Science, 149(3683) : 510-515, July 30, 1965

Networks of Scientific Papers

The pattern of bibliographic references indicates the nature of the scientific research front.

Derek J. de Solla Price

This article is an attempt to describe in the broadest outline the nature of the total world network of scientific papers. We shall try to picture the network which is obtained by linking each published paper to the other papers directly associated with it. To do this, let us consider that special relationship which is given by the citation of one paper by another in its footnotes or bibliography. I should make it clear, however, that this broad picture tells us something about the papers themselves as well as something about the practice of citation. It seems likely that many of the conclusions we shall reach about the network of papers would still be essentially true even if citation became much more or much less frequent, and even if we considered links obtained by subject indexing rather than by citation. It happens, however, that we now have available machine-handled citation studies, of large and representative portions of literature, which are much more tractable for such analysis than any topical indexing known to me. It is from such studies, by Garfield (*1, 2*), Kessler (*3*), Tukey (*4*), Osgood (*5*), and others, that I have taken the source data of this study.

Incidence of References

First, let me say something of the incidence of references in papers in serial publications. On the average, there are about 15 references per paper and, of these, about 12 are to other serial publications rather than to books, theses, reports, and unpublished work. The average, of course, gives us only part of the picture. The distribution (see Fig. 1) is such that about 10 percent of the papers contain no references at all; this notwithstanding, 50 percent of the references come from the 85 percent of the papers that are of the "normal" research type and contain 25 or fewer references apiece. The distribution here is fairly flat; indeed about 5 percent of the papers fall in each of the categories of 3, 4, 5, 6, 7, 8, 9, and 10 references each. At the other end of the scale, there are review-type papers with many references each. About 25 percent of all references come from the 5 percent (of all papers) that contain 45 or more references each and average 75 to a paper, while 12 percent of the references come from the "fattest" category—the 1 percent (of all papers) that have 84 or more references each and average about 170 to a paper. It is interesting to note that the number of papers with n references falls off in this "fattest" category as $1/n^2$, up to many hundreds per paper.

These references, of course, cover the entire previous body of literature. We can calculate roughly that, since the body of world literature has been growing exponentially for a few centuries (*6*), and probably will continue at its present rate of growth of about 7 percent per annum, there will be about 7 new papers each year for every 100 previously published papers in a given

The author is Avalon Professor of the History of Science, Yale University, New Haven, Connecticut. This article is based on a paper presented 17 March 1964 at the National Bureau of Standards, Washington, D.C., in a Symposium on Statistical Methods for Mechanized Documentation. Part of this research was supported by grant GN-299 from the National Science Foundation.

field. An average of about 15 references in each of these 7 new papers will therefore supply about 105 references back to the previous 100 papers, which will therefore be cited an average of a little more than once each during the year. Over the long run, and over the entire world literature, we should find that, on the average, *every scientific paper ever published is cited about once a year.*

Incidence of Citations

Now, although the total number of citations must exactly balance the total number of references, the distributions are very different. It seems that, in any given year, about 35 percent of all the existing papers are not cited at all, and another 49 percent are cited only once ($n = 1$) (see Fig. 2). This leaves about 16 percent of the papers to be cited an average of about 3.2 times each. About 9 percent are cited twice; 3 percent, three times; 2 percent, four times; 1 percent, five times; and a remaining 1 percent, six times or more. For large n, the number of papers cited appears to decrease as $n^{2.5}$ or $n^{3.0}$. This is rather more rapid than the decrease found for numbers of references in papers, and indeed the number of papers receiving many citations is smaller than the number carrying large bibliographics. Thus, only 1 percent of the cited papers are cited as many as six or more times each in a year (the average for this top 1 percent is 12 citations), and the maximum likely number of citations to a paper in a year is smaller by about an order of magnitude than the maximum likely number of references in the citing papers. There is, however, some parallelism in the findings that some 5 percent of all papers appear to be review papers, with many (25 or more) references, and some 4 percent of all papers appear to be "classics," cited four or more times in a year.

What has been said of references is true from year to year; the findings for individual cited papers, however, appear to vary from year to year. A paper not cited in one year may well be cited in the next, and one cited often in one year may or may not be heavily cited subsequently. Heavy citation appears to occur in rather capricious bursts, but in spite of that I suspect a strong statistical regularity. I would conjecture that results to date could be explained by the hypotheses that

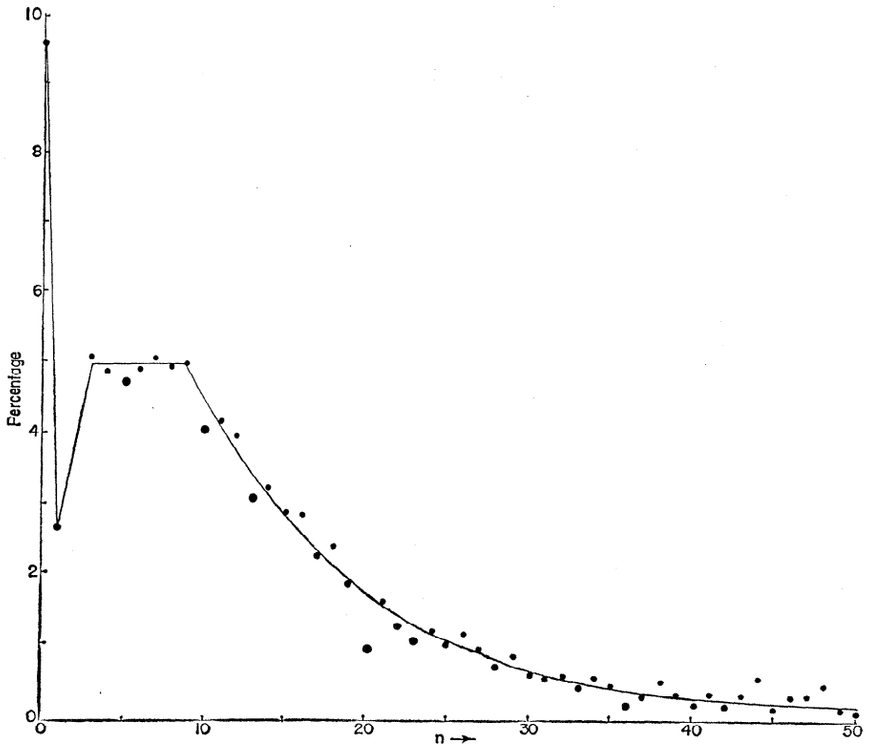

Fig. 1. Percentages (relative to total number of papers published in 1961) of papers published in 1961 which contain various numbers (n) of bibliographic references. The data, which represent a large sample, are from Garfield's 1961 *Index* (*2*).

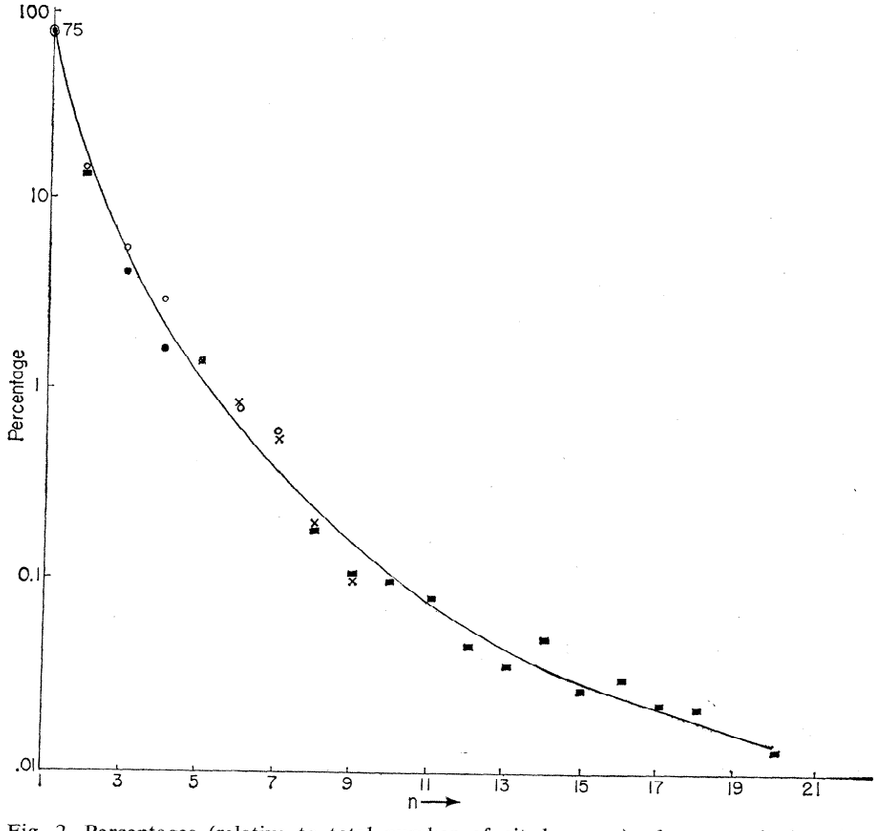

Fig. 2. Percentages (relative to total number of cited papers) of papers cited various numbers (n) of times, for a single year (1961). The data are from Garfield's 1961 *Index* (*2*), and the points represent four different samples conflated to show the consistency of the data. Because of the rapid decline in frequency of citation with increase in n, the percentages are plotted on a logarithmic scale.

every year about 10 percent of all papers "die," not to be cited again, and that for the "live" papers the chance of being cited at least once in any year is about 60 percent. This would mean that the major work of a paper would be finished after 10 years. The process thus reaches a steady state, in which about 10 percent of all published papers have never been cited, about 10 percent have been cited once, about 9 percent twice, and so on, the percentages slowly decreasing, so that half of all papers will be cited eventually five times or more, and a quarter of all papers, ten times or more. More work is urgently needed on the problem of determining whether there is a probability that the more a paper is cited the more likely it is to be cited thereafter. It seems to me that further work in this area might well lead to the discovery that classic papers could be rapidly identified, and that perhaps even the "superclassics" would prove so distinctive that they could be picked automatically by means of citation-index-production procedures and published as a single *U.S.* (or *World*) *Journal of Really Important Papers.*

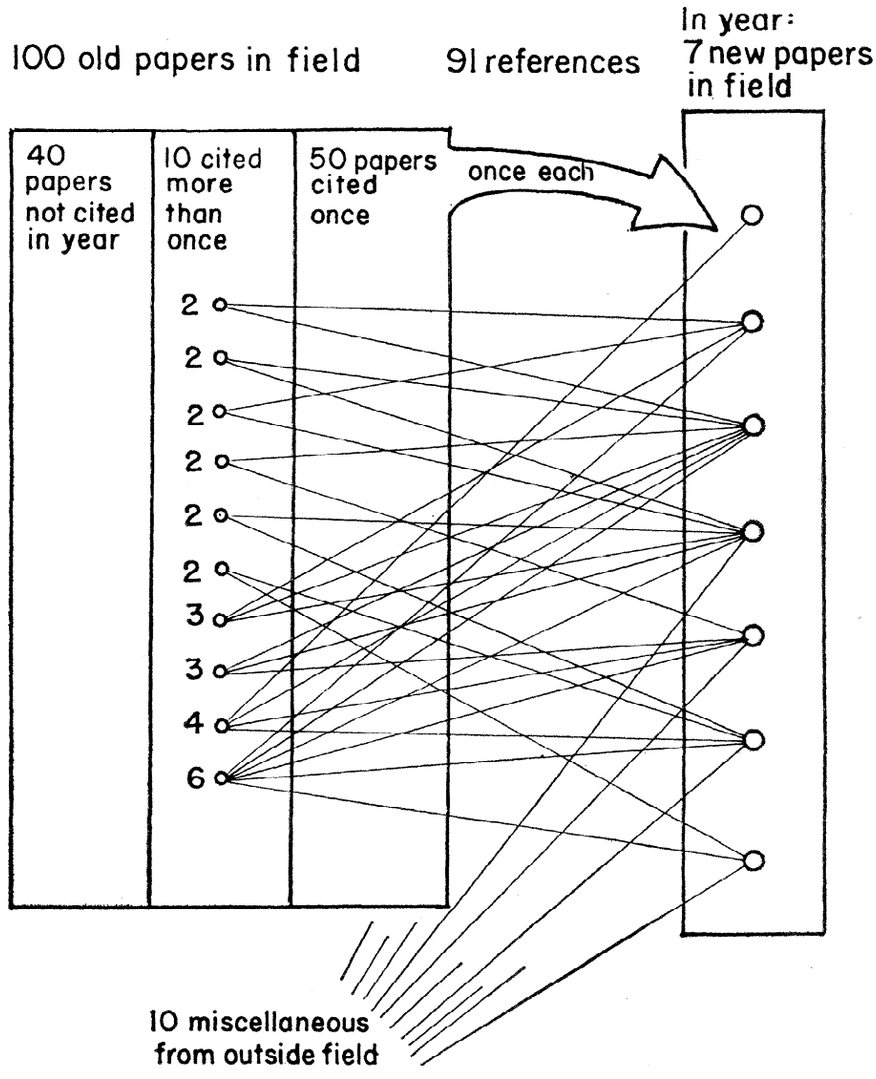

Fig. 3. Idealized representation of the balance of papers and citations for a given "almost closed" field in a single year. It is assumed that the field consists of 100 papers whose numbers have been growing exponentially at the normal rate. If we assume that each of the seven new papers contains about 13 references to journal papers and that about 11 percent of these 91 cited papers (or ten papers) are outside the field, we find that 50 of the old papers are connected by one citation each to the new papers (these links are not shown) and that 40 of the old papers are not cited at all during the year. The seven new papers, then, are linked to ten of the old ones by the complex network shown here.

Unfortunately, we know little about any relationship between the number of times a paper is cited and the number of bibliographic references it contains. Since rough preliminary tests indicate that, for much-cited papers, there is a fairly standard pattern of distribution of numbers of bibliographic references, I conjecture that the correlation, if one exists, is very small. Certainly, there is no strong tendency for review papers to be cited unusually often. If my conjecture is valid, it is worth noting that, since 10 percent of all papers contain no bibliographic references and another, presumably almost independent, 10 percent of all papers are never cited, it follows that there is a lower bound of 1 percent of all papers on the number of papers that are totally disconnected in a pure citation network and could be found only by topical indexing or similar methods; this is a very small class, and probably a most unimportant one.

The balance of references and citations in a single year indicates one very important attribute of the network (see Fig. 3). Although most papers produced in the year contain a near-average number of bibliographic references, half of these are references to about half of all the papers that have been published in previous years. The other half of the references tie these new papers to a quite small group of earlier ones, and generate a rather tight pattern of multiple relationships. Thus each group of new papers is "knitted" to a small, select part of the existing scientific literature but connected rather weakly and randomly to a much greater part. Since only a small part of the earlier literature is knitted together by the new year's crop of papers, we may look upon this small part as a sort of growing tip or epidermal layer, an active research front. I believe it is the existence of a research front, in this sense, that distinguishes the sciences from the rest of scholarship, and, because of it, I propose that one of the major tasks of statistical analysis is to determine the mechanism that enables science to cumulate so much faster than nonscience that it produces a literature crisis.

An analysis of the distribution of publication dates of all papers cited in a single year (Fig. 4) sheds further light on the existence of such a research front. Taking [from Garfield (*2*)] data for 1961, the most numerous count

available, I find that papers published in 1961 cite earlier papers at a rate that falls off by a factor of 2 for every 13.5-year interval measured backward from 1961; this rate of decrease must be approximately equal to the exponential growth of numbers of papers published in that interval. Thus, the chance of being cited by a 1961 paper was almost the same for all papers published more than about 15 years before 1961, the rate of citation presumably being the previously computed average rate of one citation per paper per year. It should be noted that, as time goes on, there are more and more papers available to cite each one previously published. Therefore, the chance that any one paper will be cited by any other, later paper decreases exponentially by about a factor of 2 every 13.5 years.

For papers less than 15 years old, the rate of citation is considerably greater than this standard value of one citation per paper per year. The rate increases steadily, from less than twice this value for papers 15 years old to 4 times for those 5 years old; it reaches a maximum of about 6 times the standard value for papers 2½ years old, and of course declines again for papers so recent that they have not had time to be noticed.

Incidentally, this curve enables one to see and dissect out the effect of the wartime declines in production of papers. It provides an excellent indication, in agreement with manpower indexes and other literature indexes, that production of papers began to drop from expected levels at the beginning of World Wars I and II, declining to a trough of about half the normal production in 1918 and mid-1944, respectively, and then recovering in a manner strikingly symmetrical with the decline, attaining the normal rate again by 1926 and 1950, respectively. Because of this decline, we must not take dates in the intervals 1914–25 and 1939–50 for comparison with normal years in determining growth indexes.

The "Immediacy Factor"

The "immediacy factor"—the "bunching," or more frequent citation, of recent papers relative to earlier ones—is, of course, responsible for the well-known phenomenon of papers being considered obsolescent after a decade. A numerical measure of this factor can be derived and is particularly useful. Calculation shows that about 70 percent of all cited papers would account for the normal growth curve, which shows a doubling every 13.5 years, and that about 30 percent would account for the hump of the immediacy curve. Hence, we may say that the 70 percent represents a random distribution of citations of all the scientific papers that have ever been published, regardless of date, and that the 30 percent are highly selective references to recent literature; the distribution of citations of the recent papers is defined by the shape of the curve, half of the 30 percent being papers between 1 and 6 years old.

I am surprised at the extent of this immediacy phenomenon and want to indicate its significance. If all papers followed a standard pattern with respect to the proportions of early and recent papers they cite, then it would follow that 30 percent of all references in all papers would be to the recent research front. If, instead, the papers

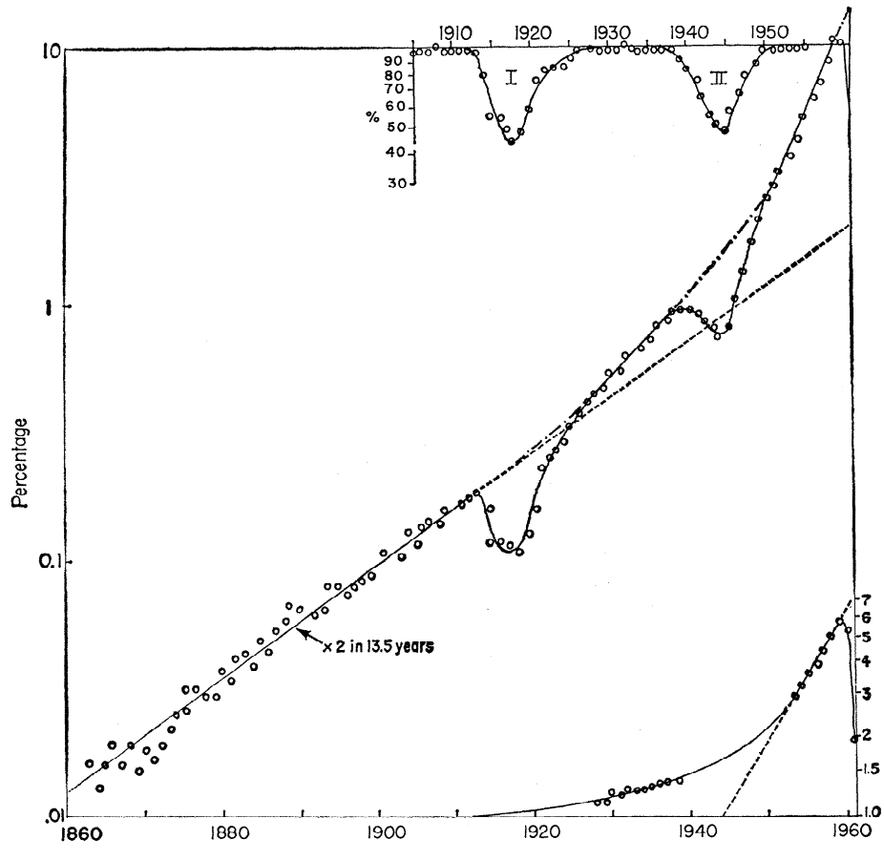

Fig. 4. Percentages (relative to total number of papers cited in 1961) of all papers cited in 1961 and published in each of the years 1862 through 1961 [data are from Garfield's 1961 *Index* (2)]. The curve for the data (solid line) shows dips during world wars I and II. These dips are analyzed separately at the top of the figure and show remarkably similar reductions to about 50 percent of normal citation in the two cases. For papers published before World War I, the curve is a straight line on this logarithmic plot, corresponding to a doubling of numbers of citations for every 13.5-year interval. If we assume that this represents the rate of growth of the entire literature over the century covered, it follows that the more recent papers have been cited disproportionately often relative to their number. The deviation of the curve from a straight line is shown at the bottom of the figure and gives some measure of the "immediacy effect." If, for old papers, we assume a unit rate of citation, then we find that the recent papers are cited at first about six times as much, this factor of 6 declining to 3 in about 7 years, and to 2 after about 10 years. Since it is probable that some of the rise of the original curve above the straight line may be due to an increase in the pace of growth of the literature since World War I, it may be that the curve of the actual "immediacy effect" would be somewhat smaller and sharper than the curve shown here. It is probable, however, that the straight dashed line on the main plot gives approximately the slope of the initial falloff, which must therefore be a halving in the number of citations for every 6 years one goes backward from the date of the citing paper.

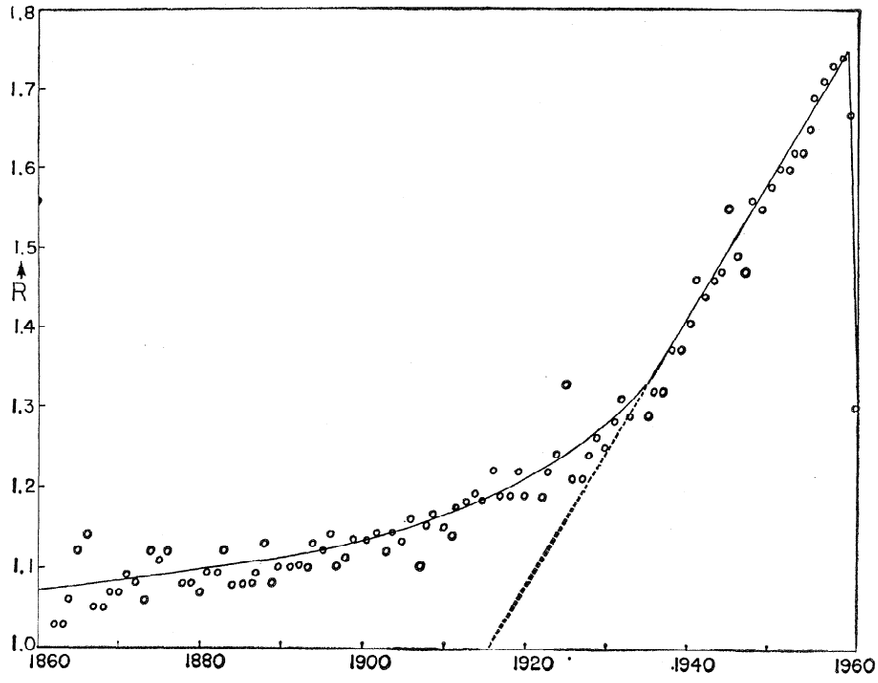

Fig. 5 (top left). Ratios of numbers of 1961 citations to numbers of individual cited papers published in each of the years 1860 through 1960 [data are from Garfield's 1961 *Index* (*2*)]. This ratio gives a measure of the multiplicity of citation and shows that there is a sharp falloff in this multiplicity with time. One would expect the measure of multiplicity to be also a measure of the proportion of available papers actually cited. Thus, recent papers cited must constitute a much larger fraction of the total available population than old papers cited.

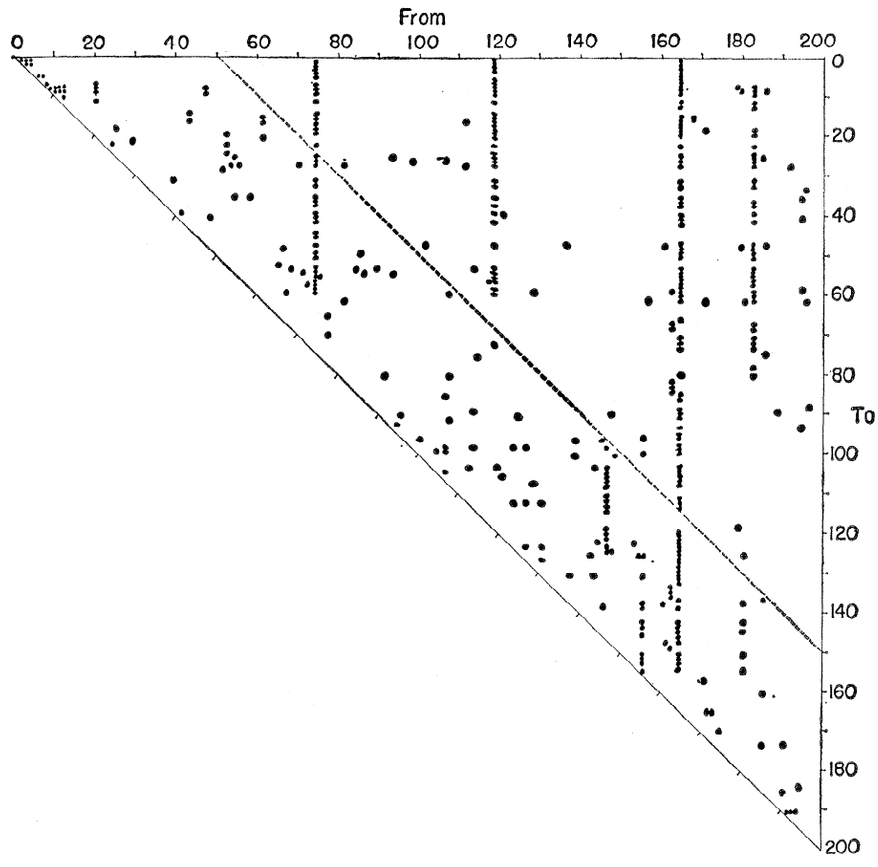

Fig. 6. Matrix showing the bibliographical references to each other in 200 papers that constitute the entire field from beginning to end of a peculiarly isolated subject group. The subject investigated was the spurious phenomenon of N-rays, about 1904. The papers are arranged chronologically, and each column of dots represents the references given in the paper of the indicated number rank in the series, these references being necessarily to previous papers in the series. The strong vertical lines therefore correspond to review papers. The dashed line indicates the boundary of a "research front" extending backward in the series about 50 papers behind the citing paper. With the exception of this research front and the review papers, little background noise is indicated in the figure. The tight linkage indicated by the high density of dots for the first dozen papers is typical of the beginning of a new field.

cited by, say, half of all papers were evenly distributed through the literature with respect to publication date, then it must follow that 60 percent of the papers cited by the other half would be recent papers. I suggest, as a rough guess, that the truth lies somewhere between—that we have here an indication that about half the bibliographic references in papers represent tight links with rather recent papers, the other half representing a uniform and less tight linkage to all that has been published before.

That this is so is demonstrated by the time distribution: much-cited papers are much more recent than less-cited ones. Thus, only 7 percent of the papers listed in Garfield's 1961 *Index* (*2*) as having been cited four or more times in 1961 were published before 1953, as compared with 21 percent of all papers cited in 1961. This tendency for the most-cited papers to be also the most recent may also be seen in Fig. 5 (based on Garfield's data), where the number of citations per paper is shown as a function of the age of the cited paper.

It has come to my attention that R. E. Burton and R. W. Kebler (*7*) have already conjectured, though on somewhat tenuous evidence, that the periodical literature may be composed of two distinct types of literature with very different half-lives, the classic and the ephemeral parts. This conjecture is now confirmed by the present evidence. It is obviously desirable to explore further the other tentative finding of Burton and Kebler that the half-lives, and therefore the relative proportions of classic and ephemeral literature, vary considerably from field to field: mathematics, geology, and botany being strongly classic; chemical, mechanical, and metallurgical engineering and physics strongly ephemeral; and chemistry and physiology a much more even mixture.

Historical Examples

A striking confirmation of the proposed existence of this research front has been obtained from a series of historical examples, for which we have been able to set up a matrix (Fig. 6). The dots represent references within a set of chronologically arranged papers which constitute the entire literature in a particular field (the field happens to be very tight and closed over the interval under discussion). In such a matrix there is high probability of citation in a strip near the diagonal and extending over the 30 or 40 papers immediately preceding each paper in turn. Over the rest of the triangular matrix there is much less chance of citation; this remaining part provides, therefore, a sort of background noise. Thus, in the special circumstance of being able to isolate a "tight" subject field, we find that half the references are to a research front of recent papers and that the other half are to papers scattered uniformly through the literature. It also appears that after every 30 or 40 papers there is need of a review paper to replace those earlier papers that have been lost from sight behind the research front. Curiously enough, it appears that classical papers, distinguished by full rows rather than columns, are all cited with about the same frequency, making a rather symmetrical pattern that may have some theoretical significance.

Two Bibliographic Needs

From these two different types of connections it appears that the citation network shows the existence of two different literature practices and of two different needs on the part of the scientist. (i) The research front builds on recent work, and the network becomes very tight. To cope with this, the scientist (particularly, I presume, in physics and molecular biology) needs an alerting service that will keep him posted, probably by citation indexing, on the work of his peers and colleagues. (ii) The random scattering of Fig. 6 corresponds to a drawing upon the totality of previous work. In a sense, this is the portion of the network that treats each published item as if it were truly part of the eternal record of human knowledge. In subject fields that have been dominated by this second attitude, the traditional procedure has been to systematize the added knowledge from time to time in book form, topic by topic, or to make use of a system of classification optimistically considered more or less eternal, as in taxonomy and chemistry. If such classification holds over reasonably long periods, one may have an objective means of reducing the world total of knowledge to fairly small parcels in which the items are found to be in one-to-one correspondence with some natural order.

It seems clear that in any classification into research-front subjects and taxonomic subjects there will remain a large body of literature which is not completely the one or the other. The present discussion suggests that most papers, through citations, are knit together rather tightly. The total research front of science has never, however, been a single row of knitting. It is, instead, divided by dropped stitches into quite small segments and strips. From a study of the citations of journals by journals I come to the conclusion that most of these strips correspond to the work of, at most, a few hundred men at any one time. Such strips represent objectively defined subjects whose description may vary materially from year to year but which remain otherwise an intellectual whole. If one would work out the nature of such strips, it might lead to a method for delineating the topography of current scientific literature. With such a topography established, one could perhaps indicate the overlap and relative importance of journals and, indeed, of countries, authors, or individual papers by the place they occupied within the map, and by their degree of strategic centralness within a given strip.

Journal citations provide the most readily available data for a test of such methods. From a preliminary and very rough analysis of these data I am tempted to conclude that a very large fraction of the alleged 35,000 journals now current must be reckoned as merely a distant background noise, and as very far from central or strategic in any of the knitted strips from which the cloth of science is woven.

References and Notes

1. E. Garfield and I. H. Sher, "New factors in the evaluation of scientific literature through citation indexing," *Am. Doc.* **14**, 191 (1963); ———. *Genetics Citation Index* (Institute for Scientific Information, Philadelphia, 1963). For many of the results discussed in this article I have used statistical information drawn from E. Garfield and I. H. Sher, *Science Citation Index* (Institute for Scientific Information, Philadelphia, 1963), pp. ix, xvii–xviii.
2. I wish to thank Dr. Eugene Garfield for making available to me several machine printouts of original data used in the preparation of the 1961 *Index* but not published in their entirety in the preamble to the index.
3. I am grateful to Dr. M. M. Kessler, Massachusetts Institute of Technology, for data for seven research reports of the following titles and dates: "An Experimental Study of Bibliographic Coupling between Technical Papers" (November 1961); "Bibliographic Coupling Between Scientific Papers" (July 1962); "Analysis of Bibliographic Sources in the *Physical Review* (vol. 77, 1950, to vol. 112, 1958) (July 1962); "Analysis of Bibliographic Sources in a Group of Physics-Related Journals" (August 1962); "Bibliographic Coupling Extended in Time: Ten Case Histories" (August 1962); "Concerning the Probability that a Given Paper will be Cited" (November 1962); "Comparison of the Results of Bibliographic Coupling and Analytic Subject Indexing" (January 1963).
4. J. W. Tukey, "Keeping research in contact with the literature: Citation indices and beyond," *J. Chem. Doc.* **2**, 34 (1962).
5. C. E. Osgood and L. V. Xhignesse, *Characteristics of Bibliographical Coverage in Psychological Journals Published in 1950 and 1960* (Institute of Communications Research, Univ. of Illinois, Urbana, 1963).
6. D. J. de Solla Price, *Little Science, Big Science* (Columbia Univ. Press, New York, 1963).
7. R. E. Burton and R. W. Kebler, "The 'half-life' of some scientific and technical literatures," *Am. Doc.* **11**, 18 (1960).