Reprinted in Essays of an Information Scientist, Vol:4, p.621-633, 1979-80 Originally published in The Citation Key Papers in Information Science, P.196-210, Ed. B.C. Griffith, White Plains, NY: Knowledge

The Citation Cycle

Derek de S. Price, Yale University
2036 Yale Station, New Haven, CT 06520

This paper had its origin in a chart, long resident on my office wall. Frequently revised, corrected and redrawn, it seems to have taken on a life of its own and some utility to colleagues and students who acquired copies at various stages during more than ten years. The basic idea is to exhibit an interlocking metabolic complex of bibliometric (and scientometric) parameters in a comprehensive and integrated structure after the manner of the Nitrogen Cycle and other such paraphernalia beloved of organic chemists and ecologists. The data for this cycle are those drawn from the largest collection we have of machine-handled and automatically counted bibliographic items — the *Science Citation Index* (SCI) which has been published by the Institute for Scientific Information (ISI) since its foundation by Eugene Garfield in 1961. In biblio- and scientometrics it is often fatal and invariably debilitating to do your own counting. Beyond the tedious work and expense there is a hidden danger that one might well falsify the investigation by artifacts of definition and selection, so it is far better to use unobtrusive indicators produced by people who didn't know you were going to use them thus.[1] Much of my research in this area has been fed by a steady diet from the cutting room floor of printouts produced by ISI partly in their direct function of producing what is not only primarily a bibliographical aid but also *the* chief bibliographical service for scientists. The other part has been composed of special printouts generated by their admirable curiosity about their own processes, for which I am truly grateful.

An incidental advantage of this parasitic nourishment of my work is that the data, most of which are now conveniently published on an annual basis in the preambles to the *Social Science Citation Index*, the *Science Citation Index*, and the *Who is Publishing in Science* volumes, cover a large range of that which is implicated in the available corpus of both bibliometric and scientometric research theories. The citation cycle therefore embodies many of the elements of theory which are treated in the scholarly literature in our fields,[2] and it thus provides a sort of overview and coherent conspectus of a framework for the theories.

A tour of the Citation Cycle begins (see Figure 1), as does the formation of a citation index, with the selection of the Source Journals and the Items (usually research papers, but the more general term is useful) which are contained in them. The *selection* of journals is crucial to the success of a citation index because it is a strategy quite different from the usual librarian's striving for completeness. Though one may well start from an

Reprinted from: **Griffith B C,** ed. *Key papers in information science.*
White Plains, NY: Knowledge Industry Publications, 1980. p. 195-210.

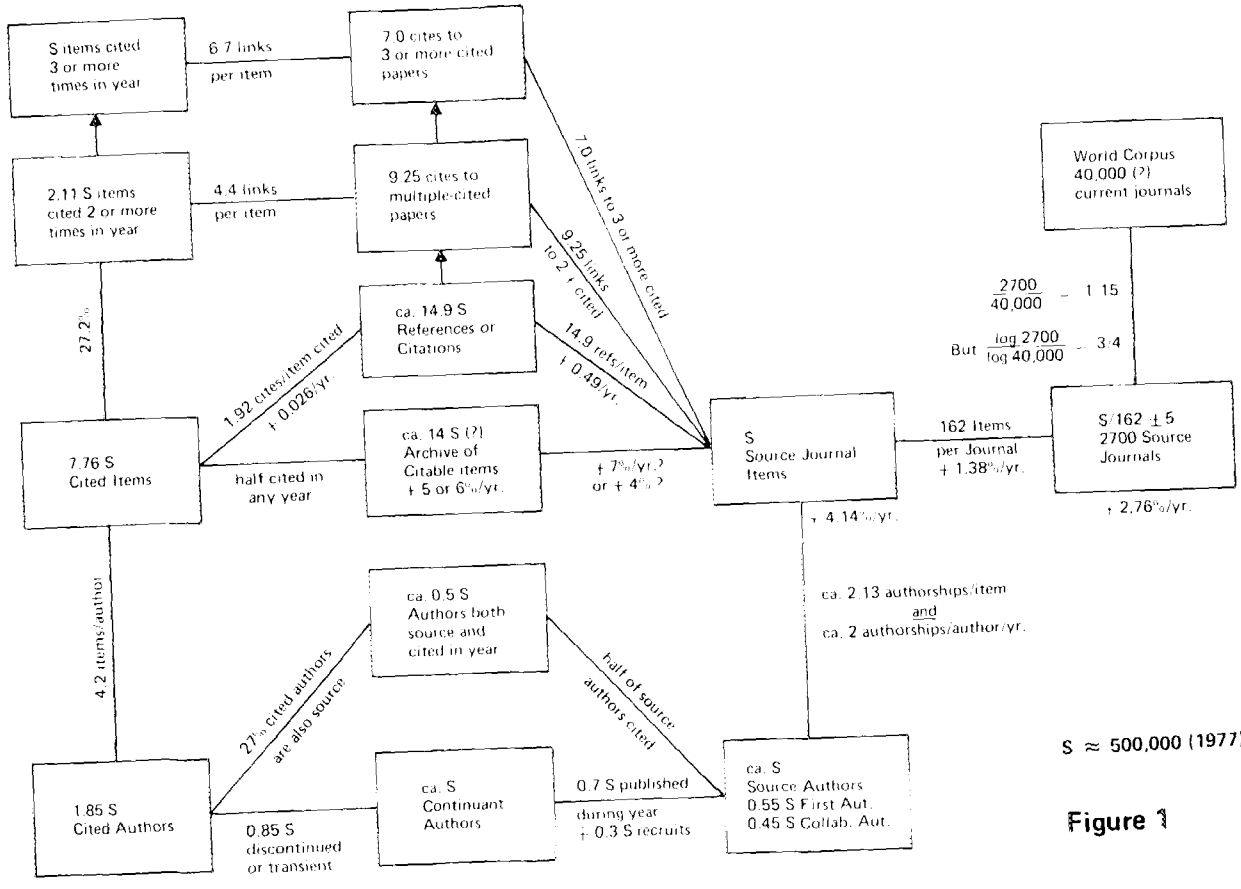
S items cited 3 or more times in year
6.7 links per item
7.0 cites to 3 or more cited papers
2.11 S items cited 2 or more times in year
4.4 links per item
9.25 cites to multiple-cited papers
7.0 links to 3 or more cited
9.25 links to 2 + cited
27.2%
1.92 cites/item cited + 0.026/yr.
ca. 14.9 S References or Citations
14.9 refs/item + 0.49/yr.
7.76 S Cited Items
half cited in any year
ca. 14 S (?) Archive of Citable items + 5 or 6%/yr.
+ 7%/yr.? or + 4%?
S Source Journal Items
+ 4.14%/yr.
162 Items per Journal + 1.38%/yr.
S/162 ± 5 2700 Source Journals
+ 2.76%/yr.
World Corpus 40,000 (?) current journals
2700/40,000 = 1.15
But log 2700/log 40,000 = 3.4
4.2 items/author
ca. 0.5 S Authors both source and cited in year
27% cited authors are also source
half of source authors cited
ca. 2.13 authorships/item and ca. 2 authorships/author/yr.
1.85 S Cited Authors
0.85 S discontinued or transient
ca. S Continuant Authors
0.7 S published during year + 0.3 S recruits
ca. S Source Authors 0.55 S First Aut. 0.45 S Collab. Aut.
S ≈ 500,000 (1977)

Figure 1

attempt to include all significant journals within some definition from all countries and all fields as sources, the ultimate test is provided as feedback from the journals which are cited by such sources. For many years the list of cited journals has provided a higher criticism of which journals to accept and which to reject as sources. Some journals may be so esoteric or so local that the citations they receive are from themselves. Others may have purposes of news and current awareness rather than the communication of citable knowledge and be for that reason almost uncited even by themselves. Then again some of the most cited journals may be extinct or living under a new name, or they may use the archaic practice of incorporating references in the body of the text where it is too expensive to employ key-punchers to excavate them.

If ISI chose its ca. 2700 source journals at random they would be only about 6.7 percent of the (maybe) 40,000 scientific and technical journals extant in the world, and hence they would contain only a comparable fraction of the current source literature. If ISI were perfectly successful, as no doubt they are not quite, in skimming only cream, they would get as sources just those source journals which were the most cited. In that case one can apply the powerful principle of Bradford's approximation to the distribution law of cumulative advantage in journals;[3] cumulating citations from the most-cited journals downwards, the total of citations is proportional to the logarithm of the number of journals included. This is much more realistic and it has the advantage, as it should, that the result is not at all sensitive to the count of all the world's journals — a ball-park estimate will serve. The result of this estimate is that the SCI now includes log2700/log40000 = 0.75 of all cited papers. Thus although it is derived from only 1/15 of the source papers, it includes 3/4 of the cited literature. As a corollary we may now claim that if the data in our Citation Cycle are multiplied by 4/3 they will give the world data for the cited corpus.

The 2700 source journals did not come all at once. The first few numbers of annual publications were based on about 600 journals and then in 1964-67 there was a period of expansion and revision (see Figure 2). From 1969 onwards the number of journals has been expanding at an exponential growth of 2.76% a year (derived from a regression of the logarithm of the number). This is much smaller than estimates of the world growth of scientific literature, 6-7% a year, so we are dealing with a relatively unchanging core of journals. The number of source articles in these journals is now about 500,000 and it has been growing since 1969 at a rate of 4.14% a year; it follows that on the average the journals have become slightly fatter at a rate of 1.38% a year. Apart from this slow change we can say that although there is considerable variation in size between journals, on the average each contains about 162 ± 5 source items/year (see Figure 3). Note the sharp drop in average size during the 1964-68 expansion as smaller journals are added. This is an interesting size, for it is equal in magnitude to an average invisible college of co-workers, usually 100-200

people, each writing about one paper a year in any of the major subdisciplines into which science is divided.[4] One might conjecture that a journal comes into being to serve such an internally communicating group of researchers, and then in the normal process of aging as the invisible college grows and produces new groups by fission some of the journals survive as media for aggregates of the living subfields.

The next stage of the tour of the Citation Cycle connects the number of Source Items (we shall designate this as S henceforth) with the authors of those items. In the dim distant past of science, from the late seventeenth century when scientific journals began until about World War I, when collaborative authorship was a rather rare event, the norm was that an

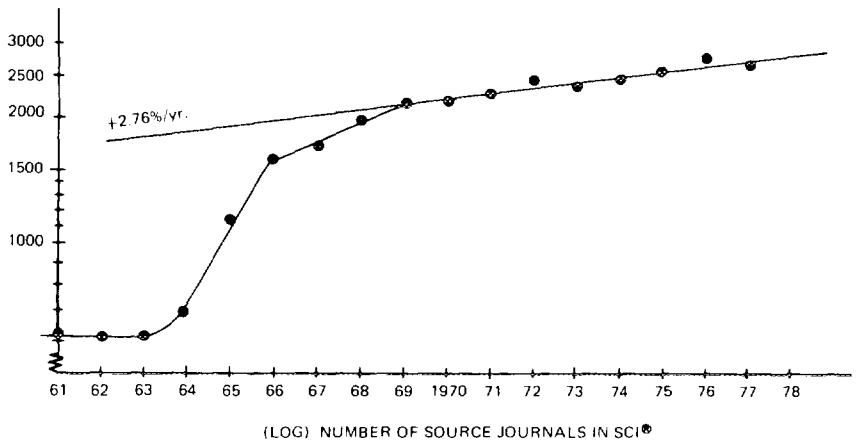

Figure 2

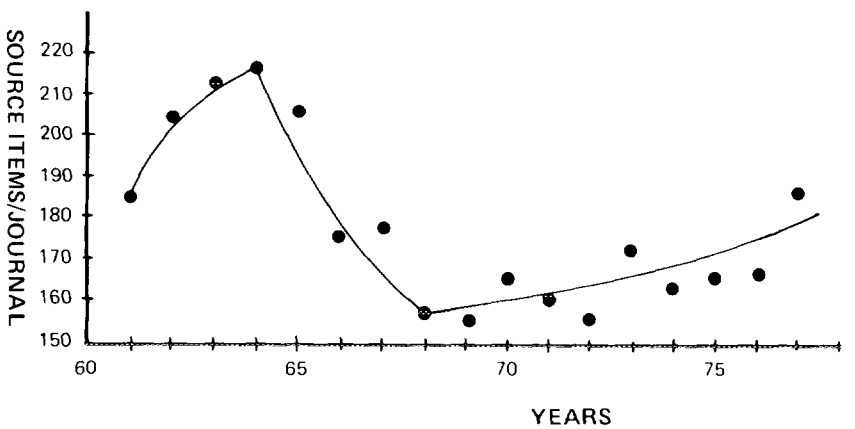

Figure 3

active researcher produced about one scientific paper per year.[5] Professionals tend to have a discretionary period in their life style which runs on an annual basis, and in harmony with our annual reporting of activity the normal life cycle of a project tends to be adjusted to this calendar cycle. What has happened since that period, and with great rapidity in the time since World War II, is that scientific authors collaborate increasingly so that in most scientific fields there is an average of two names or a little more per paper. What is happening is that the developing entrepreneurial tradition of channeling research support funding through a principal investigator permits that person to purchase subsidiary authors in effect. The result is that the number of authors per paper has become a rather good indicator of the extent of grant support in the field. Cancer and heart disease research is highly collaborative, pure mathematics much less so; it may be that in fields that need big team work the grants have to run high, but the effect may just as well be the other way round in causality. At all events, even though it now takes two authors to produce a paper, the output in papers has stayed constant, for now instead of each author getting out one paper per year, the team of two on the average produces two papers per year. The result is that the number of Source Authors is also S, and to be more precise there will be amongst them 0.55 S primary authors and 0.45 S secondary authors. Also to be a little more precise, there are now 2.13 authorships associated with each paper, across all fields. It should be noted that this group of statistics varies quite a lot from field to field, perhaps even from country to country. There are some fields like systematic taxonomy in natural history, or certain parts of organic chemistry where a paper may correspond to only a few weeks' work, and there are fields like astrophysics where an ordinary research contribution may be of two years' duration or longer to make a single paper—such goes the size of atoms of knowledge in various disciplines.

For the next stage in the tour we enter the domain of citations. Each paper includes a list of articles to which it refers. The references are usually at the end of the paper or footnotes on the page, and in the formation of the SCI these are key-punched into the computer record to be sorted into a citation index, alphabetic by cited author. Although the source items include everything from those totally devoid of references, e.g., news items and pontificating remarks, to those with hundreds or thousands of references in a bibliography, on the average there are about 14 references from each of the source items. In fact, cumulative advantage theory shows that what is really happening is not to be thought of as the new papers making reference back to the old; it is the old papers that are throwing off citations every year and thereby making occasion for the new literature. At all events, the average number of references in a paper is determined by the size of the available archive of literature in that field. Indeed the number of references per paper must be a small constant (less than one) plus the natural logarithm of the size of the archive. The natural logarithm of one million is about 14 and that is why the number of references is what it is.

For the Social Science Citation Index (SSCI) the corresponding number is about 11 references per paper, which is what would happen for an archive of about 60,000 papers in each field. Both in the SCI and in the SSCI, the number of references per paper has been increasing as the archive has grown. For the SCI there has been an increase of just less than half a reference per year (0.49) and for the SSCI the value is 0.62 per year. For the SCI the relative growth in number of references is about 3.5% a year and for the SSCI about 5.5%, corresponding to rates of growth of the archive at these values. Though both are lower than the traditional 7% per year growth rate of all scientific literature that we used to assume, they

Source and Citation Data from SCI and SSCI

SCIENCE CITATION INDEX®

	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977
Source Journals	613	605	610	700	1146	1573	1711	1968	2180	2192	2277	2425	2364	2443	2540	2717	2655
Source Items*	113	124	129	152	236	274	304	309	341	362	364	378	407	401	419	451	495
Refs/Cites*	1370	1486	1558	1790	2925	3074	3387	3699	3850	4108	4380	4459	5017	5232	5536	6177	7398
Items Cited*	890	895	970	1092	1617	1821	1994	2139	2262	2340	2450	2597	2730	2818	3006	3246	3776
Authors Cited*	258	267	281	324	439	474	510	547	601	620	646	688	711	730	772	813	908
Cites/Item Cited	1.52	1.63	1.58	1.60	1.65	1.65	1.66	1.70	1.67	1.73	1.76	1.76	1.81	1.83	1.81	1.87	1.92
Cites/Author Cited	5.23	5.67	5.44	5.38	6.07	6.36	6.51	6.52	6.28	6.52	6.57	6.65	6.95	7.05	7.05	7.68	8.01
Items/Author Cited	3.44	3.36	3.44	3.36	3.68	3.85	3.92	3.91	3.76	3.77	3.79	3.78	3.84	3.85	3.90	4.00	4.17
Refs/Source Item	12.1	12.0	12.1	11.8	12.4	11.2	11.1	12.0	11.3	11.4	12.0	12.3	12.3	13.0	13.2	13.7	14.9

*Thousands

Data from prefaces of SCI® and SSCI™

SOCIAL SCIENCES CITATION INDEX®

	1970	1971	1972	1973	1974	1975	1976
Source Journals	1000	1030	970	1052	1278	1232	1517
Source Items*	73	80	73	70	83	98	127
Refs/Cites*	618	644	604	633	872	1025	1372
Items Cited*	423	436	400	415	576	686	925
Authors Cited*	166	169	158	165	230	253	336
Cites/Item Cited	1.28	1.33	1.36	1.36	1.36	1.33	1.33
Cites/Author Cited	3.27	3.42	3.39	3.41	3.40	3.68	3.68
Items/Author Cited	2.55	2.57	2.53	2.51	2.50	2.71	2.77
Refs/Source Item	8.69	8.06	8.25	9.06	10.50	10.44	10.81

match the growth rate of source articles reasonably well. One must suppose that the ISI corpus is now growing at little more than half the historic long-term growth rate of the literature in the past century or so.

The references back from the source papers fall upon the available archive of papers already published. As we shall see, only about half of this archive is cited at all in any particular year, but of those papers that are cited a large majority, 72.8%, are cited once only. Of the remaining papers about half are cited just twice, and though the number of papers falls off very rapidly at about the inverse cube of the number of citations, there are still 1/400 of the items with more than 20 citations per year. Since some few

heavily cited items with several thousand citations a year exist — the so-called Method Papers and Reference Books — this tail of the distribution may represent a highly significant part of the citation behavior. Cumulative advantage theory accounts very well for the observed distribution. A fundamental parameter is the number of citations per cited paper[6] which varies slowly, as does the number of references per source item, with the logarithm of the available archive. There are now about 1.92 citations per cited item, and this is increasing linearly at 0.026 per year (see Figure 4). The corresponding figure for the SSCI is 1.33 citations per item cited, but as yet any secular increase appears to be masked by the settling down of the source selection which is still in its first few years.

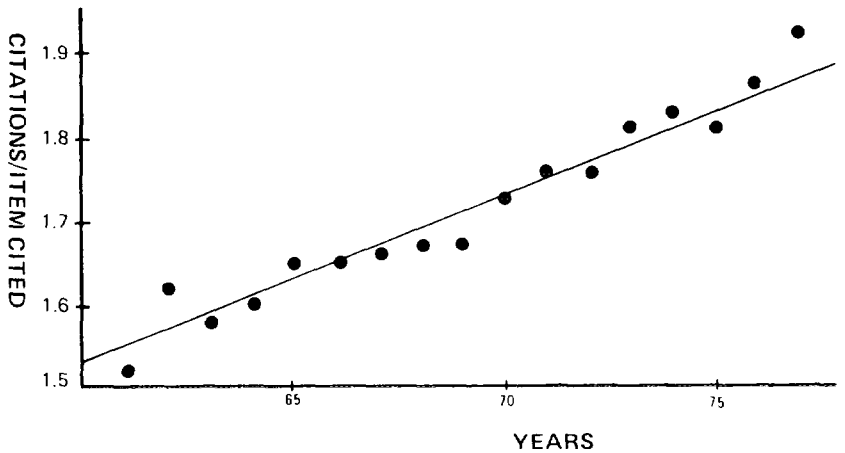

Figure 4

As a result of this multiplicity of citation, the 14.9 S references from the S source items fall upon 14.9 S / 1.92 = 7.76 S cited items. For the SSCI the corresponding figure is 10.8 S / 1.33 = 8.12 S cited items. It should be remembered that although all source items are from journals, the cited items include also a significant proportion of books, monographs, etc. Even so, the cited items could be only a minority of the archive available for citation, since at a growth rate of 7% the archive must be ca. 14 S, and for the empirical growth rate of 4.14% for source items the archive would be 24 S. Even at random, the probability of an archival item being cited at all should be in the range 0.33 to 0.57 and with a Poisson Distribution the citation hits per item cited would be in the range 1.18 -1.31. The significantly higher empirical figures show that cumulative advantage works very forcibly to increase the number of highly cited items beyond those that would occur with random events.

Since the cited items are sorted alphabetically by author it is easy to make a distribution of the number of citations per cited author, or better still, the average number of cited items per cited author. At present this

parameter has a value of about 4.2 for the SCI and 2.8 for the SSCI. In the former case we have enough years of data to establish a trend (see Figure 5); there seems to have been considerable perturbation of the parameter during the 1964-68 reorganization, but since 1969 the parameter has been increasing about 1% a year probably due more to the secular increase in collaborativeness rather than to any real increase in productivity of paper producing. Since we have 7.76 S cited items in the SCI there will be 7.76 S / 4.2 = 1.85 S cited authors, and for the SSCI there will be 8.12 S / 2.8 = 2.90 S cited authors.

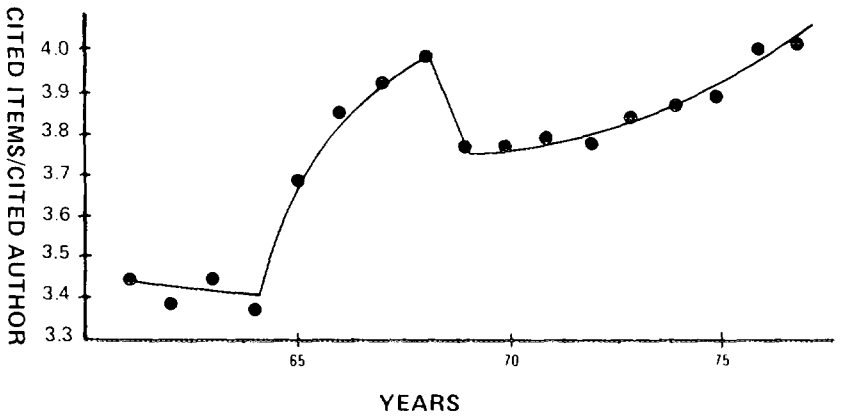

Figure 5

At this point in the tour of the Citation Cycle we may complete a loop by examining the relationship between the cited authors and the source authors. A collating of source and cited indexes shows that for both the SCI and the SSCI only about half of the source authors in any year are also cited. This doubtless corresponds to the fact that about half of the year's source authors are collaborating graduate students and junior faculty without a backlog of papers of which they are the first author available for citation. The 0.55 S first authors in the sources are therefore to be compared with the 1.85S that are cited in the SCI and the 2.90 cited in the SSCI. It follows that those active in the year are 30% of the SCI stock and 19% of the SSCI stock.

Another, more accurate way of looking at the relationship is to note that we know from an independent investigation of a small slice of the SCI for a long period[7] that only some of the collaborative authors are newcomers. In fact, of the S source authors, 70% are continuants who publish for an extended period, and 30% are newcomers. Further, for the continuants we know that in any year they have a probability of 0.7 of making a publication. It follows that the S source authors imply the existence of the same number S continuants together with 0.3 S newcomers. The S continuants may now be compared with the cited authors, and we derive

immediately that for the SCI some 0.85/1.85 = 46%, and for the SSCI some 1.90/2.90 = 66%, of the cited authors must have become discontinued by the current date. Many of the authors who once published, particularly those who published only transiently, are no longer cited; only a few are retired or deceased. It is worth noting as an overall figure that the number of cited authors in the SCI is just under a million, and in the SSCI about 112,000.

Having made one circuit of the Citation Cycle by the comparison of source and cited authors we may make another from the comparison of source and cited items. They have already been compared above through the medium of considering the available accrued corpus. We now look at structural relationships of the network of references/citations which, as has long been evident,[8] knit the new layer of source papers to a small selection of highly active papers in the accrued corpus. Items that are cited only once in the index are, so to speak, only tacked on to the source item that cites them, and they cannot relate two source papers or be related to any other cited paper except through this. Multiple-cited papers are comparatively rare, constituting about 27.2% of those in an annual index. Since we have 7.76 S cited items in the SCI there must be 2.11 S multiple-cited items which are connected to the S source items by about 7.63 links of reference/citation; there are therefore 7.63 links per source item and 7.63/2.11 = 3.6 links per multiple-cited item. Going to next higher level of papers cited three or more times it turns out that the number of such papers is approximately equal to S, and the number of links at this level will be about 5.5 for each source *or* multiple-cited paper. For the SSCI there is less referencing, a small corpus, and hence a lower level of miltiple citation. For those papers cited twice or more we have about 1.3 S which are connected to the S source papers by 4.2 S links of reference/citation. These parameters enable us to establish the way in which the corpus of papers is knitted together by its links into a structure of source papers overlaying a similarly structured corpus of source papers.

A first visualization of the implied structure may be had by cutting out the very highly referencing bibliography-like sources and also the very highly cited method-like cited papers as well as those which are singly-cited and cannot therefore contribute more than a tacking-on process. In this case to a first crude approximation we may suppose there to be roughly equal numbers of source and multiple-cited papers connected by about four links to and from each respectively. We can visualize the source papers as lying on the intersections of a rectangular grid on a thin sheet which overlaps a similar grid of cited papers on a thick sheet representing several years of accretion of former sources (see Figure 6). Each point on the thin sheet is directly linked to the neighboring four on a complementary place of the thick sheet and vice versa. In this convention we may now see that the bibliography and method papers may be reinserted as extensive areas, rather than points, that each blanket a whole region in the

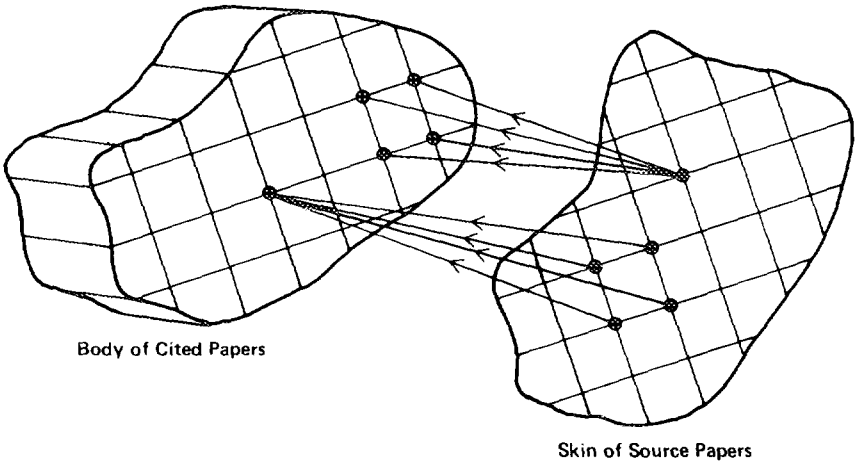

Figure 6

other sheet (see Figure 7). Clearly the general form of this picture can be extended to include the moderately referencing and cited papers, too, and we may make the depiction dynamic by supposing the thick corpus of cited papers to be formed from an onion-like accretion of annual shells growing out from a nucleus laid down in the distant past.

As a next stage in this visualization we note that if there were exactly four links per item the pattern of linkage might be represented by making

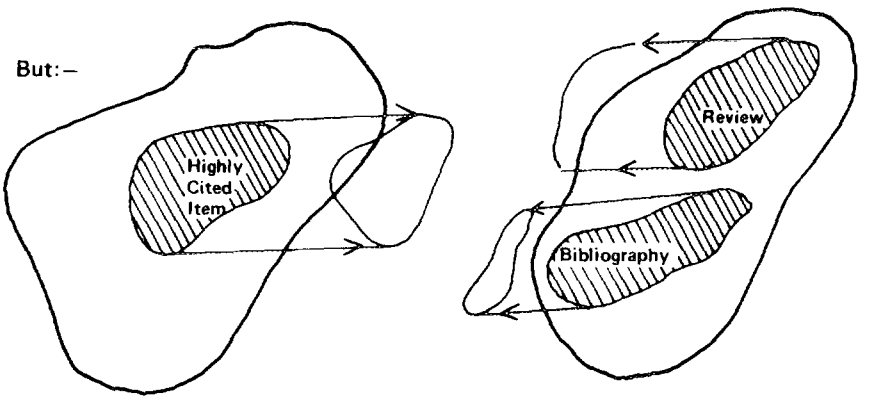

Figure 7

each intersection of a square lattice represent an item and the four lines running to it as the links. If each of the alternating source and cited items (denoted as S and C in Figure 8) had *exactly* four links the result would be a perfect lattice. If four is only a statistical mean, the corresponding lattice with various numbers of links would look rather like a very torn and deformable fishing net (see Figure 9), and if this is not envisaged in a

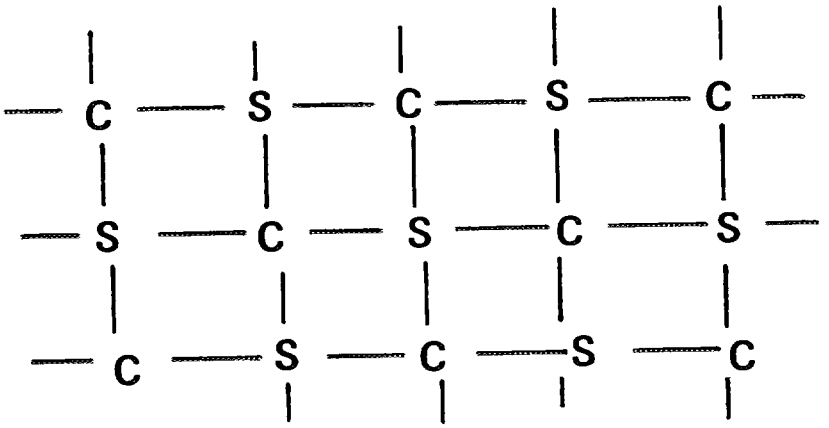

Figure 8

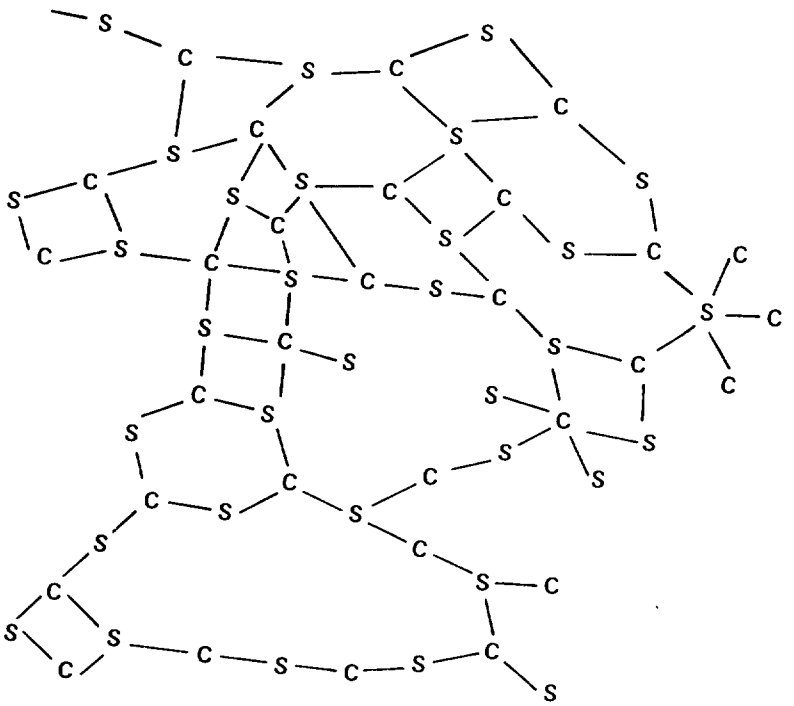

Figure 9

three-dimensional analog the result must look rather like the structure that is built into the network linkage of the corpus of science.

This property of the corpus now makes it possible to model relational structure of what has been called "subject space."[9] It is this space that is approximately mapped by the Griffith and Small[10] technique of cocitation analysis or that of Kessler in his bibliographic linkage which corresponds to co-referencing structure. What is implied is that we have built into the Citation Cycle not only the quantitative modeling but also a structural scheme. In a strong sense this structure provides a natural and automatic "indexing" of the entire corpus of scientific literature, and it seems evident that many of the recall/relevance trade-off problems of actual indexing arise from a conflict between this built-in structure and that imposed by the arbitrary structure of the classifier. Not the least of the problems must be that an essentially two-dimensional skin of source papers, or a three-dimensional corpus of cited papers (with time as the extra dimension) must be traversed by a classification scheme which, like the telephone book or the Dewey decimal system is essentially a one-dimensional traversing of the map.

1. Another advantage: this paper acknowledges no support whatsoever from any agency or foundation, but then, no time wasted either from preparing and submitting proposals.

2. For a general survey of the bibliometrics of citation see: Narin, Francis. *Evaluative Bibliometrics.* Computer Horizons, Inc., Project No. 704R, March 31, 1976; and, Garfield, Eugene. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities.* New York, Wiley-Interscience, 1979; and, Hjerppe, Roland. *An Outline of Bibliometrics and Citation Analysis.* Stockholm, The Royal Institute of Technology Library, October 1978.

3. Price, D.de S. "A General Theory of Bibliometric and other Cumulative Advantage Processes," *Journal of the American Society for Information Science,* 27,5/6:292-306, 1976; and, Price, D. de S. "Cumulative Advantage Urn Games Explained: A Reply to Kantor," *Journal of the American Society for Information Science,* 29:4, 204-206, 1978. For a recent large-scale empirical test of the Bradford approximation see: Drott, M. Carl and Belver C. Griffith. "An Empirical Examination of Bradford's Law and the Scattering of Scientific Literature," *Journal of the American Society for Information Science,* 29: 238-246, September 1978.

4. Price, D. de S. and Donald deB. Beaver. "Collaboration in an Invisible College," *American Psychologist,* 21:1011-1018, November 1966.

5. For a history of scientific coliaboration see: Beaver, D. DeB. and R. Rosen, "Studies in Scientific Collaboration, Part I. The Professional Origins of Scientific Collaboration, Part I. The Professional Origins of

Scientific Co-authorship," *Scientometrics*, 1: 65-84, 1978; Beaver, D. deB. and R. Rosen, "Studies in Scientific Collaboration, Part II. Scientific Co-authorship, Research Productivity and Visibility in the French Scientific Elite, 1799-1830," *Scientometrics*, 1:133-149, 1979; and Beaver, D. deB. and R. Rosen, "Studies in Scientific Collaboration, Part III. Professionalization and the Natural History of Modern Scientific Co-authorship," *Scientometrics* (in press).

6. Garfield, E. "Is the Ratio Between Number of Citations & Publications Cited a True Constant?" *Current Contents* 6:editorial, February 9, 1976.

7. Price, D. de S. and S. Gursey. "Studies in Scientometrics. Part I. Transience and Continuance in Scientific Authorship," *International Forum on Information and Documentation*, International Federation for Documentation, Moscow 1:2: 17-24, 1976; and Price, D. de S. and S. Gursey. "Studies in Scientometrics. Part II. The Relation Between Source Author and Cited Author Populations," *International Forum on Information and Documentation*, Moscow 1:3: 19-22, 1976.

8. Price, D. de S. "Networks of Scientific Papers," *Science* 149, 510-515, 1965.

9. Meincke, Peter P.M. and Pauline Atherton. "Knowledge Space: A Conceptual Basis for the Organization of Knowledge," *Journal of the American Society for Information Science*, 27: 18-24, Jan.-Feb. 1976 and, McGill, Michael J. "Knowledge and Information Spaces: Implications for Retrieval Systems," *Journal of the American Society for Information Science*, 27: 205-210, July-August 1976.

10. Small, H. and B.C. Griffith. "The Structure of Scientific Literatures I: Identifying and Graphing Specialties," *Science Studies*, 4:17-40, 1974; Griffith, B.C. and H.G. Small. "The Structure of Scientific Literature II: The Macro- and Micro-Structure of Science," *Science Studies*, 4: 339-365; Small, H.G. "A Co-citation Model of a Scientific Specialty: A Longitudinal Study of Collagen Research," *Social Studies of Science*, 7:139-166, 1977; and, Small, H. and Edwin Greenlee. Citation Context Analysis of a Co-citation Cluster: Recombinant-DNA," (to be published).