Scientific impact quantity and quality: Analysis of two sources of bibliographic data

Richard K. Belew
Cognitive Science Dept.
Univ. California – San Diego
La Jolla CA 92093-0515 USA

10 April 2005

arXiv#: CoRR/0504036

Abstract

Attempts to understand the consequence of any individual scientist's activity within the long-term trajectory of science is one of the most difficult questions within the philosophy of science. Because scientific publications play such as central role in the modern enterprise of science, bibliometric techniques which measure the "impact" of an individual publication as a function of the number of citations it receives from subsequent authors have provided some of the most useful empirical data on this question. Until recently, Thompson/ISI has provided the only source of large-scale "inverted" bibliographic data of the sort required for impact analysis. In the end of 2004, Google introduced a new service, GoogleScholar, making much of this same data available. Here we analyze 203 publications, collectively cited by more than 4000 other publications. We show surprisingly good agreement between data citation counts provided by the two services. Data quality across the systems is analyzed, and potentially useful complementarities between are considered. The additional robustness offered by multiple sources of such data promises to increase the utility of these measurements as open citation protocols and open access increase their impact on electronic scientific publication practices.

1 Background

Bibliometric analysis of scientific publications goes back to at least the 1970s [11, 13, 5]; similar analysis of judicial opinions has been done by Shepards/LexisNexis for more than a hundred years. The Institute for Scientific Information has made an industry of providing citation data to libraries since the mid-1960s; the products are currently available as part of Thomson/ISI (*ISI*). *ISI* reports that they currently index 16,000 journals, books and proceedings [6]. While far from exhaustive (ISI estimates that of the 2000 new journals reviewed annually, only 10% are selected), the service cites "Bradford's Law" that a relatively small number of sources capture the bulk of significant scientific results. All articles appearing in selected publications have their bibliographies manually transcribed, and "inverted bibliographies" pointing from a (earlier) cited work to all (subsequent) citing publications is generated to support users' searches. Critically, the translation of these bibliographies into distinct records involves a great deal of *manual* effort.

May has reported extensive analyses of British scientific activity in comparison with other countries, primarily based on *ISI* 's data [9, 10]. "The database has many shortcomings and biases, but overall it gives a wide coverage of most fields." [10, p. 793] His critique of shortcomings in this data is useful:

> Some problems have to do with the compilation of the database. It includes citations of books and chapters in edited books, but it does not include the citations in such publications. Other publications, such as government and other agency reports and working papers, are essentially omitted. It does not cover all significant scientific journals.... Papers that describe technical methods may attract thousands of reflexive citations, while path-breaking papers may be cited only slightly for many years. Review articles can mask the primary papers they review. Citation patterns vary among fields.... Spectacular scientific errors may attract many citations.... Self-citation (which accounts for at least 10% of all citations) may bias some of the results. [10, Footnote 3]

Some of these issues (e.g., having to do with the sources being compiled) can be expected to altered by new forms of electronic scientific publication, but others (e.g., self-citation) are likely to be more intrinsic to scientific authoring processes. It is for this reason that Google's recent announcement of

their Scholar.Google(beta) (*GoogleScholar*) service is welcome, as a second, independent source of similar data.

While specifics concerning Google's operation are difficult to come by, it is reasonable to assuem that the process relies on more *automatic*, algorithmic procedures than those used by *ISI* . Linkage structure among Web pages is analogous in important ways to scientific publication [4, 8]. These links are captured by Web crawling algorithms as both "citing" pages (i.e., Web pages with HTML anchors pointing to other Web pages) and "cited" pages are visited, a feature exploited by Google's original "PageRank" retrieval algorithm [12]. *GoogleScholar* attempts to bring similar analyses to academic publication, despite the fact that these source documents are often much less accessible.

2 Methods

Given an author's name[1], both *ISI* and *GoogleScholar* provide search facilities that return a list of publications putatively authored by this individual, together with the number of times each of these publications has been cited by other publications discovered by the service. Six academics were selected at random and used as "probe" queries with both systems. [2] Complete bibliographies of all publications by these authors were manually reconciled against 203 references to these publications returned by one or both systems, and then analyzed in detail. Cumulatively, *ISI* discovered 4741 such references, *GoogleScholar* found 4045.

Because standards and format of bibliographic citations vary widely across different publications, the process of reconciling citation strings from different papers to the same target publication is problematic, whether via *ISI* 's manual process or Google's automatic one. It is common, therefore, to find the same publication has been treated as more than one record.[3]

For example, manual inspection reveals that a single publication in the "Proceedings of the 12th Annual Conference of ACM's Special Interest Group in Information Retrieval (SIGIR)" is listed as twelve separate records by *ISI* ; these are shown in Table 1. While most citations to this target publi-

[1] Translation of an author's name into search query string(s) can be ambiguous. In these experiments both first letter, and first letter with the middle initial together with full last name was used as the author's name.

[2] These academics were all drawn from a single, particularly interdisciplinary academic department.

[3] The alternative type of error, where citations to multiple, distinct publications are confounded as part of the citation record of a single entry, is more difficult to identify

PubYear	CiteString	NCitations
1989	12 ANN INT ACM SIGIR	1
1989	12 ANN INT C RES DEV	1
1989	12TH P ANN INT ACM S 11	14
1989	12TH P INT C RES DEV	1
1989	ACM SIGIR INT C RES	1
1988	JUN P ACM SIGIR 88 G 11	1
1989	P 11 INT ACM SIGIR C	1
1989	P 12 ANN INT ACM SIG	2
1989	P 12 ANN INT ACM SIG 11	16
1989	SIGIR 89 11	2
1989	SIGIR FORUM 23 11	1
1990	SIGOIS B 11 48	1

Table 1: Citation variations for same publication

cation have been conveniently collected with respect to two of these records, such noisy data makes impact analysis difficult. In these experiments, a publication's "impact" is defined as the number of citations found to any of the variations resolved to the published work, i.e., the sum is taken across all records (manually) identified as referencing the same publication.

3 Results

Figure 1 shows how well both systems aggregate individual citations that in fact to refer to the same published paper. This shows the cumulative probability that one, two, or more publications listed as distinct to by both systems in fact refer to the same publication. For example, it shows that more than 60% of the articles are represented as unique entries within *ISI* 's listing while 85% of them are unique with *GoogleScholar* . None of the articles had more than five separate listings within *GoogleScholar* , while 13% had five or more entries in *ISI* 's system (e.g., the example shown in Table 1 had 12).

Overlap between the two sources of data was relatively small. Of the 203 citations analyzed, only 78 publications received at least one cited reference from each system. However, for this subset the general pattern of agreement was quite good. Figure 2 shows the number of citations reported by *GoogleScholar* and *ISI* for the subset of 78 publications. Note that the num-

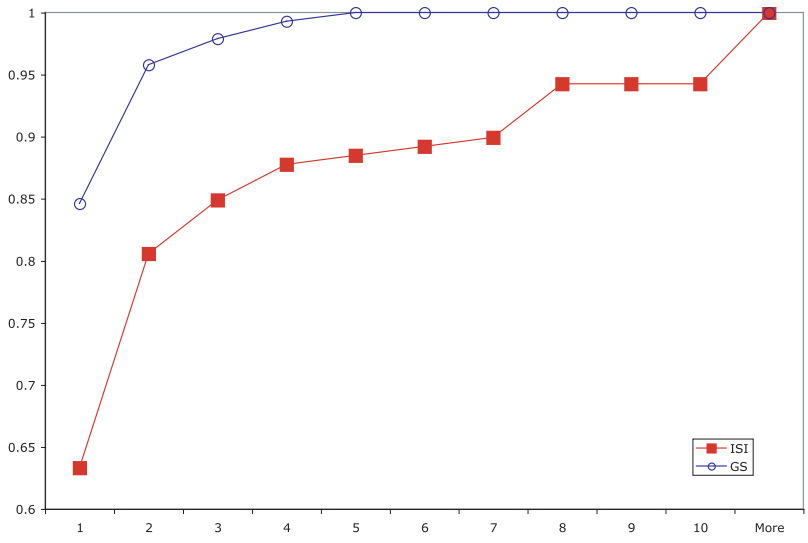

Figure 1: Redundant citation noise

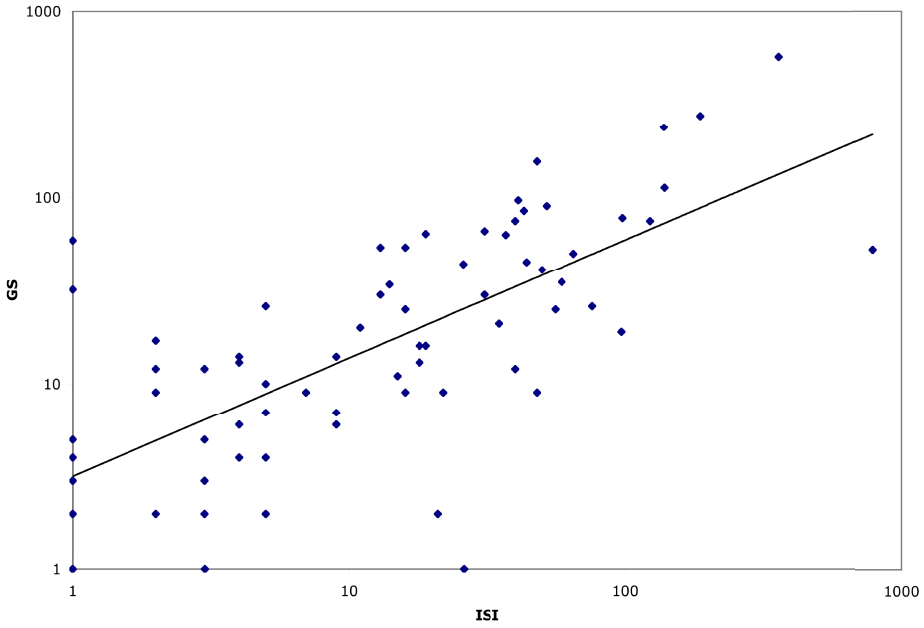

Figure 2: Correlation of *GoogleScholar* and *ISI* citation counts

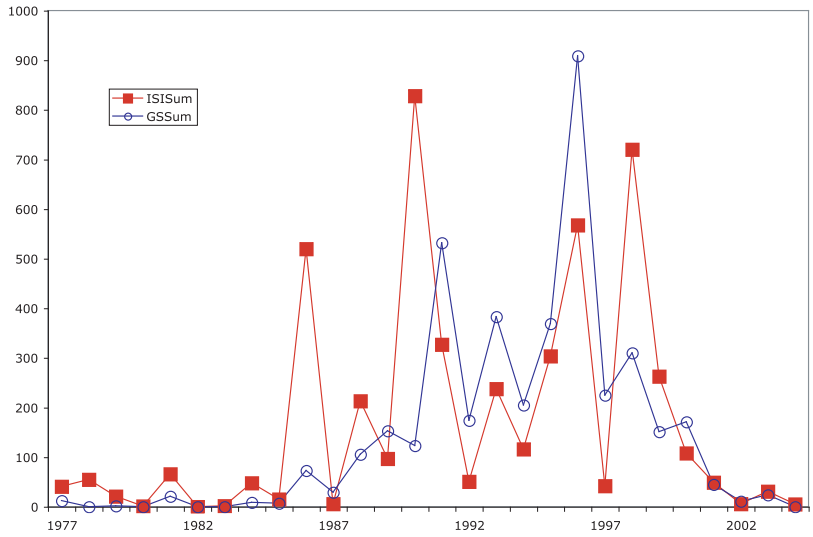

Figure 3: Temporal distribution of citations

ber of citations is plotted on a log-log scale, reflecting the well-known power law distribution of citation reference [14]. Based on this sample, there seems good evidence ($r^2 = 0.5023, t = 8.872, \rho > 0.005$) for a power law relation ($GS = 3.1718 * ISI^{0.6359}$) relating the number of citations reported by the two services.

Figure 3 shows the cumulative number of citations reported by publication year of the cited work. An alternative criterion for considering the match between systems is to define a "miss" to be a publication for which one service has identified three or more citations, but which the other service does not capture whatsoever. Figure 4 shows missing citations, found by one service but not the other, again distributed by publication year. *GoogleScholar* seems competitive in terms of coverage for materials published in the last twenty years; before then *ISI* seems to dominate.

Coverage with respect to the two systems can also be analyzed by other dimensions of the publications, including publication venue and author. Figure 5 aggregates publications into four categories: conference publications,

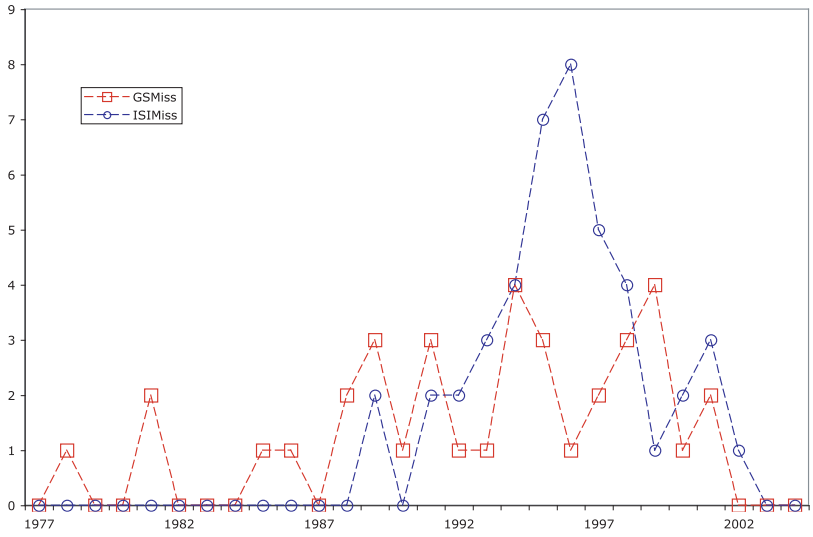

Figure 4: Temporal distribution of missing citations

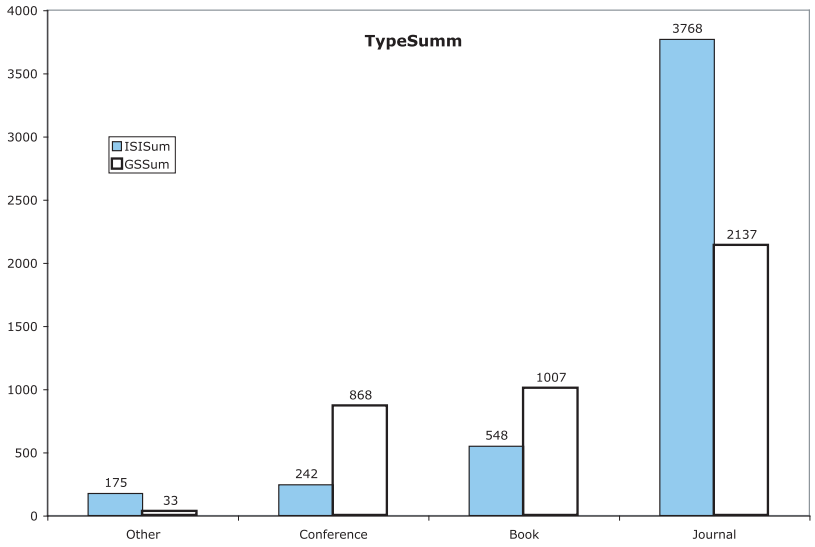

Figure 5: Coverage by publication type

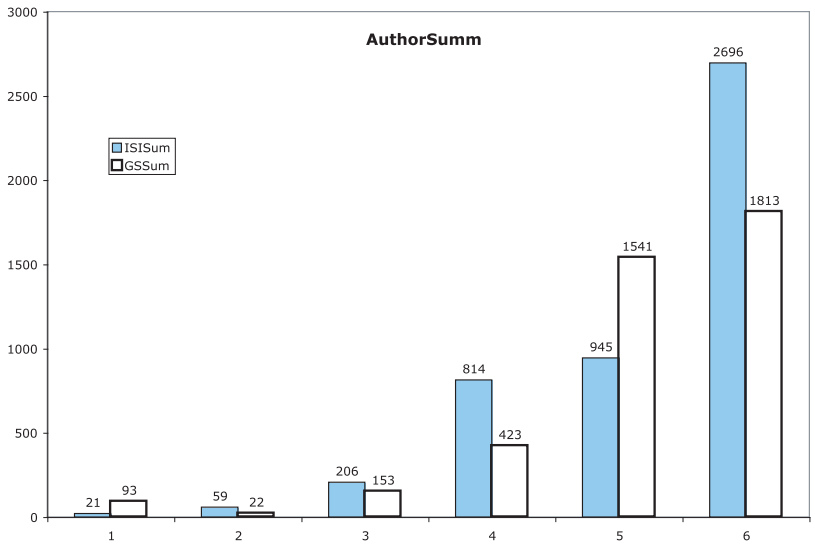

Figure 6: Coverage for individual authors

books (or book chapters), journal articles, and other forms of publications (e.g, technical reports, dissertations, etc.); χ^2 tests confirm the distributions are distinct. Publications in books (as noted by May, above) and conference proceedings are much more likely to be available via *GoogleScholar* ; conversely, journal articles are better indexed via *ISI* . If citations are summarized with respect to the six authors analyzed, Figure 6 shows that some authors are better represented with respect one service as opposed to another. Such variation is to be expected, given that some authors, via the publication venues through which they typically report, will be more or less well-covered by one service or another. Again, χ^2 tests confirm the distributions are distinct.

4 Summary

Evaluating academics' performance, as individuals or as part of larger social groups, in terms of the number of publications they produce is common practice. The ability to quantify their "impact" in terms of the number of other publications that subsequently choose to cite their work arguably provides a more refined and relevant measure. Such data is subject, however, to confounding factors ranging from noise in the process of collating and "inverting" bibliographic references through intrinsic features of scientific publication (e.g., self-citation). The results presented above are therefore reassuring in that new evidence provided by *GoogleScholar* provides the first independent confirmation of impact data previously available only from *ISI* . However, analysis across both systems also shows significant variations with respect to the two dimensions (authorship and publication type) considered; other dimensions of variation are certain to exist. This analysis also revealed some problems common to both systems. For example, both services support only simple ASCII encodings of author names which are likely to lose important character markup (available via Unicode representations) which can be especially problematic for authors with foreign names.

Critically, new services within selected disciplines [1, 2], changing standards regarding exchange of "open citation" information [3], in combination with increased pressure for public access to scientific publications [15], may soon make some operational difficulties associated with impact analysis obsolete. In the interim, academic deans, science policy advisors and anyone else relying on citation count data are cautioned that any individual measurement requires more context. In the longer term, the increased availability of statistics like bibliographic impact makes it increasingly important to understand how publication and citation activities, within both scientific publication and Web publishing more generally, can be included as part of more holistic evaluations of intellectual contribution [7].

References

[1] portal.acm.org. Association for Computing Machinery's Digital Library Portal.

[2] www.computer.org/publications/dlib/. Institute for Electrical Engineering Society's Digital Library.

[3] www.crossref.org. CrossRef is an association of scholarly and professional publishers that cooperate to provide reference links.

[4] R. K. Belew. *Finding Out About: A cognitive perspective on search engine technologies and the WWW*. Cambridge Univ. Press, 2000.

[5] E. Garfield. *Essays of an information scientist*. ISI Press, Philadelphia, PA, 1986.

[6] E. Garfield. Using the impact factor. *Current Contents*, July 18 1994. also available at www.isinet.com/essays/journalcitationreports/8.html/.

[7] J. Grant, R. Cottrell, F. Cluzeau, and G. Fawcett. Evaluating 'payback' on biomedical research from papers cited in clinical guidelines: applied bibliometric study. *BMJ*, 320:1107–1111, 22 Apr 2000.

[8] S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 280(5360):98–100, 1998.

[9] R. M. May. Government funding of research and development. *Science*, 278(5339):878–880, 31 Oct 1997.

[10] R. M. May. The scientific wealth of nations. *Science*, 275(5301):793–796, 7 Feb 1997.

[11] R.K. Merton. *The sociology of science*. Univ. Chicago Press, Chicago, 1973.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. google.stanford.edu/ backrub/pageranksub.ps, 1998.

[13] D. J. de Solla Price. *Little Science, Big Science...and Beyond*. Columbia University Press, 1986.

[14] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B*, 4(2):131–134, 1998.

[15] E. A. Zerhouni. Information access: NIH public access policy. *Science*, 306(5703):1895–, 2004.