Essays of an Information Scientist, Vol:3, p.234-246, 1977-78 Current Contents, #39, p.7-19, September 26, 1977

On the Use of Citations in Studying Scientific Achievements and Communication*

Also see Introduction : p.232 and Reprint by D. Edge on p.240, Essays, Vol:3

Belver C. Griffith, M. Carl Drott
Graduate School of Library Science, Drexel University
Philadelphia, Pennsylvania 19104

Henry G. Small
Institute for Scientific Information
325 Chestnut Street, Philadelphia, Pennsylvania 19106

Rather than trying to reply in kind to some recent, slightly polemical, criticisms[1] of the use of citations, we will discuss the assumptions underlying the use of citations in the study of science. We shall attempt to explicate the principles underlying our work, with certain technical problems, and end with a brief panegyric on research programs and approaches to study of science and scientific communication.

Any value of citations and the restrictions on such value stems from the combined operation of several assumptions, three massive qualifications, and one critically important conjecture. The operation of these principles *differs* widely from field to field and from application to application. In discussing these principles we would note that the assumptions are very weak ones, and that their power derives from the stochastic nature of the world, laying the groundwork for true quantification only with very large files. For example, we can see a possible restriction in mathematics, where the literature base is small and each article contains few references. In such a field citations must be far more robust than in molecular biology, where the reverse holds on both dimensions.

Let's look now at four assumptions:

I. A document x cited by document y is more likely to be judged as related in content to document y than one not cited.

If we were to eponymize, this might be called the Garfield[2]-Kessler[3] Assumption, since it underlies both the remarkable *Science Citation Index®* service and the benchmark research at MIT. It's a hard assumption to gainsay.

II. If there are two documents x_1 and x_2, and x_1 is cited by document y and x_2 is not so cited, x_1 is more likely

---

*Reprinted from: *Society for Social Studies of Science Newsletter* 2:9-13, Summer, 1977.

to have been of use in the preparation of document y than x_2.

Let's continue to honor our intellectual forebears and designate this the Gross[4]-Price[5] Assumption, after the first major user of, and the user who has extracted amazing intellectual power from this assumption. Note again the extreme modesty of this assumption.

III. If documents cite documents in common, they are more likely to be judged as related in content than documents which do not cite any document in common.[3]
IV. If documents x_1 and x_2 are cited by document y, they are more likely to be judged as related to one another in content than to document x_3, which is not so co-cited with x_1 and x_2.

This is, of course, the co-citation assumption and with a friendly, generous spirit of self-eponymization—already a tradition in this field—the first author would call this the Small[6]-Marshakova[7]-Griffith[8] Assumption, after the first "mappers" to use this assumption.

These modest assumptions lay the groundwork for quantification and introduce other considerations, in particular, a series of necessary qualifications. These qualifications are massive, and give the user of citations fair warning that use is fraught with danger. (The "conjecture" is, however, quite powerful and offers the researcher hope.)

I. Citation measures are only the by-product of a file, and their quality is directly related to the dimensions of that file and the care taken to develop the file.
II. A series of complex social, psychological and bibliographical factors intervene between any intentions of the author to acknowledge precedent work or to recognize any form of similarity.

And, less terrifying:

III. Citation measures critically reflect the scale of the literature, and slightly independently, the scale and pace of research activity, as well as norms and institutions within the specialty and discipline.

These very strong restrictive qualifications regarding the use of citation measures have been repeated as admonitions,[9] but *violated* again and again in the literature. Later, we shall argue, however, that only rather serious violations matter.

What specifically do these qualifications mean?

(1) A file that is developed by the investigator must be fully described bibliographically and its dimensions must be justified as part of the study.

(2) If *SCI*® or *SSCI*™ is used, the investigator should be sensitive to the continuing improvement in coverage, particularly with regard to recent volumes of *SSCI* and, according to Derek Price, for the *SCI* prior to 1967.

(3) The difficulties encountered with developing counts for individuals have been described elsewhere,[10] and have become in certain areas of physics extremely difficult where the authorship includes 20-70 persons. For cognoscenti, we list typical considerations: homonyms, fractional authorships, self-citations, alphabetical as opposed to attributive ordering of authors, the precise relation of the citation to the content of the citing document, and multiple spellings, or arrangements, of the author(s)' name(s). (Ironically, Derek Price is a principal victim of the last. *)

(4) The investigator using *SCI* should be warned that errors normally originate in the citing document and, therefore exhibit far more creativity than typographical errors inherent in clerical operations, many of which are automatically eliminated.[11]

(5) The psychological and social factors have to do with habits, conventions, and perceptions of the scientist, his research group, his specialty, etc. Accordingly, they can be either an object of study or a bother, depending on the goals of the investigator. Also, different measures may be affected differentially. (The relative contributions to hepatic research of Baruch Blumberg and Alfred Prince, as recognized by the Nobel award to Blumberg, were reflected in total number of authorships among clustered documents on Australia Antigen, but not by number of citations to each author's most-cited article.)

(6) Bibliographical factors, alone or in conjunction with social and psychological factors, perturb citation measures. The pure case is the work of Karl Marx, where citations to a numerically small set of documents have been exploded into chaos by differently dated, different language editions (*SSCI*). Complex factors are introduced by the lack of any *necessary* correspondence between the content of dis-

---

*Derek J. DeSolla Price is often cited as Desolla P, Desolla DJ, Desollaprice D, or DJ in addition to the more logical variations on Price D, or DDS, or JDS, etc.

covery, the documents reporting the discovery, and a consensual perception of those documents by the research community. From the Australia Antigen example we can select three documents: one, a seminal finding that there may be a relation between the presence of Australia Antigen and hepatitis; the second, five years later, reporting a frequent association; and a third, slightly later, reporting 100% association within reasonable experimental error. The three taken as a group constituted the "crucial experiment" for that specialty. Citations to the first soared on appearance of the second and third papers—a formal finding only interpretable by recourse to the content of the papers. But in this case, community consensus, not the greatest credit, went to the second paper. One can easily see that any particular pattern of citing these papers might be intimately bound to the style of the citing author. The omission, or near omission, of any particular paper in citations can only be interpreted in terms of the pattern of citations and the content of related papers.

(7) Much of the above, as well as differences in scale and differences in custom of citations, renders the count of citations to an individual paper a fragile measure of the value of the paper. Instead, *the investigator should use all means at his disposal to determine the degree of consensus within the relevant community represented by a citation count and the nature of that consensus.* While the "nature" of consensus may suggest another cop-out and our rushing to content to bail out the measure, we believe it possible to turn this concept into a formally derived measure (by combining citation counts with clustering) that considers both co-citation and the total frequency of citation for each document. To give a rough idea, while papers which report standard laboratory methods in biomedicine are highly cited, they do not connect to groups of other papers and no group of new papers grows about them. On the other hand, the Nobel-award winning work of Baltimore, Blumberg, and Temin was related to existing documents and became the center of developing clusters.

(8) The scientist of science must regard his assumptions regarding the existence of a type of literature or the relevance of a particular paper to a particular literature as an *hypothesis,* which, we venture, is likely to be incorrect.

> Beginning with Parker, Paisley and Garrett,[12] investigators have been guided by strong hypotheses, independent of the inherent structure of the literature. (They sought the structure of the communication research literature, and found instead sociology and psychology.) Even greater strangeness may arise through uncritical acceptance of document assignment to specialties by bibliographies.

In all the above, we hope we make clear that the measures which are likely to be the most fragile are counts for the individual document and for the individual scientist. These are the measures which create the greatest sensitivity within the scientific community; and in part, our intention in writing this piece is to argue for the development and substitution of more sophisticated approaches.

We now turn to a conjecture of great amiability. It appears strong but may resist convincing proof.

> The quality and quantity of the scientific literature "channelizes." That is, a combination of social and probabilistic mechanisms ensure that most documents of a discipline, and nearly all documents of the highest quality, appear in a limited number of sources (i.e., journals in the natural sciences). Furthermore, all such important sources may be readily recognized and ranked along this quality dimension by citation counts.

This idea certainly dates from the Garfield[13]-Bourne[14] controversy over the number of scientific journals which should be covered by information services, say, 3000 ± 1000 (Garfield) as opposed to 30,000 ± 2000 (Bourne); these ideas are implicit in Price's writing, too.[15] The work of Narin on individual disciplines fully explores and supports this principle empirically.[16] However, the power of the mechanisms involved has not been emphasized sufficiently, nor has the central importance of this conjecture to citation research been indicated.

The likely truth and power of this conjecture is essential; if all journals (60,000) voted equally, research would be impossible and of course, *SCI* and *SSCI* would lose much of their usefulness as information services. Perhaps most strangely, the power of this conjecture has permitted persons, totally ill-equipped methodologically, to do valid work simply because they cannot avoid the "centers" of literatures.

This conjecture can be tested rather simply. Results of even a two-hour study within a departmental physics library—given expertise in sampling—would be sufficient for the highest ranked journals to be ordered

highly reliably and with perhaps 70-80% accuracy, as compared to a full study of our best data, the *Journal Citation Reports*.[17] Our only personal misgivings about use of this conjecture focus on the difficulty of finding starting points without language or national bias.

This conjecture, which appears correct but must again be approached and used with caution, completes the set of principles which, for us, underlie citation research.

REFERENCES

1. **Edge E.** Quantitative measures of communication in science. *International Symposium on Quantitative Methods in the History of Science: Proceedings.* Berkeley, California, August 25-27, 1976.
2. **Garfield E.** Citation indexes for science. *Science* **122**:108-11, 1955.
3. **Kessler M M.** Bibliographic coupling between scientific papers. *American Documentation* **14**:10-25, 1963.
4. **Gross P L K & Gross E M.** College libraries and chemical education. *Science* **66**:385-9, 1927.
5. **Price D J D.** Networks of scientific papers. *Science* **149**:510-15, 1965.
6. **Small H G.** Co-citation in the scientific literature; a new measure of the relationship between two documents. *Journal of the American Society for Information Science* **24**:265-9, 1973.
7. **Marshakova I V.** Bibliographic coupling system based on cited references. *Nauchno-Tekhnicheskaya Informatsiya Seriya* 2(6):3-8, 1973.
8. **Griffith B C, Small H G, Stonehill J A & Dey S.** The structure of scientific literatures. 2. Toward a macro- and microstructure for science. *Science Studies* **4**:339-65, 1974.
9. **Cole J R & Cole S.** Measuring the quality of sociological research; problems in the use of *Science Citation Index*. *American Sociologist* **6**:23-9, 1971.
10. ----------. *Social stratification in science.* Chicago: University of Chicago Press, 1973.
11. **Sher I H, Garfield E & Elias A W.** Control and elimination of errors in ISI services. *Journal of Chemical Documentation* **6**:132, 1966.
12. **Parker E B, Paisley W J & Garrett R.** *Bibliographic citations as unobtrusive measures of scientific communication.* Stanford, California: Stanford University Institute for Communication Research, October, 1967.
13. **Garfield E.** Significant journals of science. *Nature* **264**:609-15, 1976.
14. **Bourne C P.** The world's technical journal literature: an estimate of volume, origin, language, field, indexing, and abstracting. *American Documentation* **13**:159-68, 1962.
15. **Price D J D.** *Little science, big science.* New York: Columbia University Press, 1963.
16. **Narin F, Carpenter M P & Berlt N C.** Inter-relationships of scientific journals. *Journal of the American Society for Information Science* **23**:323-31, 1972.
17. **Garfield E.** *Journal citation reports: a bibliometric analysis of references processed for the 1975* Science Citation Index. *Science Citation Index 1976 Annual,* Vol. 9. Philadelphia: Institute for Scientific Information® , 1977.