Networks of Scientific Papers

- **Derek J. De Solla Price**
- **1965, Science**
- **Link to paper**

Overview:

- There exists an "immediacy effect" wherein more recently published papers are cited most often.
- Goal: Describe the global network of scientific papers
- This is done by linking papers together based on their citation links

Incidence of References

- It is stressed that the conclusions drawn should not change even if the rate of citation increases or decreases – and should remain even if we linked papers by subject index rather than by citation.
    - See below for the distribution

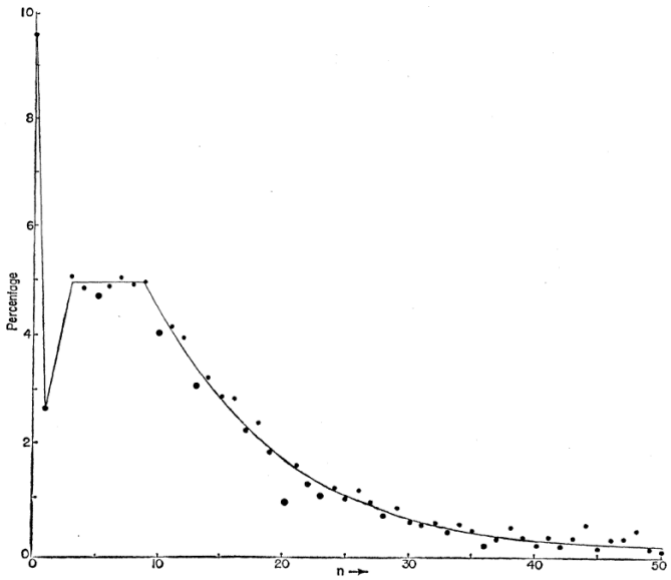

Fig. 1. Percentages (relative to total number of papers published in 1961) of papers published in 1961 which contain various numbers (n) of bibliographic references. The data, which represent a large sample, are from Garfield's 1961 *Index* (2).

    - The number of papers with n references falls of in the "fattest" category as $1/n2$
    - "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"*
        - See paper for details on how he determines this

Incidence of Citations

- While the total number of citations must exactly balance the total number of references, the distributions between the two are quite different.
- Yearly averages…
  - 35% of papers are not cited at all
  - 49% are cited only once
  - 16% are cited on average 3.2x
  - 9% are cited 2x
  - 3% → 3x
  - 2% → 4x
  - 1% → 5x
- For large *n* the # of papers cited appears to decrease as $n^{2.5}$ or $n^{3.0}$
  - This is much more rapid than the decrease for numbers of references in papers

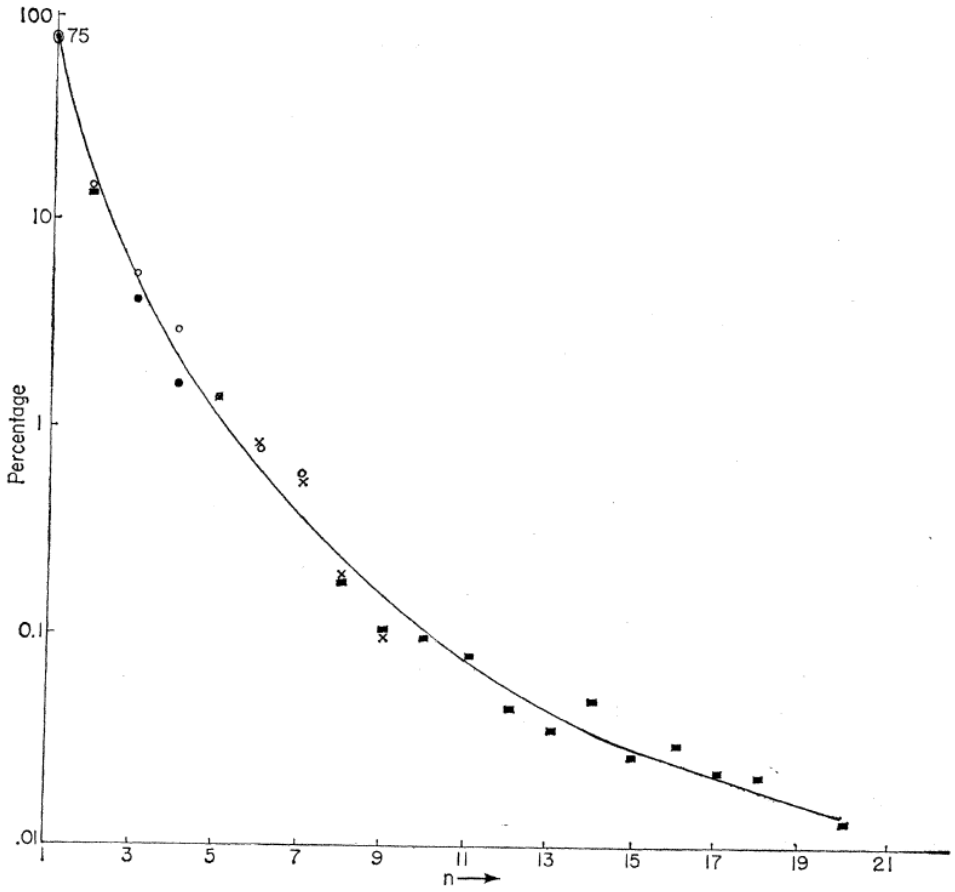

Fig. 2. Percentages (relative to total number of cited papers) of papers cited various numbers (n) of times, for a single year (1961). The data are from Garfield's 1961 *Index* (2), and the points represent four different samples conflated to show the consistency of the data. Because of the rapid decline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

A Tight Knit Network of Cited Papers

- Although most papers produced in a year contain about 13 references, half of them are references to about half of all the papers that have been published in previous years.
- The other half of the references tie the new papers to another small group of earlier ones.
- "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."
    - This smaller group represents the "active research front"

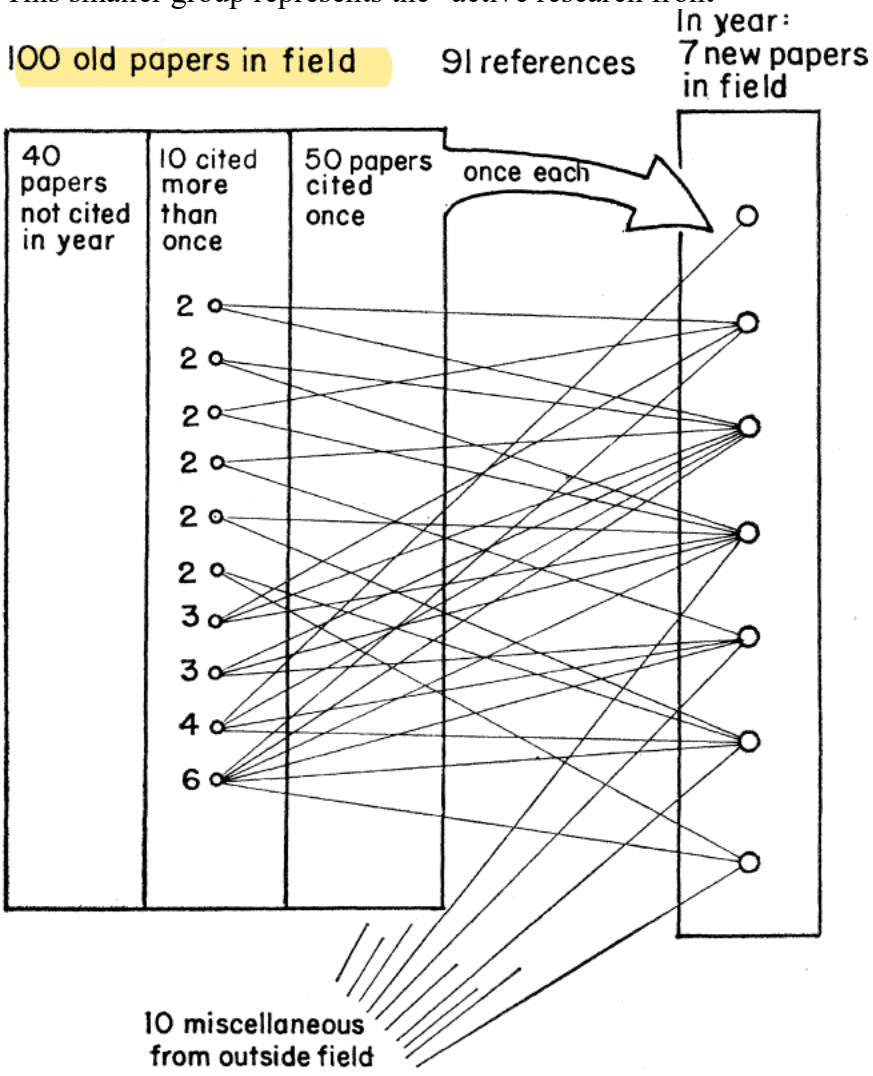

Fig. 3. Idealized representation of the balance of papers and citations for a given "almost closed" field in a single year. It is assumed that the field consists of 100 papers whose numbers have been growing exponentially at the normal rate. If we assume that each of the seven new papers contains about 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.

Analysis of Publication Dates

- Papers published in 1961 cite earlier papers at a rate that falls off by a factor of 2 for every 13.5-year interval.
    - This must be approximately equal to the exponential growth of numbers of papers published in that interval
    - Thus, the chance of being cited reduces by a factor of two every 13.5 years
    - As a result, all papers published more than 15 years earlier than 1961 have approx. the chance of being published
- Papers published less than 13.5 years earlier, however, have a much greater chance of being published

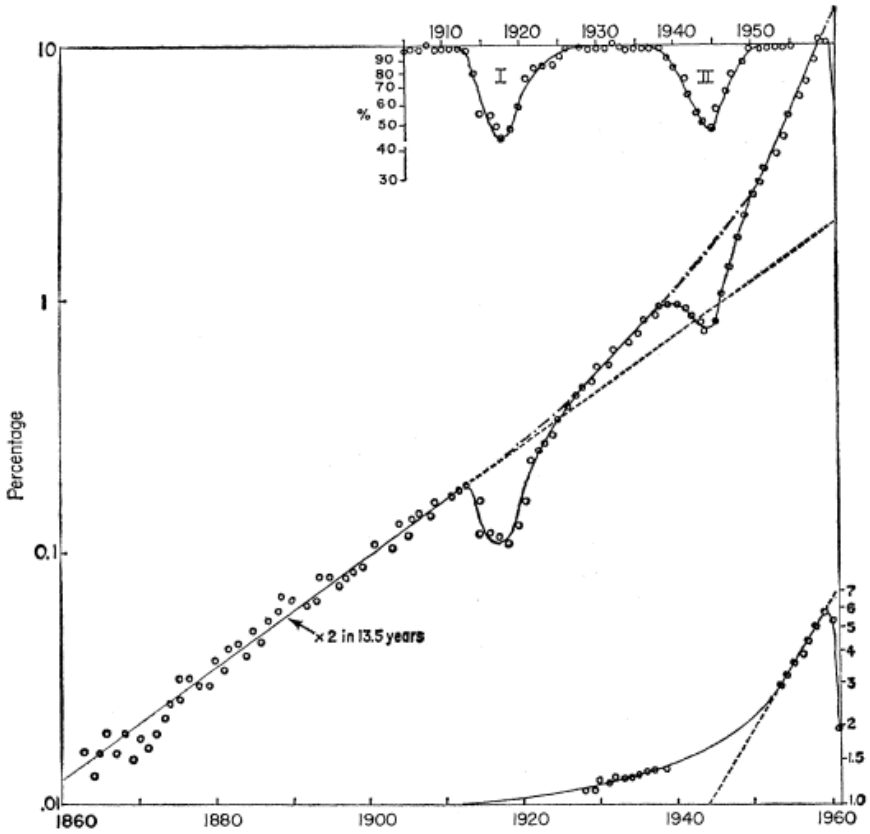

Fig. 4. Percentages (relative to total number of papers cited in 1961) of all papers cited in 1961 and published in each of the years 1862 through 1961 [data are from Garfield's 1961 *Index* (*2*)]. The curve for the data (solid line) shows dips during world wars I and II. These dips are analyzed separately 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.

Immediacy Factor

- Immediate factor – the "bunching" or more frequent citation of recent papers relative to earlier ones
- 70% of all cited papers would account for the normal growth curve – which shows a doubling every 13.5 years
    - 30% would account for the hump of the immediacy curve
        - These represent highly selective references to recent literature

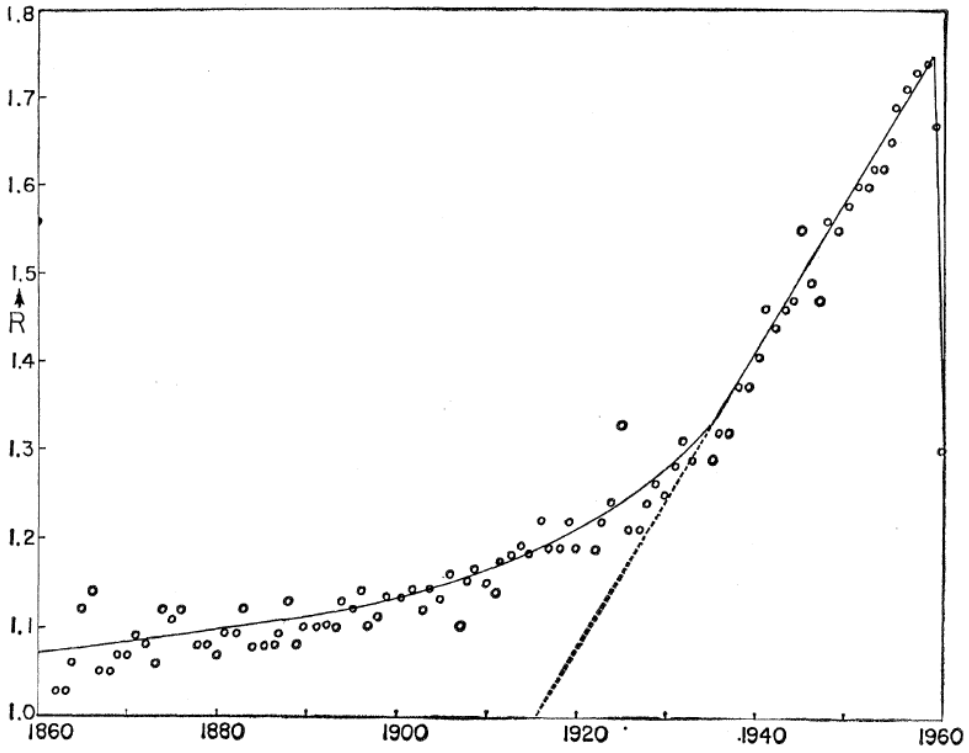

Fig. 5 (top left). Ratios of numbers of 1961 citations to numbers of individual cited papers published in each of the years 1860 through 1960 [data are from Garfield's 1961 *Index* (*2*)]. This ratio gives a measure of the multiplicity of citation and shows that there is a sharp falloff in this multiplicity 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.

- The above graph simply shows the "immediacy effect"
    - That most citations are provided to the most recent papers.
    - However, there is a sharp drop for the most recent papers as there has not been enough time for them to acquire citations
- This evidence shows us that that there are two types of papers – the "classic" and the "ephemeral"
- The "half-lives" of these types of papers vary considerably from field to field

Historical Examples…

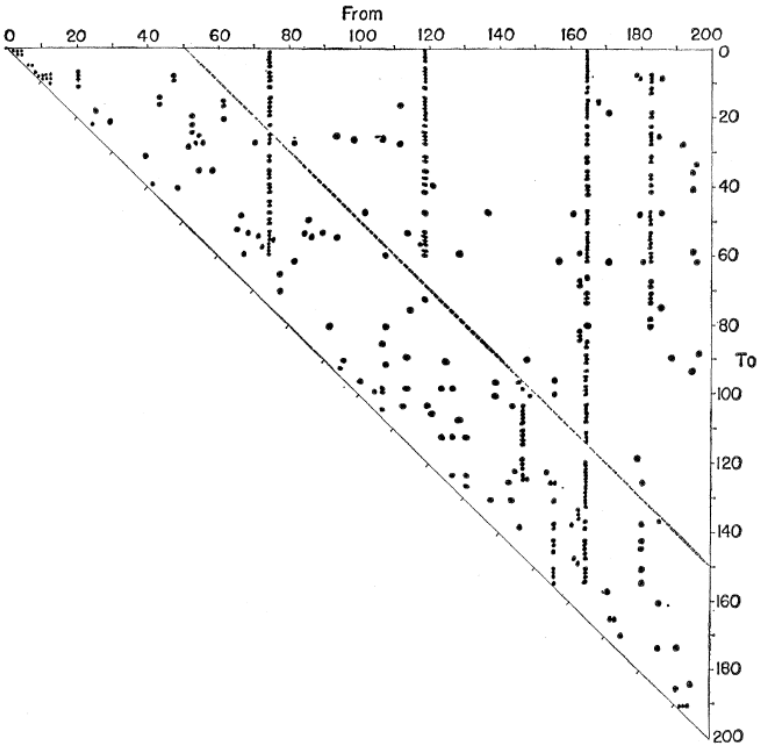

Fig. 6. Matrix showing the bibliographical references to each other in 200 papers that constitute the entire field from beginning to end of a peculiarly isolated subject group. The subject investigated was the spurious phenomenon of N-rays, about 1904. The papers are arranged chronologically, and each 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.

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.