *PPS* is referred to as the production technology or production frontier.

Table 1. Values of input and output indicators across 39 countries/territories.

No.	*Countries/Territories*	*Output*		*Input*	
		Papers	*Citations*	*Research Funding (PPP)*	*Researcher (FTE)*
1	Argentina	8136	40201	4592.313295	50340
2	Austria	12843	100412	9971.246479	37113.8
3	Belgium	18876	152731	9739.425206	42685.77
4	Canada	59025	427079	24756.76203	157360
5	Chile	5795	31737	1172.833167	6082.9
6	China	162794	846720	247808.3033	1318086
7	Czech Republic	9866	55662	4659.446488	30681.59
8	Denmark	13608	124330	6934.707773	37944.1
9	Estonia	1509	10731	733.5776566	4511
10	Finland	10761	82802	7897.729287	40002.61
11	France	67407	480151	53310.69922	249086.3
12	Germany	95935	738284	96971.46462	338608
13	Greece	10819	62818	2006.921474	24674.25
14	Hungary	5934	36137	2721.690282	23019
15	Iceland	815	9013	317.6389104	2258.3
16	Ireland	7438	57682	3169.659323	15172
17	Israel	12478	88753	9306.312467	49797
18	Italy	55338	385416	25780.80141	106151.3
19	Japan	77453	429710	148389.2294	656651
20	Luxembourg	678	4480	660.3865084	3031
21	Mexico	10490	46668	8058.470588	46124.96
22	Netherlands	33845	302477	14597.91748	58447.26
23	New Zealand	8181	50974	1766.588573	16300
24	Norway	10825	78889	5064.393225	27228
25	Poland	21057	91097	6409.165974	64132.8
26	Portugal	10789	66489	4152.692178	50061.2
27	Romania	6927	24373	1725.931612	16080
28	Russia	29072	85915	35192.07719	447579
29	Singapore	9950	82648	6922.39777	33718.5
30	Slovakia	3083	13861	921.2876157	15325.9
31	Slovenia	3776	17682	1429.743722	8774
32	South Africa	9477	48450	4652.174133	20115.06
33	South Korea	45588	222201	58379.65416	288901
34	Spain	50677	332172	20106.98571	130234.9
35	Sweden	21568	172220	13366.28061	48589
36	Taiwan	27283	129286	26184.28683	134047.7
37	Turkey	23920	72981	11301.84442	72108.6
38	UK	100895	784071	39217.4483	251357.6
39	USA	364548	2774572	429143	1252948

Data sources: Input: OECD Statistics. http://www.oecd-ilibrary.org/statistics; Output: InCites. http://incites.isiknowledge.com/Home.action.

Definition 1: The efficient frontier of *PPS* is defined as follows:

$$EF = \{(X,Y) \in PPS | there\ is\ no\ (\bar{X},\bar{Y}) \in PPS\ such\ that\ (-\bar{X},\bar{Y}) > (-X,Y)\} \quad (4)$$

Note: This unobservable production frontier is called the true efficient frontier hereinafter. When there is only a single output, the production frontier is known in the economic literature as the production function. DMUs, which are technically efficient, operate on the frontier, while technically inefficient DMUs operate at points in the interior of the *PPS*. Thus it is rational to rank DMUs according to their distance to the true frontier.

The core idea of classic DEA is to identify first the production frontier. DMUs on the frontier are regarded as efficient. DMUs not situated on the frontier are compared with their peers or projections on the frontier to measure their relative efficiency. All DMUs on the frontier are considered to represent the best practices and have the same level of performance.

Let $\{(x_j, y_j)|j = 1, \dots, n\}$ be a group of observed input and output data. Based on such observations, DEA models construct a piecewise linear production frontier, a non-parametric estimate of the unobservable true frontier. Then DEA models measure the efficiency of a DMU via its distance to the estimated frontier. Using radial measurement and input orientation, we have the following input-based CCR-DEA model (Charnes et al., 1978):

$$\theta_c^* = min\,\theta$$
$$s.t.\begin{cases} \sum_{j=1}^{n} x_{ij}\lambda_j \le \theta x_{i0}, i = 1, \dots, m, \\ \sum_{j=1}^{n} y_{rj}\lambda_j \ge y_{r0}, r = 1, \dots, s, \\ \lambda_j \ge 0, j = 1, \dots, n. \end{cases} \quad (5)$$

where $\lambda_j \ge 0$ are the multipliers of inputs and outputs. Here θ_c^* measures the degree of efficiency by radial measurement under the assumption of constant RTS.

If we assume that the production technology satisfies the variable returns to scale assumption, we have the following input-based BCC-DEA model (Banker et al., 1984):

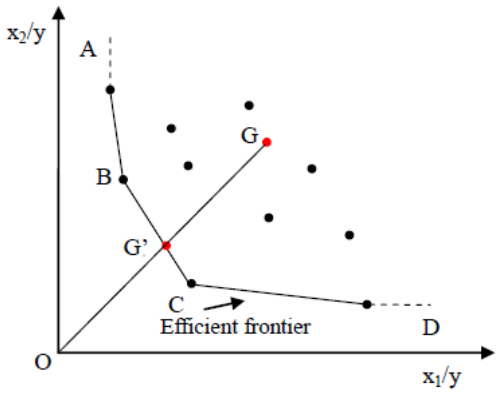
$$\theta_b^* = min\,\theta$$
$$s.t.\begin{cases} \sum_{j=1}^{n} x_{ij}\lambda_j \le \theta x_{i0}, i = 1, \dots, m, \\ \sum_{j=1}^{n} y_{rj}\lambda_j \ge y_{r0}, r = 1, \dots, s, \\ \sum_{j=1}^{n} \lambda_j = 1, \\ \lambda_j \ge 0, j = 1, \dots, n. \end{cases} \quad (6)$$

where θ_b^* measures the degree of efficiency by radial measurement under the assumption of variable returns to scale. It should be noted that Model (6) differs from Model (5) only regarding the constraint $\sum_{j=1}^{n} \lambda_j = 1$, which yields that the variable RTS assumption is satisfied.

Obviously, if $\theta_c^* = 1$ in model (5) or $\theta_b^* = 1$ in Model (6), then the DMU is situated on the efficient frontier in CCR-DEA or BCC-DEA, respectively.

We visualize the frontier of a DEA model in Figure 1, using two inputs (x_1 and x_2) and one output (y). The piecewise linear line ABCD defines the efficient frontier of the existing observations. For example, for point G, representing a DMU, its efficiency score can be calculated as the ratio of distance OG' to distance OG.

We now give an example to illustrate the detection of the efficient frontier and the evaluation of DMUs using a DEA model. We suppose there are six DMUs with two inputs and a single output. In Table 2, hypothetical data is given.



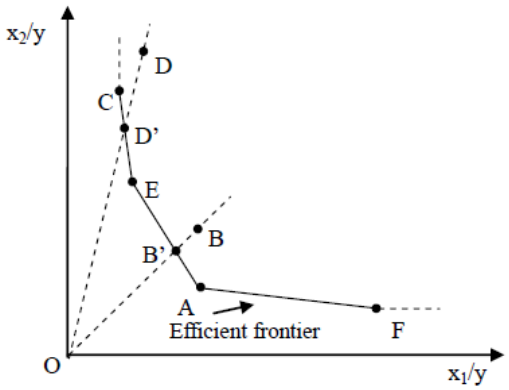


Figure 1. Efficient Frontier of a DEA model. Figure 2. Efficient Frontier and DMUs.

First, for comparison, we expand the inputs and output of each DMU proportionally and let the output of each DMU be 120 (Table 3).

Table 2. 6 DMUs with 2 inputs and a single output.

DMUs	DMU_1	DMU_2	DMU_3	DMU_4	DMU_5	DMU_6
Output (y)	120	8	24	40	120	24
Input 1 (x_1)	19	1	1	2	10	8
Input 2 (x_2)	10	1	6	15	17	1

We show these six DMUs in Figure 2 (which gives projections in input space) using points A-F to denote DMU_1-DMU_6.

Table 3. Expanded DMUs with 2 inputs and single output.

DMUs	DMU_1	DMU_2	DMU_3	DMU_4	DMU_5	DMU_6
Output(y)	120	120	120	120	120	120
Input 1(x_1)	19	15	5	6	10	40
Input 2(x_2)	10	15	30	45	17	5

We use a piecewise linear curve to link points C, E, A, F and merge it with the horizontal and vertical lines from point F and C, respectively, to obtain the piecewise linear convex hull, which is the efficient frontier produced from this DEA model. Points C, E, A, F are on the efficient frontier and their efficiencies are all unity. On the contrary, points B and D are inside the convex hull, so these two DMUs are inefficient compared with their peers or projections (points B' and D') on the efficient frontier. Taking point B as example, the DEA model uses the ratio of distance OB' to the distance OB to measure point Bs relative efficiency.

Decomposition of countries/territories based on multi-level frontiers in DEA

In the preceding section, we showed how the effective frontier can be detected. If we remove the efficient DMUs on the frontier, we can use the DEA model again to obtain a new frontier. We do this repeatedly in order to decompose DMUs into different levels. This process is illustrated in Figure 3. In this figure, the first tier of the efficient frontier is the piecewise line ABCD (Efficient frontier – tier1), on which the DMUs with the best level of efficiency are located. After we remove the DMUs on the Efficient frontier – tier1, we rerun the DEA model, obtaining the DMUs on the efficient frontier – tier2 as the second group, and so on.

This process is iterated until there is no DMU left, and the grading of the DMUs ends. The efficient frontier in Figure 1 is the same as the efficient frontier– tier1 in Figure 3.

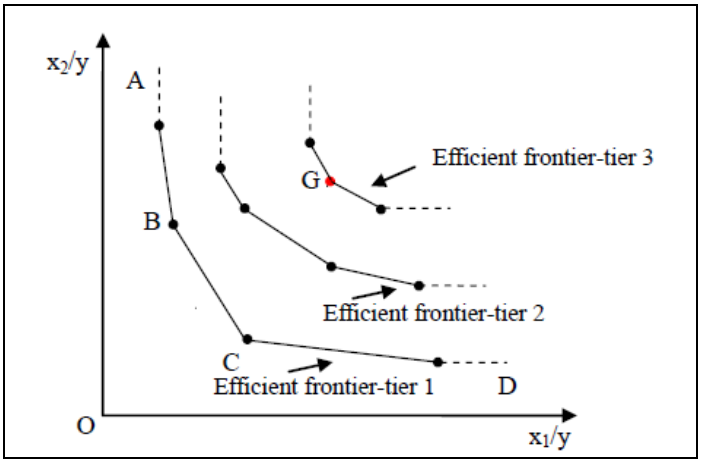


Figure 3. Multi-level efficient frontiers of a DEA model.

In earlier works, DEA frontiers have been used either to measure the relative efficiency of the DMUs (e.g., Charnes et al., 1978; Cook and Seiford, 2009) by comparing them with their peers or projections on the frontier, or to estimate the RTS by the frontier's shape (Banker et al., 2004). To the best of our knowledge, no research similar to the research reported in this paper has used multi-level frontiers in DEA models to decompose DMUs into different grades to reflect different levels of performance.

In the process of decomposing the DMUs into different grades, we need to ensure that a given DMU can only be assigned to one level to avoid conflicts. An efficient frontier is a convex hull. This implies that if a point belongs to F_k it cannot belong to any other F_{k+l} (if it exists, where l is a positive integer). Indeed a point on the frontier is a convex linear combination of efficient points on the frontier. If point P would belong to F_k and F_{k+l} this would mean that P is a convex linear combination of points that do not belong to F_k, which is not possible. Thus, one country/territory can only be assigned to one level.

Results

The BCC-DEA model was applied to produce multi-level efficient frontiers, and these were used to decompose the countries/territories of the study into different grades. Table 4 reports the levels of the countries/territories for the three experiments: two inputs & two outputs, two inputs & the first output (papers), and two inputs & the second output (citations).

We first consider the case of two inputs and two outputs. The results show that Chile, Greece, Iceland, Italy, Netherlands, UK and USA are the first level countries in the sense of efficiency of S&T resource utilization (Table 4). Mexico is the least efficient unit among the 39 countries/territories and belongs to the last level (Tier 6).

We reused the multi-level efficient frontiers in the BCC-DEA model on the 39 countries/territories with two inputs and the first output (papers) to decompose the countries/territories into different grades. We can see that now Chile, Greece, Iceland, Italy, Netherlands, UK and USA are the most efficient countries/territories (Table 4). Mexico, Finland, Israel and Singapore have with the lowest efficiencies.

We also used the multi-level efficient frontiers in the BCC-DEA model on the 39 countries/territories with two inputs and the second output (Citations), which is shown in table 4. Also in this case Chile, Greece, Iceland, Netherlands, UK and USA are first level countries, and Italy has moved into Tier 2. The latter means that Italy performs better for papers than for citations. Mexico and Turkey are in the last tier, Tier 7. It is interesting to see

that Turkey is in Tier 3 in the case of two inputs and two outputs while in Tier 7 in the case of two inputs and the second output, which means that the citation performance of Turkey is considerably worse than its performance for papers.

Table 4. Levels of the countries/territories.

No.	*Countries /Territories*	*two inputs & two outputs*	*two inputs & first output(paper)*	*two inputs & second output(citation)*
1	Chile	1	1	1
2	Greece	1	1	1
3	Iceland	1	1	1
4	Netherlands	1	1	1
5	UK	1	1	1
6	USA	1	1	1
7	Italy	1	1	2
8	Canada	2	2	2
9	China	2	2	2
10	Estonia	2	2	2
11	Germany	2	2	2
12	Luxembourg	2	2	2
13	New Zealand	2	2	2
14	Spain	2	2	2
15	Belgium	2	2	3
16	Slovakia	2	2	3
17	Sweden	2	2	3
18	Poland	2	2	4
19	Ireland	2	3	2
20	Denmark	2	4	3
21	France	3	3	3
22	Slovenia	3	3	3
23	Japan	3	3	4
24	Romania	3	3	4
25	South Africa	3	3	4
26	Turkey	3	3	7
27	Norway	3	4	4
28	Portugal	3	4	4
29	Austria	3	5	4
30	South Korea	4	4	4
31	Hungary	4	4	5
32	Taiwan	4	4	5
33	Czech Republic	4	5	6
34	Israel	4	6	5
35	Singapore	4	6	5
36	Argentina	5	5	6
37	Russia	5	5	6
38	Finland	5	7	6
39	Mexico	6	8	7

Figure 4 corresponds to Table 4 and visualizes the levels of the countries/territories when using two inputs and two outputs, two inputs and the first output (paper), and two inputs and the second output (citation). From this figure, it is clear that some countries/territories (e.g.,

Argentina, Belgium, Czech Republic, Turkey) belong to a lower level in the case of two inputs & the second output (citations) compared to the case of two inputs & the first output (papers), which indicates that these countries perform more efficient for papers than for citations. Inversely, some countries (e.g., Austria, Denmark, Finland) perform more efficient for citations than for papers.

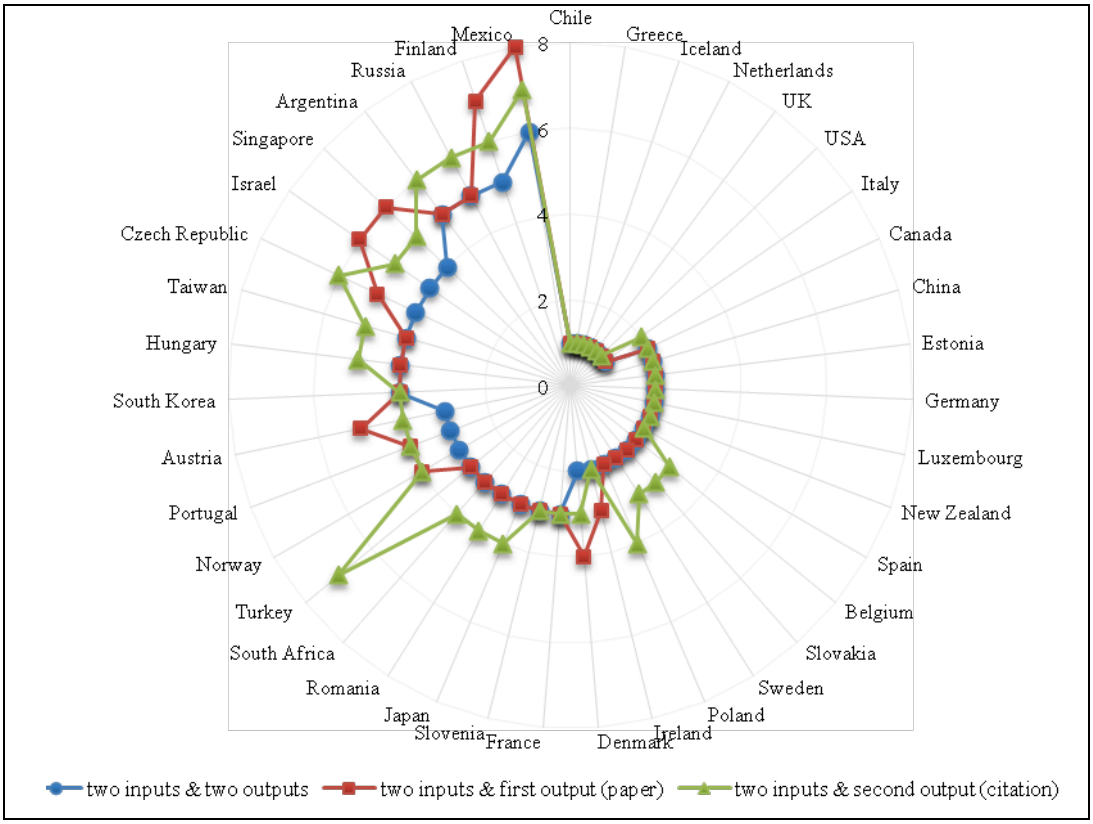


Figure 4. Visualisation of the levels of the countries/territories.

It is surprising that Greece and Chile are rated first level countries together with S&T-developed countries like USA and UK. For papers as output, we can verify this result using the ratios Papers to Researcher and Papers to Research Funding. From Table 5, we can see that Greece and Chile perform very well for these two ratios. On the contrary, we can see China, Japan and South Korea have low performance compared to other countries. We believe that a reason for this is that researchers from these countries publish relatively frequently in domestic journals that are not covered by WoS. We do not tabulate the values of the corresponding two ratios for citations, but it turned out that Chile and Greece perform well also with respect to these ratios.

Discussion and conclusions

In this paper we have shown that multi-level frontiers of DEA can be used to decompose countries/territories into different levels, reflecting the efficiency of S&T resource utilization of the countries/territories. The approach put forward is not restricted to the grading of countries/territories. It can also be used to grade, for instance, journals and research institutions based on properly selected indicators. In case of no explicit inputs, e.g., when journals should be graded, we can assume that there is single constant input, which is equal to unity for all observations (e.g., Yang et al. 2014b).

There are two main advantages of the grading approach proposed in this paper. First, it is a nonparametric and recursive approach, which needs no a priori information such as indicator

weights and threshold values for different grading levels. Second, the observations within the same level are indifferent in the sense of efficiency of resource utilization. The main disadvantage of the approach is that in some cases there are too few indicators (single input and single output). Under such circumstances, it might be the case that each level includes exactly one observation (in our case, exactly one DMU). Thus, the approach is more suitable for grading observations with multiple input and output indicators.

For future research, we would like to investigate the multiple DEA frontiers regarding weight restrictions in DEA models. There are at least four types of restrictions on the weights of input and output variables (e.g., Allen et al., 1997), and the efficient frontiers will vary accordingly and show different properties. Furthermore, this grading approach can be easily extended to the classification of scientific journals, research institutions, etc.

Table 5. Ratios of Papers to Researcher and Research Funding.

No.	*Countries/Territories*	*Papers/Researcher*	*Papers/Research Funding*	*No.*	*Countries/Territories*	*Papers/Researcher*	*Papers/Research Funding*
1	Argentina	0.1616	1.7717	21	Mexico	0.2274	1.3017
2	Austria	0.3460	1.2880	22	Netherlands	0.5791	2.3185
3	Belgium	0.4422	1.9381	23	New Zealand	0.5019	4.6310
4	Canada	0.3751	2.3842	24	Norway	0.3976	2.1375
5	Chile	0.9527	4.9410	25	Poland	0.3283	3.2855
6	China	0.1235	0.6569	26	Portugal	0.2155	2.5981
7	Czech Republic	0.3216	2.1174	27	Romania	0.4308	4.0135
8	Denmark	0.3586	1.9623	28	Russia	0.0650	0.8261
9	Estonia	0.3345	2.0570	29	Singapore	0.2951	1.4374
10	Finland	0.2690	1.3625	30	Slovakia	0.2012	3.3464
11	France	0.2706	1.2644	31	Slovenia	0.4304	2.6410
12	Germany	0.2833	0.9893	32	South Africa	0.4711	2.0371
13	Greece	0.4385	5.3908	33	South Korea	0.1578	0.7809
14	Hungary	0.2578	2.1803	34	Spain	0.3891	2.5204
15	Iceland	0.3609	2.5658	35	Sweden	0.4439	1.6136
16	Ireland	0.4902	2.3466	36	Taiwan	0.2035	1.0420
17	Israel	0.2506	1.3408	37	Turkey	0.3317	2.1165
18	Italy	0.5213	2.1465	38	UK	0.4014	2.5727
19	Japan	0.1180	0.5220	39	USA	0.2910	0.8495
20	Luxembourg	0.2237	1.0267				

Acknowledgements. We would like to acknowledge the support of the National Natural Science Foundation of China (NSFC, No.71201158).

References

Abbott, M. & Doucouliagos, C. (2003). The efficiency of Australian universities: a data envelopment analysis. *Economic of Education Review*, *22*(1), 89-97.

Allen, R., Athanassopoulos, A., Dyson, R.G., & Thanassoulis, E. (1997). Weights restrictions and value judgements in data envelopment analysis: Evolution, development and future directions. *Annals of Operations Research*, *73*, 13-34.

Aristovnik, A. (2012). The relative efficiency of education and R&D expenditures in the new EU member states. *Journal of Business Economics and Management*, *13*(5), 832-848.

Avkiran, N.K. (2001). Investigating technical and scale efficiencies of Australian universities through data envelopment analysis. *Socio-Economic Planning Sciences*, *35*(1), 57-80.

Bandyopadhyay, A. (2013). *Ranking of Business School Journals: A Rating Guide for Researchers.* Retrieved August 23, 2014 from: http://mpra.ub.uni-muenchen.de/49608/1/MPRA_paper_49608.pdf.

Banker, R.D. (1993) Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. *Management Science*, *39*(10), 1265–1273.

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.

Banker, R.D., Cooper, W.W., Seiford, L.M., Thrall, R.M., & Zhu, J. (2004). Returns to scale in different DEA models. *European Journal of Operational Research*, *154*, 345-362.

Bonaccorsi, A., & Daraio, C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics*, *63*(1), 87-120.

Brandt, T., & Schubert, T. (2013). Is the university model an organizational necessity? Scale and agglomeration effects in science. *Scientometrics*, *94*(2), 541-565.

Carrington, R., Coelli, T., & Rao, P.D.S. (2005). The performance of Australian universities: Conceptual issues and preliminary results. *Economic Papers-Economic Society of Australia*, *24*(2), 145-163.

CAS (2006). *Report on Comprehensive Quality Evaluation of Institutes in Chinese Academy of Sciences (CAS)*. Retrieved August 23, 2014 from http://www.bps.cas.cn/kjpj/xgzl/200905/t20090527_232560.html.

Charnes, A., Cooper, W.W., & Rhodes, E.L. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429-444.

Chen, C.P., Hu, J.L., & Yang, C.H. (2011). R&D efficiency of multiple innovative outputs: The role of the national innovation system. *Innovation: Management, policy & practice*, *13*, 341-360.

Cook, W.D., & Seiford, L.M. (2009). Data Envelopment Analysis (DEA)-Thirty years on. *European Journal of Operational Research*, *192*, 1-17.

Cooper, W.W., Seiford, L.M., & Tone, K. (2007). *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software (Second Edition)*. New York: Springer.

Flegg, A.T., Allen, D.O., Field, K., & Thurlow, T.W. (2004). Measuring the Efficiency of British Universities: A Multi-Period Data Envelopment Analysis. *Education Economics*, *12*(3), 231-249.

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.

Harvey, C., Kelly, A., Morris, H., & Rowlinson, M. (2010). *Academic Journal Quality Guide—Version 4*. London: Association of Business Schools.

Harvey, C., Morris, H., & Kelly, A. (2007a). *Academic Journal Quality Guide: Context, Purpose and Methodology*.London: Association of Business Schools.

Harvey, C., Morris, H., & Kelly, A. (2007b). *Academic Journal Quality Guide*. London: Association of Business Schools.

Harvey, C., Morris, H., & Kelly, A. (2008). *Academic Journal Quality Guide Version 2: Context, Purpose and Methodology*. London: Association of Business Schools.

Johnes, G., & Johnes, J. (1992). Apples and oranges: The aggregation problem in publications analysis. *Scientometrics*, *25*(2), 353-365.

Kao, C., & Lin, Y.C. (1999). Comparing university libraries of different university size. *Libri*, *49*(3), 150-158.

Kempkes, G., & Loikkanen, H.A. (1998). The efficiency of German universities: Some evidence from nonparametric and parametric methods. *Applied Economics*, *42*, 2063-2079.

Lee, H.Y., & Park, Y.T. (2005). An international comparison of R&D efficiency: DEA approach. *Asian Journal of Technology Innovation*, 13(2), 207-222.

Lee, H.Y., Park, Y.T., & Choi, H. (2009). Comparative evaluation of performance of national R&D programs with heterogeneous objectives: A DEA approach. *European Journal of Operational Research*, *196*(3), 847-855.

REIST-2 (1997). *European Commission, Second European Report on S&T Indicators.* (EUR 17639). Brussels: Luxembourg.

Rousseau, S. & Rousseau, R. (1997). Data envelopment analysis as a tool for constructing scientometric indicators. *Scientometrics, 40*(1), 45-56.

Rousseau, S., & Rousseau, R. (1998). The scientific wealth of European nations: taking effectiveness into account. *Scientometrics, 42*(1), 75-87.

Roy, S., & Nagpaul, P.S. (2001). A quantitative evaluation of relative efficiencies of research and development laboratories: A data envelopment analysis approach. In: (M. Davis & C.S. Wilson. Eds.) *Proceedings of the 8th International Conference on Scientometrics & Informetrics* (pp. 629-638). Sydney (Australia): BIRG.

Sharma, S., & Thomas, V.J. (2008). Inter-country R&D efficiency analysis: an application of data envelopment analysis. *Scientometrics*, *76*(3), 483-501.

Shim, W., & Kantor, P.B. (1998). A novel economic approach to the evaluation of academic research libraries. *Proceedings of the ASIS Annual Meeting*, *35*, 400-410.

Smith, R. (1996). Peer review: A flawed process in the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4), 178-182.

Smith, P. (1997). Model misspecification in data envelopment analysis. *Annals of Operations Research*, *73*, 233-252.

Schubert, T. (2014). Are there scale economies in scientific production? On the topic of locally increasing returns to scale. *Scientometrics*, (to appear), DOI 10.1007/s11192-013-1207-1.

Sueyoshi, T., & Goto, M. (2013). A use of DEA-DA to measure importance of R&D expenditure in Japanese information technology industry. *Decision Support Systems*, *54*, 941-952.

Wang, E.C., & Huang, W.C. (2007). Relative efficiency of R&D activities: A cross-country study accounting for environmental factors in the DEA approach. *Research Policy*, *36*(2), 260-273.

Wolszczak-Derlacz, J., & Parteka, A. (2011). Efficiency of European public higher education institutions: a two-stage multicountry approach. *Scientometrics*, *89*, 887-917.

Worthington, A.C. (2001). An empirical survey of frontier efficiency measurement techniques in education. *Education Economics*, *9*(3), 245-268.

Worthington, A.C., & Lee, B.L. (2008). Efficiency, technology and productivity change in Australian universities, 1998-2003. *Economics of Education Review*, *27*(3), 285-298.

Yang, G.L., Yang, L.Y., Liu, W.B., Li, X.X., & Fan, C.L. (2013a). Directional returns to scale of biological institutes in Chinese Academy of Sciences. In: (J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, H. Moed, Eds.). *Proceedings of ISSI 2013* (pp. 551-566).Vienna: Austrian Institute of Technology (AIT).

Yang, G.L., Rousseau, R., Yang, L.Y., & Liu, W.B. (2014a). A study on directional returns to scale. *Journal of Informetrics*, *8*(3), 628-641.

Yang, G.L., Shen, W.F., Zhang, D.Q., & Liu, W.B. (2014b). Extended utility and DEA models without explicit input. *Journal of the Operational Research Society*, *65*, 1212–1220.

Yang, L.Y., Zhou, Q.J., & Yue, T. (2013b). China's Science: The Overall Development and the Balance of Disciplinary Structure—Statistics and Analysis of SCI-indexed Papers in 2012. *Science Focus*, 8(1), 23-50. (In Chinese).

Yang, H.-H., & Chang, C.-Y. (2009). Using DEA window analysis to measure efficiencies of Taiwan's integrated telecommunication firms. *Telecommunications Policy*, *33*(1-2), 98-108.

Zhong, W., Yuan, W., Li, S.X., & Huang, Z.M. (2011). The performance evaluation of regional R&D investments in China: An application of DEA based on the first official China economic census data. *Omega-the International Journal of Management Science*, *39*(4), 447-455.

Continuous, Dynamic and Comprehensive Article-Level Evaluation of Scientific Literature

Xianwen Wang[1], Zhichao Fang[1] and Yang Yang[1]

[1] xianwenwang@dlut.edu.cn, fzc0225@dlut.edu.cn, yangyang0477@mail.dlut.edu.cn
WISE Lab, Faculty of Humanities and Social Sciences, Dalian University of Technology, Dalian 116085 (China)

Abstract

Current research assessment is built on the basis of core-journals-selection system. Journal evaluation is not equal to article evaluation, evaluating scientists, institutions and countries based on article-level evaluation is more reasonable than the current journal-based evaluation. Different from the current research evaluation tools and databases, e.g., ESI and Nature Index, in this study, we propose the idea of continuous, dynamic and comprehensive article-level-evaluation based on article-level-metrics data. Different kinds and sources of metrics are integrated into a comprehensive indicator, to quantify both the long-term academic and short term societal impact of the article. At different phases after the publication, the weights of different metrics are dynamically adjusted to mediate the long term and short-term impact of the paper. Using the sample data, we collect the metrics data over two years for each sample article, and make empirical study of the article-level-evaluation method. The original data and interactive visualization of this research is available at http://xianwenwang.com/research/ale/.

Conference Topic

Altmetrics; Indicators; Science policy and research assessment

Introduction

For decades, citation has been regarded as the sole indicator to evaluate the impact of a paper, a paper that is cited more frequently means the research results gained more recognition. However, citations need a long time (often over two years) to accumulate. In many situations, e.g., funding decisions, hiring tenure and promotion, people need to make evaluations for newly published papers. Alternatively, some people begin to use journal based metrics, e.g., Journal Impact Factor, as an alternative way to quantify the qualities of individual research articles (Alberts, 2013). There are many debates about the abuse of Impact Factor (Bordons, Fernández, & Gomez, 2002; Garfield, 2006; Opthof, 1997; PLoS_Medicine_Editors, 2006; Seglen, 1997), applying Journal Impact Factor to assess the research excellence is not the most appropriate way. In addition, only tracking citation metrics could not tell the whole story about the influence of a paper. Besides citation, the impact of scientific papers could be reflected with article usage (browser views and pdf downloads), captures (bookmarks and readership), online mentions (blog posts, social media discussions and news reports) (Priem, Taraborelli, Groth, & Neylon, 2010). Therein, the idea of altmetrics comes into being. Different from citation, which puts particular emphasis on describing the academic impact of articles, altmetrics is based on data gathered from social media platforms and focuses on the societal impact (Kwok, 2013; Sud & Thelwall, 2014; Zahedi, Costas, & Wouters, 2014). Compared with the long time for papers to reach their citation peaks, it takes a short period for newly published articles to peak for altmetric scores. In summary, citation is an indicator to measure the long-term academic impact, when the indicator of altmetrics reflects short term societal impact. Neither citations nor altmetrics individually could fully indicate the complete impact of a paper, we cannot accurately conjecture the results of one metric by the results of another.

It is necessary to find a way to quantify both the academic and societal impact together, and mediate the long term and short-term impact of the paper. Some publishers have already listed

the different types of metrics for an individual article, e.g., PLOS, when some altmetrics tools and services are also available, e.g., Impact Story, Altmetric.com, Plum Analytics, etc. Although altmetric score from altmetric.com is a weighted count that integrates different online mentions of the paper. If we go further on this way, taking all available metrics (e.g., citation, usage, online attention, etc.) into consideration to design a comprehensive metric, which could be used to evaluate the complete impacts of articles.

Based on the calculated total impacts, the comprehensive metric makes it possible to rank articles on a unified dimension, which solo academic or societal impact indicator could not.

The absence of evaluating data source

According to the official statement of Web of Science, it is designed for researchers to "find high-impact article". Nowadays, with the absence of specialized evaluating data source, Web of Science has been adopted by many scientometrics researchers and institutions as the primary data source of article evaluation. In some countries, e.g., China, articles indexed in Science Citation Index/Social Science Citation Index or not is a very important criterion to judge the quality of the research.

However, applying Web of Science to assess the research performance and research excellence is not a good choice. Web of Science is designed and created on the basis of journal selection, it collectively index journals cover-to-cover. However, articles published in the same journal, the same issue, have totally different impacts. Even for those high impact factor journals, there are many articles have few citations.

We check the articles published in 2000 and indexed in Science Citation Index Expanded, as Table 1 shows. For example, 2901 of the total 13660 articles in Chemical Engineering have never been cited. For the area of Condensed Matter Physics, the zero-citation percentage is 10.91%, for the area of Biochemistry, Molecular Biology, the zero-citation percentage is 3.23%.

Table 1. Number of Zero-citation articles in 2000 indexed in Science Citation Index Expanded.

	Total	*Zero-citation*	*Percentage*
Engineering, Chemical	13660	2901	21.24
Physics, Condensed Matter	21974	2397	10.91
Biochemistry, Molecular Biology	42710	1380	3.23

There are also some publishers regard Web of Science as a profit-making tool. For example, *Academic Journals* charges a US$550-$750 manuscript handling fee from the author for each accepted article (http://www.harzing.com/esi_highcite.htm). Among which, several ISI-listed journals publish more than 1,000 articles per year, e.g., in 2007, *African Journal of Business Management* only published 28 articles, in 2010, it published 446, when in 2011, as many as 1350 articles were published by this single journal. Thomson Reuters has the mechanism to review the exiting journal coverage constantly, some journals that have become less useful would be deleted. However, this kind of mechanism does not apply to the articles, even some journals are deleted from the coverage, numerous low-quality papers published by these journals are still indexed in Web of Science.

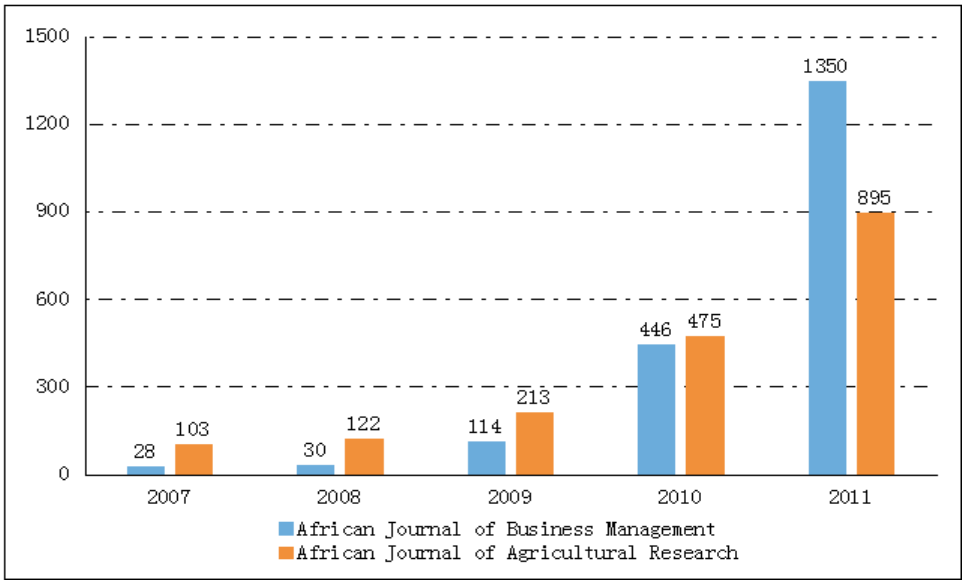


Figure 1. Rapid growth of yearly indexed articles of two journals.

With the same idea of Web of Science, Nature Publishing Group (NPG) introduced the Nature Index in November 2014, which is "a database of author affiliation information collated from research articles published in an independently selected group of 68 high-quality science journals" (Nature, 2014). The 68 journals are selected by a group of professors and validated by 2,800 responses to a large-scale survey, when these 68 journals account for approximate 30% of total citations to natural science journals (http://www.nature.com/press_releases/nature-index.html).

Based on journal article publication counts and citation data from Thomson Scientific databases (mainly from Web of Science), ISI/Thomson (now Thomson Reuters) proposed Essential Science Indicators (ESI), which is an in-depth analytical tool and also a database where citations are analyzed, so that scientists, journals, institutions, and countries can be ranked and compared, for example, most cited scientists rankings, institutions rankings and countries rankings. Ranking in ESI is made by the citations, it has nothing to do with the Impact Factors of journals, which means that whichever journal the paper is published in, citations is the only factor to be taken into account. Although ESI set a relatively low selection criterion for newly published papers (http://www.in-cites.com/thresholds-highly-cited.html), using cited times to evaluate is not a good choice.

Compared to 8670 journals covered by Science Citation Index Expanded, the journals selected by Nature Index is so much less, which makes Nature Index become an elite database. The aim of Nature Index is "intended to be one of a number of metrics to assess research excellence and institutional performance" (http://www.natureindex.com/faq). However, we think journal-based database is not appropriate for research evaluation, including research excellence and institutional performance, which should be on the basis of article-level metrics. Because of the great influence of Nature Publishing Group, the Nature Index will definitely make great changes to the academia and research evaluation system.

It is necessary to make changes to the current evaluating way of scientific literature. In this research, our purpose is to design a new method, through which the continuous, dynamic and comprehensive evaluation of scientific literature could be made. This new method will be valuable to the research community. With this evaluating method and system, we could make a better evaluation of articles, scientists, journals, institutions, and even countries.

Design a new evaluation way

Considering both academic and societal impact of a paper

As mentioned above, the impact of a paper could be measured by citation, article usage and online mentions, etc., as Table 2 shows.

Table 2. Types and metrics of the impact of a paper.

Type	*Metric*
Article usage	browser views (abstract, full-text), pdf downloads
Captures	bookmarks (CiteUlike), readers (Mendeley)
Online mentions	blog posts, news reports, likes (Facebook), shares (Facebook), Tweets, +1 (Google plus)
Citations	citations

The Issue 6, Volume 8 of PLOS Computational Biology is selected as our research object. It was published in June 2012, and includes 46 research articles.

In November 2012, PLOS began to provide a regular report covering a wide range of article-level-metrics covering all of its journals via the platform http://article-level-metrics.plos.org/. In this research, the cumulative article-level-metrics data for the entire PLOS corpus are harvested from the PLOS ALM platform. From October 2012 to October 2014, PLOS has provided the ALM reports for 8 times, when the provided date are Oct. 10, 2012, Dec. 12, 2012, Jan. 8, 2013, Apr. 11, 2013, May. 20, 2013, Aug. 27, 2013, Mar. 10, 2014 and Oct. 1, 2014. Factor analysis is employed to study the metrics data of the 46 articles, Table 3 shows the results of the data extracted from the ALM report of Oct. 2014.

Table 3. Rotated Component Matrix.

	Factor 1: Academic impact	*Factor 2: Societal impact*
CiteUlike	0.775	
Mendeley	0.856	
HTML views	0.692	0.672
PDF downloads	0.917	
Scopus	0.751	
Facebook		0.745
Twitter		0.709

Note. Factor loadings < .5 are suppressed

7 metrics data of Oct. 10, 2012 are factor analyzed by using principal component analysis with Varimax (orthogonal) rotation. The analysis yields two factors explaining a total of 73.709% of the variance for the entire set of variables. Factor 1 is labeled academic impact to the high loadings by the following items: CiteUlike bookmarks, Mendeley readership, PDF downloads and Scopus citations. This first factor explained 48.691% of the variance. The second factor derived is labeled societal impact. This factor is labeled as such due to the high loadings by the two indicators of Facebook and Twitter. The variance explained by this factor is 25.018%. For the indicator of HTML views, the both factor loadings are greater than 0.65, which means that browser HTML views has both academic and societal impact.

The Altmetric score is a quantitative measure of the attention that a scholarly article has received. It is a weighted count of the different online platform sources (newspaper stories, tweets, blog posts, comments) that mention the paper. Downloads, citations and reader counts

from Mendeley or CiteULike are not used in the score calculation. So, Altmetric score could be regarded as a comprehensive indicator that measures the societal impact of paper partially.

Dual function of societal impact

The value of societal metrics is not only reflected by the social effects of the diffusing of the knowledge embodied in the literature, but also reflected by the possible additional academic impact caused by social online attention.

Social media make the research achievements and scientific discoveries spread to the general public, which is just the goal of scientific researches. From the other hand, wide spreading of scientific literature could lead to more scholarly citations. The mechanism from online attention to citation is very complicated, but social attention do have the potentiality to contribute some extra citations to a paper (Wang, Liu, Fang, & Mao, 2014; Wang, Mao, Zhang, & Liu, 2013).

Dynamic patterns of article-level metrics

For the 46 selected articles published in June 2012, we sum the metrics data at the 8 time periods separately, as Figure 2 shows. Different metrics show different dynamic evolution patterns. In October 2012, when the articles had been published for about 4 months, there is few citations. The curve of citations begins a sharp rise at the phase of May 2013, one year after the publication. However, for the Facebook and Twitter data, the two curves have almost reached their summits at the very first phase. During the next periods, there is little increase for the Facebook and Twitter data. And for the views data, which is placed on the secondary Y axis in Figure 2, the situation is somehow between the citations and Facebook/Twitter. At the first phase, there is considerable data. During the following 7 periods, there is a steady growth trend for the curve of views.

Dynamic patterns for the different metrics are distinct. Social attention comes to go, citation takes a long time to know, when article view also comes fast but keeps a steady growth.

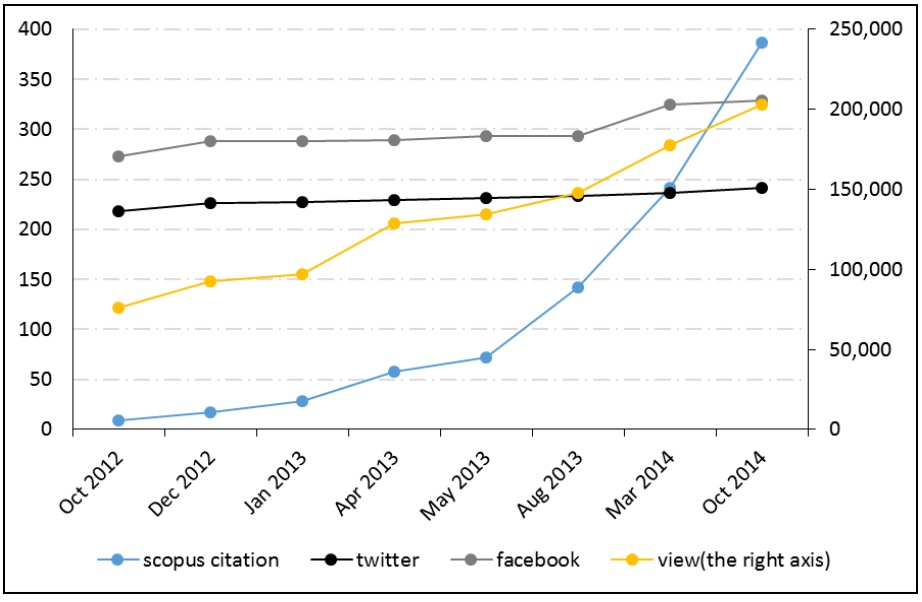


Figure 2. Temporal trend of different metrics of 46 articles published in June 2012.

Article-level evaluation based on Article-level-metrics

In the era of print, the article could not be separated from the whole issue. For example, libraries could provide the borrowing statistical data, however, it's difficult to know which

single article or articles readers are interested in. In the digital era, the situation has been changed greatly. Metrics data for each article are easy to know, including the views, downloads, altmetric score and citations. Of course, some data are easy for publishers to know but not released to public. As early in March 2009, PLOS inaugurated a program to provide "article-level metrics" on an article across all PLOS journals. The metrics data include five main categories, which are Viewed, Cited, Saved, Discussed and Recommended. Following PLOS, more and more publishers began to provide detailed article-level metrics data for readers and researchers. For example, in October 2012, Nature began to provide a real-time online count of article-level metrics for its published research papers, including citation data, news mentions, blog posts and details of sharing through social networks, such as Facebook and Twitter (http://www.nature.com/news/nature-metrics-1.11681). In 2014, the article-level metrics data are also available for PNAS and Science. The growing article-level metrics dataset provides us with the possibility to design a new evaluating way to make article-level evaluation.

Problems need to be solved

The first problem is there are too many indicators need to be considered. Citation has been regarded as the single indicator for the past tens of years, nowadays there are much more indicators which are worth being considered, including article views, bookmarks and readership, online discussion, news reports and citations, etc. So many indicators mean a lot of dimensions of the impact, different papers may have different values for the indicators, for example, paper A has been downloaded many times but retweeted few times, when paper B may has opposite situation, so it is very difficult to compare the impact of these two articles, especially when these articles are newly published.

Could these so many indicators be synthesized to one single comprehensive indicator, which could reflect the most of information of the original data and make the papers in diverse situations comparable?

The second problem is the dynamic adjustment of the results. At different phases after publication, the same indicator may have different effects on the impact of the paper. For the newly published articles, because the citations are generally low, it is difficult to judge the qualities and compare the new articles. At the early phase, it is a better choice to use article usage data, online mention data to make evaluation of the newly published articles. As time goes by, the evaluation is gradually dominated by citation metrics, which means that citation would play the most important role in the evaluation when the article has been published for a relatively long time. To solve these two problems, we propose the idea of designing a comprehensive indicator to reflect all the impacts of an article. The weights of the indicators at different phases should be adjusted dynamically due to the change of relative importance of metrics, just like Table 4 shows.

To integrate different metrics into a comprehensive indicator, the first problem needs to be solved is weighting. Here we use Analytic hierarchy process (AHP) to calculate the weights of different metrics. The AHP methodology was developed by Thomas L. Saaty in the 1970s (Saaty, 1980). It allows users to assess the relative weight of multiple criteria in an intuitive manner, so it has both advantages of quantitative criteria and qualitative judgment provided by the users. Using pairwise comparisons (X is more important than Y), the relative importance (priority) of one criterion over another can be expressed. To calculate the weights for the different criteria, a pairwise comparison matrix needs to be created. The matrix is a matrix A, where m is the number of evaluation criteria considered, denotes the entry in the *i*th row and the *j*th column of matrix. Each entry of the matrix represents the importance of the *i*th criterion relative to the *j*th criterion. If the cell value in the entry is greater than 1, then the *i*th criterion is more important than the *j*th criterion, and vice versa. If two criteria have the

same importance, then the cell value in the entry is 1. The relative importance between two criteria is measured according to a numerical scale from 1 to 9 or 1/9 to 1.

Table 4. Relative importance of metrics at different phases.

Phase	*Relative importance*	*Selection standard*
1 (0-6 months)	PDF downloads > HTML views > Twitter > Facebook > Mendeley > CiteUlike > Citation	Top 80% of all articles of same month and subject
2 (6 months-2 years)	PDF downloads > HTML views > Mendeley > CiteUlike > Citation > Twitter > Facebook	Top 70% of all articles of same month and subject
3 (2 -5 years)	Citation > Mendeley > CiteUlike > PDF downloads > HTML views > Twitter > Facebook	Top 50% of all articles of same year and subject
4 (5 years-)	Citation > Mendeley > CiteUlike > PDF downloads > HTML views > Twitter > Facebook	Top 30% of all articles of same year and subject

According to the definition of relative importance of different metrics, we need to construct different pairwise comparison matrixes at different phases. The pairwise comparison matrix at phase 1 is shown in Table 5. The higher the weight is, the more important the corresponding criterion becomes, which is represented by the cell value in the matrix. For example, the values in the cells where the row of CiteUlike, the column of HTML views and PDF downloads intersect are less than 1, moreover, the ratio of CiteUlike and PDF downloads is less than the ratio of CiteUlike and HTML views, it means that at phase 1, CiteUlike is less important than HTML views, and much less important than PDF downloads.

Table 5. Pairwise Comparison Matrix at phase 1.

	CiteUlike	*Mendeley*	*HTML views*	*PDF downloads*	*Citation*	*Facebook*	*Twitter*
CiteUlike	1	1	1/4	1/6	4	1/4	1/6
Mendeley		1	1/4	1/6	4	1/4	1/6
HTML views			1	1/4	6	3	2
PDF downloads				1	9	4	3
Citation					1	1/4	1/7
Facebook						1	1/2
Twitter							1

At phase 4, there is much change in the relative importance of the metrics, as Table 6 shows. CiteUlike and Mendeley become more important than HTML views, so the cell values get greater than 1. At this phase, citation is the most important criterion.

In this study, the weights and CI values of AHP models are calculated by a CGI system (http://www.isc.senshu-u.ac.jp/~thc0456/EAHP/AHPweb.html). The results are shown in Table 7.

In Figure 3, we show the change of the weights of metrics. At Phase 1 and 2, the metric of PDF downloads has the greatest weight. From Phase 1 to 4, the curve of PDF downloads shows a downward trend, when the weight of citation is upward.

Empirical Study

The weights in Table 7 are applied to calculate the comprehensive scores of the metrics data of the 46 articles. Metrics data of Oct. 10, 2012 is calculated with the weights of phase 1,

Table 6. Pairwise Comparison Matrix at phase 4.

	CiteUlike	*Mendeley*	*HTML views*	*PDF downloads*	*Citation*	*Facebook*	*Twitter*
CiteUlike	1	1	3	2	1/7	3	2
Mendeley		1	3	2	1/7	3	2
HTML views			1	1/4	1/9	1	1
PDF downloads				1	1/6	1	1
Citation					1	4	3
Facebook						1	1/2
Twitter							1

Table 7. Weights of AHP models at different phases.

	CiteUlike	*Mendeley*	*HTML views*	*PDF downloads*	*Citation*	*Facebook*	*Twitter*
Phase 1	0.0477	0.0477	0.1996	0.3901	0.0234	0.1109	0.1806
Phase 2	0.1723	0.1723	0.1182	0.2108	0.1321	0.0828	0.1116
Phase 3	0.1514	0.1514	0.0481	0.0921	0.3979	0.0644	0.0947
Phase 4	0.1269	0.1269	0.0455	0.0809	0.4819	0.0570	0.0810

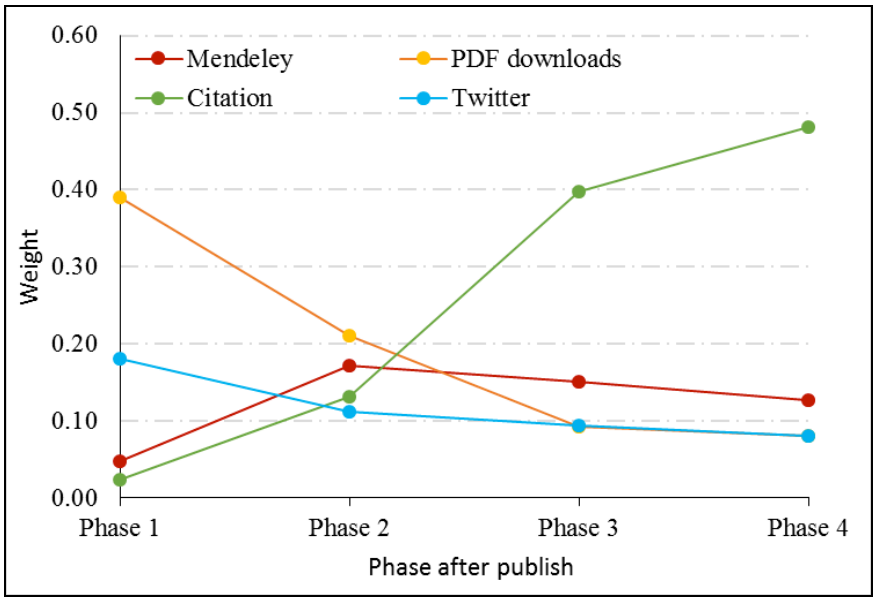


Figure 3. The change of the weights of different metrics.

when weights of phase 2 and 3 are used for metrics data of Aug. 27, 2013 and Oct. 1, 2014 separately. All the original metrics data are normalized to the range of 0-1. The normalized value of e_i for variable E in the ith row is calculated as:

$$Normalized\ (e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}}$$

Where $E_{\min}$ and $E_{\max}$ are the minimum and maximum value for variable E correspondingly.

In Table 8, the values of 7 metrics are original data, when the scores are calculated with the normalized data instead of the original metrics data.

Table 8. Top 25% articles with greatest score at 3 phases.

phase	*rank*	*doi*	*citeulike*	*mendeley*	*html*	*pdf*	*citation*	*facebook*	*twitter*	*score*
	1	1002358	16	81	5060	1733	3	8	12	0.7906
	2	1002543	14	0	4041	871	0	2	31	0.5653
	3	1002590	0	18	4302	469	0	73	11	0.4413
	4	1002561	3	37	3579	721	0	0	9	0.3671
	5	1002519	3	17	2516	648	0	0	13	0.3146
1	6	1002538	3	6	1777	394	0	22	15	0.2603
	7	1002541	13	24	1794	354	0	3	12	0.2456
	8	1002527	3	12	1818	373	0	6	14	0.2305
	9	1002572	6	18	2045	489	0	0	6	0.2248
	10	1002588	0	13	1809	454	1	0	7	0.1989
	11	1002531	4	20	1519	522	1	2	1	0.1865
	1	1002358	16	170	11720	3236	30	7	14	0.8579
	2	1002543	16	72	5389	1103	1	2	34	0.4739
	3	1002561	3	79	9669	1242	5	2	11	0.3408
	4	1002541	15	57	3609	665	3	4	13	0.3395
	5	1002590	1	36	6024	627	1	91	13	0.2622
2	6	1002531	8	39	3389	912	11	3	1	0.2552
	7	1002519	3	39	5515	1262	1	0	13	0.2419
	8	1002572	6	44	3273	754	2	0	6	0.2006
	9	1002538	3	14	3155	668	4	22	15	0.1889
	10	1002577	2	25	5063	1141	2	0	5	0.1816
	11	1002527	3	21	3266	638	1	6	14	0.1641
	1	1002358	18	324	19909	4651	73	23	14	0.8942
	2	1002543	16	95	6071	1241	1	2	36	0.3113
	3	1002541	16	91	4896	824	11	4	13	0.2931
	4	1002531	9	77	5670	1229	26	3	1	0.2874
	5	1002561	4	121	11231	1577	21	2	11	0.2866
3	6	1002588	0	56	6112	1314	19	3	8	0.1849
	7	1002572	9	62	3803	910	6	0	6	0.1707
	8	1002519	3	69	8233	1653	6	0	13	0.1692
	9	1002590	1	42	7101	904	3	90	13	0.1690
	10	1002555	3	31	5048	701	13	22	4	0.1531
	11	1002562	7	58	2840	529	10	0	0	0.1476

Note: (1) Because of the limited layout space, the first half of the doi is omitted. For example, for the doi 10.1371/journal.pcbi.1002358, we only keep 1002358 in Table 8.

(2) Detailed information of Table 8 is available at http://xianwenwang.com/research/ale

Table 8 lists the top 11 (top 25% of 46) articles of each phase. At phase 1, when the 46 articles had been published for 4 months, article 10.1371/journal.pcbi.1002358 has 16

CiteUlike bookmarks, 81 Mendeley readers, 5060 HTML views, 1733 PDF downloads and 3 Scopus citations, etc., when the comprehensive score of this article is 0.7906, ranks top 1. At phase 2, the values of the metrics of Mendeley, HTML views, PDF downloads and Scopus citations have risen sharply, but not for the metrics of Facebook and Twitter, when the score is 0.8579 and still ranks top 1. From phase 1 to 2 and 3, there is much change for the top 11 articles. The ranks of some articles rise, when others may fall. For example, article 10.1371/journal.pcbi.1002538 ranks 6th at phase 1, downs to 9 at phase 3, and is disappeared from the top 11 at phase 3; article 10.1371/journal.pcbi.1002531 ranks 11 at phase 1, and rises to top 4 at phase 3.

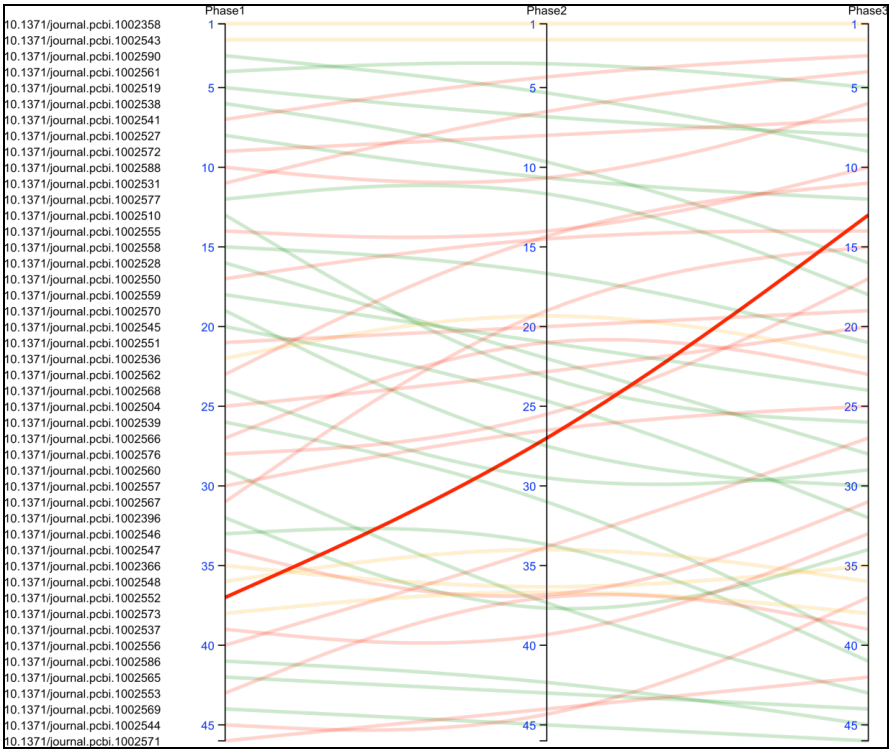


Figure 4. Dynamic changes according to the ranking at different phases.[1]

The dynamic changes of the scores and rankings of the 46 articles from phase 1 to 3 are shown in Figure 4. The DOIs of 46 articles are listed on the leftmost column, and ranked according to the scores at phase 1. The position of article at the certain phase is decided by the ranking of score at that phase. 46 articles could be only compared at the same phase. Articles at different phases, and even the same article at different phases are not comparable. As shown in Figure 4, if the rank of an article from phase 1 to 3 shows an upward trend, it is displayed with a red curve, there are 20 papers with red curves. We use green curve to represent the downward trend, there are also 20 papers with green curves. Otherwise, if the rank of the article has not changed, the color of the curve is yellow, there are 6 yellow curves. In Figure 4, one red curve with dramatic upward trend is highlighted, indicating that the performance of this paper is rising. The doi of this article is 10.1371/journal.pcbi.1002552, it only ranks 37 at phase 1, rises to 28 at phase 2 and continue to rise to 13 at phase 3.

According to the rankings calculated by the comprehensive metric, articles with the highest impact are selected into the database. There are different selection standards at different phases, as Table 4 shows. As time goes on, the data of the original indicators become

[1] An interactive version of Figure 4 is available at http://xianwenwang.com/research/ale/dynamic.html

increasingly sufficient, the accuracy of the results becomes higher. Due to the dynamic changes of the rankings of articles, the database is also dynamic, it ensures the articles included are always has the highest impact at each phase. It would be much easier for researchers to index the high quality articles through the dynamic database.

Discussion

In the 1950s, people read papers from printed journals. A group of articles are bundled together to form an issue of journal, it is difficult to separate single article from the whole issue, which is the carrier of articles. For example, if we want to know which paper the readers are interested in when they borrow the journal from the library, which seems to be an extremely difficult task. At that time, journal evaluation is the most important and basic issue. SCI is designed on the basis of core journals selection, specialized indicators and tools are proposed to evaluate journals, e.g., Impact Factor and Journal Citation Reports.

Compared to fifty years ago, scholarly communicating ways have changed a lot. With the advent and fast development of computers, internet and digital libraries, the transformation from print to electronic publishing is accelerating, just as the digital music revolution set music free from the carriers of cassette tape and CD, the concept of printed journals or even journals in the conventional sense is not important any more. Actually, for some new journals, articles are not organized and published by issues and volumes, e.g., PLOS ONE, Scientific Reports, eLIFE and Peer J, etc.

It is necessary to make changes to the current research evaluation way rooted in the journal selection system. We should be aware of that journal evaluation is not equal to article evaluation, evaluating scientists, institutions and countries based on article-level evaluation is more reasonable than the current journal-based evaluation. It would "be better to measure the performance of countries and institutions on the basis of individual papers, rather than on the journals in which they are published" (Haunschild & Bornmann, 2015). In order to make better assessment of research performance and research excellence, we propose the idea of article level evaluation system and database. Using metrics data at different time periods of 46 articles in one issue, we make empirical test of the article level evaluation method.

Firstly, the basic function of this evaluation system is to assess the qualities of articles. Based on article level evaluation, it is also available to assess the research excellence of scientists, journals, institutions and countries. For example, how many articles tracked in phase 3 and 4 are published by one specific institution? What are the top institutions in one specific field? Secondly, both scholarly and societal impact of articles are taken into account. Thirdly, using the article usage data and online mention data, we can make evaluation of newly published papers. At different phases after publication, the comprehensive score of the paper is calculated with different weights of metrics, so the score and rank of a paper in different phases change.

To accomplish this, the biggest problem needs to be solved is the availability of metrics data. The citation data could be obtained from Web of Science, Scopus, Google Scholar, etc. The online attention data, e.g., social media, news reports, Mendeley readership is also available from various but certain data sources. However, for the article usage data, only part of academic publishers and journals provide usage data to public, including Nature Publishing Group, Science, PLOS, Taylor & Francis, ACM Digital Library, IEEE Xplore Digital Library, etc. (Wang, Mao, Xu, & Zhang, 2013). For many others, e.g., Elsevier, Sage and Wiley, they may provide the metrics data of each article to some specific users and subscribers, but not free to public. If we want to evaluate all the papers whatever the publishers are, metrics data from publishers is indispensable.

With the movement from print to electronic publishing and the diversification of article-level-metrics, it is time to make change to the current research evaluation system. To better assess

scientists' research and satisfy the evaluation needs in many situations, ranging from funding decisions to hiring tenure and promotion, we need to build an article-level-evaluation system.

Limitation

In this study, we interpret the idea of building such a kind of system and make empirical study using a relative small size dataset, and we only track the metrics data of the sample articles in the last two years. To build the article-level-evaluation system is not an easy job, of course there are lots of problems need to be solved, including a bigger dataset, longer time period, more detailed metrics and maybe more scientific weighting methods, but we think it is the right way to make assessment of research, we are moving on the right direction.

Acknowledgments

The work was supported by the project of ''National Natural Science Foundation of China'' (61301227), the project of "Growth Plan of Distinguished Young Scholar in Liaoning Province" (WJQ2014009), and the project of "the Fundamental Research Funds for the Central Universities" (DUT15YQ111).

References

Alberts, B. (2013). Impact factor distortions. *Science*, *340*(6134), 787-787.

Bordons, M., Fernández, M. T., & Gomez, I. (2002). Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics*, *53*(2), 195-206.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, 295(1), 90-93.

Haunschild, R. & Bornmann, L. (2015). Publishing: Criteria for Nature Index questioned. *Nature*, *517*(7532), 21-21.

Kwok, R. (2013). Research impact: Altmetrics make their mark. *Nature*, 500(7463), 491-493.

Nature (2014). Launch of the Nature Index. Retrieved December 15, 2014 from: http://www.nature.com/news/launch-of-the-nature-index-1.16310

Opthof, T. (1997). Sense and nonsense about the impact factor. *Cardiovascular Research*, *33*(1), 1-7.

PLoS_Medicine_Editors. (2006). The impact factor game. *PLoS Medicine*, 3(6), e291.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. In (Vol. 2014). Retrieved June 14, 2015 from: http://altmetrics.org/manifesto.

Saaty, T. L. (1980). *The analytic hierarchy process: planning, priority setting, resources allocation*. New York: McGraw.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, *314*(7079), 497.

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics*, *98*(2), 1131-1143.

Wang, X., Liu, C., Fang, Z., & Mao, W. (2014). From attention to citation, what and how does altmetrics work? *arXiv preprint arXiv:1409.4269*.

Wang, X., Mao, W., Xu, S., & Zhang, C. (2013). Usage history of scientific literature: Nature metrics and metrics of nature publications. *Scientometrics*, *98*(3), 1923-1933.

Wang, X., Mao, W., Zhang, C., & Liu, Z. (2013). The diffusion of scientific literature in web. In S*TI 2013 – 18th International Conference on Science and Technology Indicators* (pp. 415-426). Berlin.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, *101*(2), 1-23.

Interdisciplinarity and Impact: Distinct Effects of Variety, Balance, and Disparity

Jian Wang[1], Bart Thijs[1] and Wolfgang Glänzel[1]

{Jian.Wang, Bart.Thijs, Wolfgang.Glanzel}@kuleuven.be
[1] KU Leuven (Belgium)

Abstract

Interdisciplinary research is increasingly recognized as the solution to today's challenging scientific and societal problems, but the relationship between interdisciplinary research and scientific impact is still unclear. This paper studies the relationship between interdisciplinarity and citations at the paper level. Different from previous literature compositing various aspects of interdisciplinarity into a single indicator, this paper uses factor analysis to uncover distinct aspects of interdisciplinarity and investigates their independent dynamics with scientific impact. Three uncovered factors correspond to variety, balance, and disparity. Subsequently, we estimate Poisson models with journal fixed effects and robust standard errors to investigate the relationship between these three factor and citations. We find that the number of citations (1) increase at an increasing rate with variety, (2) decrease with balance, and (3) increase at a decreasing rate with disparity. These findings have important implications for interdisciplinarity research and science policy.

Conference Topic

Science policy and research assessment

Introduction

Interdisciplinary research has been increasingly viewed as the remedy for the challenging contemporary scientific and societal problems. As important ideas often transcend the scope of a single discipline, interdisciplinary research is the key to accelerate scientific discoveries and solve societal problems. Given the normative interest in and the policy push for interdisciplinary research, it's important to empirically investigate the consequences of interdisciplinary research. Bibliometric studies have explored the relationship between interdisciplinary research and citation impact, but findings are mixed. For example, Steele and Stier (2000) found a positive effect of interdisciplinarity on citation impact for environmental sciences papers, where interdisciplinarity was measured as the disciplinary diversity of the cited references. Rinia, van Leeuwen, van Vuren, and van Raan (2001) studied physics programs in the Netherlands and operationalized interdisciplinarity as the ratio of non-physics publications. They found significantly negative correlations between interdisciplinarity and non-normalized citation-based metrics, but correlations became insignificant when field-normalization took place. Levitt and Thelwall (2008) found that interdisciplinary papers received fewer citations in life and physical sciences but not in social sciences, and interdisciplinary papers were defined as papers published in journals assigned to multiple subject categories. Larivière and Gingras (2010) analyzed all Web of Science (WoS) articles published in 2000, measured interdisciplinarity as the percentage of its cited references to other disciplines, and found an inverted *U*-shaped relationship between interdisciplinarity and citations.

One possible explanation for these conflicting results pertains to their different choices of the interdisciplinarity measure. On the one hand, a number of interdisciplinarity indicators have been proposed, at various levels (e.g., paper, journal, institution, and fields) and using various bilometric information (e.g., disciplinary memberships of authors, published journals, or cited references). On the other hand, the concept of interdisciplinarity remains an abstract and complex one (Wagner et al., 2011). One useful conceptualization is to view interdisciplinarity

as the diversity of disciplines invoked in the research (Porter & Rafols, 2009; Stirling, 1998, 2007). Furthermore, diversity has three distinct components (Stirling, 2007, p. 709):

> *Variety is the number of categories into which system elements are apportioned. It is the answer to the question: 'how many types of thing do we have?'*
>
> *Balance is a function of the pattern of apportionment of elements across categories. It is the answer to the question: 'how much of each type of thing do we have?'*
>
> *Disparity refers to the manner and degree in which the elements may be distinguished. It is the answer to the question: 'how different from each other are the types of thing that we have?'*

Many studies have devoted to compositing all aspects of interdisciplinarity into one single indicator. However, this paper adopts an opposite approach: we decompose different aspects of interdisciplinarity and explore their unique relationships with citation impact, at the individual paper level. Given that interdisciplinarity is an abstract and multidimensional concept, there might not be a straightforward answer to the question of whether interdisciplinary research draws higher impact. Instead, we should ask the question in another way: what kinds of interdisciplinarity have positive/negative relationships with citation impact? In addition, nuanced understanding of the divergent dynamics underlying different aspects of interdisciplinarity is also important for informing interdisciplinary research and science policy.

Data and methods

We analyzed all the journal articles published in 2001 indexed in the Thomson Reuters Web of Science Core Collection (WoS). Only articles were analyzed, while all other document types such as reviews and letters were excluded. The year 2001 was chosen so that studied papers could have a sufficiently long period to accumulate their citations (Wang, 2013).

Interdisciplinarity measures

Following previous literature, we constructed interdisciplinarity measures for each individual articles based on the disciplinary profile of its cited references, since referencing to prior literature in various disciplines indicates drawing and integrating knowledge pieces from these disciplines. Specifically, we constructed interdisciplinarity measures based on the WoS subject categories (SCs) referenced by each article. Interdisciplinarity measures constructed in this paper are listed in Table 1, which have been commonly used in the literature (Leydesdorff & Rafols, 2011; Rafols et al., 2012; Stirling, 2007). Because the last two interdisciplinarity measures cannot be constructed if the focal article references fewer than two subject categories, we excluded these articles from the analysis. Nevertheless, regressions using the whole dataset for the other measures yielded consistent results. In total, our data have 646,669 papers.

Factor analysis

We used factor analysis to uncover components underlying these interdisciplinarity measures. The first step was to determine the number of factors to retain. A classic approach is Kaiser's eigenvalue greater than one rule (Kaiser, 1960). The idea is that the retained factor should explain more variance than the original standardized variables. Horn's parallel analysis

Table 1. Interdisciplinarity measures.

Measure	*Description*
Ratio of references to other subject categories	
Number of referenced subject categories	n
1 – Gini	$1 - \frac{\sum(2i - n - 1)x_i}{n \sum x_i}$ where i is the index, x_i is the number of references to the i-th subject category, and subject categories are sorted by x_i in non-decreasing order.
Simpson index	$1 - \sum {p_i}^2$ where $p_i = x_i/X$, and $X = \sum x_i$
Shannon entropy	$-\sum p_i log\ (p_i)$
Average dissimilarity between referenced subject categories	$\frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}$ where d_{ij} is the dissimilarity between subject category i and j. Specifically, $d_{ij} = 1 - s_{ij}$, where s_{ij} is the cosine similarity between subject category i and j based on their co-citation matrix.
Rao-Stirling diversity	$\sum_{i \neq j} p_i p_j d_{ij}$

modified Kaiser's rule, where the criterion for each eigenvalue is different and also superior to one, and these criteria are obtained from a Monte-Carlo simulation (Horn, 1965). Cattell's scree test provided a graphical strategy: plotting the eigenvalues against the component numbers and searching for the elbow point (Cattell, 1966). However it does not yield a definitive number of factors to retain, which still relies on subjective judgments of the researcher. Recently, Raiche, Walls, Magis, Riopel, and Blais (2013) developed numerical solutions for Cattell's scree test: (1) the optimal coordinate solution for the location of the scree and (2) the acceleration factor solution for the location of the elbow. We implemented all these methods to determine the number of factors. After determining the number of factors to retain, we extracted these factors using the varimax rotated principal components method. In addition, the number of referenced subject categories is highly skewed, so its nature logarithm was used in the factor analysis.

Regression analysis

To study the relationship between interdisciplinarity and citation impact at the article level, we ran regressions, using the number of long-term citations (in a 13-year time window from 2001 to the end of 2013) as the dependent variable and the interdisciplinarity measures and extracted factors as explanatory variables.

For all our regressions, we incorporated journal fixed effects to control for (1) unobserved topic/subfield heterogeneities at a very refined level and (2) journal reputation effects (Judge et al., 2007). Therefore, we estimated the within-journal effects, in other words, we were evaluating the association between interdisciplinarity and citations among papers published in the same journal. In addition, the following variables were incorporated as controls: the number of authors, the number of countries, the number of pages, and the number of references. The numbers of authors, pages, and references are skewed so that their natural

logarithms were used in regression analyses. The number of countries is still highly skewed after logarithm transformation, so we created a dummy variable, international: 1 if the paper has authors from more than one country, and 0 otherwise. In our sample, about 19% of the papers are internationally coauthored.

Because citation counts are over-dispersed count variables, we used Poisson regression with robust standard errors, following previous literature (Hall & Ziedonis, 2001; Hottenrott & Lopes-Bento, In Press; Somaya, Williamson, & Zhang, 2007). An alternative is the negative binomial model. However, because the Poisson model is in the linear exponential class, Gourieroux, Monfort, and Trognon (1984) have shown that the Poisson estimator and the robust standard errors are consistent so long as the mean is correctly specified even under misspecification of the distribution, but the negative binomial estimator is inconsistent if the true underlying distribution is not negative binomial. Therefore, we adopted the Poisson model with robust standard errors in our empirical analysis. Furthermore, we incorporated journal fixed effects. Such fixed effects Poisson models can be fitted by conditioning out the individual fixed effects (Hausman, Hall, & Griliches, 1984).

Results

Decomposing interdisciplinarity

We used the following variables in the factor analysis: log number of referenced subject categories, ratio of references to other subject categories, 1 – Gini, Simpson index, Shannon entropy, average dissimilarity between referenced subject categories, and Rao-Stirling diversity. The first three eigenvalues are greater than 1, so 3 factors should be retained according to Kaiser's rule. Horn's parallel analysis also suggests 3 factors. Raiche's nongraphic solutions for Cattell's scree test lead to conflicting conclusions: the optimal coordinate approach suggests 3 factors, while the acceleration factor approach suggests 1 factor to retain. Considering (1) the consensus between the classic Kaiser's rule and Horn's parallel analysis, (2) the divergence in this recent nongraphic solution for Cattell's scree test, and (3) that the optimal coordinate solution actually agrees with the more conventional approaches. We decided to retain 3 factors. Subsequently, we extracted 3 factors using the varimax rotated principal components method, and the cumulative proportion variance explained is 0.89. Factor loadings are reported in Table 2. Simpson index and Shannon entropy have the highest loading on the first factor, which reflects the variety aspect of disciplinary diversity. 1 – Gini has the highest loading on the second factor, which reflects balance, and the average dissimilarity between referenced subject categories has the highest loading on the third factor, which reflects disparity. The results are also in line with Harrison and Klein (2007) that Simpson index and Shannon entropy reflect more on variety, while Gini reflects more on unbalance.

Table 2. Factor loading.

	Factor 1	*Factor 2*	*Factor 3*
ln(referenced SCs)	0.78	-0.59	0.15
Ratio oth-disc refs	0.67	0.35	-0.17
1 – Gini	-0.07	0.94	0.05
Simpson	0.93	-0.11	0.18
Shannon	0.91	-0.32	0.18
Avg dissimilarity	0.09	0.00	0.95
Rao-Stirling	0.77	0.04	0.59

Data sourced from Thomson Reuters Web of Science Core Collection.

Interdisciplinarity and impact

We first estimated the fixed effects Poisson models using the citation counts as the dependent variable and original interdisciplinarity measures as the independent variables (Fig. 1A, regression table not reported). The divergent results suggest that the low consensus in previous literature regarding the relationship between interdisciplinarity and citation impact may be partially explained by their different choice of the interdisciplinarity measures.

Table 3 reports fixed effects Poisson models using the extracted interdisciplinarity factors as independent variables. Variety, balance, and disparity are the three extracted factors, and they follow the standard normal distribution with mean equals to 0 and standard deviation equals to 1. Holding that the papers are published in the same journal, with the same number of authors, pages and references, and have the same status in terms of whether being internationally coauthored, the expected number of citations increases by 1.48% as variety increases by 1 standard deviation (column 1), decreases by 2.45% as balance increases by 1 standard deviation (column 3), and increases by 5.77% as disparity increases by 1 standard deviation. Squared terms are subsequently added to test the non-linearity in these relationships. On the one hand, the square terms of variety and disparity are significant, suggesting nonlinear relationships. On the other hand, the squared term of balance is insignificant, suggesting a simply linear relationship. Fig. 1B plots the estimated number of citations with variety, balance, and disparity, based on column 2, 4, and 6 in Table 3, respectively. Again, for these estimations, we fix journal fixed effect at 0, international at 0, and all other variables at their mean.

We observe that long-term citations increase at an increasing rate with variety, which is in line with the information processing perspective that cognitive variety is very important for creative and innovative work (Lee, Walsh, & Wang, In Press; Page, 2007; Simonton, 2003). For interdisciplinary research, integrating knowledge from more disciplines contributes to potentially more broadly useful outcomes.

We also observe a negative relationship between balance and citation impact, which is also in line with Uzzi, Mukherjee, Stringer, and Jones (2013) that a paper with both higher novelty and conventionality are more likely to be a top cited paper. In other words, a paper is more likely to be top cited if it is embedded at the core of a discipline (drawing most of its prior knowledge/references from one discipline) while at the same time borrows some knowledge from some remote disciplines. However, the reason for this negative association between long-term citations and balance is still unclear. On the one hand, it could be that interdisciplinary research driving evenly by different disciplinary logics is more likely to fail in integrating these logics into something useful. Therefore, having one disciplinary core and simultaneously borrowing knowledge from other disciplines is a more effective research strategy, compared with drawing knowledge evenly from multiple disciplines. On the other hand, it could be that the current science system is biases against balanced interdisciplinary research. There are anecdotes that balanced interdisciplinary research which truly transcend disciplinary boundaries is difficult to evaluate and more likely to be unnoticed, simply because most scientists are trained within a discipline and unable to realize its value, although such balanced interdisciplinary research is very novel and broadly useful.

In addition, we observe that the number of citations increases with disparity but at a decreasing rate. This is in line with the combinatorial novelty literature that combining more remote disciplines is more novel than combining neighboring disciplines (Lee et al., In Press; Uzzi et al., 2013). Furthermore, there is a rather complex dynamics between novelty and impact. On the one hand, novelty is important for generating impact. On the other hand, a highly novel paper might not be useful or helpful for other scientists to further build on it, and therefore would fail to generate high impact (Latour & Woolgar, 1986; Merton, 1973; Whitley, 2000). We do observe that that the marginal return from disparity is decreasing. It's

possible that the effect of disparity on long-term citations might turn into a negative one after certain point, but this threshold is about six standard deviations above the mean, which is beyond the maximum disparity value in our data.

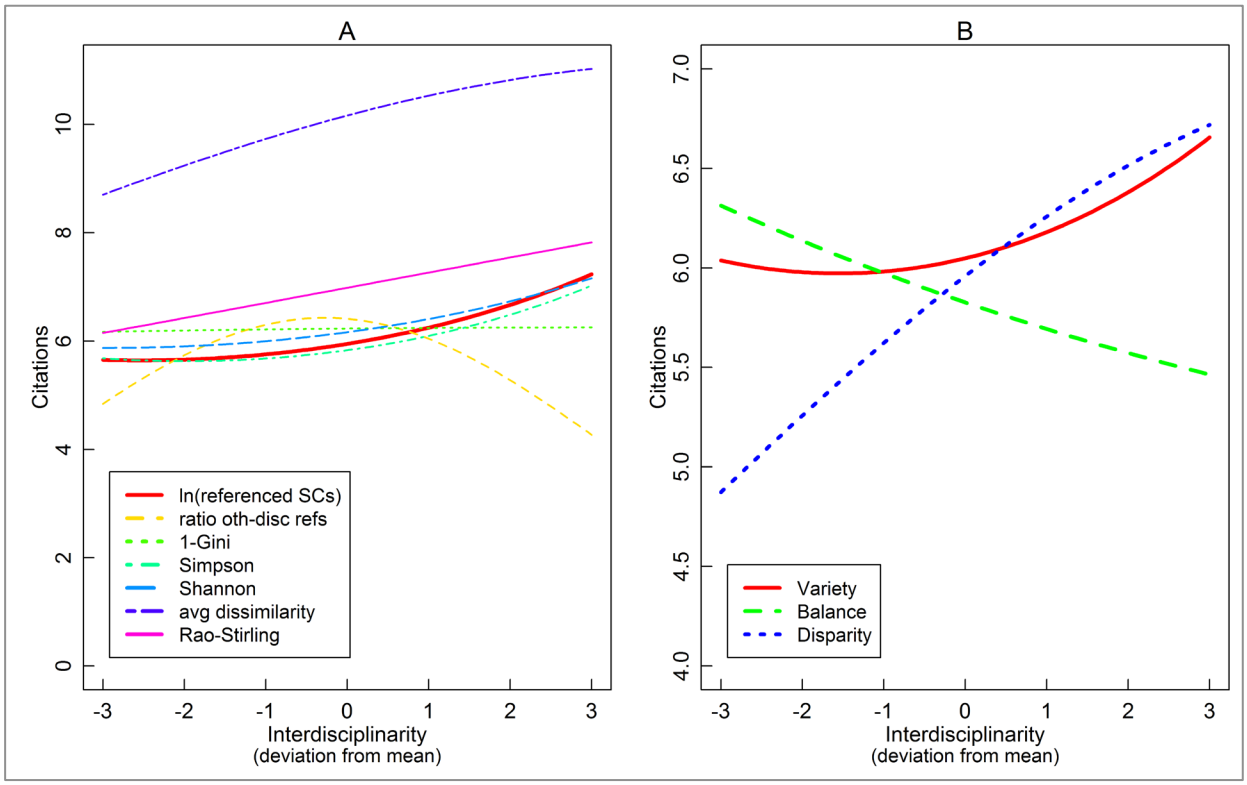


Figure 1. Interdisciplinarity and citations. Data sourced from Thomson Reuters Web of Science Core Collection.

Conclusions

This paper studies three different aspects of interdisciplinarity and investigates their distinct relationships with citation impact. The factor analysis extracts three main factors underlying various interdisciplinarity measures, and these three factors correspond to variety, balance, and disparity. Regression analysis further uncovers their different relationships with long-term citation impact: citations (1) increase at an increasing rate with variety, (2) decrease with balance, and (3) increase at a decreasing rate with disparity.

This paper contributes to future interdisciplinarity research and science policy. First, we advocate the idea of using different interdisciplinarity measures in different contexts. This paper demonstrates that various interdisciplinarity measures bear non-identical relationships with citation impact. Interdisciplinarity is an abstract and multidimensional concept, and different aspects of interdisciplinarity may (1) respond to certain individual, team, or institutional factors in completely different ways, and (2) have unique consequences in terms of usefulness or impact. Furthermore, various theories which might shed light on interdisciplinarity research have their own unique focuses. For example, the information processing perspective focuses on cognitive variety, while the combinatorial novelty literature emphasizes disparity. Therefore, it's important to choose a suitable interdisciplinarity measure consistent with the invoked theory and focal research question.

Table 3. Fixed effects Poisson models: interdisciplinarity and long-term impact (N = 646223).

	Citations							
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
ln(authors)	0.1588***	0.1586***	0.1600***	0.1600***	0.1590***	0.1586***	0.1578***	0.1575***
	(0.0105)	(0.0105)	(0.0106)	(0.0106)	(0.0110)	(0.0110)	(0.0107)	(0.0107)
International	-0.0009	-0.0008	-0.0013	-0.0013	-0.0025	-0.0025	-0.0023	-0.0022
	(0.0130)	(0.0130)	(0.0130)	(0.0130)	(0.0135)	(0.0135)	(0.0133)	(0.0133)
ln(pages)	0.4054***	0.4055***	0.4022***	0.4019***	0.3958***	0.3963***	0.3965***	0.3965***
	(0.0295)	(0.0295)	(0.0295)	(0.0294)	(0.0301)	(0.0302)	(0.0300)	(0.0300)
ln(refs)	0.3021***	0.3013***	0.2868***	0.2871***	0.3056***	0.3045***	0.2855***	0.2836***
	(0.0078)	(0.0077)	(0.0105)	(0.0105)	(0.0082)	(0.0083)	(0.0118)	(0.0119)
Variety	0.0148*	0.0162*					0.0137^{+}	0.0154^{+}
	(0.0061)	(0.0064)					(0.0078)	(0.0083)
Variety^2		0.0052*						0.0044^{+}
		(0.0026)						(0.0026)
Balance			-0.0245**	-0.0241**			-0.0194^{+}	-0.0194^{+}
			(0.0074)	(0.0073)			(0.0106)	(0.0108)
Balance^2				0.0009				0.0021
				(0.0033)				(0.0030)
Disparity					0.0577***	0.0535***	0.0528***	0.0488***
					(0.0075)	(0.0074)	(0.0088)	(0.0087)
Disparity^2						-0.0045^{+}		-0.0036
						(0.0025)		(0.0025)
Journal fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Log pseudolikelihood	-8642990	-8642683	-8642595	-8642588	-8629711	-8629503	-8628738	-8628365
χ^2	2946***	2957***	2967***	2961***	4450***	4438***	4552***	4807***

Cluster-robust standard errors in parentheses.
*** p<.001, ** p<.01, * p<.05, $^{+}$ p<.10.
Data sourced from Thomson Reuters Web of Science Core Collection.

Second, this paper suggests a more refined policy agenda for encouraging interdisciplinary research. This paper pushes forward the research on the relationship between interdisciplinarity and scientific impact: from a dichotomous question of whether interdisciplinary research draws higher impact towards a more complicated question about differentiated dynamics underlying different aspects of interdisciplinarity. Answers to this more complicated question is also important for more effective science policies. As science increasingly deals with boundary-spanning problems, various policy and funding initiatives have been developed to encourage interdisciplinary research, such as the US National Science Foundation (NSF) solicited interdisciplinary programs, the US National Institutes of Health (NIH) common fund's interdisciplinary research program, European Research Council (ERC) synergy grants, and UK Research Councils' cross-council funding agreement. However, interdisciplinarity is an abstract and multidimensional concept, and nuanced understanding of these different dimensions and their consequences are important for effective policies. Specifically, the positive relationship between variety and citation impact demonstrates the benefits of cognitive variety for creative work. Therefore, policy and funding initiatives can encourage research across more disciplinary boundaries and integrating knowledge from more disciplines. Furthermore, the positive relationship between disparity and citation impact also suggests potential improvements from encouraging interdisciplinary research across more remotely connected disciplines. However, since the positive marginal effect is decreasing, the policy might not want to push too far. It's possible that disparity effect on citations might turn into a negative one when the disparity is too high, that is, integrating disciplines too far apart may fail to find a common ground to produce something useful. In addition, the negative relationship between balance and citation impact may suggest that the most effective interdisciplinary research strategy in terms of generating impact is to have one disciplinary core and simultaneously borrow knowledge from some other disciplines, instead of drawing knowledge evenly from multiple disciplines without a disciplinary core. It's possible that research driving evenly by different disciplinary logics fails to integrate these logics into something useful. On the other hand, this might also suggest that balanced interdisciplinary research is biased against in the current discipline-based science system, in which scientists are mostly trained within a single discipline and therefore fail to realize the value of balanced interdisciplinary work which truly transcends interdisciplinary bounties. However, further research is required to better understand this problem. Specifically, to claim the bias against balanced interdisciplinary research, we need to estimate the unbiased should-be scientific impact first and then compare it with the observed citations. To recommend policies encouraging unbalanced instead of balanced interdisciplinary research, we would also need to test the usefulness or value of the papers directly, instead of only examining citation counts.

Acknowledgments

The authors thank You-Na Lee, Diana Hicks, Paula Stephan, and Reinhilde Veugelers for their helpful comments and suggestions.

References

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica, 52*(3), 701-720.

Hall, B.H. & Ziedonis, R. H. (2001). The patent paradox revisited: an empirical study of patenting in the US semiconductor industry, 1979-1995. *Rand Journal of Economics, 32*(1), 101-128.

Harrison, D.A. & Klein, K.J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review, 32*(4), 1199-1228.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents R&D relationship. *Econometrica, 52*(4), 909-938.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

Hottenrott, H. & Lopes-Bento, C. (In Press). Quantity or quality? Knowledge alliances and their effects on patenting. *Industrial and Corporate Change*.

Judge, T. A., Cable, D.M., Colbert, A.E., & Rynes, S.L. (2007). What causes a management article to be cited: Article, author, or journal? *Academy of Management Journal, 50*(3), 491-506.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141-151.

Larivière, V. & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126-131.

Latour, B. & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Lee, Y.-N., Walsh, J.P., & Wang, J. (In Press). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*.

Levitt, J. M. & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science and Technology, 59*(12), 1973-1984.

Leydesdorff, L. & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics, 5*(1), 87-100.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago, IL: University of Chicago Press.

Page, S. E. (2007). *The difference: how the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Porter, A. & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics, 81*(3), 719-745.

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.

Raiche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 9*(1), 23-29.

Rinia, E., van Leeuwen, T.N., van Vuren, H.G., & van Raan, A.F.J. (2001). Influence of interdisciplinarity on peer-review and bibliometric evaluations in physics research. *Research Policy, 30*(3), 357-361.

Simonton, D.K. (2003). Scientific creativity as constrained stochastic behavior: the integration of product, person, and process perspectives. *Psychological Bulletin, 129*(4), 475-494.

Somaya, D., Williamson, I.O., & Zhang, X.M. (2007). Combining patent law expertise with R&D for patenting performance. *Organization Science, 18*(6), 922-937.

Steele, T.W., & Stier, J. C. (2000). The impact of interdisciplinary research in the environmental sciences: a forestry case study. *Journal of the American Society for Information Science, 51*(5), 476-484.

Stirling, A. (1998). On the economics and analysis of diversity. *SPRU Electronic Working Papers, 28*.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707-719.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*(6157), 468-472.

Wagner, C.S., Roessner, J.D., Bobb, K., Klein, J.T., Boyack, K.W., Keyton, J., . . . Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics, 5*(1), 14-26.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851-872.

Whitley, R. (2000). *The Intellectual and Social Organization of the Sciences* (2nd ed.). Oxford, UK; New York, NY: Oxford University Press.

The Evaluation of Scholarly Books as a Research Output. Current Developments in Europe

Elea Giménez-Toledo[1], Jorge Mañana-Rodríguez[1], Tim Engels[2], Peter Ingwersen[3], Janne Pölönen[4], Gunnar Sivertsen[5], Frederik Verleysen[6] and Alesia Zuccala[7]

{elea.gimenez, jorge.mannana}@cchs.csic.es
[1]Centro de Ciencias Humanas y Sociales, ÍLIA Research Group, CSIC, Albasanz Street, 28037, Madrid (Spain)

[2]*tim.engels@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Political and Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp (Belgium); Antwerp Maritime Academy, Noordkasteel Oost 6, B-2030 Antwerp (Belgium)

[3]clb798@iva.ku.dk, [7]*spl465@iva.ku.dk*
[3, 7]Royal School of Library and Information Science, University of Copenhagen (Denmark)

[4]*janne.polonen@tsv.fi*
Publication Forum, Federation of Finnish Learned Societies, Snellmaninkatu 13, 00170 Helsinki (Finland)

[5]*gunnar.sivertsen@nifu.no*
NIFU Nordic Institute for Studies in Innovation, Research and Education, PO Box 5183 Majorstuen, 0302 Oslo (Norway)

[6]*frederik.verleysen@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Political and Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp (Belgium)

Abstract

The relevance and value of books in scholarly communication from both sides, the scholars who chose this format as a communication channel and the instances assessing the scholarly and scientific output is undisputed. Nevertheless, the absence of worldwide comprehensive databases covering the items and information needed for the assessment of this type of publication has urged several European countries to develop custom-built information systems for the registration of books, weighting procedures and funding allocation practices enabling a proper assessment of books and book-type publications. For the first time, these systems make the assessment of books as a research output feasible. This paper resumes the main features of the assessment systems developed in five European countries / regions (Spain, Denmark, Flanders, Finland and Norway), focusing on the processes involved in the collection and processing of data on books, weighting, as well as their application in the context of research funding assessment.

Conference Topic

Science policy and research assessment and/or University policy and institutional rankings

Introduction

Scholarly books are key for the communication of research outputs in Social Sciences and Humanities (Hicks, D., 2004; Thompson, 2002; Engels, Ossenbklok & Spruyt, 2012). At the same time, performance-based assessment and funding allocation systems, as well as evaluation exercises at an individual level are widespread throughout Europe, affecting all instances of universities and research institutions (Hicks, D., 2012; Frølich, N., 2011). Despite developments such as Book Citation Index (Adams & Testa, 2011) there still exist a clear need for comprehensive databases collecting 'quality' indicators for books and book publishers. Quality in books is a multi-faceted concept and translating it into indicators is a

difficult task, in many occasions closely oriented to the specific research and assessment policies of each country. This diversity at the policy level is matched by an intrinsic heterogeneity of scholarly books themselves (e.g. disciplines, languages, formats, peer review and other editorial standards, etc.). In the past, the vast variety of books has made their reliable and comprehensive registration notoriously difficult and, consequently, their inclusion in research assessments unrewarding. By introducing the information systems presented in this paper, five European countries/regions have sought to redress the balance.

Objectives

The aim of this paper is to compare different approaches for assessing books across Europe. To do so, the context of each assessment exercise -where books evaluation occurs- is presented. The existence of valid peer review processes, the prestige of book publishers or the division in tiers according to the quality of the communication channel and the specific features of each discipline are some of the elements on which Spain, Denmark, Flanders, Finland and Norway have developed assessment systems for books. These developments are the result of applied research and also the object of a research-in-progress. This paper summarizes the main features of the current registration and assessment systems developed in the five countries in their present state. After a detailed discussion of each system, preliminary conclusions are presented, as well as a perspective on possible future developments.

Results

Scholarly Book's evaluation practices at the micro level

Spain

Scholarly books are taken into account in various assessment processes on the research outputs of scholars. As an example, both ANECA and CNEAI (Spanish assessment agencies) include various aspects of books and book publishers among their assessment criteria at the individual level. One of them is the prestige of the publisher (the latest, being CNEAI Resolution of November 26, 2014, but included as quality criteria various years backwards). Given the lack of specific data on the prestige of book publishers, the Research Group on Scholarly Books (ÍLIA) at CSIC developed Scholarly Publishers Indicators (SPI) on the grounds of the research conducted in previous years (Giménez-Toledo & Román Román, 2009). SPI ranks the perceived prestige of book publishers in the social sciences and humanities (SSH), both Spanish and non-Spanish, according to the scores resulting from an extensive survey to Spanish lecturers, researchers and scholars specializing in all fields of SSH. The system is based on more than 3,000 usable responses in 2012 and almost 3,000 in 2013. The responses are given to the question of which are the first, second or third (and from first to tenth in the 2013 edition) most prestigious book publishers in the responder's field; only specialists with positive assessment of their research are susceptible of being included among the respondents. Once collected, the responses are summarized using a simple weighting algorithm based on the share of scores in each position (1^{st}, 2^{nd}, etc.). The results are summarized in an indicator: ICEE. This indicator serves as a ranking item, both at the general level and specifically for each discipline, since the assigned weights are related to each discipline's distribution of scores (Giménez-Toledo, Tejada-Artigas & Mañana Rodríguez, 2012). The weighting procedure involves no arbitrary intervention from its designers and permits certain normalization per discipline. The ranking is publicly available at (http://ilia.cchs.csic.es/SPI/) and the users can access both discipline-level and general rankings for Spanish and non-Spanish publishers.

The main advantage of this system is the wide population on which it is based (more than 11,000 experts), while the main disadvantage lies in the difficulty to control for possible bias in the surveying process. The ranking was first used for assessment purposes in 2013 and is increasingly being included in the current evaluation framework as a reference for the assessment of SSH books and book chapters, together with other criteria. It is important to note that SPI is a reference tool for assessment exercises. It is meant to inform, not to perform, the research evaluation.
SPI also includes interactive charts as well as a 'specialization profile' of publishers obtained from the DILVE database (collecting the editorial production of Spanish publishers). Specialization is a point where evaluation agencies may focus their attention. In progress is the research into the use of different peer review systems with the use of surveys to book publishers as well as information about the transparency of their websites. These are qualitative indicators which aim is to serve as supporting information in the assessment processes.

Book's evaluation practices at meso or macro-level

Denmark

The performance indicator model (BFI/BRI, the Bibliometric Research Indicator) was started up in 2009. For each year 68 groups of academics selected by the Danish Research Agency from the Danish universities list all available knowledge resources and assign points to peer reviewed journals, publishers and conferences that publish scientific material authored by Danish academics from the previous year. Each of the 68 groups represents an academic field or specialty. The bibliometric research indicator takes into account published peer reviewed research and review articles, monographs as well as anthology and proceedings papers published by the Danish research institutions, which provide the input metadata for the system. In the period 2008–2012 proceedings (and anthology) papers were assigned .75 points. Journal articles received 1.0 point in Level 1 journals and 3.0 points 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 2013 proceedings papers and articles receive similar points as journal articles, depending on the level of the conference or publisher, as assessed by the relevant academic group. Monographs are assessed according to two publisher levels, Level 1 (5 points) and Level 2 (8 points). Anthology papers and chapters receive 0.5 and 2 points depending on publisher level. For each document the points are fractionalized (min 0.1) according to number of collaborating universities, including non-Danish universities. The model encourages collaboration by multiplying the institutional fraction by 1.25. The previous year's cumulated points per university is used to distribute a substantial portion (in 2013 it was 25%) of public basic 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, Social Sciences, Natural Sciences or Medicine/Health sciences via the Danish Research Agency's web page (https://bfi.fi.dk/). The intermediate or more detailed publication point distributions and document lists per unit and department will be publicly accessible from 2015. This is in difference to Norway where no multiplication of fractions takes place and all the documents and their point assignments are transparent as well as publicly accessible through an open access database. In the Finnish system and in Belgium the Flemish BOF-key applies whole counting at the institutional level (Debackere & Glänzel, 2004; Engels, Ossenblok & Spruyt, 2012). The output of the Danish BRI system can, as a spin-off, be used for assessment purposes. See also Ingwersen & Larsen (2014).

Flanders (Belgium)

The Flemish Academic Bibliographic Database for the Social Sciences and Humanities ('Vlaams Academisch Bestand voor de Sociale en Humane Wetenschappen', or VABB-SHW) has been developed to allow for the inclusion of the peer reviewed academic publication output in the Social Sciences and Humanities (SSH) in the regional performance-based research funding model. As such, in 2015 the VABB-SHW accounts for 6.62% of the University Research Fund (or BOF), distributing over 150 million euro per year over the five universities. As the BOF-key is also re-used for the distribution of other research funding, the actual impact of the VABB-SHW is even greater. In a secondary role, the VABB-SHW supports research assessments at various levels. As all information in the VABB-SHW is available to both the universities and the Flemish national science foundation (FWO), data is harvested and integrated into each institution's repository. In a third role, the VABB-SHW's comprehensive publication coverage (peer reviewed or otherwise) allows for in-depth research on publication practices in the SSH (Engels, Ossenblok, & Spruyt, 2012; Verleysen, Ghesquière, & Engels, 2014). The database covers the comprehensive publication output of academic research in 16 SSH disciplines and 3 general categories. Three types of book publications are included: 1° monographs, 2° edited books, 3° book chapters, weighted 4, 1 and 1 for the funding model, respectively. Journal articles also receive a weight of 1 and proceedings papers a weight of 0.5. No prestige levels are distinguished. For funding calculation, a ten-year timeframe is used. For research purposes, coverage extends back to the year 2000. For books, four aggregation levels are in use: 1° publisher names (as collections of ISBN-roots), 2° book series, 3° books published in Flanders and labeled as Guaranteed Peer Reviewed Content (GPRC-label (Verleysen & Engels, 2013), and 4° individual books identified as peer reviewed by the Authoritative Panel ('Gezaghebbende Panel' or GP, a committee of full professors installed by the government and responsible for decisions regarding the content of the VABB-SHW). The information system is fed through a yearly upload (May 1st) of all SSH publications from the two preceding years newly registered in the five universities' academic bibliographies. Data is managed at the Flemish Centre for R&D monitoring (ECOOM), University of Antwerp, through its custom-built Brocade library services (http://en.wikipedia.org/wiki/Brocade_Library_Services). Each individual publication receives a unique identifier, contributing to maximum granularity and reliability of the data both for funding calculation as well as for retrieval and research. Consolidation processes making use of algorithmic identification allow a systematic de-duplication of records that are submitted more than once. Publications are identified algorithmically at the publisher, series or journal level by their ISBN-prefix or ISSN. Each year all new publishers, series, books and journals are classified by the Authoritative Panel as peer reviewed and presenting new content (or not). At the public interface www.ecoom.be/en/vabb, online access is provided to the database itself, lists of publishers, journals and series, explanation of procedures, FAQ's, and background information.

Finland

In Finland, the use of publications in the performance based funding model is based on two components: the publication metadata consisting of the entire output of universities, and a quality index of outlets. Universities have their own registries of publications, including peer-reviewed and non-peer-reviewed articles in journals, conferences and anthologies, as well as monographs. Universities report their publication data, with full bibliographic details, once a year to the ministry of education and culture (Puuska 2014). The publication data is processed (including deduplication) at CSC - IT Centre for Science, which is a company owned by the ministry. The bibliographic details of publications are matched against the list of serials, conferences and book publishers classified in three quality levels by 23 expert panels

coordinated by the Federation of Finnish Learned Societies (FFLS). This quality index of outlets is called Julkaisufoorumi (JUFO) -luokitus (Publication Forum Classification). The universities' publication metadata collected by the ministry is known as OKM-julkaisuaineisto (MinEdu publication data).

In the Publication Forum classification, published for the first time in 2012, the level 2 comprises 20 % of the leading serials and conferences and 10% of the leading book publishers (Auranen & Pölönen, 2012). Most peer-reviewed outlets belong to the level 1, and those that fail to meet the criteria of scientific publication channel are listed as the level 0. For serials there is also a level 3, in which are classified 25% of the level 2 titles, but in the funding model it is not differentiated from the level 2. Updated classifications have been published in the beginning of 2015. In the new classification, as in Denmark, the level 2 serials and conferences comprise at most 20% share of the world production of articles in each panel's field. The level 3 was added also for book publishers. The new classifications will be applied on articles and books published in 2015. The classification of book publishers is used specifically to determine the level of monographs and articles in anthologies when the publication does not come out in a book series or the series has not been classified. The main rule is that the Finnish book series are classified, while those of foreign book publishers are not classified separately.

In the current funding model for 2015 and 2016, which still uses the 2012 Publication Forum classifications, 13% of all budget-funding is allocated on basis of publications (Ministry of Education and Culture, 2014). The peer-reviewed articles in journals, conferences and anthologies published in the level 0 channels will have the weighting coefficient 1, those of the level 1 have the coefficient 1.5, and for the level 2 and 3 channels the coefficient is 3. The weighting coefficient of non-peer-reviewed (scholarly, professional and general public) articles is 0.1 regardless of outlet. Weighting coefficient of peer-reviewed monographs is four times higher than that of articles: 4 in the level 0, 6 in the level 1, and 12 in the level 2. For non-peer-reviewed monographs, as well as all edited volumes, the weight is 0.4. There is no fractionalization of co-publications at the institutional or author level. The Ministry has instituted a working-group to determine the weights and calculation method of publications used in the funding model from 2017 onwards.

The MinEdu publication data, which covers Finnish universities output since 2010, is openly available through Vipunen-portal (www.vipunen.fi) for statistics, as well as Juuli-portal (www.juuli.fi) for browsing the publication information. The quality index of outlets is openly available on the Publication Forum website (www.tsv.fi/julkaisufoorumi).

Norway

The Norwegian model (Sivertsen, 2010; Sivertsen & Larsen, 2012) consists of three main elements: 1) A national database containing comprehensive and unified bibliographic metadata for the peer reviewed literature in all areas of research; 2) a publication indicator making field-specific publishing traditions comparable in the same measurement; and 3) a performance based funding model.

The national database is called CRISTIN (Current Research Information System in Norway). It is shared by all research organizations in the public sector: universities, university colleges, university hospitals, and independent research institutes. The institutions provide quality-assured and complete bibliographic about articles in journals and series (ISSN), articles in books (ISBN), and books (ISBN) that can be included according to a definition of peer-reviewed scholarly literature.

The indicator is based on a division of publication channels (journals, series, book publishers) in two levels: level 1 and level 2. Level 2 contains the most selective international journals, series and book publishers and may not contain more than 20 per cent of the publications

worldwide in each field of research. Articles in journals and series are given 1 point on level 1 and 3 points on level 2. Articles in books (with ISBN only) are given 0.7 1 points on level 1 and 1 point on level 2. Monographs are given 5 points in level 1 and 8 points on level 2. The points are fractionalized in the level of institutions according to the institution's share of contributing authors.

Although less than two per cent of the total expenses reallocated by the use of the indicator in Norway, it has attracted a lot of attention among researchers and resulted in increased productivity (Aagaard et al., 2014).

Conclusions

One of the first conclusions which stand out is the lack of use of citation metrics in any of the five systems. This might be the result of a lack of fit, lack of acceptance or the irrelevance as a quality indicator for books of the traditional measures for journals. Another element is the incomprehensiveness for many scholarly fields of the current citation indexes. Equally remarkable is the clear convergence as regards criteria and procedures among the Nordic countries and Flanders, not only in the registration of books, but also in the funding and/or assessment policies making use of book data. For assessments, in Northern Europe data is used mainly at the institutional level, despite its collection and registration being nationally coordinated in the context of a performance-based research funding system. This is clearly not the case for Spain, where data is used for assessments at the individual level, while university budgets are not calculated in a performance-based, centralized system. Also, the different policies show great divergences regarding the much higher weight given to scholarly books in the Nordic systems, while in Spain the tendency is just the opposite (more weight is given to papers than is to books). It is also remarkable that the most frequently used aggregation level is that of book publishers, although in the case of Flanders the Guaranteed Peer Reviewed Content-label allows for the inclusion of individual books in the regional system as well, while Finland currently counts with a Peer Review Mark similar to the already mentioned, making feasible that possibility. This involves that the expected coherence in the practices underlying to the concept of quality is sufficient at the level of book publishers, since the congruent use of this level of aggregation (from which the positioning in tiers of each individual contribution is derived) is common to all systems analyzed. Nevertheless, future developments may well see a stronger interest in the registration of book data at lower aggregation levels as well (e.g. that of the book series), as this evidently implies a more fine-grained approach to the comprehensive registration and the validation in assessments of books. In Spain, that specific level of aggregation (book series) is the object of a current initiative by UNE (University Presses Union) in collaboration with three research teams.

Finally, it will be interesting to see whether the on-going internationalization of research and the growing collaboration between scholars worldwide will contribute to a greater harmonization at the European level of the assessment systems for books and book publishers. Such developments could indeed provide scholars with new opportunities to assert the (often under-rated) value of their books, although some hypotheses regarding the role of the book in the scholarly communication shall be addressed in the close future.

Acknowledgements

This research is partially the result of the project 'Evaluación de editoriales científicas (españolas y extranjeras) de libros en Ciencias Humanas y Sociales a través de la opinión de los expertos y del análisis de los procesos HAR2011-30383-C02-01 (Ministerio de Economía y Competitividad. Plan Nacional de I+D+I).

Table 1. Comparison of the main features of the information systems for the assessment of books.

ITEM	SPI	BFI/BRI*	VABB-SHW	MinEdu Data/JUFO	CRISTIN
Country	Spain	Denmark	Flanders	Finland	Norway
Reasons for its development	Assessment at the individual level and research evaluation (unknown uses at institutional level)	Research funds allocation among universities and measures of research activities at institutional levels.	Inclusion of the peer reviewed scholarly publication output in the regional performance-based research funding model.	Funding allocation, research information and quality promotion.	Research information and fund allocation in the public sector. National statistics.
Object of study/ aggregation level	Book publishers / specialization from book-level information.	Book publishers, books and book parts (anthologies); journal articles and proceeding papers.	Book publishers, book series, GPRC**-labeled books published in Flanders and individual books assessed by the Authoritative Panel.	Book publishers and monographic series / peer reviewed monographs and articles in books at university level.	Bibliographic references to all scholarly publications in books, book articles and journal papers.
Stage	Already published and applied in assessment.	Already published and applied in assessment and funding since 2009.	Applied for funding allocation and institution-level assessment since 2010.	Published in 2012 and applied in funding since 2015.	Applied in assessment and funding since 2005.
Coverage	All Spanish and non-Spanish book publishers mentioned by experts in each field.	All scholarly publishers worldwide with publications from Danish scholars since 2009.	The comprehensive peer reviewed publication output of academic research in the Social Sciences and Humanities since 2000.	National and international scholarly book publishers and Finnish book series	All scholarly publishers worldwide with publications from Norwegian scholars since 2004.
Information feeding the system	Survey to experts and book publishers / database analysis.	Metadata for scholarly publications from all Danish universities.	Yearly upload from the academic bibliographies of the five Flemish universities, of all newly registered publications of the previous two years.	Metadata for universities' scholarly publications and new additions suggested by researchers	Metadata for scholarly publications from all Norwegian institutions in (CRISTIN).
Information processing	Votes from respondents are summarized in the ICEE indicator. DILVE database is statistically analyzed. Surveys to book publishers are summarized. Done by ILIA research group (CSIC).	Quality level assessments of publishers and journals by 67 topical peer groups plus a central coordination council, providing authoritative lists from which each publication is assigned a score by the system.	Data input from the universities processed by ECOOM / University of Antwerp Scientific steering and assessment of publication channels by a central Authoritative Panel.	In order to assign weight to universities' publications in the funding model, the metadata of publications is collected and matched against the list of serials, conferences and book publishers classified in quality levels by 23 panels.	Input from the institutions of metadata for individual publications is connected to a centrally monitored dynamic register of approved scholarly publication channels (journals, series, and book publishers)
Operative results	Ranking of book publisher's prestige / specialization charts / peer review info.	Annual number of publications and number of publication points per university and per larger academic topic.	A growing database of 125,000 scholarly peer reviewed and other publications. Publicly available lists of assessed book publishers, book series, journals and conference proceedings.	List of quality-classified outlets and database of universities' all publications from 2011 that can be analyzed by type, field and outlet.	A database of so far 70,000 scholarly publications that can be analyzed by type, field, language, institution, and publication channel
Use for research assessment and aggregation level	Used at the individual level by ANECA and CNEAI, two Spanish assessment agencies.	Funding allocation in the following year; Institutional level; also used as promotion or 'extras' factor (local incentive). Individual level in the future.	Funding allocation to five universities; support of internal assessments at individual universities, and assessments by the Flemish national science foundation (FWO)	Funding allocation to universities; internal assessment and planning at universities (also funding allocation); use for assessment at individual level is discouraged.	Funding allocation, stats for field and/or institution research evaluation, administrative information at institutions and annual reports.
Public availability	Yes (from 2012)	Yes (from 2015)	Yes	Yes	Yes (from 2004)
Book / paper weighting	Approx. 1 to 3 (as defined by assessment agencies, but not by SPI)	From 5 to 8 and from 0.5 to 2 (anthology items) and from 1 to 3.	From 4 to 1 and from 1 to 0.5	From 0.4 to 12 and from 0.1 to 3.	From 8 to 3 and from 3 to 1.

* BFI/BRI = Bibliometric Forskningsindokator / Bibliometric Research Indicator, **GPRC = Guaranteed Peer Reviewed Content

References

Aagaard, K., C.W. Bloch, & J.W. Schneider. (2014). Impacts of Performance-based Research Funding Systems: the case of the Norwegian Publication Indicator. *Research Evaluation,* (forthcoming).

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* (pp. 13–18). Durban, South Africa: ISSI, Leiden University and University of Zululand.

Auranen, O., & Pölönen, J. (2012). *Classification of scientific publication channels: Final report of the Publication Forum project (2010–2012)*. Federation of Finnish Learned Societies: http://www.tsv.fi/files/yleinen/publication_forum_project_final_report.pdf.

Debackere, K., & Glänzel, W. (2004). Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key. *Scientometrics*, *59*(2), 253-276.

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.

Frølich, N. (2011). Multi-layered accountability. Performance-based funding of universities. *Public Administration*, *89*(3), 840-859.

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation*, *22*(1), 64–77. doi:10.1093/reseval/rvs036.

Giménez-Toledo, E. & Román-Román, A. (2009). Assessment of humanities and social sciences monographs through their publishers: a review and a study towards a model of evaluation. *Research Evaluation*, 18(3), 201-213.

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

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41(2), 251-261.

Ingwersen, P. & Larsen, B. (2014). Influence of a performance indicator on Danish research production and citation impact 2000–12. *Scientometrics*, DOI 10.1007/s11192-014-1291-x.

Ministry of Education and Culture (2014). *Greater incentives for strengthening quality in education and research: A proposal for revising the funding model for universities as of 2015*: http://www.minedu.fi/export/sites/default/OPM/Julkaisut/2014/liitteet/tr07.pdf?lang=fi.

Puuska, H.-M. (2014). *Scholarly Publishing Patterns in Finland: A comparison of disciplinary groups*. University of Tampere.

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*(1), 22–28.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics, 91*(2), 567- 575.

Thompson, J.W. (2002). 'The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship'. *Libri*, *52*, 121–36.

UNE. (2014). España crea un sello de calidad para reconocer la excelencia científica del proceso editorial de las colecciones publicadas por las universidades. http://www.une.es/Ent/Items/ItemDetail.aspx?ID=9610

Verleysen, F. T., Ghesquière, P., & Engels, T. C. E. (2014). The objectives, design and selection process of the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB-SHW). In W.Blockmans, et al., (Eds.). *The use and abuse of bibliometrics* (pp. 115-125). Academiae Europaea; Portland Press.

Verleysen, F.T. & Engels, T.C.E. (2013). A label for peer-reviewed books. *Journal of the American Society for Information Science and Technology*, *64*, 428-430.

Zuccala, A., Costas, R., & van Leeuwen, T.N. (2010). Evaluating research departments using individual level bibliometrics. In: *Eleventh International Conference on Science and Technology Indicators* (p. 314).

Publications or Citations – Does it Matter? Beneficiaries in Two Different Versions of a National Bibliometric Performance Model, an Existing Publication-based and a Suggested Citation-based Model

Jesper W. Schneider

jws@ps.au.dk
The Danish Centre for Studies in Research and Research Policy, Department of Political Science, Aarhus University (Denmark)

Abstract

The paper discusses the adoption of the Norwegian Publication Model in a Danish context and examines arguments for supplementing or substitution the current mechanism where reward is based on publication activity with one based on citations. Based on national publication data from 2009 from the Danish model, belonging to the science and technology research area, and corresponding citation data, we examine the Danish universities' relative input when it comes to publications and subsequently examine the relative output from these publications, i.e., the "returns on investment" from the model, either the current publication points, or the alternative, citations. Findings support the claims that high-performing units would benefit more from a citation-based approach, but at the same time also show, contrary to what was conjectured, that in the present case the same university also benefits the most from the current publication model. Based on the findings, we discuss the publication versus citation-based models, or hybrids between them, and argue that citation-based models in performance-based funding context are harder to influence and most likely will support already existing cumulative advantages.

Conference Topic

Science policy and research assessment

Introduction

In recent decades several countries have introduced performance-based research funding among their universities (Hicks, 2012). The performance-based research funding systems (PRFS) vary considerably between countries, from panel-based peer review evaluations, to systems based on citation or publication metrics, or various hybrids of these three basic forms (see Hicks, 2012). Generally, peer review systems are considered superior to systems based on bibliometric indicators (see Gläser & Laudel, 2007). Nevertheless, large-scale panel evaluations are very expensive, and several *post hoc* comparisons between panel results and citation metrics, for example from the UK Research Assessment Exercises, suggest that the latter could be an effective, and cost-effective, supplement or even substitute to peer reviews (e.g., Oppenheim, 1996; Moed, 2008). Among PRFS based on bibliometric indicators, citation-based systems are considered by some to be superior due to the assumption that citation indicators to some extent are able to measure aspects of research quality by focusing on impact (Gläser & Laudel, 2007). But citation indicators also have obvious deficiencies especially when implemented in PRFS which in principle are supposed to cover all fields of research (Schneider, 2009). It is well-known that citation indicators are not equally valid across all fields of research and even where relevant, coverage in the citation databases is also restricted (Moed, 2005). Consequently, PRFS based on citation indicators severely restricts the measurable outcome of research basically to journal articles indexed in one of the two major citation databases. But there are other issues with citation indicators which can be considered inadequate when used in PRFS, especially when such systems are supposed to (re)distribute funding on a regular basis, most often annually, and at the same time also give

universities (and their researchers) incentives to improve performance (e.g., Gläser & Laudel, 2007; Schneider, 2009). Citation indicators reflect research done in the past often a considerable number years prior to the actual funding year. It is also very difficult to directly influence citations when conceived of as an incentive system, in fact the well-known cumulative advantages could be detrimental to such an incentive system if it is supposed to be fair for all involved (Merton, 1988). Such features are seen by some as undesirable if PRFS as supposed to cover all research fields with their different publication traditions, and be able to reflect recent research performance in a dynamic model, as well as give transparent behavioural incentives to change performance (Schneider, 2009; Hicks, 2012).

PRFS based on publication activity have been introduced as an alternative to citation-based systems (Butler, 2002; Schneider, 2009). There are some apparent "benefits" with publication-based systems compared to citation-based systems. They can reflect short-term research activity making them more up-to-date when it comes to redistributing funding. In principle they can encompass all desired publication types and they can provide straightforward behavioural incentives. But it is important to emphasise that the two approaches measure different constructs. It would be naïve to suppose that incentives directed at publication behaviour, i.e., quantity and/or supposed status of the publication outlet, encompass the same aspects of perceived "quality" that citation impact is thought to reflect (Schneider, 2009). Experiences from Australia testify to this. In a succession of papers, Linda Butler demonstrated how researchers in Australia responded when funding, at least partially, was linked to publication counts undifferentiated by any measure of supposed "quality" in the early 1990s (e.g., Butler, 2003a; Butler, 2003b). Australian publication output increased considerably with the highest percentage increase in lower impact journals. For a consecutive number of years, this lead to a general drop in overall citation impact for Australia. Since Butler's documentation of the adverse effects, the experience from Australia has stood as a "warning" for what would most likely happen if funding was linked to publication activity. Nonetheless, in the early 2000s a so-called "quality reform" of the higher education sector in Norway introduced a PRFS where publication activity again was linked to funding. The main political intention with the model was in fact to encourage more research activity and thereby also more publication activity, and preferably more international publication activity, in the university sector[1].

The so-called *Norwegian Publication Model* (NPM) is interesting in in relation to PRFS. Obviously, the designers of the NPM were well-aware of the adverse behavioural effects documented in the Australian case. As a consequence, a slightly more sophisticated model was developed (Schneider, 2009; Sivertsen, 2010). A primacy of the model was to reflect the encouragement to publish in international outlets (i.e., international journals and academic book publishers) and at the same time to counter so-called adverse publication effects like the Australian case, where researchers seek to publish more but with less effort. Hence, a differentiated publication model was constructed where publication channels were classified on two levels. Level one comprises in principle all scholarly eligible publication channels, where eligibility criteria are some basic norms such as a standard external peer review process. Level two, is an exclusive number of publication channels, which are deemed to be leading in a field and preferably with an international audience. Level two is exclusive in as much as the number of publication channels designated at any given time to this level should produce roughly one-fifth of the publications produced in a field "world-wide". Correspondingly, three different types of scholarly publications are included in the model: journal publications (articles and reviews), articles in books (contributions to anthologies and

[1] http://www.uhr.no/documents/Rapport_fra_UHR_prosjektet_4_11_engCJS_endelig_versjon_av_hele_oversettelsen.pdf.

conference papers) and books. A two dimensional point system was implemented where the different publication types yield different points within the same level and between the two levels depending on the outlet status. Hence, the basic idea behind this two-tiered classification system is that publications on level two receive more publication points than publications on level one. Finally, publication points are fractioned $1/n$ so that an institution eventually receives $1/n$ points depending on their number of contributing authors.

Eventually the annual sum of publication points for an institution is exchanged for funds, where the exchange rate is determined by the amount of money available for redistribution and the total number of publication points in the system in a given year. A noticeable assumption in the NPM is that publication behaviour, publication activity and publication types across all fields can be treated identically. Consequently, all research fields' eligible research publications are included in the model, which for example means that a level one journal article with one author is worth the same in physics and literature studies. It is assumed that the differentiated point system together with fractionalized counting will level out the major differences in publication behaviour between the fields and also to some extent will discourage researchers to speculate in "easy publications" resulting in a levelling out effect at the aggregate level. Consequently, in the Norwegian PRFS funding is competitive not only between institutions but also across all fields. Hence, the subject composition within and between the research institutions is interesting as performance improvement in one major area, in principle can lead to improved funding at the expense of another major area due to the basic zero-sum situation.

The NPM has recently been "adopted" in several European countries, for example in Denmark, Finland and Flanders (Hicks, 2012; Verleysen, Ghesquière & Engels, 2014). In the present paper we look at the "adoption" of the indicator in Denmark and examine the overall distributional consequences of focusing on publication activity and not impact.

It is important to accentuate that in Norway the publication model was to a large extent developed to support overall political goals, i.e., more international research activity. As it were, Norway's internationalization in research and general citation impact, were considerably lower, than for example Denmark, at the time of the introduction of the model. Since then Norway's international publication output has risen considerably, albeit rise in citation impact has been meagre (e.g., Aagaard, Bloch & Schneider, 2015). Nonetheless, the NPM was developed and implemented with a legitimate goal which to some extent seems to have been achieved seen from the national policy perspective.

During a reform of the Danish research funding system in the mid-2000s it was decided to implement a PRFS officially in order to enlarge competition among universities for funding, although the board of university rectors probably more saw it as management tool that should legitimize their overall research activity to the public (Schneider & Aagaard, 2012). The political process leading to the "adoption" of the NPM in Denmark is complex and documented in Aagaard (2011). It is not totally clear why the choice fell upon the NPM, although its coverage of all areas, transparency and clear incentive system were no doubt deemed viable, yet some actors actually indicated that it would probably be "the one that would cause the least damage" (Aagaard, 2011). Most interesting, contrary to Norway, there were no immediate strategies or goals for research and publication behaviour behind the "adoption" of the NPM in Denmark.

Denmark was the first country to adopt the NPM at a time when the model was still in its infancy in Norway and little empirical evidence of its potential effects was available. The NPM was adopted with very few moderations, as if the model was a one-size fit all package suitable for all contexts. Most notably, the simple two-tiered classification system was kept and considerations about expanding or adapting the classification to a Danish context were not done. Nevertheless, some seemingly minor moderations turned out to be imperative,

including a maximum fractionalization of contributions at 1/10th; but perhaps most important, performance-based publication activity was locked between the major research areas: science and technology, health sciences, social sciences and humanities. Consequently, in the Danish adoption of the NPM, funding is not competitive across areas only within areas. Further, politically it was decided to more or less keep the old annual allocation model between the areas which effectively meant that a publication point, contrary the Norwegian PRRS, have different monetary values across the four main research areas. This is an extremely important deviation from NPM and it gives rise to some questions about the Danish adoption of the NPM, popularly known by the acronym BFI (bibliometric research indicator).

One can argue that the model is transparent, seemingly coherent and all-inclusive when it comes to research areas. All areas are measured with same indicator. But since competition is restricted to within areas and as a consequence publication points have different values across areas, one could also ask why the model still assumes equality of publication practices across areas? And to go further, with the locking of the competition to within areas, there is basically no reason why fields where citation analysis could be a reasonable and indeed preferred indicator could implement such devices either in combination with a publication model or alone. Of course the latter would muddle the overall model, although it would probably satisfy many of the critics of the publication-based model, arguing for more emphasis on impact.

Indeed, the Technical University of Denmark (DTU) has been an ardent critic of the adoption of the NPM in Denmark. A common argument goes: Why implement an incentive model that reward publication activity in international outlets when "we" already do that and do it well? More generally the critics stated that the behavioural goals with the model in Norway were irrelevant in a Danish context, because Denmark, contrary to Norway, has 1) for decades consistently been among the top five highest performing countries when it comes to impact; 2) has consistently four of its eight universities in the top 200 of the Leiden Ranking[2]; and 3) the Danish research system has had a long trajectory of internationalization (e.g., Karlsson & Persson, 2012). According to DTU, what should be procured and rewarded is impact and not publication activity. While the argument is relevant, it is also self-serving. DTU happens to be the highest performing Danish university when it comes to impact and is ranked in the top 50 of the Leiden Ranking. DTU has a very strong focus upon science and technology and close to no medical, social or humanistic research activities. Also, DTU has the lowest student to researcher ratio in Denmark. Obviously, DTU would fit very-well to a model based on citations. DTU has continued the criticism over the years claiming that they are the actually "losers" in the current Danish PRFS. According to DTU, universities are reward for quantity and not "quality" which should always be the focus in research. Why risk the current impact status by increasing output for some marginal gains? This cannot be a national interest.

So goes the argument - what we examine in this paper is to what extent the argument holds. Who benefits from the current Danish publication-based model and is DTU the current "losers"? What would be the differences if a citation-based approach was applied instead?

The aim of the analysis is to examine the universities' "return on investment". We take a simple approach where we examine the relative input of the universities when it comes to publications and subsequently examine the relative output from these publications, i.e., the rewards in the model, either the current publication points, or the alternative, citations. We keep the analysis simple using basically a zero-sum approach, like the current model, where gains somewhere mean losses elsewhere.

[2] www.leidenranking.com

The next section briefly presents the data and main methods and indicators used for the analyses. The subsequent section presents main results, and the final section contains a brief discussion of the findings.

Data and methods

The paper examines the first full publication year (2009) used for redistributing funds in the Danish model. We are able to measure the citation impact of the Danish journal publications from 2009 and make comparisons between the Danish universities and examine their potential gains and/or losses by using either differentiated publication counts or citations. We compare publication counts and points derived from the BFI model between Danish universities, and we likewise compare the impact between these universities for the 2009 journal publications indexed in Web of Science (WoS). As argued in the introduction section, locking the main research areas in principle means that the current publication-based model could be adapted to specific behaviours and wishes, or even supplemented or exchanged with a citation approach, in the individual areas, although citations would only be relevant in the areas: science and technology and medical and health sciences. In this paper we focus the analysis on the main research area of science and technology. We do this because the issue concerning citation impact versus publication activity raised by DTU is directly linked to this area due to DTUs research profile. We have done a corresponding analysis for the medical and health sciences but due to limited space we will not address them in this paper.

The publication activity in 2009 in the main research area of science and technology is around 8700 publications of all types eligible in the BFI model, books constituted 2%, articles in books 19% and journal articles 79%. It is reasonable to argue that (international) journal publication is the primary publication activity in this area, which means that citation analysis of eligible articles is a sensible endeavour. However, as the area includes some fields known to have their main publication activity in conference proceedings (i.e., articles in books), we do scrutinize the influence of proceedings papers on the total number of BFI points acquired for the individual universities and discuss that in relation to the citation analysis where proceedings papers are excluded. Notice, we do not include conference papers in the citation analysis due to the meagre quality of the current proceedings citation indices.

All journal publications published in 2009 reported by the universities to the BFI-indicator were extracted from the BFI database. Subsequently, paper titles were extracted, and so were first author names and journal names. These parameters were used to match the publications with Danish WoS journal publications from 2009 using CWTS's in house version of WoS. Eligible publication types are research articles and reviews. The match rate is 77% of the initial journal articles. Among the non-matched publications were non-English language articles, as well as false positive articles, articles not eligible for the BFI model, but still succeeded in accruing points.

As indicated in the introduction section, the BFI model applies a fractional counting method at the institutional level where articles are fractioned up to $1/10^{th}$ among the participating institutions. We do not apply the exact same counting formula for the WoS publications going into the citation analysis. Here we simply do a straightforward fractional counting on the institutional level. As will be clear from the results section, this small deviance had no practical relevance on relative publication shares.

We use standard CWTS citation indicators from the Leiden Ranking (www.leidenranking: P_{frac} (fractionalized publications), TNCS (total number of normalized citations), MNCS (mean normalized citation score) and PPtop10% (proportion of papers for a unit among the 10 percent most cited in the database) (Waltman et al., 2012).

Eight universities are included in the Danish PRFS. The universities differ considerable in both subject/faculty composition and size. We have two "old" universities basically covering

all four main research areas included in the BFI model: Copenhagen University (KU) and Aarhus University (AU). These universities are also the largest universities in Denmark with long research traditions and strong science faculties. University of Southern Denmark (SDU) is a younger university, but its subject/faculty composition is basically a reflection of KU and AU, although the size is considerably lower. Roskilde University (RU) and Aalborg University (AAU) are even younger, from the mid-1970s. These universities have regional obligations with a substantial emphasis on teaching. Nevertheless, both universities have developed unique research profiles, both universities have focused on interdisciplinary research, where RU has a strong focus on the social sciences and AAU has focused strongly on engineering. Both universities have science and technology faculties, albeit at RU the size is only comparable to a large department. The Information-Technology University is the youngest and smallest university in Denmark. Their focus is mainly outside the science and technology areas but we include them here for numbers to add up. Likewise, Copenhagen Business School (CBS) is also included for matters of completeness in the analyses, their publication activity in the science and technology area are scanty. Finally, as discussed in the introduction, the Technical University of Denmark (DTU) is basically a "mono-faculty" university, albeit its activities are spread between science and technology. It is important to emphasise that while the university is known for primarily educating engineers, it has a considerable research activity in what would be considered basic natural science fields as well. In fact DTU can be dated back to the early nineteenth century where it was part of Copenhagen University, making it the second oldest university in Denmark. We recapitulate, DTU has been particularly dissatisfied with the Danish PRFS arguing that - for them at least - citations would be a more appropriate and valid performance-based indicator. In the next section we examine the consequences of this claim.

We calculate basic statistics based on individual articles both for the publication-based model and the simple citation approach we apply. As stated in the introduction, we take a simple approach where we examine the relative input of the universities when it comes to publication shares and subsequently examine the relative "rewards" the universities archives from these publications, i.e., the output in the model, either shares of the total publication points, or the alternative, shares of the total number of citations. Also, we keep the analysis simple using basically a zero-sum approach, like the current PRFS, where gains somewhere mean losses elsewhere.

Results

Table 1 below shows the eight universities' total number of matched fractionalized WoS publications belonging to the science and technology area, as well as their accumulated number of normalized citations after four years. Notice, these are fractionalized WoS publications, the absolute number of publications is 6,117.

Table 1 also shows relative citation performance for the eight universities using the MNCS and PPtop10% field normalized indicators.

The three main actors measured by volume is not surprisingly KU (32.9%), DTU (28.7%) and AU (21.4%), the volumes for AAU and SDU are considerably lower, both universities have a share of 7.2% of the total volume. DTU has the largest number of normalized citations among the eight universities. It is noticeable that DTU's share of citations (34.8%) is markedly higher than their share of publications (28.7%). Obviously, this is also reflected in the relative citation indicators. The MNCS at 1.66 is considerably higher than the average of the database and a score that would rank DTU among the top 30 in the Leiden Ranking if we only focused on science and technology, and among the top 50 for all fields combined.

Table 1. Science and technology: Number of fractionalized publications in WoS, total number of citations and relative citation indicators.

	WoS pubs (P_{frac})	TNCS	MNCS	Share of total P_{frac}	Share of total no. of NCS	PPtop10%
AAU	225.3	284.4	1.26	7.2%	6.6%	12.3%
AU	673.0	874.5	1.30	21.4%	20.3%	14.6%
CBS	13.2	12.2	0.93	0.4%	0.3%	
DTU	904.9	1498.8	1.66	28.7%	34.8%	17.0%
ITU	11.5	9.1	0.79	0.4%	0.2%	
KU	1035.9	1281.0	1.24	32.9%	29.7%	13.4%
RU	56.9	61.3	1.08	1.8%	1.4%	10.7%
SDU	227.7	284.8	1.25	7.2%	6.6%	15.8%
Total	3148.2	4306.3		100%	100%	100%

Interestingly, we also see that the minor universities, SDU and AAU, have relative citation indicator scores comparable to the larger universities KU and AU. In fact, SDU has more of their 2009 publications among the 10% most cited in the database compared to KU and AU. Overall, these results confirm what we suspect and are essentially the basis for the argument about including citations in the BFI model advanced by DTU.

In order to examine "return on investment", i.e., the institutions' reward for their publication input, we have calculated their share of BFI publications and BFI points for 2009 for the science and technology area, as well as the shares of fractionalized WoS publications and the total number of field normalized (fractionalized) citations. We thereby assume that shares of BFI points and shares of normalized citations can be treated equally. In the final discussion section we reflect upon this. We do, however, think that the straightforward approach taken can give a rudimentary indication of potential differences in "returns" for the individual institutions if one was to apply a citation based approach instead of or as a supplement to the current differentiated publication-based indicator in the science and technology area.

Table 2 below shows the shares of BFI publications and BFI points, where all publication types used in the science and technology fields are included (e.g., also conference proceedings), as well as shares of fractionalized WoS journal articles and normalized citations.

Table 2. Science and technology: Distribution and shares of BFI-points, BFI-publications, plus fractionalized publications from WoS and total number of normalized citations; notice <u>all</u> BFI-publication types are included.

	BFI-points	BFI-publications (P)	Share of BFI-points	Share of total BFI P	Share of P_{frac} (WoS)	Share of total no. of TNCS
AU	1814.9	1766	19.1%	20.4%	21.4%	20.3%
CBS	6.9	6	0.1%	0.1%	0.4%	0.3%
DTU	2854.1	2378	30.1%	27.5%	28.7%	34.8%
ITU	117.4	107	1.2%	1.2%	0.4%	0.2%
KU	2730.9	2457	28.8%	28.4%	32.9%	29.7%
RUC	185.9	157	2.0%	1.8%	1.8%	1.4%
SDU	571.0	572	6.0%	6.6%	7.2%	6.6%
AAU	1203.6	1219	12.7%	14.1%	7.2%	6.6%
	9484.8	8662	100%	100%	100%	100%

Table 3 below shows the same variables as Table 2, but in this case we *only* use the BFI publication type journal articles and the points derived from these articles. Table 3 is included for comparison because the citation analysis in reality only deals with journal articles. Notice, the BFI journal articles include non-WoS indexed articles, which give points in the indicator, however, the numbers are very low, the coverage of the science area in WoS is very high.

Table 3. Science and technology: Science and Technology: Distribution and shares of BFI-points, BFI-publications, plus fractionalized publications from WoS and total number of normalized citations; notice only the BFI-publication type journal article is included.

	BFI-points (journals only)	BFI-publications (P) (journals only)	Share of BFI-points (journals only)	Share of total BFI P (journals only)	Share of P_{frac} (WoS)	Share of total no. of NCS
AU	1526.2	1515	21.9%	23.5%	21.4%	20.3%
CBS	6.9	6	0.1%	0.1%	0.4%	0.3%
DTU	2007.4	1663	28.8%	25.8%	28.7%	34.8%
ITU	53.3	39	0.8%	0.6%	0.4%	0.2%
KU	2166.8	2047	31.1%	31.8%	32.9%	29.7%
RUC	139.5	126	2.0%	2.0%	1.8%	1.4%
SDU	420.1	442	6.0%	6.9%	7.2%	6.6%
AAU	657.5	596	9.4%	9.3%	7.2%	6.6%
Total	6977.7	6434	100%	100%	100%	100%

For analytical and illustrative reasons we plot the results from Table 2 and 3 in Figures 1 and 2 below. Figure 1 shows the results based on all BFI publication types, whereas Figure 2 shows the results where only BFI journal articles are included.

The figures are simple plots were the shares of the total number of publications (i.e., both BFI publications and fractionalized publications from WoS) for the eight universities constitute the x-axis, this is the "input", i.e. what the individual institutions "invested" in the Danish performance-based model for science and technology in 2009. The y-axis shows the shares of BFI points and citations, this is the "output", i.e. the institutions' "return on their investment" in the Danish performance-based model for science and technology in 2009. The axes are symmetrical and the diagonal shows the point where the institution has the same relative share of input (publications) and output (BFI points or citations). The distance from the university to the diagonal suggests whether input is larger than the return (output), which means that the institution will be below the diagonal, or the return (output) is larger, in which case the university is placed above the diagonal. Further, each university is plotted two times, one for the BFI data and one for the WoS citation data. Significant changes between these two representations for a university up and down the diagonal, suggest that the university receives a substantial number of BFI points from publication types other than journal articles. Notice in order to avoid confusion when examining the figures, shares of BFI publications on the x-axis should be compared to shares of BFI points on the y-axis, and likewise shares of WoS publications on the x-axis should be compared with shares of citations on the y-axis.

It is clear from Figure 1 that RU, CBS and ITU are not interesting for the current analysis as their numbers and shares are too low. We are interested in the other five universities, which all have a faculty of some size within science and technology. Interestingly, from Figure 1, where *all* BFI publication types are included, we can see that DTU actually has a larger output than input with a ratio of 1.09. This is somewhat unexpected and contrary to the conjecture that DTU is not gaining much from the current model. If we then turn to the

citation analysis, then we can see an even larger distance from the diagonal to DTU, compared to the BFI data, but also all other universities. The ratio is 1.25, so in line with the previous findings, DTUs WoS publications receive considerably more citations than the other Danish universities in 2009 but also the average paper in the WoS database. If a citation-based indicator of some sort were constructed where points were given based on citations, as implied in the arguments from DTU, then it seems that DTU would benefit from such a model, obviously conditioned on how it was designed. However, the most interesting finding here is perhaps that DTU within the science and technology area also seems to be the largest beneficiary when it comes to BFI points earned per input publication. Notice, like the current PRFS, we also treat it as a zero-sum game. If all universities improve then we have status quo. As it is in Figure 1, only DTU seems to really benefit from the citation approach. While KU seems to be in balance with the BFI data, they experience a smaller drop in returns on their input in the citation approach. Perhaps the most remarkable result from Figure 1 is the dramatic drop on the diagonal between BFI data and WoS citation data for AAU. We return to this below.

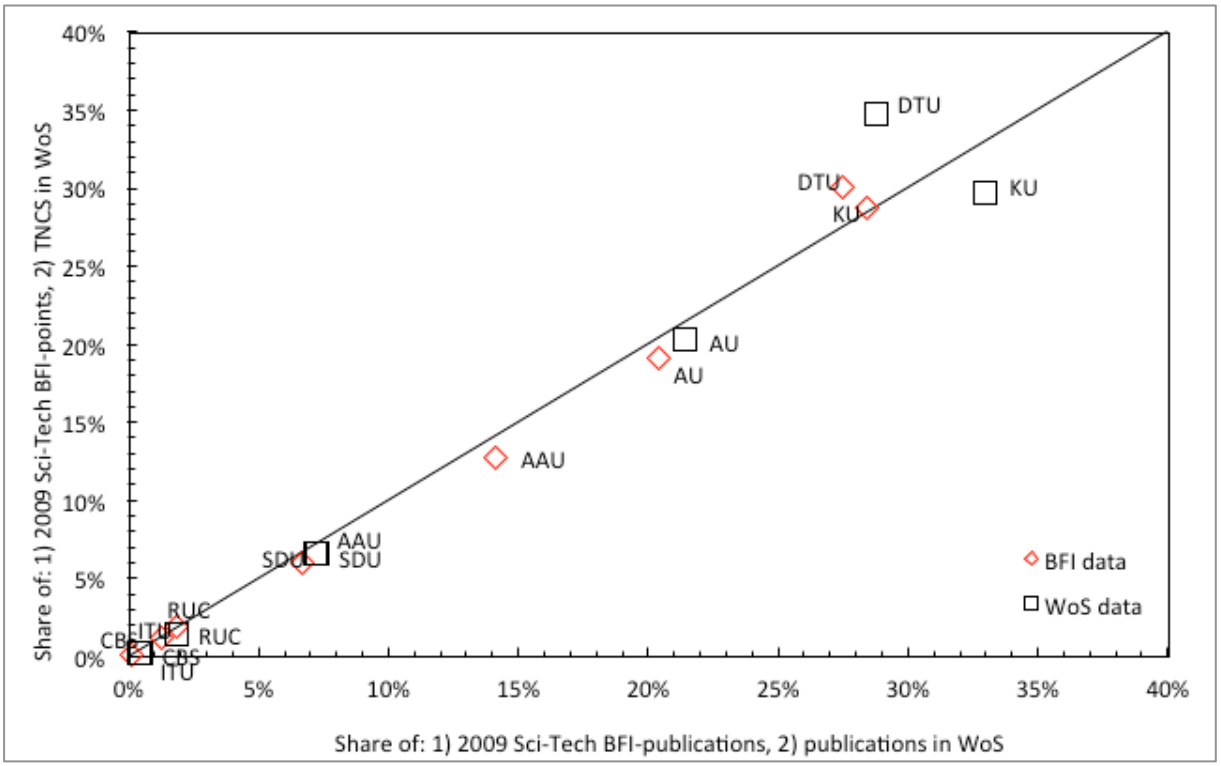


Figure 1. Science and technology: 1) Shares of 2009 Sci-Tech BFI publications as a function of shares of 2009 BFI-points; and 2) Shares of 2009 Sci-Tech WoS publications as a function of shares of total number of citations to these publications; notice BFI data includes all publication types.

Figure 2 depicts the same analysis but this time we have reduced the BFI data to include only journal publications in order to compare like with like, i.e., BFI journal data with WoS journal data. Obviously, the WoS data are identical to Figure 1, what is changing is the relative shares of BFI data (i.e., shares of publications and shares of points). There are some minor repositions, but the two major differences are the large drop on the diagonal for AAU and the corresponding smaller drop above the diagonal for DTU. Notice, the input-output is in balance for AAU, whereas DTU still has a substantial "return on investments" when it comes BFI journal data.

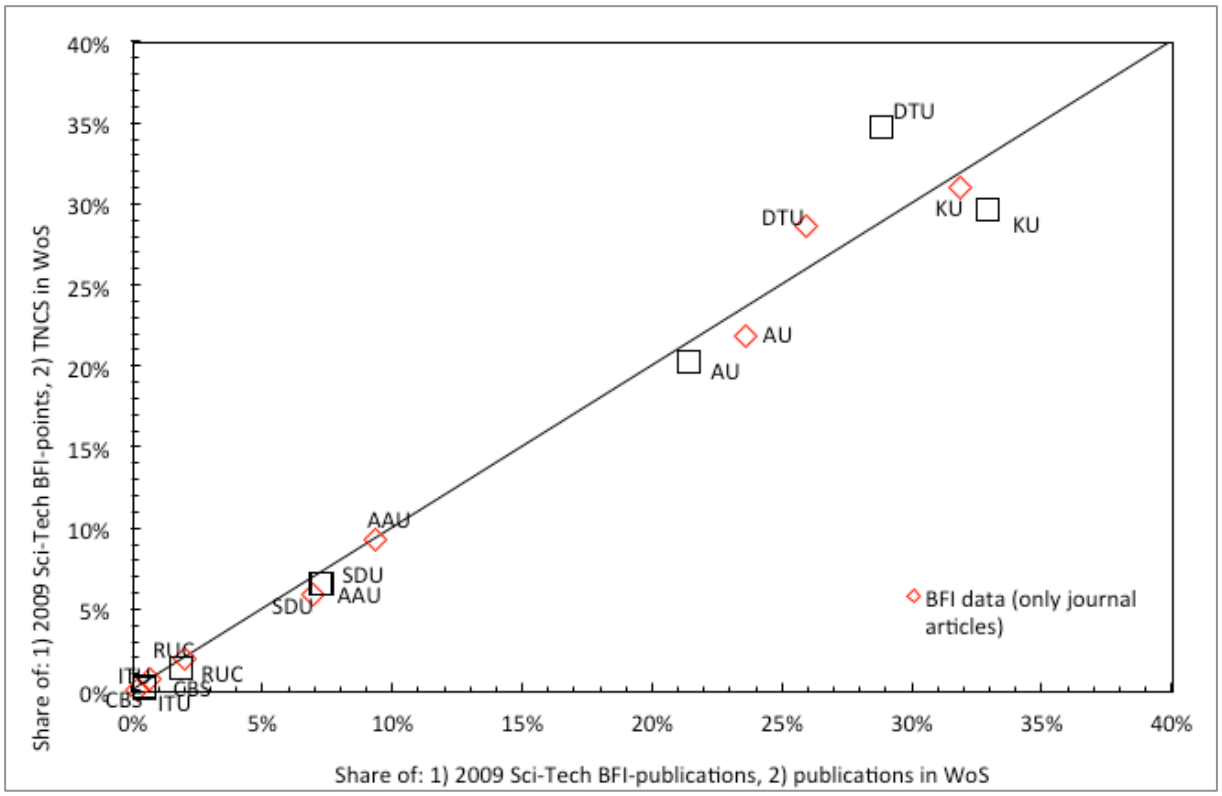


Figure 2. Science and technology: 1) Shares of 2009 Sci-Tech BFI publications as a function of shares of 2009 BFI-points; and 2) Shares of 2009 Sci-Tech WoS publications as a function of shares of total number of citations to these publications; notice BFI data only includes the publication type journal articles.

The drop of AAU along the diagonal was foretold in the WoS data in Figure 1. Here we saw a considerable distance between the BFI data when they included all publication types and the restricted WoS journal data needed for the citation analysis. For obvious reasons, this gap has been shortened considerably in Figure 2 since both data sets are restricted to journal articles.

The discrepancy in Figure 1 and the drop in Figure 2 are caused by the deviant publication profile for AAU compared to the other four universities with substantial publication activity in the science and technology area. Interestingly, 41% of the BFI publication activity in 2009 for AAU is in the category "articles in books", which in this case essentially means conference papers, and 49% is journal articles. For a comparison, 21% of DTUs activity is in "articles in books" and 70% in journal articles. These are both universities with strong focus on the technical sciences where publication in conference proceedings is very important. To contrast these profiles, the three other universities, KU, AU and SDU, all have more traditional science faculties and their relative publication activity in "articles in books" is 9%, 9% and 14% respectively. For these universities, due to their strong focus on science and less focus on technology, journal publication is the main activity 83% for KU, 86% for AU and 77% for SDU. However, we can also see that DTU does indeed have a strong science focus judged from their strong journal publication profile.

Considering the impetus for DTU to argue for a citation model, it is interesting to notice that while DTU clearly has the highest citation performance among the eight universities based on the 2009 journal publications, as we expected, they also have the highest performance when it comes to BFI publication points. Indeed, it seems that DTU would benefit even more in the science and technology area if they were to be rewarded for their relative share of the total number of citations, but contrary to the expected and suggested, DTU also benefit the most when it comes shares of BFI publication points compared to their relative input in the science

and technology area. DTU seems not only to be the most efficient when it comes to citations, this is also the case when it comes to BFI publication points. For example, the size of KUs activity in the science and technology area is larger than DTUs, but DTUs average point per publication is 1.20 for both of the above-mentioned analyses, considerably higher than KUs at 1.11.

Discussion

The main immediate findings in the present case study is that DTU will most probably benefit from a citation model, but perhaps more important, that they also seem to be the relatively most efficient university when it comes to BFI publication points. What are the more general implications of these findings seen in relation the current spread of the NPM to a number of European countries? The Danish case is special because competition is locked within the main areas this opens up for adapted models across areas including citation models where relevant. In Sweden a citation model is currently in use encompassing all fields. This is undesirable for several reasons; one of them is clearly demonstrated in this analysis, the desire to embrace all major publication behaviours, one of the rationales for the original NPM. A citation model alone restricts data to journal articles indexed in one of the two major citation databases. It was clear from Figure 1, that a university with an emphasis on technical sciences, like AAU, will be reduced in relative size when it comes to sharing the output.

The NPM is a differentiated publication indicator where points are graded for where you publish. Incentives to improve performance are clear and straightforward. Citation indicators reflect short term impact upon the scientific communication system. Citation indicators are retrospective and quite stable. It is very difficult to directly try to improve performance when it comes to impact. While one can argue that a publication-based model support the publish and perish culture with the ever increasing publication pressure, one could also argue that a citation model at the university level, due to its stability or conservative nature, and the fact that preferential attachment is at play for some universities, most likely would give cumulative advantages to those "who already have plenty", and potential changes brought about by incentives, are certainly not a short term phenomena.

There have been suggestions in Denmark to meet some of the requirements from DTU to focus more on citation impact. In order to keep the existing differentiated publication model intact, suggestions have been presented to bring in a third level especially in relation to journal outlets. This should be a category for the few hyped journals and publishing in these should be rewarded more lavishly. There may be good reasons for extending the levels in the model, but it is a flawed argument to claim to compensate wishes for more focus on impact by rewarding publication activity in "high impact" outlets. As it is well-known, article citation rates and journal citation impact have meagre correlations and the latter is a rather poor predictor of the former (Seglen, 1997).

A citation-based indicator or a hybrid indicator based on both publications and citations can be conceived in many ways, the question is whether the former or the latter is desirable. As discussed in the introduction, publication activity and citation impact are two different phenomena with substantially different prospects when it comes to incentives and behavioural adjustments. In the present analysis we could of course have experimented with more sophisticated citation-based approaches, for instance by constructing a mirror of the current publication-based model, where an arbitrary system allocates points according to which percentile group in the citation distribution they belonged to. We actually did that with a three-tiered point system, both the results were in line with the ones presented here.

As it is, based on the 2009 data, the BFI model in Denmark seems to work. Claims of more focus on citation impact seem only to speed up the cumulative advantage for "those who

already have" and at the same downgrade the influence of certain publication behaviours and muddling the transparent incentive structure.

References

Aagaard, K. (2011). Kampen om basismidlerne. Historisk institutionel analyse af basisbevillingsmodellens udvikling på universitetsområdet i danmark. PhD, Aarhus University, Aarhus.

Aagaard, K., Bloch, C., & Schneider, J. W. (2015). Impacts of performance-based research funding systems: The case of the Norwegian publication indicator. Research Evaluation, 24(2), 106-117.

Butler, L. (2002). A list of published papers is no measure of value - the present system rewards quantity, not quality - but hasty changes could be as bad. Nature, 419(6910), 877-877.

Butler, L. (2003a). 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. (2003b). Modifying publication practices in response to funding formulas. Research Evaluation, 12(1), 39-46.

Gläser, J., & Laudel, G. (2007). The social construction of bibliometric evaluations. In R. Whitley & J. Gläser (Eds.), The changing governance of the sciences (Vol. 26, pp. 101-123): Springer Netherlands.

Hicks, D. (2012). Performance-based university research funding systems. Research Policy, 41(2), 251-261.

Karlsson, S. & Persson, O. (2012). The swedish production of highly cited papers Vetenskabsrådets lilla rapportserie. Stockholm, SWE.

Merton, R. K. (1988). The Matthew effect in science, ii: Cumulative advantage and the symbolism of intellectual property. Isis, 79(4), 606-623.

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

Moed, H. F. (2008). UK research assessment exercises: Informed judgments on research quality or quantity? Scientometrics, 74(1), 153-161.

Oppenheim, C. (1996). Do citations count? Citation indexing and the research assessment exercise (rae). Serials: The Journal for the Serials Community, 9(2), 155-161.

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.

Schneider, J. W., & Aagaard, K. (2012). "Stor ståhej for ingenting" - den danske bibliometriske indikator. In K. Aagaard & N. Mejlgaard (Eds.), Dansk forskningspolitik efter årtusindskiftet (pp. 229-260). Aarhus: Aarhus Universitetsforlag.

Seglen, P. O. (1997). Citations and journal impact factors: Questionable indicators of research quality. Allergy, 52(11), 1050-1056.

Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. ISSI Newsletter (International Society for Scientometrics and Informetrics), 6(1), 22-28.

Verleysen, F. T., Ghesquière, P., & Engels, T. (2014). The objectives, design and selection process of the Flemish academic bibliographic database for the social sciences and humanities (vabb-shw). In W. Blockmans, L. Engwall & D. Weaire (Eds.), Bibliometrics: Use and abuse in the review of research performance (pp. 117-127). London: Portland Press Ltd.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., 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.