UNIVERSITY OF WEST BOHEMIA

Faculty of Applied Sciences

HABILITATION THESIS

Plzeň, 2013 Dalibor FIALA

KIV-FAV-ZČU

Modern Informetric Methods for the Evaluation of Scientific Research

Habilitation Thesis

Dalibor FIALA

December 2013

To Anna and Veronika

Abstract

Informetrics is a relatively new scientific discipline linking computer science with information science and using various data mining, statistical, and graph-theoretical approaches and methods to measure information. It may be regarded as a general field of science that comprises scientometrics, bibliometrics, webometrics, and other –metrics fields that have all seen an enormous growth in recent years. In fact, in a time when the need for scientific advancement and technological innovation is immense, but funding sources are limited, measuring the quality of research outputs has become indispensable and recognized by many. Indeed, the recently founded *Journal of Informetrics* has immediately become one of the fastest growing high-impact journals in the Journal Citation Reports® database by Thomson Reuters, which clearly demonstrates the importance of informetrics as a research domain.

This habilitation thesis presents the research I have conducted in the last several years with the aim of developing new methods to evaluate scientific research output more fairly and whose main results I have published in leading journals of the field. First, a new method based on the PageRank algorithm by Google is presented that detects the most influential researchers by analyzing citation as well as collaboration networks and by taking into account the time of publications and collaborations (published in *Journal of Informetrics*). Second, a large-scale bibliometric analysis of a huge data collection gained from the CiteSeer digital library is carried out in order to determine the scientific production and impact of countries in computer science (appeared in *Information Processing and Management*). Third, the same digital library is used to find the most prominent computer science researchers by applying 12 different ranking methods (published in *Scientometrics*). Fourth, a new scientometric indicator is introduced that is based on the h-index and can grow as well as decline (in press in *Journal of the American Society for Information Science and Technology*). Fifth, a large-scale analysis of institutional suborganizations in the field of library and information science is carried out (published in *Information*). And sixth, the differences between CiteSeer and its successor, CiteSeer^{X}, in terms of coauthorship networks are investigated (appeared in *Journal of Theoretical and Applied Information Technology*).

The work presented in this habilitation thesis is a significant contribution to the advancement of informetrics, particularly in the Czech Republic, where the field is almost unknown.

Acknowledgements

I would like to thank professors Jiří Šafařík and Karel Ježek without whom this work would have never been possible. I am also grateful to all of my family for their love and support, in particular to my father who introduced me to the world of computers when I was nine years old. And, last but not least, I thank God for all those crazy things to explore!

Dalibor Fiala Plzeň, 17 December 2013

Table of Contents

Introduction

What is informetrics?

The evaluation of scientific research output has become crucial in recent years as the budgets of science funding bodies (governments, foundations, etc.) have become tight, but the need for research and innovations (with a view to increase competitiveness) has been ongoing or even growing. Therefore, it has become clear that it is absolutely necessary to identify high quality research that should be prioritized in receiving funding and also poor quality research whose funding is no more effective. The key concept here is to promote the advancement of science as efficiently as possible, i.e. to maximally increase the effort/award rate from the point of view of financing science. The scientific field concerned with the measurements of science is called *scientometrics* and along with the related domains *bibliometrics* and *webometrics* it forms the basis of an emerging field called ***informetrics***. Informetrics stands on the boundary between computer science and information science and is currently a hot topic among researchers (Bar-Ilan, 2008). It is well documented by the fast recognition of the recently founded *Journal of Informetrics* (in 2007) as one of the leading journals in information science.

Science Evaluation

Science evaluation (one of the main informetric applications) is possible at various levels and can be easily transformed into the evaluation of individual researchers, research teams, institutions, or even countries. This evaluation is mostly based on the assessment of research productivity (number of publications) and research impact (number of citations). In the research productivity assessment, not only publication numbers are considered but also the reputation of publication sources. This leads us to the evaluation of the impact of journals and conferences. In this context, an important scientometric indicator of journal quality is the journal impact factor. It has been criticized since it was first introduced in the 1970s (Archambault and Larivière, 2009; Rossner et al., 2007), but it still plays a tremendous role in the evaluation of journal quality and research output. Perhaps one of the reasons for its ongoing usage is its

simplicity (Hubbard and McVeigh, 2011): it is the number of citations to a journal's articles that appeared in two preceding years from journal articles published in a specific year divided by the number of that journal's articles that appeared in the two preceding years. This approach is actually a simple relative citation counting and has a great deal of flaws some of which will be further discussed in the Open Problems.

Bollen et al. (2006) applied the recursive PageRank algorithm used in the Google search engine (Brin and Page, 1998) to a citation network of journals and found big differences between the journal quality rankings generated in this way and those based on the standard impact factor. The PageRank algorithm, which can be applied to any directed graph, considers not only the number of citations a node receives but also the quality of citing nodes. A quality citing node has itself many citations from other quality nodes. Therefore, the quality of nodes is defined recursively and is also referred to as *prestige* in contrast to *popularity* represented by simple citation counts. Unlike a popular journal (or researcher, institution, country, etc.), a prestigious journal (researcher) may be cited less but from more prestigious journals (researchers). This finesse can never be captured by simple citation counts. Although these higher-order methods have been long used on the Web to detect quality Web sites (see also Kleinberg, 1999), they are still relatively little used in research assessment. A notable exception is the recent introduction of the Eigenfactor score (and Article Influence) in the Web of Science[1] database and of SCImago Journal Rank (and Source Normalized Impact per Paper) in Scopus[2] to evaluate the impact of journals.

The situation with the assessment of individual researchers is even worse. Until recently, the main scientometric indicator of a scientist's research impact was the number of publications and citations the researcher had in the Web of Science database by Thomson Reuters. But with the appearance of other sources of bibliographic and citation data such as Scopus, ACM DL[3], DBLP[4], CiteSeer[5], and Google Scholar[6], the number of metrics assessing the significance of individuals has grown as well. In 2005 Hirsch proposed the h-index that combined both the productivity and impact of an individual researcher in a single number (Hirsch, 2005). The index is defined as follows: if we have a set of publications ordered by

[1] *apps.isiknowledge.com*

[2] *www.scopus.com*

[3] *dl.acm.org*

[4] *dblp.uni-trier.de*

[5] *citeseer.ist.psu.edu*

[6] *scholar.google.com*

the number of times they are cited in descending order, the index *h* is the largest number *h* such that there are *h* publications having at least *h* citations each. Thus, a scholar with an h-index of 20 has published 20 papers at least (productivity) and has received no less than 400 citations (impact). The h-index attained a great popularity and was mathematically analyzed (Egghe, 2007; Wu et al., 2011) and praised (Vanclay, 2007) but it was also soon discovered that various corrections were needed (Costas and Bordons, 2007; Bornmann and Daniel, 2007; Egghe, 2011).

For instance, the h-indices of two researchers from different research fields or sub-fields are incomparable because publication and citation practice may vary to a great extent between those two fields. Also, it would be unfair to consider the h-indices of two scientists the same if one of the researchers always publishes with a large group of co-authors and the other researcher only publishes alone. In addition, author self-citations can inflate the h-index (Schreiber, 2007), etc. To remedy this situation, many h-index variants have been proposed (Alonso et al., 2009; Jin et al., 2007). One of the most popular variants is also the g-index defined by Egghe (Egghe, 2006; Schreiber, 2010). The h-index and other metrics based on it can be applied to any set of publications, for instance aggregated by institutions, countries, or journals (Schubert, 2007; Schubert and Glänzel, 2007). Thus, they seem to be a suitable tool for science evaluation. However, there is a large number of these indicators and they are still relatively very new so further research is needed in order to exactly identify their strengths and weaknesses or even find the optimal informetric indicator for the evaluation of scientific research (Podlubny and Kassayova, 2006).

Previous Work

The use of PageRank has been recently extended from journal citation networks also to other network types. Chen et al. (2007) applied it to find outstanding physics publications and Ma et al. (2008) used PageRank to rank publications and countries. Ding et al. (2009) computed PageRank and weighted PageRank for authors in an author co-citation network. Similar experiments were also conducted for an author citation graph (Ding, 2011) and a co-authorship graph (Yan and Ding, 2011). Citation weighting and a time factor is included in the *prestige rank* for publications conceived by Yan and Ding (2010) which uses the Article Influence score. In general, PageRank-based methods seem to be a promising tool in the scientific output evaluation.

In my previous work (Fiala et al., 2008), I modified the standard PageRank algorithm to better reflect the nature of bibliographic networks by taking into account not only the in-

formation on citations of researchers but also on collaborations among them. The key idea was that not all citations should weigh equally – a citation from a colleague should be considered less valuable than a citation from a foreign scholar. I later extended this model to also include the time information of citations and collaborations (Fiala, 2012b). In the new model only collaborations occurring prior to a citation decrease its value while the number of common publications of the citing and cited author appearing after a citation have absolutely no effect on the citation value. However, the number of common publications was not the only factor influencing citation weights – I introduced 14 new scientometric indicators in total and tested them on a large collection of citation and collaboration data (Fiala, 2011; Fiala, 2012a).

Open Problems

After a preliminary review of the current informetric literature, we have determined the following principal challenges for informetric researchers:

- life-time achievement vs. current performance (influence)
- account of co-authorship in both publications and citations
- differences between scientific fields
- different behaviour of researchers at different stages of their career
- honorary co-authorship or "ghost" co-authorship
- self-citations and citation cliques
- "the rich get richer" effect and time dependency of citations
- impact factor flaws and other journal quality metrics

A successful solution of any of the above issues would significantly advance the state of the art in the field of informetrics and would substantially contribute to the creation of a more objective, more robust, and a fairer system of scientific research output evaluation. The challenges are described in more detail below but I am aware that trying to tackle all of them at once would be too ambitious. Therefore, I suppose that only some of the issues will be addressed in my future work (see Conclusions).

Life-time achievement vs. current performance (influence). Scientometric indicators such as citation counts or h-index generally play in favour of more senior researchers (or older institutions, etc.) because these simply have had more time to publish and collect citations. Therefore, the current metrics indicate a kind of lifetime achievement instead of current (most recent) performance. There is a great need for an "age normalization" factor to be able to

fairly compare researchers of different ages. Also, the new indicator should be able to grow as well as decline – it should be dynamic.

Account of co-authorship in both publications and citations. Standard citation counts or even h-index do not take into account the number of co-authors a publication has. In this way, a researcher that has published twenty papers is considered more productive than another one who has published just ten articles even if the first has always published with a team of collaborators and the latter is the sole author of all of his/her papers. Also, citations to papers always count the same although the numbers of authors in the cited (or even citing) articles differ significantly. Some authors feel that this is unjust. Thus, informetricians seek to develop new methods that correct the standard indicators for multiple co-authorship.

Differences between scientific fields. Different fields of science have different publication and citation practices. Therefore, it is impossible and unfair to compare researchers and institutions specialized in distinct research domains or subdomains by raw metrics. It is necessary to develop techniques that enable one to correct scientometric indicators for differences between scientific domains and subdomains.

Different behaviour of researchers at different stages of their career. In addition to the "scientific age" of a researcher, the stage of his/her professional career is also important. Junior researchers may tend to publish with their more experienced colleagues to learn lessons from them even if they themselves have abilities to conduct research and publish. Conversely, mid-career researchers may be tempted to prove their skills by writing more single-author papers, while late-career scientists generously let junior colleagues co-publish with them to gain experience. This different behaviour results in different numbers of publications and co-authors at distinct career stages and a fair research output evaluation should take it into account.

Honorary co-authorship or "ghost" co-authorship. The role of co-authors in research publications can vary dramatically. Some co-authors may take part in the research work to the same extent as the primary author; others just contribute by revising a paper draft or by discussing the design or results of a study. Honorary co-authors have contributed neither to conducting research nor to publishing a manuscript. They are present among authors for strategic purposes only, for instance to support a paper's acceptance in a single blind review process. On the other hand, "ghost" co-authors are missing in the author list even though they have significantly contributed to the research. One of the possible reasons may be the need to avoid

a conflict of interests. Both co-authorship types are ethical issues that an objective assessment of scientific productivity and impact should be able to detect and take into account.

Self-citations and citation cliques. As the main instrument to identify quality research are still considered citations to its results, there is a growing pressure on scientists to be cited. Low citation numbers often mean low chances to receive grants or to be promoted. Therefore, scholars may feel tempted to cite themselves to a greater extent than admissible, which is normally easily detected by standard scientometric indicators, or to cite their friends who, in turn, cite them later. If the friends are not immediate collaborators and are not from the same institution, such citation cliques (or even "loops" if there are also other authors in the path between the citing and cited researcher) cannot be discovered by current scientometric methods. More advanced techniques and metrics should be developed that will be able to cope with this kind of citation cheating.

"The rich get richer" effect and time dependency of citations. It is well known that highly cited publications attract more attention than poorly cited papers and, therefore, they also receive more citations than their less successful counterparts. Or, in other words, the more citations a paper has, the bigger the chances that it will be cited. Thus, the number of citations an article receives (in a specific year, for instance) should be corrected for the probability of being cited based on the current number of citations. Similarly, as citations are always directed towards older publications, following a chain of citations leads us inevitably to the past. On the other hand, researchers might tend to cite more recent articles on a given topic when searching in a bibliographic database. This nature of citations and citation behaviour should be further studied to evaluate research more objectively and reliably.

Impact factor flaws and other journal quality metrics. The mythical journal impact factor has a few well-known flaws, one of which is the way its numerator and denominator are computed (The PLoS Medicine Editors, 2006). Whereas in the numerator there are citations to all "items" in a journal (including research articles, reviews, letters, editorials, notes, etc.), in the denominator, there is just the number of "citable items", i.e. research articles and reviews. Of course, the number of these "citable items" influences the resulting impact factor to a great extent. The problem is how to decide what is "citable" and what is not because the differences between the various article types may not always be evident. To overcome some of the difficulties with the journal impact factor, other journal quality metrics have been proposed such as Eigenfactor (Bergstrom, 2007) or SCImago Journal Rank (González-Pereira et al., 2010).

A lot of in-depth research is needed to explore their possible advantages over the impact factor or to design such impact factor modifications that will reflect journal quality in a more robust manner.

Structure of the Thesis

This thesis contains six recent high-impact journal articles and two other journal articles on informetric issues along with a brief nonspecialist commentary for each of them:

- Fiala, D. (2012b). Time-aware PageRank for bibliographic networks. ***Journal of Informetrics***, 6(3), 370-388.
- Fiala, D. (2012a). Bibliometric analysis of CiteSeer data for countries. ***Information Processing and Management***, 48(2), 242-253.
- Fiala, D. (2011). Mining citation information from CiteSeer data. ***Scientometrics***, 86(3), 553-562.
- Fiala, D. (2013). Current Index: A proposal of a dynamic rating system for researchers. ***Journal of the American Society for Information Science and Technology***. DOI: 10.1002/asi.23049. (in press)
- Fiala, D. (2013). Suborganizations of institutions in library and information science journals. ***Information***, 4(4), 351-366.
- Fiala, D. (2013). From CiteSeer to CiteSeer^{X}: Author rankings based on coauthorship networks. ***Journal of Theoretical and Applied Information Technology***, 58(1), 191-204.

The first article lies at the core of this thesis and its great importance to the informetric community has been honoured by its appearance in the second highest-ranked journal in its category (2012 Journal Citation Reports®, Thomson Reuters, 2013). It presents a novel modification of the well-known PageRank algorithm that takes into account the time information inherent in bibliographic networks. The other two articles and the last one are scientometric studies of the different aspects of the CiteSeer digital library or its successor. The fourth paper is a breaktrough proposal of a dynamic evaluation scheme for the rating of researchers' scientific performance and the fifth one is a pioneering scientometruic analysis of institutional suborganizations publishing library and information science journals. I am the only author of all the articles above, which allows me speaking in the first person in all the text below.

Despite the late Jan Vlachý (1937-2010) being one of the pioneers in scientometrics, informetrics as such is quite immature and underdeveloped in the Czech Republic. In this respect, the studies presented in this habilitation thesis are a significant contribution to the advancement of the field and surely a major contribution from the Czech perspective.

Article 1

I consider the following article as the core of this habilitation thesis and as my main contribution to the field of informetrics at all. It was published in the *Journal of Informetrics* in 2012. This journal was founded in 2007 solely for the purpose of promoting and consolidating the informetric discipline and currently it is the second highest-ranked journal in the category "Information Science & Library Science" with an impact factor of 4.153 (2012 Journal Citation Reports®, Thomson Reuters, 2013).

In fact, the article is an extension and generalization of an earlier publication (Fiala et al., 2008), in which a method was proposed to evaluate (and rank) researchers based on the citation network of their publications. The underlying algorithm is the famous PageRank by Brin and Page used in their Web search engine named Google. This algorithm was first applied in the Web domain to identify important Web pages, but it was later employed in many other situations, where the problem addressed could be modelled as a (directed) graph, including bibliographic networks. The original PageRank was based on the structure of the Web graph (with Web pages as nodes and links between them as directed edges) and it was recursive – it considered a Web page as important if many important Web pages linked to it. Of course, importance can be gained by one inlink from a very important document or by many inlinks from unimportant documents. Both possibilities are legitimate. In any case, the graph structure is essential. It enables a recursive method (assessing prestige) to see "farther" than a non-recursive one and, as an example, to value more a page inlinked by a single document than a page inlinked by ten other documents, which could never be accomplished by a non-recursive technique such as in-degree (assessing popularity). But bibliographic networks are more complex than the Web graph:

- First, there can be more types of them such as citation, collaboration (coauthorship), or co-citation networks to name the most essential ones. Citation networks themselves can have various nodes connected by directed edges. The nodes, for instance, can be publications, their authors, or journals in which they appear. The basic citation is a

reference from a publication to another publication and all other citation types (between authors, journals, institutions, countries, etc.) can be derived from this. A collaboration (or coauthorship) network represents how individual actors (authors, institutions, countries) collaborate, e.g. whether or not they publish joint papers, and its structure can vary considerably depending on the application. And finally, a co-citation network has papers (or authors or journals) as nodes and (undirected) edges between them when they are co-cited (in a paper).

- Second, as a result of the existence of so many different bibliographic networks, the information inherent in each of these networks can be combined and used to assign weights to the edges in the citation graph (of authors, for example). In fact, virtually all of the aforementioned bibliographic networks can be edge-weighted. In contrast, edges in the Web graph are unweighted. Consequently, these properties of bibliographic networks enable one to modify the standard PageRank formula and to inject additional information into it. An example of such additional information used to weight an edge (citation) between two authors can be the number of joint publications of these two authors. (Actually, I defined seven items of additional information.) This add-on information enables one to distinguish the "value" of a citation. A citation from a colleague is certainly less valuable than a citation from a complete stranger. Therefore, edges (citations) between authors with many common publications are assigned smaller weights than edges between authors with no co-published papers.
- And third, there is another significant factor that is not at all present in the Web graph structure – time. On the other hand, the time factor plays a significant role in the bibliographic networks. More specifically, papers are published in certain years (or even months of years) and these years can be associated with the edges in coauthorship as well as in citation graphs. Moreover, when calculating the weight of a citation based on the number of common publications of two authors, the papers published before the citation was made should influence the citation weight, but the papers published afterwards should not. In other words, the citation weight should be computed on the basis of bibliographic networks as they looked like at the time of the citation and not at the time of the network evaluation.

The last aspect is an extension of the first idea to combine citation and collaboration networks (Fiala et al., 2008) and was introduced in the following article (Fiala, 2012b). Whereas the former article proposed seven factors to include in the PageRank formula, the latter presented a time-aware variant for each of them resulting in 14 various PageRank modifications.

Preprint of: Fiala, D. (2012). Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, 6(3), 370-388.

Time-aware PageRank for bibliographic networks

Dalibor Fiala

University of West Bohemia, Univerzitní 8, 30614 Plzeň, Czech Republic
Phone: +420 377 63 24 29, fax: +420 377 63 24 01, email: dalfia@kiv.zcu.cz

Abstract: In the past, recursive algorithms, such as PageRank originally conceived for the Web, have been successfully used to rank nodes in the citation networks of papers, authors, or journals. They have proved to determine prestige and not popularity, unlike citation counts. However, bibliographic networks, in contrast to the Web, have some specific features that enable the assigning of different weights to citations, thus adding more information to the process of finding prominence. For example, a citation between two authors may be weighed according to whether and when those two authors collaborated with each other, which is information that can be found in the co-authorship network. In this study, we define a couple of PageRank modifications that weigh citations between authors differently based on the information from the co-authorship graph. In addition, we put emphasis on the time of publications and citations. We test our algorithms on the Web of Science data of computer science journal articles and determine the most prominent computer scientists in the 10-year period of 1996 – 2005. Besides a correlation analysis, we also compare our rankings to the lists of ACM A. M. Turing Award and ACM SIGMOD E. F. Codd Innovations Award winners and find the new time-aware methods to outperform standard PageRank and its time-unaware weighted variants.

Keywords: PageRank, citations, collaboration, time, salient researchers, computer science.

1. Introduction

When Brin and Page made public their PageRank algorithm in 1998 (Brin and Page, 1998), they would probably hardly have imagined what an enormous impact on computer science this would have in the decade to come. They presented a straightforward method of computing the importance of Web pages using the link structure of the (then still relatively new) World Wide Web. The same concept of "authoritativeness" of Web pages was, at approximately the same time, invented independently by Kleinberg (1999). The idea was surprisingly simple: if a link from one Web page to another one can be viewed as a vote then popular pages will have many in-links. In addition, if those in-links come from pages that themselves have many in-links, popularity becomes prestige. It was soon discovered that this recursive

technique (applied successfully by Google) could be used to evaluate (rank) nodes in any (directed) graph. Bibliographic citation networks of papers, authors, journals, institutions, or even countries are good examples of such graphs and some studies making use of PageRank or related methods to find prominent players in them are touched upon in Section 2. However, researchers also felt the need to weigh the edges in bibliographic citation networks (unlike the original PageRank which was unweighted) based on the differences between the Web graph and the citation networks. First, bibliographic networks often contain information that can add value to citations, e.g. citation counts or co-authorship information. Second, unlike the Web graph, bibliometric networks include time, e.g. publication (citation) years, which could also help weigh citations more discriminately. And third, bibliometric networks are static in that citations always point from newer to older publications and they can never be removed. Fiala et al. (2008) addressed the first problem. They assigned more or less weight to the edges in a citation graph of authors based on the information from a co-authorship graph. The principal assumption was that a citation from a colleague should contribute less to the prestige of the cited author than a citation from a "foreign" researcher. On the other hand, this penalization could be mitigated in some circumstances, for instance, if the number of common publications of those two authors is relatively small compared to the total number of publications by the authors. Time in terms of publication (and thus citation or collaboration) years was ignored in this scenario. However, it is clear that if a citation is made before any common papers are published, it should not be considered a "friendly" citation from a colleague. This problem is addressed in this article. As for the third issue (static citations), some proposals to solve this problem are mentioned in the section on related work.

The principal objectives of the research reported in this paper are as follows:

- Define "time-aware" modifications of the "bibliographic PageRank" formula based on the work by Fiala et al. (2008) that take into account the time (year) when articles are published and citations are made.
- Apply the time-aware as well as the original (time-unaware) bibliographic PageRank variants to a large citation network of computer science researchers to find out the most prominent computer scientists.
- Compare the rankings of researchers generated by the new methods with each other as well as with other established bibliometric techniques in terms of a correlation analysis and a confrontation with the lists of ACM A. M. Turing Award (Turing Award) and ACM SIGMOD E. F. Codd Innovations Award (Codd Award) winners.

This article is organized in the following way: after introducing PageRank and our research goals in Section 1, related work on measuring computer science and various modifications of PageRank is reviewed in Section 2. Afterwards, in Section 3, we describe in detail an extension to the standard PageRank that is suitable for bibliographic networks and that can exploit the time information present in them. Section 4 is concerned with the data to which we applied the novel methods and then we discuss the experimental results in Section 5. Finally, we draw the main conclusions and outline our future work in Section 6.

2. Related work

This section on related work consists of three main paragraphs. The first one is concerned with previous bibliometric work on computer science, which has, somewhat surprisingly, been relatively little explored in the past. The second paragraph enumerates the principal studies known to the author that have sought to add weights to the basic PageRank formula and, finally, research into time-based weighting of PageRank is presented in the third paragraph.

2.1. Computer science

Bar-Ilan (2010) studied how publication and citation counts of some highly cited computer science researchers changed after conference proceedings papers had been added to the Web of Science (WoS). Franceschet (2010) investigates prestige, popularity, and productivity of computer science researchers with regard to journal versus conference papers. He defines a prestigious computer scientist as an ACM A. M. Turing Award winner. Wainer et al. (2011) studied how many publications by computer science researchers are not indexed by established bibliographic databases compared to other scientific fields and concluded that, on average, 66% of a computer scientist's work is unknown to the Web of Science. Bibliometric studies on computer science based on the data from the CiteSeer digital library are presented by Fiala (2011, in press).

2.2. PageRank and weighted PageRank

Bollen et al. (2006) assigned weights based on the number of citations to the edges in the citation network of journals and computed weighted PageRanks for the journals. Chen et al. (2007) calculate PageRank of papers from a set of physics journals. Different weighting and normalization schemes were applied to PageRank by Bergstrom (2007) and González-Pereira et al. (2010) to compute journal prestige. The corresponding scores are called Eigenfactor (or Article Influence when related to papers) and SCImago Journal Rank (SJR), respectively.

Ding (2011) computes weighted PageRank for authors in the information retrieval field. She assigns weights based on the number of publications or citations to nodes rather than edges, and experiments with various damping factors in the PageRank formula. A similar study for author co-citation networks is conducted by Ding et al. (2009). Yan and Ding (2011) explore co-authorship networks in the informetrics field. They calculate PageRank for authors with different damping factors and draw the conclusion that the damping factor does not have much influence on ranking in this type of network. They also define a weighted PageRank in which more weight is assigned to authors with more citations. Ma et al. (2008) computed PageRank for papers in the field of biochemistry and molecular biology. Xing and Ghorbani (2004) defined the "weighted PageRank" by multiplying the rank of each in-linking node by two factors: the in-degree of the current node divided by the sum of in-degrees of the nodes linked to by the in-linking node, and the out-degree of the current node divided by the sum of out-degrees of the nodes linked to by the in-linking node. This enabled more rank to be transferred to more "popular" nodes, i.e. to those that had relatively numerous in-links and/or out-links. The authors reported some success compared to the standard PageRank in obtaining more relevant results from a (very) small set of Web pages. Their approach does not seem reasonable in the case of citation networks of papers or authors because it is not clear why a paper (author) should be rewarded for citing many other papers (authors), i.e. the out-degree factor is doubtful. If just the in-degree factor was retained, their method would somewhat resemble the work by Ding (2011). Sidiropoulos and Manolopoulos (2005) adapted PageRank for publication citation networks in that they gave less weight to the citations from more distant publications (in terms of graph path). They were also the first to compare new ranking methods with established awards such as the ACM SIGMOD E. F. Codd Innovations Award.

2.3. Weighted PageRank considering time

Walker et al. (2007) ranked publications in two distinct citation networks of physics papers. They included the age of publications in the PageRank algorithm by favouring citations from more recent articles. They also experimentally verified the until then theoretical concept that the average path length of a random surfer following citations between research publications is only around two. Yan and Ding (2010) also bring time into play when they give more weight to more recent citations (i.e. to the citations from publications that appear shortly after the cited papers). In addition, they more heavily weight citations from prestigious articles, but their prestige (article influence score) is not computed recursively in a self-contained way (like PageRank) but rather taken from a citation database. In their "TimedPageRank", Yu et

al. (2004) simply decrease the weight of a citation exponentially with the citation age using a base (decay rate) of 0.5. For the prediction of popularity a paper will enjoy in future years, they apply an "ageing factor" as well that linearly declines a paper's TimedPageRank with the paper's age.

In summary, all the authors of the above studies on (time-)weighted PageRank report its superiority to the standard PageRank but, at the same time, find a high correlation of various PageRank variants and other bibliometric measures such as citation counts. None of the studies, however, has combined time information from both the citation and collaboration graphs to rank computer science researchers via the "time-aware" PageRank described in this paper.

3. Methods

The methods of time-aware PageRank described in this paper are based on the techniques used by Fiala et al. (2008) by including the time factor in their PageRank modifications that take into account not only citations between authors but also other information such as the number of common publications between two authors linked by a citation. The key concept was that citations between authors should not be weighted the same but should rather be based on a number of factors reflecting the behaviour of authors. For instance, a citation between two authors who often collaborate with each other is considered less valuable than that between two authors who have never co-authored a single publication. We invite the reader to get more explanations and see examples in Fiala et al. (2008). In the following paragraphs, we will re-define "the bibliographic PageRank" from our previous work and expand it with time aspects so that it allows for the computation of "time-aware bibliographic PageRank".

3.1. Definitions

To understand Figure 1, let $G^P = (P \cup A, E^P, \boldsymbol{T}^P)$ be an undirected, edge-weighted, bipartite graph (co-authorship graph), $P \cup A$ a set of vertices ($P = \{p_1, ..., p_n\}$ a set of publications, $A = \{a_1, ..., a_m\}$ a set of authors), $E^P \subseteq P \times A$ a set of edges, and $\boldsymbol{T}^P$ an $n \times m$ matrix of non-negative weights – publication years. Each edge $\{p, a\} \in E^P$, $p \in P$, $a \in A$ means that author a has (co-)authored publication p that appeared in year $\boldsymbol{T}^P_{pa}$. (If $\boldsymbol{T}^P_{ij} = 0$ then there is no such edge $\{i, j\}$ in E^P.) Let $G^C = (P, E^C, \boldsymbol{T}^C)$ be a directed edge-weighted graph (publication citation graph), $P = \{p_1, ..., p_n\}$ a set of vertices (the same set of publications), $E^C \subseteq P \times P$ a set of edges (citations between publications), and $\boldsymbol{T}^C$ an $n \times n$ matrix of non-negative weights – citation years. Each edge $\{p_1, p_2\} \in E^C$, $p_1 \in P$, $p_2 \in P$ means that publication p_1 from year

$\boldsymbol{T}^C_{p1p2}$ cites publication p_2. (If $\boldsymbol{T}^C_{ij} = 0$ then there is no such edge $\{i, j\}$ in E^C.) Now, we will combine the two graphs G^P and G^C into one more graph we will further work with. Let $G = (A, E)$ be a directed edge-weighted graph (author citation graph), $A = \{a_1, ..., a_m\}$ a set of vertices (the same set of authors) and $E \subseteq A \times A$ a set of edges (citations between authors). For every $p \in P$ let $A_p = \{a \in A: \exists\{p,a\} \in E^P\}$ be the set of authors of publication p. For each (a_1,a_2), $a_1 \in A$, $a_2 \in A$, $a_1 \neq a_2$ where there exists $(p_1,p_2) \in E^C$ such that $\{p_1,a_1\} \in E^P$ and $\{p_2,a_2\} \in E^P$ and $A_{p1} \cap A_{p2} = \varnothing$ (i.e. no common authors in citing and cited publications are allowed) there is an edge $(a_1,a_2) \in E$ (no parallel edges are admitted). Thus, $(a_1,a_2) \in E$ if and only if $\exists(p_1,p_2) \in E^C \wedge \exists\{p_1,a_1\} \in E^P \wedge \exists\{p_2,a_2\} \in E^P \wedge A_{p1} \cap A_{p2} = \varnothing \wedge a_1 \neq a_2$.

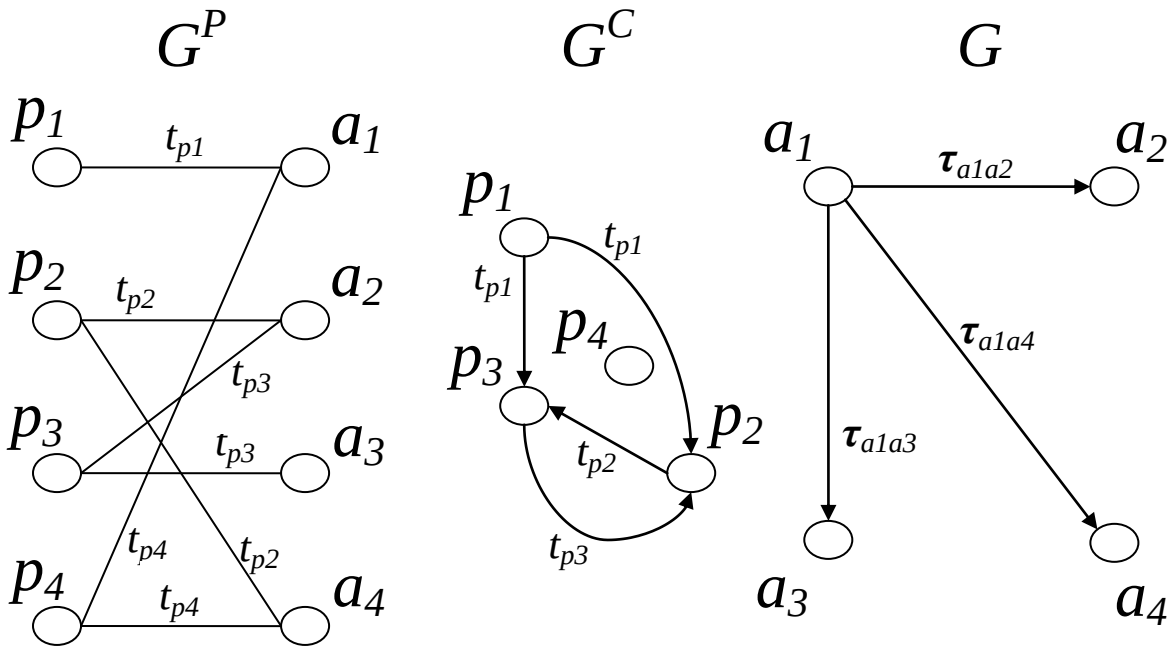


Fig. 1 Example of a co-authorship (G^P), publication citation (G^C), and author citation (G) graph

Before assigning weights to the edges in E, we further define:

- $w_{u,v} = |C|$ where $C = \{p_1 \in P: \exists\{p_1,u\} \in E^P \wedge \exists\{p_2,v\} \in E^P \wedge \exists\{p_1,p_2\} \in E^C \wedge p_1 \neq p_2\}$, as the number of citations from u to v;
- $f^t_{u,v} = |P^t_u| + |P^t_v|$ where $P^t_i = \{p \in P: \exists\{p,i\} \in E^P \wedge \boldsymbol{T}^P_{pi} < t\}$, as the number of publications by u appearing before year t plus the number of publications by v appearing before year t (called *publicationsT*); if $t = \infty$ (i.e. time is not taken into account), $f^t_{u,v}$ becomes $f_{u,v}$ (time-unaware, called *publications*);

- $c^t_{u,v} = |CP^t|$ where $CP^t = \{p \in P: \exists\{p,u\} \in E^P \wedge \exists\{p,v\} \in E^P \wedge \boldsymbol{T}^P_{pu} < t \wedge \boldsymbol{T}^P_{pv} < t\}$, as the number of common publications by *u* and *v* published before year *t* (called *collaborationT*); if $t = \infty$, $c^t_{u,v}$ becomes $c_{u,v}$ (called *collaboration*);
- $hd^t_{u,v} = |ADC^t_u| + |ADC^t_v|$ where $ADC^t_i = \{a \in A: \exists p \in P$ such that $\{p,a\} \in E^P \wedge \{p,i\} \in E^P \wedge \boldsymbol{T}^P_{pa} < t \wedge \boldsymbol{T}^P_{pi} < t\}$, as the number of all distinct co-authors of *u* in the papers published before year *t* plus the number of all distinct co-authors of *v* in the papers published before year *t* (called *allDistCoauthorsT*)); if $t = \infty$, $hd^t_{u,v}$ becomes $hd_{u,v}$ (called *allDistCoauthors*);
- $h^t_{u,v} = |ADC^t_u| + |ADC^t_v|$ where ADC^t_i is defined as above, but is a multiset, as the number of all co-authors of *u* in the papers published before year *t* plus the number of all co-authors of *v* in the papers published before year *t* (called *allCoauthorsT*); if $t = \infty$, $h^t_{u,v}$ becomes $h_{u,v}$ (called *allCoauthors*);
- $td^t_{u,v} = |DCA^t|$ where $DCA^t = \{a \in A: \exists p \in P$ such that $\{p,a\} \in E^P \wedge \{p,u\} \in E^P \wedge \{p,v\} \in E^P \wedge \boldsymbol{T}^P_{pu} < t \wedge \boldsymbol{T}^P_{pv} < t\}$, as the number of distinct co-authors in the common publications by *u* and *v* appearing before year *t* (called *distCoauthorsT)*; if $t = \infty$, $td^t_{u,v}$ becomes $td_{u,v}$ (called *distCoauthors*);
- $t^t_{u,v} = |DCA^t|$ where DCA^t is defined as above, but is a multiset, as the number of co-authors in the common publications by *u* and *v* appearing before year *t* (called *allDistCoauthorsT)*; if $t = \infty$, $t^t_{u,v}$ becomes $t_{u,v}$ (called *allDistCoauthors*);
- $g^t_{u,v} = f^t_{u,v} - |SP^t_u| - |SP^t_v|$ where $SP^t_i = \{p \in P: \{p,i\} \in E^P \wedge d_{G^P}(p) = 1 \wedge \boldsymbol{T}^P_{pi} < t\}$, as the number of publications by *u* that appeared before year *t*, where *u* is not the only author, plus the number of publications by *v* that appeared before year *t*, where *v* is not the only author (called *allCollaborationsT*); if $t = \infty$, $g^t_{u,v}$ becomes $g_{u,v}$ (called *allCollaborations*).

3.2. Time-aware PageRank

Now, we associate a vector of weight pairs $\boldsymbol{\tau}_{uv} = ((c^{t1}_{u,v}, b^{t1}_{u,v})^1, (c^{t2}_{u,v}, b^{t2}_{u,v})^2, \ldots, (c^{tk}_{u,v}, b^{tk}_{u,v})^k)$ with each edge $(u, v) \in E$, where $k = w_{u,v}$ (the number of citations from author *u* to author *v*) and *t1…tk* are the citation years selected as all those non-zero elements $\boldsymbol{T}^C_{ij}$, where $i \in P_u$ and $j \in P_v$, and we denote $P_a = \{p \in P: \exists\{p,a\} \in E^P\}$ as the set of publications of every author $a \in A$. $w_{u,v}$ and $c^t_{u,v}$ are described above, and $b^t_{u,v}$ can be equal to one of the seven following values according to the semantics of edge weights we want to stress: a) zero, b) $f^t_{u,v}$, c)

$h^t_{u,v}$, d) $hd^t_{u,v}$, e) $g^t_{u,v}$, f) $t^t_{u,v}$, g) $td^t_{u,v}$. We then define the rank $R(u)$ for author u as follows, bearing in mind that the superscipt i means an index in vector $\boldsymbol{\tau}$ and not a year:

$$\mathrm{R}(u) = \frac{1-d}{|A|} + d \sum_{(v,u)\in E} \mathrm{R}(v) \frac{\sum_{i=1}^{w_{v,u}} \frac{1}{\frac{c^i_{v,u}+1}{b^i_{v,u}+1} \sum_{(v,j)\in E} 1}}{\sum_{(v,k)\in E} \sum_{i=1}^{w_{v,k}} \frac{1}{\frac{c^i_{v,k}+1}{b^i_{v,k}+1} \sum_{(v,j)\in E} 1}}. \quad (1)$$

If we wish to ignore time (i.e. publication and citation years) and set all the coefficients *t1*…*tk* to infinity, vector $\boldsymbol{\tau}_{uv}$ takes the form $((c_{u,v}, b_{u,v})^1, (c_{u,v}, b_{u,v})^2, \ldots, (c_{u,v}, b_{u,v})^k)$ and Eq. (1) can be re-written as

$$\mathrm{R}(u) = \frac{1-d}{|A|} + d \sum_{(v,u)\in E} \mathrm{R}(v) \frac{\frac{w_{v,u}}{\frac{c_{v,u}+1}{b_{v,u}+1} \sum_{(v,j)\in E} w_{v,j}}}{\sum_{(v,k)\in E} \frac{w_{v,k}}{\frac{c_{v,k}+1}{b_{v,k}+1} \sum_{(v,j)\in E} w_{v,j}}}, \quad (2)$$

which is exactly how the time-unaware modifications of PageRank were defined by Fiala et al. (2008). These modifications penalized citations by colleagues (influence of c) but relaxed the penalty in some circumstances such as a great number of co-authors (influence of b). Now we can easily show how Eq. (2) can be further reduced to the standard PageRank formula. First, we set all b's to zero and take into account only the collaboration coefficients c:

$$\mathrm{R}(u) = \frac{1-d}{|A|} + d \sum_{(v,u)\in E} \mathrm{R}(v) \frac{\frac{w_{v,u}}{(c_{v,u}+1) \sum_{(v,j)\in E} w_{v,j}}}{\sum_{(v,k)\in E} \frac{w_{v,k}}{(c_{v,k}+1) \sum_{(v,j)\in E} w_{v,j}}}. \quad (3)$$

Second, we disregard the co-authorship information by setting all c's to zero and obtain the weighted PageRank formula, in which the edges in the author citation graph G are weighted with w's:

$$\mathrm{R}(u) = \frac{1-d}{|A|} + d \sum_{(v,u)\in E} \mathrm{R}(v) \frac{\frac{w_{v,u}}{\sum_{(v,j)\in E} w_{v,j}}}{\sum_{(v,k)\in E} \frac{w_{v,k}}{\sum_{(v,j)\in E} w_{v,j}}}. \quad (4)$$

And third, we set all the edge weights w in G to 1 and receive a standard PageRank formula which is equivalent to that by Brin and Page (1998):

$$\mathrm{R}(u) = \frac{1-d}{|A|} + d \sum_{(v,u)\in E} \mathrm{R}(v) \frac{\frac{1}{D_{out}(v)}}{\sum_{(v,k)\in E} \frac{1}{D_{out}(v)}}, \tag{5}$$

where d is the damping factor (set to 0.9 in our experiments) and $D_{out}(v)$ is the out-degree of vertex v. The damping factor represents the probability of following a link from the current node in the graph. Brin and Page (1998) set it to 0.85, Walker et al. (2007) recommend it to be 0.5 for publication citation networks. However, there is no consensus yet on what the damping factor should be in author citation graphs.

The edge weights are pre-computed so the convergence of the PageRank modifications above does not differ from the standard PageRank (depending on d). In our experiments (see Section 5), the rankings became relatively stable after 20 to 30 iterations and we always continued to 50 iterations at most.

3.3. Example

In Figure 2 there is a simplified example of an author citation graph E with four nodes u, q, v, r, three edges (u,q), (u,v), and (u,r). and three weight vectors $\boldsymbol{\tau}_{uq}$, $\boldsymbol{\tau}_{uv}$, and $\boldsymbol{\tau}_{ur}$ assigned to them.

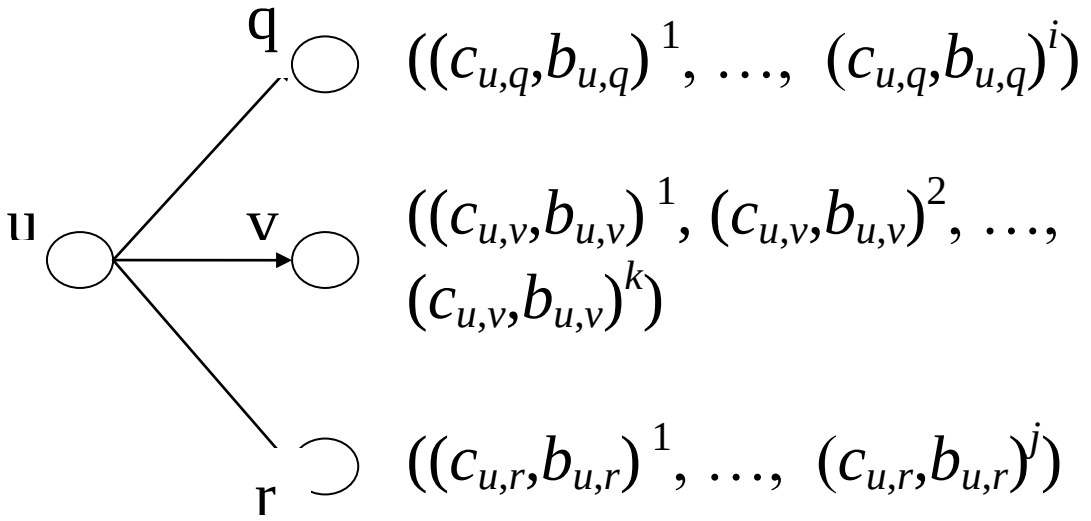


Fig. 2 Example of an author citation graph E with weight vectors $\boldsymbol{\tau}$

Now, let us suppose that *u* cites *v* three times, in 1980, 1990, and 2000. For the sake of simplicity, we will assume that all the coefficients *b* are equal to zero, i.e. we will solely rely on the citations and collaborations between authors. We will consider two cases. In the first case, *c* is 0 in 1980 (i.e., the number of common publications of *u* and *v* before 1980 was 0), 1 in 1990, and 2 in 2000. In the second case, *c* is 2 in 1980, 1990, and 2000 (see Figure 3). The interpretation of the scenarios might be the following: when author *u* first cited author *v* in 1980, they did not know each other yet (scenario 1 on the left-hand side of Figure 3). When *u* cited *v* for the second time ten years later, they were colleagues already and had written one common publication in the meantime. At the time of the third citation in 2000, their co-authorship relation was even stronger because they already had two common publications (still scenario 1). This is quite different from scenario 2 on the right-hand side of Figure 3, in which *u* and *v* probably know each other well in 1980 when *u* cites *v* for the first time as they already had two common publications at that time. However, they did not write any more articles together and their collaboration count *c* remains unchanged in 1990 and 2000 when *u* repeatedly cites *v*. If we ignore the citation years, the contribution (or weight) of the citations from *u* to *v* is 3/2 in both scenarios, which is the nominator in Eq. (2) if all *b*'s are 0. But, somehow, we feel that it is unjust and that the citation in 1980 should be weighted more if the authors do not know each other (left) than if they had already published together (right). Similarly, but perhaps less strictly, it happens in 1990 if the co-authorship relation between the authors is weaker (left) and stronger (right). Therefore, the co-authorship and other information entering the PageRank computation should always reflect the time of citation. This is exactly what we do in our time-aware PageRank modifications and formalize it in Eq. (1).

$v \Rightarrow u$	c
1980	0
1990	1
2000	2

$$\frac{w}{c} = \frac{3}{2}$$

$v \Rightarrow u$	c
1980	2
1990	2
2000	2

$$\frac{1}{0+1} + \frac{1}{1+1} + \frac{1}{2+1} = \frac{11}{6} \qquad \frac{1}{2+1} + \frac{1}{2+1} + \frac{1}{3+1} = 1$$

Fig. 3 Example of time-unaware (left) and time-aware (right) citation weighting

At the bottom of Figure 3, we can see the time-aware contributions of the individual citations. In 1980 the contribution is 1 (left) and ⅓ (right), in 1990 it is ½ (left) and ⅓ (right), and, finally, in 2000 it is ⅓ in both cases. Thus, the total weight of citations in scenario 1 is 11/6, almost twice as much as that in scenario 2. Therefore, we may feel that the time-aware weighting has brought more justice to the prestige computation.

4. Data

To conduct practical experiments with the new evaluation method (time-aware PageRank), we needed to acquire some real-world data. For this purpose, we decided to download publication data from the Web of Science database, which is a well established data source for bibliometric studies. As we were only interested in the field of computer science, it was first necessary to determine the field boundaries. Since WoS does not enable the science domain to be specified in a straightforward way, we were forced to limit ourselves to publications appearing in journals classified as computer science sources. To compile such a list of relevant journals, in March 2011, we consulted the Journal Citation Reports® Science Edition 2009 (the most recent JCR at that time) in the following seven computer science subject categories: artificial intelligence, cybernetics, hardware & architecture, information systems, interdisciplinary applications, software engineering, and theory & methods. The list contained 426 journals whose names we could use in the search queries submitted programmatically to the WoS web services via their API. The time period we were interested in was the decade at the turn of the century: 1996 – 2005. Name changes of journals in that period were not taken into account. Unfortunately, the "lite" version of WoS web services does not allow the specifying of what document types are to be retrieved, nor is the document type information available in the documents retrieved. Therefore, we simply downloaded meta data from the Science Citation Index on all documents (of any type) published in those 426 journals in the years 1996 to 2005. In this way, we obtained 205 780 "core" documents (strictly stated, their meta data such as title, authors, source, year, etc.). The next step was to find citations to these core documents from documents published up to December 31, 2010. To this end, we submitted further queries to WoS web services to determine citing documents for each individual core document. We found 1 569 057 citations from a total of 643 302 citing documents. Of the citing documents, only 91 728 were core documents for which all meta data were available. As a result, we were concerned with the analysis of 276 957 citations between core documents. In the core documents themselves, there were 187 016 different authors (disambiguated just by their surnames and given names' initials) with 1 471 312 citations between them (without self-

citations). The results discussed in Section 5 are based on the author citation graph. Detailed statistics of the data retrieved from WoS will be given in a separate article.

The data collection we have chosen has an obvious limitation: it is biased towards computer scientists who prefer publishing their research in journals, although it is well known that computer science research is presented at conferences to a greater extent than other fields of science (Bar-Ilan, 2010; Franceschet, 2010; Wainer et al., 2011). On the other hand, computer science journal articles receive more citations on average than conference papers (Franceschet, 2010) and we can expect that with a growing pressure on the visibility of papers and a faster journal editorial process, both of which we have been witnessing in recent years, the need for publishing computer science research in journals will increase.

5. Results and discussion

Table 1 shows the standings of the top 50 researchers as calculated by the "basic" methods – citation counts, in-degree, HITS, standard PageRank and weighted PageRank. By definition, citation counts are always greater or equal to in-degree. Since authors are not disambiguated, some names evidently represent more people with the same name as we can easily convince ourselves using a bibliographic database, e.g. in the case of "Jain, AK" or "Tanaka, K". On the other hand, some other names are apparently unique, e.g. "Kanade, T". The top authors by citations, in-degree, and HITS are very much the same with "Jain, AK", "Pentland, A", "Duin, RPW", and "Kanade, T" always appearing at the top. The interpretation of "Sapiro, G" being more highly ranked than "Kanade, T" in citations but more lowly ranked in in-degree is that "Sapiro, G" received more citations than "Kanade, T" but from fewer authors than "Kanade, T" did. Top-ranked authors by PageRank and by weighted PageRank are different from the first three rankings but similar to each other, with "Srinivasan, GR" and "Murley, PC" being at the very top.

Table 1 Top 50 researchers by citations, in-degree, HITS, and (weighted) PageRank

	Citations		In-degree		HITS	PageRank	Weighted PR
1	Jain, AK	3 103	Jain, AK	1 912	Jain, AK	Srinivasan, GR	Srinivasan, GR
2	Pentland, A	1 140	Pentland, A	851	Pentland, A	Murley, PC	Murley, PC
3	Duin, RPW	1 103	Duin, RPW	769	Belhumeur, PN	Tang, HHK	Ziegler, JF
4	Sapiro, G	1 036	Kanade, T	757	Duin, RPW	Freeman, LB	Freeman, LB
5	Kanade, T	1 026	Gupta, A	681	Kriegman, DJ	Ziegler, JF	Tang, HHK
6	Tanaka, K	1 018	Breiman, L	636	Kanade, T	Leinen, P	Leinen, P
7	Belhumeur, PN	971	Sapiro, G	634	Kikinis, R	Bey, J	Bey, J
8	Kriegman, DJ	964	Jain, R	631	Ayache, N	Juang, JG	Juang, JG
9	Scholkopf, B	959	Ayache, N	624	Jain, R	Juang, HG	Juang, HG
10	Breiman, L	952	Picard, RW	624	Smeulders, AWM	Korec, I	Curtis, HW
11	Viergever, MA	937	Belhumeur, PN	623	Kittler, J	Cegielski, P	Montrose, CJ
12	Kikinis, R	933	Viergever, MA	602	Maes, F	Wiener, N	Muhlfeld, HP
13	Wang, HO	933	Kittler, J	598	Vandermeulen, D	Muses, C	OGorman, TJ
14	Osher, S	920	Kriegman, DJ	596	Sapiro, G	Litkowski, KC	Ross, JM
15	Bates, DW	917	Kikinis, R	585	Hespanha, JP	McTavish, DG	Korec, I
16	Hyvarinen, A	896	Scholkopf, B	569	Suetens, P	Gazarik, MJ	Wiener, N
17	Jain, R	868	Hyvarinen, A	564	Duncan, JS	Kamen, EW	Cegielski, P
18	Muller, KR	868	Cox, IJ	562	Wells, WM	Prou, JM	Taber, AH
19	Calderbank, AR	866	Yu, PS	560	Viergever, MA	Wagneur, E	Walsh, JL
20	Tse, DNC	866	Lee, J	558	Picard, RW	Fidelman, U	Muses, C
21	Picard, RW	864	Muller, KR	544	Gupta, A	Ristow, GH	Litkowski, KC
22	Ayache, N	855	Huang, TS	542	Santini, S	Myers, JS	McTavish, DG
23	Gupta, A	852	Black, MJ	530	Huang, TS	Sampson, G	Gazarik, MJ
24	Kittler, J	838	Burges, CJC	515	Hawkes, DJ	Thomason, A	Kamen, EW
25	Yu, PS	802	Smeulders, AWM	508	Hill, DLG	Yngve, VH	Prou, JM
26	Hill, DLG	800	Osher, S	498	Poggio, T	Behbehani, J	Wagneur, E
27	Bezdek, JC	798	Szeliski, R	495	Moghaddam, B	Robinson, DL	Renegar, J
28	Tarokh, V	781	Bates, DW	489	Worring, M	Schwarzer, S	Fidelman, U
29	Hawkes, DJ	766	Oja, E	485	Sclaroff, S	Wachmann, B	Ristow, GH
30	Bro, R	764	Duncan, JS	482	Marchal, G	Wang, WY	Simon, DR
31	Black, MJ	757	Manjunath, BS	481	Manjunath, BS	Curtis, HW	Robinson, DL
32	Cox, IJ	748	Foster, I	480	Black, MJ	Montrose, CJ	Myers, JS
33	Duncan, JS	733	Zhu, SC	479	Zhu, SC	Muhlfeld, HP	Sampson, G
34	Shortliffe, EH	729	Santini, S	476	Collignon, A	OGorman, TJ	Thomason, A
35	Cimino, JJ	717	Suetens, P	475	Scholkopf, B	Ross, JM	Yngve, VH
36	Shahar, Y	713	Jain, A	471	Baluja, S	Taber, AH	Vazirani, U
37	Yager, RR	712	Flynn, PJ	470	Rowley, HA	Walsh, JL	Bernstein, E
38	Amari, S	711	Bezdek, JC	465	Mao, JC	Russell, CA	Schwarzer, S
39	Huang, TS	710	Thrun, S	461	Grimson, WEL	Chin, B	Wachmann, B
40	Suetens, P	710	Wells, WM	459	Prince, JL	Enger, TA	Wang, WY
41	Oja, E	709	Kim, J	457	Taylor, CJ	Hosier, P	Russell, CA
42	Musen, MA	692	Shortliffe, EH	455	Muller, KR	Klein, WA	Chin, B
43	Maes, F	689	Poggio, T	453	Jain, A	LaFave, LE	Enger, TA
44	Vandermeulen, D	689	Malik, J	452	Kimmel, R	Messina, B	Hosier, P
45	Wells, WM	685	Schapire, RE	452	Cox, IJ	Nicewicz, M	Klein, WA
46	Kimmel, R	683	Sejnowski, TJ	450	Jolesz, FA	Orro, JM	LaFave, LE
47	Zhu, SC	681	Maes, F	449	Matas, J	Scott, TS	Messina, B
48	Lee, J	668	Vandermeulen, D	449	Vailaya, A	Sullivan, TD	Nicewicz, M
49	Jain, A	666	Kumar, V	448	Rueckert, D	Sussman, RJ	Orro, JM
50	Schapire, RE	664	Jennings, NR	447	Ma, WY	Sykes, AJ	Scott, TS

5.1. Time-aware versus time-unaware rankings

As far as the rankings by the "advanced" methods (both time-aware and time-unaware) are concerned, the top 50 researchers in each ranking are shown in Tables A.1 to A.3 in the appendix. There are 14 rankings in total, with seven pairs of rankings, one of which is always the time-aware variant of the other: collaboration, publications, co-authors, distinct co-authors, all collaborations, all co-authors, and all distinct co-authors. The top-ranked authors by all methods are very much the same, e.g. with "Srinivasan, GR", "Murley, PC", and "Ziegler, JF" in high positions in each ranking. In fact, how similar are the individual rankings as a whole? Tables 2, 3, and 4 examine this aspect. In Table 2 we can see how the time-aware methods are correlated with each other. The table is symmetric and presents Spearman's rank correlation coefficients for each pair of time-aware rankings. The coefficients, which are all significant at the 0.01 level two-tailed, vary between 0.97 and 1 and suggest a very high correlation of all time-aware rankings. Similarly, very high Spearman's rank correlation coefficients can be observed in Table 3, which is non-symmetric and shows the correlation between time-aware and time-unaware PageRank variants. The most interesting figures are on the diagonal, where we can see how much new information is added if we use a time-aware variant instead of a standard PageRank modification. Provided that the lower the correlation achieved, the more new information is added using a time-aware method, the method taking into account all co-authors of an author prior to a citation (*allCoauthorsT*) instead of without regard to the citation time (*allCoauthors*) seems to work best. Table 4 is symmetric again. This time it shows how all the time-unaware rankings correlate with one another. The highest correlation can be observed with citations versus in-degree (0.997), the lowest correlation with HITS versus *allCoauthors* (0.730). All in all, HITS is relatively less correlated (0.74) with all the PageRank-based methods, but it is very highly positively correlated (0.93) with both of the first-order methods – citations and in-degree. As for the PageRank variants, their correlation coefficients with citations and in-degree are all around 0.83 and they are quite close to each other with correlations about 0.99. All the Spearman's rank correlation coefficients are significant at the 0.01 level two-tailed. The correlation between citations, in-degree, HITS, and (weighted) PageRank on the one side and the time-aware PageRanks on the other is not shown in Table 4, but the coefficients would be quite similar to those for the time-aware PageRanks regarding the high correlation between the time-aware and time-unaware rankings in Table 3.

Table 2 Spearman's rank correlation coefficients of time-aware rankings

	collaborati-onT	publicati-onsT	allCo-authorsT	allDistCo-authorsT	allCollabo-rationsT	co-authorsT	distCo-authorsT
collaborationT	**1**	0.975977	0.968838	0.973377	0.975813	0.999031	0.999095
publicationsT	0.975977	**1**	0.990443	0.990434	0.996298	0.976097	0.976147
allCoauthorsT	0.968838	0.990443	**1**	0.995957	0.992759	0.969081	0.969119
allDistCoauthorsT	0.973377	0.990434	0.995957	**1**	0.992303	0.973480	0.973527
allCollaborationsT	0.975813	0.996298	0.992759	0.992303	**1**	0.975932	0.975982
coauthorsT	0.999031	0.976097	0.969081	0.973480	0.975932	**1**	0.999905
distCoauthorsT	0.999095	0.976147	0.969119	0.973527	0.975982	0.999905	**1**

Table 3 Spearman's rank correlation coefficients of both kinds of rankings

	collaborati-onT	publicati-onsT	allCo-authorsT	allDistCo-authorsT	allCollabo-rationsT	co-authorsT	distCo-authorsT
collaboration	**0.999923**	0.975964	0.968850	0.973384	0.975817	0.999022	0.999074
publications	0.993022	**0.971791**	0.964273	0.968614	0.971584	0.994887	0.994677
allCoauthors	0.985647	0.963150	**0.955729**	0.960214	0.963047	0.987923	0.987665
allDistCoauthors	0.990937	0.969841	0.962263	**0.966590**	0.969608	0.993019	0.992810
allCollaborations	0.993664	0.972378	0.964848	0.969202	**0.972109**	0.995404	0.995199
coauthors	0.998322	0.975931	0.968821	0.973308	0.975756	**0.999304**	0.999151
distCoauthors	0.998737	0.975948	0.968946	0.973337	0.975784	0.999652	**0.999572**

Table 4 Spearman's rank correlation coefficients of time-unaware rankings

	Cites	InDeg	HITS	PR	PR weigh-ted	colla-borati-on	publi-cati-ons	all-Coaut hors	allDist-Coaut hors	allCol-labo-rations	coauth ors	dist-Coaut hors
Cites	**1**	0.9973	0.9269	0.8353	0.8322	0.8318	0.8295	0.8235	0.8277	0.8301	0.8322	0.8324
InDeg	0.9973	**1**	0.9284	0.8364	0.8311	0.8308	0.8283	0.8225	0.8266	0.8289	0.8311	0.8313
HITS	0.9269	0.9284	**1**	0.7538	0.7448	0.7445	0.7405	0.7301	0.7378	0.7415	0.7449	0.7450
PR	0.8353	0.8364	0.7538	**1**	0.9956	0.9956	0.9900	0.9831	0.9880	0.9906	0.9945	0.9950
PR weigh-ted	0.8322	0.8311	0.7448	0.9956	**1**	0.9998	0.9936	0.9864	0.9916	0.9943	0.9987	0.9990
collabo-ration	0.8318	0.8308	0.7445	0.9956	0.9998	**1**	0.9928	0.9853	0.9906	0.9934	0.9982	0.9986
publicati-ons	0.8295	0.8283	0.7405	0.9900	0.9936	0.9928	**1**	0.9958	0.9989	0.9997	0.9959	0.9956
all-Coauthors	0.8235	0.8225	0.7301	0.9831	0.9864	0.9853	0.9958	**1**	0.9972	0.9953	0.9894	0.9890
allDist-Coauthors	0.8277	0.8266	0.7378	0.9880	0.9916	0.9906	0.9989	0.9972	**1**	0.9986	0.9943	0.9939
allColla-borations	0.8301	0.8289	0.7415	0.9906	0.9943	0.9934	0.9997	0.9953	0.9986	**1**	0.9964	0.9961
coauthors	0.8322	0.8311	0.7449	0.9945	0.9987	0.9982	0.9959	0.9894	0.9943	0.9964	**1**	0.9996
dist-Coauthors	0.8324	0.8313	0.7450	0.9950	0.9990	0.9986	0.9956	0.9890	0.9939	0.9961	0.9996	**1**

5.2. ACM A. M. Turing Award winners

In a further experiment, we wanted to compare the rankings obtained by the various methods with a "true" human-made baseline ranking of some kind. In the computer science domain, such a "ranking" can be made of the list of ACM A. M. Turing Award laureates. Even though the list of award winners is actually not a ranking, it enables one to compare computer-generated lists of authoritative researchers with the scientists considered prestigious by their peers and has been successfully used in several comparative studies in the past (e.g. Sidiropoulos and Manolopoulos, 2005 or Fiala et al., 2008). Table 5 shows the ranks of Turing Award winners from the years 1991 to 2010 (the past 20 years) produced by all of the 19 ranking methods described above. "Hartmanis, J" (1993), "Dahl, O" and "Nygaard, K" (2001), "Naur, P" (2005), and "Thacker, C" (2009) do not appear anywhere in the rankings and their rows are blank. The first two columns in Table 5 comprise citations and in-degree (the most frequently used research evaluation method) followed by HITS, PageRank, and weighted PageRank. Then there is a block of seven time-unaware PageRank modifications and a set of their seven time-aware counterparts. The ranks generated by the recursive techniques (from HITS onwards) were computed after fifty iterations (with the Spearman's rho between the rankings of two consecutive iterations being very close to 1 after just a few iterations) and are less important than the summary figures at the bottom of the table.

These numbers are the best rank, worst rank, average rank, medium rank, and standard deviation. Obviously, the lower the numbers the "better" the ranking in that it places the Turing Award winners higher (low ranks mean high positions). Therefore, an optimum ranking (with respect to the Turing Award) would place the awardees at ranks 1 to 23 (without those five researchers omitted) thus having a best rank of 1, worst rank of 23, average and median rank of 12 and a standard deviation of 6.63. Standard PageRank (in a darker column) achieves better indicators (except for the worst rank and standard deviation) than both citations and in-degree and much better than HITS, but its weighted variant does not seem to perform more efficiently. As far as the time-unaware PageRank modifications are concerned, their best ranks are better than citations or in-degree have but similar or worse than those of (weighted) PageRank. The same holds for the average rank and standard deviation. On the other hand, worst ranks and median ranks are almost always better than PageRank has. Approximately the same conclusions may be drawn for the time-aware modifications of PageRank with two exceptions: notably better average ranks were yielded by *allCoauthorsT* and especially by *allDistCoauthorsT*, i.e. by the methods that take into account the number of all co-authors of

Table 5 ACM A. M. Turing Award winners (1991 – 2010) and their ranks

Year	Winner	citations	in-degree	HITS	PageRank	weighted PageRank	collaboration	publications	all-Coauthors	allDist-Coauthors	allCol-laborations	coauthors	dist-Coauthors	collaborationT	publicationsT	all-CoauthorsT	allDist-CoauthorsT	allCol-laborationsT	coauthorsT	dist-CoauthorsT
1991	Milner, R	16 590	12 612	50 753	12 234	11 814	11 428	17 335	22 964	22 025	17 123	14 437	14 152	11 488	14 080	13 682	12 330	16 113	13 527	13 274
1992	Lampson, B	6 842	7 698	18 864	4 295	5 504	5 407	6 744	7 556	6 901	6 701	5 839	5 736	5 438	5 294	6 299	5 628	5 946	5 551	5 517
1993	Hartmanis, J																			
1993	Stearns, RE	24 668	26 864	42 424	12 148	13 400	12 562	17 878	22 796	22 549	17 635	11 883	13 196	13 058	13 711	13 877	14 027	13 718	13 102	13 862
1994	Feigenbaum, EA	35 599	31 932	29 363	7 737	7 089	7 431	3 278	2 268	2 357	3 261	5 464	5 436	7 108	8 909	8 223	7 628	8 332	7 182	7 179
1994	Reddy, R	14 257	14 998	11 659	7 786	8 412	8 309	10 095	11 248	10 713	9 961	8 805	8 744	8 335	9 652	8 526	7 990	11 168	8 659	8 621
1995	Blum, M	8 934	7 618	33 665	6 456	5 745	5 682	8 227	9 954	9 004	8 086	6 058	5 956	5 656	9 367	9 769	9 120	9 255	5 913	5 830
1996	Pnueli, A	7 631	8 704	20 785	6 062	7 066	7 324	4 522	3 879	3 738	4 448	4 515	4 178	7 395	6 269	4 127	3 748	5 803	3 770	3 689
1997	Engelbart, D	73 807	74 009	58 367	72 511	72 129	72 108	72 588	72 651	72 505	72 569	72 139	72 133	72 115	62 265	68 327	67 279	68 912	72 117	72 117
1998	Gray, J	1 139	1 514	3 469	814	1 103	1 112	674	462	406	668	639	602	1 083	914	626	573	878	1 180	1 165
1999	Brooks, FP	2 609	3 459	13 669	2 937	3 429	3 376	4 194	4 435	4 294	4 177	3 747	3 646	3 386	3 956	3 294	2 997	4 479	3 553	3 519
2000	Yao, AC	21 696	23 911	36 410	13 812	11 711	11 487	22 671	28 340	25 165	17 860	12 916	12 558	11 604	15 398	17 259	15 557	14 827	12 085	12 018
2001	Dahl, O																			
2001	Nygaard, K																			
2002	Rivest, RL	17 719	19 912	27 857	18 337	24 745	24 682	26 376	27 077	26 454	26 254	25 228	25 038	24 697	16 265	17 081	17 190	15 847	24 926	24 824
2002	Shamir, A	12 309	11 345	16 354	978	971	1 027	876	1 448	1 043	873	926	914	1 010	1 916	2 222	2 032	1 725	941	939
2002	Adleman, LM	2 975	3 204	23 617	90	75	73	240	545	297	222	101	96	73	844	937	811	751	85	81
2003	Kay, A	74 842	74 991	75 210	77 939	77 442	77 428	76 884	75 607	76 689	76 985	77 341	77 357	77 415	74 852	74 667	75 163	74 126	77 433	77 437
2004	Cerf, VG	18 173	20 372	40 052	24 344	24 738	24 354	29 368	30 313	29 937	29 143	25 971	25 792	24 510	25 235	18 882	16 927	26 393	25 091	24 993
2004	Kahn, RE	16 450	18 672	37 702	19 355	23 906	23 521	27 978	29 155	28 679	27 675	24 920	24 776	23 676	26 755	15 643	13 908	25 437	24 197	24 130
2005	Naur, P																			
2006	Allen, F	6 308	7 178	8 569	28 841	29 024	29 057	28 691	29 631	28 773	28 540	27 920	28 315	29 084	25 829	14 864	13 726	25 413	28 568	28 544
2007	Clarke, EM	683	869	5 869	1 080	1 205	1 182	1 594	1 828	1 622	1 562	1 346	1 291	1 177	1 025	835	809	941	1 291	1 254
2007	Emerson, EA	4 859	3 974	14 513	6 082	4 627	4 604	5 282	5 932	5 451	5 226	4 666	4 659	4 625	3 582	4 083	4 133	3 480	4 595	4 577
2007	Sifakis, J	9 186	10 953	18 865	16 937	18 788	24 930	7 965	5 824	5 504	7 855	7 581	6 832	24 880	23 630	13 580	12 455	19 228	5 985	5 844
2008	Liskov, B	11 662	9 802	30 434	9 165	10 230	9 902	12 936	13 732	13 007	12 686	10 942	10 790	9 914	8 631	6 424	5 632	8 381	10 725	10 653
2009	Thacker, C																			
2010	Valiant, LG	15 980	16 523	32 513	10 186	9 711	9 935	8 420	8 501	8 281	8 239	8 839	9 105	9 994	8 444	13 876	13 002	12 312	9 172	9 354
	Best rank	**683**	**869**	**3 469**	**90**	**75**	**73**	**240**	**462**	**297**	**222**	**101**	**96**	**73**	**844**	**626**	**573**	**751**	**85**	**81**
	Worst rank	**74 842**	**74 991**	**75 210**	**77 939**	**77 442**	**77 428**	**76 884**	**75 607**	**76 689**	**76 985**	**77 341**	**77 357**	**77 415**	**74 852**	**74 667**	**75 163**	**74 126**	**77 433**	**77 437**
	Average rank	**17 605**	**17 875**	**28 304**	**15 658**	**16 211**	**16 388**	**17 166**	**18 093**	**17 626**	**16 859**	**15 749**	**15 709**	**16 423**	**15 949**	**14 657**	**14 029**	**16 238**	**15 637**	**15 627**
	Median rank	**12 309**	**11 345**	**27 857**	**9 165**	**9 711**	**9 902**	**8 420**	**9 954**	**9 004**	**8 239**	**8 805**	**8 744**	**9 914**	**9 367**	**9 769**	**9 120**	**11 168**	**8 659**	**8 621**
	Rank std. dev.	**19 284**	**19 239**	**17 121**	**19 844**	**19 861**	**19 926**	**20 107**	**20 166**	**20 269**	**20 084**	**20 002**	**20 030**	**19 929**	**18 182**	**18 462**	**18 466**	**18 726**	**20 006**	**20 008**

both the citing and cited author prior to a citation. All in all, there are many better indicators than PageRank achieved and these are highlighted. There are more of them in the time-aware methods than in the time-unaware ones. (The ratio is 21 to 14.) A graphical representation of the results in Table 5 is displayed in Figure 4 (the award winners without ranks do not appear there).

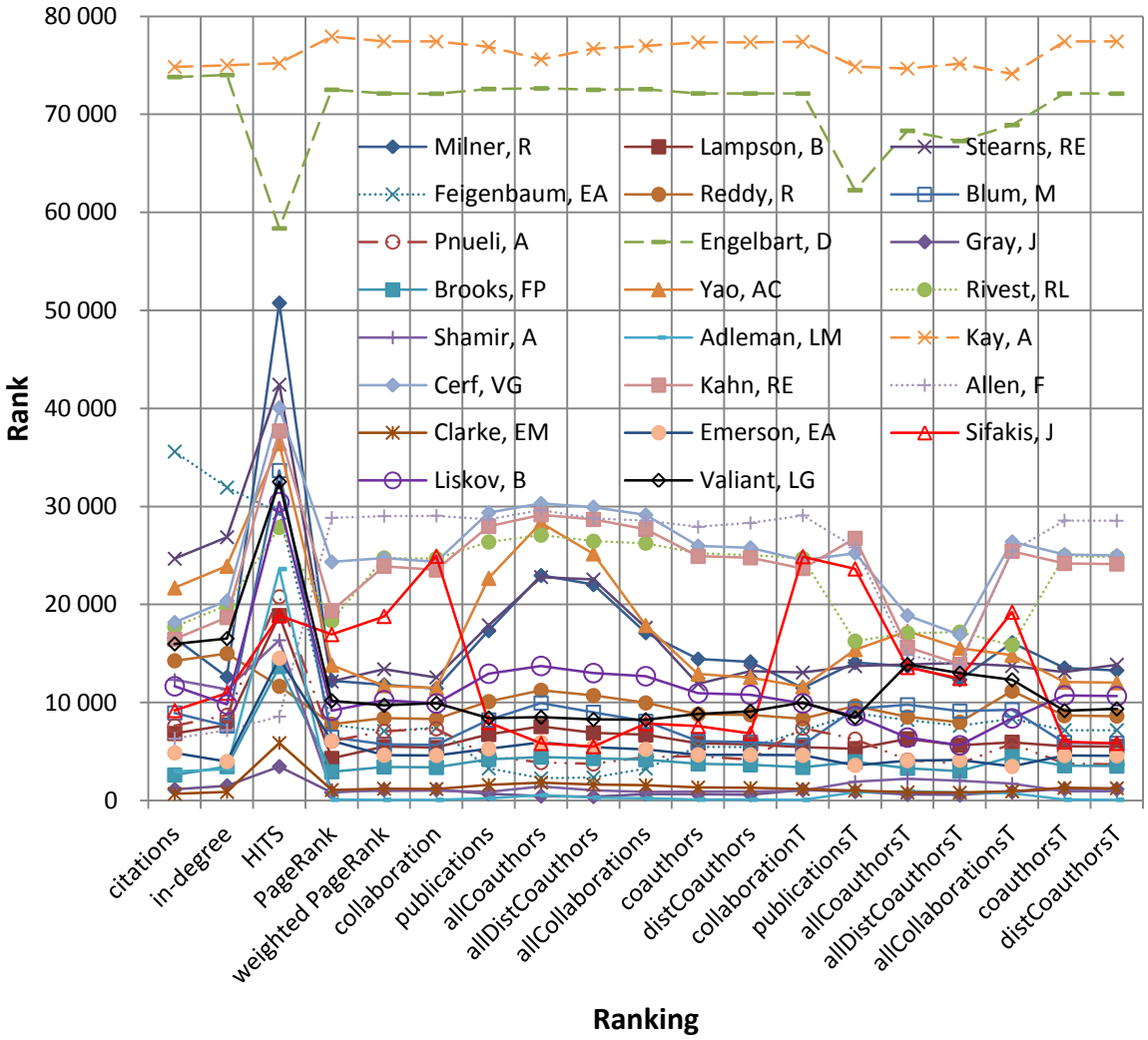


Fig. 4 ACM A. M. Turing Award winners and their ranks in different rankings

In Figure 4 we can see a general slight shift towards better (lower) ranks when moving from left to right, i.e. from citations and in-degree across recursive methods and PageRank modifications to the time-aware variants of PageRank. This would suggest that the time-aware PageRank does reflect prestige perceived by humans (expressed by awards) better than common indicators such as citation counts or the standard PageRank and its weighted variations. Of course, there are some outliers in contradiction with this trend such as "Sifakis, J" and the sudden worsening of his rank with *collaborationT* and *publicationsT* or the overall bad per-

formance of HITS for almost all of the authors, but this may also be interpreted as a feature of that particular ranking. For instance, the relatively bad ranks of "Sifakis, J" reveal that he has relatively frequently collaborated with the researchers citing him (both *collaboration* and *collaborationT*) and that he has written a great number of publications but rather after he was cited, thus having a good rank in *publications* and a bad rank in *publicationsT*. Some other authors, such as "Kay, A" or "Engelbart, D" are very badly ranked by almost all of the methods. This may be caused by the fact that they did not publish in journals in the time period under investigation. And indeed, they both have only three publications in our data set. But as we pointed out earlier, the individual ranks are less important and not discussed here than the overall trend, in which time-aware PageRanks seem to be closer to the "true" ranking than the other indicators.

5.3. ACM SIGMOD E. F. Codd Innovations Award winners

To bring additional evidence that would document the superiority of the time-aware methods over the time-unaware ones, we take advantage of yet another award – ACM SIGMOD E. F. Codd Innovations Award. The award winners from the years 1992 to 2011 are shown in Table 6 along with the ranks achieved in various rankings. ("Bayer, R" was not present in our data and, therefore, was not ranked.) Again, the ranks generated by the standard PageRank are in a darker column and the aggregate indicators yielded by both the time-unaware and time-aware rankings outperforming PageRank are highlighted. For instance, all seven worst ranks by time-aware methods outperform PageRank, but only one time-unaware worst rank does. In total, 24 time-aware indicators are better than PageRank compared to only 8 time-unaware ones. Also in Figure 5 we can see that *allCoauthorsT* and *allDistCoauthorsT* generally produce better ranks for the award winners. The worst ranked researchers, "Kitsuregawa, M" and "Selinger, P", published relatively few journal articles in the time period under study (14 and 3, respectively), but there is no such gap between them and the other laureates as in Figure 4.

The better performance of the time-aware methods over their time-unaware counterparts is further documented in Figures 6 and 7. In Figure 6, the solid blue lines represent best ranks (MIN), worst ranks (MAX), average ranks (AVG), median ranks (MED), and standard rank deviations (DEV) of the time-unaware (standard) PageRank modifications and the dashed red lines represent the time-aware PageRank variants. As for the Turing Award, three dashed lines are below their solid counterparts – MAX, DEV, and AVG. This means that from the point of view of these three indicators the time-aware methods outperform the time-unaware

Table 6 ACM SIGMOD E. F. Codd Innovations Award winners and their ranks

Year	Winner	citations	in-degree	HITS	PageRank	weighted PageRank	collaboration	publications	allCoauthors	allDistCoauthors	allCollaborations	coauthors	distCoauthors	collaborationT	publicationsT	allCoauthorsT	allDistCoauthorsT	allCollaborationsT	coauthorsT	distCoauthorsT
1992	Stonebraker, M	13 816	11 821	15 821	18 112	23 700	23 559	26 392	28 690	27 276	26 242	25 528	25 366	23 671	15 176	11 997	11 690	15 303	23 676	23 661
1993	Gray, J	1 514	1 139	3 469	814	1 103	1 112	674	462	406	668	639	602	1 083	914	626	573	878	1 180	1 165
1994	Bernstein, PA	560	429	2 301	758	808	828	904	990	895	912	877	818	818	849	845	848	885	807	792
1995	DeWitt, DJ	14 670	15 354	20 509	25 963	28 031	27 983	26 237	24 510	25 481	26 211	28 510	28 296	28 058	26 351	25 610	24 888	25 787	27 985	27 888
1996	Mohan, C	12 167	14 863	10 582	9 724	8 948	8 832	10 235	11 758	11 237	10 188	9 415	9 371	8 908	12 515	8 763	6 809	12 227	8 926	8 889
1997	Maier, D	7 954	6 859	10 832	8 760	8 655	9 595	2 077	1 350	1 290	2 032	1 936	1 847	8 568	5 184	4 445	4 357	4 826	8 848	8 788
1998	Abiteboul, S	3 054	3 348	14 794	3 934	3 986	4 361	795	563	722	820	2 271	2 817	4 602	2 391	2 386	2 316	2 280	2 938	3 322
1999	Garcia-Molina, H	1 007	936	2 003	2 442	2 720	2 654	3 842	4 395	4 010	3 790	3 024	2 929	2 659	1 205	1 070	1 009	1 132	2 858	2 829
2000	Agrawal, R	533	395	1 592	458	569	551	807	955	819	789	633	610	553	419	414	323	391	586	595
2001	Bayer, R																			
2002	Selinger, P	70 765	70 420	60 364	58 514	56 330	55 759	63 652	71 922	69 288	63 339	59 105	58 678	55 958	51 998	52 043	51 543	53 608	57 701	57 587
2003	Chamberlin, D	44 810	43 091	41 095	34 935	38 619	37 837	44 033	50 969	49 257	44 038	43 336	42 732	37 616	38 401	34 662	32 568	37 800	42 433	42 013
2004	Fagin, R	1 413	1 033	1 251	327	423	392	1 352	2 504	1 795	1 261	632	562	383	686	655	641	696	540	500
2005	Carey, MJ	5 285	4 209	6 911	5 556	6 328	6 363	6 397	5 028	5 361	6 483	6 299	6 113	6 298	5 919	6 015	5 311	5 728	6 516	6 487
2006	Ullman, JD	16 518	16 855	7 118	24 271	23 886	23 664	27 892	33 144	28 701	27 544	24 246	24 541	23 699	11 647	11 295	11 307	11 580	24 132	23 972
2007	Widom, J	2 676	2 284	4 440	2 250	2 216	2 540	990	872	774	965	776	732	2 162	1 381	1 530	1 464	1 352	2 314	2 254
2008	Vardi, MY	1 369	1 605	16 783	1 066	1 178	1 214	556	534	689	553	1 184	1 203	1 219	727	769	828	713	1 185	1 168
2009	Kitsuregawa, M	75 050	74 905	49 529	70 797	70 681	70 576	72 726	74 800	73 536	72 636	71 108	70 852	70 590	61 126	59 358	59 795	60 820	70 792	70 707
2010	Dayal, U	31 806	29 660	25 204	44 691	45 438	45 368	45 022	43 704	44 106	44 948	45 463	45 436	45 389	36 150	33 212	35 192	35 105	45 321	45 339
2011	Chaudhuri, S	408	268	871	637	790	765	1 143	1 498	1 277	1 117	897	842	770	185	236	231	172	813	800
	Best rank	**408**	**268**	**871**	**327**	**423**	**392**	**556**	**462**	**406**	**553**	**632**	**562**	**383**	**185**	**236**	**231**	**172**	**540**	**500**
	Worst rank	**75 050**	**74 905**	**60 364**	**70 797**	**70 681**	**70 576**	**72 726**	**74 800**	**73 536**	**72 636**	**71 108**	**70 852**	**70 590**	**61 126**	**59 358**	**59 795**	**60 820**	**70 792**	**70 707**
	Average rank	**16 072**	**15 762**	**15 551**	**16 527**	**17 074**	**17 050**	**17 670**	**18 876**	**18 259**	**17 607**	**17 152**	**17 071**	**17 000**	**14 380**	**13 470**	**13 247**	**14 278**	**17 345**	**17 303**
	Median rank	**5 285**	**4 209**	**10 582**	**5 556**	**6 328**	**6 363**	**3 842**	**4 395**	**4 010**	**3 790**	**3 024**	**2 929**	**6 298**	**5 184**	**4 445**	**4 357**	**4 826**	**6 516**	**6 487**
	Rank std. dev.	**22 588**	**22 424**	**16 814**	**20 885**	**20 940**	**20 777**	**22 605**	**24 306**	**23 668**	**22 551**	**21 895**	**21 779**	**20 822**	**18 614**	**18 007**	**18 043**	**18 653**	**21 316**	**21 259**

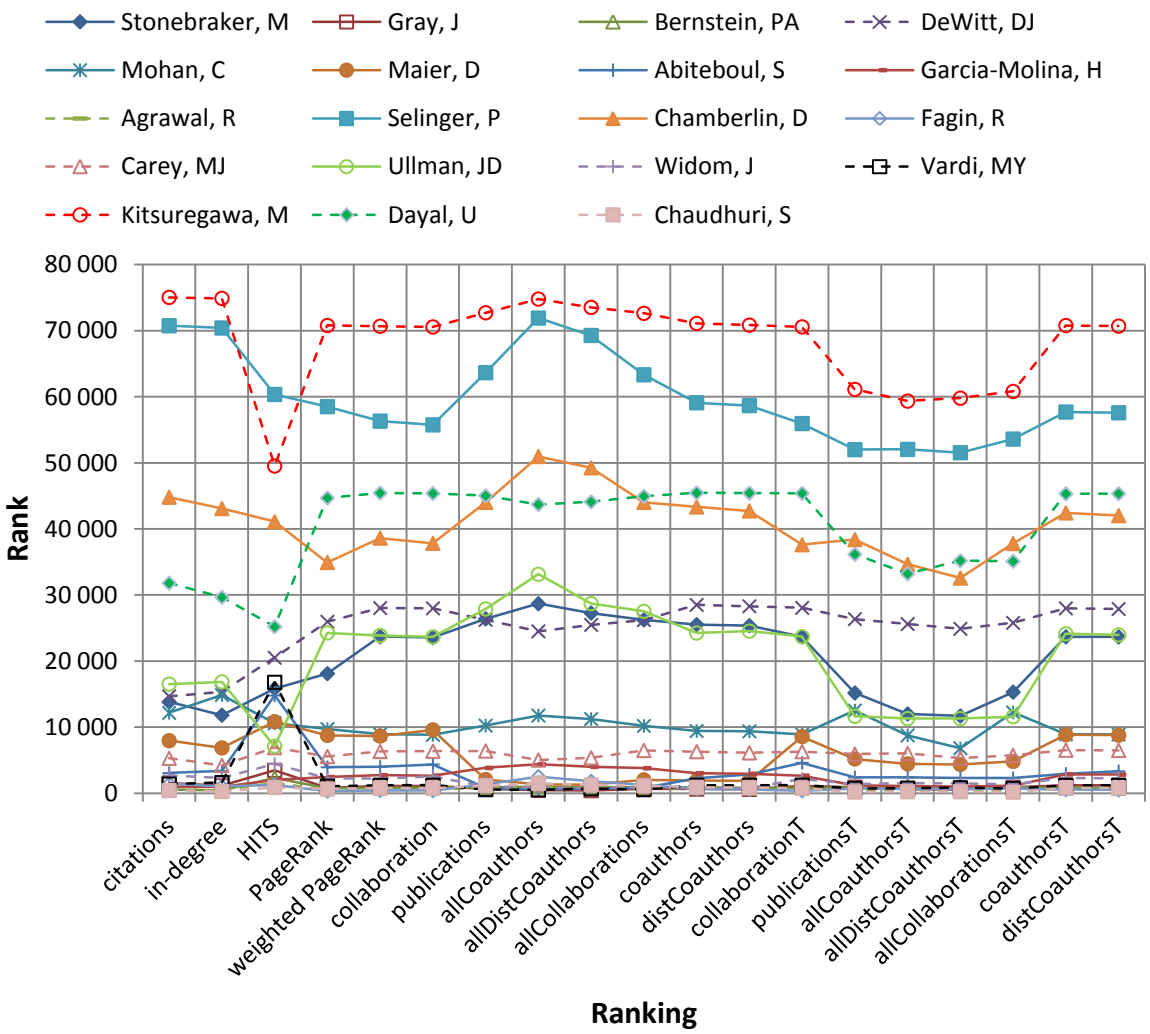


Fig. 5 ACM SIGMOD E. F. Codd Innovations Award winners and their ranks in different rankings

ones by generating lower (i.e. better) ranks for the awardees. As far as the Codd Award is concerned, even four indicators speak in favour of the time-aware methods – MAX, DEV, AVG, and MIN. The only indicator that is worse with both method types is the median rank (MED), which is, however, not very distinct as the solid and dashed lines lie close to each other. In Figure 7 box plots of the time-aware and time-unaware rankings are presented for each pair of rankings. In the case of both awards we can observe that the boxes of the time-aware rankings tend to be placed more towards lower (better) ranks than those of the time-unaware rankings.

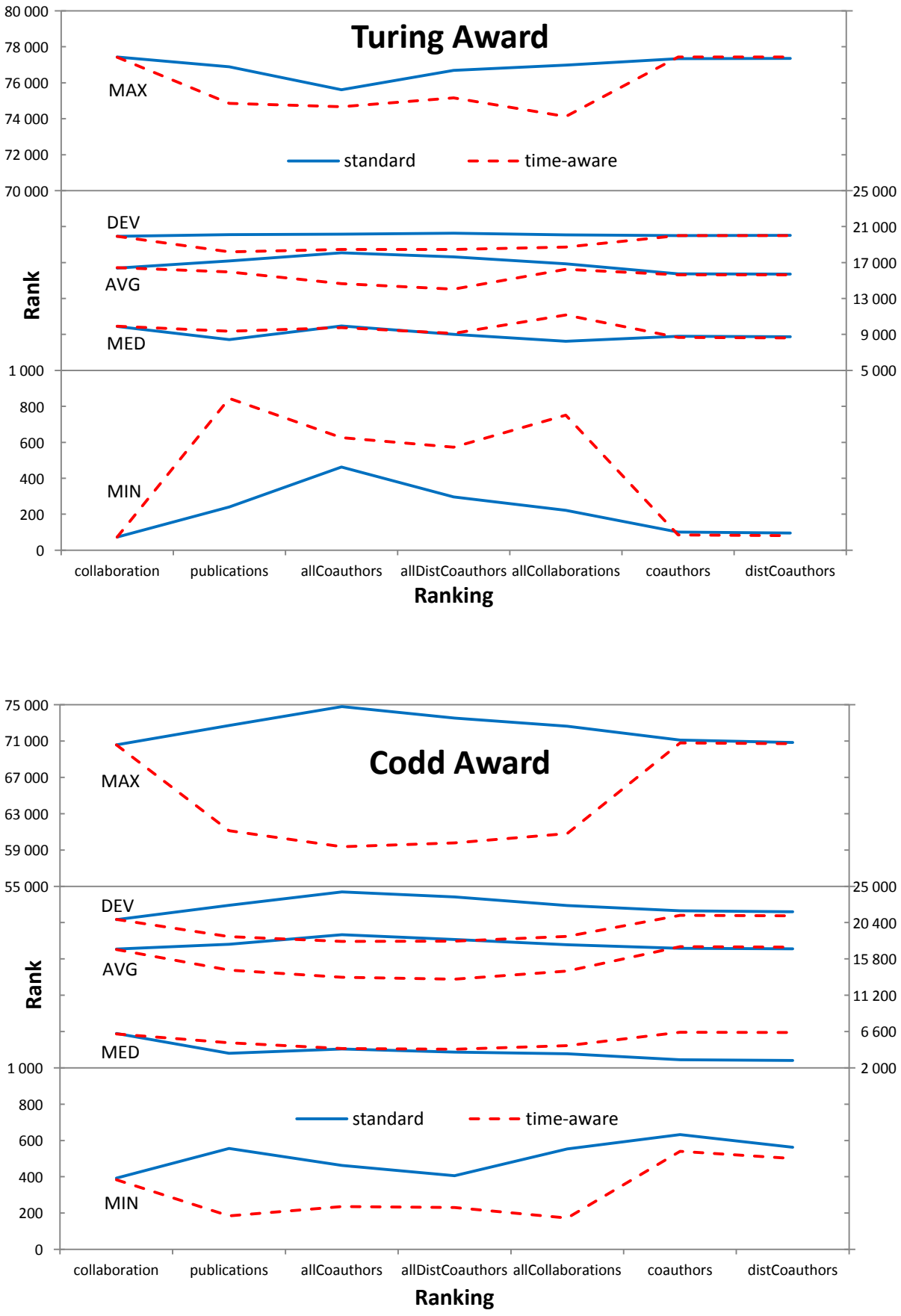


Fig. 6 Aggregate indicators of time-unaware (standard) and time-aware rankings

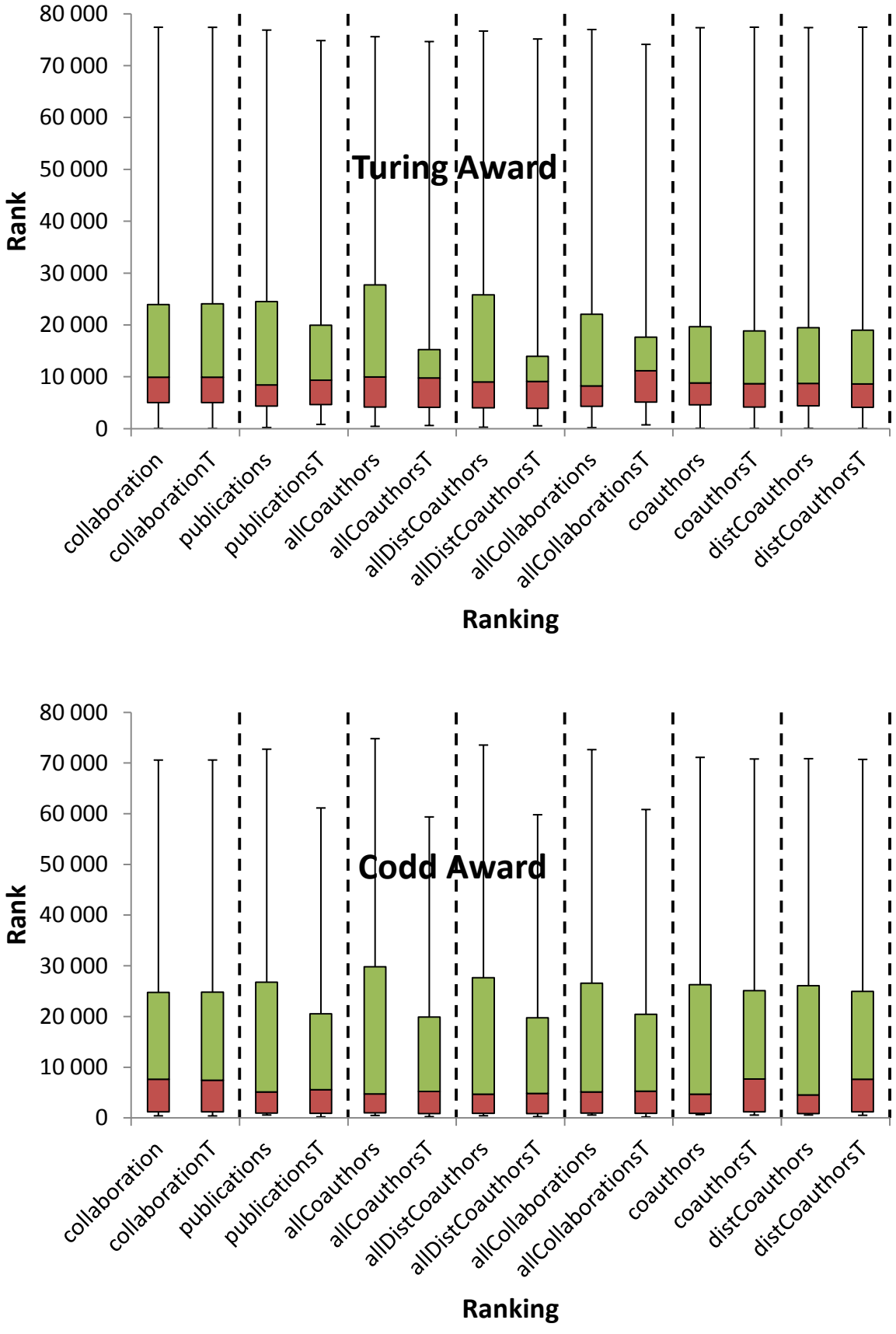


Fig. 7 Box plots of time-unaware and time-aware rankings

6. Conclusions and future work

Algorithms based on the recursive technique called PageRank (Brin and Page, 1998), which was first applied to the Web graph in order to determine the significance of Web pages, have been successfully used in many other situations since then. These methods enable one to evaluate nodes in any directed graphs and rank them according to their importance. In bibliometrics, citation networks of papers or authors, among others, can represent such directed graphs in which the nodes are papers (or authors) and the edges are citations between them. The prominence of researchers has long been detected by first-order methods such as simple citation counts, but it has been shown that popularity, not prestige, is often reflected by citation numbers. On the contrary, higher-order (recursive) methods such as PageRank are able to find prestigious actors that may have fewer citations but from prestigious sources. Also, PageRank-like ranking methods for bibliographic networks can take advantage of the additional information that is not present in a Web graph to weight edges in the network, e.g. co-authorship (Fiala et al., 2008) or time data (Walker et al., 2007, Yan and Ding, 2010, or Yu et al., 2004). Fiala et al. (2008) assigned different weights to the edges in a citation network of authors bearing in mind that a citation from a colleague was less valuable than that from a foreign researcher, but they did not distinguish whether the possible collaboration occurred before the citation was made or afterwards. In this article, we have made an attempt to remedy this situation. The main contributions of the research presented in this paper are as follows:

- We extended the model by Fiala et al. (2008) to incorporate the time of publications (and citations) in their "bibliographic PageRank" to create a "time-aware PageRank" for bibliographic networks. In this model, citations between researchers weight differently depending on a number of factors such as the number of common publications and whether or not they were published before a citation was made.
- We applied seven time-aware PageRank variants along with their time-unaware counterparts and five other common ranking methods (citations, in-degree, HITS, PageRank, and weighted PageRank) to the Web of Science data for computer science journal articles from the period 1996 – 2005 in order to find the most influential computer scientists publishing their work in journals in the decade at the turn of the century.
- We conducted a thorough correlation analysis of the time-aware rankings themselves as well as of the time-aware and time-unaware rankings and other bibliometrics measures such as citations or in-degree. We also compared all the 19 rankings with the lists

of ACM A. M. Turing Award laureates from the years 1991 – 2010 and ACM SIGMOD E. F. Codd Innovations Award winners from the years 1992 - 2011.

Based on our experiments, we achieved the following main results:

- All the 19 rankings are significantly highly positively correlated with each other. The very lowest correlation (around 0.74 of Spearman's rho) was found between HITS authorities and the other PageRank modifications. As for the new time-aware PageRanks, the lowest correlation (0.956), and thus the most added information when compared to its time-unaware counterpart, was observed between the variants in which the number of all co-authors in all publications of both the citing and cited authors are considered.
- The most prominent computer scientists contributing to WoS-indexed journals in the decade 1996 – 2005 detected by citations, in-degree, and HITS are "Jain, AK", "Pentland, A", and "Duin, RPW", whereas those determined by PageRank and all its variants are "Srinivasan, GR", "Murley, PC", and "Ziegler, JF".
- As far as the award winners are concerned, they generally receive better ranks in the time-aware rankings (as can be seen in Figures 4 and 5), but it is impossible to proclaim the "best" ranking because each individual ranking brings an improvement in some aspect (see Tables 5 and 6). However, compared to the standard (unweighted) PageRank in terms of several statistical indicators, the time-aware variants outperform the time-unaware ones (see Tables 5 and 6 and Figures 6 and 7).

For the time-aware PageRank modifications to be more effective, a greater citation window would probably be needed. This would result in a larger number of citations and collaborations of authors in different years. Then, the time-aware and time-unaware rankings should diverge from each other even more than in this study. Therefore, we would like to examine data spanning a greater time period in our future work on this promising topic. Other possibilities of adding more information to citations' weights would include investigating citation loops between authors and assigning less weight to the citations of authors who cite each other.

Acknowledgements

This work was supported by the European Regional Development Fund (ERDF), project "NTIS - New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090. The related software may be found at http://textmining.zcu.cz/

downloads/tarank.php. Thanks are due to the anonymous reviewers for their insightful comments.

References

Bar-Ilan, J. (2010). Web of Science with the Conference Proceedings Citation Indexes: The case of computer science. *Scientometrics*, 83(3), 809-824.

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, 68(5), 314-316.

Bollen, J., Rodriguez, M.A., & Van De Sompel, H. (2006). Journal status. *Scientometrics*, 69(3), 669-687.

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th World Wide Web Conference*, Brisbane, Australia, 107-117.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1), 8-15.

Codd Award. http://www.sigmod.org/sigmod-awards/sigmod-awards#innovations.

Ding, Y. (2011). Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*, 62(2), 236-245.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11), 2229-2243.

Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks. *Scientometrics*, 76(1), 135-158.

Fiala, D. (2011). Mining citation information from CiteSeer data. *Scientometrics*, 86(3), 553-562.

Fiala, D. (in press). Bibliometric analysis of CiteSeer data for countries. *Information Processing and Management*, doi:10.1016/j.ipm.2011.10.001.

Franceschet, M. (2010). The Role of Conference Publications in CS. *Communications of the ACM*, 53(12), 129-132.

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

Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5), 604-632.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing and Management*, 44(2), 800-810.

Sidiropoulos, A., & Manolopoulos, Y. (2005). A citation-based system to assist prize awarding. *SIGMOD Record*, 34(4), 54–60.

Turing Award. http://awards.acm.org/homepage.cfm?srt=all&awd=140.

Wainer, J., Goldenstein, S., & Billa, C. (2011). Invisible Work in Standard Bibliometric Evaluation of Computer Science. *Communications of the ACM*, 54(5), 141-148.

Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 6, art. no. P06010.

Xing, W., & Ghorbani, A. (2004). Weighted PageRank algorithm. In *Proceedings of the 2nd Annual Conference on Communication Networks and Services Research*, Fredericton, Canada, 305-314.

Yan, E., & Ding, Y. (2010). Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*, 61(8), 1635-1643.

Yan, E., & Ding, Y. (2011). Discovering author impact: A PageRank perspective. *Information Processing and Management*, 47(1), 125-134.

Yu, P.S., Li, X., & Liu, B. (2004). On the Temporal Dimension of Search. In *Proceedings of the 13th World Wide Web Conference*, New York, USA, 1180-1181.

Appendix A

Table A.1 Top 50 researchers by both kinds of rankings (part 1)

	collaboration	collaborationT	publications	publicationsT	allCoauthors	allCoauthorsT
1	Srinivasan, GR	Srinivasan, GR	Srinivasan, GR	Srinivasan, GR	Sudan, M	Srinivasan, GR
2	Murley, PC	Murley, PC	Ziegler, JF	Jain, AK	Ziegler, JF	Murley, PC
3	Tang, HHK	Ziegler, JF	Verdu, S	Murley, PC	Verdu, S	Jain, AK
4	Freeman, LB	Freeman, LB	Sudan, M	Ziegler, JF	Srinivasan, GR	Ziegler, JF
5	Ziegler, JF	Tang, HHK	Shamai, S	Freeman, LB	Sapiro, G	Freeman, LB
6	Leinen, P	Leinen, P	Murley, PC	Tang, HHK	Osher, S	Tang, HHK
7	Bey, J	Bey, J	Freeman, LB	Sudan, M	Shamai, S	Sudan, M
8	Juang, JG	Juang, JG	Tse, DNC	Shamai, S	Jain, AK	Calderbank, AR
9	Juang, HG	Juang, HG	Jain, AK	Tse, DNC	Tse, DNC	Renegar, J
10	Korec, I	Korec, I	Osher, S	Calderbank, AR	Bartlett, PL	Shamai, S
11	Curtis, HW	Curtis, HW	Sapiro, G	Cimino, JJ	Lin, YB	Pentland, A
12	Montrose, CJ	Montrose, CJ	Tarokh, V	Pentland, A	Kschischang, FR	Tse, DNC
13	Muhlfeld, HP	Muhlfeld, HP	Vardy, A	Kanade, T	Cimino, JJ	Osher, S
14	OGorman, TJ	OGorman, TJ	Kschischang, FR	Breiman, L	Vardy, A	Sapiro, G
15	Ross, JM	Ross, JM	Tang, HHK	Tarokh, V	Shortliffe, EH	Cimino, JJ
16	Wiener, N	Wiener, N	McEliece, RJ	Sapiro, G	Scholkopf, B	Gupta, A
17	Cegielski, P	Cegielski, P	Cimino, JJ	Sejnowski, TJ	Bates, DW	Sejnowski, TJ
18	Taber, AH	Taber, AH	Leinen, P	Gupta, A	McEliece, RJ	Viergever, MA
19	Walsh, JL	Walsh, JL	Lapidoth, A	Verdu, S	Tarokh, V	Kikinis, R
20	Muses, C	Muses, C	Arora, S	Lee, J	Arora, S	Kanade, T
21	Litkowski, KC	Litkowski, KC	Oja, E	Osher, S	Bro, R	Tarokh, V
22	McTavish, DG	McTavish, DG	Schapire, RE	Leinen, P	Duin, RPW	Lee, J
23	Gazarik, MJ	Gazarik, MJ	Bartlett, PL	Jain, R	Oja, E	Schapire, RE
24	Kamen, EW	Kamen, EW	Mesiar, R	Jordan, MI	Zuckerman, D	Yu, PS
25	Prou, JM	Prou, JM	Yager, RR	Yu, PS	Lapidoth, A	Freund, RM
26	Wagneur, E	Wagneur, E	Bro, R	Viergever, MA	Marzetta, TL	Alon, N
27	Ristow, GH	Ristow, GH	Marzetta, TL	MacKay, DJC	Mesiar, R	Amari, S
28	Fidelman, U	Fidelman, U	Forney, GD	Yager, RR	Overhage, JM	Motwani, R
29	Simon, DR	Simon, DR	Zuckerman, D	Schapire, RE	Freeman, LB	Bates, DW
30	Renegar, J	Renegar, J	Bey, J	Amari, S	Schapire, RE	McDonald, CJ
31	Robinson, DL	Robinson, DL	Warmuth, MK	Alon, N	Forney, GD	Jordan, MI
32	Myers, JS	Myers, JS	Shortliffe, EH	Vardy, A	Shahar, Y	Verdu, S
33	Sampson, G	Sampson, G	Scholkopf, B	Feige, U	Shu, CW	Paxson, V
34	Thomason, A	Thomason, A	Helleseth, T	Richardson, TJ	Kimmel, R	Vardy, A
35	Yngve, VH	Yngve, VH	Lin, YB	Bey, J	Musen, MA	Jain, R
36	Vazirani, U	Vazirani, U	Amari, S	Motwani, R	Yager, RR	Hill, DLG
37	Bernstein, E	Bernstein, E	Sharir, M	Goldreich, O	Smola, AJ	Scholkopf, B
38	Wang, WY	Schwarzer, S	Duin, RPW	Renegar, J	Warmuth, MK	Muller, KR
39	Schwarzer, S	Wachmann, B	Hochwald, BM	Szeliski, R	Amari, S	Ross, JM
40	Wachmann, B	Wang, WY	Kimmel, R	Picard, RW	Williamson, RC	Curtis, HW
41	Russell, CA	Russell, CA	Jordan, MI	Kittler, J	Long, PM	Montrose, CJ
42	Chin, B	Chin, B	Long, PM	Bartlett, PL	Linder, T	Muhlfeld, HP
43	Enger, TA	Enger, TA	Calderbank, AR	Paxson, V	Helleseth, T	OGorman, TJ
44	Hosier, P	Hosier, P	Williamson, DP	Hyvarinen, A	Maass, W	Bartlett, PL
45	Klein, WA	Klein, WA	Shahar, Y	Sharir, M	Campbell, KE	Leinen, P
46	LaFave, LE	LaFave, LE	Freund, Y	Hochwald, BM	Chlamtac, I	Towsley, D
47	Messina, B	Messina, B	Shu, CW	Tanaka, K	Jordan, MI	Kim, J
48	Nicewicz, M	Nicewicz, M	Renegar, J	Black, MJ	Greenes, RA	Willinger, W
49	Orro, JM	Orro, JM	Szegedy, M	Arora, S	Williamson, DP	Chute, CG
50	Scott, TS	Scott, TS	Maass, W	Kim, J	Fang, YG	Breiman, L

Table A.2 Top 50 researchers by both kinds of rankings (part 2)

	allDistCoauthors	allDistCoauthorsT	allCollaborations	allCollaborationsT
1	Ziegler, JF	Srinivasan, GR	Srinivasan, GR	Srinivasan, GR
2	Srinivasan, GR	Murley, PC	Ziegler, JF	Jain, AK
3	Sudan, M	Ziegler, JF	Verdu, S	Murley, PC
4	Freeman, LB	Freeman, LB	Sudan, M	Ziegler, JF
5	Osher, S	Tang, HHK	Murley, PC	Freeman, LB
6	Sapiro, G	Jain, AK	Shamai, S	Tang, HHK
7	Verdu, S	Sudan, M	Freeman, LB	Sudan, M
8	Shamai, S	Renegar, J	Jain, AK	Shamai, S
9	Jain, AK	Calderbank, AR	Tse, DNC	Tse, DNC
10	Tse, DNC	Pentland, A	Osher, S	Renegar, J
11	Kschischang, FR	Shamai, S	Sapiro, G	Calderbank, AR
12	McEliece, RJ	Gupta, A	Tang, HHK	Tarokh, V
13	Tarokh, V	Tse, DNC	Tarokh, V	Kanade, T
14	Arora, S	Freund, RM	Vardy, A	Pentland, A
15	Vardy, A	Sapiro, G	Kschischang, FR	Cimino, JJ
16	Cimino, JJ	Kanade, T	McEliece, RJ	Sapiro, G
17	Zuckerman, D	Alon, N	Leinen, P	Sejnowski, TJ
18	Bates, DW	Sejnowski, TJ	Cimino, JJ	Osher, S
19	Shortliffe, EH	Lee, J	Arora, S	Gupta, A
20	Marzetta, TL	Cimino, JJ	Oja, E	Verdu, S
21	Oja, E	Osher, S	Schapire, RE	Jordan, MI
22	Bartlett, PL	Kikinis, R	Lapidoth, A	Lee, J
23	Bro, R	Leinen, P	Bartlett, PL	Viergever, MA
24	Schapire, RE	Ross, JM	Marzetta, TL	Yu, PS
25	Forney, GD	Curtis, HW	Bro, R	Freund, RM
26	Scholkopf, B	Montrose, CJ	Bey, J	Schapire, RE
27	Murley, PC	Muhlfeld, HP	Forney, GD	Vardy, A
28	Lapidoth, A	OGorman, TJ	Warmuth, MK	Leinen, P
29	Leinen, P	Motwani, R	Zuckerman, D	Amari, S
30	Amari, S	Schapire, RE	Scholkopf, B	Alon, N
31	Overhage, JM	Tarokh, V	Sharir, M	Motwani, R
32	Szegedy, M	Viergever, MA	Helleseth, T	Goldreich, O
33	Shu, CW	Jordan, MI	Shortliffe, EH	Bartlett, PL
34	Williamson, DP	Paxson, V	Duin, RPW	Szeliski, R
35	Warmuth, MK	Kim, J	Amari, S	Kittler, J
36	Duin, RPW	Amari, S	Hochwald, BM	Breiman, L
37	Lin, YB	Vardy, A	Lin, YB	Feige, U
38	Kimmel, R	Bey, J	Kimmel, R	Hochwald, BM
39	Shahar, Y	Lakshman, TV	Jordan, MI	Bey, J
40	Helleseth, T	Feige, U	Calderbank, AR	Sharir, M
41	Jordan, MI	Yu, PS	Mesiar, R	Black, MJ
42	Campbell, KE	Arora, S	Williamson, DP	Jain, R
43	Long, PM	Breiman, L	Shu, CW	Towsley, D
44	Sharir, M	Willinger, W	Long, PM	Lakshman, TV
45	Musen, MA	Vera, JR	Renegar, J	Kim, J
46	Freund, Y	Tanaka, K	Freund, Y	Paxson, V
47	Bey, J	Shor, PW	Shahar, Y	Richardson, TJ
48	Hochwald, BM	Verdu, S	Szegedy, M	Tanaka, K
49	Calderbank, AR	Jain, R	Breiman, L	Kikinis, R
50	Greenes, RA	Hill, DLG	Bates, DW	MacKay, DJC

Table A.3 Top 50 researchers by both kinds of rankings (part 3)

	coauthors	coauthorsT	distCoauthors	distCoauthorsT
1	Srinivasan, GR	Srinivasan, GR	Srinivasan, GR	Srinivasan, GR
2	Ziegler, JF	Murley, PC	Ziegler, JF	Murley, PC
3	Freeman, LB	Ziegler, JF	Freeman, LB	Ziegler, JF
4	Murley, PC	Freeman, LB	Murley, PC	Freeman, LB
5	Tang, HHK	Tang, HHK	Tang, HHK	Tang, HHK
6	Leinen, P	Leinen, P	Leinen, P	Leinen, P
7	Bey, J	Bey, J	Bey, J	Bey, J
8	Juang, JG	Juang, JG	Juang, JG	Juang, JG
9	Juang, HG	Juang, HG	Juang, HG	Juang, HG
10	Wiener, N	Wiener, N	Wiener, N	Wiener, N
11	Korec, I	Curtis, HW	Korec, I	Curtis, HW
12	Cegielski, P	Montrose, CJ	Cegielski, P	Montrose, CJ
13	Curtis, HW	Muhlfeld, HP	Curtis, HW	Muhlfeld, HP
14	Montrose, CJ	OGorman, TJ	Montrose, CJ	OGorman, TJ
15	Muhlfeld, HP	Ross, JM	Muhlfeld, HP	Ross, JM
16	OGorman, TJ	Korec, I	OGorman, TJ	Korec, I
17	Ross, JM	Cegielski, P	Ross, JM	Cegielski, P
18	Renegar, J	Taber, AH	Renegar, J	Taber, AH
19	Sudan, M	Walsh, JL	Muses, C	Walsh, JL
20	Schapire, RE	Muses, C	Simon, DR	Muses, C
21	Simon, DR	Renegar, J	Litkowski, KC	Renegar, J
22	Muses, C	Litkowski, KC	McTavish, DG	Litkowski, KC
23	Litkowski, KC	Simon, DR	Gazarik, MJ	McTavish, DG
24	McTavish, DG	McTavish, DG	Kamen, EW	Simon, DR
25	Vazirani, U	Gazarik, MJ	Prou, JM	Gazarik, MJ
26	Bernstein, E	Kamen, EW	Wagneur, E	Kamen, EW
27	Taber, AH	Prou, JM	Sudan, M	Prou, JM
28	Walsh, JL	Wagneur, E	Taber, AH	Wagneur, E
29	Ristow, GH	Fidelman, U	Walsh, JL	Fidelman, U
30	Fidelman, U	Ristow, GH	Ristow, GH	Ristow, GH
31	Gazarik, MJ	Vazirani, U	Fidelman, U	Vazirani, U
32	Kamen, EW	Bernstein, E	Vazirani, U	Bernstein, E
33	Prou, JM	Robinson, DL	Bernstein, E	Robinson, DL
34	Wagneur, E	Myers, JS	Schapire, RE	Myers, JS
35	Bennett, CH	Sampson, G	Bennett, CH	Sampson, G
36	Shamai, S	Thomason, A	Robinson, DL	Thomason, A
37	Osher, S	Yngve, VH	Breiman, L	Yngve, VH
38	Breiman, L	Sudan, M	Myers, JS	Sudan, M
39	Jain, AK	Bennett, CH	Sampson, G	Wang, WY
40	Tarokh, V	Wang, WY	Thomason, A	Schwarzer, S
41	Myers, JS	Breiman, L	Yngve, VH	Wachmann, B
42	Sampson, G	Schwarzer, S	Jain, AK	Russell, CA
43	Thomason, A	Wachmann, B	Shamai, S	Chin, B
44	Yngve, VH	Russell, CA	Schwarzer, S	Enger, TA
45	Robinson, DL	Chin, B	Wachmann, B	Hosier, P
46	Calderbank, AR	Enger, TA	Calderbank, AR	Klein, WA
47	Sapiro, G	Hosier, P	Tarokh, V	LaFave, LE
48	Freund, Y	Klein, WA	Freund, RM	Messina, B
49	Schwarzer, S	LaFave, LE	McEliece, RJ	Nicewicz, M
50	Wachmann, B	Messina, B	Behbehani, J	Orro, JM

Article 2

The next article deals with CiteSeer, which is a digital library (with a Web search interface) covering mainly computer science and related fields. CiteSeer possesses a Web spider that crawls the Web at more or less regular time intervals and collects potentially relevant publicly available documents, i.e. PDF and PostScript files that look like computer science research papers. The recognition whether or not a document is a research paper is done in quite a simple way – whenever a document is structured similarly to a scientific article (e.g., it contains an introduction, sections on related work, methods, data, results, conclusions and a list of references), it is considered a research paper. Moreover, to include computer science literature only, it must be aware of the well-known computer science paper repositories and it must employ a classifier to categorize documents. In addition to crawling the World Wide Web, CiteSeer also accepts Web addresses of institutional repositories or of individual users' repositories to retrieve documents from.

Besides converting PDF and PostScript files into plain text, classifying documents into research papers and non-research publications, and categorizing research papers into computer science articles and non-computer science articles, CiteSeer needs to perform various information retrieval tasks with the papers collected. Most importantly, each article shall be provided with metadata describing its title, authors, authors' addresses and affiliations, abstract, and cited references. These metadata accompany the full text of each paper in CiteSeer. All the processes (crawling, collecting, converting, parsing, classifying, and creating metadata) are carried out automatically using many machine learning techniques. This, on the one hand, allows for huge amounts of data to be processed in very short time intervals and at very low costs, but, on the other, enables errors inherent to automated text processing techniques to emerge. Such errors may, for instance, include the recognition of "PhD student" as an author's name or "ACM Fellow" as an author's address.

As a result, CiteSeer data have been very rarely used in bibliometric analyses. This stands in a stark contrast to manually created bibliographic databases like Web of Science and

Scopus. These databases, which primarily serve as publication and citation indices of general scientific literature, are manually created and maintained employing a great deal of human labour. Therefore, their expansion is relatively slow compared to the exponential growth of scientific research output and very costly. On the other hand, they are expected to be (almost) error-free and their data are commonly used as data sources for numerous bibliometric studies. Thus, the main challenge is to show that the freely available data from CiteSeer can be successfully used for scientometric purposes as a complement to the subscription-based databases Web of Science and Scopus. In the following study, I demonstrate this by analyzing CiteSeer data and measuring the research productivity and performance of countries. I use a couple of established scientometric indicators and compare the results from CiteSeer to those from Web of Science and Scopus, which I retrieved manually[7]. I conclude that CiteSeer data can be used in bibliometric research and include a list of the top 30.countries by out-degree and references in Table 1 as opposed to in-degree and citations described in the next article.

	Country	Out-degree	Country	References
1	USA	80	USA	297640
2	United Kingdom	69	Germany	189568
3	Germany	68	France	115969
4	France	67	United Kingdom	107670
5	Canada	61	Italy	92254
6	Italy	60	Canada	85547
7	Netherlands	57	Netherlands	57844
8	Spain	56	Switzerland	46971
9	Sweden	55	Australia	45955
10	Australia	53	Japan	45548
11	Austria	53	Spain	42700
12	Japan	53	Sweden	33541
13	Belgium	50	Israel	32384
14	Greece	50	Belgium	28134
15	Switzerland	50	Austria	28111
16	Portugal	49	Greece	21191
17	Denmark	48	Finland	19510
18	Finland	48	Brazil	18251
19	Israel	48	Denmark	18041
20	Brazil	47	India	17744
21	Singapore	45	Hong Kong	17198
22	Hong Kong	42	Portugal	16503
23	New Zealand	42	Singapore	15461
24	Czech Republic	41	Taiwan	13316
25	Norway	41	China	11850
26	India	40	Korea	10192
27	Ireland	40	Ireland	8676
28	Russia	40	Norway	6677
29	China	39	New Zealand	6506
30	Poland	38	Hungary	6000

Table 1: Top 30 countries in CiteSeer by out-degree and references

[7] Nowadays, Scopus data for countries can be obtained from http://scimagojr.com/.

Preprint of: Fiala, D. (2012). Bibliometric analysis of CiteSeer data for countries. *Information Processing and Management*, 48(2), 242-253.

Bibliometric analysis of CiteSeer data for countries

Dalibor Fiala

University of West Bohemia, Univerzitní 8, 30614 Plzeň, Czech Republic
Phone: +420 377 63 24 29, fax: +420 377 63 24 01, email: dalfia@kiv.zcu.cz

Abstract: This article describes the results of our analysis of the data from the CiteSeer digital library. First, we examined the data from the point of view of source top-level Internet domains from which the data were collected. Second, we measured country shares in publications indexed by CiteSeer and compared them to those based on mainstream bibliographic data from the Web of Science and Scopus. And third, we concentrated on analyzing publications and their citations aggregated by countries. This way, we generated rankings of the most influential countries in computer science using several non-recursive as well as recursive methods such as citation counts or PageRank. We conclude that even if East Asian countries are underrepresented in CiteSeer, its data may well be used along with other conventional bibliographic databases for comparing the computer science research productivity and performance of countries.

Keywords: CiteSeer, CiteSeer^{X}, citations, shares, countries, Internet domains.

1. Introduction

CiteSeer (CiteSeer) is a vast free Web digital library and search engine of mainly computer science papers that have been automatically acquired from various Web sites, stored, and analyzed to allow for searching and exploring its bibliographic data. Despite its free on-line as well as off-line availability and well structured data, it has been relatively rarely used in bibliometric studies particularly due to fears of incomplete and erroneous machine-generated data. We refer to the work by Fiala (2011) where a detailed overview of CiteSeer's features in the context of other established bibliographic databases is given.

The purpose of this study is to show: a) where CiteSeer has got its data (i.e. which Web domains it has visited to obtain them), b) which countries have contributed most to its digital library (in terms of the number of papers published by authors from these countries), and c) which countries have the most influence (in terms of citedness of "their" publications). We have thoroughly analyzed the CiteSeer data file from December 13, 2005 and have made

a quick look at the newer data provided by CiteSeer^X (CiteSeer^X) which replaced CiteSeer in April 2010 but is still a beta version at the time of writing this article (May 2011).

2. Related work

There have been a number of studies of research productivity (publications) and impact (citations) at the level of countries in recent years. There is a growing need for such scientometric indicators because they often reflect the quality of science policy in a specific country and may have influence on changes in science funding. From the many research papers discussing this topic, let us mention just one of the most recent by Albarrán et al. (2010), which compares the United States to the European Union in a detailed way in various fields of science.

While quite a lot of research efforts have been devoted to bibliometrics of chemistry, biology, or humanities, relatively few scientometric studies have been concerned with the field of computer science. Bakri & Willett (2011) measure the performance of computer science research in Malaysia and Gupta et al. (2011) analyze the research output of Indian computer science. Wainer et al. (2009) compared the Brazilian computer science production to twelve other countries. Ma et el. (2008) did not limit their analysis to a particular country but evaluated the computer science research performance of universities around the globe and Guan & Ma (2004) evaluated China and five other countries. Different sources of bibliographic data for the scientometric evaluation of computer science publications were examined by Bar-Ilan (2010) and by Franceschet (2010). The latter author also presents an overview of literature comparing citation data from various data sources for a specific scientific field. Furthermore, Franceschet (2010b) investigated the influence of computer science journal and conference papers on the scientific community.

Unlike our paper, most of the articles above have mainly exploited the well-known and manually-maintained bibliographic database Web of Science (Web of Science) or its variants. As far as CiteSeer as a data source is concerned, some researchers have already used it for bibliometric purposes: Zhou et al. (2007) explored CiteSeer documents to discover temporal communities of collaborating authors in the domains of databases and machine learning. On the other hand, Hopcroft et al. (2004) tracked evolving communities in the whole CiteSeer paper citation graph. An et al. (2004) conducted a component analysis of the CiteSeer paper citation graph in several research domains and CiteSeer^X data were used by Wu et al. (2010) in order to enhance collaborative networks with topic information. Zhao & Strottman (2007) and Zhao & Logan (2002) analyzed co-citations in CiteSeer documents in the XML research field and a similar study for computer graphics was reported by Chen (2000). Bar-Ilan (2006)

used CiteSeer data for a citation analysis of the works of a famous mathematician. A kind of citation analysis for acknowledgements was also performed by Giles & Councill (2004). Feitelson & Yovel (2004) examined citation ranking lists obtained from CiteSeer and predicted future rankings of authors.

Our study is the first of its kind that attempts to measure the productivity and impact of computer science research conducted by countries by analyzing CiteSeer data.

3. Data

The last CiteSeer data originate from December 2005 and they contain roughly 717 thousand publications with 1.8 million references within CiteSeer. On the other hand, CiteSeer^X (data from March 2011) provides more than 1.3 million publications with almost 15 million references within CiteSeer^X. This means that the citation graph with publications as nodes and references as edges has become much denser over the past six years – the mean number of references in a publication increased from 2.5 in 2005 to 11.2 in 2011.

Let us have a look at a few obvious differences between CiteSeer (CS) / CiteSeer^X (CS^X) and Web of Science / Scopus (Scopus) – two well-known databases of scientific literature. Both CiteSeer and CiteSeer^X collect (or collected) its data in the same way: they crawl the Web starting from some seed pages submitted by their engineers or by individual users (authors) and pick up freely accessible documents (mostly PDF or PostScript files) that have the potential to be research papers in computer science, mathematics, or related fields. Web crawling as well as information extraction (titles, author names, references, etc.) occurs automatically, without human intervention. The contents of CiteSeer and CiteSeer^X depend generally on the content and structure of the Web. On the other hand, both Web of Science and Scopus use a great deal of human labour to receive publications (mainly journal issues and conference proceedings) and to index them. Unlike CiteSeer and CiteSeer^X, WoS and Scopus cover all scientific fields. Which publication sources are indexed and which are not is decided by the editorial boards of both "human-made" databases. Another big difference between CiteSeer and CiteSeer^X on one side and WoS and Scopus on the other is that the first two are free whereas the latter two are subscription-based.

4. Methods

4.1 Data collection

Data collection methods were different for CiteSeer and for CiteSeer^X. For CiteSeer, there was a single archive data file created in December 2005 (the most recent CiteSeer data) that

we merely downloaded from the CiteSeer Web site and unpacked into 2 GB of 72 XML-like files. As for CiteSeer^X, we were forced to use one of the harvesting tools referenced on its Web site to gain off-line access to its current repository. The harvest itself took a few days in March 2011 and resulted in a regular 3.7 GB XML file which we further split up into 73 files to process them more smoothly in main memory. We developed software[8] that parsed the data files and stored information about publications, authors, and citations in a relational database. We were then able to query the database and obtain the information presented in the following sections. The software also had capabilities to compute more complex values such as HITS and PageRank.

4.2 Internet domains and countries

Gathering statistics about Internet top-level domains (TLD) is quite smooth and accurate given that the "source" property for each document is almost always present and error free. The situation gets considerably worse when we try to assemble similar statistical data for the distribution of countries whose authors produced the publications collected by CiteSeer. As far as CiteSeer^X is concerned, unfortunately, it does not provide any information on the addresses or affiliations of the authors of its publications – not only for "new" publications, but also for "old" publications for which this information is present in CiteSeer. Therefore, we could not use CiteSeer^X data for our experiments with countries. Let us hope that future versions of CiteSeer^X (the current one is still a beta) will have such information included.

4.3 Missing data and name unification

In CiteSeer, there is a problem with missing data. For almost each document, there are authors assigned to it but only for some of the authors there is also an address affiliated with him/her. Strictly said, from the total of 1.66 million authors (without any name unification or disambiguation), we had no address information at our disposal for about 690 thousand or 42% of them, let alone the accuracy of such information.

Thus, to obtain the data shown later in Figure 2, we proceeded in the following way: We discarded publications without any address information for any of its authors. This resulted in only 439 thousand being kept. (For these publications, one author at least had some address information included.) Then, we tried to unify country names used in the addresses. This task consisted in obtaining a list of countries and territories owning a top-level Internet domain. After some cleansing, 243 countries or territories were left. Next, we attempted to

[8] http://textmining.zcu.cz/downloads/sciento.php

unify country names by replacing common synonymic variants of each of those 243 countries with one standard name.

For instance, in the case of the United States of America, we had to count in names like "United States", "U.S.A.", "U.S.A", "U.S.", "USA", or "US". Since U.S. postal addresses often do not contain any mention of "USA" or its variants and only display the name or abbreviation of a federal state such as "California" or "CA", we also needed to take this into account and counted such occurrences as "USA". Other types of unification included considering often independently appearing entities such as England, Scotland, Wales and Northern Ireland as one country (United Kingdom) or, in contrast, keeping territories of one country separate such as Hong Kong, Taiwan, and Macau from China or Reunion and Martinique from France. Finally, we processed international postal country codes in the addresses as well, thus yielding Czech Republic for an address "CZ-30416" with respect to the prefix "CZ-" as an example.

4.4 Comparison with the Web of Science and Scopus

Since the CiteSeer data we examined were from December 2005, we restricted our analysis to a 10-year period from 1996 to 2005. This decade is the most probable one, in which CiteSeer was collecting its documents. Moreover, Scopus itself does not generally capture citations to documents published before 1996, which is also a good reason for 1996 as a decade's start with regard to possible future comparisons of citations. In September 2010, we were querying on-line Web services of both WoS and Scopus and generated the rankings in Tables 3 and 4. As for WoS, we opted to limit our search to the "Science Citation Index Expanded" database, to the "article" document type, and to the publications from the journals included in the seven computer science subject categories of the Journal Citation Reports® Science Edition 2009. In this way, we arrived at the total of 148 838 publications, which is 100% for the relative shares in Table 3. As far as Scopus is concerned, querying was easier in that the subject area (computer science) could be specified directly in the query and the exact results number was always disclosed. The final 325 614 "article" documents form 100% for the relative shares in Table 4. Due to the search limits of both WoS and Scopus, it was sometimes necessary to split up "big" queries into subqueries and to combine their results.

Alternatively, WoS as well as Scopus provide programming interfaces that enable submitting queries and obtaining results without needing to interact with their Web front-ends. However, the basic APIs included in the subscription do have queries and results restrictions that are similar to those on their Web sites.

4.5 Citations and recursive indicators

In addition to measuring shares of individual countries in the publications indexed by CiteSeer, we wished to determine the influence of countries by examining citations they receive. Thus, we derived a citation graph of countries from the citation graph of publications. In the directed publication citation graph, there were 717 thousand nodes (publications) and 1.76 million edges (citations between publications). This accounts for roughly 2.45 citations per paper so, obviously, many citations (or references) are missing in CiteSeer. Let us recall that addresses of publications' authors were normalized by the approach described earlier. We aggregated citations by the country of the source and target publication. If there were more countries associated with a publication, a couple of citations came into being. We removed self-citations of countries as well.

Besides first-order methods such as in-degree and citations, there are recursive techniques as well that not only count citations but take also into account whether the citing node itself is frequently cited. Some of these methods are HITS introduced by Kleinberg (1999), PageRank defined by Brin and Page (1998), or weighted PageRank (e.g., Fiala et al., 2008). We applied these methods to the normalized country citation graph from CiteSeer and present the country rankings obtained in Table 6.

5. Results and discussion

5.1 Internet domains

One of the properties of each document item indexed by CiteSeer is its source. This is the URL (a Web page) from which the document has originally been downloaded. We were interested in the distribution of Internet top-level domains (TLD) among the sources of CiteSeer documents. This would reveal what regions of the Web the CiteSeer Web crawler has visited and to what extent. It might also help explain a possible bias in publication and citations shares of individual countries discovered later.

Figure 1 shows the shares of top twenty top-level Internet domains as sources of CiteSeer and CiteSeer^{X} documents. The charts are quite similar - approximately one third of all publications originate from *.edu* servers, followed by *.de*, *.uk*, *.fr*, and *.com* with the most notable change for *.org*, which grew from 3.62% to 9.42% between 2005 and 2011. Although *.edu*, *.com*, and *.org* domains do not necessarily mean U.S. Web sites, we shall not be too far from the truth if we count them along with *.gov* as U.S. sites and claim that about a half of all CiteSeer documents have been gathered in the United States with a small increase by several

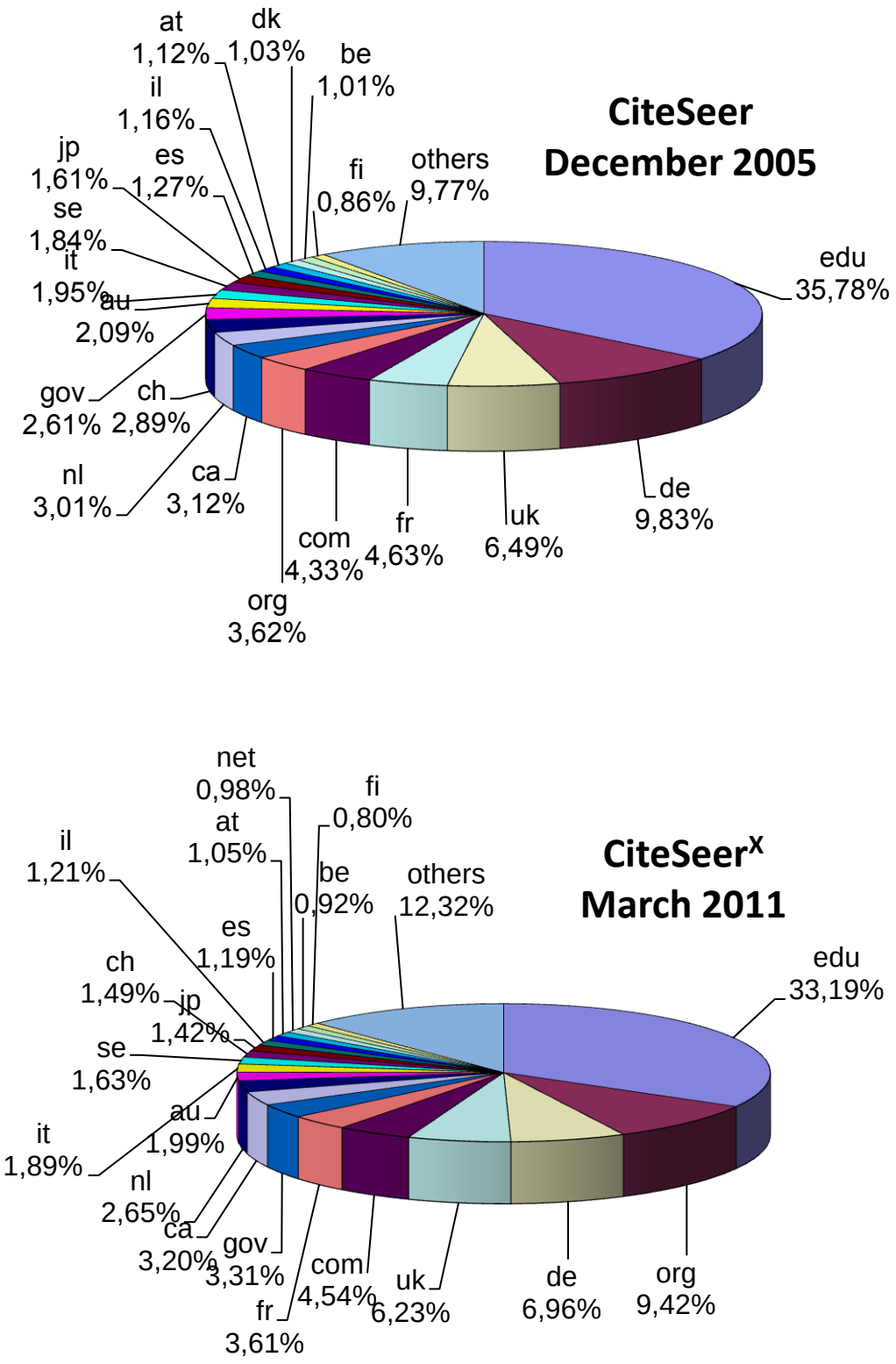


Fig. 1 Shares of Internet domains from which CiteSeer and CiteSeer^{X} documents have been collected

Table 1 Top 100 Internet top-level domains (TLD) by publications in CiteSeer compared to CiteSeer[X]

Dec 2005				March 2011			Dec 2005				March 2011		
No.	TLD	# Pub.	%	No.	# Pub.	%	No.	TLD	# Pub.	%	No.	# Pub.	%
1	edu	256 433	35.78	1	442 756	33.19	51	ua	273	0.04	68	280	0.02
2	de	70 446	9.83	3	92 821	6.96	52	ar	226	0.03	52	739	0.06
3	uk	46 544	6.49	4	83 049	6.23	53	hr	197	0.03	55	546	0.04
4	fr	33 172	4.63	6	48 190	3.61	54	cy	175	0.02	63	338	0.03
5	com	31 051	4.33	5	60 570	4.54	55	yu	165	0.02	58	488	0.04
6	org	25 922	3.62	2	125 650	9.42	56	uy	146	0.02	69	175	0.01
7	ca	22 368	3.12	8	42 671	3.20	57	ee	137	0.02	56	536	0.04
8	nl	21 544	3.01	9	35 411	2.65	58	ir	135	0.02	54	548	0.04
9	ch	20 686	2.89	13	19 908	1.49	59	bg	116	0.02	60	369	0.03
10	gov	18 694	2.61	7	44 179	3.31	60	co	109	0.02	81	79	0.01
11	au	14 976	2.09	10	26 547	1.99	61	ve	105	0.01	72	123	0.01
12	it	13 976	1.95	11	25 188	1.89	62	info	91	0.01	41	2 507	0.19
13	se	13 178	1.84	12	21 721	1.63	63	lv	65	0.01	71	155	0.01
14	jp	11 522	1.61	14	18 911	1.42	64	my	65	0.01	59	462	0.03
15	es	9 092	1.27	16	15 851	1.19	65	py	54	0.01	169	0	0.00
16	il	8 287	1.16	15	16 162	1.21	66	to	54	0.01	66	285	0.02
17	at	8 056	1.12	17	14 013	1.05	67	is	52	0.01	67	284	0.02
18	dk	7 360	1.03	21	10 250	0.77	68	lt	52	0.01	53	581	0.04
19	be	7 270	1.01	19	12 261	0.92	69	ps	51	0.01	65	286	0.02
20	fi	6 145	0.86	20	10 705	0.80	70	lu	46	0.01	75	109	0.01
21	kr	4 791	0.67	29	6 404	0.48	71	mt	30	0.00	77	106	0.01
22	gr	4 336	0.60	22	9 077	0.68	72	mk	27	0.00	87	50	0.00
23	pt	4 229	0.59	24	7 604	0.57	73	lb	26	0.00	72	123	0.01
24	no	3 977	0.55	27	6 697	0.50	74	ma	26	0.00	79	95	0.01
25	br	3 973	0.55	31	6 109	0.46	75	ph	25	0.00	76	107	0.01
26	cz	3 844	0.54	30	6 305	0.47	76	gb	24	0.00	93	24	0.00
27	ie	3 522	0.49	26	6 708	0.50	77	nu	21	0.00	80	91	0.01
28	hk	3 470	0.48	23	7 759	0.58	78	et	18	0.00	106	14	0.00
29	net	2 847	0.40	18	13 091	0.98	79	aero	15	0.00	109	11	0.00
30	mil	2 527	0.35	35	4 054	0.30	80	fm	15	0.00	62	345	0.03
31	nz	2 427	0.34	28	6 448	0.48	81	id	15	0.00	78	96	0.01
32	pl	2 202	0.31	34	4 417	0.33	82	sa	15	0.00	57	514	0.04
33	tw	2 056	0.29	33	4 981	0.37	83	biz	10	0.00	88	49	0.00
34	mx	1 978	0.28	42	2 301	0.17	84	cu	10	0.00	98	20	0.00
35	hu	1 905	0.27	37	3 805	0.29	85	name	10	0.00	64	290	0.02
36	sg	1 725	0.24	32	5 572	0.42	86	rs	10	0.00	102	15	0.00
37	in	1 423	0.20	25	7 342	0.55	87	tc	10	0.00	99	19	0.00
38	cn	1 265	0.18	38	3 396	0.25	88	ws	9	0.00	84	65	0.00
39	tr	1 208	0.17	40	2 800	0.21	89	mu	6	0.00	113	9	0.00
40	ru	1 176	0.16	44	1 892	0.14	90	mo	5	0.00	85	61	0.00
41	cl	1 054	0.15	47	1 657	0.12	91	om	5	0.00	122	5	0.00
42	si	801	0.11	43	1 900	0.14	92	li	4	0.00	113	9	0.00
43	za	785	0.11	45	1 735	0.13	93	tv	4	0.00	96	21	0.00
44	int	621	0.09	39	3 256	0.24	94	ac	3	0.00	110	10	0.00
45	th	474	0.07	51	844	0.06	95	af	3	0.00	126	4	0.00
46	us	462	0.06	36	3 954	0.30	96	cx	3	0.00	100	18	0.00
47	sk	459	0.06	49	1 439	0.11	97	pg	3	0.00	169	0	0.00
48	su	447	0.06	61	353	0.03	98	ae	2	0.00	86	52	0.00
49	cc	333	0.05	48	1 630	0.12	99	am	2	0.00	119	7	0.00
50	ro	277	0.04	50	1 014	0.08	100	ge	2	0.00	102	15	0.00

percentage points from 2005 to 2011. In 2005, only 25 documents had no source URL affiliated with them and they are included in those almost 10% of "other" domains. In 2011, this number is considerably higher – almost 17 thousand – and the share of "other" domains is as much as 12%. A complete list of the top 100 CiteSeer source domains is available in Table 1 with their respective ranks and shares in CiteSeer^X. After a quick look at the table, we may notice that a couple of non-country TLDs have significantly increased their shares such as *.org* (moving from rank 6 to rank 2), *.net* (from 29 to 18), or *.info* (from 62 to 41) while the main country-code TLDs remain relatively stable or even slightly decline. There is one remarkable exception, *.in*, which increases its rank from 37 to 25 and its share from 0.20% to 0.55% between the years 2005 and 2011. In this context, it is interesting to see that the postion of *.cn* (38) remains unchanged in both CiteSeer and CiteSeer^X.

Nowadays, most open access repositories are located within North America and Europe (Repository66) and, therefore, it is logical that even Asian researchers might prefer placing their manuscripts in the repositories of these regions, which further increases the prevalence of American and European top-level Internet domains crawled by CiteSeer.

5.2 Countries

After unifying country names in the available addresses as described in Section 4.3, we tried to assign all 439 thousand publications to one or more country depending on how many authors from which countries they had. About 25 thousand publications could not be assigned to any country, i.e. it was impossible to make use of the information in their address field to identify a standard country by the above approach. Thus, only 414 thousand documents (58% of 717 thousand) were finally assigned to one or more country. We counted the assignments to countries and found out country shares that are demonstrated relatively as well as absolutely in Figure 2 and in Table 2. Note, however, that the relative shares in Figure 2 differ from those presented in Table 2.

The relative shares in Figure 2 sum up to 100% constituted by a total of 449 thousand publication-country assignments, which is not equal to 414 thousand publications due to international co-authorships. (Albarrán et al. (2010) call the publication-country assignments "extended articles".) Even though the number of such assignments is only less than 10% greater than that of publications, it does not necessarily imply a relatively low number of international publications in CiteSeer. We may rather assume that addresses in international papers are more difficult to be processed by a machine (CiteSeer) and, therefore, they are often missing or erroneous and do not appear in our cleansed data.

In Figure 2, the top twenty most represented countries take almost 93% of "extended articles". The first country is the United States with a four-fold greater share (42.59%) than the second most "prolific" country – Germany (10.65%). At the third position, there is a tie between France and the United Kingdom (both 5.35%). As a remarkable point, two developing countries have entered the Top 20 – India and Brazil with shares of 0.67% and 0.64%, respectively. The number (or share) of publications not assigned to any country is not visible in Figure 2.

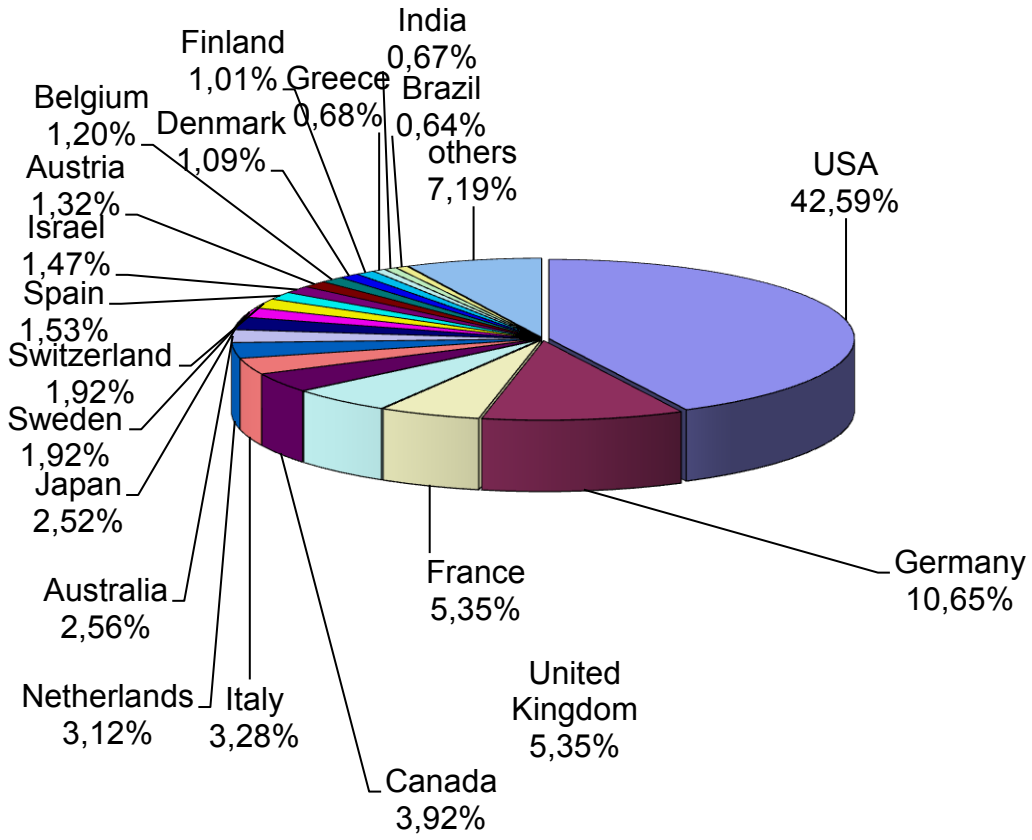


Fig. 2 Shares of countries to which publications are assigned in CiteSeer

The relative shares in Table 2 are smaller than those in Figure 2 because the base (100%) is much larger – 717 thousand, which is the original number of CiteSeer documents. These relative shares are important for they help us compare CiteSeer publication shares with those from the Web of Science and Scopus where the number of all documents can be determined, but the number of publication-country assignments is unknown. The absolute numbers in Table 2 are the numbers of publications assigned to a country and they were input in Figure 2. If, hypothetically, each CiteSeer article was assigned to exactly one country, the sum of counts in

Table 2 Top 100 countries by publications in CiteSeer

Rank	Country	Public.	Share	Rank	Country	Public.	Share
1	USA	191 363	26.70%	51	Belarus	119	0.02%
2	Germany	47 866	6.68%	52	Venezuela	114	0.02%
3	France	24 052	3.36%	53	Egypt	107	0.01%
4	United Kingdom	24 042	3.35%	54	Latvia	102	0.01%
5	Canada	17 630	2.46%	55	Uruguay	96	0.01%
6	Italy	14 718	2.05%	56	Serbia and Mont.	94	0.01%
7	Netherlands	14 022	1.96%	57	Lithuania	93	0.01%
8	Australia	11 496	1.60%	58	Lebanon	66	0.01%
9	Japan	11 328	1.58%	59	Tunisia	66	0.01%
10	Sweden	8 639	1.21%	60	Colombia	60	0.01%
11	Switzerland	8 611	1.20%	61	Malta	60	0.01%
12	Spain	6 876	0.96%	62	Armenia	55	0.01%
13	Israel	6 616	0.92%	63	Iceland	55	0.01%
14	Austria	5 934	0.83%	64	Panama	53	0.01%
15	Belgium	5 411	0.75%	65	Vietnam	44	0.01%
16	Denmark	4 882	0.68%	66	Cuba	42	0.01%
17	Finland	4 533	0.63%	67	Morocco	39	0.01%
18	Greece	3 038	0.42%	68	Macau	37	0.01%
19	India	3 002	0.42%	69	Pakistan	36	0.01%
20	Brazil	2 889	0.40%	70	Indonesia	34	0.00%
21	Portugal	2 650	0.37%	71	Saudi Arabia	34	0.00%
22	Russia	2 351	0.33%	72	Puerto Rico	32	0.00%
23	Hong Kong	2 238	0.31%	73	Philippines	31	0.00%
24	Norway	2 215	0.31%	74	Kuwait	30	0.00%
25	Singapore	1 897	0.26%	75	Algeria	25	0.00%
26	Taiwan	1 808	0.25%	76	Bangladesh	24	0.00%
27	New Zealand	1 703	0.24%	77	Costa Rica	23	0.00%
28	China	1 600	0.22%	78	Jordan	21	0.00%
29	Poland	1 564	0.22%	79	Kenya	14	0.00%
30	Czech Republic	1 453	0.20%	80	Liechtenstein	14	0.00%
31	South Korea	1 450	0.20%	81	Macedonia	14	0.00%
32	Hungary	1 423	0.20%	82	Nigeria	14	0.00%
33	Ireland	1 366	0.19%	83	Moldova	13	0.00%
34	Mexico	1 071	0.15%	84	Oman	11	0.00%
35	Turkey	775	0.11%	85	Cameroon	9	0.00%
36	Slovenia	659	0.09%	86	Jamaica	9	0.00%
37	Chile	489	0.07%	87	Martinique	9	0.00%
38	South Africa	472	0.07%	88	Netherlands Antilles	9	0.00%
39	Romania	450	0.06%	89	Sri Lanka	9	0.00%
40	Argentina	445	0.06%	90	Reunion	8	0.00%
41	Thailand	335	0.05%	91	United Arab Emirates	8	0.00%
42	Ukraine	306	0.04%	92	Uzbekistan	8	0.00%
43	Bulgaria	299	0.04%	93	Ethiopia	7	0.00%
44	Cyprus	285	0.04%	94	Vatican	7	0.00%
45	Slovakia	250	0.03%	95	Bahrain	6	0.00%
46	Luxembourg	242	0.03%	96	Fiji	6	0.00%
47	Iran	215	0.03%	97	Guinea	6	0.00%
48	Croatia	149	0.02%	98	Mozambique	6	0.00%
49	Estonia	141	0.02%	99	Nicaragua	6	0.00%
50	Malaysia	131	0.02%	100	Uganda	6	0.00%

Table 2 would be approximately 717 thousand and the total share 100% (the rest after rank 100 is negligible). If each document was assigned to two or more countries (i.e. all papers are internationally co-authored), the sum of counts would be more than 717 thousand and the total share more than 100 %. A further discussion of the results in Table 2 will follow in the next section along with a comparison to the Web of Science and Scopus.

5.3 Comparison with the Web of Science and Scopus

To get a clue how reliable CiteSeer data are and to see how distant or close to other well-known bibliographic data sources they are, it was necessary to perform a couple of comparisons and measurements. Based on the amount of available information on publication shares of countries from the previous section, we decided to compare these country shares to those

Table 3 Top 30 computer science countries by Web of Science in 1996 – 2005

Rank	Cite-Seer	Country	Publications	Share	Citations	Average citations	h-index
1	1	*USA*	52 579	35.33%	904 339	17.20	258
2	4	*United Kingdom*	11 515	7.74%	160 691	13.95	125
3	9	*Japan*	8 902	5.98%	72 379	8.13	82
4	2	*Germany*	8 554	5.75%	114 075	13.34	108
5		China	8 348	5.61%	92 050	11.03	86
6	5	*Canada*	7 630	5.13%	102 609	13.45	105
7	3	*France*	7 159	4.81%	97 801	13.66	102
8		Taiwan	6 690	4.49%	66 762	9.98	76
9	6	*Italy*	6 587	4.43%	76 837	11.66	87
10		South Korea	4 753	3.19%	42 720	8.99	65
11	12	*Spain*	4 421	2.97%	50 272	11.37	76
12	8	*Australia*	4 196	2.82%	54 625	13.02	82
13	7	*Netherlands*	3 503	2.35%	55 459	15.83	88
14	19	*India*	3 103	2.08%	27 613	8.90	55
15	13	*Israel*	3 014	2.03%	46 385	15.39	82
16		Singapore	2 695	1.81%	32 015	11.88	66
17		Russia	2 246	1.51%	7 879	3.51	33
18	18	*Greece*	2 153	1.45%	20 283	9.42	50
19	15	*Belgium*	1 849	1.24%	29 343	15.87	65
20	11	*Switzerland*	1 838	1.23%	37 542	20.43	78
21	10	*Sweden*	1 766	1.19%	23 825	13.49	57
22	20	*Brazil*	1 449	0.97%	14 601	10.08	46
23		Poland	1 440	0.97%	15 948	11.08	50
24	17	*Finland*	1 408	0.95%	23 137	16.43	59
25	14	*Austria*	1 357	0.91%	17 065	12.58	51
26		Turkey	1 284	0.86%	13 160	10.25	44
27	16	*Denmark*	1 045	0.70%	16 645	15.93	53
28		Hong Kong	858	0.58%	10 909	12.71	47
29		Ireland	806	0.54%	8 202	10.18	38
30		Hungary	791	0.53%	8 072	10.20	41

obtained from the Web of Science and Scopus – two established manually maintained bibliographic databases. The goal was to create rankings of countries by the number of "their" publications in the field of computer science and to compare them to the CiteSeer ranking in Table 2.

Table 4 Top 30 computer science countries by Scopus in 1996 - 2005

Rank	Cite-Seer	Country	Publications	Share	Citations	Average citations	h-index
1	1	*USA*	87 591	26.90%	1 731 096	19.76	360
2		China	26 004	7.99%	149 019	5.73	104
3	4	*United Kingdom*	21 545	6.62%	292 929	13.60	163
4	9	*Japan*	21 231	6.52%	141 346	6.66	106
5	2	*Germany*	18 125	5.57%	213 144	11.76	143
6	3	*France*	14 570	4.47%	187 746	12.89	136
7	5	*Canada*	13 001	3.99%	191 347	14.72	135
8	6	*Italy*	12 133	3.73%	147 608	12.17	117
9		South Korea	10 370	3.18%	84 225	8.12	91
10		Taiwan	10 238	3.14%	106 810	10.43	95
11	12	*Spain*	8 035	2.47%	87 291	10.86	94
12	8	*Australia*	7 105	2.18%	96 481	13.58	103
13	19	*India*	5 997	1.84%	58 432	9.74	80
14	7	*Netherlands*	5 966	1.83%	93 431	15.66	110
15		Hong Kong	5 382	1.65%	78 625	14.61	94
16		Russia	5 177	1.59%	16 783	3.24	45
17	13	*Israel*	4 767	1.46%	81 874	17.18	108
18		Singapore	4 230	1.30%	51 347	12.14	79
19	18	*Greece*	3 932	1.21%	38 669	9.83	66
20	10	*Sweden*	3 916	1.20%	69 242	17.68	85
21	11	*Switzerland*	3 618	1.11%	75 824	20.96	111
22	15	*Belgium*	3 479	1.07%	55 409	15.93	86
23		Poland	3 165	0.97%	25 992	8.21	57
24	17	*Finland*	2 867	0.88%	37 645	13.13	73
25	20	*Brazil*	2 860	0.88%	24 543	8.58	55
26		Turkey	2 496	0.77%	23 679	9.49	57
27	14	*Austria*	2 371	0.73%	27 242	11.49	66
28	16	*Denmark*	1 818	0.56%	26 444	14.55	64
29		Portugal	1 527	0.47%	15 513	10.16	50
30		Hungary	1 500	0.46%	16 459	10.97	50

In addition to article counts, we also found out numbers of citations to the articles, average citations per article, and h-indices as defined by Hirsch (2005) for individual countries. In both Table 3 and Table 4, countries are ordered descendingly by the number of publications and the countries from the top 20 CiteSeer countries (see Table 2) are marked with their CiteSeer rank in the second column. When looking at the rankings, we may immediately note that three East Asian countries (mainland China, South Korea, and Taiwan) are underrepresented in CiteSeer. Both WoS and Scopus place them in the Top 10 whereas in CiteSeer

they are at ranks around 30. The corresponding top-level Internet domains *.cn*, *.kr*, and *.tw* in Table 1 are also relatively lowly ranked, which might suggest that CiteSeer did not crawl these Web regions so extensively as it should have regarding their real scientific productivity in computer science. Otherwise, we cannot see any striking discrepancies between CiteSeer on one side and WoS and Scopus on the other.

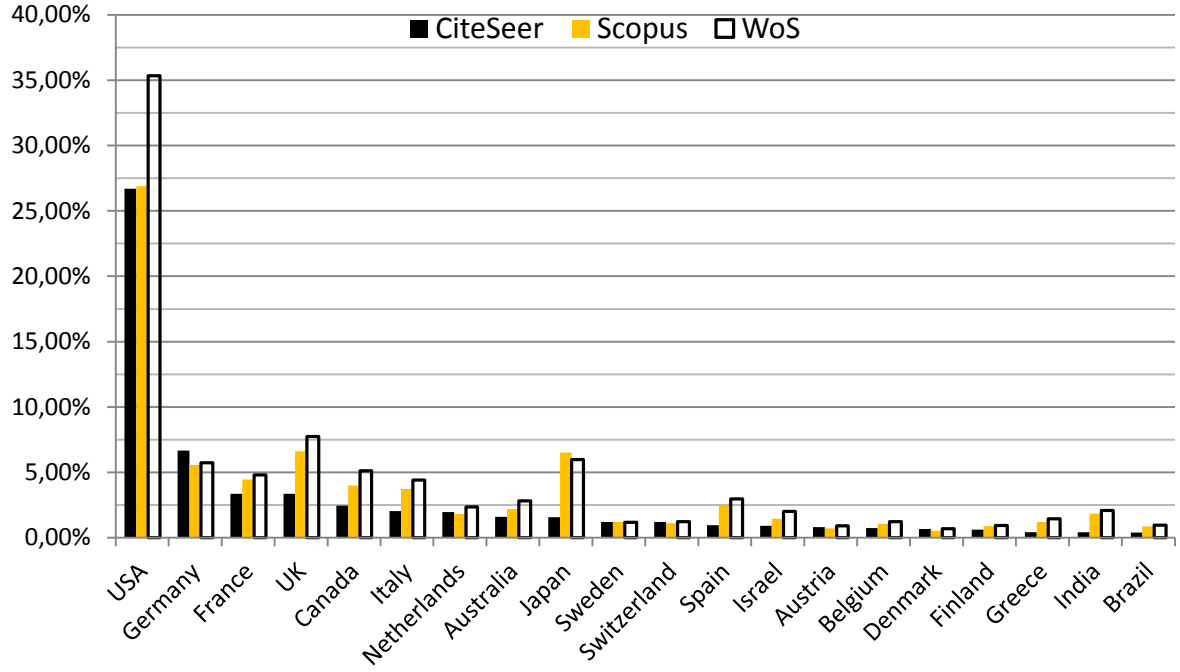


Fig. 3 Publication shares of top 20 CiteSeer countries in Scopus and WoS

Publication shares of the top 20 CiteSeer countries in CiteSeer, WoS, and Scopus are shown in Figure 3. There are no evident outliers or differences either, except perhaps for a greater USA share in WoS. In Figure 4, we show Spearman's rank correlation coefficients between the rankings of CiteSeer and Scopus, CiteSeer and WoS, and Scopus and WoS for the top 10, 20, 30, 40, and 50 CiteSeer countries. All the coefficients are significant at the 0.01 level (two-tailed) except those around 0.65 in the top ten, which are significant at the 0.05 level. Not surprisingly, the rankings from Scopus and WoS are always very highly positively correlated (0.96 – 0.99). But as for CiteSeer, it is also positively correlated with the highest correlation being about 0.86 in the top 50. We may conclude that the ranking by publications from CiteSeer (Table 2) is relevant and quite competitive compared to the rankings from both WoS and Scopus. As there is no simple way of obtaining the total count of citations to all computer science publications published from 1996 to 2005 from the Web sites of WoS and Scopus, which would be necessary to determine the relative citation shares in Tables 3 and 4, we do not present a comparison plot similar to Figure 3 for citations. But we do show, in analogy to

Figure 5, how citation-based rankings correlate with each other in Figure 5. As we can see, the rankings of countries based on citations from CiteSeer correlate quite positively (0.79 – 0.90) with those from Scopus and WoS. All the coefficients in Figure 5 are significant at the 0.01 level (two-tailed).

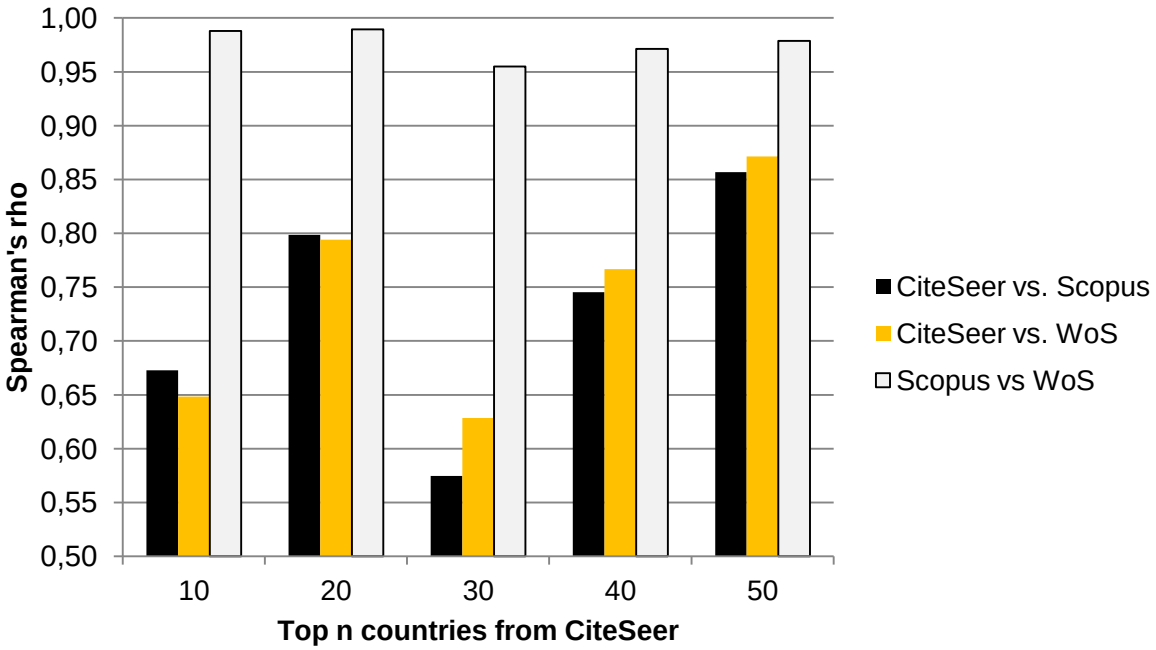


Fig. 4 Correlations of country publication rankings of CiteSeer, Scopus, and WoS

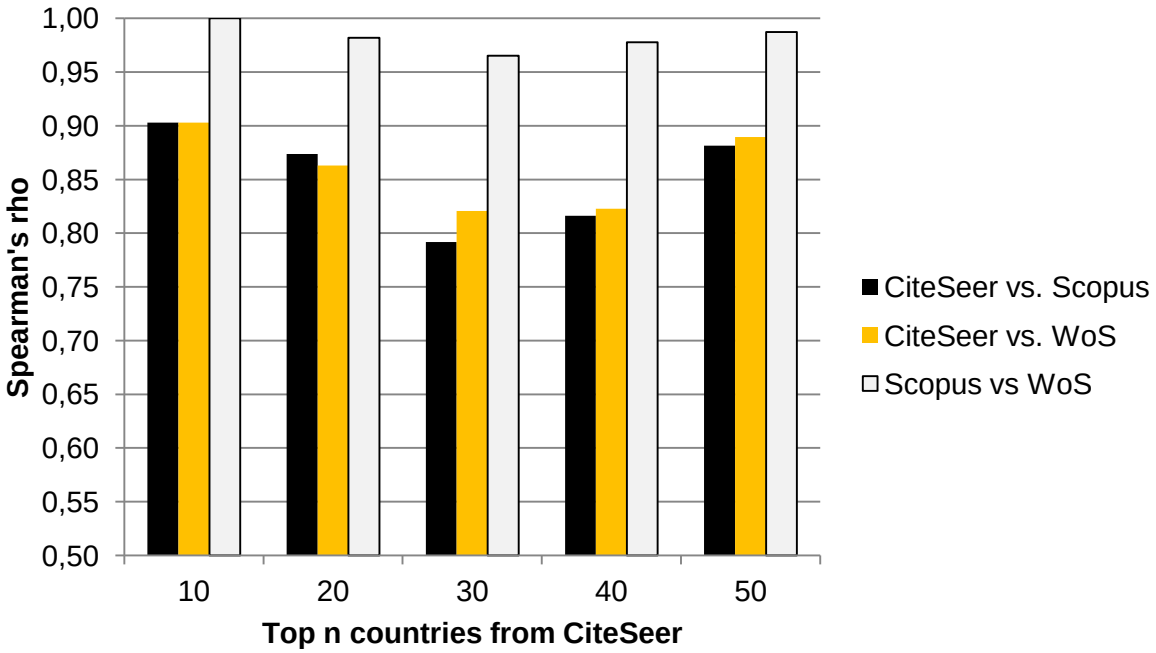


Fig. 5 Correlations of country citation rankings of CiteSeer, Scopus, and WoS

5.4 Citations and recursive indicators

Finally, the resulting directed graph of citations between countries had 243 nodes (countries) and 2 472 edges (citations between them). There were no parallel edges in the graph. Instead, a weight was assigned to each edge denoting from how many parallel edges the edge was created. The sum of weights in the whole graph was about 1.5 million.

In Table 5, we can see the top 80 countries ordered descendingly by their in-degree in the country citation graph. In the first case ("In-degree") the edge weights are all set to one, in the second case ("Citations") they are left as they are. Both rankings place USA, Germany, and the United Kingdom at the top with approximately 48%, 8%, and 6% of all citations, respectively. The rank four in In-degree is tied by Canada and France with the same number of citing countries (74) but, in total, France is cited more often by foreign countries and is positioned ahead of Canada in Citations. A similar behaviour may be observed with several other countries. The country rankings in Table 6 were obtained by applying recursive techniques, but despite their much higher computational costs they do not seem to provide any striking new information, though. We found the five rankings in Tables 5 and 6 to be very highly positively correlated with each other with Spearman's ρ between 0.97 and 1 (all significant at the 0.01 level two-tailed).

6. Conclusions and future work

We have presented a thorough study of CiteSeer data with focus on countries and territories with which authors of publications indexed by CiteSeer are affiliated. The main contributions of the study are the following:

- We show from which parts of the Web CiteSeer and CiteSeer^X gathered its documents in terms of shares of top-level Internet domains in article sources.
- We analyze country shares in CiteSeer publications. (Unfortunately, CiteSeer^X does not have the information needed for this kind of analysis.)
- We compare the CiteSeer ranking to country shares of computer science publications from the Web of Science and Scopus to test the reliability of the productivity ranking.
- We submit CiteSeer data to a citation analysis and determine the most influential countries in terms of in-degree, citations, HITS, PageRank, and weighted PageRank.

Table 5 Top 80 countries by in-degree and citations in CiteSeer

In-degree						Citations					
R.	Country	In	R.	Country	In	R.	Country	Cites	R.	Country	Cites
1	USA	98	41	Slovakia	26	1	USA	728 289	41	Romania	641
2	Germany	82	42	Chile	24	2	Germany	122 389	42	Chile	590
3	United Kingdom	75	43	Jordan	22	3	United Kingdom	89 933	43	Jordan	425
4	Canada	74	44	Argentina	21	4	France	82 632	44	Slovakia	416
5	France	74	45	Bahrain	21	5	Canada	76 148	45	Thailand	409
6	Australia	66	46	South Africa	21	6	Italy	52 570	46	South Africa	328
7	Netherlands	66	47	Bulgaria	20	7	Netherlands	42 252	47	Venezuela	321
8	Switzerland	66	48	Croatia	18	8	Israel	33 701	48	Bahrain	246
9	Italy	64	49	Estonia	18	9	Switzerland	33 185	49	Croatia	222
10	Israel	63	50	Venezuela	18	10	Japan	32 433	50	Estonia	190
11	Japan	62	51	Uruguay	15	11	Australia	27 484	51	Ukraine	183
12	Sweden	62	52	Egypt	14	12	Belgium	21 356	52	Bulgaria	179
13	Spain	58	53	Lebanon	14	13	Sweden	21 211	53	Uruguay	179
14	Austria	55	54	Serbia & Mt.	14	14	Austria	13 975	54	Panama	165
15	Denmark	55	55	Lithuania	13	15	Finland	13 953	55	Lebanon	147
16	Finland	54	56	Latvia	12	16	Spain	13 543	56	Iceland	141
17	Singapore	53	57	Malta	12	17	Denmark	12 744	57	Egypt	138
18	Belgium	52	58	Panama	11	18	India	10 882	58	Iran	119
19	Greece	50	59	Belarus	10	19	Greece	7 304	59	Lithuania	108
20	India	48	60	Fiji	10	20	Singapore	6 165	60	Latvia	103
21	Hong Kong	45	61	Iceland	10	21	Mexico	5 618	61	Fiji	97
22	Portugal	45	62	Bangladesh	9	22	Hong Kong	5 419	62	Serbia & Mt.	81
23	Russia	45	63	Iran	9	23	Portugal	5 398	63	Macau	62
24	Brazil	43	64	Pakistan	8	24	Brazil	5 056	64	Belarus	55
25	Taiwan	42	65	Ukraine	8	25	Taiwan	3 828	65	Pakistan	54
26	China	40	66	Saudi Arabia	7	26	South Korea	3 413	66	Saudi Arabia	50
27	New Zealand	40	67	Moldova	6	27	Russia	3 218	67	Liechtenstein	42
28	Poland	40	68	Macau	5	28	Norway	3 008	68	Kuwait	40
29	Ireland	39	69	Morocco	5	29	New Zealand	2 978	69	Moldova	35
30	Hungary	38	70	Costa Rica	4	30	Ireland	2 952	70	Bangladesh	23
31	Mexico	37	71	Kuwait	4	31	Hungary	2 816	71	Reunion	21
32	Norway	37	72	Vietnam	4	32	China	2 385	72	Vietnam	21
33	Czech Republic	36	73	Armenia	3	33	Poland	1 696	73	Costa Rica	18
34	Cyprus	35	74	Colombia	3	34	Slovenia	1 389	74	Armenia	16
35	South Korea	34	75	Indonesia	3	35	Cyprus	1 162	75	Indonesia	15
36	Turkey	34	76	Tunisia	3	36	Turkey	1 089	76	Monaco	14
37	Slovenia	33	77	Antarctica	2	37	Luxembourg	920	77	Morocco	13
38	Luxembourg	29	78	Congo	2	38	Czech Republic	837	78	Tunisia	12
39	Thailand	27	79	Ethiopia	2	39	Argentina	721	79	Antarctica	10
40	Romania	26	80	Jamaica	2	40	Malta	649	80	Colombia	9

Table 6 Top 80 countries by HITS, PageRank and weighted PageRank in CiteSeer

HITS		PageRank		Weighted PageRank	
R. Country	R. Country	R. Country	R. Country	R. Country	R. Country
1 USA	41 Slovakia	1 USA	41 Thailand	1 USA	41 Romania
2 Germany	42 Chile	2 Canada	42 Chile	2 Germany	42 Chile
3 UK	43 Jordan	3 Germany	43 Jordan	3 UK	43 Malta
4 Canada	44 Argentina	4 UK	44 South Africa	4 France	44 Thailand
5 France	45 Bahrain	5 France	45 Argentina	5 Canada	45 Jordan
6 Netherlands	46 South Africa	6 Israel	46 Bahrain	6 Italy	46 Venezuela
7 Italy	47 Bulgaria	7 Italy	47 Bulgaria	7 Netherlands	47 South Africa
8 Australia	48 Croatia	8 Switzerland	48 Venezuela	8 Israel	48 Bahrain
9 Switzerland	49 Venezuela	9 Netherlands	49 Croatia	9 Japan	49 Croatia
10 Japan	50 Estonia	10 Australia	50 Estonia	10 Switzerland	50 Estonia
11 Sweden	51 Uruguay	11 Japan	51 Uruguay	11 Australia	51 Bulgaria
12 Israel	52 Egypt	12 Sweden	52 Egypt	12 Sweden	52 Panama
13 Spain	53 Lebanon	13 Austria	53 Lebanon	13 Belgium	53 Iceland
14 Finland	54 Serbia & Mt.	14 Spain	54 Serbia & Mt.	14 Austria	54 Lebanon
15 Denmark	55 Lithuania	15 Denmark	55 Lithuania	15 Finland	55 Uruguay
16 Austria	56 Latvia	16 Belgium	56 Latvia	16 Spain	56 Egypt
17 Belgium	57 Malta	17 Finland	57 Malta	17 Denmark	57 Ukraine
18 Singapore	58 Panama	18 Greece	58 Panama	18 India	58 Lithuania
19 Greece	59 Iceland	19 India	59 Iceland	19 Greece	59 Iran
20 India	60 Belarus	20 Singapore	60 Belarus	20 Mexico	60 Fiji
21 Hong Kong	61 Fiji	21 Russia	61 Fiji	21 Singapore	61 Latvia
22 Russia	62 Iran	22 Portugal	62 Iran	22 Hong Kong	62 Serbia & Mt.
23 Portugal	63 Bangladesh	23 Hong Kong	63 Bangladesh	23 Brazil	63 Liechtenstein
24 Taiwan	64 Ukraine	24 Brazil	64 Pakistan	24 Portugal	64 Pakistan
25 Brazil	65 Pakistan	25 New Zealand	65 Ukraine	25 Taiwan	65 Saudi Arabia
26 Ireland	66 Saudi Arabia	26 Taiwan	66 Saudi Arabia	26 New Zealand	66 Belarus
27 Poland	67 Moldova	27 Poland	67 Moldova	27 South Korea	67 Macau
28 China	68 Morocco	28 Ireland	68 Morocco	28 Russia	68 Vietnam
29 New Zealand	69 Macau	29 China	69 Kuwait	29 Norway	69 Kuwait
30 Norway	70 Costa Rica	30 Norway	70 Macau	30 Hungary	70 Moldova
31 Hungary	71 Kuwait	31 Hungary	71 Costa Rica	31 Ireland	71 Monaco
32 Mexico	72 Vietnam	32 Mexico	72 Vietnam	32 China	72 Reunion
33 South Korea	73 Armenia	33 Czech Rep.	73 Armenia	33 Poland	73 Costa Rica
34 Czech Rep.	74 Indonesia	34 Cyprus	74 Indonesia	34 Slovenia	74 Indonesia
35 Turkey	75 Tunisia	35 South Korea	75 Tunisia	35 Cyprus	75 Tunisia
36 Cyprus	76 Colombia	36 Turkey	76 Colombia	36 Turkey	76 Morocco
37 Slovenia	77 Reunion	37 Slovenia	77 Ethiopia	37 Slovakia	77 Armenia
38 Luxembourg	78 Liechtenstein	38 Luxembourg	78 Liechtenstein	38 Czech Rep.	78 Vatican
39 Thailand	79 Neth. Antilles	39 Romania	79 Reunion	39 Luxembourg	79 Bangladesh
40 Romania	80 Ethiopia	40 Slovakia	80 Puerto Rico	40 Argentina	80 Colombia

Based on our analysis, we have obtained the following key results:

- Both CiteSeers collected computer science papers mainly from North American domains, followed by the domains of developed European and Asian countries. The top domains are *.com*, *.de*, *.edu*, *.fr*, *.org*, and *.uk*.
- United States is by far the greatest producer of computer science research papers although West European countries are, relatively at least, very competitive. Germany, France, and the United Kingdom can be named as a few examples.
- CiteSeer rankings of countries by publications and citations are very similar to those generated by the Web of Science or Scopus with a notable difference that CiteSeer apparently underestimates the potential of mainland China, South Korea, and Taiwan.
- Recursive techniques such as PageRank do not provide much new information on the influence of countries compared to simple citation counts. More or less, they confirm that popularity and prestige are close terms in the rankings of countries.

The study presented in this paper is the first of its kind that seeks to determine the most influential countries in computer science by analyzing the free CiteSeer digital library data. It complements the paper by Fiala (2011), which is concerned with individual authors in CiteSeer. From the papers listed in the literature review, the research conducted by Wainer et al. (2009) is closest to ours in that it evaluates the scientific output in computer science of several (thirteen) countries. However, it just examines publications from the Web of Science and Scopus from 2001 to 2005 and is not at all concerned with citations. Even less countries (six) are explored by Guan & Ma (2004) for the period of 1993 - 2002. Both studies, in accordance with our results, document a clear superiority of the USA over the rest of the world in computer science research. Unfortunately, there seems to be no previous complex computer science study for countries with which we could compare our findings.

Although CiteSeer data are far from complete and precise (in our experience, some 10% of the existing information might be erroneous), we may conclude that CiteSeer is a free digital library of valuable data and may be successfully used in bibliometric studies, possibly along with other well-known bibliographic databases, as we have shown in this paper. Let us underline in this place that the results we present depend solely on the content and quality of CiteSeer data. If other regions of the Web had been crawled, if Asian paper repositories had been preferred by authors (see Section 5.1), or if the information extraction from papers done by CiteSeer had been more precise and complete, the outcomes of our analysis could have been different. Let us hope in this respect that CiteSeer^{X} will acquire data in a more standard-

ized and transparent way and that it will enrich its metadata with the information on addresses and affiliations as well. Our future work on CiteSeer will concentrate on the citation analysis of institutions and on other reliability measures of CiteSeer data as well as on exploring further differences between the data in CiteSeer and CiteSeer^{X}.

Acknowledgements

This work was supported by the European Regional Development Fund (ERDF), project "NTIS - New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090. Many thanks are due to the anonymous reviewers for their useful comments.

References

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.

An, Y., Janssen, J., & Milios, E. E. (2004). Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, 6(6), 664-678.

Bakri, A., & Willett, P. (2011). Computer science research in Malaysia: A bibliometric analysis. *Aslib Proceedings: New Information Perspectives*, 63(2-3), 321-335.

Bar-Ilan, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing and Management*, 42(6), 1553-1566.

Bar-Ilan, J. (2010). Web of Science with the Conference Proceedings Citation Indexes: The case of computer science. *Scientometrics*, 83(3), 809-824.

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th World Wide Web Conference*, Brisbane, Australia, 107-117.

Chen, C. (2000). Domain visualization for digital libraries. In *Proceedings of the International Conference on Information Visualization (IV2000)*, London, UK, 261-267.

CiteSeer. http://citeseer.ist.psu.edu.

CiteSeer^{X}. http://citeseerx.ist.psu.edu.

Feitelson, D. G., & Yovel, U. (2004). Predictive ranking of computer scientists using CiteSeer data. *Journal of documentation*, 60(1), 44-61.

Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks. *Scientometrics*, 76(1), 135-158.

Fiala, D. (2011). Mining citation information from CiteSeer data. *Scientometrics*, 86(3), 553-562.

Franceschet, M. (2010). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, 83(1), 243-258.

Franceschet, M. (2010b). The role of conference publications in CS. *Communications of the ACM*, 53(12), 129-132.

Giles, C. L., & Councill, I. G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51), 17599-17604.

Guan, J., Ma, N. (2004). A comparative study of research performance in computer science. *Scientometrics*, 61(3), 339-359.

Gupta, B. M., Kshitij, A., & Verma, C. (2011). Mapping of Indian computer science research output, 1999-2008. *Scientometrics*, 86(2), 261-283.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572.

Hopcroft, J., Khan, O., Kulis, B., & Selman, B. (2004). Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(suppl. 1), 5249-5253.

Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5), 604-632.

Ma, R., Ni, C., & Qiu J. (2008). Scientific research competitiveness of world universities in computer science. *Scientometrics*, 76(2), 245–260.

Repository66. http://maps.repository66.org.

Scopus. http://www.scopus.com.

Wainer, J., Xavier, E. C., & Bezerra, F. (2009). Scientific production in computer science: A comparative study of Brazil and other countries. *Scientometrics*, 81(2), 535-547.

Web of Science. http://apps.isiknowledge.com.

Wu, C.-L., & Koh, J.-L. (2010). Hierarchical topic-based communities construction for authors in a literature database. *Lecture Notes in Computer Science*, 6097, 514-524.

Zhao, D., & Logan, E. (2002). Citation analysis using scientific publications on the Web as data source: A case study in the XML research area. *Scientometrics*, 54(3), 449-472.

Zhao, D., & Strotmann, A. (2007). Can citation analysis of web publications better detect research fronts? *Journal of the American Society for Information Science and Technology*, 58 (9), 1285-1302.

Zhou, D., Councill, I., Zha, H., & Giles, C. L. (2007). Discovering temporal communities from social network documents. In *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM'07)*, Omaha, Nebraska, USA, 745-750.

Article 3

The next article is also concerned with CiteSeer[9]. This time, however, the study is focused on individual authors rather than on countries. The aim of the analysis was to create citation and collaboration graphs of authors from CiteSeer data and to identify the best performers in terms of various scientometric indicators including citations, HITS, and PageRank and its variants adopted from Fiala et al. (2008), To compare the 12 resulting author rankings, I also employed a previously used methodology to test the ranks of authors who won the ACM E. F. Codd Innovations Award[10]. Simply said, the best ranking is the one that places best the award winners. The main conclusions are that large enough citation and collaboration graphs of authors can be generated from CiteSeer data that enable bibliometric analyses and that simple citation counts or in-degree rank award winners better than PageRank and its modifications, from which it is impossible to unambiguously choose the best one. This analysis is the first large-scale bibliometric study of author citations based on CiteSeer data.

CiteSeer data are much larger than DBLP data analyzed by Fiala et al. (2008). There are more than 1.8 million citations between 717 thousand publications. Some publications (about 333 thousand) are entirely isolated – neither do they cite, nor are they cited by other publications. On the other hand, roughly 149 thousand publications cite and are cited at the same time. Of course, there are publications that cite but are not cited and vice versa. These and other relations can be seen in Figure 1. From this publication citation graph the resulting directed graph of citations between authors (author citation graph) was constructed that had then some 411 thousand vertices (authors) and 4.8 million weighted edges (citations). As with publications, some authors (171 thousand) are isolated from the rest while other authors cite or are cited by others (111 thousand in Figure 2). The relation of those who cite and are not cited to those who are cited but do not cite is approximately 3 : 1.

[9] See Article 2 for more information on CiteSeer.

[10] http://www.sigmod.org/sigmod-awards/sigmod-awards#innovations

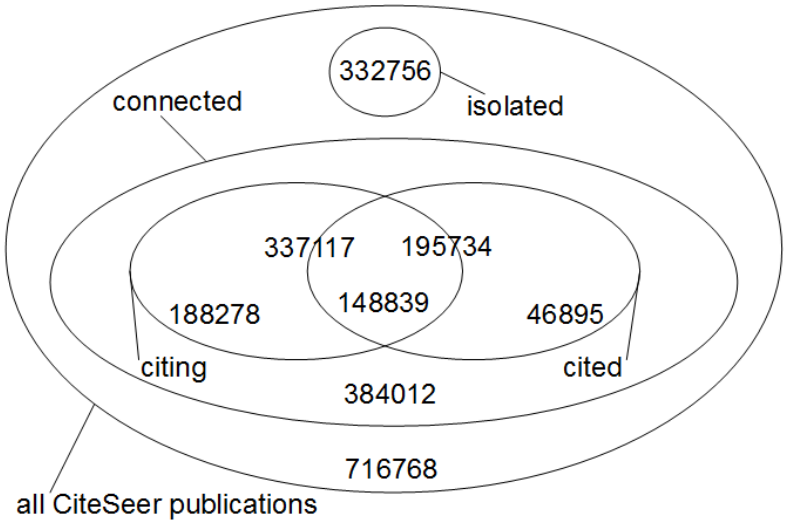


Figure 1: Numbers of citing and cited CiteSeer publications

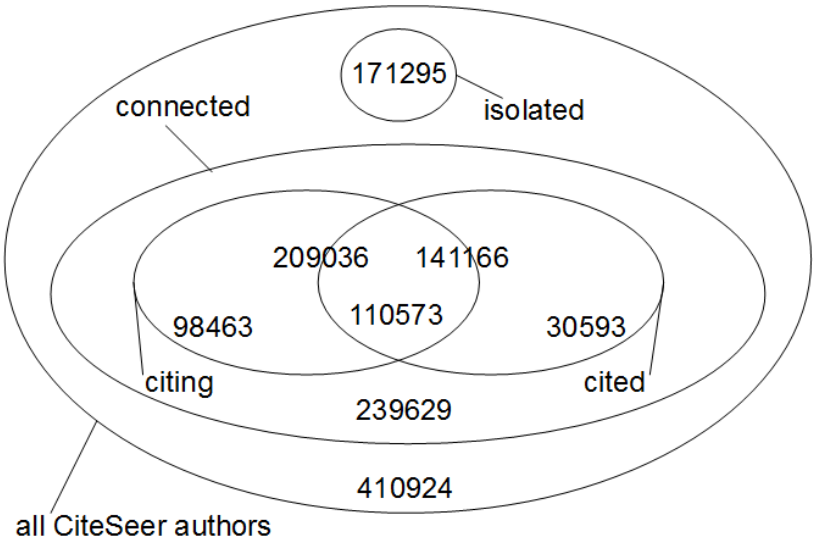


Figure 2: Numbers of citing and cited CiteSeer authors

Neither Figure 1 nor Figure 2 did appear in the final version of the article. In addition, there is an online supplement to this article (available on the journal's Web site only), which appears in print as Tables 4 – 7 attached to the following article's main text.

Preprint of: Fiala, D. (2011). Mining citation information from CiteSeer data. *Scientometrics*, 86(3), 553-562.

Mining citation information from CiteSeer data

Dalibor Fiala

University of West Bohemia, Univerzitní 8, 30614 Plzeň, Czech Republic

Phone: 00420 377 63 24 29, fax: 00420 377 63 24 01, email: dalfia@kiv.zcu.cz

Abstract: The CiteSeer digital library is a useful source of bibliographic information. It allows for retrieving citations, co-authorships, addresses, and affiliations of authors and publications. In spite of this, it has been relatively rarely used for automated citation analyses. This article describes our findings after extensively mining from the CiteSeer data. We explored citations between authors and determined rankings of influential scientists using various evaluation methods including citation and in-degree counts, HITS, PageRank, and its variations based on both the citation and collaboration graphs. We compare the resulting rankings with lists of computer science award winners and find out that award recipients are almost always ranked high. We conclude that CiteSeer is a valuable, yet not fully appreciated, repository of citation data and is appropriate for testing novel bibliometric methods.

***Keywords**: CiteSeer, citation analysis, rankings, evaluation.*

Introduction

Data from CiteSeer have been surprisingly little explored in the scientometric literature. One of the reasons for this may have been fears that the data gathered in an automated way from the Web are inaccurate – incomplete, erroneous, ambiguous, redundant, or simply wrong. Also, the uncontrolled and decentralized nature of the Web is said to simplify manipulating and biasing Web-based publication and citation metrics. However, there have been a few attempts at processing the CiteSeer data which we will briefly mention.

Zhou et al. (2007) have investigated documents from CiteSeer to discover temporal social network communities in the domains of databases and machine learning. On the other hand, Hopcroft et al. (2004) track evolving communities in the whole CiteSeer paper citation graph. An et al. (2004) have constructed article citation graphs in several research domains by querying CiteSeer and have explored them in terms of components. Popescul et al. (2003) have classified CiteSeer articles into categories by venues. Šingliar and Hausknecht (2006) cluster CiteSeer papers by topics based on their references to authors. Author co-citation analysis of CiteSeer documents in the XML research field has been conducted by Zhao and Strotmann (2007) and Zhao and Logan (2002) and in computer graphics by Chen (2000). Bar-Ilan (2006) has used CiteSeer for a citation analysis of the works of a famous mathematician. A kind of citation analysis, but this time for acknowledgements, has also been performed by

Giles and Councill (2004). Chakrabarti and Agarwal (2006) use CiteSeer data in their experiments with learning ranking functions for real-world entity-relation graphs. Feitelson and Yovel (2004) have examined citation ranking lists obtained from CiteSeer and predicted future rankings of authors.

Most of the research activities mentioned above have been concerned with just a small part of the CiteSeer database, limited to a specific scientific field or even venue (conference or journal). Very few have dealt with the CiteSeer citation graph as a whole as we do in this study whose research questions are the following: What is the nature of CiteSeer data? Can sufficiently large citation and co-authorship graphs for publications and authors be constructed out of them? If yes, can we, based on those graphs, generate realistic rankings of salient researchers? In the rest of this paper, we will first describe the methods we work with, present the basic features of CiteSeer and its data and then show that we can answer yes to the last two questions.

Methods

In our previous work (Fiala et al. 2008 and Ježek et al. 2008), we have built on top of the well-known PageRank concept by Brin and Page (1998) and have modified this ranking function originally devised for the Web graph so as to evaluate author significance based on the citation as well as collaboration networks. The key concept is that a citation from a colleague is less valuable than that from a foreign researcher. Thus, cited authors should be penalized for the frequency of collaboration (co-authorship) with authors citing them. To add more information to the citation graph, we defined several parameters to weight its edges more discriminatively than purely by citation counts. These parameters, calculated from the collaboration graph, are the following:

a) $c_{u,v}$ is the number of common publications by authors *u* and *v* (i.e. the number of their collaborations, code-named COLLABORATION),
b) $f_{u,v}$ is the number of publications by author *u* plus the number of publications by author *v* (i.e. the total number of publications by those two authors, code-named ALL_PUBLICATIONS),
c) $h_{u,v}$ is the number of all co-authors (including duplicates) in all publications by author *u* plus the number of all co-authors (including duplicates) in all publications by author *v*, code-named ALL_COAUTHORS,

d) $hd_{u,v}$ is the number of all distinct co-authors in all publications by author *u* plus the number of all distinct co-authors in all publications by author *v*, code-named ALL_DIST_COAUTHORS,

e) $g_{u,v}$ is the number of publications by author *u* where *u* is not the only author plus the number of publications by author *v* where *v* is not the only author (i.e. the total number of collaborations by those two authors, code-named ALL_COLLABORATIONS),

f) $t_{u,v}$ is the number of co-authors (including duplicates) in common publications by authors *u* and *v*, code-named COAUTHORS,

g) $td_{u,v}$ is the number of distinct co-authors in common publications by authors *u* and *v*, code-named DIST_COAUTHORS.

Note that we make no distinction between authoring and co-authoring a publication. In either case, an author has published the publication. Also, for the sake of simplicity of parameters *h*, *hd*, *t*, and *td*, authors are considered as co-authors of themselves. For a much more detailed theoretical background as well as a practical example, we refer the reader to the article by Fiala et al. (2008).

Data

CiteSeer[11] gathers information mainly about computer science publications by crawling the World Wide Web, downloading, and automatically analyzing potential scientific publications (mostly PDF or PS files) and provides access to it via a Web interface and downloadable XML-like files that can be further processed by machines. The information in these XML files typically includes publication title, authors, their affiliations and addresses, abstract, and references. For our experiments, we chose the CiteSeer data files from December 13, 2005. These are the most recent data files prior to transforming CiteSeer into CiteSeer^X, which is dubbed "the next generation CiteSeer" and which is still in a beta version.

Possible data sources

CiteSeer is just one of many of bibliographic databases the most widely used of which are presented in Table 1. We may divide the databases into two groups according to their free availability or the way they are created and maintained. ACM Portal[12] consisting of the ACM

[11] http://citeseer.ist.psu.edu

[12] http://portal.acm.org

Digital Library and of the ACM Guide along with Scopus[13] and Web of Science[14] are commercial subscription-based services (although some limited free access is provided by ACM) whereas CiteSeer, DBLP[15], and Google Scholar[16] are free for everyone with an Internet connectivity. On the other hand, CiteSeer and Google Scholar are automated systems while the databases of ACM Portal, DBLP, Scopus, and Web of Science are created and maintained mostly manually needing much human labour.

Table 1 Feature matrix of the main bibliometric systems as of October 4, 2010

	ACM Portal	CiteSeer^x	DBLP	Google Scholar	Scopus	Web of Science
Free	partly	yes	yes	yes	no	no
Automated	no	yes	no	yes	no	no
# records	1.59 mil.	32.23 mil.	1.46 mil.	NA	42.74 mil.	45.68 mil.
All bibl. data downloadable	no	yes	yes	no	no	no
Reference linking	yes	yes	partly	no	yes	yes
Citation linking	yes	yes	partly	yes	yes	yes
# citations for a publication	yes	yes	partly	yes	yes	yes
# citations for an author	yes	indirectly	partly indirectly	indirectly	yes	yes
domain coverage	computer science	computer science	computer science	general	general	general

As for the scope of the individual databases, the number of records in Table 1 means actually the number of all bibliographic records in the database, i.e. the number of research papers indexed plus the number of articles cited by the papers indexed that are not in the database. For instance, the ACM Digital Library contains 290 thousand documents; 1.59 million records are available in the ACM Guide. CiteSeer^X actually owns 1.67 million documents only. DBLP is somewhat different – it is not a document repository, it merely stores bibliographic records so there is no need to make a distinction between documents and records. Google Scholar does not reveal any details about its database so, with certainty, we can only say that, in October 2010, it provides about 8.94 million results as a response to the query "the". (We are looking for documents containing the most frequent English word.) . Some of the results are documents but some of them are cited references only. Thus, Google Scholar currently

[13] http://www.scopus.com

[14] http://apps.isiknowledge.com

[15] http://dblp.uni-trier.de

[16] http://scholar.google.com

provides access to no less than 9 million bibliographic records. Until now, solely estimates of the relative size of Google Scholar have been made by comparing its overlap with other bibliographic databases. Most of the papers on this topic are listed by Franceschet (2010).

While the absolute size of Google Scholar is unknown, a little bit more can be said about the documents it indexes – Meho and Yang (2007) report over 30 different document types in a sample of Google Scholar records such as journal articles, conference papers, dissertations, theses, technical reports, etc. (A similar earlier study by Goodrum et al. (2001) identified the following main document types in CiteSeer – journal articles, conference proceedings, technical reports, and books.) Indeed, regarding the same approach to obtaining documents by crawling the World Wide Web and looking for anything that looks like a research paper (a computer science research paper in the case of CiteSeer), one might expect that the document types covered by both Google Scholar and CiteSeer are almost the same.

Finally, the two huge human-made repositories of scientific literature, Scopus and Web of Science, make both available over 40 million records. Those 42.74 million records in Scopus can be really retrieved, for instance by searching for articles with an arbitrary title ("%"). If we restrict the search to articles published since 1996, we get the actual number of full-text documents in the database – 22.37 million. This number (of full-text documents) cannot be found out from the Web of Science.

Of the six databases, only CiteSeer and DBLP provide a full access to their bibliographic data in the form of one or more XML-like files. Unlike DBLP (see Fiala et. al. 2008), CiteSeer data records are substantially more linked by citations. The free availability of downloadable XML data and the high density of the citation graph are the key features that make CiteSeer the best tool for automated bibliometric and citation analyses despite its errors.

The other features in Table 1 describe more or less the user interface friendliness of the databases. In some of them, the user can go directly to the cited articles by clicking on the references in a paper (*reference linking*) or to the citing articles of the current paper (*citation linking*). We can get citation counts for an author directly or indirectly by counting citations to its publications (*citations for a publication* and *citations for an author*). These features are very limited in DBLP as it contains very few links between publications. The last aspect is the domain coverage of the databases – ACM Portal, CiteSeer, and DBLP cover mainly computer science whereas Google Scholar, Scopus, and Web of Science are general services. Let us recall that this paper deals with CiteSeer (and not CiteSeer^{X}) and that the relevant information in Table 1 is true for both of them except for the number of records.

Citation graphs

CiteSeer data are much larger than DBLP data analyzed by Fiala et al. (2008). There are more than 1.8 million citations between 717 thousand publications. We took the publication citation graph as it was and constructed an author citation graph out of it. The only data preprocessing we performed was transforming author names into upper case, removing duplicate authors, parallel edges, and self-citations. The resulting directed graph G of citations between authors has then some 411 thousand vertices (authors) and 4.8 million weighted edges (citations).

We made no attempt at disambiguating authors and publications, which is a complicated and time-consuming task. Thus, one author name may represent many real people and a single researcher may be referred to with several names, e.g. "Jack Dongarra" and "Jack J. Dongarra" at positions 9 and 13 of the first ranking in Table 6 (Online Resource 1). Also, automatic name recognition in CiteSeer produces errors and may identify absurd words as author names, e.g. "Senior Member" or "Student Member" at positions 2 and 4 of the first ranking in the same table. As for publications, there may also be duplicates and other inaccuracies. It is unclear whether CiteSeer groups all similarly looking publications found on the Web into one and if so, with what precision this happens. Nevertheless, if this was not the case, one might easily bias CiteSeer citation counts by placing many copies of particular articles all over the Web. We can expect as well that small typos in paper titles may wrongly result in new or missing publications, etc.

All in all, computer-generated Web-based bibliographic data like in CiteSeer are always less reliable than those created by humans like in DBLP. This is one of the reasons why they have been so little used in bibliometric studies so far. On the other hand, they are much larger and much more up-to-date and we believe that the democratic, decentralized, and self-controlled nature of the Web itself makes it very difficult to manipulate Web-based bibliographic citations significantly and systematically. Zhao (2005) indicates that citation analyses based on CiteSeer may be as valid as those based on conventional data sources. Therefore, analyzing CiteSeer data makes sense and can bring new bibliometric insights into recent computer science publications.

Results

In the following tables and figures, we present the results of applying twelve different ranking methods to the amended citation graph of authors described earlier. The first five rankings are

by pure citation counts (*Cites*), in-degree of author citation graph nodes (*InDeg*), HITS authorities (*HITS* - see Kleinberg 1999), PageRank (*PR*), and weighted PageRank (*w*). Next, we computed the previously defined parameters *c*, *f*, *g*, *h*, *hd*, *t*, and *td* from the collaboration graph, incorporated them into the PageRank formula (for details, see Fiala et al. 2008) and obtained rankings *a*) – *g*) corresponding to the numbering in section Methods.

Rankings

In addition to computing the ranks of all authors in the citation graph by each ranking method, we also compared each ranking with the list of ACM SIGMOD E. F. Codd Innovations Award winners (http://www.sigmod.org/awards) like Sidiropoulos and Manolopoulos (2005) to see how well they correlate with human-made charts of influential computer scientists. In Tables 2 and 3, we can see the ranks by all methods of 18 researchers awarded from 1992 to 2009. One of the researchers, Patricia Selinger, does not appear in any ranking. She is not present in the CiteSeer data we analyzed. For all rankings, we calculated three simple metrics characterizing the aggregate rank achieved by the awardees – worst rank, average rank, and median rank. The assumption is that the smaller are these values, the better is the ranking. In fact, an optimal ranking (including Patricia Selinger) equivalent to the human-made list in terms of these metrics, would have a worst rank of 18, an average rank of 9.5, and a median rank of 9.5.

Table 2 ACM Innovations Award winners and their ranks (part 1)

Year	Author	Cites	InDeg	HITS	PR	w
1992	Michael Stonebraker	137	87	170	35	36
1993	Jim Gray	194	132	132	287	367
1994	Philip Bernstein	1 477	1 767	2 055	4 884	4 749
1995	David DeWitt	27	38	84	75	43
1996	C. Mohan	2 634	2 419	3 996	4 945	4 958
1997	David Maier	458	284	604	375	521
1998	Serge Abiteboul	22	54	322	123	69
1999	Hector Garcia-Molina	14	14	58	89	63
2000	Rakesh Agrawal	3	9	112	41	15
2001	Rudolf Bayer	29 834	26 272	19 969	43 206	48 897
2002	Patricia Selinger					
2003	Don Chamberlin	5 497	4 577	4 474	7 162	9 125
2004	Ronald Fagin	512	587	1 160	701	774
2005	Michael Carey	161	163	220	308	306
2006	Jeffrey D. Ullman	228	205	476	609	575
2007	Jennifer Widom	7	15	103	81	29
2008	Moshe Vardi	217	326	1 622	447	441
2009	Masaru Kitsuregawa	16 497	12 603	7 972	27 477	42 133
	Worst rank	**29 834**	**26 272**	**19 969**	**43 206**	**48 897**
	Average rank	**3 407**	**2 915**	**2 561**	**5 344**	**6 653**
	Median rank	**217**	**205**	**476**	**375**	**441**

Table 3 ACM Innovations Award winners and their ranks (part 2)

Year	Author	a	b	c	d	e	f	g
1992	Michael Stonebraker	33	81	103	85	75	40	40
1993	Jim Gray	335	917	1 238	698	879	479	396
1994	Philip Bernstein	4 871	2 858	2 280	2 907	2 914	4 462	4 642
1995	David DeWitt	55	46	49	30	45	22	42
1996	C. Mohan	4 877	5 357	5 502	5 269	5 340	5 327	5 095
1997	David Maier	537	169	128	117	161	446	473
1998	Serge Abiteboul	76	43	47	36	42	44	66
1999	Hector Garcia-Molina	78	23	22	45	18	34	76
2000	Rakesh Agrawal	17	15	19	15	14	17	17
2001	Rudolf Bayer	48 600	52 676	54 482	51 648	52 522	49 505	49 098
2002	Patricia Selinger							
2003	Don Chamberlin	8 880	13 497	18 963	12 341	13 129	9 879	9 236
2004	Ronald Fagin	838	419	457	476	416	658	795
2005	Michael Carey	310	620	689	430	580	314	312
2006	Jeffrey D. Ullman	560	427	349	547	388	415	588
2007	Jennifer Widom	43	24	23	28	21	20	30
2008	Moshe Vardi	507	100	114	144	106	349	443
2009	Masaru Kitsuregawa	42 179	44 500	44 869	44 072	44 531	42 659	42 558
	Worst rank	**48 600**	**52 676**	**54 482**	**51 648**	**52 522**	**49 505**	**49 098**
	Average rank	**6 635**	**7 163**	**7 608**	**6 993**	**7 128**	**6 745**	**6 700**
	Median rank	**507**	**419**	**349**	**430**	**388**	**415**	**443**

The baseline ranking PR appears in a coloured column. It has a median rank of 375 which is outperformed only by ranking *c*) – ALL_COAUTHORS and by the both first-order methods *Cites* and *InDeg*. Its average rank 5 344 is forth best after *HITS*, *InDeg*, and *Cites*. The same holds for its worst rank 43 206. *HITS* is, somewhat surprisingly, the best ranking method as for the worst and the average rank. However, this is particularly thanks to the relatively high ranks (small numbers) for Rudolf Bayer and Masaru Kitsuregawa in comparison to the other rankings. On the other hand, it is the second worst ranking in terms of the median rank. Only method *a*) – COLLABORATION is worse in this respect.

A graphical presentation of the results in Tables 2 and 3 is given in Figure 1. Rakesh Agrawal, Jennifer Widom, and Hector Garcia-Molina are always top-ranked. While Rakesh Agrawal obtains the highest median rank of 16 and Hector Garcia-Molina never falls off the Top 100, Jennifer Widom's result is remarkable in that she received the award only in 2007 and thus could not attract citations after her nomination (CiteSeer data are from 2005). The rank series are quite stable – there are no evident outliers except a slight deterioration by *HITS* for the better ranked authors.

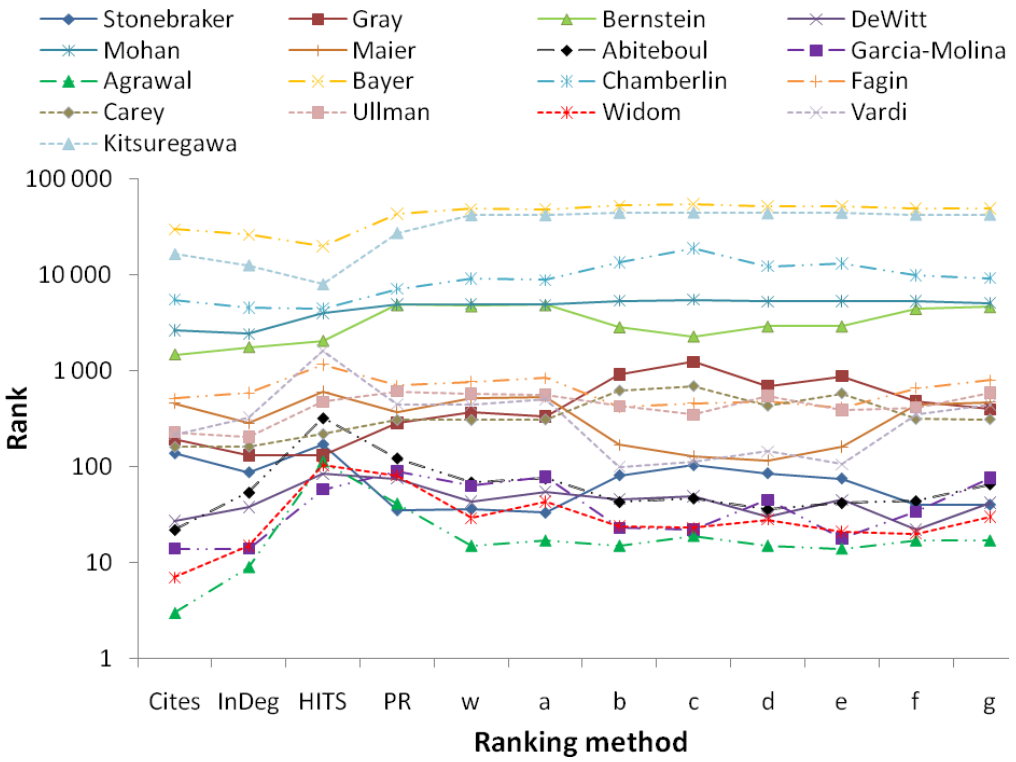


Fig. 1 ACM Innovations Award winners and their ranks

A complete overview of top 40 scientists in all rankings may be found in Tables 4 through 7 (Online Resource 1) with award recipients printed in bold. A simple look at the tables reveals that the number of award winners varies between 5 in *Cites* and *f* (COAUTHORS) or 4 in *InDeg* and *d* (ALL_DIST_CO-AUTHORS) and 1 in *PR* or even 0 in *HITS*. This suggests that as far as the top of each ranking is concerned, any improved PageRank (with some additional information from the collaboration graph) is closer to the real-world perception of a researcher's significance than the standard PageRank but is still at best as good as common (and far less computationally expensive) first-order methods based on simple citation counts.

The above tables may also be used for a prediction of future ACM SIGMOD E. F. Codd Innovations Award winners if we choose scientists active in the database field. Regarding the fact that citation and in-degree rankings have the largest overlap with the true list of awardees (see Table 2) and after consulting Scopus about the fields of interest of the top-ranked authors in Table 4 (Online Resource 1), Ramakrishnan Srikant and Christos Faloutsos seem to be the hot candidates. Scott Shenker, Sally Floyd, and Van Jacobson appear almost always among the top researchers in each ranking but as their interests do not focus on databases, they should be considered as candidates for other awards.

Conclusions

Current tools for analyzing social networks in the scientific community concentrate mainly on established citation indices such as ISI Web of Science or Scopus. These databases were originally not conceived to allow for a direct machine processing and, therefore, information scientists treat them manually or semi-manually. This approach results in very time-consuming analyses of relatively little data. On the other hand, the data from open access Web services such as CiteSeer are still rather underestimated as they are computer-generated and hence error-prone. However, their potential is great as their accuracy and completeness get higher and the general need for large and up-to-date bibliographic and citation databases grows.

In this paper, we present the results of our experiments with CiteSeer data. We show that sufficiently large citation and collaboration graphs for publications and authors can be created from these data. We analyze the citation graph of publication authors and present twelve rankings of the most influential researchers. In addition to common ranking methods such as counting citations or in-degree, we apply variations of the standard PageRank formula that combine information from both the citation and collaboration graphs. With respect to CiteSeer's drawbacks such as missing or wrong data, we argue that author rankings based on CiteSeer are realistic enough (by comparing them with true award recipients) so that they might be carefully used along with other data sources for the prediction of future computer science award winners. We conclude that CiteSeer, due to its free availability and well-structured large-scale data, is very well suited for citation analyses and testing of bibliometric methods despite its inherent errors. This work is the most comprehensive analysis of author citations based on CiteSeer data that we are aware of.

The remaining research issues are particularly the reliability of CiteSeer data, a more in-depth analysis of the CiteSeer collaboration graph, and differences between CiteSeer and CiteSeer^X. These and other topics including retrieving addresses, affiliations, and countries from CiteSeer shall be discussed in future studies.

Acknowledgements

This work[17] was supported in part by the Ministry of Education of the Czech Republic under Grant 2C06009. Many thanks go to the anonymous reviewers for their useful hints and comments and to Karel Ježek for his support of this project.

[17] The related software may found at http://textmining.zcu.cz/downloads/sciento.php.

References

AN, Y., JANSSEN, J., MILIOS, E. E. (2004). Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, 6(6): 664—678.

BAR-ILAN, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing and Management*, 42(6): 1553—1566.

BRIN, S., PAGE, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the 7th World Wide Web Conference, Brisbane, Australia, pp. 107-117.

CHAKRABARTI, S., AGARWAL, A. (2006). Learning parameters in entity relationship graphs from ranking preferences. *Lecture Notes in Computer Science*, 4213: 91-102.

CHEN, C. (2000). Domain visualization for digital libraries. Proceedings of the International Conference on Information Visualization (IV2000), London, UK, pp. 261-267.

FEITELSON, D. G., YOVEL, U. (2004). Predictive ranking of computer scientists using CiteSeer data. *Journal of documentation*, 60(1): 44-61.

FRANCESCHET, M. (2010). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, 83(1): 243-258.

FIALA, D., ROUSSELOT, F., JEŽEK, K. (2008). PageRank for bibliographic networks. *Scientometrics*, 76(1): 135-158.

GILES, C. L., COUNCILL, I. G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51): 17599-17604.

GOODRUM, A. A., MCCAIN, K. W., LAWRENCE, S., GILES, C. L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing and Management*, 37(5): 661-675.

HOPCROFT, J., KHAN, O., KULIS, B., SELMAN, B. (2004). Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(suppl. 1): 5249-5253.

JEŽEK, K., FIALA, D., STEINBERGER, J. (2008). Exploration and Evaluation of Citation Networks. Proceedings of the 12th International Conference on Electronic Publishing, Toronto, Canada, pp.351-362.

KLEINBERG, J. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5): 604-632.

MEHO, L. I., YANG, K. (2007). Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13): 2105–2125.

POPESCUL, A., UNGAR, L. H., LAWRENCE, S., PENNOCK, D. M. (2003). Statistical relational learning for document mining. Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), Melbourne, Florida, USA, pp. 275-282.

SIDIROPOULOS, A., MANOLOPOULOS, Y. (2005). A citation-based system to assist prize awarding. *SIGMOD Record*, 34 (4): 54–60.

ŠINGLIAR, T., HAUSKRECHT, M. (2006). Noisy-OR Component Analysis and its Application to Link Analysis. *Journal of Machine Learning Research*, 7: 2189–2213.

ZHAO, D. (2005). Challenges of scholarly publications on the Web to the evaluation of science - A comparison of author visibility on the Web and in print journals. *Information Processing & Management*, 41(6): 1403-1418.

ZHAO, D., LOGAN, E. (2002). Citation analysis using scientific publications on the Web as data source: A case study in the XML research area. *Scientometrics*, 54(3): 449-472.

ZHAO, D., STROTMANN, A. (2007). Can citation analysis of web publications better detect research fronts? *Journal of the American Society for Information Science and Technology*, 58 (9): 1285-1302.

ZHOU, D., COUNCILL, I., ZHA, H., GILES, C. L. (2007). Discovering temporal communities from social network documents. Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM'07), Omaha, Nebraska, USA, pp. 745-750.

Table 4 Top 40 CiteSeer authors for each ranking (part 1)

	Cites		InDeg		HITS
1	Scott Shenker	15 343	Scott Shenker	4 745	Scott Shenker
2	Sally Floyd	13 076	Senior Member	4 253	M. Frans Kaashoek
3	**Rakesh Agrawal**	12 988	Van Jacobson	4 114	Van Jacobson
4	Van Jacobson	12 654	Sally Floyd	3 941	Hari Balakrishnan
5	Deborah Estrin	11 097	M. Frans Kaashoek	3 775	Sally Floyd
6	M. Frans Kaashoek	10 358	Deborah Estrin	3 686	Lixia Zhang
7	**Jennifer Widom**	9 702	Lixia Zhang	3 570	Deborah Estrin
8	Hari Balakrishnan	9 496	Student Member	3 457	Robert Morris
9	Don Towsley	9 476	**Rakesh Agrawal**	3 455	Steven McCanne
10	Lixia Zhang	9 418	Hari Balakrishnan	3 425	Don Towsley
11	Ian Foster	9 218	John K. Ousterhout	3 394	Hui Zhang
12	Thomas A. Henzinger	8 110	Ian Foster	3 010	Vern Paxson
13	Willy Zwaenepoel	8 091	Don Towsley	2 988	Larry L. Peterson
14	**Hector Garcia-Molina**	7 777	**Hector Garcia-Molina**	2 847	Willy Zwaenepoel
15	Vern Paxson	7 581	**Jennifer Widom**	2 834	Ion Stoica
16	Steven McCanne	7 551	Steven McCanne	2 773	Y H. Katz
17	Robert Morris	7 525	Robert Morris	2 596	Thomas E. Anderson
18	Hui Zhang	7 310	Hui Zhang	2 532	John Kubiatowicz
19	Senior Member	7 157	Vern Paxson	2 482	Henry M. Levy
20	Carl Kesselman	6 989	Willy Zwaenepoel	2 458	Mark Handley
21	Ramakrishnan Srikant	6 683	Randal E. Bryant	2 371	Richard Karp
22	**Serge Abiteboul**	6 575	Carl Kesselman	2 356	Eric A. Brewer
23	Randal E. Bryant	6 463	Takeo Kanade	2 340	Peter Druschel
24	Ken Kennedy	6 383	Thorsten Von Eicken	2 329	David Culler
25	David B. Johnson	6 213	Fachbereich Informatik	2 274	Brian N. Bershad
26	John K. Ousterhout	6 078	Ramakrishnan Srikant	2 235	David Karger
27	**David J. DeWitt**	5 949	Henry M. Levy	2 234	Srinivasan Seshan
28	Bart Selman	5 926	Thomas E. Anderson	2 232	Steven D. Gribble
29	Student Member	5 875	M. P. Vecchi	2 223	Stefan Savage
30	Larry L. Peterson	5 479	C. D. Gelatt	2 223	John K. Ousterhout
31	Rajeev Alur	5 459	S. Kirkpatrick	2 223	Jim Kurose
32	Anoop Gupta	5 445	David B. Johnson	2 204	M. Satyanarayanan
33	Christos Faloutsos	5 418	David Culler	2 196	Jeffrey C. Mogul
34	Thorsten Von Eicken	5 362	David E. Culler	2 192	David B. Johnson
35	Monica S. Lam	5 285	Larry L. Peterson	2 169	David E. Culler
36	Walter Willinger	5 237	David Harel	2 152	Thorsten Von Eicken
37	M. Satyanarayanan	5 192	Jack Dongarra	2 113	Thomas Anderson
38	Ion Stoica	5 175	**David J. DeWitt**	2 113	Peter B. Danzig
39	Jim Kurose	5 150	Anoop Gupta	2 101	Sylvia Ratnasamy
40	Marc Levoy	5 055	Y H. Katz	2 070	Venkata N. Padmanabhan
	Missed: 137. M. Stonebraker, 161. M. Carey, 194. J. Gray, 217. M. Vardi, 228. J. Ullman, 458. D. Maier, 512. R. Fagin, 1 477. P. Bernstein, 2 634. C. Mohan, 5 497. D. Chamberlin, 16 497. M. Kitsuregawa, 29 834. R. Bayer		Missed: 54. S. Abiteboul, 87. M. Stonebraker, 132. J. Gray, 163.M. Carey, 205. J. Ullman, 284.D. Maier, 326. M. Vardi, 587. R. Fagin,1 767. P. Bernstein, 2 419. C. Mohan, 4 577. D. Chamberlin, 12 603. M. Kitsuregawa, 26 272. R. Bayer		Missed: 58. H. Garcia-Molina, 84. D. DeWitt, 103. J. Widom, 112. R. Agrawal, 132. J. Gray, 170. M. Stonebraker, 220. M. Carey, 322. S. Abiteboul, 476. J. Ullman, 604. D. Maier, 1 160. R. Fagin, 1 622. M. Vardi, 2 055. P. Bernstein, 3 996. C. Mohan, 4 474. D. Chamberlin, 7 972. M. Kitsuregawa, 19 969. R. Bayer

Table 5 Top 40 CiteSeer authors for each ranking (part 2)

	PR	w	a
1	G. J. Chaitin	John K. Ousterhout	G. J. Chaitin
2	Ashok Chandra	G. J. Chaitin	John K. Ousterhout
3	John K. Ousterhout	Van Jacobson	Gregory J. Chaitin
4	Whitfield Diffie	Whitfield Diffie	Whitfield Diffie
5	Martin E. Hellman	Gregory J. Chaitin	Martin E. Hellman
6	S. Kent	Martin E. Hellman	Van Jacobson
7	Jack J. Dongarra	Scott Shenker	Ashok Chandra
8	Randolph Bentson	Ashok Chandra	S. Kent
9	Scott Shenker	S. Kent	Scott Shenker
10	Van Jacobson	Sally Floyd	Randal E. Bryant
11	Michael Burrows	Lixia Zhang	Sally Floyd
12	Randal E. Bryant	Randal E. Bryant	Lixia Zhang
13	George W. Furnas	Jack J. Dongarra	Jack J. Dongarra
14	Ken Thompson	Deborah Estrin	Deborah Estrin
15	Dennis M. Ritchie	**Rakesh Agrawal**	Randolph Bentson
16	Stephen C. Johnson	Randolph Bentson	Michael Burrows
17	Lixia Zhang	Michael Burrows	**Rakesh Agrawal**
18	Butler W. Lampson	Vern Paxson	George W. Furnas
19	D. Balenson	George W. Furnas	Ken Thompson
20	Adi Shamir	Adi Shamir	Adi Shamir
21	A. Brewer	Ken Thompson	Dennis M. Ritchie
22	Senior Member	Steven McCanne	Vern Paxson
23	Gregory J. Chaitin	Dennis M. Ritchie	D. Balenson
24	Sally Floyd	Domenico Ferrari	Eli Biham
25	Neil Immerman	Eli Biham	L. Adleman
26	Electronic Eric	D. Balenson	A. Shamir
27	Eli Biham	Ken Kennedy	R. L. Rivest
28	Deborah Estrin	M. Frans Kaashoek	A. Brewer
29	Jacob Ziv	**Jennifer Widom**	Steven McCanne
30	C. D. Gelatt	L. Adleman	Domenico Ferrari
31	M. P. Vecchi	R. L. Rivest	Raj Jain
32	S. Kirkpatrick	A. Shamir	Kwangjo Kim
33	In R. D. Levine	A. Brewer	**Michael Stonebraker**
34	M. Tribus	Don Towsley	M. Frans Kaashoek
35	**Michael Stonebraker**	Thomas A. Henzinger	Neil Immerman
36	L. Adleman	**Michael Stonebraker**	Don Towsley
37	R. L. Rivest	Raj Jain	Stephen C. Johnson
38	A. Shamir	Senior Member	Electronic Eric
39	Kwangjo Kim	Neil Immerman	Butler W. Lampson
40	Wayne Jouberty	Kwangjo Kim	Senior Member
	Missed: 41. R. Agrawal, 75. D. DeWitt, 81. J. Widom, 89. H. Garcia-Molina, 123. S. Abiteboul, 287. J. Gray, 308. M. Carey, 375. D. Maier, 447. M. Vardi, 609. J. Ullman, 701. R. Fagin, 4 884. P. Bernstein, 4 945. C. Mohan, 7 162. D. Chamberlin, 27 477. M. Kitsuregawa, 43 206. R. Bayer	43. D. DeWitt, 63. H. Garcia-Molina, 69. S. Abiteboul, 306. M. Carey, 367. J. Gray, 441. D. Maier, 575. J. Ullman, 774. R. Fagin, 4 749. P. Bernstein, 4 958. C. Mohan, 9 125. D. Chamberlin, 42 133. M. Kitsuregawa, 48 897. R. Bayer	Missed: 43. J. Widom, 55. D. DeWitt, 76. S. Abiteboul, 78. H. Garcia-Molina, 310. M. Carey, 335. J. Gray, 507. M. Vardi, 537. D. Maier, 560. J. Ullman, 838. R. Fagin, 4 871. P. Bernstein, 4 877. C. Mohan, 8 880. D. Chamberlin, 42 179. M. Kitsuregawa, 48 600. R. Bayer

Table 6 Top 40 CiteSeer authors for each ranking (part 3)

	b	c	d
1	Scott Shenker	Scott Shenker	Senior Member
2	Senior Member	Senior Member	Scott Shenker
3	Deborah Estrin	Student Member	Student Member
4	Student Member	Deborah Estrin	Sally Floyd
5	Sally Floyd	Sally Floyd	Jonathan Rees
6	Van Jacobson	Ken Kennedy	Van Jacobson
7	Oded Goldreich	Jack Dongarra	Deborah Estrin
8	Bart Selman	Van Jacobson	Ian Foster
9	Jack Dongarra	Bart Selman	Ken Kennedy
10	Ken Kennedy	Jonathan Rees	K. K. Ramakrishnan
11	Thomas A. Henzinger	Thomas A. Henzinger	Lixia Zhang
12	Lance Fortnow	Oded Goldreich	Bart Selman
13	Jack J. Dongarra	Lance Fortnow	Lance Fortnow
14	Moni Naor	Ian Foster	Jack Dongarra
15	**Rakesh Agrawal**	Toby Walsh	**Rakesh Agrawal**
16	K. K. Ramakrishnan	Don Towsley	Vern Paxson
17	Philip Wadler	K. K. Ramakrishnan	Moni Naor
18	Vern Paxson	Jack J. Dongarra	Thomas A. Henzinger
19	Don Towsley	**Rakesh Agrawal**	Jack J. Dongarra
20	Michael I. Jordan	Michael I. Jordan	Oded Goldreich
21	Lixia Zhang	Lixia Zhang	John K. Ousterhout
22	Toby Walsh	**Hector Garcia-Molina**	Don Towsley
23	**Hector Garcia-Molina**	**Jennifer Widom**	Steven McCanne
24	**Jennifer Widom**	Moni Naor	Randal E. Bryant
25	Jonathan Rees	David Culler	Hari Balakrishnan
26	Ian Foster	Philip Wadler	Michael I. Jordan
27	Randal E. Bryant	M. Frans Kaashoek	Pat Hanrahan
28	John K. Ousterhout	Steven McCanne	**Jennifer Widom**
29	Baruch Awerbuch	Baruch Awerbuch	Baruch Awerbuch
30	Steven McCanne	Y H. Katz	**David J. DeWitt**
31	Madhu Sudan	Vern Paxson	Y H. Katz
32	M. Frans Kaashoek	Pat Hanrahan	Toby Walsh
33	Noam Nisan	Randal E. Bryant	Ion Stoica
34	Hui Zhang	Hari Balakrishnan	Noam Nisan
35	Hari Balakrishnan	Robert E. Schapire	Tomaso Poggio
36	Robert E. Schapire	Anoop Gupta	**Serge Abiteboul**
37	David Culler	Christos Faloutsos	David Culler
38	Christos Faloutsos	Joseph M. Hellerstein	Hui Zhang
39	Y H. Katz	Mark D. Hill	Nancy Lynch
40	Tomaso Poggio	John K. Ousterhout	Philip Wadler
	Missed: 43. S. Abiteboul, 46. D. DeWitt, 81. M. Stonebraker, 100. M. Vardi, 169. D. Maier, 419. R. Fagin, 427. J. Ullman, 620. M. Carey, 917. J. Gary, 2 858. P. Bernstein, 5 357. C. Mohan, 13 497. D. Chamberlin, 44 500. M. Kitsuregawa, 52 676. R. Bayer	Missed: 47. S. Abiteboul, 49. D. DeWitt, 103. M. Stonebraker, 114. M. Vardi, 128. D. Maier, 349. J. Ullman, 457. R. Fagin, 689. M. Carey, 1 238. J. Gray, 2 280. P. Bernstein, 5 502. C. Mohan, 18 963. D. Chamberlin, 44 869. M. Kitsuregawa, 54 482. R. Bayer	Missed: 45. H. Garcia-Molina, 85. M. Stonebraker, 117. D. Maier, 144. M. Vardi, 430. M. Carey, 476. R. Fagin, 547. J. Ullman, 698. J. Gray, 2 907. P. Bernstein, 5 269. C. Mohan, 12 341. D. Chamberlin, 44 072. M. Kitsuregawa, 51 648. R. Bayer

Table 7 Top 40 CiteSeer authors for each ranking (part 4)

	e	f	g
1	Scott Shenker	Scott Shenker	Scott Shenker
2	Senior Member	Sally Floyd	Van Jacobson
3	Deborah Estrin	Van Jacobson	John K. Ousterhout
4	Student Member	John K. Ousterhout	G. J. Chaitin
5	Sally Floyd	Lixia Zhang	Sally Floyd
6	Van Jacobson	Deborah Estrin	Whitfield Diffie
7	Bart Selman	G. J. Chaitin	Gregory J. Chaitin
8	Oded Goldreich	Whitfield Diffie	Martin E. Hellman
9	Ken Kennedy	Martin E. Hellman	Ashok Chandra
10	Jack Dongarra	Jonathan Rees	Jonathan Rees
11	Thomas A. Henzinger	Gregory J. Chaitin	S. Kent
12	Lance Fortnow	Steven McCanne	Lixia Zhang
13	Jack J. Dongarra	Ashok Chandra	Randal E. Bryant
14	**Rakesh Agrawal**	Jack J. Dongarra	Jack J. Dongarra
15	Moni Naor	S. Kent	Deborah Estrin
16	K. K. Ramakrishnan	Randal E. Bryant	K. K. Ramakrishnan
17	Don Towsley	**Rakesh Agrawal**	**Rakesh Agrawal**
18	**Hector Garcia-Molina**	Ken Kennedy	Michael Burrows
19	Michael I. Jordan	Vern Paxson	Randolph Bentson
20	Lixia Zhang	**Jennifer Widom**	Vern Paxson
21	**Jennifer Widom**	K. K. Ramakrishnan	Steven Mccanne
22	Toby Walsh	**David J. DeWitt**	Ken Kennedy
23	Ian Foster	M. Frans Kaashoek	George W. Furnas
24	Jonathan Rees	Domenico Ferrari	Ken Thompson
25	Baruch Awerbuch	Michael Burrows	Adi Shamir
26	Vern Paxson	Thomas A. Henzinger	Dennis M. Ritchie
27	Philip Wadler	Randolph Bentson	Senior Member
28	Steven McCanne	Robert E. Schapire	Eli Biham
29	Randal E. Bryant	Adi Shamir	Domenico Ferrari
30	M. Frans Kaashoek	Hari Balakrishnan	**Jennifer Widom**
31	John K. Ousterhout	Senior Member	Thomas A. Henzinger
32	Madhu Sudan	George W. Furnas	D. Balenson
33	Hui Zhang	Hui Zhang	M. Frans Kaashoek
34	Hari Balakrishnan	**Hector Garcia-Molina**	Butler W. Lampson
35	Noam Nisan	Eli Biham	A. Brewer
36	David Culler	Bart Selman	Neil Immerman
37	Christos Faloutsos	Philip Wadler	L. Adleman
38	Y H. Katz	Ken Thompson	A. Shamir
39	Robert E. Schapire	Don Towsley	R. L. Rivest
40	Tomaso Poggio	**Michael Stonebraker**	**Michael Stonebraker**
	Missed: 42, S. Abiteboul, 45. D. DeWitt, 75. M. Stonebraker, 106. M. Vardi, 161. D. Maier, 388. J. Ullman, 416. R. Fagin, 580. M. Carey, 879. J. Gray, 2 914. P. Bernstein, 5 340. C. Mohan, 13 129. D. Chamberlin, 44 531. M. Kitsuregawa, 52 522. R. Bayer	Missed: 44. S. Abiteboul, 314. M. Carey, 349. M. Vardi, 415. J. Ullman, 446. D. Maier, 479. J. Gray, 658. R. Fagin, 4 462. P. Bernstein, 5 327. C. Mohan, 9 879. D. Chamberlin, 42 659. M. Kitsuregawa, 49 505. R. Bayer	Missed: 42. D. DeWitt, 66. S. Abiteboul, 76. H. Garcia-Molina, 312. M. Carey, 396. J. Gray, 443. M. Vardi, 473. D. Maier, 588. J. Ullman, 795. R. Fagin, 4 642. P. Bernstein, 5 095. C. Mohan, 9 236. D. Chamberlin, 42 558. M. Kitsuregawa, 49 098. R. Bayer

Article 4

The following short paper is a first attempt at coping with the problem I call "lifetime achievement versus current performance" in the Introduction. The paper has been accepted for publication in the prestigious journal *Journal of the American Society for Information Science and Technology* and has been published ahead of print at the time of writing this thesis (December 2013). The key concept is that common scientometric indicators such as the times cited or the h-index play generally in favour of more senior scientists because these metrics use time windows of the same lengths as researchers' careers. Therefore, an "older" researcher always has a larger publication/citation window than a "younger" scientist and has more time to write papers and collect citations to be reflected in his scientometric indicator. Thus, this indicator is likely to be higher than that of a "younger" scholar simply by definition. (Of course, the "scientific" age of a researcher may be different from the biological age.) This concept of "lifetime achievement" or "all-career" indicators may be perceived as unfair towards junior researchers whose current scientific performance (quantity and quality) might in reality be much superior to their indicators based on the established scientometric metrics.

It is a similar situation as if a tennis player or a chess player who won a couple of tournaments ten years ago was still ranked first in the ranking of his peers and the (possibly junior) players with the best current performance winning tournaments in the current year or in the year before were still positioned far behind this number one (senior) player. This situation would certainly be unsustainable and it is also the reason why many sports federations use explicit (like in tennis) or implicit (like in chess) "time windows" in the evaluation of their players' performance. By this sports analogy, I propose a 3-year publication/citation window for the assessment of researchers' performance expressed by their h-index achieved during the most recent three calendar years. I call the resulting indicator the *Current Index*. Its greatest advantage over the standard (or "all-career") h-index is that it is dynamic: it can grow as well as decline in the course of a researcher's career whereas the standard h-index can never decline. Both the Current Index and the standard h-index can stagnate, though. However,

for the Current Index to stagnate a researcher must, generally speaking, remain active – he must go on publishing and keep on being cited. Otherwise, his Current Index will soon decrease – no later than in a 3-years' time when the Current Index will finally drop down to zero. On the other hand, a researcher's h-index will stagnate if the researcher is no more cited and it can even grow if the researcher is no more active (he stops publishing) but keeps receiving citations (to his older publications). This is certainly in contradiction with the general perception of a fair and objective assessment of the current performance in any kind of human activity and is definitely resolved by the Current Index that can grow, decline, and stagnate in the course of a researcher's career.

Of course, the Current Index does not correct the h-index for some of its other well-known deficiencies such as self-citations or multi-authorship, but these may be fixed in a way similar to the h-index corrections proposed in the literature so far. The biggest question is how it will be accepted by the scientific community, which is rather conservative and might argue that science is not tennis or chess. This is certainly true, but I believe that some kind of dynamic scientometric indicators in the evaluation of researchers is inevitable and that they will appear in informetrics sooner or later, anyway.

Preprint of: Fiala, D. (2013). Current Index: A proposal of a dynamic rating system for researchers. *Journal of the American Society for Information Science and Technology*. DOI: 10.1002/asi.23049. (in press)

Current Index: A proposal of a dynamic rating system for researchers

Dalibor Fiala

University of West Bohemia, Department of Computer Science and Engineering

Univerzitní 8, 30614 Plzeň, Czech Republic

Phone: +420 377 63 24 29, fax: +420 377 63 24 02, email: dalfia@kiv.zcu.cz

Abstract: An index is proposed that is based on the h-index and a 3-year publication/citation window. When updated regularly, it shows the current scientific performance of researchers rather than their life-time achievement as indicated by common scientometric indicators. In this respect, the new rating scheme resembles established sports ratings such as in chess or tennis. By the example of ACM SIGMOD E. F. Codd Innovations Award winners and Priestley Medal recipients, we illustrate how the new rating can be represented by a single number and visualized.

Keywords: Rating, ranking, h-index, citations, publications, time.

Introduction

Hirsch proposed the h-index that combined both the productivity and impact of an individual researcher in a single number (Hirsch, 2005). The index is defined as follows: if we have a set of publications ordered by the number of times they are cited in descending order, the index *h* is the largest number *h* such that there are *h* publications having at least *h* citations each. Thus, a scholar with an h-index of 20 has published 20 papers at least (productivity) and has received no less than 400 citations (impact). The h-index attained a great popularity and was mathematically analyzed and praised, but it was also soon discovered that various corrections were needed. For instance, the h-indices of two researchers from different research fields or subfields are incomparable because publication and citation practice may vary to a great extent between those two fields. Also, it would be unfair to consider the h-indices of two scientists the same if one of the researchers always publishes with a large group of co-authors and the other researcher only publishes alone. In addition, author self-citations can inflate the h-index, etc. To remedy this situation, many h-index variants have been proposed, but their description is not the concern of this short paper that does not aim at the shortcomings above.

The h-index and other metrics based on it can be applied to any set of publications, for instance aggregated by institutions, countries, or journals, but this paper deals with individual researchers.

Scientometric indicators such as citation counts or h-index generally play in favour of more senior researchers because these simply have had more time to publish and collect citations. Therefore, the current metrics indicate a kind of lifetime achievement instead of current (most recent) performance, which is reflected in many sports ratings. There is a great need for an "age normalization" factor to be able to fairly compare researchers of different ages. Also, the new indicator should be able to grow as well as decline – it should be dynamic. We will introduce a dynamic indicator of scientific performance that will not only increase in time but also decrease according to the current publication activity and citation reputation. A model of such an indicator can be the ħ-index (h bar), which, contrary to the h-index, can decrease in time (Hirsch, 2010). But a decrease can only occur if the researcher under examination publishes new articles. If he/she stops publishing, the ħ-index (as well as all other related metrics) will never decline – it can only remain the same or grow. Our "*Current Index*" is able to change over time (increase as well as decrease) even if the scientist under study is not active because the new indicator considers a 3-year time window for both publications and citations and, therefore, reflects current performance rather than life-time achievement. This feature is common in many established sports rating systems such as in chess (FIDE Ratings, ratings.fide.com) or tennis[18], where the rating scheme is not biased towards more senior players. But we must be cautious with the *Current Index* as a researcher's performance is not always quantitatively countable and clear-cut compared to an athlete's performance. Therefore, whether the proposed scheme is a good "rating" mechanism for the evaluation of researchers needs to be debated.

Methods and data

In October 2012 we collected publication and citation data of all twenty ACM SIGMOD Edgar F. Codd Innovations Award[19] winners from Scopus. The Codd Award has been awarded annually since 1992 for outstanding contributions in the field of databases. We wanted to determine the winners' *Current Index* in the years 2003 – 2012. As *Current Index* (*CI*) uses a 3-year publication/citation window, the actual data collection time span was 2000 – 2011. For

[18] ATP Rankings, http://www.atpworldtour.com/Rankings/Singles.aspx

[19] Codd Award, http://www.sigmod.org/sigmod-awards/sigmod-awards#innovations

instance, the 2012 rating of a researcher is based on the papers published by him/her in the period 2009 – 2011 and on the citations received by these papers in the same period. Similarly, the 2003 rating takes into account the articles published in 2000 – 2002 and the citations to these articles in 2000 – 2002. For the sake of simplicity, we considered all document types and did not discard self-citations. Table 1 shows the results of our data collection for Héctor García-Molina, who won the Codd Award in 1999. In the rating year 2003, he published 29 publications that were cited 24 times, thus producing an h-index of 3 (denoted as h3-index in Table 1). Then, regarding the h3-index and the citation count, his *CI(2003)* is 3_{24}. It is a single compound number consisting of the h3-index and the citation count as its subscript. By analogy, the ratings of García-Molina are 4_{43} in 2004, 7_{140} in 2005, and so on.[20] The interpretation may be that from a modest starting point in 2003 he quickly reached his top form in 2005 and then gradually worsened his performance with a low in 2010 to finally achieve a good shape in 2011 and 2012 again. So far, the number of publications has not been involved because it does not seem practical to integrate it (perhaps as a superscript) in the rating score. Instead, it will be kept separately and used only as a further criterion to differentiate between researchers whose rating is the same. All in all, regularly updated (possibly on a yearly basis) h3-indices and citation counts (together as *Current Index*) and publication numbers (as a tiebreak score) represent a dynamic rating system changing in time that ranks researchers in a scientific discipline based on their current impact. The yearly ranks of García-Molina in the small set of twenty Codd Award winners are shown in the sixth column of Table 1. In the very last column of Table 1, the standard h-index known at the time of each specific rating year appears as "career h-index". For example, in 2003 the career h-index is based on publications and citations before 2003 (in Scopus), in 2004 it is based on publications and citations before 2004, etc. As we can see, unlike the h3-index that can grow and decline, the "career h-index" is non-decreasing since it represents life-time achievement. It does definitely not reflect current performance.

[20] If needed, the ratings can indeed be stored and manipulated as single decimal numbers with a fixed number of decimal digits, e.g. 3.0024, 4.0043, or 7.0140.

TABLE 1. Personal rating record of Héctor García-Molina.

Year	Rating	h3-index	Citations	Publications	Rank	Career h-index
2003	3_{24}	3	24	29	3	16
2004	4_{43}	4	43	42	2	19
2005	7_{140}	7	140	51	1	22
2006	6_{146}	6	146	53	3	26
2007	5_{82}	5	82	39	3	29
2008	5_{72}	5	72	28	7	32
2009	5_{86}	5	86	27	2	35
2010	4_{103}	4	103	34	3	37
2011	6_{196}	6	196	35	1	40
2012	6_{167}	6	167	30	1	43

Results and discussion

Once the rating score has been defined, it can be visualized. Figure 1 displays the rating progress charts of García-Molina and three other arbitrarily chosen Codd Award winners – Jeffrey D. Ullman (awarded in 2006), Serge Abiteboul (1998), and Rakesh Agrawal (2000). The blue line represents the h3-index progress and the red vertical bars the citation count achieved in the specific rating year with the actual number displayed next to them. One can immediately grasp when the researchers were "in good shape" (using the sports terminology) and when they were not. The size of the citation bars is proportional to the largest citation count in each researcher's chart and rescaled for a comfort look. Also, the upper bounds of each chart's Y-axis vary so as to visualize the progress of each scientist as well as possible. On the other hand, if the researchers should be compared to one another, the upper bounds of Y-axes and the sizes and scaling factors of citation bars may be adjusted accordingly for an immediate comparison. Note that publication counts are not present in the charts. They could possibly be displayed as bars below the h3-index line opposite the citation bars (which would then have to be depicted above the line only), but this would probably clutter up the charts with too much information.

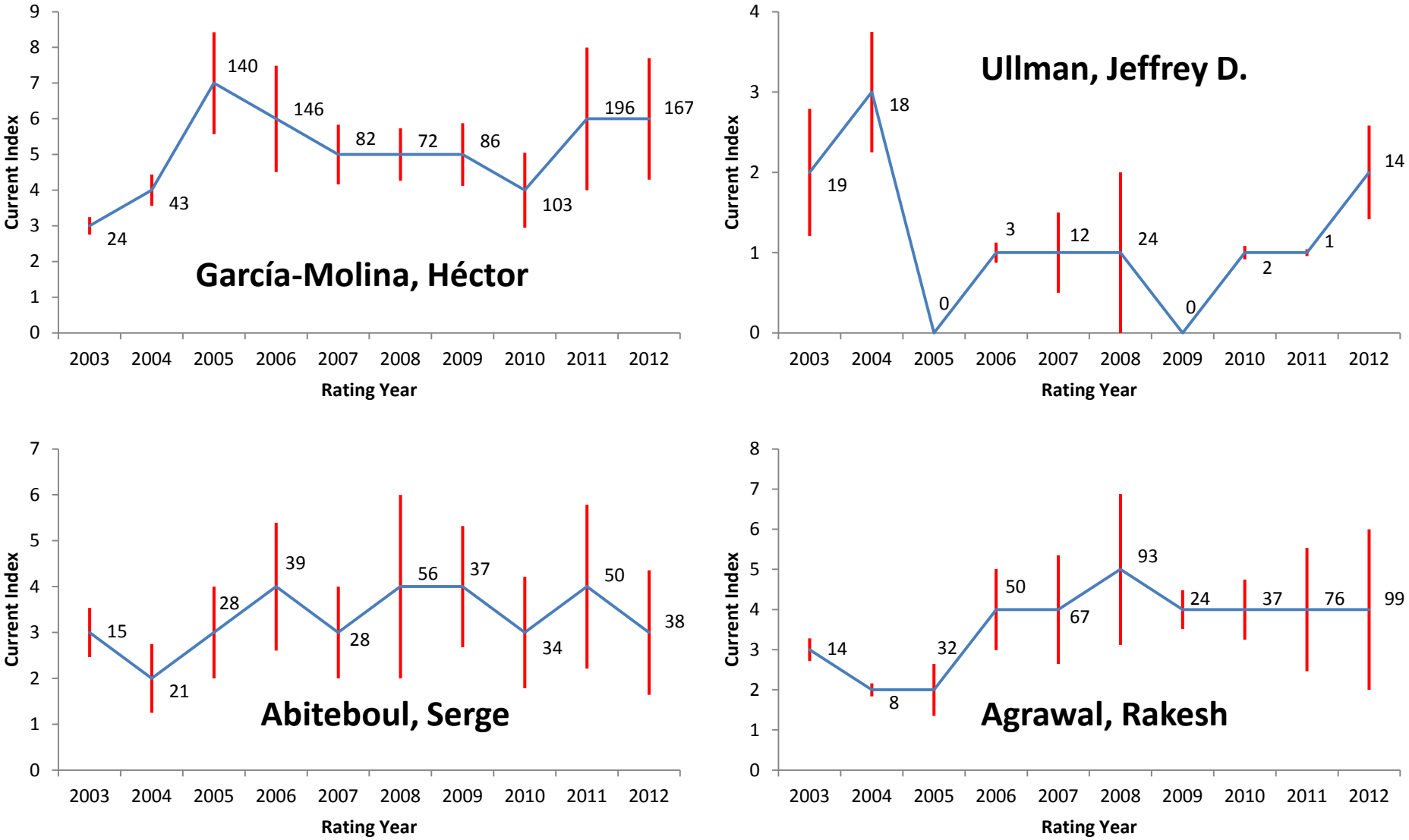


FIG. 1. Rating progress charts of four arbitrary Codd Award winners.

If the researchers' ratings in a scientific domain under study are updated once a year, an annual rating table may look like Table 2, in which researchers' 2012 ratings, ranks, and score changes in contrast to the previous year 2011 are shown. The look of the table is somewhat inspired by Live Chess Ratings (www.2700chess.com) in that positive changes (ranking or rating increase) in comparison to the previous rating year are marked in green with "↑" and "+" signs and negative changes (ranking or rating decline) in red with "↓" and "-" signs. The zero changes of ranks and ratings (strictly said, their constituents – h3-index, citation count, and publication count) that did not change from the previous rating year are not explicitly displayed, however. The table is well arranged to see quickly that, for example, Philip A. Bernstein (rank 14, rating 2_{25}) has dropped by 11 places since last year by decreasing his h3-index from 6 to 2, receiving 63 citations less and publishing 2 papers less than in the 2011 rating. On the other hand, García-Molina (rank 1, rating 6_{167}) maintained his rank but somewhat decreased his rating by keeping his h3-index and losing some citations and publications. Besides their names, the researchers in the table are indicated by their Scopus Author ID so that they can be identified unambiguously within Scopus. Of course, this rating table (Table 2) includes a very small number of researchers for whom the ratings could be computed manually. If researchers from a whole scientific field should be rated, the annual rating table would have to be generated automatically by means of computer programs. Since Scopus or Web of Science have data and software to produce world-wide scientometric indicators, they could easily integrate such annual field-specific rating tables of researchers within their products as a built-in feature. In fact, even researchers from various fields of science could be rated together if the underlying scientometric indicators (h-index, citation count, and publication number) and the time window length are corrected for the differences in publication and citation practices in those fields.

TABLE 2. 2012 *Current Index* ratings of all Codd Award winners.

Rank	Chg.	Name	Scopus ID	Rating	h3-index	Chg.	Citations	Chg.	Publications	Chg.
1		García-Molina, Héctor	7005594983	6_{167}	6		167	-29	30	-5
2	↑2	Chaudhuri, Surajit	7402978010	5_{58}	5		58	-9	26	-1
3	↓1	Stonebraker, Michael R.	7005476233	4_{157}	4	-2	157	+57	14	-1
4	↑1	Agrawal, Rakesh	7201475122	4_{99}	4		99	+23	18	+7
5	↑7	Dayal, Umeshwar	7006545572	4_{63}	4	+1	63	+38	48	+5
6	↑3	Vardi, Moshe Y.	7005334525	4_{60}	4		60	+20	36	+3
7		Dewitt, David J.	7101912578	3_{125}	3	-1	125	+61	5	-1
8	↓2	Widom, Jennifer	7006676535	3_{51}	3	-1	51	-13	11	+3
9	↓1	Abiteboul, Serge	7005292791	3_{38}	3	-1	38	-12	21	+3
10		Fagin, Ronald	7005757964	3_{32}	3	-1	32		14	+1
11	↑2	Carey, Michael J.	7202744401	3_{29}	3	+1	29	+2	12	+2
12	↓1	Gray, Jim O M	7404300349	2_{69}	2	-1	69	-392	2	-3
13	↑2	Kitsuregawa, Masaru	7005566641	2_{34}	2	+1	34	+19	47	-4
14	↓11	Bernstein, Philip A.	7102505937	2_{25}	2	-4	25	-63	16	-2
15	↑1	Ullman, Jeffrey D.	7004490091	2_{14}	2	+1	14	+13	7	+3
16	↓2	Maier, David	7103065333	1_{9}	1	-1	9		11	-2
17		Selinger, Patricia Griffiths	6701317222	1_{2}	1	+1	2	+2	3	+2
18	↑1	Mohan, Chander K J	7102973829	0_{0}	0		0		2	+2
19		Chamberlin, Donald D.	7005587366	0_{0}	0		0		0	
19	↓2	Bayer, Rudolf	7201391304	0_{0}	0		0		0	-1

The 3-year time window appears reasonable in the database field (and probably also in many other fields), but it can be adjusted to get a good balance between currency and sufficiency of publication/citation information in the research disciplines, where it is necessary. The time window length also influences how fast non-active researchers obtain a zero rating. With a 3-year time window, two consecutive years of inactivity can still yield a non-zero rating that grows or declines from the previous year. Alternatively to the rating progress charts in Figure 1, where the actual ratings can be seen, also ranking progress charts with researchers' ranks might be presented. This is commonplace in tennis, where the actual ratings are much less important than players' ranks. However, we believe that a researcher's current scientific performance is better reflected by a rating (rather than a rank), similarly to chess.

One might argue that if the publication and citation windows are the same (they are in a complete overlay), publications near the end of the publication (and citation) window have less time to collect citations than publications from the beginning of the time window. This is certainly true, but if that property is the same for all researchers in a scientific field, it may still be fair to compare the scientists using the same (3-year) publication/citation window. Alternatively, we propose two other time windows and present the ratings of García-Molina based on them in Table 3. The first variant is a 2-year publication window (rating year minus 4 and rating year minus 3) and a 4-year citation window (the four years preceding the rating year). For example, in 2003 the publication window is 1999 and 2000 and the citation window is the period 1999 – 2002. In this case, papers published in 2000 have a shorter citation window than those published in 1999, but all papers have two years at least (2001 and 2002) to gather citations (denoted as h4-index in Table 3). In the second alternative, the time window is defined as above, but there is a sliding 3-year citation window, e.g. in 2003 we look at publications from 1999 and their citations in 1999 – 2001 and at publications from 2000 and their citations in 2000 – 2002 (denoted as h4'-index in Table 3). Using this definition, all publications have "equal" conditions to obtain citations. (Of course, it can still happen that a paper published in January has a longer citation window than another paper published in December of the same year.)

One of the reviewers argued that there was a problem with the *Current Index* ignoring citations occurring within the time window to papers outside of (i.e. prior to) the time window and suggested that also "current" citations to "earlier" papers should contribute to the index. This is actually represented by a fixed 3-year citation window and a floating (ever-growing) publication window. While it is true that reflecting earlier work may bring more justice to the

rating and that inactive researchers may still have a non-zero rating with this approach, another problem arises: the rating loses its dynamics. This is demonstrated by the last five columns of Table 3 and denoted as h3'-index. As the numbers of publications used to calculate the index form a non-decreasing series by definition, the h3'-index will typically decline quite rarely or not decline at all as we can see with García-Molina whose rank remains static as well. In fact, 9 of the 20 researchers (45%) under study never experienced a decrease of their h3'-index compared to only 1 out of 20 (5%) whose *CI* never declined. Also, only 14% of all changes in the h3'-index were decreases whereas 34% of all changes in the *CI* were declines giving the rating equal chances to grow, fall, or stagnate. Moreover, the different natures of the h3'-index and *CI* are documented with quite uncorrelated rankings of scientists in various years with an average Pearson correlation coefficient of 0.286. Therefore, h3'-index cannot be used as a dynamic rating system.

We computed Spearman's rank correlation coefficients between the ten *Current Index* rankings from 2003 to 2012 of twenty E. F. Codd Award winners in 1992 – 2011 and the respective rankings based on the other two definitions of the time window (h4-index and h4'-index) and found a strong positive correlation ranging from 0.741 in 2009 to 0.971 in 2008 (with all coefficients being significant at the 0.01 level two-tailed). Regarding this high correlation and the simplicity and intuitive notion of the 3-year publication/citation window, it may be preferable to the other two time windows, especially with small-scale manual rating calculations, e.g. using the Scopus website. However, for automatic large-scale calculations based on off-line data, different publication/citation window definitions might also be considered. To show that *Current Index* works in other research disciplines as well, we computed ratings of twenty Priestley Medal[21] recipients in 1992 – 2011 (awarded by the American Chemical Society) and present the annual rating table for 2012 in Table 4. Unlike the database researchers in Table 2, there is a greater number of zero-rated chemistry researchers in Table 4, which may indicate that the Priestley Medal is more often conferred to scientists who are at the end of their careers or even no longer active.

[21] Priestley Medal, http://webapps.acs.org/findawards/detail.jsp?ContentId=CTP_004545

TABLE 3. Alternative ratings of Héctor García-Molina using different publication/citation windows.

Year	Rating	h4	Citations	Publications	Rank	Rating	h4‘	Citations	Publications	Rank	Rating	h3’	Citations	Publications	Rank
2003	5_{67}	5	67	19	1	4_{47}	4	47	19	2	13_{554}	13	554	161	1
2004	4_{41}	4	41	17	4	3_{36}	3	36	17	3	14_{734}	14	734	183	1
2005	6_{101}	6	101	20	1	5_{94}	5	94	20	1	16_{969}	16	969	200	1
2006	10_{290}	10	290	34	1	9_{204}	9	204	34	1	19_{1269}	19	1269	215	1
2007	9_{271}	9	271	39	1	7_{182}	7	182	39	1	21_{1541}	21	1541	223	1
2008	7_{152}	7	152	31	3	5_{105}	5	105	31	4	22_{1763}	22	1763	229	1
2009	6_{132}	6	132	22	7	5_{105}	5	105	22	6	23_{1929}	23	1929	242	1
2010	4_{123}	4	123	14	7	4_{94}	4	94	14	5	24_{2041}	24	2041	257	1
2011	5_{179}	5	179	19	4	5_{155}	5	155	19	3	25_{2196}	25	2196	264	1
2012	9_{343}	9	343	28	1	8_{262}	8	262	28	1	27_{2306}	27	2306	273	1

TABLE 4. 2012 *Current Index* ratings of 20 Priestley Medal winners.

Rank	Chg.	Name	Scopus ID	Rating	h3-index	Chg.	Citations	Chg.	Publications	Chg.
1		Whitesides, George M.	36038822100	17_{890}	17	-1	890	-107	105	
2		Somorjai, Gábor A.	35396886300	14_{682}	14		682	-155	67	-10
3	↑1	Bard, Allen J.	35350527400	12_{371}	12		371	-30	56	-2
4	↓1	Zewail, Ahmed H.	7004914740	11_{394}	11	-1	394	-17	53	-3
5	↑2	Oláh, George Andrew	36045924000	7_{232}	7		232	+99	48	+8
6	↓1	Corey, Elias James	7202254852	7_{163}	7	-2	163	-29	21	-11
7	↓1	Zare, Richard N.	35355951800	7_{151}	7		151	-38	46	+5
8		Breslow, Ronald C.	24443481400	7_{143}	7	+2	143	+64	20	+5
9	↑1	Frederick Hawthorne, M. Frederick	7102788963	3_{23}	3		23	+11	17	+3
10	↑1	Hoffman, Darleane C.	7402222195	3_{10}	3	+1	10	-2	4	-2
11	↓2	Albert Cotton, F. Albert	7201778183	2_{14}	2	-2	14	-26	9	-5
12		Djerassi, Carl	35600002000	1_{1}	1	+1	1	+1	2	-1
13		Anderson, Paul S.	7404425321	0_{0}	0		0		0	
14		Barton, Derek HR R	9272689500	0_{0}	0		0		0	
15		Basolo, Fred	7007150726	0_{0}	0		0		0	
16		Eliel, Ernest L.	7004876653	0_{0}	0		0		0	
17		Good, Mary L.	7202187884	0_{0}	0		0		0	
18		Parry, Robert W.	7101830447	0_{0}	0		0		0	
19		Simmons, Howard E.	35576856300	0_{0}	0		0		0	
19		Vandenberg, Edwin J.	7003531370	0_{0}	0		0		0	

Concluding remarks

Stefani (2011) reports that out of 159 international sports federations under study, only 60 (or 38%) have no rating system at all. On the other hand, 84 sports federations (53%) use an accumulative rating system, in which points are gathered over a time window. In these systems, senior and junior players have the same starting positions, which allows for an immediate comparison of players of different seniority. This stands in a stark contrast to the current scientometric indicators of the performance of researchers such as h-index or citation counts that accrue non-decreasingly over the researchers' careers and, therefore, are biased towards more senior scientists. To overcome this problem, we have introduced the *Current Index* which is an h-index based on a 3-year publication/citation window combined with a citation count for that time period. If the *Current Index* is equal for two or more researchers, the number of papers published in the specific period is used as a tiebreak criterion. We have shown that if the rating is updated regularly (possibly on a yearly basis), it may present a dynamic rating framework in which researchers' ratings (and ranks) can grow as well as decline in time according to their most recent performance like in many sports rating systems. A researcher's *Current Index* can be presented as a single (compound) number and its development is easy to visualize by means of a rating progress chart. Although we demonstrated the rating system on a very small set of researchers (ACM SIGMOD Edgar F. Codd Innovations Award winners and Priestley Medal recipients), it may be used to rate researchers in a whole scientific field or even across various fields if appropriate correction measures, which reflect different publication and citation patterns in those fields, are taken. However, these large-scale ratings cannot be performed manually, but the annual rating tables could be easily integrated within Scopus or Web of Science as a built-in feature. In addition, as athletes' life span and peak time can be very different from those of scientists, more evidence is needed before using the proposed mechanism to truly rate researchers.

Acknowledgements

This work was supported by the European Regional Development Fund (ERDF), project "NTIS - New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090. Many thanks are due to the reviewers for their insightful comments.

References

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569-16572.

Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. Scientometrics, 85(3), 741-754.

Stefani, R. (2011). The methodology of officially recognized international sports rating systems. Journal of Quantitative Analysis in Sports, 7(4), art. no. 10.

Article 5

There have been numerous scientometric studies on the research production and impact of institutions and some references to them will be given in the following article. The paper itself differs from such studies in that it does not analyze institutions as whole organizations but, instead, deals with suborganizations of these organizations. I adopted the term "suborganization" from the source XML files of a Web of Science data set acquired from Thomson Reuters and used in the research described in Article 1. The suborganizations of an institution (a university, a company, or a governmental body) are typically faculties, schools, divisions, sections, departments, groups, laboratories, etc. Some of the previously mentioned suborganizations (e.g. a faculty) would fall immediately below the main (or primary) organization (such as a university) in the institutional hierarchy and would itself consist of other suborganizations (such as departments), which could be called sub-suborganizations in the strict sense of word. I simply call the first group *level-1 suborganizations* and the second group *level-2 suborganizations*. There can even be higher-level suborganizations as, for instance, a scenario like *university – faculty - department – research group* is not uncommon in researchers' affiliations.

This leads us to the main problem with institutional suborganizations and also possibly to an explanation of the rarity of informetric analyses dealing with suborganizations – inconsistent author affiliations and addresses. Let us enumerate the main obstacles when quantitatively studying suborganizations of institutions that I sometimes call just *departments* for the sake of simplicity and brevity:

- Not all institutions are hierarchically organized. They can have a completely flat structure and thus have no suborganizations at all.
- Authors do not write their affiliations in a consistent way. They may sometimes start with a university followed by a department and other times the opposite is true.

- There are even more ambiguities, inconsistencies, errors, and omissions in the names of departments (including wrong translations) than in the names of institutions and no bibliographic database treats this problem appropriately and reliably.

These issues may well be the reason why the following study is one of the first large-scale investigations into the research performance of institutional suborganizations. I analyzed almost 47 thousand journal articles that appeared between 1991 and 2010 and were indexed in the Social Sciences Citation Index of the Web of Science database by Thomson Reuters in the category "Information Science & Library Science" (ISLS). Unlike Article 1, I did not acquire XML records of the articles to explore, but I manually downloaded text files with article records from the online version of Web of Science and parsed them in an automated way. This makes the problem of ambiguities and errors even more severe, but I believe that the results presented are valuable and not counterintuitive as shown in Table 2 on the top 30 departments sorted by their h-index based on the citation network of departments publishing articles in ISLS journals from 1991 to 2010.

	Department	h-index
1	Indiana Univ; Sch Lib & Informat Sci	24
2	Leiden Univ; Ctr Sci & Technol Studies	24
3	Univ Minnesota; Carlson Sch Management	22
4	Univ Sheffield; Dept Informat Studies	22
5	Harvard Univ; Sch Med	21
6	Rutgers State Univ; Sch Commun Informat & Lib Studies	20
7	Univ Georgia; Terry Coll Business	20
8	Univ Illinois; Grad Sch Lib & Informat Sci	18
9	Wolverhampton Univ; Sch Comp & Informat Technol	18
10	Penn State Univ; Sch Informat Sci & Technol	17
11	Univ Tampere; Dept Informat Studies	17
12	Univ Western Ontario; Fac Informat & Media Studies	17
13	City Univ London; Dept Informat Sci	16
14	Univ British Columbia; Fac Commerce & Business Adm	16
15	Univ Maryland; Robert H Smith Sch Business	16
16	Univ Tennessee; Sch Informat Sci	16
17	Univ N Carolina; Sch Lib & Informat Sci	15
18	Drexel Univ; Coll Informat Sci & Technol	14
19	Indiana Univ; Kelley Sch Business	14
20	Univ Calif Irvine; Grad Sch Management	14
21	Univ Maryland; Coll Lib & Informat Serv	14
22	Florida State Univ; Coll Business	13
23	Georgia State Univ; Coll Business Adm	13
24	Harvard Univ; Sch Publ Hlth	13
25	Hong Kong Univ Sci & Technol; Sch Business & Management	13
26	Queens Univ; Sch Business	13
27	Univ Pittsburgh; Katz Grad Sch Business	13
28	Univ Pittsburgh; Sch Informat Sci	13
29	Univ So Calif; Marshall Sch Business	13
30	Univ Warwick; Warwick Business Sch	13

Table 2: Top 30 departments by h-index based on the department citation network

Preprint of: Fiala, D. (2013). Suborganizations of institutions in library and information science journals. *Information*, 4(4), 351-366.

Suborganizations of institutions in library and information science journals

Dalibor Fiala

Department of Computer Science and Engineering, University of West Bohemia,

Univerzitní 8, Plzeň 30614, Czech Republic; E-Mail: dalfia@kiv.zcu.cz;

Tel.: +420-377-63-2429; Fax: +420-377-63-2402

Abstract: In this paper, we analyze Web of Science data records of articles published from 1991 to 2010 in library and information science (LIS) journals. We focus on addresses of these articles' authors and create citation and collaboration networks of departments which we define as the first suborganization of an institution. We present various rankings of departments (e.g., by citations, times cited, PageRank, publications, *etc*.) and highlight the most influential of them. The correlations between the individual departments are also shown. Furthermore, we visualize the most intense citation and collaboration relationships between "LIS" departments (many of which are not genuine LIS departments but merely affiliations of authors publishing in journals covered by the specific Web of Science category) and give examples of two basic research performance distributions across departments of the leading universities in the field.

Keywords: departments; ranking; PageRank; citations; collaborations

1. Introduction and Related Work

Bibliometric studies can roughly be conducted at three levels—individual researchers (micro-level), institutions (meso-level), and countries (macro-level). Of course, these "basic" levels can have their own sublevels (e.g., regions of a country) or they can be grouped into supralevels (such as continents). There have been many bibliometric analyses at various levels, but we can feel that at the meso-level those analyses have mainly concentrated on institutions as such or that they have not really been large-scale, *i.e.*, involving tens or hundreds of thousands of items to analyze. This study tries to bridge this gap in the field of library and information science (LIS) by analyzing several tens of thousands of bibliographic records at the meso-level and concentrating on the suborganizations of institutions. An institution (or the primary organization) usually has an organizational structure comprising some suborganizations (level 1) that themselves may consist of other suborganizations (level 2). The depth of this hierarchy may vary—some institutions have a relatively flat structure, while other hierarchies may include suborganizations of even higher levels. A typical academic institution (a university) may be divided into faculties, schools, departments, laboratories, and

research groups, which are difficult to capture in scientometric studies due to the inconsistent way they are present (or absent) in authors' addresses. As we will show later on, we will call level-1 suborganizations "departments" for the sake of simplicity. The main research questions of this study are the following: (a) Do Web of Science (WoS) data contain enough information to analyze the scientific performance and collaboration of the departments with which authors of journal articles in the LIS research area are affiliated? (hereafter called "LIS" departments); (b) What are the most intense citations and collaborations between "LIS" departments? and (c) Which "LIS" departments are the most highly ranked by various indicators based on publications from 1991–2010? Responses to these questions will be given in the next sections.

Bibliometric analysis of library and information science institutions has a long history in the United Kingdom. For instance, Bradley *et al*. [1] measured the publication patterns of the Department of Information Studies at the University of Sheffield, Holmes and Oppenheim [2] analyzed the citation impact of British LIS departments, and Oppenheim [3] ranked British LIS schools by citation impact. Seng and Willet [4] conducted a citation analysis of a small number of LIS departments in the UK and LIS departments in the UK were investigated by Webber [5]. British LIS departments were also analyzed webometrically—by Thomas and Willet [6] and by Arakaki and Willet [7]. As for other regions of the world, Aina and Mooko [8] analyzed a small set of top African LIS researchers and defined the centers of the African LIS research. Another tiny group of LIS publications was investigated by Herrero-Solana and Ríos-Gómez [9] to identify the most productive Latin American universities and departments. Meho and Spurgin [10] ranked American LIS schools by the visibility of their faculty in various databases and Yazit and Zainab [11] reported on the publication productivity in LIS of some Malaysian institutions. There have been two large-scale studies in which Yan and Sugimoto [12] explored citation patterns of various LIS institutions and He *et al*. [13] explored tens of thousands of LIS publications, but both of them remained at the institutional level. This study is the only large-scale one at the departmental level and the visualization tools used in this article are discussed by Shannon *et al*. [14].

2. Data and Methods

In November 2012 we manually queried the Web of Science web interface to obtain records of all articles published in the period 1991–2010 and indexed in the Social Sciences Citation Index in the research area "Information Science & Library Science" (ISLS). We were interested in the "article" document type only. In this way, we acquired plain text metadata on 46,800 journal articles. (Saving to plain text took about 50 min because a maximum of 500 records can be saved at once by anyone with a Web of Science subscription.) These metadata typically include an article's title, journal name, volume, issue, pagination, and year as well as its authors' names, addresses, times cited count and some other information. An example of a journal record is presented in Figure 1. As we can see, only some of the cited references (CR) can be identified unambiguously—in this case with a digital object identifier (DOI). The re-

maining references can be identified using the volume, issue, and pagination or cannot be identified at all. To create a citation network from the article records retrieved (a basic, root, or seed set of articles), we need one more tool.

Figure 1. A sample journal article record.

```
PT J
AU Schreiber, M
AF Schreiber, Michael
TI A new family of old Hirsch index variants
SO JOURNAL OF INFORMETRICS
LA English
DT Article
DE Hirsch index; g-Index; Performance evaluation; Citations; Ranking; Generalized mean
ID SCIENTIFIC-RESEARCH OUTPUT; H-INDEX; EGGHES G; PHYSICISTS
AB The Hirsch index h and the g index proposed by Egghe as well as the f index and the t index.
C1 Tech Univ Chemnitz, Inst Phys, D-09107 Chemnitz, Germany.
RP Schreiber, M (reprint author), Tech Univ Chemnitz, Inst Phys, D-09107 Chemnitz, Germany.
EM schreiber@physik.tu-chemnitz.de
CR Burrell QL, 2009, SCIENTOMETRICS, V79, P79, DOI 10.1007/s11192-009-0405-3
  Egghe L, 2006, SCIENTOMETRICS, V69, P131, DOI 10.1007/s11192-006-0144-7
  Egghe L., 2006, ISSI NEWSLETTER, V2, P8
  Egghe L, 2008, J AM SOC INF SCI TEC, V59, P1304, DOI 10.1002/asi.20823
  Hirsch JE, 2005, P NATL ACAD SCI USA, V102, P16569, DOI 10.1073/pnas.0507655102
  Prathap G, 2006, CURR SCI INDIA, V91, P1439
  Rousseau R., 2006, SIMPLE MODELS CORRES
  Schreiber M, 2007, ANN PHYS-BERLIN, V16, P640, DOI [10.1002/andp.200710252,
10.1002/andp.20071025]
  Schreiber M, 2010, J AM SOC INF SCI TEC, V61, P169, DOI 10.1002/asi.21218
  SCHREIBER M, 2010, ANN PHYS BERLIN
  Schreiber M, 2008, J AM SOC INF SCI TEC, V59, P1513, DOI 10.1002/asi.20856
  Schubert A, 2007, SCIENTOMETRICS, V70, P201, DOI 10.1007/s11192-007-0112-x
  Tol RSJ, 2009, SCIENTOMETRICS, V80, P317, DOI 10.1007/s11192-008-2079-7
  Woeginger GJ, 2009, J AM SOC INF SCI TEC, V60, P1267, DOI 10.1002/asi.21061
NR 14
TC 2
Z9 2
PU ELSEVIER SCIENCE BV
PI AMSTERDAM
PA PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS
SN 1751-1577
J9 J INFORMETR
JI J. Informetr.
PD OCT
PY 2010
VL 4
IS 4
BP 647
EP 651
DI 10.1016/j.joi.2010.05.002
PG 5
WC Information Science & Library Science
SC Information Science & Library Science
GA 647KA
UT WOS:000281616200020
ER
```

Therefore, in the next step, we used the Web Services Lite application programming interface (API) to retrieve the records of articles citing the articles in the basic set. This API is available for free to anyone with a Web of Science subscription after registration. In total, we got 175,139 citing article records. The information contained in the citing article records is somewhat less abundant than in the plain text seed article records. In particular, any author address information is missing. On the other hand, citing article records are structured in a similar way as XML records. See Figure 2 for an example of a citing article record. In the example, an article with ID (UT) 000283981500004 is cited by an article with ID 000283981500001. These IDs can then be matched with "UT WOS" in seed article records (see bottom of Figure 1) and, as a result, a complete citation network of the articles in the root set can be constructed. This citation graph had 94,836 edges, *i.e.*, slightly over 54% of all citations were citations within the seed set.

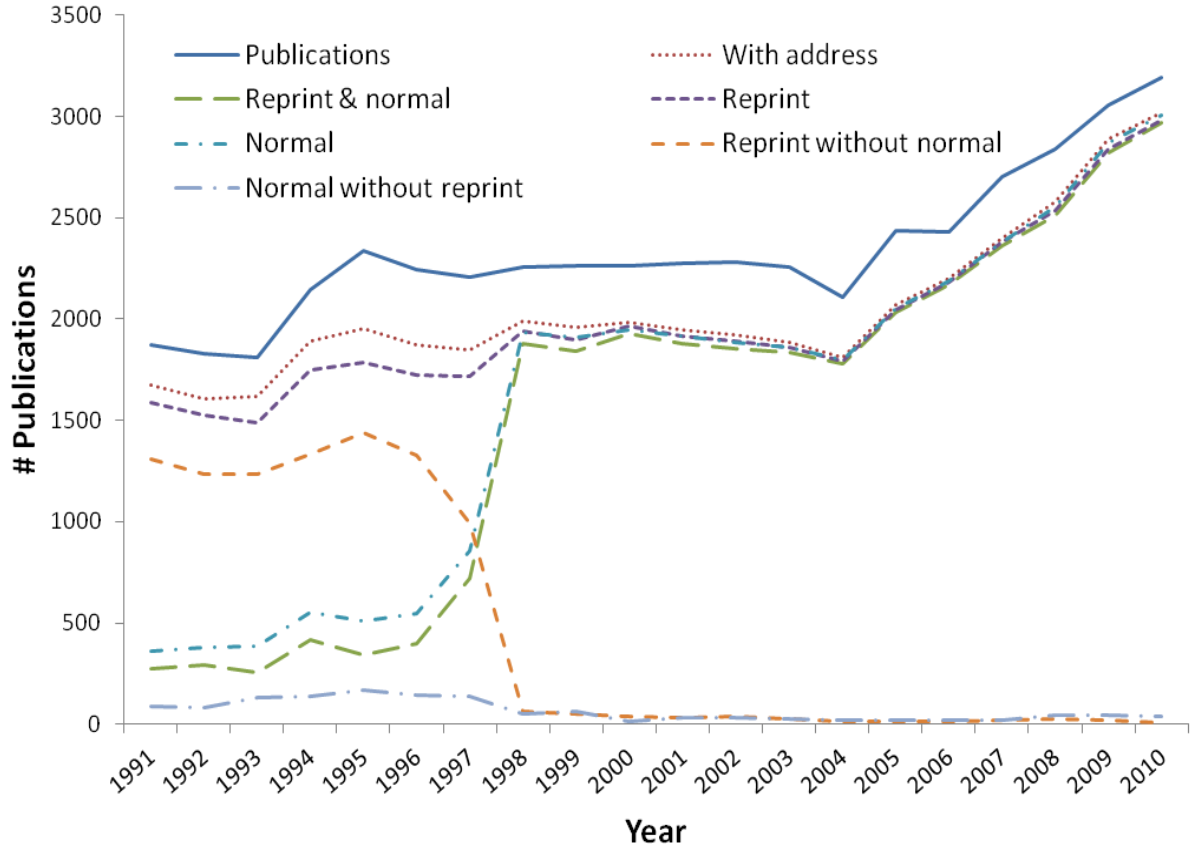
Figure 2. A sample citing article record.


```
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
<soap:Body><ns2:citingArticlesResponse xmlns:ns2="http://woksearchlite.cxf.wokmws.thomsonreuters.com">
<return><parent><authors><label>Authors</label><values>Urquhart, C</values><values>Thomas, R</values>
<values>Ovens, J</values><values>Lucking, W</values><values>Villa, J</values></authors>
<source><label>Issue</label><values>4</values></source><source><label>Pages</label><values>277-285
</values></source><source><label>Published.BiblioDate</label><values>DEC</values></source><source>
<label>Published.BiblioYear</label><values>2010</values></source><source><label>SourceTitle</label>
<values>HEALTH INFORMATION AND LIBRARIES JOURNAL</values></source>
<source><label>Volume</label><values>27</values></source><title><label>Title</label>
<values>Planning changes to health library services on the basis of impact assessment</values></title>
<UT>000283981500004</UT></parent><queryID>54</queryID><records><authors><label>Authors</label>
<values>Grant, MJ</values></authors><source><label>Issue</label><values>4</values></source>
<source><label>Pages</label><values>259-261</values></source>
<source><label>Published.BiblioDate</label><values>DEC</values></source><source>
<label>Published.BiblioYear</label><values>2010</values></source><source><label>SourceTitle</label>
<values>HEALTH INFORMATION AND LIBRARIES JOURNAL</values></source>
<source><label>Volume</label><values>27</values></source><title><label>Title</label>
<values>Writing for publication: ensuring you find the right audience for your paper</values></title>
<UT>000283981500001</UT></records><recordsFound>1</recordsFound><recordsSearched>3126432
</recordsSearched></return></ns2:citingArticlesResponse></soap:Body></soap:Envelope>
```


Since this paper is concerned with departments, the research depends on the extent to which affiliations and addresses of article authors are systematically present in the records we analyzed. There is no genuine affiliation information in the records, but there is often information on authors' addresses denoted with C1 and RP like in Figure 1. RP means a "reprint address", which is the address of the corresponding author (usually, but not always, the first author), and C1 is a field containing authors' addresses. Reprint and "normal" addresses may sometimes be the same, for instance when there is one author only. In total, almost 88% of publications had some address information associated with them and 65% had both reprint and normal address. 85% of publications had a reprint address and 68% had one normal address at least, but the latter percentage was quite different in various years under study as can be seen from Figure 3. While the share of publications with some address information has been about 90% throughout the period, the number of publications with one normal address at

least has only had a similar share since 1998. Before 1998 there was a high percentage of publications having a reprint but no normal address (from 45% to 70%), but this was almost negligible in later years and so was the number of articles having a normal address but no reprint address in the whole period 1991–2010.

Figure 3. Numbers of publications with different types of addresses.



As can be seen in Figure 1, addresses have a relatively clear structure starting with an institution followed with suborganizations (from bigger to smaller ones) and ending with a city and a country. Organizations (institutions) and suborganizations are written using standardized abbreviations and are delimited with commas as are cities and countries. In our experience, reprint addresses often include also other information such as street names and numbers or state or province names, *etc*. This additional information can distort the common address pattern "institution, suborganizaiton1, …, suborganizationN, city (+ZIP), country", but based on our experiments with random address samples and a manual checking of the pattern correctness, the pattern is violated in a few percent of cases even if reprint addresses are included. As a result, we made an approximation and considered all addresses in all publications in the period 1991–2010 as having an institution as their first item, a city and a country as their last item, and suborganizations in between. The number of suborganizations can vary as shown in Table 1. In the data under study, an institution (main organization) can have up to seven suborganizations associated with it, but most affiliations consist of an institution and its suborganization. Thus, before all the experiments whose results will be reported in the

next section, we retained suborganization 1 in each address and discarded the other suborganizations of higher levels. We will call the couple "institution; suborganization 1" a "department" because this is typically what is represented by that.

Table 1. Examples of various suborganizations of an institution.

Organization	Suborganization 1	Suborganization 2	Suborganization 3
Indiana Univ			
Indiana Univ	Sch Lib & Informat Sci		
Indiana Univ	Sch Business	Decis & Informat Syst Dept	
Indiana Univ	Sch Med	Dept Med	Div Gen Med & Geriatr

3. Results and Discussion

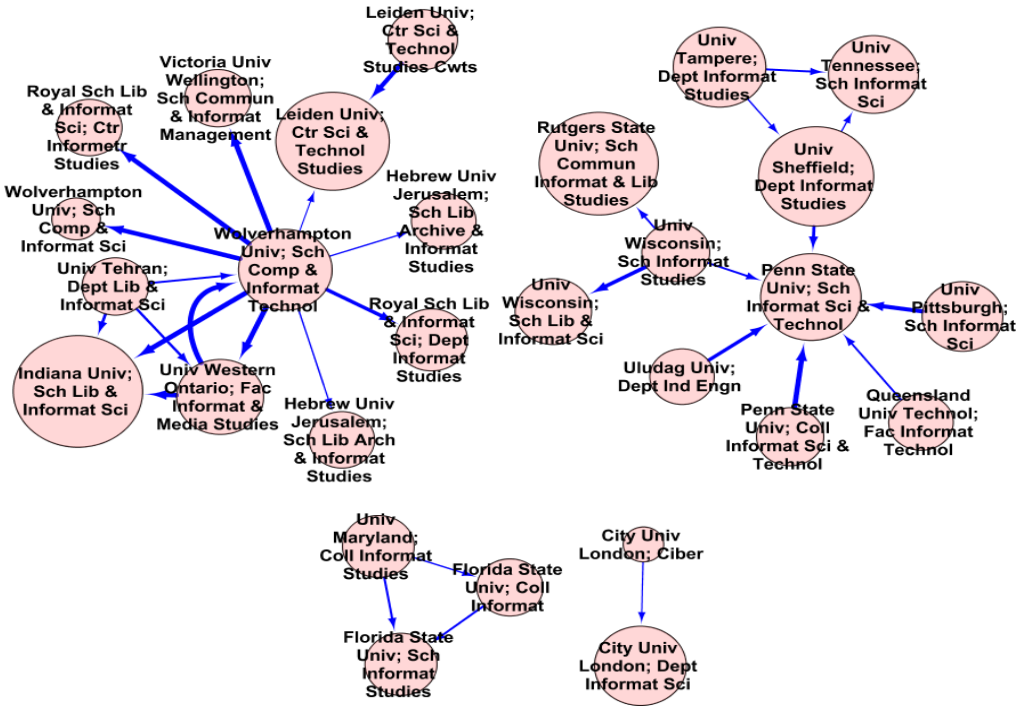
The citation graph of departments we obtained had 18,291 nodes and 154,744 edges. The graph is directed and the edges are weighted with an average weight of 2.62 per edge. The total sum of edge weights in the graph (404,755) is the total number of citations between departments. In Table 2 we can see the departments that received the most citations: "Indiana Univ; Sch Lib & Informat Sci", "Leiden Univ; Ctr Sci & Technol Studies", and "Univ Sheffield; Dept Informat Studies". However, the numbers of publications by which the departments are represented (see the last column in Table 2) vary significantly so "Leiden Univ; Ctr Sci & Technol Studies" with 3722 citations and 84 publications is actually relatively more cited than "Indiana Univ; Sch Lib & Informat Sci" with 4334 citations and 243 publications (44 citations per publication compared to 18). But the measure of citations per publication is obviously biased towards departments with fewer publications. For instance, the relatively most cited department in Table 2 is "Lib Hungarian Acad Sci; Bibliometr Serv" (position 33) with 165 citations per publication.

As far as the citations between individual departments are concerned, we can see the most intense of them in Figure 4. The size of nodes is based on the "times cited" (see below for an explanation) of a department and the thickness of edges depends on the number of citations from one department to another. We can notice that there are two big components—one centred around "Wolverhampton Univ; Sch Comp & Informat Technol" and the other one around "Penn State Univ; Sch Informat Sci & Technol". The most intense citations as such are those from "Wolverhampton Univ; Sch Comp & Informat Technol" to "Indiana Univ; Sch Lib & Informat Sci", "Victoria Univ Wellington; Sch Commun & Informat Management", and "Univ Western Ontario; Fac Informat & Media Studies". There are also intra-institutional citations such as from "Wolverhampton Univ; Sch Comp & Informat Technol" to "Wolverhampton Univ; Sch Comp & Informat Sci" or from "Penn State Univ; Coll Informat Sci & Technol" to "Penn State Univ; Sch Informat Sci & Technol", but these may sometimes be self-citations of departments that changed their names or whose names are used inconsistently. These errors are inherent in the Web of Science data and they could be removed only by means of a huge amount of manual effort. In total, we found that 4.3% of all citations were intra-institutional.

Table 2. Top 40 "library and information science (LIS)" departments by citations.

	Department	Citations	Publications
1	Indiana Univ; Sch Lib & Informat Sci	4334	243
2	Leiden Univ; Ctr Sci & Technol Studies	3722	84
3	Univ Sheffield; Dept Informat Studies	3606	195
4	Rutgers State Univ; Sch Commun Informat & Lib Studies	3413	144
5	Penn State Univ; Sch Informat Sci & Technol	3361	56
6	Univ Maryland; Robert H Smith Sch Business	3013	52
7	Univ Minnesota; Carlson Sch Management	2835	71
8	Univ Tennessee; Sch Informat Sci	2661	118
9	Drexel Univ; Coll Informat Sci & Technol	2288	101
10	Univ Tampere; Dept Informat Studies	2285	96
11	City Univ London; Dept Informat Sci	2162	192
12	Univ Western Ontario; Fac Informat & Media Studies	2125	138
13	Wolverhampton Univ; Sch Comp & Informat Technol	2068	109
14	Univ British Columbia; Fac Commerce & Business Adm	1821	26
15	Univ Illinois; Grad Sch Lib & Informat Sci	1710	167
16	Queens Univ; Sch Business	1651	24
17	Univ N Carolina; Sch Lib & Informat Sci	1630	102
18	Harvard Univ; Sch Med	1516	143
19	Univ Georgia; Terry Coll Business	1484	38
20	Florida State Univ; Coll Business	1447	36
21	Univ Virginia; Mcintire Sch Commerce	1413	18
22	Syracuse Univ; Sch Informat Studies	1273	162
23	Georgia State Univ; Coll Business Adm	1266	24
24	Univ Calif Irvine; Grad Sch Management	1261	25
25	Univ Wisconsin; Sch Lib & Informat Sci	1195	71
26	Royal Sch Lib & Informat Sci; Dept Informat Studies	1158	31
27	Univ Pittsburgh; Sch Informat Sci	1150	84
28	Univ So Calif; Marshall Sch Business	1139	28
29	City Univ Hong Kong; Dept Informat Syst	1064	64
30	Univ N Texas; Sch Lib & Informat Sci	1053	60
31	Univ Calif Los Angeles; Grad Sch Educ & Informat Studies	1015	42
32	Univ S Florida; Coll Business Adm	992	17
33	Lib Hungarian Acad Sci; Bibliometr Serv	991	6
34	Katholieke Univ Leuven; Steunpunt O&o Stat	984	20
35	Univ Arkansas; Sam M Walton Coll Business	973	11
36	Florida State Univ; Sch Informat Studies	971	53
37	Csic; Cindoc	966	32
38	Georgia State Univ; Dept Comp Informat Syst	966	27
39	Univ Wisconsin; Sch Lib & Informat Studies	946	74
40	Univ N Carolina; Kenan Flagler Business Sch	926	13

Figure 4. Most intense citations between "LIS" departments.



The citations shown in Table 2 are based on the citation graph of departments, which was generated from the core 46,800 publication records retrieved. Citations from publications outside of this core are not counted in, but they are included in the "Times Cited" indicator which is present in each publication record retrieved (TC in Figure 1). The ranking of departments by times cited looks different than that in Table 2 and the top departments are presented in Table 3. The best three departments are "Univ Minnesota; Carlson Sch Management", "Harvard Univ; Sch Med", and "Univ Maryland; Robert H Smith Sch Business". Again, departments with fewer publications often have higher times cited counts. An extreme case is "Univ So Calif; Knowledge Syst Lab" with one publication only and the largest times cited in Table 3. Note that the times cited count is not always greater than or equal to citations because both indicators are based on different citation graphs—the citation graph of articles and the citation graph of departments, respectively. Imagine a department affiliated with one article only that is merely cited once from an article with which three distinct departments are affiliated. In that case the cited department's times cited count is 1 and its citations indicator is 3. Thus the ranks of individual departments in both rankings can differ significantly. For example, "Univ So Calif; Knowledge Syst Lab" is ranked 10th by times cited but 396th by citations or "Lib Hungarian Acad Sci; Bibliometr Serv" is 33th by citations but 155th by times cited. Anyway, the interpretation may be that "Univ So Calif; Knowledge Syst Lab" is relatively more cited by researchers from other scientific fields than from the community of library and information science whereas "Lib Hungarian Acad Sci; Bibliometr Serv" is relatively more cited from within the community than from outside of it. There is also one highly ranked "department" by times cited, namely "The Scientist; 3600 Market St", which is wrongfully identified as such from frequent addresses associated with "The Scientist" journal articles in WoS data and

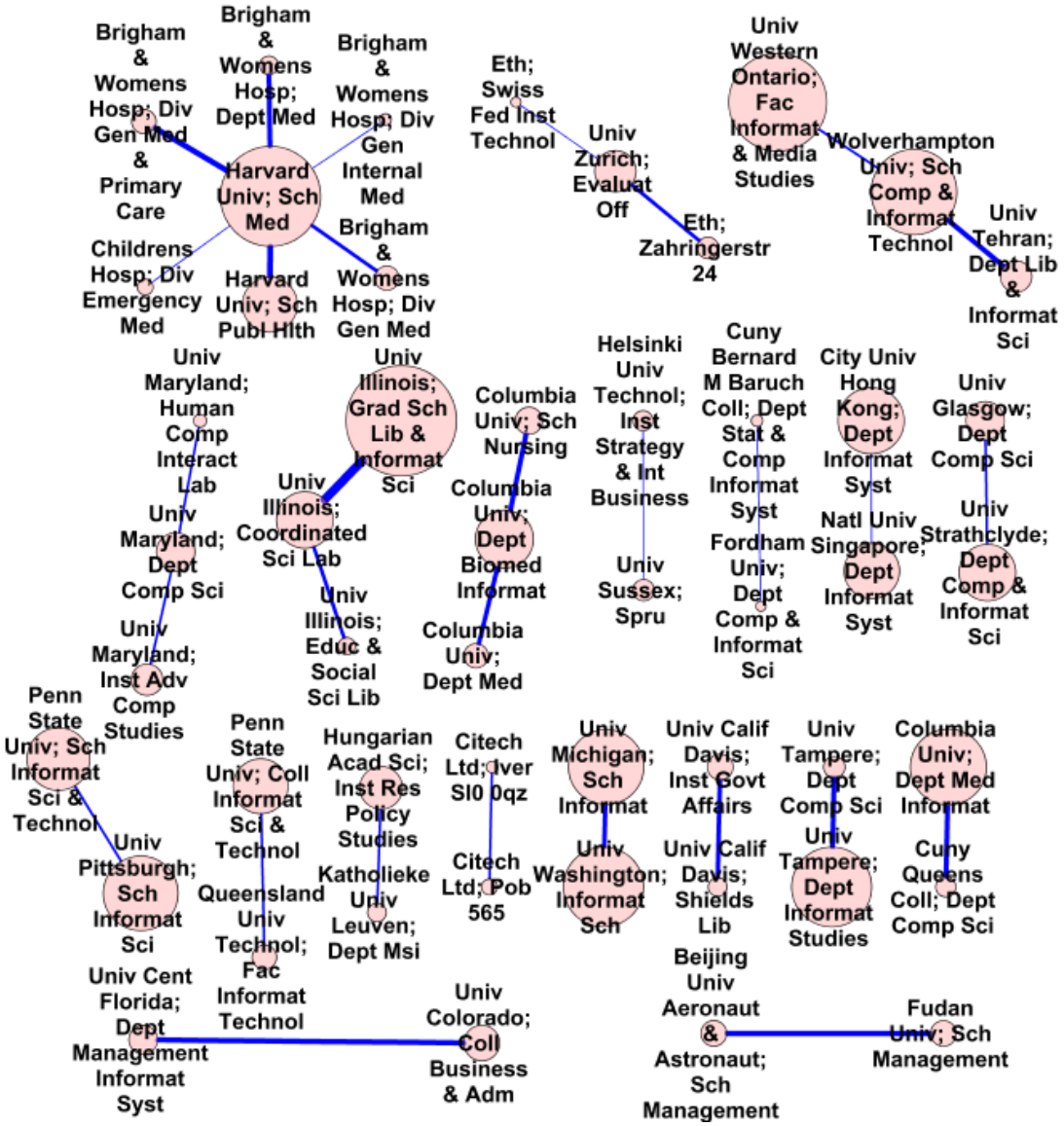
which is ranked very low by citations. Nevertheless, the correlation between the department rankings by citations and by times cited is still rather high as will be shown later on. By the way, many of the present departments are not genuine LIS departments, but are affiliations of authors publishing in journals categorized as ISLS by WoS showing the multidisciplinarity of this field. On the other hand, some LIS research is also published in other WoS categories not covered by this study.

Table 3. Top 40 "LIS" departments by times cited.

	Department	Times Cited	Publications
1	Univ Minnesota; Carlson Sch Management	4756	71
2	Harvard Univ; Sch Med	4051	143
3	Univ Maryland; Robert H Smith Sch Business	3860	52
4	Indiana Univ; Sch Lib & Informat Sci	3475	243
5	Queens Univ; Sch Business	3070	24
6	Rutgers State Univ; Sch Commun Informat & Lib Studies	2950	144
7	Univ Virginia; McIntire Sch Commerce	2942	18
8	The Scientist; 3600 Market St	2922	569
9	Univ Sheffield; Dept Informat Studies	2761	195
10	Univ So Calif; Knowledge Syst Lab	2696	1
11	Leiden Univ; Ctr Sci & Technol Studies	2673	84
12	Univ Arkansas; Sam M Walton Coll Business	2169	11
13	Univ British Columbia; Fac Commerce & Business Adm	2167	26
14	Univ Georgia; Terry Coll Business	2022	38
15	Penn State Univ; Sch Informat Sci & Technol	2017	56
16	Florida State Univ; Coll Business	2008	36
17	Georgia State Univ; Coll Business Adm	1967	24
18	Harvard Univ; Sch Publ Hlth	1707	38
19	Univ Illinois; Grad Sch Lib & Informat Sci	1669	167
20	Wolverhampton Univ; Sch Comp & Informat Technol	1669	109
21	Univ Tampere; Dept Informat Studies	1621	96
22	City Univ London; Dept Informat Sci	1580	192
23	Drexel Univ; Coll Informat Sci & Technol	1535	101
24	Univ Tennessee; Sch Informat Sci	1488	118
25	City Univ Hong Kong; Dept Informat Syst	1459	64
26	Univ So Calif; Marshall Sch Business	1446	28
27	Georgia State Univ; Robinson Coll Business	1421	25
28	Univ Calif Irvine; Grad Sch Management	1385	25
29	Univ Western Ontario; Fac Informat & Media Studies	1332	138
30	Univ S Florida; Coll Business Adm	1233	17
31	Univ N Carolina; Sch Lib & Informat Sci	1180	102
32	Syracuse Univ; Sch Informat Studies	1178	162
33	Stanford Univ; Sch Med	1162	76
34	Univ Penn; Wharton Sch	1141	49
35	Georgia State Univ; Dept Comp Informat Syst	1076	27
36	Brigham & Womens Hosp; Div Gen Med & Primary Care	1074	16
37	Univ N Carolina; Kenan Flagler Business Sch	1064	13
38	McGill Univ; Fac Management	1063	20
39	Univ Western Ontario; Sch Business Adm	1056	2
40	Carnegie Mellon Univ; Grad Sch Ind Adm	1046	15

We did not make an attempt to disambiguate and/or unify the names of institutions and suborganizations, but we used them as they were in WoS data. Instead, we tried to estimate the share of possible duplicate departments. The easiest way to do so was to calculate the similarities of all department names in three random samples of 500 departments using a well known algorithm and then manually check the department pairs whose similarity reached a certain threshold. The determined share of duplicate departments was always below 1%. Thus, we believe that the absence of name disambiguation and unification (which is a very time-consuming task) does not significantly affect the results of this study.

Figure 5. Most intense collaborations between "LIS" departments.



Apart from citations, we can also inspect collaboration patterns. The most intense collaborations between departments are depicted in Figure 5, where the node size depends on the publication count of a department and the edge thickness depends on the number of collaborations. The three most intense collaborations occur between "Univ Illinois; Coordinated Sci

Lab" and "Univ Illinois; Grad Sch Lib & Informat Sci" (an intra-institutional collaboration), "Brigham & Womens Hosp; Div Gen Med & Primary Care" and "Harvard Univ; Sch Med", and "Harvard Univ; Sch Med" and "Harvard Univ; Sch Publ Hlth" (also an intra-institutional collaboration). "Harvard Univ; Sch Med" is the "centre" of the biggest community in Figure 5 collaborating with four "Brigham & Womens Hosp" departments, with another "Harvard Univ" department, and with "Childrens Hosp; Div Emergency Med". The share of intra-institutional interactions is substantially greater with collaborations than with citations—we found that almost 22% of all 22,569 collaborations were intra-institutional. As for the strength of the relationship between citations and collaborations, it does not seem meaningful to draw any conclusions from our data since only about 6% of collaborations occurred more than once and only about 1.5% of citations occurred more than ten times.

In addition to the rankings by citations or times cited, we created also other rankings of "LIS" departments based on other indicators: *Publications* (by the number of publications), *Indegree* (like citations but with all weights in the citation graph of departments set to 1), *AvgTimesCited* (average times cited per publication), *HindexByTimesCited* (h-index as defined by Hirsch [15] and based on times cited), *HindexByEdges* (based on citations within the graph), *HITS* [16], *PageRank* [17], and *Weighted PageRank* [18]. From these other eight rankings we only show the top 40 departments by PageRank and weighted PageRank in Table 4 and Spearman's rank correlations between all the rankings in Table 5 (all significant at the 0.01 leveltwo-tailed).

The PageRank and weighted PageRank rankings are the most highly correlated rankings of all with a rank correlation coefficient of 0.996 and also the first difference in the rankings is at rank 5, where there is "Haifa Univ; Dept Geog" by PageRank and "Univ Minnesota; Carlson Sch Management" by the weighted PageRank. Otherwise, the rankings in Table 4 are quite similar to each other but less so to the ranking by citations (correlation about 0.83) and even less to the ranking by times cited (around 0.69). PageRank-like algorithms (and also HITS) are iterative recursive methods dependent on the structure of the citation graph of departments and, therefore, they are much more related to citations than to times cited. Although the top departments shown in Table 4 do not resemble those in Tables 2 and 3, the overall rankings are still quite strongly correlated with all other rankings except *Publications*. The least correlation we found between *Publications* and *AvgTimesCited*—only about 0.2 *Publications* is also the most distant ranking from all others with an average correlation of 0.483.

Finally, to conclude the section on results, in Table 6 we present examples of the most influential departments (by times cited) of four leading universities having the greatest times cited counts in our LIS data set. These universities are "Univ Maryland", "Indiana Univ", "Georgia State Univ", and "Univ Minnesota". We can notice that there are basically two types of performance distribution at institutions—either there is one dominant department like "Carlson Sch Management" at "Univ Minnesota" or "Robert H Smith Sch Business" at "Univ Maryland" or, to a lesser extent, "Sch Lib & Informat Sci" at "Indiana Univ", or there are several comparably well performing departments like "Coll Business Adm", "Robinson Coll Business", and "Dept Comp Informat Syst" at "Georgia State Univ". Even if this example is

small, we can assume that all influential institutions whose research influence is investigated at the level of departments can fit into one of these two basic performance distribution schemes.

Table 4. Top 40 "LIS" departments by PageRank and weighted PageRank.

	PageRank	**Weighted PageRank**
1	Inst Studies Res & Higher Educ; Munthes Gt 29	Inst Studies Res & Higher Educ; Munthes Gt 29
2	Norwegian Radium Hosp; Inst Canc Res	Norwegian Radium Hosp; Inst Canc Res
3	Univ Missouri; Med Informat Grp	Univ Missouri; Med Informat Grp
4	Univ Missouri; Program Hlth Serv Management	Univ Missouri; Program Hlth Serv Management
5	Haifa Univ; Dept Geog	Univ Minnesota; Carlson Sch Management
6	Univ Maryland; Dept Geog	Indiana Univ; Sch Lib & Informat Sci
7	Enea; Cr Casaccia	Haifa Univ; Dept Geog
8	Univ Washington; Coll Educ	Univ Hull; Inst European Publ Law
9	Washington State Univ; Edward R Murrow Sch Commun	Univ Hull; Sch Law
10	Cornell Univ; Coll Agr & Life Sci	Rutgers State Univ; Sch Commun Informat & Lib Studies
11	Cornell Univ; Coll Vet Med	Enea; Cr Casaccia
12	Univ Hull; Sch Law	Univ Maryland; Dept Geog
13	Univ Hull; Inst European Publ Law	Univ Washington; Coll Educ
14	Univ Minnesota; Carlson Sch Management	Univ Sheffield; Dept Informat Studies
15	Enea; Res Ctr Casaccia	Cornell Univ; Coll Vet Med
16	Univ Hamburg; Inst Ethnol	Queens Univ; Sch Business
17	Univ Calabria; Ctr Ingn Econ & Sociale	Leiden Univ; Ctr Sci & Technol Studies
18	Enea; Ente Nuove Tecnol Energia Ambiente	Cornell Univ; Coll Agr & Life Sci
19	Indiana Univ; Sch Lib & Informat Sci	Washington State Univ; Edward R Murrow Sch Commun
20	Rutgers State Univ; Sch Commun Informat & Lib Studies	Univ British Columbia; Fac Commerce & Business Adm
21	Queens Univ; Sch Business	Penn State Univ; Sch Informat Sci & Technol
22	Univ Vermont; Sch Business Adm	Univ Illinois; Grad Sch Lib & Informat Sci
23	Univ Sheffield; Dept Informat Studies	Univ Maryland; Robert H Smith Sch Business
24	Univ Virginia; Mcintire Sch Commerce	Harvard Univ; Sch Med
25	Leiden Univ; Ctr Sci & Technol Studies	Enea; Res Ctr Casaccia
26	Univ Maryland; Hlth Sci Lib	Univ Tennessee; Sch Informat Sci
27	Univ Illinois; Grad Sch Lib & Informat Sci	Univ Vermont; Sch Business Adm
28	Univ Michigan; Alfred Taubman Med Lib	Univ Virginia; Mcintire Sch Commerce
29	Univ Texas; Grad Sch Business	Univ Penn; Wharton Sch
30	Harvard Univ; Sch Med	Univ Tampere; Dept Informat Studies
31	Natl & Univ Lib Iceland; Interlib Loans Document Delivery Dept	Univ Calif Irvine; Grad Sch Management
32	Reykjavik Univ; European Documentat Ctr	Univ Maryland; Hlth Sci Lib
33	Georgia State Univ; Coll Business Adm	Georgia State Univ; Coll Business Adm
34	Univ Western Ontario; Sch Business Adm	Univ Georgia; Terry Coll Business
35	Univ Calif Irvine; Grad Sch Management	Carnegie Mellon Univ; Grad Sch Ind Adm
36	Univ British Columbia; Fac Commerce & Business Adm	City Univ London; Dept Informat Sci
37	Syracuse Univ; Sch Informat Studies	Univ Michigan; Alfred Taubman Med Lib
38	Univ Michigan; Head Hlth Sci Lib	Univ N Carolina; Sch Lib & Informat Sci
39	Oregon State Univ; Dept Journalism	Drexel Univ; Coll Informat Sci & Technol
40	Carnegie Mellon Univ; Grad Sch Ind Adm	Syracuse Univ; Sch Informat Studies

Table 5. Spearman's rank correlation coefficients between various rankings.

	Avg Times Cited	Citations	Indegree	Publications	Times Cited	Hindex By Edges	Hindex ByTimes Cited	HITS	PR	PR weighted
Avg TimesCited	1	0.7009	0.7055	**0.2045**	0.9513	0.6944	0.7048	0.6785	0.6358	0.6340
Citations	0.7009	1	**0.9908**	0.4360	0.7641	0.7805	0.6355	0.9604	0.8300	0.8342
Indegree	0.7055	**0.9908**	1	0.4270	0.7653	0.7790	0.6312	0.9623	0.8416	0.8385
Publications	**0.2045**	0.4360	0.4270	1	0.4561	0.4974	0.6126	0.4052	0.3917	0.3981
TimesCited	0.9513	0.7641	0.7653	0.4561	1	0.7765	0.8224	0.7352	0.6918	0.6923
Hindex ByEdges	0.6944	0.7805	0.7790	0.4974	0.7765	1	0.7879	0.7648	0.6881	0.6887
Hindex ByTimes Cited	0.7048	0.6355	0.6312	0.6126	0.8224	0.7879	1	0.6153	0.5978	0.6011
HITS	0.6785	0.9604	0.9623	0.4052	0.7352	0.7648	0.6153	1	0.8020	0.7999
PR	0.6358	0.8300	0.8416	0.3917	0.6918	0.6881	0.5978	0.8020	1	**0.9958**
PR weighted	0.6340	0.8342	0.8385	0.3981	0.6923	0.6887	0.6011	0.7999	**0.9958**	1

Table 6. Top 20 "LIS" departments of four leading universities by times cited.

Univ Maryland		Indiana Univ	
Robert H Smith Sch Business	3860	Sch Lib & Informat Sci	3475
Rh Smith Sch Business	755	Kelley Sch Business	1035
Coll Lib & Informat Serv	597	Sch Med	709
Coll Informat Studies	565	Sch Business	254
Asian Div	480	Dept Telecommun	227
Coll Business & Management	407	Slis	221
Dept Decis & Informat Technol	387	Grad Sch Business	213
Dept Comp Sci	221	Ctr Social Informat	142
Coll Lib & Informat Sci	151	Sch Publ & Environm Affairs	141
Inst Adv Comp Studies	145	Kelly Sch Business	121
Dept Informat Syst	136	Sch Informat	89
Dept Geog	123	Sch Educ	70
Sch Med	100	Regenstrief Inst Hlth Care	62
Human Comp Interact Lab	96	Sch Journalism	51
Amer Use Time Project	72	Dept Geog	44
Joint Program Survey Methodol	69	Inst Commun Res	38
Ctr Comp Sci	65	Dept Instruct Syst Technol	35
College Pk	62	Dept Polit Sci	26
Rh Smith Sch	62	Dept Amer Studies	24
Hlth Sci Lib	58	Roudebush Va Med Ctr	21

Table 6. *Cont.*

Georgia State Univ		Univ Minnesota	
Coll Business Adm	1967	Carlson Sch Management	4756
Robinson Coll Business	1421	Curtis L Carlson Sch Management	609
Dept Comp Informat Syst	1076	Dept Informat & Decis Sci	100
J Mack Robinson Coll Business	697	Sch Journalism & Mass Commun	91
Comp Informat Syst Dept	675	Mis Res Ctr	66
Robinbson Coll Business	220	Dept Geog	46
Dept Management	210	Sch Law	40
Ctr Proc Innovat & Comp Informat Syst	194	Digital Technol Ctr	38
Ctr Proc Innovat	119	Informat & Decis Sci Dept	35
Coll Business	77	Biomed Lib	32
J Mack Robinson Coll Business Adm	45	Dept Psychol	30
Cis Dept	40	E Asian Lib	24
Business Adm	36	Coll Educ & Human Dev	23
Dept Comp Informat Ssyt	36	St Paul Campus Lib	18
Dept Commun	34	Dept Comp Sci & Engn	17
Policy Res Ctr	24	Sch Med	17
Coll Educ	12	Sch Nursing	14
Pullen Lib	11	1445 Gortner Ave	13
William Russell Pullen Lib	11	Sci & Engn Lib	13
Dept Sociol	8	Walter Lib 108	13

4. Conclusions and Future Work

Most large-scale scientometric research at the meso-level is concerned with primary research organizations (institutions), but only few studies analyze the scientific impact and collaboration of the suborganizations of these institutions. These suborganizations can be called schools, departments, divisions, laboratories, *etc*. and they themselves may be divided into further suborganizations of lower levels in the organizational hierarchy of an institution. Varying organizational structures along with ambiguities in the names of suborganizations may be the reason of the lack of large-scale scientometric analyses at the level of departments. This article tries to bridge this gap in the field of library and information science. The main contributions of this study are the following:

- We analyzed the bibliographic records of 46,800 journal articles indexed in the Web of Science category "Information Science & Library Science" that were published between 1991 and 2010.
- We created citation and collaboration networks of level-1 suborganizations that we call departments and we visualized the most intense citations and collaborations between departments.
- We produced various rankings of "LIS" departments using ten well-known methods and computed the correlations between these rankings.

The main findings of our study confirm the sufficiency of WoS data and are as follows:

- Almost 88% of publications had some address information associated with them, but prior to 1998 only few publications had other than reprint addresses included.
- "Indiana Univ; Sch Lib & Informat Sci" is the best department in terms of citations and "Univ Minnesota; Carlson Sch Management" is ranked first by times cited.
- Most cited of all departments is "Indiana Univ; Sch Lib & Informat Sci" by "Wolverhampton Univ; Sch Comp & Informat Technol" and the most intense departmental collaboration occurs between "Univ Illinois; Coordinated Sci Lab" and "Univ Illinois; Grad Sch Lib & Informat Sci".

In our future work on the scientific performance and collaboration at the level of departments, we would like focus on other fields of science, other publication sources (e.g., conference proceedings), and other time periods.

Acknowledgments

This work was supported by the European Regional Development Fund (ERDF), project "NTIS—New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Bradley, S.J.; Willett, P.; Wood, F.E. Publication and citation analysis of the department of information studies, University of Sheffield, 1980–1990. *J. Inf. Sci.* **1992**, *18*, 225–232.
2. Holmes, A.; Oppenheim, C. Use of citation analysis to predict the outcome of the 2001 research assessment exercise for unit of assessment (UoA) 61: Library and information management. *Inf. Res.* **2001**, *6*. Available online: http://InformationR.net/ir/6-2/paper103.html (accessed on 5 October 2013).
3. Oppenheim, C. The correlation between citation counts and the 1992 research assessment exercise ratings for British library and information science university departments. *J. Doc.* **1995**, *51*, 18–27.
4. Seng, L.B.; Willett, P. The citedness of publications by United Kingdom library schools. *J. Inf. Sci.* **1995**, *21*, 68–71.
5. Webber, S. Information science in 2003: A critique. *J. Inf. Sci.* **2003**, *29*, 311–330.
6. Thomas, O.; Willett, P. Webometric analysis of departments of librarianship and information science. *J. Inf. Sci.* **2000**, *26*, 421–428.
7. Arakaki, M.; Willett, P. Webometric analysis of departments of librarianship and information science: A follow-up study. *J. Inf. Sci.* **2009**, *35*, 143–152.

8. Aina, L.O.; Mooko, N.P. Research and publication patterns in library and information science. *Inf. Dev.* **1999**, *15*, 114–119.
9. Herrero-Solana, V.; Ríos-Gómez, C. Producción latinoamericana en biblioteconomía y documentación en el social science citation index (SSCI) 1966–2003. *Inf. Res.* **2006**, *11*, 21–45, (in Spanish).
10. Meho, L.I.; Spurgin, K.M. Ranking the research productivity of library and information science faculty and schools: An evaluation of data sources and research methods. *J. Am. Soc. Inf. Sci. Technol.* **2005**, *56*, 1314–1331.
11. Yazit, N.; Zainab, A.N. Publication productivity of Malaysian authors and institutions in LIS. *Malays. J. Libr. Inf. Sci.* **2007**, *12*, 35–55.
12. Yan, E.; Sugimoto, C.R. Institutional interactions: Exploring social, cognitive, and geographic relationships between institutions as demonstrated through citation networks. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62*, 1498–1514.
13. He, B.; Ding, Y.; Yan, E. Mining patterns of author orders in scientific publications. *J. Informetr.* **2012**, *6*, 359–367.
14. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504.
15. Hirsch, J.E. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569–16572.
16. Kleinberg, J.M. Authoritative sources in a hyperlinked environment. *J. ACM* **1999**, *46*, 604–632.
17. Brin, S.; Page, L. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.***1998**, *30*, 107–117.
18. Fiala, D. Time-aware PageRank for bibliographic networks. *J. Informetr.* **2012**, *6*, 370–388.

Article 6

The motivation for writing the following article was the fact that the digital library CiteSeer (closely studied in Articles 2 and 3) was definitely upgraded to CiteSeer^X (the "next-generation CiteSeer") in April 2010[22] and the new version was announced to be superior to the old one in various aspects. I was primarily interested in the coverage of both CiteSeers' databases (how many articles were indexed) and in the quality of the metadata on these articles. The metadata quality was of great concern to me since as I noted in my earlier articles, CiteSeer's automatically generated metadata on autonomously collected data from mainly computer science papers were prone to errors and this might well have been the reason for the reluctance to include CiteSeer data in scientometric studies. I originally intended to parse CiteSeer^X metadata in the same way as those of CiteSeer, import them into a relational database and carry out the same analysis of the citation network of authors as in Article 3 to detect the most prestigious researchers using various ranking methods. This prestige analysis is expected to yield good indicators of metadata quality in terms of author rankings, which can be compared to the established ways of the recognition of researchers' excellence such as awarding prizes and medals. But after inspecting and comparing the metadata of both CiteSeers, I soon discovered that this approach was impossible here due to a different structure of both metadata sets as shown in Figure 3. While CiteSeer's records (articles) are linked to by elements of type "References", which enables a standard citation network of papers (and authors) to be created, there are no such clear references in CiteSeer^X with merely "relations" being present, which may mean both citations and references, but also other kinds of relationship between papers. If we do not wish to get citation data from the creators of CiteSeer^X per email upon request (which I personally do not consider a publicly accessible way), the only possibility of obtaining a citation graph of authors from the freely available data on the web-

[22] I identified this date by regularly checking CiteSeer's web address citeseer.ist.psu.edu and found out that it was finally merged with citeseerx.ist.psu.edu at about this time.

site was to transform the undirected coauthorship graph by directing its edges appropriately as has been previously described in the literature. The analysis results were interesting anyway and supported the hypothesis that both the coverage and quality of CiteSeer^X improved.

```
<record>
<header>
<identifier>oai:CiteSeerPSU:2</identifier>
<datestamp>1997-11-01</datestamp>
<setSpec>CiteSeerPSUset</setSpec>
</header>
<metadata>
<oai_citeseer:oai_citeseer xmlns:oai_citeseer="http://copper.ist.psu.edu/oai/oai_citeseer/" xmlns:dc
   <dc:title>The Graham Scan Triangulates Simple Polygons</dc:title>
   <oai_citeseer:author name="Xianshu Kong"></oai_citeseer:author>
   <oai_citeseer:author name="Hazel Everett"></oai_citeseer:author>
   <oai_citeseer:author name="Godfried Toussaint"></oai_citeseer:author>
   <dc:subject>Xianshu Kong,Hazel Everett,Godfried Toussaint The Graham Scan Triangulates Simple Pol
   <dc:description>The Graham scan is a fundamental backtracking technique in computational geometry
   <dc:contributor>The Pennsylvania State University CiteSeer Archives</dc:contributor>
   <dc:publisher>unknown</dc:publisher>
   <dc:date>1997-11-01</dc:date>
   <oai_citeseer:pubyear>1991</oai_citeseer:pubyear>
   <dc:format>ps</dc:format>
   <dc:identifier>http://citeseer.ist.psu.edu/2.html</dc:identifier>
   <dc:source>http://www-cgrl.cs.mcgill.ca/~godfried/publications/tri.scan.ps.gz</dc:source>
   <dc:language>en</dc:language>
   <oai_citeseer:relation type="References">
      <oai_citeseer:uri>oai:CiteSeerPSU:97473</oai_citeseer:uri>
   </oai_citeseer:relation>
   <oai_citeseer:relation type="References">
      <oai_citeseer:uri>oai:CiteSeerPSU:154288</oai_citeseer:uri>
   </oai_citeseer:relation>
   <dc:rights>unrestricted</dc:rights>
</oai_citeseer:oai_citeseer>
</metadata>
</record>
```

```
<record>
  <header>
    <identifier>oai:CiteSeerXPSU:10.1.1.1.1485</identifier>
    <datestamp>2009-05-24</datestamp>
  </header>
  <metadata>
    <oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:oai_dc="http://www.openarchives.org/
      <dc:title>DEEM: a Tool for the Dependability Modeling and Evaluation</dc:title>
      <dc:creator>A. Bondavalli</dc:creator>
      <dc:creator>I. Mura</dc:creator>
      <dc:creator>S. Chiaradonna</dc:creator>
      <dc:creator>S. Poli</dc:creator>
      <dc:creator>F. Sandrini</dc:creator>
      <dc:subject>Processes</dc:subject>
      <dc:description>Multiple-Phased Systems, whose operational life can be partitioned in a set of
      <dc:contributor>CiteSeerX</dc:contributor>
      <dc:publisher>IEEE Computer Society</dc:publisher>
      <dc:date>2009-05-24</dc:date>
      <dc:date>2007-11-19</dc:date>
      <dc:date>2000</dc:date>
      <dc:format>application/pdf</dc:format>
      <dc:type>text</dc:type>
      <dc:identifier>http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.1485</dc:identifier>
      <dc:source>http://bonda.cnuce.cnr.it/Documentation/Papers/file-BMCFPS00-DSN2000-76.pdf</dc:sour
      <dc:language>en</dc:language>
      <dc:relation>10.1.1.47.2594</dc:relation>
      <dc:relation>10.1.1.58.2039</dc:relation>
      <dc:rights>Metadata may be used without restrictions as long as the oai identifier remains atta
    </oai_dc:dc>
  </metadata>
</record>
```

Figure 3: Sample metadata records in CiteSeer (top) and CiteSeer^X (bottom)

Preprint of: Fiala, D. (2013). From CiteSeer to CiteSeer^X: Author rankings based on coauthorship networks. *Journal of Theoretical and Applied Information Technology*, 58(1), 191-204.

From CiteSeer to CiteSeer^X: Author rankings based on coauthorship networks

Dalibor Fiala

University of West Bohemia,

Department of Computer Science and Engineering, Czech Republic

E-mail: dalfia@kiv.zcu.cz

ABSTRACT

CiteSeer was a digital library and a search engine gathering its mainly computer science research papers from the World Wide Web. After a few years of stagnation, it was definitely replaced with a new version called CiteSeer^X in April 2010. As both CiteSeers provide(d) freely available metadata on the articles they index(ed), it is possible to analyze two different data sets to see the differences between CiteSeer and CiteSeer^X. More specifically, we examined the article metadata from CiteSeer (downloaded in December 2005) and from CiteSeer^X (harvested in March 2011) with a view of creating rankings of prestigious computer scientists. Since the free article metadata acquired from the Web site of CiteSeer^X differ from those in CiteSeer in that they do not systematically include cited references, the only possibility of creating such rankings is to base them on the coauthorship networks in both CiteSeers. In this study, we produce these rankings using 12 different ranking methods including PageRank and its variants, compare them with the lists of ACM A. M. Turing Award and ACM SIGMOD E. F. Codd Innovations Award winners and conclude that the rankings generated from CiteSeer^X data outperform those from CiteSeer.

Keywords: *CiteSeer, CiteSeer^X, Coauthorships, Citations, Researchers, PageRank*

1. INTRODUCTION AND RELATED WORK

CiteSeer [1] was a digital library and a search engine specialized mainly in computer science literature that gathered its content by autonomously crawling the World Wide Web and downloading and parsing potentially relevant documents [2]. After some time of running in parallel with a new version, finally, in April 2010, the "old" CiteSeer officially ceased to exist and was replaced by the new CiteSeer^X [3], which is, however, still in a beta version at the time of writing this paper (May 2013). In fact, the old URL redirects to the new one now. Anyway, in the last years of its existence, CiteSeer was no more updated. On the other hand, CiteSeer^X has been continuously updated since its creation until now. Although there have been enough studies based on CiteSeer data, some of which will be mentioned in the related work section, research dealing with CiteSeer^X has been somewhat rare so far, probably partly due to the relative novelty and presumed immaturity of CiteSeer^X. Also, even though the nature of CiteSeer data invites bibliometric analyses, there have been few of them, perhaps as a result of the presence of errors in the data that have been created using automated text pro-

cessing tools. In spite of this, some papers have reported a successful usage of CiteSeer data for bibliometric purposes (see more on this in the following paragraphs).

This study tries to analyze the freely available article metadata of CiteSeer and CiteSeer^X (obtainable from their respective Web sites) and to answer the following main research questions: a) What is the structure of these article metadata of CiteSeer and CiteSeer^X and what are the basic characteristics of the coauthorship networks generated from them? b) Can the coauthorship networks of CiteSeer and CiteSeer^X be used to rank computer scientists? c) And, if yes, which CiteSeer generates better rankings if they are compared to the lists of prestigious computer science award winners (ACM A. M. Turing Award and ACM SIGMOD E. F. Codd Innovations Award)?

Numerous studies have explored CiteSeer or CiteSeer^X data for non-bibliometric purposes, mainly to test various graph-theoretic approaches. An et al. [4] analyzed the citation graph of CiteSeer (then called ResearchIndex) in terms of connectivity. Chakrabarti and Agarwal [5] made use of CiteSeer citation data to test their unified ranking model on real-world graphs. Chakrabarti et al. [6] utilized the CiteSeer corpus and query logs to test new techniques of personalized PageRank computation on entity-relation graphs. Hopcroft et al. [7] tracked evolving communities of computer science research papers by exploring the CiteSeer citation graph from 1998 and 2001. Joorabchi and Mahdi [8] used CiteSeer documents to evaluate the performance of their automatic classification of research papers according to a standard library classification scheme. Popescul et al. [9] employed CiteSeer data to train and test their new classifier that categorized research papers into publication venues. Šingliar and Hauskrecht [10] performed a component analysis of a partial CiteSeer citation graph. Zhou et al. [11] used thousands of CiteSeer documents in the construction of a real-world network to test their graph partitioning algorithm for the discovery of temporal communities of computer science researchers. Chen et al. [12] proposed a system based on the coauthorship network of CiteSeer^X to recommend potential collaborators. He et al. [13] designed a recommender system suggesting cited references for a given article based on the many citation contexts available in CiteSeer^X. Abstracts from CiteSeer^X documents were employed in the construction of hierarchical topic-based communities of authors by Wu and Koh [14].

Fewer studies have been bibliometric. CiteSeer was used as one of the data sources providing citation data for the citation analysis of the works of a famous mathematician by Bar-Ilan [15]. Feitelson and Yovel [16] took advantage of CiteSeer's citation counts of highly cited researchers in their predictive model of future citation-based ranks of researchers. Giles

and Councill [17] investigated acknowledgements in the papers of the CiteSeer archive including its citation graph and determined the most acknowledged entities as well as their citation counts. Goodrum et al. [18] analyzed the most cited documents in the CiteSeer database and found out their publication type and age, among others. Zhao [19] explored the CiteSeer citation graph in the XML research field and identified highly productive and influential scientists. Zhao and Logan [20] carried out a similar study and concluded that citation analysis based on CiteSeer (at least in the XML domain) is as valid as that based on established data sources. And, finally, Zhao and Strotmann [21], again in the XML research field, conducted an author co-citation analysis of CiteSeer documents and compared the results with an analysis based on ISI Science Citation Index. Krumov et al. [22] constructed a coauthorship network from CiteSeer^X data and examined the relation of coauthorship patterns to the impact of scientific publications.

Unlike our research, most of the above studies have not dealt with the CiteSeer citation or coauthorship graph as a whole – they have been mostly concerned with a part of it only. Furthermore, none of them has analyzed CiteSeer as well as CiteSeer^X at the same time. In this context, this study is unique in that it examines the whole coauthorship graphs of both CiteSeers. It is an extension to our previous work, in which a citation analysis of the whole CiteSeer citation graph with a view of identifying prominent computer scientists was carried out [23] and a bibliometric analysis of all CiteSeer metadata aimed at finding the most productive and influential countries in computer science was conducted [24]. The usefulness of coauthorships in the assessment of researchers was shown by Yan and Ding [25] who determined the impact of authors in the informetrics research community by applying the PageRank algorithm to a coauthorship network. For the evaluation of the author rankings resulting from our analyses, we use the same technique (comparing the rankings with the lists of computer science award winners) as in other studies [23, 26-28].

2. DATA AND METHODS

In the present study, we examined two data sets – CiteSeer and CiteSeer^X. Because CiteSeer was no more updated in the last years of its existence, the most recent data file that we could obtain was from December 2005. On the other hand, CiteSeer^X has been continuously updated since its creation until now and we took a snapshot if its metadata in March 2011. Thus, there is a roughly six-year age difference in the two data files, the analysis of which we present in this study. We downloaded CiteSeer metadata straight from its Web site

as an archive file and we harvested CiteSeer^X metadata from its Open Archives Initiative collection [29]. The freely available metadata for each article in CiteSeer generally include its title, abstract, authors, authors' addresses and affiliations, source URL, document format and language, cited references, and publication year and download date. However, addresses and affiliations, references, and publication years are often missing, incomplete, or erroneous. On the other hand, the article metadata harvested from CiteSeer^X include information on the document publisher, but addresses and affiliations are entirely absent and references (or citations) do not appear systematically.

In total, there were 716768 "core" (i.e., with article full texts) publication records in CiteSeer and 1334000 "core" publication records in CiteSeer^X. Thus, the number of records almost doubled between 2005 and 2011. As complete citations between publications are not available in the CiteSeer^X metadata we had (unlike CiteSeer), the only possibility of constructing comparable author citation graphs from both CiteSeers is to base them on the coauthorship networks (similarly to Yan and Ding, 2011) that can be easily built from both metadata sets. From a coauthorship (or collaboration) network with publications and their respective authors, we can obtain a graph of authors, in which every two coauthors of a publication are connected with an undirected edge. To avoid parallel edges in the case of many publications being written by the same coauthors, the edge will be assigned a weight denoting the number of joint publications. Next, each undirected edge is replaced with two oppositely directed edges both retaining the original weight. As a result, a citation graph of authors based on the collaboration network has been created. The basic statistics of such author citation graphs generated from the article metadata of CiteSeer and CiteSeer^X can be seen in figure 1.

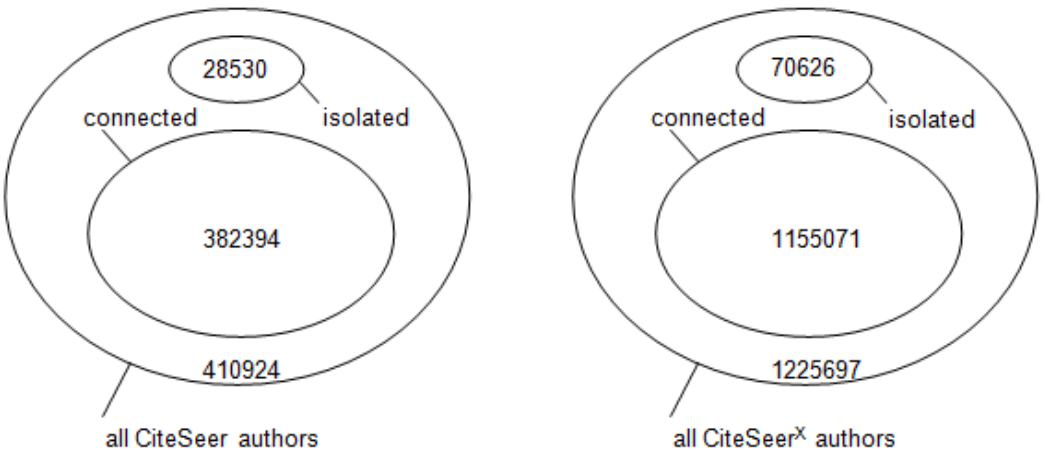


Figure 1: Basic Statistics of the Coauthorship Graphs in Both CiteSeers

Without disambiguation or duplicates removal, we found a total of 1 663 044 author records in CiteSeer and 3837226 in CiteSeer^X (not visible in figure 1). After transforming author names into upper case, we identified 410924 "distinct" authors in CiteSeer and 1225697 "distinct" authors in CiteSeer^X. These are the actual numbers of nodes in the author citation graphs. We must underline that name unification and disambiguation is a very tedious and time-consuming task and is not the concern of this research. We examine the data from CiteSeer "as is", without any pre- or postprocessing and this may have influence on the rather high per-author citation counts below. Prior to the elimination of parallel edges in the author citation graphs, there were 4764960 citations (formerly collaborations) between authors in CiteSeer (11.6 per author) and 16023138 in CiteSeer^X (13.1 per author) excluding self-citations of all authors. After eliminating the parallel edges, there were 2466446 and 9607486 edges left, which were assigned weights as described above. As for the authors, their number tripled between 2005 and 2011, but the percentage of isolated authors remained almost the same (7% and 6%, respectively) compared to the total number of authors. "Connected authors" are those who cite or are cited, which is equivalent here, because the citation graph is based on symmetric collaborations. Finally, we can conclude that the linkage density of the CiteSeer coauthorship graphs did not change between 2005 and 2011.

To analyze the citation graphs, we decided to apply the same 12 ranking methods used also by Fiala [23], which were described in detail in another paper [27]. In this section, we will briefly summarize the rationale of these methods. In the citation analysis, we can basically choose from simple (first-order, non-recursive) methods such as citation counts (in fact, a "weighted" in-degree) or in-degree ("unweighted") or from more complicated (higher-order, recursive) methods such as HITS [30] or the notoriously known PageRank [31], which were originally conceived for the World Wide Web but later also applied to other network types such as author citation networks to identify influential actors. The "standard" PageRank (*PR*, by Brin and Page) can be modified so as to better reflect the features of bibliographic networks. For instance, the formerly unweighted edges can be assigned weights that denote the number of citations between two authors and thus give rise to a "weighted PageRank" (*PR-W*). The weighted PageRank formula can be further extended with some additional information such as the number of collaborations (*PR-C*), publications (PR-P), all coauthors (*PR-AC*), all distinct coauthors ((*PR-ADC*), all collaborations (*PR-AColl*), coauthors (*PR-CA*), or distinct coauthors (*PR-DCA*) that can all have influence on the weight of the directed edge between two authors. Thus, we get 12 ranking methods in total (*Cites*, *InDeg*, *HITS*, *PR*, *PR-*

W, *PR-C*, *PR-P*, *PR-AC*, *PR-ADC*, *PR-AColl*, *PR-CA*, and *PR-DCA*), all of which will be used in our analysis. (For all the PageRank-like methods, we used a damping factor *d* of 0.9, a Spearman correlation-based convergence criterion and a maximum of 50 iterations.)

3. RESULTS AND DISCUSSION

We were interested in the changes that occurred in the CiteSeer data from 2005 to 2011. First, we had a look at the distribution of publications based on the number of their authors. Figure 2 shows such a histogram. There we can observe some similarities and discrepancies between the two CiteSeers. For instance, both digital libraries have a significant amount of publications with no authors and this amount remains relatively the same. The cause of this may be the inability of the underlying algorithms to correctly identify author names. From this point of view, the parsing quality does not seem to improve over the years. The most frequent number of authors per paper is two in both cases, but there is a difference in the second most frequent number – this is one author in CiteSeer but three authors in CiteSeer^X. There may be several reasons for this phenomenon including the general increase in the average number of authors per paper in computer science between 2005 and 2011 or the concentration of CiteSeer^X on a specific subfield of computer science with a higher number of authors. However, finding a precise explanation was not the aim of this study.

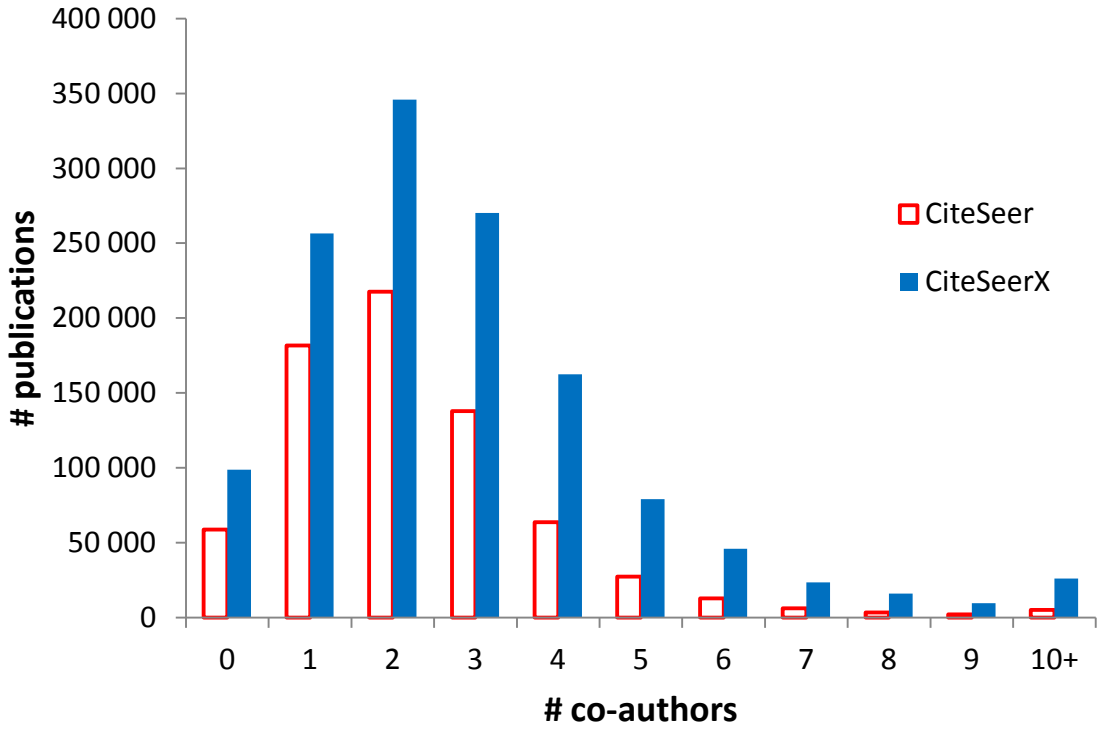


Figure 2: Coauthor Distribution of Publications in Both CiteSeers

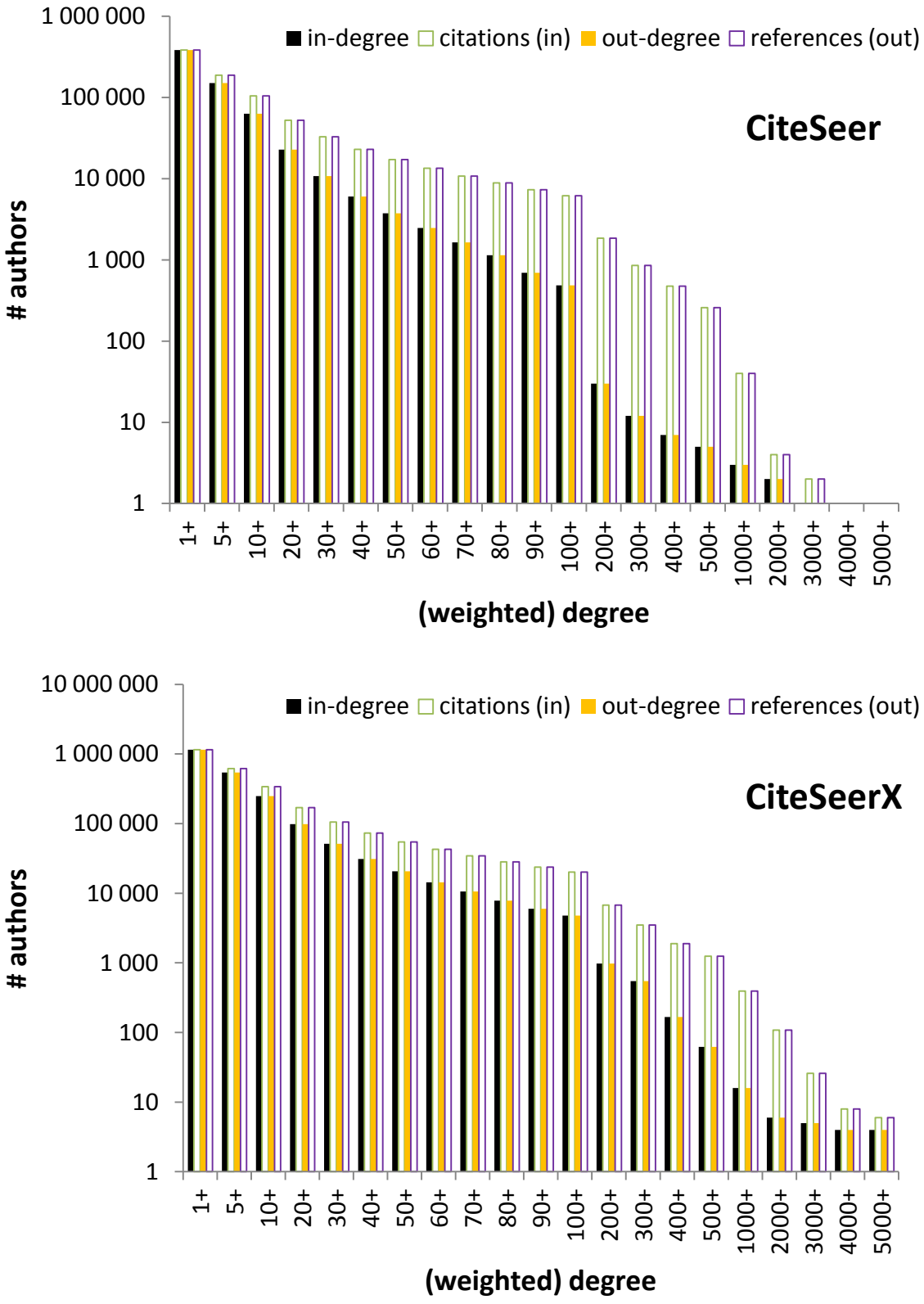


Figure 3: Distribution of Authors by (Weighted) In- and Out-degree in Both CiteSeers

As far as the "density" of the graph of citations between publications is concerned, a great deal is revealed from the cumulative histogram charts in figure 3. The bars represent authors (i.e., graph nodes) with a specific magnitude of (weighted) in-degree or (weighted) out-degree. All the indicators are always larger in CiteSeer^X due to the overall greater number of nodes and edges in the graph. We call the weighted in-degree "citations" and the weighted out-degree "references". Evidently, for a weighted degree, the weights of in-coming (or out-going) edges are summed up. Since the directed graphs under study are based on symmetric collaborations, in-degrees and weighted in-degrees are equal and so are out-degrees and weighted out-degrees. The charts use a logarithmic Y-axis scale to better display bars in their tales. Thus, for instance, some 0.13% of authors have an in-degree of 100 or more in CiteSeer, whereas it is 0.41% in CiteSeer^X. Also, CiteSeer^X includes some authors that have more than 5000 citations, but CiteSeer does not. What authors are the most cited in both CiteSeers is shown in table 1.

Table 1 presents the top 40 authors by citations and in-degree in CiteSeer and CiteSeer^X. (Names in italics cannot be printed in full due to space limitations.) As we can see, there is a lot of noise in the results due to errors in the metadata. As a consequence, the most cited "researchers" turn out to be "Senior Member", "Student Member", or "Ph. D" in both CiteSeers, which are the words frequently occurring close to proper names on papers' title pages that were incorrectly parsed and classified as such. Nevertheless, some well known computer science researchers' names (such as "Jack Dongarra" or "Ian Foster") appear in the top 40 results from CiteSeer. In CiteSeer^X, less known scientists are in the top results, e.g. "R. R. Barton". An interesting extension to table 1 is table 2, in which the top 40 authors determined by three other methods (HITS, PageRank, and weighted PageRank) are presented. The HITS ranking differs the most from the others – it contains no noise and its researchers are mostly unknown. On the other hand, the PageRank and weighted PageRank rankings are noisy and include well known as well as little known computer science authors such as "Jack Dongarra", "Ian Foster", "Takeo Kanade", "R. R. Barton", or "Vladik Kreinovich".

As it is impossible to show all the 12 rankings in full, we focused our attention to two sets of researchers whose ranks generated by all the methods are visualized in the charts in figure 4 and in figure 5. In the first set, there are ACM A. M. Turing Award ("Nobel Prize" in computer science) winners from the years 1991 - 2010. In the second, there are ACM SIGMOD E. F. Codd Innovations Award winners ("Nobel Prize" in databases) from 1992 to 2011. The time spans for both prizes were selected as the last 20 available years at the time of

our experiments. All the charts are displayed on the logarithmic scale and lower ranks mean better ranks (e.g. a rank of 10 is better than a rank of 100). By looking at the charts, we can immediately see a striking feature in all of them – the award winners generally receive bad ranks by HITS. This is supported by the fact we observed in table 2 – no well known researchers were placed at the top by HITS. Another clearly visible property of all the charts is the very good performance of simple citation counts (*Cites*). In principle, the award winners achieve good ranks by citation counts and, therefore, citations can be considered a "good" ranking in contrast to the much more computationally expensive HITS. And finally, PageRank (*PR*), itself also a computationally expensive method, performs comparably to citations but better than HITS and some of its variants are of the same quality or even slightly better than the standard PageRank (most notably *PR-W* for Codd Award winners in CiteSeer^X, see the lower chart in figure 5). All the three findings are in accordance with those reported by Fiala [23] on the normal author citation graph of CiteSeer. As for the individual scientists, the best ranked Turing Award winners (according to their median rank) are "Pnueli" and "Rivest" in CiteSeer and "Gray" and "Rivest" in CiteSeer^X and the best ranked Codd Award winners (according to their median rank) are "Garcia-Molina" and "Stonebraker" in CiteSeer and "Garcia-Molina" and "Widom" in CiteSeer^X. (Awardees whose names were absent in the data are missing in the charts. These are "Selinger" for the Codd Award in CiteSeer, "Feigenbaum", "Yao", "Nygaard", "Naur", and "Allen" for the Turing Award in CiteSeer and "Allen" for the Turing Award in CiteSeer^X.)

Table 1: Top 40 Authors by Citations and In-degree in CiteSeer (CS) and CiteSeer^X (CS^X)

CS	Citations	CS^X		CS	In-degree	CS^X	
Senior Member	4390	Ph. D	28641	Senior Member	2570	Ph. D	10811
Student Member	3676	Senior Member	23136	Student Member	2185	Senior Member	10305
Fachbereich Informatik	2515	Prof Dr	21032	Ph. D	1795	Student Member	7771
Ph. D	2513	Student Member	17173	Fachbereich Informatik	823	Prof Dr	7114
Michael H. Bohlen	1898	Email Alerting	6105	Prof Dr	780	Email Alerting	3843
Kristian Torp	1895	J Neurophysiol	5128	Mathematisch Centrum	481	Jr.	2876
Christian S. Jensen (Codirector	1883	The Erwin	4960	Copyright Stichting	480	Et Al	1797
Richard T. Snodgrass (Codirector	1883	Jr.	4845	G. W. Evans	393	United States	1686
Heidi Gregersen	1880	H. Wahl	3467	H. B. Nembhard	393	J Neurophysiol	1639
Alex Waibel	1877	R. R. Barton	3397	P. A. Farrington	393	The Erwin	1488
Jack Dongarra	1795	V. Kekelidze	3258	D. T. Sturrock	392	Key Words	1149
Christian S. Jensen	1446	M. Martini	3255	Associate Member	311	*Technische Universität*	1146
Sudha Ram	1410	A. Gonidec	3204	Computer Science	287	*Schrödinger International*	1112
Deborah Estrin	1380	A. Ceccucci	3190	Forest Service	282	Forest Service	1110
Curtis E. Dyreson	1360	L. Gatignon	3180	Key Indicators	282	Computer Science	1054
Dieter Pfoser	1344	*Schrödinger International*	3179	E. Dvorkin (Eds	273	R. R. Barton	1009
Giedrius Slivinskas	1288	A. Gianoli	3079	Ian Foster	267	IEEE Computer Society	959
Renato Busatto	1272	A. Norton	3079	S. Idelsohn	265	Fachbereich Informatik	941
Janne Skyt	1244	W. Bartel	3076	Thme Rseaux Et Systmes	256	Prof Dr. -ing	818
Douglas C. Schmidt	1235	V. Falaleev	3054	Rwth Aachen	253	M. Sc	752
Mathematisch Centrum	1228	W. Kubischta	3051	Ecole Normale	248	Supervisor Prof	731
Copyright Stichting	1227	D. Cundy	3050	Jack Dongarra	248	Editorial Board	728
Hector Garcia-Molina	1166	A. Belousov	3039	Sophia Antipolis	244	Associate Member	698
Sebastian Thrun	1159	G. Bocquet	3039	Arthur C. Smith	239	Ipan Mohanty	673
Michael Stonebraker	1154	P. Hristov	3032	Member IEEE	220	Wildlife Service	664
Bongki Moon	1153	N. Molokanova	3018	P. L. Frabetti	216	Lt Col	663
H. Niemann	1104	F. Petrucci	2997	Alle Rechte Vorbehalten	214	Assoc Prof	662
J. Engler	1075	A. Zinchenko	2996	Vladik Kreinovich	211	Member IEEE	659
Prof Dr	1066	P. Dalpiaz	2996	Sun Microsystems	209	III	657
P. Doll	1052	E. Barrelet	2976	IEEE Computer Society	206	Ulrich H. E. Hansmann	638
D. Heck	1049	V. Boudry	2964	M. Martini	197	Gutachter Prof	631
Ian Foster	1033	P. L. Frabetti	2943	Christian S. Jensen	196	Olav Zimmermann (Editors	626
K. Daumiller	1028	V. Brisson	2940	*Technische Hochschule*	196	Sophia Antipolis	609
G. W. Evans	1024	Et Al	2927	Andrei Shleifer	194	B. Biller	608
H. B. Nembhard	1024	M. Savrié	2909	INRIA Rocquencourt	193	J. A. Joines	604
P. A. Farrington	1024	P. Baranov	2848	A. Ceccucci	192	J. D. Tew	603
D. T. Sturrock	1020	M. Velasco	2824	Mario Gerla	189	J. Shortle	603
K. Bekk	1020	K. Bekk	2820	Politecnico Di Milano	189	M. -h. Hsieh	603
H. Bozdog	1013	H. Bozdog	2790	D. Cundy	188	Principal Investigator	603
Don Towsley	1005	D. Bruncko	2763	Ron Kikinis	188	S. G. Henderson	603

Table 2: Top 40 Authors by HITS, PageRank, and Weighted PR in CiteSeer (CS) and CiteSeer^X (CS^X)

CS	HITS	CS^X	CS	PageRank	CS^X	CS	PageRank (weighted)	CS^X
D. Cundy	H Collaboration		Senior Member	Ph. D		Senior Member	Ph. D	
H. Wahl	A. Belousov		Student Member	Senior Member		Student Member	Senior Member	
A. Ceccucci	V. Boudry		Ph. D	Student Member		Ph. D	Prof Dr	
V. Kekelidze	V. Brisson		Fachbereich Informatik	Prof Dr		Fachbereich Informatik	Student Member	
G. Bocquet	D. Bruncko		Prof Dr	Email Alerting		Prof Dr	Email Alerting	
A. Gianoli	A. Babaev		Mathematisch Centrum	Jr.		Mathematisch Centrum	Jr.	
P. L. Frabetti	G. Buschhorn		Copyright Stichting	The Erwin		Copyright Stichting	The Erwin	
L. Gatignon	W. Bartel		Key Indicators	United States		Jack Dongarra	J Neurophysiol	
N. Doble	E. Barrelet		G. W. Evans	Et Al		G. W. Evans	United States	
A. Gonidec	P. Baranov		H. B. Nembhard	Key Words		H. B. Nembhard	*Schrödinger International*	
B. Gorini	B. Delcourt		P. A. Farrington	*Schrödinger International*		P. A. Farrington	Et Al	
G. Barr	S. Egli		D. T. Sturrock	Computer Science		D. T. Sturrock	Forest Service	
J. Duclos	A. De Roeck		Forest Service	*Technische Universität*		Computer Science	*Technische Universität*	
A. Lacourt	G. Eckerlin		Associate Member	Forest Service		Alex Waibel	R. R. Barton	
D. Schinzel	V. Efremenko		Computer Science	J Neurophysiol		Turku Centre	Fachbereich Informatik	
M. Martini	E. Elsen		Arthur C. Smith	IEEE Computer Society		Vladik Kreinovich	Key Words	
A. Norton	Ch. Berger		Vladik Kreinovich	Fachbereich Informatik		Douglas C. Schmidt	Prof Dr. -ing	
B. Panzer-Steindel	F. Eisele		E. Dvorkin (Eds	R. R. Barton		Forest Service	Computer Science	
Yu. Potrebenikov	G. Cozzika		S. Idelsohn	Supervisor Prof		Key Indicators	Vladik Kreinovich	
A. Lai	J. Cvach		Member IEEE	M. Sc		Don Towsley	Assoc Prof	
W. Kubischta	M. Fleischer		Ecole Normale	Prof Dr. -ing		*Technische Hochschule*	M. Sc	
P. Grafstrom	A. Fedotov		Thme Rseaux Et Systmes	Associate Member		Deborah Estrin	Wildlife Service	
P. Hristov	L. Favart		Rwth Aachen	Member IEEE		E. Dvorkin (Eds	IEEE Computer Society	
A. Zinchenko	J. Ferencei		Key Words	*Schrodinger International*		Ian Foster	J. A. Joines	
H. Taureg	W. Braunschweig		Sophia Antipolis	III		Sebastian Thrun	B. Biller	
G. Tatishvili	G. Franke		Jack Dongarra	Assoc Prof		S. Idelsohn	J. D. Tew	
D. Madigojine	D. Clarke		Anthony M. Santomero	Vladik Kreinovich		Hector Garcia-Molina	J. Shortle	
F. Petrucci	L. Goerlich		M. Asce	Open Access		Mario Gerla	M. -h. Hsieh	
S. Palestini	E. Gabathuler		Ian Foster	Principal Investigator		Takeo Kanade	S. G. Henderson	
P. Dalpiaz	B. Andrieu		Turku Centre	Lt Col		Kang G. Shin	*Schrodinger International*	
M. Lenti	M. Erdmann		Takeo Kanade	Sophia Antipolis		Andrew B. Kahng	Lt Col	
I. Mikulec	G. Flügge		IEEE Computer Society	Gutachter Prof		Rwth Aachen	Jack Dongarra	
M. Savrie	J. Formánek		*Technische Hochschule*	John Wiley		Jason Cong	Supérieure Lyon	
D. Marras	R. Gerhards		Mario Gerla	David		Calton Pu	École Normale	
N. Molokanova	J. Gayler		Marie Curie	Civil Justice		Michael H. Bohlen	Supervisor Prof	
W. Funk	J. Feltesse		INRIA Rocquencourt	E. Onate		Manuela Veloso	Terrence J. Sejnowski	
C. Cheshkov	G. Bernardi		Scientiarum Fennicae	John		David E. Goldberg	Member IEEE	
O. Vossnack	J. Bürger		Politecnico Di Milano	J. A. Joines		Daniel Thalmann	Takeo Kanade	
R. Sacco	S. Burke		Sun Microsystems	Wildlife Service		Kristian Torp	Civil Justice	
V. Falaleev	U. Bassler		Alle Rechte Vorbehalten	Editorial Board		Heidi Gregersen	Ian Foster	

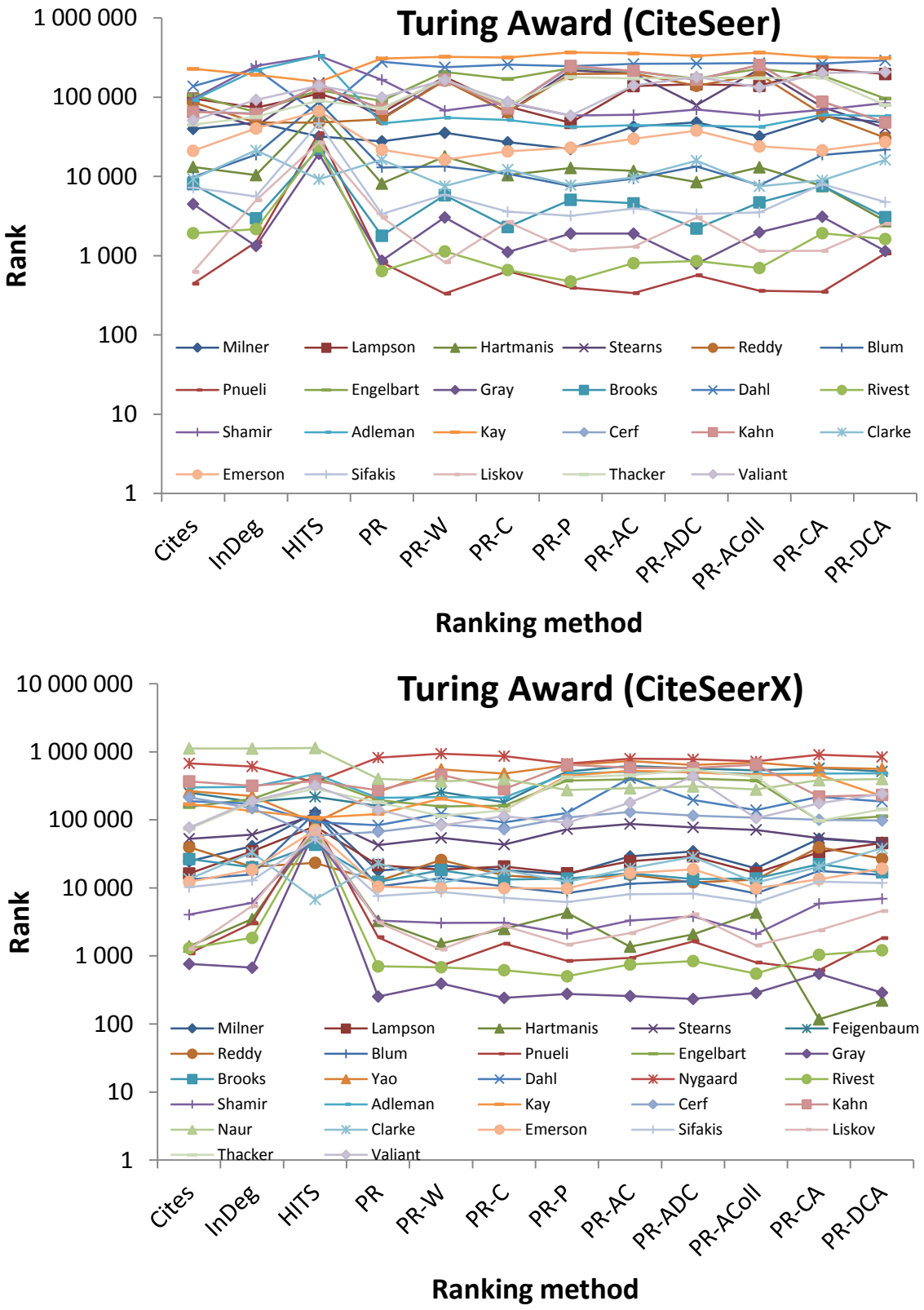


Figure 4: Ranks of Turing Award Winners by Various Methods in Both CiteSeers

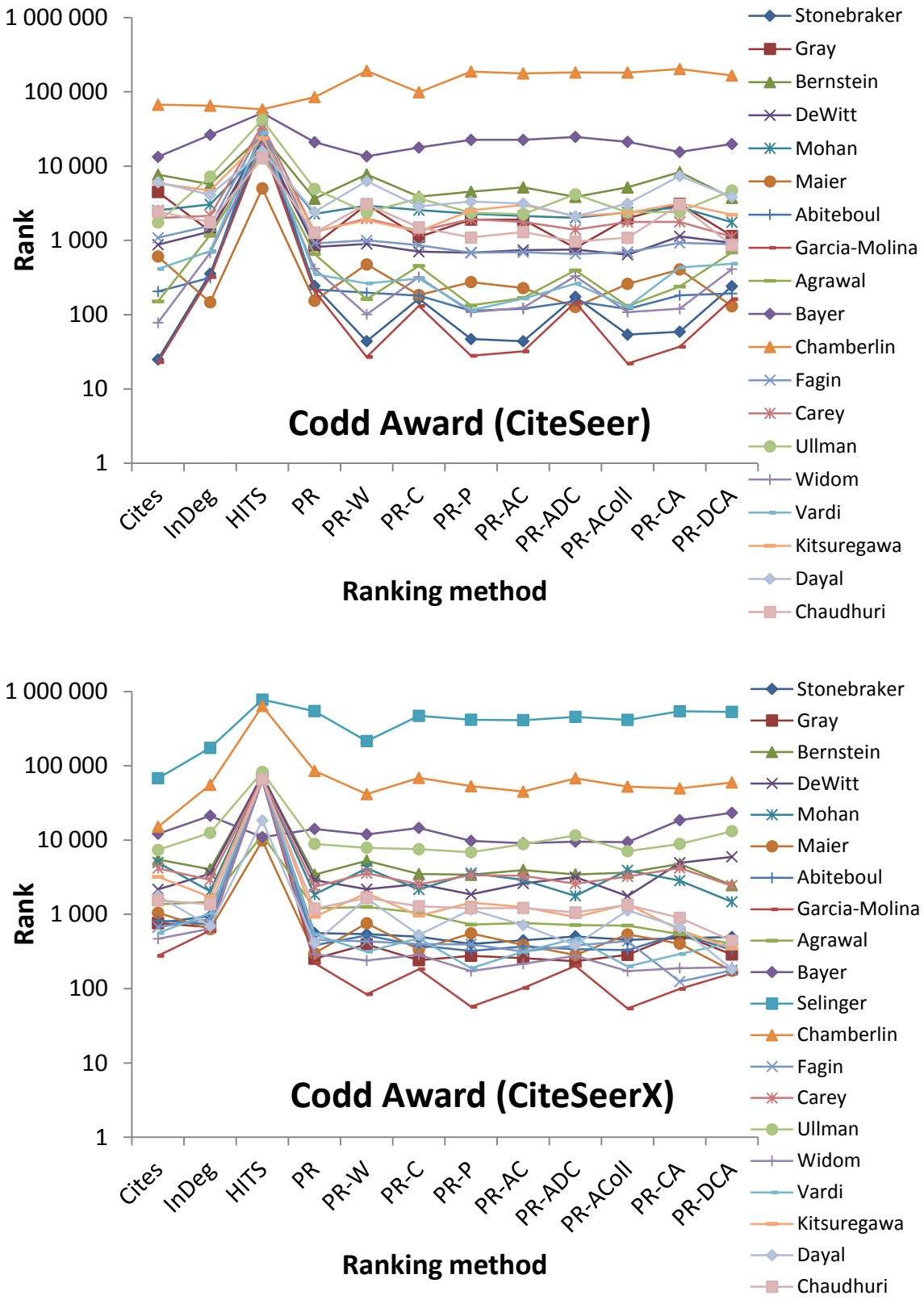


Figure 5: Ranks of Codd Award Winners byVarious Methods in Both CiteSeers

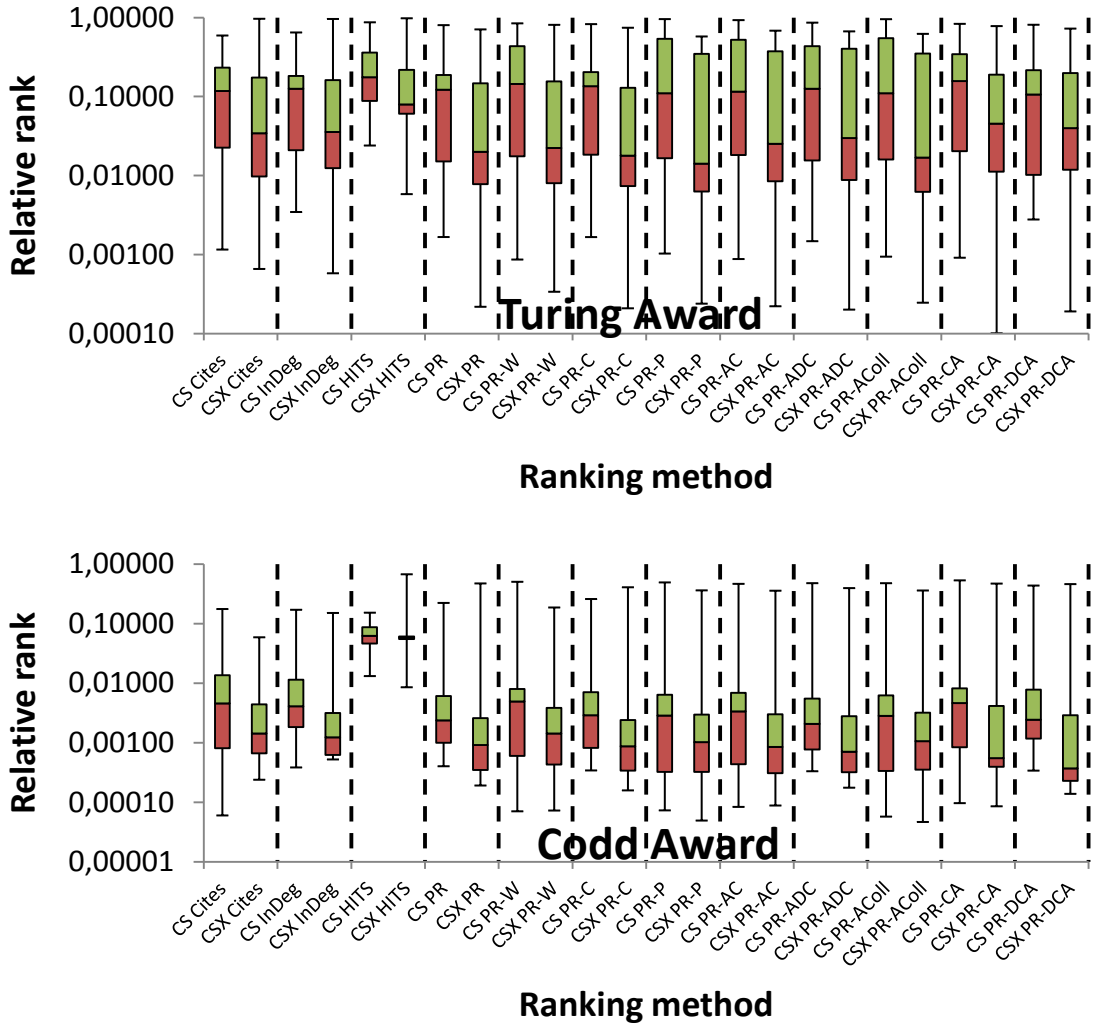


Figure 6: Boxplots of Relative Ranks Generated by Various Methods for Award Winners in Both CiteSeers

To answer the research question which of the two CiteSeers is better suited to evaluate computer science researchers, let us have a look at figure 6 and figure 7, in which charts comparing the ranks of Turing Award and Codd Award winners based on both CiteSeers are presented. figure 6 shows two boxplot charts (with the Y-axis on the logarithmic scale) depicting the relative ranks generated for the award winners by 12 methods in each CiteSeer. Thus, there are 24 different rankings for each of the awards. Relative ranks instead of absolute ranks are needed because the total number of researchers in CiteSeer and CiteSeer^X differs as explained earlier. In general, the ranks based on CiteSeer^X tend to be better (i.e., closer to 0) than those based on CiteSeer as we can see from the boxplots. We can also observe that the relative median rank of Turing Award winners in both CiteSeers roughly falls within top 10% and the relative median rank of Codd Award winners in both CiteSeers roughly falls within top 1% (except HITS). This might suggest that the coverage of general computer science literature (including theoretical computer science relevant to the Turing Award) in both Cite-

Seers is weaker than the coverage of database literature (relevant to the Codd Award). Another explanation may be that the Turing Award is a more life-time achievement prize than the Codd Award and that the main body of work of Turing Award winners was published in the years out of the scope of both CiteSeers. Similarly, the relative average and median ranks produced by 12 methods from two CiteSeer data sets for the winners of two awards are displayed in the charts in figure 7. Here the ranks of Turing Award winners based on CiteSeer^X are always clearly better than CiteSeer-based ranks and the ranks of Codd Award winners based on CiteSeer^X are generally better than those in CiteSeer with the most notable exception being the relative average rank by HITS. As the basic characteristics of the coauthorship networks of both CiteSeers are similar (except for their size), the cause of the better ranks in CiteSeer^X seems to be its broader coverage of the relevant computer science literature.

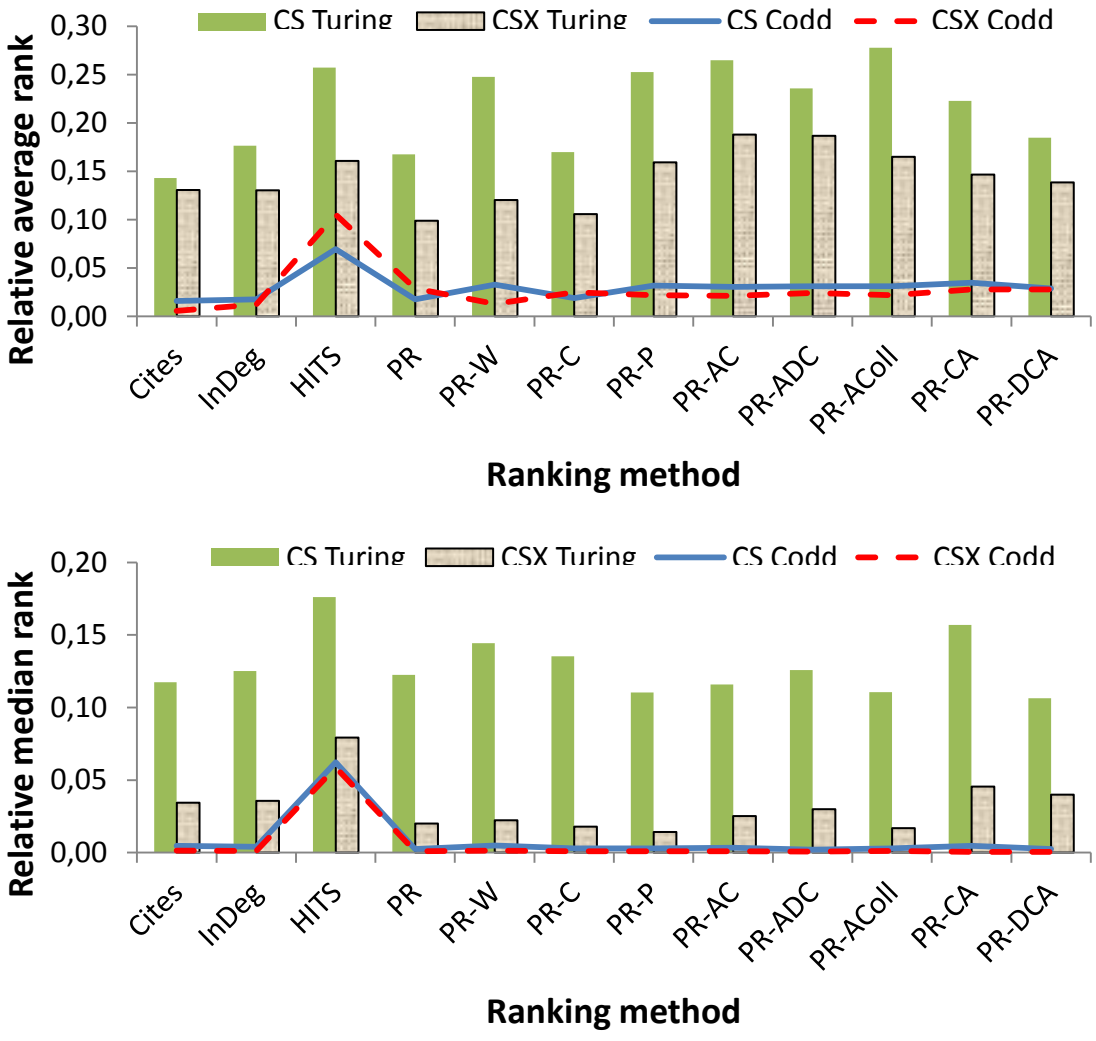


Figure 7: Relative Ranks by Various Methods for Award Winners in Both CiteSeers

4. CONCLUSIONS AND FUTURE WORK

CiteSeer and its current (yet still beta) version CiteSeer^X is a digital library and a search engine for computer science literature**,** whose article metadata have been successfully used for various purposes in the past. Some of the studies based on its data have been of bibliometric nature investigating its citation or coauthorship graphs. This paper belongs to such studies. Whereas CiteSeer has been discontinued and its most recent data come from December 2005, CiteSeer^X has been continuously updated until now. This research is concerned with CiteSeer^X data harvested from its Open Archives Initiative collection in March 2011. The number of articles covered by CiteSeer^X almost doubled between 2005 and 2011 and, unfortunately, the structure of the metadata on these articles freely obtainable from the respective Web sites changed considerably. These modifications do not enable the 2011 data to be analyzed in the same way as the 2005 data. The greatest difference is the general lack of the information on cited references in the article metadata. This fact excludes the possibility of a direct analysis of the CiteSeer^X citation graph acquired in this way. As a result, only its coauthorship network can be examined. The main contributions of this research are the following:

- We compared the structure of the article metadata in CiteSeer and CiteSeer^X freely available via their Web sites and constructed coauthorship (or author collaboration) networks from both data sets.
- We treated the coauthorship networks as citation graphs (according to the model of Yan and Ding [25]) and created rankings of researchers using 12 different ranking methods such as citation counts, HITS, PageRank, or its variations.
- We concentrated on the ranks achieved by the winners of the ACM A. M. Turing Award from the years 1991 – 2010 and by the winners of the ACM SIGMOD E. F. Codd Innovations Award from the years 1992 – 2011 and compared the rankings in both CiteSeers.

We thereby obtained the following main results:

- The coauthorship graphs of both CiteSeers have similar characteristics, apart from their sizes (see figure 1, figure 2, and figure 3).
- The basic properties of the individual rankings based on coauthorship networks are the same as of those previously reported that were based on citation networks, which may indicate the usefulness of coauthorship networks for the ranking of researchers (see figure 4 and figure 5).

- The relative ranks of both Turing Award and Codd Award winners based on CiteSeer^X are generally better than CiteSeer-based ranks presumably resulting from the broader coverage of the relevant computer science literature in CiteSeer^X (see figure 6 and figure 7).

In the future, a natural continuation of this research would be the acquisition of the complete CiteSeer^X citation graph and its thorough analysis. It would be interesting to see how different the researcher rankings are between CiteSeer and CiteSeer^X (based on their citation graphs) and between CiteSeer^X (based on the citation graph) and CiteSeer^X (based on the coauthorship graph).

Acknowledgements. This work was supported by the European Regional Development Fund (ERDF), project "NTIS – New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

REFERENCES:

[1] CiteSeer, http://citeseer.ist.psu.edu.

[2] S. Lawrence, C.L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing", *IEEE Computer*, Vol. 32, No. 6, 1999, pp. 67-71.

[3] CiteSeer^X, http://citeseerx.ist.psu.edu.

[4] Y. An, J. Janssen, and E.E. Milios, "Characterizing and mining the citation graph of the computer science literature", *Knowledge and Information Systems*, Vol. 6, No. 6, 2004, pp. 664–678.

[5] S. Chakrabarti and A. Agarwal, "Learning parameters in entity relationship graphs from ranking preferences", *Lecture Notes in Computer Science*, Vol. 4213, 2006, pp. 91–102.

[6] S. Chakrabarti, A. Pathak, and M. Gupta, "Index design and query processing for graph conductance search", *VLDB Journal*, Vol. 20, No. 3, 2011, pp. 445-470.

[7] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Tracking evolving communities in large linked networks", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, 2004, pp. 5249–5253.

[8] A. Joorabchi and A.E. Mahdi, "An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata", *Journal of Information Science*, Vol. 37, No. 5, 2011, pp. 499-514.

[9] A. Popescul, L.H. Ungar, S. Lawrence, and D.M. Pennock, "Statistical relational learning for document mining", *Proceedings of the Third IEEE International Conference on Data Mining*, Melbourne (USA), 2003, pp. 275–282.

[10] T. Šingliar and M. Hauskrecht, "Noisy-OR component analysis and its application to link analysis", *Journal of Machine Learning Research*, Vol. 7, 2006, pp. 2189–2213.

[11] D. Zhou, I. Councill, H. Zha, and C.L. Giles, "Discovering temporal communities from social network documents", *Proceedings of the Seventh IEEE International Conference on Data Mining*, Omaha (USA), 2007, pp. 745–750.

[12] H.-H. Chen, L. Gou, X. Zhang, and C.L. Giles, "CollabSeer: A search engine for collaboration discovery", *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, Ottawa (Canada), 2011, pp. 231-240.

[13] Q. He, D. Kifer, J. Pei, P. Mitra, and C.L. Giles, "Citation recommendation without author supervision", *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, Hong Kong (China), 2011, pp. 755-764.

[14] C.-L. Wu and J.-L. Koh, "Hierarchical topic-based communities construction for authors in a literature database", *Lecture Notes in Computer Science*, Vol. 6097, 2010, pp. 514–524.

[15] J. Bar-Ilan, "An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes", *Information Processing and Management*, Vol. 42, No. 6, 2006, pp. 1553–1566.

[16] D.G. Feitelson and U. Yovel, "Predictive ranking of computer scientists using CiteSeer data", *Journal of Documentation*, Vol. 60, No. 1, 2004, pp. 44-61.

[17] C.L. Giles and I.G. Councill, "Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 51, 2004, pp. 17599–17604.

[18] A.A. Goodrum, K.W. McCain, S. Lawrence, and C.L. Giles, "Scholarly publishing in the Internet age: A citation analysis of computer science literature", *Information Processing and Management*, Vol. 37, No. 5, 2001, pp. 661–675.

[19] D. Zhao, "Challenges of scholarly publications on the Web to the evaluation of science: A comparison of author visibility on the Web and in print journals", *Information Processing and Management*, Vol. 41, No. 6, 2005, pp. 1403–1418.

[20] D. Zhao and E. Logan, "Citation analysis using scientific publications on the Web as data source: A case study in the XML research area", *Scientometrics*, Vol. 54, No. 3, 2002, pp. 449–472.

[21] D. Zhao and A. Strotmann, "Can citation analysis of web publications better detect research fronts?", *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 9, 2007, pp. 1285–1302.

[22] L. Krumov, C. Fretter, M. Müller-Hannemann, K. Weihe, and M.-T. Hütt, "Motifs in co-authorship networks and their relation to the impact of scientific publications", *European Physical Journal B*, Vol. 84, No. 4, 2011, pp. 535-540.

[23] D. Fiala, "Mining citation information from CiteSeer data", *Scientometrics*, Vol. 86, No. 3, 2011, pp. 553-562.

[24] D. Fiala, "Bibliometric analysis of CiteSeer data for countries", *Information Processing and Management*, Vol. 48, No. 2, 2012, pp. 242-253.

[25] E. Yan and Y. Ding, "Discovering author impact: A PageRank perspective", *Information Processing and Management*, Vol. 47, No. 1, 2011, pp. 125-134.

[26] A. Sidiropoulos and Y. Manolopoulos, "A citation-based system to assist prize awarding", *SIGMOD Record*, Vol. 34, No. 4, 2005, pp. 54–60.

[27] D. Fiala, F. Rousselot, and K. Ježek, "PageRank for bibliographic networks", *Scientometrics*, Vol. 76, No. 1, 2008, pp. 135-158.

[28] D. Fiala, "Time-aware PageRank for bibliographic networks", *Journal of Informetrics*, Vol. 6, No. 3, 2012, pp. 370-388.

[29] CiteSeer^X OAI, http://citeseerx.ist.psu.edu/oai2.

[30] J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Vol. 46, No. 5, 1999, pp. 604-632.

[31] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, 1998, pp. 107-117.

Conclusions

And future work

Informetrics is a young scientific field comprising scientometrics, bibliometrics, webometrics, cybermetrics and other –metrics disciplines. It lies at the crossroads of computer science, information science, and social sciences. At present, the Czech research output in this domain is almost negligible as can be seen in the following Table 3. This table was generated on 17 December 2013 from the Web of Science (WoS) citation databases by Thomson Reuters to retrieve article numbers and from the 2012 Journal Citation Reports® (Thomson Reuters, 2013) to obtain journal impact factors and categories. Table 3 presents ten journals that, in my view, are entirely or partially concerned with informetrics or related topics. All of the journals are categorized as "Information Science & Library Science" (JCR Social Sciences Edition) and, in addition, six of them are further classified into "Computer Science, Information Systems" (JCR Science Edition) and one into "Computer Science, Interdisciplinary Applications" (JCR Science Edition). As we can see, the total number of articles (including reviews) published in these journals since they have been indexed in the citation database is well over 13 thousand.

Journal title	Impact factor	Indexed since	Articles	Czech articles	Czech %
Annual Review of Information Science and Technology	2.174	1977	339	0	0.00
Aslib Proceedings	0.432	1965	1 929	0	0.00
Information Processing & Management	0.817	1975	2 032	17	0.84
Journal of Documentation	1.138	1945	1 385	1	0.07
Journal of Information Science	1.238	1979	1 413	0	0.00
Journal of Informetrics	4.153	2007	369	2	0.54
Journal of the American Society for Information Science and Technology	2.005	2001	1 977	4	0.20
Online Information Review	0.939	2000	627	0	0.00
Research Evaluation	1.074	2000	366	0	0.00
Scientometrics	2.133	1978	3 268	17	0.52
			13 705	**41**	**0.30**

Table 3: Journals concerned with informetric topics

However, the number of articles having one Czech coauthor at least is 41 only, which accounts for 0.3% of the total production in the field. In this context, those four journal articles of mine already indexed in WoS constitute one tenth (or 9.8%) of the overall Czech research production in this scientific discipline and with another article (Article 4) appearing soon in WoS, this share will even encrease. Therefore, informetrics (and also research evaluation, which is one of its main applications) should be further promoted and cultivated in the Czech Republic. I have done so in my previous work (presented in this thesis) and I envisage to do so in my future work as well.

Since it would be too ambitious to try to tackle all of the open problems mentioned in the introduction to this habilitation thesis, my aim is to concentrate primarily (but not exclusively) on the following issues: a) life-time achievement vs. current performance (influence), b) self-citations and citation cliques, and c) impact factor flaws and other journal quality metrics. The approaches to solve these problems will be outlined in the next paragraphs.

Life-time achievement vs. current performance (influence). I will continue seeking to introduce a dynamic indicator of scientific performance that will not only increase in time but also decrease according to the current publication activity and citation reputation. A model of such an indicator can be the ħ-index (h bar), which, contrary to the h-index, can decrease in time (Hirsch, 2010). But a decrease can only occur if the researcher under examination publishes new articles. If he/she stops publishing, the ħ-index (as well as all other related metrics) will never decline – it can only remain the same or grow. I will address this problem by considering "real time" and not just "publication time". In this way, the "real time indicator" will be able to change over time even if the scientist under study will no longer be active. The new indicator will consider a time window for both publications and citations and will, therefore, reflect current performance rather than life-time achievement. A first attempt at such an indicator is the Current Index presented in Article 4, but more experiments are needed to test various time window sizes and the indicator's properties in different research areas.

Self-citations and citation cliques. I will develop and test new algorithms that will be able to detect indirect self-citations and citation loops, i.e. citations leading from a researcher to the same researcher via one or more other researchers. As a result, the edges in a directed graph of citations between authors will be weighted according to their presence or absence in citation loops. These citation loops can come into being either deliberately upon agreement of several scientists involved or naturally as a product of normal research work. My new algorithms will analyze a citation network and find citation loops of up to a certain length. In addi-

tion, we will need to bring time information to the citation analysis for the order in which citations occurred to be respected. I believe that the resulting scientometric indicator(s) will better represent the "true" impact of a researcher's scientific work corrected for his/her social interaction.

Impact factor flaws and other journal quality metrics. I will experiment with current journal quality metrics, compare them, identify their strengths and weaknesses and propose some improvements to overcome their drawbacks. More specifically, I will test how journal impact factor rankings change if the time window for publications and citations is modified, if a median or modus is used in the impact factor equation instead of a mean, if only citations to "citable items" are counted in the numerator of the impact factor equation, or if author self-citations and journal self-citations are left out from the computation. In addition, I will determine how the journal impact factor and other PageRank-based metrics (including Eigenfactor, Article Influence, SCImago Journal Rank, and Source Normalized Impact per Paper) are correlated and how sensitive they are to small changes in the input data. Based on the analysis and experimental results, I will identify the best technique to determine journal quality and/or propose alternative journal quality indicators.

References

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). h-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4), 273-289.

Archambault, E., & Larivière, V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics*, 79(3), 635-649.

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

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, 68(5), 389 314–316.

Bollen, J., Rodriguez, M. A., & Van De Sompel, H. (2006). Journal status. *Scientometrics*, 69(3), 669–687.

Bornmann, L., & Daniel, H.-D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology*, 58(9), 1381-1385.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 107-117.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1), 393 8–15.

Costas, R., & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, 1(3), 193-203.

Ding, Y. (2011). Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*, 62(2), 236–245.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11), 2229–2243.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131-152.

Egghe, L. (2007). Dynamic h-index: The Hirsch index in function of time. *Journal of the American Society for Information Science and Technology*, 58(3), 452-454.

Egghe, L. (2011). A disadvantage of h-type indices for comparing the citation impact of two researchers. *Research Evaluation*, 20(4), 341-346.

Fiala, D. (2011). Mining citation information from CiteSeer data. *Scientometrics*, 86(3), 553-562.

Fiala, D. (2012a). Bibliometric analysis of CiteSeer data for countries. *Information Processing and Management*, 48(2), 242-253.

Fiala, D. (2012b). Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, 6(3), 370-388.

Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks. *Scientometrics*, 76(1), 135-158.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569–16572.

Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3), 741-754.

Hubbard, S. C., & McVeigh, M. E. (2011). Casting a wide net: The journal impact factor numerator. *Learned Publishing*, 24(2), 133-137.

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

Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 52(6), 855-863.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604–632.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing and Management*, 44(2), 800–810.

Podlubny, I., & Kassayova, K. (2006). Towards a better list of citation superstars: Compiling a multidisciplinary list of highly cited researchers. *Research Evaluation*, 15(3), 154-162.

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

Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *EPL*, 78(3).

Schreiber, M. (2010). A case study of the modified g index: Counting multi-author publications fractionally. *Journal of Informetrics*, 4(4), 636-643.

Schubert, A. (2007). Successive h-indices. *Scientometrics*, 70(1), 201-205.

Schubert, A., & Glänzel, W. (2007). A systematic analysis of hirsch-type indices for journals. *Journal of Informetrics*, 1(3), 179-184.

The PLoS Medicine Editors (2006). The impact factor game: It is time to find a better way to assess the scientific literature. *PLoS Medicine*, 3(6), 707-708.

Vanclay, J. K. (2007). On the robustness of the h-index. *Journal of the American Society for Information Science and Technology*, 58(10), 1547-1550.

Yan, E., & Ding, Y. (2010). Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*, 61(8), 1635–1643.

Yan, E., & Ding, Y. (2011). Discovering author impact: A PageRank perspective. *Information Processing and Management*, 47(1), 125–134.

Wu, J., Lozano, S., & Helbing, D. (2011). Empirical study of the growth dynamics in real career h-index sequences. *Journal of Informetrics*, 5(4), 489-497.

Appendix - Author's Publications and Their Citations

[1] Fiala, D. (2013). Testing ranking algorithms on CiteSeer data. ***Global Journal of Advanced Engineering Technologies***, 2(4), 176-180. ISSN: 2277-6370. Link.

[2] Fiala, D. (2013). From CiteSeer to CiteSeer^X: Author rankings based on coauthorship networks. ***Journal of Theoretical and Applied Information Technology***. ISSN: 1992-8645.

[3] Fiala, D. (2013). Suborganizations of Institutions in Library and Information Science Journals. ***Information***, 4(4), 351-366. ISSN: 2078-2489. DOI: 10.3390/info4040351.

[4] Dostal, M., Nykl, M., Ježek, K., & Fiala, D. (2013). Linked Data and PageRank-based Classification. In *Proceedings of IADIS International Conference on Theory and Practice in Modern Computing*, pp. 61–64, Prague, Czech Republic. ISBN: 978-972-8939-94-6. Link.

[5] Fiala, D. (2013). Extracting information from CiteSeer's textual data. ***Journal of Theoretical and Applied Information Technology***, 56(2), 176-182. ISSN: 1992-8645. Link.

[6] Fiala, D. (2013). Science Evaluation in the Czech Republic: The Case of Universities. ***Societies***, 3(3), 266-279. ISSN: 2075-4698. DOI: 10.3390/soc3030266. Cited by 1:

1. Parinov, S. I., Kogalovsky, M. R., & Nevolin, I. V. (2013). Европейский опыт оценки научной результативности и его использование в Российской академии наук. *Technical report*, Russian Academy of Sciences, Moscow, Russia. Link.

[7] Fiala, D. (2013). Current Index: A Proposal for a Dynamic Rating System for Researchers. ***Journal of the American Society for Information Science and Technology***. ISSN: 1532-2890. IF (2012): **2.005**. DOI: 10.1002/asi.23049. (in press)

[8] Dostal, M., Fiala, D., & Ježek, K. (2013). Semantic Markup for Web Applications. In *Proceedings of the International Conference on Computer, Communication, Information Sciences, and Engineering*, pp. 506 – 509, Paris, France. ISSN: 2010-376X. Link.

[9] Fiala, D. (2012). Time-aware PageRank for bibliographic networks. ***Journal of Informetrics***, 6(3), 370-388. ISSN: 1751-1577. IF (2012): **4.153**. DOI: 10.1016/j.joi.2012.02.002. Cited by 1:

1. Meng., Q., & Kennedy, P. J. (2013). Discovering influential authors in heterogeneous academic networks by a co-ranking method. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, pp. 1029-1036, San Francisco, CA, USA. ISBN: 978-1-4503-2263-8. DOI: 10.1145/2505515.2505534.

[10] Fiala, D. (2012). Bibliometric analysis of CiteSeer data for countries. ***Information Processing and Management***, 48(2), 242-253. ISSN: 0306-4573. IF (2012): **0.817**. DOI: 10.1016/j.ipm.2011.10.001. Cited by 1:

1. Xie, Z., & Willett, P. (2013). The development of computer science research in the People's Republic of China 2000–2009: a bibliometric study. ***Information Development***, 29(3), 251-264. ISSN: 0266-6669. IF (2012): **0.375**. DOI: 10.1177/0266666912458515.

[11] Fiala, D. (2011). Mining citation information from CiteSeer data. ***Scientometrics***, 86(3), 553-562. ISSN: 0138-9130. IF (2011): **1.966**. DOI: 10.1007/s11192-010-0326-1. Cited by 4:

1. Alzahrani, S., Palade, V., Salim, N., & Abraham, A. (2012). Using structural information and citation evidence to detect significant plagiarism cases in scientific publications. ***Journal of the American Society for Information Science and Technology***, 63(2), 286-312. ISSN: 1532-2882. IF (2012): **2.005**. DOI: 10.1002/asi.21651.

2. Cabanac, G. (2012). Shaping the landscape of research in information systems from the perspective of editorial boards: A scientometric study of 77 leading journals. ***Journal of the American Society for Information Science and Technology***, 63(5), 977-996. ISSN: 1532-2882. IF (2012): **2.005**. DOI: 10.1002/asi.22609.

3. Nykl, M., & Ježek, K. (2012). Varianty použití PageRanku pro citační analýzu. In *Proceedings of DATAKON 2012*, pp. 87 - 97, Mikulov, Czech Republic. ISBN: 978-80-553-1049-7. Link.

4. Kaya, M., & Alhajj, R. (2013). Development of multidimensional academic information networks with a novel data cube based modeling method. ***Information Sciences***. IF (2012): **3.643**. DOI: 10.1016/j.ins.2013.11.012. (in press)

[12] Fiala, D. (2009). Web Mining Methods for the Detection of Authoritative Sources: Theory and Practice. VDM Verlag, Saarbrücken. ISBN: 978-3639173376. Link.

[13] Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks. ***Scientometrics***, 76(1), 135-158. ISSN: 0138-9130. IF (2008): **2.328**. DOI: 10.1007/s11192-007-1908-4. Cited by 18:

1. Vellino, A. (2008). The effect of PageRank on the collaborative filtering recommendation of journal articles. *Technical report*, Canada Institute for Scientific and Technical Information, National Research Council, Ottawa, Ontario, Canada. Link.

2. Liu, Q.-B., & Xu, J. (2008). Personalized ranking of scientific publications using link analysis. ***Zhongshan Daxue Xuebao/Acta Scientiarum Naturalium Universitatis Sunyatseni***, 47(6), 87-92. ISSN: 0529-6579. Link.

3. Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. ***Journal of the American Society for Information Science and Technology***, 60(11), 2229-2243. ISSN: 1532-2882. IF (2009): **2.300**. DOI: 10.1002/asi.21171.

4. Li, J., & Willett, P. (2009). ArticleRank: A PageRank-based alternative to numbers of citations for analysing citation networks. ***Aslib Proceedings: New Information Perspectives***, 61(6), 605-618. ISSN: 0001-253X. IF (2009): **0.595**. DOI: 10.1108/00012530911005544.

5. Vellino, A. (2009). Recommending journal articles with PageRank ratings. *Technical report*, Canada Institute for Scientific and Technical Information, National Research Council, Ottawa, Ontario, Canada. Link.

6. Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. ***Journal of the American Society for Information Science and Technology***, 60(10), 2107-2118. ISSN: 1532-2882. IF (2009): **2.300**. DOI: 10.1002/asi.21128.

7. Yan, E., & Ding, Y. (2009). The use of centrality measures in scientific evaluation: a coauthorship network analysis. In *Proceedings of the 12th International Conference of the International Society for Scientometrics and Informetrics*, vol. 2, pp. 561-570, Rio de Janeiro, Brazil. ISSN: 2175-1935. Link.

8. Chikhi, N. F. (2010). Calcul de centralité et identification de structures de communautés dans les graphes de documents. *PhD dissertation*, Université Toulouse 3 Paul Sabatier, Toulouse, France. Link.

9. Sani, E., Ruffaldi, & E., Bergamasco, M. (2010). Interactive technology maps for strategic planning and research directions based on textual and citation analysis of patents. In A. Gunasekaran and M. Sandhu (Eds.), *Handbook on Business Information Systems* (pp. 487 - 514). Singapore: World Scientific Publishing Co. ISBN: 978-981-283-605-2. DOI: 10.1142/9789812836069_0020.

10. Su-Fang, L., & Li, J. (2010). Review on the mechanism of link degree distribution: preferential attachment and uniform attachment. ***Journal of Intelligence***, 29(10), 167-171. ISSN: 1002-1965. Link.

11. Fischbach, K., Putzke, J., & Schoder, D. (2011). Co-authorship networks in electronic markets research. ***Electronic Markets***, 21(1), 19-40. ISSN: 1019-6781. IF (2011): **0.784**. DOI: 10.1007/s12525-011-0051-5.

12. Gągolewski, M. (2011). Wybrane operatory agregacji i ich zastosowanie w modelu formalnym systemu oceny jakości w nauce. *PhD dissertation*, Polish Academy of Sciences, Warsaw, Poland. Link.

13. Ku, L. (2011). PatentRank Algorithm: A Study of Using Cited Time and Citation Network to Calculate U.S. Patents. ***New Technology of Library and Information Service***, 6, 14-19. ISSN: 1003-3513. Link.

14. Yan, E., & Ding, Y. (2011). Discovering author impact: A PageRank perspective. ***Information Processing and Management***, 47(1), 125-134. ISSN: 0306-4573. IF (2011): **1.119**. DOI: 10.1016/j.ipm.2010.05.002.

15. Duan, Q., & Zhu, D. (2012). Co-ranking Approach Based on the Heterogeneous Network for both Co-authorship Data and Citation Data. ***Journal of The China Society for Scientific and Technical Information***, 31(2), 189-195. ISSN: 1000-0135. Link.

16. Hong, D., & Baccelli, F. (2012). On a joint research publications and authors ranking. *Technical report*, Alcatel-Lucent Bell Labs France, Nozay, France. Link.

17. Su, C., Pan, Y., Ma, Z., Yuan, J., Yu, Z., & Guo, H. (2012). PrestigeRank and Peer Review for Evaluation of Scientific Papers. ***Journal of The China Society for Scientific and Technical Information***, 31(2), 180-188. ISSN: 1000-0135. Link.

18. Lin, L., Xu, Z., Ding, Y., & Liu, X. (2013). Finding topic-level experts in scholarly networks, ***Scientometrics***, 97(3), 797-819. ISSN: 0138-9130. IF (2012): **2.133**. DOI: 10.1007/s11192-013-0988-6.

19. Ben Jabeur, L. (2013). Leveraging social relevance: Using social networks to enhance literature access and microblog search. *PhD dissertation*, Université Toulouse 3 Paul Sabatier, Toulouse, France. Link.

[14] Ježek, K., Fiala, D., & Steinberger, J. (2008). Exploration and Evaluation of Citation Networks. In *Proceedings of the 12th International Conference on Electronic Publishing*, pp. 351-362, Toronto, Canada. ISBN: 978-0-7727-6315-0. Link. Cited by 5:

1. Vellino, A. (2008). The effect of PageRank on the collaborative filtering recommendation of journal articles. *Technical report*, Canada Institute for Scientific and Technical Information, National Research Council, Ottawa, Ontario, Canada. Link.

2. Li, J., & Willett, P. (2009). ArticleRank: A PageRank-based alternative to numbers of citations for analysing citation networks. ***Aslib Proceedings: New Information Perspectives***, 61(6), 605-618. ISSN: 0001-253X. IF (2009): **0.595**. DOI: 10.1108/00012530911005544.

3. Radziwill, N. M. (2009). Topology, evolution, and network-based continuous improvement of the quality management journal. *PhD dissertation*, Indiana State University, Terre Haute, Indiana, USA. Link.

4. Vellino, A. (2009). Recommending journal articles with PageRank ratings. *Technical report*, Canada Institute for Scientific and Technical Information, National Research Council, Ottawa, Ontario, Canada. Link.

5. Su-Fang, L., & Li, J. (2010). Review on the mechanism of link degree distribution: preferential attachment and uniform attachment. ***Journal of Intelligence***, 29(10), 167-171. ISSN: 1002-1965. Link.

[15] Fiala, D. (2007). Web Mining Methods for the Detection of Authoritative Sources. *PhD dissertation*, University of West Bohemia, Plzeň, Czech Republic, and Louis Pasteur University Strasbourg, France. Link. Cited by 1:

1. Nykl, M., & Ježek, K. (2012). Varianty použití PageRanku pro citační analýzu. In *Proceedings of DATAKON 2012*, pp. 87 - 97, Mikulov, Czech Republic. ISBN: 978-80-553-1049-7. Link.

[16] Fiala, D., Rousselot, F., & Ježek, K. (2007). Ranking algorithms for web sites. In *Proceedings of the 3rd International Conference on Web Information Systems and Technologies*, pp. 372-375, Barcelona, Spain. ISBN: 978-972-8865-78-8. Link.

[17] Fiala, D., Ježek, K., & Rousselot, F. (2007). Využití struktury webu pro vyhledávání autoritativních institucí a osob. In *Proceedings of the 6th Annual Conference ZNALOSTI*, pp. 300-303, Ostrava, Czech Republic. ISBN: 978-80-248-1279-3. Link.

[18] Fiala, D., Ježek, K., & Rousselot, F. (2006). Finding Authoritative Researchers on Academic Web Sites. In *Proceedings of the Conference of the World Academy of Science, Engineering, and Technology*, pp. 74-79, Cairo, Egypt. ISBN: 975-00803-7-8. Link.

[19] Fiala, D., Tesař, R., Ježek, K., & Rousselot, F. (2006). Extracting Information from Web Content and Structure. In *Proceedings of the 9th International Conference on Information Systems Implementation and Modelling*, pp. 133-140, Přerov, Czech Republic. ISBN: 80-86840-19-0. Link. Cited by 1:

1. Rodríguez, R., Jorge, E., Barrera, F., Harry, A., Bautista, M., & Sandra, P. (2011). Software para el filtrado de páginas web pornográficas basado en el clasificador KNN - UDWEBPORN. ***Revista Avances en Sistemas e Informática***, 8(1), 43-49. ISSN: 1657-7663. Link.

[20] Fiala, D. (2005). Web Mining and Its Applications to Researchers Support. *Technical report*, no. DCSE/TR-2005-06, University of West Bohemia, Plzeň, Czech Republic. Link.

1. Jain, A., Sharma, R., Dixit, G., & Tomar, V. (2013). Page Ranking Algorithms in Web Mining, Limitations of Existing Methods and a New Method for Indexing Web Pages. In *Proceedings of the International Conference on Communication Systems and Network Technologies*, pp. 640-645, Gwalior, India. ISBN: 978-1-4673-5603-9. DOI: 10.1109/CSNT.2013.137.

[21] Tesař, R., Fiala, D., Rousselot, F., & Ježek, K. (2005). A comparison of two algorithms for discovering repeated word sequences. In *Proceedings of the 6th International Conference on Data Mining, Text Mining, and their Business Applications*, pp. 121-131, Skiathos, Greece. ISBN: 1-84564-017-9. Link. Cited by 2:

1. Češka, Z. (2007). Využití N-gramů pro odhalování plagiátů. In *Proceedings of the Conference on information (intelligent) technologies - applications and theory*, pp. 63-66, Poľana, Slovakia. ISBN: 978-80-969184-6-1. Link.

2. Krátký, M., Bača, R., Bednář, D., Walder, J., Dvorský, J., & Chovanec, P. (2011). Index-based N-gram extraction from large document collections. In *Proceedings of the 6th International Conference on Digital Information Management*, pp. 73-78, Melbourne, Australia. ISBN: 978-145771538-9. DOI: 10.1109/ICDIM.2011.6093324.

[22] Belaïd, A., Alusse, A., Rangoni, Y., Cecotti, H., Farah, F., Gagean, N., Fiala, D., Rousselot, F., & Vigne, H. (2005). Document retro-conversion for personalized electronic reedition. In *Proceedings of the International Workshop on Document Analysis*, pp. 193-218, Kolkata, India. ISBN: 978-8177647846. Link. Cited by 1:

1. Ben Moussa, S., Zahour, A., Benabdelhafid, A., & Alimi, A. M. (2010). New features using fractal multi-dimensions for generalized Arabic font recognition. ***Pattern Recognition Letters***, 31(5), 361-371. ISSN: 0167-8655. IF (2010): **1.235**. DOI: 10.1016/j.patrec.2009.10.015.

[23] Fiala, D., & Ježek, K. (2004). Retrieving Citations on the Web. In *Proceedings of the International Conference on Knowledge Engineering and Decision Support*, pp. 481-488, Porto, Portugal. ISBN: 972-8688-24-5. Link.

[24] Fiala, D. (2003). A System for Citations Retrieval on the Web. *MSc thesis*, University of West Bohemia, Plzeň, Czech Republic. Link.

Invited Talks

Technische Universität Chemnitz, Chemnitz, Germany.
Neue informetrische Methoden zur Bewertung der Wissenschaftler. December 16, 2013.

Ostbayerische Technische Hochschule Regensburg, Regensburg, Germany.
Webtechnologien und Webdienste. December 5, 2013.

Univerza v Ljubljani, Ljubljana, Slovenia.
PageRank: from Google to citation analysis. July 10, 2013.

Technická univerzita v Košiciach, Košice, Slovakia.
PageRank: od Googlu k citační analýze. April 9, 2013.

Royal Holloway, University of London, London, United Kingdom.
Web technologies and services. November 22, 2012.

Hochschule Regensburg, Regensburg, Germany.
PageRank: v*on Google zur Zitationsanalyse*. July 4, 2012.

Research Grants

2011 – present: “NTIS – New Technologies for Information Society”, European Centre of Excellence, CZ.1.05/1.1.00/02.0090 (junior researcher), funded by the European Regional Development Fund (ERDF).

2006, 2009 – 2010: 2C06009 COT-SEWing (project participant), on the semantic Web and natural language processing methods, funded by the Ministry of Education of the Czech Republic.

2005: FRVŠ 1347/2005/G1 (principal investigator), on knowledge extraction from the Web content and topology, funded by the Ministry of Education of the Czech Republic.

2004 – 2005: RNTL Paploo (project participant), on electronic re-edition, funded by the French Ministry of Education.

Reviewing Activities

2014: programme committee member of the 8th **IEEE International Conference on Research Challenges in Information Science** (Marrakesh, Morocco).

2013: reviewer of ***Science Innovation*** (ISSN: 2328-7861)

2013: reviewer of ***Computers and Electrical Engineering*** (ISSN: 0045-7906).

2013: reviewer of ***Webology*** (ISSN: 1735-188X).

2013: reviewer of the 16th **International Conference on Text, Speech and Dialogue** (Plzeň, Czech Republic).

2012: external referee of the Romanian **National Research Council** (CNCS) in the programmes Exploratory Research Projects (PCE2012) and Postdoctoral Research Projects (PD2012) for the research areas *PE6_6 Informatics and information systems* and *SH2_13 Social studies of science and technology, S&T policies, science and society*.

2012: reviewer of the 16th **International Conference on Elecktronic Publishing** (Guimares, Portugal).

2012: reviewer of the **DATAKON** 2012 conference (Mikulov, Czech Republic).

Curriculum Vitae

Dalibor Fiala (http://www.kiv.zcu.cz/~dalfia/)

Qualifications:

- 2007, Ph.D. under joint supervision in computer science at Louis Pasteur University, Strasbourg, France and at the University of West Bohemia in Pilsen (UWB), Czech Republic
- 2003, MSc. (Ing.) in computer science at UWB

Professional experience:

- 2009 – present: assistant professor at UWB
- 2007 – 2009: software engineer at Gefasoft AG in Munich, Germany
- 2006 – 2007: assistant professor at Marc Bloch University, Strasbourg, France

Research grants:

- 2011 – present: "NTIS – New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090 (junior researcher), funded by the European Regional Development Fund (ERDF)
- 2006, 2009 – 2010: 2C06009 COT-SEWing (project participant), on the semantic Web and natural language processing methods, funded by the Ministry of Education of the Czech Republic
- 2005: FRVŠ 1347/2005/G1 (principal investigator), on knowledge extraction from the Web content and topology, funded by the Ministry of Education of the Czech Republic
- 2004 – 2005: RNTL Paploo (project participant), on electronic re-edition, funded by the French Ministry of Education

Research and teaching experience:

- Publications: 21 publications in total on data mining, Web mining, information retrieval, information science, and informetrics from which there are 10 conference proceedings papers, 10 (impacted) journal articles and 1 monograph; 8 of the publications are in the Web of Science database
- Research interests: data mining, information retrieval, and information science
- Courses taught: Web applications, database models and methods, programming structures, user interfaces

Institutions of Researchers Citing Author’s Work