A Comparison of On-Line Computer Science Citation Databases

Vaclav Petricek[1], Ingemar J. Cox[1], Hui Han[2],
Isaac G. Councill[3], and C. Lee Giles[3]

[1] University College London,
WC1E 6BT, Gower Street, London, United Kingdom
v.petricek@cs.ucl.ac.uk, ingemar@ieee.org
[2] Yahoo! Inc., 701 First Avenue, Sunnyvale, CA, 94089
huihan@yahoo-inc.com
[3] The School of Information Sciences and Technology,
The Pennsylvania State University, University Park, PA 16802, USA
igc2@psu.edu giles@ist.psu.edu

Abstract. This paper examines the difference and similarities between the two on-line computer science citation databases DBLP and CiteSeer. The database entries in DBLP are inserted manually while the CiteSeer entries are obtained autonomously via a crawl of the Web and automatic processing of user submissions. CiteSeer's autonomous citation database can be considered a form of self-selected on-line survey. It is important to understand the limitations of such databases, particularly when citation information is used to assess the performance of authors, institutions and funding bodies.

We show that the CiteSeer database contains considerably fewer single author papers. This bias can be modeled by an exponential process with intuitive explanation. The model permits us to predict that the DBLP database covers approximately 24% of the entire literature of Computer Science. CiteSeer is also biased against low-cited papers.

Despite their difference, both databases exhibit similar and significantly different citation distributions compared with previous analysis of the Physics community. In both databases, we also observe that the number of authors per paper has been increasing over time.

1 Introduction

Several public[1] databases of research papers became available due to the advent of the Web [1,22,5,3,2,4,8] These databases collect papers in different scientific disciplines, index them and annotate them with additional metadata. As such, they provide an important resource for (i) finding publications, (ii) identifying important, i.e. highly cited, papers, and (iii) locating papers that cite a particular paper. In addition, author and document citation rates are increasingly being

[1] By public, we mean that access to the database is free of charge. Commercial databases are also available, the most well-known being the science-citation index [6]

A. Rauber et al. (Eds.): ECDL 2005, LNCS 3652, pp. 438–449, 2005.

used to quantify the scientific impact of scientists, publications, journals and funding agencies.

Within the computer science community, there are two popular public citation databases. These are DBLP [5] and CiteSeer [22]. The two databases are constructed in very different ways. In DBLP, each entry is manually inserted by a group of volunteers and occasionally hired students. The entries are obtained from conference proceeding and journals. In contrast, each entry in CiteSeer is automatically entered from an analysis of documents found on the Web. There are advantages and disadvantages to both methods and we discuss these issues in more detail in the next Section.

In Section 2 we compare the two databases based on the distribution of number of authors. We reveal that there are very pronounced differences which appear to be primarily due to the absence of very many single author papers in the CiteSeer database. A probabilistic model for document acquisition is then developed that provides an intuitive explanation for this phenomenon in Section 2.2.

There have been a number of studies on the distribution of citations [23,28,19,12] and the number of collaborators [26] using other on-line databases. This literature is reviewed in Section 3. We replicate some of these studies and show that citation distributions from both DBLP and CiteSeer differ considerably from those reported in other research communities.

2 The DBLP and CiteSeer Databases

There are a number of public, on-line computer science databases [1,22,5,3,2,4]. The CS BiBTeX database [4] contains a collection of over 1.4 million references. However, only 19,000 entries currently contain cross-references to citing or cited publications. The Compuscience database [2] contains approximately 400,000 entries. The Computing Research Repository CoRR [3] contains papers from 36 areas of computer science and is now part of ArXiV [1] that covers Physics, Mathematics, Nonlinear Sciences, Computer Science and Quantitative Biology. Networked Computer Science Technical Reference Library is a repository of Computer Science Technical Reports located at Old Dominion University.

DBLP was created by Michael Ley in 1998 [5]. It currently contains over 550,000 computer science references from around 368,000 authors. CiteSeer was created by Steve Lawrence and C. Lee Giles in 1997 [22]. It currently contains over 716,797 documents.

We chose to examine DBLP and CiteSeer due to the availability of detailed citation information and their popularity.

In our analysis we focus on the difference in data acquisition and the biases that this difference introduces.

2.1 The Differences Between DBLP and CiteSeer Databases

While both the DBLP and CiteSeer databases contain computer science bibliography and citation data, their acquisition methods greatly vary. In this section we first discuss these differences in acquisition methods, then we look at the distribution of papers over time in each dataset, and after that we compare the distribution in the number of authors per paper. Section 2.2 then describes acquisition models for both DBLP and CiteSeer.

Data acquisition. At the time of writing, DBLP contains over 550,000 bibliographic entries. Papers in DBLP originally covered database systems and logic programing. Currently DBLP also includes theory of information, automata, complexity, bioinformatics and other areas. Database entries are obtained by a limited circle of volunteers who manually enter tables of contents of journals and conference proceedings. The volunteers also manually entered citation data as part of compiling the ACM anthology CD/DVDs. Corrections that are submitted to the maintainer are also manually checked before committing. Though the breadth of coverage may be more narrow than CiteSeer, DBLP tries to ensure comprehensive and complete coverage within its scope. The coverage of ACM, IEEE and LNCS is around 80–90%. The narrower focus of DBLP is partially enforced by the cost associated with manual entry. Although there is the possibility of human error in the manual process of DBLP, its metadata is generally of higher quality than automatically extracted metadata[2].

In our analysis we used a DBLP dataset consisting of 496,125 entries. From this we extracted a dataset of 352,024 papers that specified the year of publication and the number of authors. Only papers published between 1990 and 2002 were included, due to the low number of papers available outside of this range.

CiteSeer currently contains over 716,797 bibliographic entries. Automatic crawlers have the potential of achieving higher coverage as the cost of automatic indexing is lower than for manual entry. However, differences in typographic conventions make it hard to automatically extract metadata such as author names, date of publication, etc.

CiteSeer entries can be acquired in two modes. First, the publication may be encountered during a crawl[3]. In this case, the document will be parsed, and title, author and other information will be entered into the database. Second, during this parsing operation, a document's bibliography is also analyzed and previously unknown cited documents are also entered into the database.

CiteSeer is continuously updated with user submissions. Currently updates are performed every two weeks. However, it was not updated at all during the period from about March 2003 to April 2004. Prior to March 2003 crawls were made with declining regularity. As of July 2004 CiteSeer has been continuously crawling the web to find new content using user submissions, conference, and journal URLs as entry points.

In our analysis, we used a CiteSeer dataset consisting of 575,068 entries. From this we extracted a dataset of 325,046 papers that specified the year of publication and the number of authors. Once again, only papers published between

[2] This remains true, despite the recent improvement of automatic extraction algorithms by use of support vector machines [14].

[3] CiteSeer is not performing a brute force crawl of the web but crawling a set of starting pages to the depth of 4-7

1990 and 2002 were considered. It is also important to note that this dataset only contained entries that CiteSeer acquired by parsing the actual document on the Web, i.e. documents that were only cited but not actually parsed, were not included. We assume that the number of parsing errors is independent of the number of authors and does not introduce any new bias.

CiteSeer may be considered a form of self-selected on-line survey - authors may choose to upload the URL where their publications are available for subsequent crawling by CiteSeer. This self-selection introduces a bias in the CiteSeer database that we discuss later. A fully automatic scientometric system is also potentially susceptible to “shilling” attacks, i.e. authors trying to alter their citation ranking by, for example, submitting fake papers citing their work. This later issue is not discussed further here, but appears related to similar problems encountered by recommender systems [20].

Accumulation of papers per year. In order to compare the two databases, we first examined the number of publications in the two datasets for the years 1990 through 2002. These years were chosen to ensure that a sufficient number of papers per year is available in both datasets.

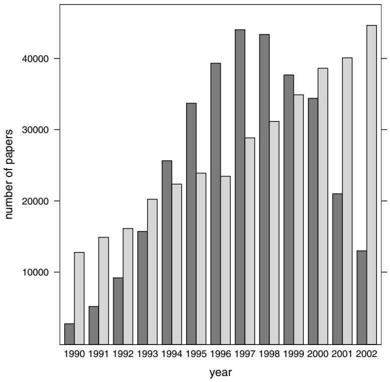


Fig. 1. Number of papers published in the years from 1990 to 2002 present in the DBLP (light) and CiteSeer (dark) databases

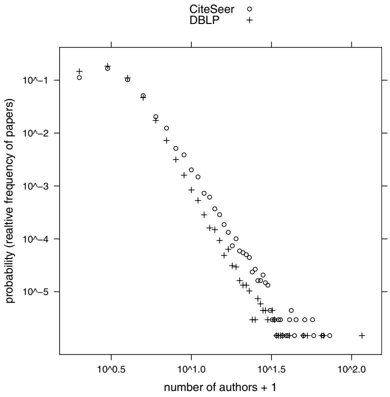
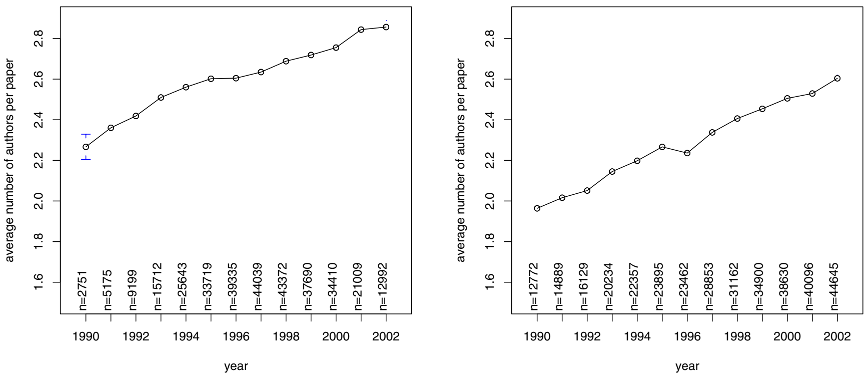


Fig. 2. Probability histogram of number of authors. (double logarithmic scale)

Figure 1 shows a considerable difference in the number of papers present in the two databases on an annual basis.

The increase in the papers per year exhibited by DBLP is probably explained by a combination of (i) the increasing number of publications each year [27,24] and (ii) an increase in the coverage of DBLP thanks to additional funding and improvement in processing efficiency[4].

[4] Personal communication with Michael Ley



(a) CiteSeer

(b) DBLP

Fig. 3. Average number of authors per paper for the years 1990 to 2002

The decrease in the number of papers per year exhibited by CiteSeer since 1997 is mainly due to (i) declining maintenance, although (ii) declining coverage (iii) intellectual property concerns (iv) dark matter effect [9] (v) end of web fever and (vi) specifics of submission process, may also have contributed.

Team size. We also examined the average number of authors for papers published between 1990 and 2002, see Figure 3. In both datasets, the average is seen to be rising. It is uncertain what is causing this rise in multi-authorship. Possible explanations include (i) funding agencies preference to fund collaborative research and/or (ii) collaboration has become easier with the increasing use of email and the Web. We observe that the CiteSeer database contains a higher number of multi-author papers.

Bias in number of authors. Figure 2 examines the relative frequency of n-authored papers in the two datasets. Note that the data is on a log-log scale. It is clear that CiteSeer has far fewer single and two authored papers. In fact, CiteSeer has relatively fewer papers published by one to three authors. This is emphasized in Figure 4 in which we plot the ratio of the frequency of n-authored papers in DBLP and CiteSeer for one to fifty authors. Here we see the frequency of single-authored papers in CiteSeer is only 77% of that ocurring in DBLP. As the number of authors increases, the ratio decreases since CiteSeer has a higher frequency of n-authored papers for $n > 3$. For $n > 30$, the ratio is somewhat random, reflecting the scarcity of data in this region. We therefore limit our analysis to numbers of authors where there are at least 100 papers in each dataset. This restricts the number of authors to less than 17.

As we see in Figure 2 the number of authors follows a power law corresponding to a line with slope approximately -0.23 for DBLP and -0.24 for CiteSeer. There is an obvious cut-off from the power law for papers with low number of authors. For CiteSeer, we hypothesize that (i) papers with more authors are more likely to be submitted to CiteSeer and (ii) papers with more authors appear on more homepages and are therefore more likely to be found by the crawler. These ideas are modeled in Section 2.2.

However none of these factors is relevant to DBLP, which also exhibits a similar drop off in single-authored papers. Other explanations may be that (i) single author papers are less likely to be finished and published, (ii) funding agencies encourage collaborative and therefore multi-authored research and (iii) it is an effect of limited number of scientists in the world [19].

2.2 DBLP and CiteSeer Data Acquisition Models

To explain the apparent bias of CiteSeer towards papers with larger numbers of authors, we develop two possible models for the acquisition of papers within CiteSeer. We also provide a simple acquisition model for DBLP.

The first CiteSeer model is based on authors submitting their papers directly to the database. The second CiteSeer model assumes that the papers are obtained by a crawl of the Web. We show that in fact, both models are equivalent.

To begin, let citeseer(i) be the number of papers in CiteSeer with i authors, dblp(i) the number of papers in DBLP with i authors and all(i) the number of papers with i authors published in all Computer Science.

For DBLP, we assume a simple paper acquisition model such that there is a probability α that a paper is included in DBLP and that this probability is independent of the number of authors.

For CiteSeer we assume that the acquisition method introduces a bias such that the probability, p(i) that a paper is included in CiteSeer is a function of number of authors of that paper. That is,

$$\mathrm{dblp}(i) = \alpha \cdot \mathrm{all}(i) \tag{1}$$

$$\mathrm{citeseer}(i) = \mathrm{p}(i) \cdot \mathrm{all}(i) = p(i) \cdot \frac{\mathrm{dblp}(i)}{\alpha} \tag{2}$$

CiteSeer Submission Model. Let $\beta \in (0,1)$ be the probability that an author submits a paper directly to CiteSeer then $p(i) = 1 - (1-\beta)^i$ where $(1-\beta)^i$ is the probability that none of the i authors submit their paper to CiteSeer.

Substituting to (2) and re-arranging, we have

$$\mathrm{r}(i) = \frac{\mathrm{dblp}(i)}{\mathrm{citeseer}(i)} = \frac{\alpha}{(1-(1-\beta)^i)} \tag{3}$$

It is clear from Equation 3 that as the number of authors, i, increases, the ratio, r(i), tends to α, i.e. we expect that the number of i-authored papers in CiteSeer will approach all(i) and thus from Equation 1 the ratio tends to α. For single authored papers, i.e. $i = 1$, we have that $\mathrm{r}(1) = \frac{\alpha}{\beta}$ and since we know that DBLP has more single-authored papers, it must be the case that $\beta < \alpha$. More generally, we expect the ratio, r(i), to monotonically decrease with the number of authors,

i, reaching an asymptote of α for large i. This is approximately observed in Figure 4, ignoring points for $n > 30$ for which there is a scarcity of data.

In Figure 5 we plot the proportion $r(i)$ for numbers of authors i where we have at least 100 papers available. We fit Equation 3 to the data in Figure 5[5]. We see the fit is not perfect suggesting that this is not the only mechanism in play.

The value to which the data points are converging for high numbers of authors is $\alpha \approx 0.3$. We have to take into account that we only used 71% of DBLP papers and 57% of CiteSeer papers in our analysis – the papers that have both year and number of authors specified. Substituting α into (4) we get the value of $\alpha' \approx 0.24$. If our model is correct, this would suggest that the DBLP database covers approximately 24% of the entire Computer Science literature.

$$\alpha' = \frac{\text{complete\_dblp}(i)}{\text{complete\_citeseer}(i)} = \frac{0.57}{0.71} \cdot \frac{\text{dblp}(i)}{\text{citeseer}(i)} = 0.8 \cdot \alpha \tag{4}$$

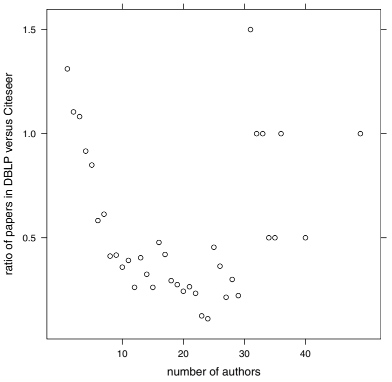


Fig. 4. Ratio of the number of authors in DBLP to CiteSeer as a function of the number of authors of a paper

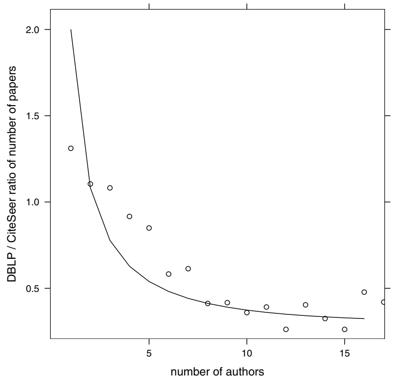


Fig. 5. Fit of model (3) for values of $\alpha = 0.3$ and $\beta = 0.15$ for numbers of authors where there are at least 100 documents in both datasets in total

CiteSeer Crawler Model. CiteSeer not only acquires papers based on direct submission by authors, but also by a crawl of the Web.

To begin, let $\delta \in (0, 1)$ be the probability that an author puts a paper on a web site (homepage for example). Then the average number of copies of an n-authored paper on the Web is $n \cdot \delta$. Let us further assume that the crawler finds each available on-line copy with a probability γ. If $\text{pp}(i, c)$ denotes the probability that there will be c copies of an an i-authored paper published on-line, then we have:

[5] Note that this is the same data as Figure 4 but restricted to $n < 17$.

authors	pp(i,c)	
1	$\mathrm{pp}(1,1) = \delta$	1 copy online
	$\mathrm{pp}(1,0) = 1 - \delta$	0 copies online
2	$\mathrm{pp}(2,2) = \delta^2$	2 copies online
	$\mathrm{pp}(2,1) = 2\delta(1 - \delta)$	1 copy online
	$\mathrm{pp}(2,0) = (1 - \delta)^2$	0 copies online
$\vdots$		
n	$\mathrm{pp}(n,c) = \binom{n}{c}\delta^c(1-\delta)^{n-c}$	c copies online of an n-authored paper

The probability, pf(c), of crawling a document with c copies online, is

$$\mathrm{pf}(c) = 1 - (1 - \gamma)^c \tag{5}$$

thus the probability that CiteSeer will crawl an n-authored document, p(n) is

$$\begin{aligned}
\mathrm{p}(n) &= \sum_{c=0}^{n} \mathrm{pp}(n,c)\,\mathrm{pf}(c) \\
&= \sum_{c=0}^{n} \mathrm{pp}(n,c)(1 - (1-\gamma)^c) \\
&= \sum_{c=0}^{n} \left(\binom{n}{c}\delta^c(1-\delta)^{n-c}\right)(1 - (1-\gamma)^c) \\
&= 1 - \sum_{c=0}^{n} \left(\binom{n}{c}\delta^c(1-\delta)^{n-c}\right)(1-\gamma)^c \qquad \text{(sum of probabilities equals 1)} \\
&= 1 - \sum_{c=0}^{n} \left(\binom{n}{c}((1-\gamma)\delta)^c(1-\delta)^{n-c}\right) \\
&= 1 - (\delta(1-\gamma) + (1-\delta))^n \qquad \text{(from binomial theorem)} \\
&= 1 - (\delta - \gamma\delta + 1 - \delta)^n \\
&= 1 - (1 - \gamma\delta)^n \qquad (6)
\end{aligned}$$

where $(1 - \gamma\delta)^n$ is the probability that no copy of an n-author paper is found by CiteSeer.

Once again, if we substitute Equation (6) in 2, we have

$$\mathrm{r}(i) = \frac{\mathrm{dblp}(i)}{\mathrm{citeseer}(i)} = \frac{\alpha}{(1 - (1 - \gamma\delta)^i)} \tag{7}$$

which is equivalent to the "submission" model of Equation 3. That is, both models lead to the same bias.

3 Prior Work

There has been considerable work in the area of citation analysis and a comprehensive review is outside of the scope of this paper. Broadly, prior citation analysis has examined a wide variety of factors including (i) the distribution of citation rates [28,23,12,19], (ii) the variation in the distribution of citation rates across research fields and geographical regions [23,15], (iii) the geographic distribution of highly cited scientists [10,11] (iv) various indicators of the scientific performance of countries [25] (v) citation biases and miscitations [17,18,29] (vi) collaboration networks [26] (vii) distribution of references in papers [30], and (viii) visualization and navigation [16,13].

The number of citations is the most widely used measure of academic performance and as such it influences decisions about distribution of financial subsidies. The study of citation distributions helps us understand the mechanics behind citations and objectively compare scientific performance.

With regard to the distribution of citations, Laherrere *et al* [19] argued that a stretched exponential[6] is suitable for modeling citation distributions as it is based on multiplicative processes and does not imply an unlimited number of authors. Redner [28] then analyzed the ISI and Physical Review databases and showed that the number of citations of highly cited papers follows a power-law. Lehmann [23] attempted to fit both a power law and stretched exponential to the citation distribution of 281,717 papers in the SPIRES [7] database and showed it is impossible to discriminate between the two models.

So far most of the research on citation distributions has come from the Physics community. Surprisingly little work has been done on computer science papers. The ISI dataset contains computer science papers but these were usually studied together with other disciplines despite the fact that their dynamics may differ. The only work the authors are aware of [26] is based on a small dataset (13000 papers) and was concerned with the distribution of the number of collaborators.

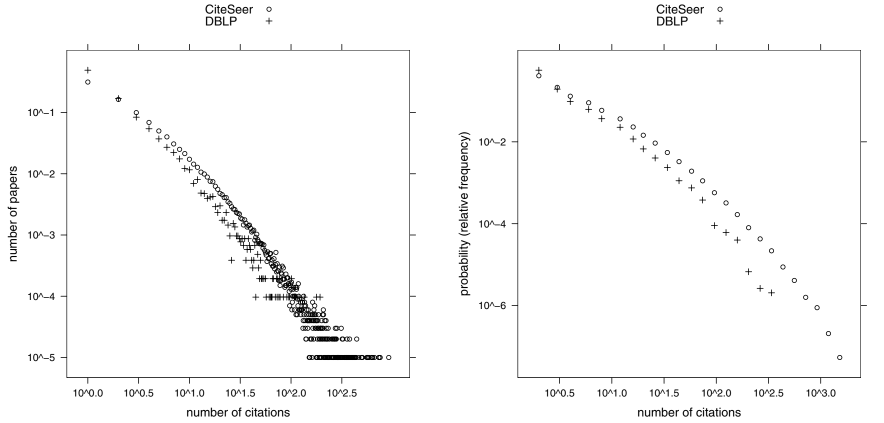
In the next section we examine the distribution of citations in both the CiteSeer and DBLP datasets.

3.1 Citation Distributions for Computer Science

Citation linking in DBLP was a one-time project performed as a part of the 'ACM SIGMOD Anthology' - a CD/DVD publication. The citations were entered manually by students paid by ACM SIGMOD. As a result DBLP now contains a significant number of new papers that have not been included in this effort. To mitigate against this distortion, we limit ourselves in both datasets to papers that have been cited at least once (CiteSeer 100,059 papers, DBLP: 10,340 papers).

Figure 6(a) compares citation distributions in CiteSeer versus DBLP. We see that DBLP contains more low cited papers than CiteSeer. We currently do not have an explanation for this phenomenon. However, it may be related to Lawrence's [21] observation that articles freely available online are more highly cited.

[6] Stretched exponential distribution has the form $exp(-(x/w)^c)$



(a) atomic histogram (b) histograms after exponentialbinning

Fig. 6. Probability histograms on double logarithmic scales for number of citations in the two datasets

Table 1. Slopes for Figure 6(b) representing the parameter of the corresponding power-laws

number	slope		
of citations	Lehmann	CiteSeer	DBLP
< 50	-1.29	-1.504	-1.876
> 50	-2.32	-3.074	-3.509

We use exponential binning (Figure 6(b)) to estimate the parameters of the citation distribution in CiteSeer and DBLP. Exponential binning is a technique where the data are aggregated in exponentially increasing ‘bins’. In this manner we obtain a higher number of samples in each bin, which reduces the noise in the data.

The slopes in Table 1 correspond to linear interpolation of exponentially binned data as displayed in Figure 6(b). Higher slopes in our datasets indicate a more uneven distribution of citations. The papers in each dataset have been divided into two groups – papers with more than and less than 50 citations.

For both datasets we obtain parameters bigger in absolute value than Lehmann [23] derived for Physics. This means that highly cited papers acquire a larger share of citations in Computer Science than in Physics.However, there is also a significant difference between CiteSeer and DBLP.

4 Conclusions

This paper compared two popular online science citation databases, DBLP and CiteSeer, which have very different methods of data acquisition. We showed that autonomous acquisition by web crawling, (CiteSeer), introduces a significant bias against papers with low number of authors (less than 4). Single author papers appear to be disadvantaged with regard to the CiteSeer acquisition method. As such, single authors, (who care) will need more actively submit their papers to CiteSeer if this bias is to be reduced.

We attempted to model this bias by constructing two probabilistic models for paper acquisition in CiteSeer. The first model assumes the probability that a paper will be submitted is proportional to the number of authors of the paper. The second model assumes that the probability of crawling a paper is proportional to the number of online copies of the paper and that the number of online copies is again proportional to the number of authors. Both models are equivalent and permit us to estimate that the coverage of DBLP is approximately 24% of the entire Computer Science literature.

We then examined the citation distributions for both CiteSeer and DBLP and observed that CiteSeer has a fewer number of low-cited papers. The citation distributions were compared with prior work by Lehmann [23], who examined datasets from the Physics community. While the CiteSeer and DBLP distributions are different, both datasets exhibit steeper slopes than SPIRES HEP dataset, indicating that highly cited papers in Computer Science receive a larger citation share than in Physics.

Acknowledgments

The authors thank Michael Ley for his assistance in understanding DBLP.

References

1. Arxiv e-print archive, `http://arxiv.org/`.
2. Compuscience database, `http://www.zblmath.fiz-karlsruhe.de/COMP/quick.htm`.
3. Corr, `http://xxx.lanl.gov/archive/cs/`.
4. Cs bibtex database, `http://liinwww.ira.uka.de/bibliography/`.
5. Dblp, `http://dblp.uni-trier.de/`.
6. Scientific citation index, http://www.isinet.com/products/citation/sci/.
7. Spires high energy physics literature database,
   http://www.slac.stanford.edu/spires/hep/.
8. Sciencedirect digital library, http://www.sciencedirect.com, 2003.
9. P. Bailey, N. Craswell, and D. Hawking. Dark matter on the web. In *Poster Proceedings of 9th International World Wide Web Conference*. ACM Press, 2000.
10. M. Batty. Citation geography: It's about location. *The Scientist*, 17(16), 2003.
11. M. Batty. The geography of scientific citation. *Environment and Planning A*, 35:761–770, 2003.
12. T. C and de Albuquerque MP. Are citations of scientific papers a case of nonextensivity?, 2000.
13. D. Cosley, S. Lawrence, and D. M. Pennock. REFEREE: An open framework for practical testing of recommender systems using researchindex. In *28th International Conference on Very Large Databases, VLDB 2002*, Hong Kong, August 20–23 2002.

14. H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines, 2003.
15. M.-J. Kim. Comparative study of citations from papers by korean scientists and their journal attributes, 1998.
16. S. Klink, M. Ley, E. Rabbidge, P. Reuther, B. Walter, and A. Weber. Browsing and visualizing digital bibliographic data, 2004.
17. J. S. Kotiaho. Papers vanish in mis-citation black hole, 1999.
18. J. S. Kotiaho. Unfamiliar citations breed mistakes, 1999.
19. J. Laherrre and D. Sornette. Stretched exponential distributions in nature and economy: 'fat tails' with characteristic scales. *The European Physical Journal B - Condensed Matter*, 2(4):525–539, 1998.
20. S. K. Lam and J. Riedl. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*, pages 393–402. ACM Press, 2004.
21. S. Lawrence. Online or invisible? *Nature*, 411(6837):521, 2001.
22. S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.
23. S. Lehmann, B. Lautrup, and A. D. Jackson. Citation networks in high energy physics. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 68(2):026113, 2003.
24. L. M. The dblp computer science bibliography: Evolution, research issues, perspectives, 2002.
25. R. M. May. The scientific wealth of nations. *Science 275 793 795*, 1997.
26. M. E. J. Newman. The structure of scientific collaboration networks, 2000.
27. D. D. S. Price. Price, d. de solla, little science, big science, columbia univ. press, new york, 1963., 1963.
28. S. Redner. How popular is your paper? an empirical study of the citation distribution. *European Physics Journal B 4 131 134*, 1998.
29. M. Simkin and V. Roychowdhury. Read before you cite!, 2002.
30. A. Vazquez. Statistics of citation networks, 2001.