

Scholarly Communication in Transition: Evidence for the Rise of a Two-Tier System

Dangzhi Zhao

School of Library and Information Studies, University of Alberta
Edmonton, Alberta, Canada T6G 2J4
dzhao@ualberta.ca

Abstract

This paper presents major findings from a research project that attempts to systematically compare Web-based and print-journal-based scholarly communication, highlighting some considerable differences between the two formats of communication, which provide evidence for the rise of a “two-tier” scholarly communication system. Implications for Open Access are also discussed.

1 Introduction

Scholarly communication is increasingly being conducted over the Internet – by exchanging email, participating in online discussion boards, and publishing papers on the Web or in electronic repositories, to name just a few formats. As a result of the changing structures and processes, we have seen renewed interest in the study of scholarly communication systems (Borgman, 2000). People have started to rethink the roles of various players in this system such as scholars, publishers and libraries. New scholarly communication models are being proposed which are becoming feasible with new technologies and which may be more efficient than traditional ones, such as the Open Access movement which promotes self-archiving of refereed papers and a disaggregated model in which institutional repositories are a central component (Chan, 2004; Crow, 2002; Lynch, 2003).

However, the experience of the Open Access movement and a low participation rate in institutional repositories clearly indicate how persistent the traditional system is and how difficult it is to impose a new system. In order to envision and promote an effective future system, we probably need to first really understand how the print-based scholarly communication system evolved (Nentwich, 2001), the types of communication that are currently taking place, and the similarities and differences between the new and the well-established formats.

We have conducted research that attempts to systematically compare Web-based and print-journal-based scholarly communication. This paper summarizes major findings of this project, highlighting some considerable differences between the two formats of communication, and finding evidence for an evolving “two-tier” scholarly communication system. Implications of these findings for Open Access are also discussed.

2 A comparative study of Web-based and print-journal-based scholarly communication

Before we summarize major findings of the study, we present here a brief description of the study to provide necessary background and context for the findings. Details can be found in Zhao (2003).

This study explored the opportunities opened up by increasingly available new data sources and tools for scholarly communication through an author citation analysis of scholarly communication patterns in the eXtensible Markup Language (XML) research field. It used data both from the Web as indexed by *CiteSeer* (<http://citeseer.ist.psu.edu/>) and from print journals as indexed by the Institute for Scientific Information (ISI)'s *Science Citation Index (SCI)*. A series of citation analyses including author visibility analysis and author co-citation analysis were conducted, and comparisons of results from the two data sources were carried out controlling for data scope and citation counting method respectively, to identify the similarities and differences between Web-based and print journal-based scholarly communication as revealed by citation analysis and to study the capacity of scientific papers published on the Web along with existing citation indexing tools for Web publications to serve as an alternative to the ISI databases as a data source for citation analysis studies.

Meanwhile, publications and characteristics of three groups of highly visible authors were examined and compared — authors highly visible both on the Web and in journals, those only in journals, and those only on the Web, to ascertain possible contributing factors to differences between the two communication formats.

Informed by Merton's normative view of science (Griffith, 1990; Cronin, 1984), which sees science, of which citing is a part, as a social activity governed by a set of norms such as “universalism”, “communism”, “disinterestedness”, and “organized skepticism” (Merton, 1942), citation analysis has proven to be a reliable and valid approach to the study of formal scholarly communication. This approach has been applied to such issues as characteristics and evolution of scholarly communities, evaluation of scholarly contributions, and diffusion of ideas. The past several decades have seen a large number of citation analysis studies of various research fields, from natural sciences to social sciences and humanities. Citation analysis results have been used widely in scientific evaluation for purposes such as tenure and promotion in academics (Borgman & Furner, 2002).

The ISI's citation indexes have been used as the data source for most of the citation analysis studies reported in the literature to date. They have contributed significantly to the wide application of a citation analysis approach in various studies and in scientific evaluation, but have also drawn considerable criticism especially when applied to the evaluation of scholars (Smith, 1981). Two major problems with these indexes are their limited and biased coverage and their “first authors only” approach to multiple authorship of cited papers — they only cover journals selected by the ISI's advisory boards of experts in each topic represented (ISI, 2004), and only index first authors of cited papers. As a result, they do not well support citation analysis studies that take into account scholars' contributions as co-authors or the perceptions of authors of other types of publications such as conference papers and technical reports. These limitations and biases have become increasingly serious as new formats of publication are emerging and collaboration has become the norm rather than exception in science.

The Internet as a powerful communication medium is changing the way information is produced, exchanged and used. Full text research papers and the corresponding search tools are becoming increasingly available on the Internet; examples include *CiteSeer* and *CiteBase* (<http://citebase.eprints.org>). These citation indexes are different from those for print journals such as the ISI databases in that, among others, papers they index use the Web as their communication medium, which affords higher speeds of communication and wider distribution of information than the journal; they cover a wider range of document types such as degree theses, technical reports, conference papers, and preprints in addition to journal articles, which represent different stages in the scholarly communication process; they contain more information about cited documents such as all authors, full titles and full names of journals or conferences, unlike the limited information provided by the ISI databases; and their source paper selection and indexing process is automatic and inclusive as opposed to manual and highly selective as in the ISI databases.

While these tools may have their own problems (Zhao & Strotmann, 2004), they do not have the problems of the ISI databases as listed above, and therefore open up opportunities for a larger variety of inquiries and more sophisticated measures of research impact and intellectual relationships (Zhao, 2003; Zhao & Logan, 2002), some of which have been proposed but not widely practised, partly due to the lack of support by available citation indexes. Consequently, more balanced and less biased results may be produced from these tools.

We explored the usefulness of one of these tools, namely the NEC Research Institute's *CiteSeer*¹ as compared with the ISI's *SCI*. *SCI* was originally designed for print journals, and the majority of journals covered by *SCI* nowadays are still print-based (in print format or having a print version), although it now also selectively indexes e-journals (ISI 2004). *CiteSeer* is a *SCI*-like tool freely available on the Web. It automatically indexes research papers of any type (journal articles, technical reports, conference papers, etc.) that is in the broadly defined computer science field and publicly available on the Web. More information about *CiteSeer* can be found in Goodrum et al (2001), Lawrence et al (1999), and Zhao & Strotmann (2004).

3 Findings from comparing results from different citation counting methods

Unlike *SCI*, *CiteSeer* indexes all authors of cited papers, which supports various citation counting methods. We took advantage of this feature and used three different citation counting methods to rank and map authors cited by papers from *CiteSeer*: straight counts by which only the first author's citation count increases by 1 when a

¹ Now a joint effort of NEC and the School of Information Science and Technology at Pennsylvania State University.

paper with N authors is cited, as well as complete counts and fractional counts by which the citation count of each of the N authors increases by 1 and $1/N$ respectively when a paper with N authors is cited².

3.1 Low correlation between author rankings by straight counts and other methods

It was found that, if we select the 100 most highly visible authors, only about one half of the selected authors are the same no matter which of the three counting methods we use, and about three quarters are the same whether we use complete or fractional counts. In terms of correlations, Pearson's r 's were calculated for the 45 authors who were common to all three lists of top ranked 100 authors³. The value is 0.917 between author rankings by fractional and complete counts, 0.654 between those obtained by fractional and straight counts, and 0.648 between those by straight and complete counts. Clearly, author rankings by fractional and complete counts are highly correlated, but are significantly different from the author ranking by straight counts.

3.2 Clearer picture of intellectual structure when counting more than first authors

Four identical specialties within the XML research field were identified by factor analysis both when counting first authors only (first-author co-citation analysis) and when taking into account the first five authors of cited papers (all-author co-citation analysis). However, these author groups were quite clear-cut on the multi-dimensional scaling (MDS) map generated from all-author co-citation analysis, but overlapped considerably on the MDS map produced via first-author co-citation analysis. In addition to these four major specialties common to both sets of results, first-author co-citation analysis identified some other areas of research.

The considerably clearer picture produced from all-author co-citation analysis is probably due to the co-citation counting method used. All-author co-citation takes into account co-authorship which, in a sense, is an even closer relationship between scholars than co-citation, and it counts co-citations received by scholars as co-authors. Thus, more links between scholars are considered in all-author co-citation analysis. As a result, related authors tend to get higher co-citation counts in all-author co-citation than in first-author co-citation analysis, tying authors within a group closer and pulling authors in different groups farther away from each other, resulting in a clearer picture.

4 Findings from comparing citation analysis results from two data sources

4.1 Small number of publications shared by the two data sources

A search conducted on Dec. 18, 2001 on "XML" or "eXtensible Markup Language" resulted in 312 distinct papers using *CiteSeer* and 374 using *SCI*. Among these, 26 were common to both *CiteSeer* and *SCI*, a very low percentage of citing papers shared by the two data sources. Clearly, in this research field, journal articles were not often made available on the Web and papers published on the Web were not well represented in *SCI*.

4.2 High correlation between author rankings by the same citation counting method

Among the highly cited 100 authors from each of the two data sources, *CiteSeer* and *SCI*, only about one half were common to both data sources. However, the correlation between the rankings resulting from the two different data sources was high when the same citation counting method, namely straight counts, was used: the Pearson's r was 0.92. The high correlation suggests that citation analysis using *CiteSeer* as a data source is just as valid in the evaluation of scholars as is citation analysis based on *SCI* data, the data source most widely used in the literature so far and widely validated in evaluation studies. That means that publications on the Web should no longer be ignored as a data source for the study of scholarly communication because they are similar to those in print journals in terms of the way they refer to earlier publications. If this can be confirmed even more strongly in the future, researchers who either might not have easy access to *SCI* data or who might investigate research areas or populations that are under-represented in *SCI* would then be able to conduct citation analysis studies using data and tools freely available on the Web and still get valid results (Zhao, 2005).

² We used a simplified version of fractional and complete counts. See Zhao (2003) for details.

³ The number 100 was used as a guideline rather than a cut-off point in selecting top ranked authors. See Zhao (2003) for details.

4.3 Cutting-edge research on the Web, and mature research areas in journals

Many of the authors in *the Semantic Web* area were ranked much higher by number of citations in *CiteSeer* than in *SCI*. Examples include Lassila (24.5 in CS vs. 35.5 in SCI), Brickley (31 vs. 53), and Fensel (34 vs. 46). This suggests that research in *the Semantic Web* area is better represented on the Web. Since this area of research was emerging, aiming to develop “the next generation” of the Web, this is evidence that research reported on the Web may be more cutting-edge than that reported in journals.

This can also be seen from the other direction. Authors in application areas of XML, such as Chemical Markup Language (CML) and XML for medical information exchange, were not as well represented in *CiteSeer* as in *SCI*. From the point of view of XML research, unlike the Semantic Web, applications are about relatively mature rather than cutting-edge technologies. Similarly, some of the founding figures in the XML field such as Bosak, Goldfarb and Berners-Lee were highly cited in *SCI* but less so in *CiteSeer*. Perhaps papers in journals were referring to a considerable extent to foundational and historical materials or to opinion papers and might therefore contain more reviews and research at earlier stages.

We can also see here that citation analysis using either *SCI* or *CiteSeer* exclusively may be biased by leaving out certain research areas — those in which scholars publish heavily in venues other than journals in the case of *SCI*, or those where Web publishing is not widely accepted and practised in the case of *CiteSeer*.

4.4 Different citing behaviour demonstrated in the two data sources

In both data sources, about forty percent of authors analysed were placed in the research area *XML data management*. Authors working in this area form a single group in the results from *SCI* while, in those from *CiteSeer*, this research area splits into two. One possible explanation we have explored (Zhao, 2004) is that the intellectual difference between the two groups is blurred in the print world by such factors as “diplomatic citing.” As Edge (1979, p. 120) observed, “adding a list of references to a paper is often a last-minute chore: colleagues, ‘trusted assessors’, referees and editors all contribute suggestions as to authors and papers that ‘ought’ to be included somewhere.” These citations are usually not among the core documents that directly contributed to the writing of a citing paper, and would therefore widen its intellectual scope from the perspective of citation analysis. Since editors or referees often come into play after drafts have been published on the Web, this type of citations would occur more frequently in the print world, which may have pulled the two database groups together. This appears to be supported by our data: reference lists in print journals were at an average about 20% longer than those on the Web (18.1 vs. 14.7 references per paper).

5 Discussion

5.1 Evidence for the rise of a two-tier scholarly communication system

Above findings suggest that research published on the Web is perhaps more at a research front than that in print journals in the XML research field. This appears to provide evidence for a “two-tier system” in scholarly communication that is believed by some scholars to be a future model of the scholarly communication system (van Raan, 2001). In this model, the first tier is predicted to be a “free space” which represents the scholarly enterprise in “real time” and is likely to feature free and timely Web-based publications, while the second tier is thought to be the world of more formal publications that is likely to continue to be dominated by journals (van Raan, 2001, p. 61). As suggested by the present study and others (Chan, 2004; Crow, 2002; Zhao, 2004), the first tier would primarily serve as an information distribution medium that improves the effectiveness and efficiency of informal communication, on which scholars rely to obtain the information they need for their research, while the second tier would primarily serve as an archive and evaluation rather than information distribution device. The faster and wider distribution of information on the Web makes it a perfect medium for initial publication of new research results in the first tier, while the journal has served well as an archive and evaluation device for a long time, which makes it natural to continue its role in the second tier.

As we concluded in earlier studies (Zhao, 2004; Zhao, 2005), if this system evolves, journals that currently do not accept papers published on the Web may have to change their policies, and all journals may eventually

implement procedures and policies to build on rather than conflict with Web publications. This may improve the efficiency of scholarly communication significantly.

5.2 Importance of citation analysis studies based on new data and tools

Different research foci were represented in different publishing media, and different results were obtained using different citation counting methods. In addition, as a “two-tier system” evolves, the study of scholarly communication patterns demonstrated in the first tier becomes increasingly important, and it becomes a more serious problem to use the “journal only” ISI databases as the only data source for the evaluation of scholars and the examination of intellectual structures using citation analysis.

These findings challenge the common practice in science evaluation and other types of citation analysis studies in which the ISI databases are used exclusively to obtain citation data on performance and other measures, and they clearly indicate that the best way to evaluate scholars and to obtain a more complete and less biased picture of scholarly communication patterns using a citation analysis approach is to combine multiple data sources, especially those that support multiple citation counting methods, including scholarly publications on the Web, in digital libraries and in institutional repositories in addition to the journal articles indexed by the ISI databases.

As shown in the present study, with the newly available data and tools such as *CiteSeer*, a clearer picture of intellectual structures can be obtained, multiple citation counting methods can be applied, and “diplomatic citing” can become less of an issue. Moreover, calculating complete counts or even fractional counts is no costlier than computing straight counts as information on all authors is readily available there. Since the use of straight counts for evaluating scholars is not so much due to the lack of awareness of the associated biases, but largely due to the convenience and low cost of calculating straight counts with the kind of support offered by the ISI databases, it follows that data and tools increasingly available on the Web such as *CiteSeer* may be able to help deal with this dilemma by supporting citation counting methods that result in clearer and less biased pictures of scholarly communication patterns without adding much cost.

5.3 Implications for Open Access

Citation analysis has proven to be a powerful approach to the study of scholarly communication, and has a long history in the study of various issues in scholarly communication. One important application is the evaluation of scholarly contributions. The past several decades have seen citation analysis results being widely used in scientific evaluation for purposes such as tenure and promotion in academics.

However, as collecting citation counts in the print world is nearly impossible without citation indexes, science evaluation based on citation analysis results has been relying heavily on the ISI databases, and consequently has been limited to simple “straight” citation counting. The well-known bias associated with straight counts, the lack of correlation between author rankings by straight counts and other counting methods, and the ability of tools like *CiteSeer* to support better counting methods as demonstrated here leave us no excuse to continue using straight counts when evaluating scholars. Furthermore, full text scholarly publications with reference lists are increasingly available digitally, which makes it possible for citation analysis studies to go beyond what the ISI databases have supported, employing more sophisticated methods and more comprehensive data sources. This may in turn contribute to a more efficient scholarly communication system in which scholarly work is evaluated based on how many users it has reached in all rather than on how many users it has reached who have published in journals indexed by certain databases such as the ISI databases.

This has important implications for open access. The slow move of journals to open access and the low faculty participation rate in institutional repositories indicate that simply promoting the benefits of new formats of scholarly communication is not enough. We may in addition need to bring open access publications into the evaluation system. Including these new formats of scholarly publications in the evaluation of science not only can contribute to obtaining a more balanced and more complete evaluation, but may also serve to promote open access and consequently a more efficient two-tier scholarly communication system. It is well known that the ISI citation indexes have served as the main data source for citation-based science evaluation, and that this has pushed scholars to publish in journals indexed there. If we use full-text open access scholarly publications as data sources for citation-based science evaluation, scholars may become more motivated to make their work available for open access knowing that it is counted in evaluations. If the ISI databases with all their limitations

can influence scholarly communication, open access publications that support more sophisticated measures can do even better. Integrating a citation analysis mechanism into institutional repositories and other open access resources might be a good start.

6 Conclusions

Scholarly communication is increasingly being conducted over the Internet, which has brought both challenges and opportunities to the traditional scholarly communication system. This paper has presented major findings from a research project that sought to systematically compare scholarly communication patterns between the Web and the print world. These findings have provided evidence for the rise of a “two-tier system” of scholarly communication, and have demonstrated the importance and feasibility of citation analysis studies of scholarly communication based on increasingly available open access publications and corresponding search tools. As an implication for open access, we have noted the need to include the new publication formats in the science evaluation system, since scholars may be more willing to publish their work in new formats such as open access journals and institutional repositories if they can rest assured that their work is counted in evaluations, just as scholars now prefer to publish in journals indexed by the ISI databases. We hope that the present study motivates further research to contribute to the transition of the scholarly communication system from one that has evolved with print journals as the centre to one that works best in a networked digital environment.

References

- Borgman, C.L. (2000). Digital libraries and the continuum of scholarly communication. *Journal of Documentation*, 56, 412-430
- Borgman, C.L., & Furner, J. (2002). Scholarly communication and bibliometrics. In *Annual Review of Information Science and Technology*, 36 (pp. 3-72). Medford, NJ: Information Today
- Chan, L. (2004). Supporting and enhancing scholarship in the digital age: the role of open-access institutional repositories. *Canadian Journal of Communication*, 29, 277-300
- Cronin, B. (1984). *The Citation Process: The Role and Significance of Citations in Scientific Communication*. London: Taylor Graham.
- Crow, R. (2002). *The case for institutional repositories: a SPARC position paper*. Retrieved Feb. 8, 2005, from The Scholarly Publishing & Academic Resources Coalition website <http://www.arl.org/sparc/IR/ir.html>
- Edge, D. (1979). Quantitative measures of communication in science: a critical review. *History of Science*, 7, 102-134
- 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, 661-675
- Griffith, B.C. (1990). Understanding science: Studies of communication and information. In C. L. Borgman (ed.), *Scholarly Communication and Bibliometrics* (pp. 31-45). Newbury Park, CA: Sage.
- Institute for Scientific Information (2004). *The ISI Database: the journal selection process*. Retrieved May 30, 2004, from <http://www.isinet.com/essays/selectionofmaterialforcoverage/199701.html/>
- Lawrence, S., Bollacker, K., & Giles, C. L. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6), 67-71
- Lynch, C. A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *ARL: A Bimonthly Report on Research Library Issues and Actions*, 226 (2003): 1-7. Retrieved April 19, 2005, from <http://www.arl.org/newsltr/226/ir.html>
- Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology*, 1, 115-126.
- Nentwich, M. (2001). (Re-)de-commodification in academic knowledge distribution? *Science Studies*, 14, 21-42
- Smith, L. C. (1981). Citation analysis. *Library Trends*, 30, 83-106.
- Van Raan, A. F. J. (2001). Bibliometrics and Internet: some observations and expectations. *Scientometrics*, 50, 59-63
- Zhao, D. (2003). *A comparative citation analysis study of Web-based and print journal-based scholarly communication in the XML research field*. Dissertation, Florida State University
- Zhao, D. (2004). Web-based and print journal-based scholarly communication in the XML research field: a look at the intellectual structure. In *Proceedings of the American Society for Information Science and Technology 2004 Annual Meeting: Managing and Enhancing Information: Cultures and Conflicts*, (pp. 72-83). Medford, NJ: Information Today
- 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. To appear in *Information Processing and Management*
- Zhao, D. & Logan, E. (2002). Citation analysis of scientific publications on the Web: A case study in XML research area. *Scientometrics*, 54, 449-472.
- Zhao, D. & Strotmann, A. (2004). Towards a Problem Solving Environment for Scholarly Communication Research. *Proceedings of the Canadian Association for Information Science 2004 Annual Conference*, June 3-5, 2004, Winnipeg, Canada