Essays of an Information Scientist, Vol:2, p.28, 1974-76 Current Contents, #7, February 13, 1974

Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents

HENRY SMALL
Institute for Scientific Information
Philadelphia, Pennsylvania 19106

A new form of document coupling called co-citation is defined as the frequency with which two documents are cited together. The co-citation frequency of two scientific papers can be determined by comparing lists of citing documents in the *Science Citation Index* and counting identical entries. Networks of co-cited papers can be generated for specific scientific specialties, and an example is drawn from the literature of particle physics. Co-citation patterns are found to differ significantly from bibliographic coupling patterns, but to agree generally with patterns of direct citation. Clusters of co-cited papers provide a new way to study the specialty structure of science. They may provide a new approach to indexing and to the creation of SDI profiles.

• Introduction

Bibliographic citations in scientific papers have been used by a variety of researchers to establish relationships among documents. Both direct citation—the citing of an earlier document by a new document—and bibliographic coupling—the sharing of one or more references by two documents—have received considerable attention (*1, 2, 3, 4*). A related measure, which has been overlooked in earlier discussions, is co-citation. Unlike bibliographic coupling which links source documents, co-citation links cited documents and is, therefore, analogous to a measure of descriptor or word association (*5*). The purpose of this paper is to define this new kind of "coupling" and to distinguish it from bibliographic coupling, using an actual example from the literature of physics. Co-citation patterns are found to differ significantly from bibliographic coupling patterns, but to agree generally with patterns of direct citation.

• Measuring Co-citation

The strength of co-citation between two cited papers or books can be easily determined from a citation index such as the *Science Citation Index* (*SCI*). Each of the two papers is located in the *Citation Index* section of the *SCI*, and their lists of citing papers are scanned. The number of identical citing items defines the strength of co-citation between the two cited papers. An identical citing item is simply a new document which has cited both earlier papers; therefore, co-citation is the frequency with which two items of earlier literature are cited together by the later literature (*6*).

To be strongly co-cited, a large number of authors must cite the two earlier works. Therefore, co-citation is a relationship which is established by the citing authors. In measuring co-citation strength, we measure the degree of relationship or association between papers as perceived by the population of citing authors. Furthermore, because of this dependence on the citing authors, these patterns can change over time, just as vocabulary co-occurrences can change as subject fields evolve. Bibliographic coupling, on the other hand, is a fixed and permanent relationship because it depends on references contained in the coupled documents. Co-citation patterns change as the interests and intellectual patterns of the field change.

When two papers are frequently co-cited, they are also necessarily frequently cited individually as well. If it can be assumed that frequently cited papers represent the key concepts, methods, or experiments in a field (*7*), then co-citation patterns can be used to map out in great detail the relationships between these key ideas. This may lead to a more objective way of modelling the intellectual structure of scientific specialties. Changes in the co-citation patterns, when viewed over a period of years, may provide clues to understanding the mechanism of specialty development.

• Example of Co-citation

An illustration of a co-citation pattern which exists in the scientific literature is shown in Figure 1. Each box in the network diagram represents a cited paper in a specialty of particle physics which might be roughly described as: "Theories of broken chiral symmetry and current algebras." (The full bibliographic citation for each of the articles in the network is given in Table 1.) The number of undirected lines which connect the papers reflects, according to the code shown, the strength of co-citation coupling. The two strongest linkages are between Lovelace '68 and Veneziano '68, and between Gell-Mann '68 and Glashow '68. The network was drawn from data given in Table 2, omitting co-citation linkages which fall below the threshold of seven.

The procedure known as "cycling" was used in identifying the papers (*8*). The highly cited paper by Gell-Mann (1968) was used as a starting point. All the papers which cited Gell-Mann '68 in the first quarter 1971 *SCI* were examined. All citations for the papers in the bibliographies of the citing papers were then located in the quarterly *SCI*. If any of the items was cited ten or more

TABLE 1. Papers in the network

BJORKEN, J. D. "Applications of the Chiral U(6)⊗U(6) Algebra of Current Densities," *Phys Rev* 148: 1467 '66

GASIOROWICZ S. and D. A. GEFFEN "Effective Lagrangians and Field Algebras with Chiral Symmetry," *Rev Mod Phys* 41: 531 '69

GELL-MANN, M., "Symmetries of Baryons and Mesons," *Phys Rev* 125: 1067 '62

GELL-MANN, M., "The Symmetry Group of Vector and Axial Vector Currents," *Physics* 1: 63 '64

GELL-MANN, M., R. J. OAKES and B. RENNER, "Behavior of Current Divergences under $SU_3 \times SU_3$," *Phys Rev* 175: 2195 '68

GLASHOW, S. L. and S. WEINBERG "Breaking Chiral Symmetry," *Phys Rev Lett* 20: 224 '68

LOVELACE, C., "A Novel Application of Regge Trajectories," *Phys Letters* B 28B: 264 '68

VENEZIANO, G., "Construction of a Crossing-Simmetric, Regge-Behaved Amplitude for Linearly Rising Trajectories," *Nuovo Cimento* 57: 190 '68

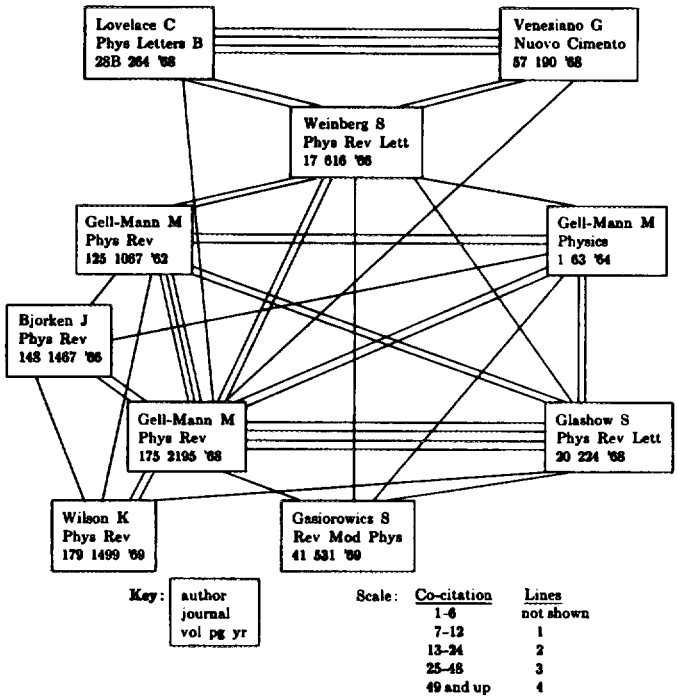
WEINBERG, S., "Pion Scattering Lengths," *Phys Rev Lett* 17: 616 '66

WILSON, K. G., "Non-Lagrangian Models of Current Algebra," *Phys Rev* 179: 1499 '69

Reprinted from Journal of the American Society for Information Science July-August 1973 Vol. 24, No. 4

FIGURE 1

Co-citation Network for Frequently Cited Papers in Particle Physics
(Data from the 1971 *SCI*)



times it became part of the cluster of key papers to be studied. Co-citation relationships among all items in the cluster were determined in the manner described earlier. The resulting network of ten papers is shown in Figure 1. It should be emphasized that while these ten papers may not represent *all* the frequently cited papers in this specialty, they probably constitute the "core" of that literature.

How can we explain the strength of these co-citation linkages? In some cases, we are dealing with relationships which are strongly recognized by physicists in this specialty. For example, *50* authors of papers in 1971 cited both Gell-Mann '68 and Glashow '68. When we examine the Gell-Mann '68 paper, we find that Gell-Mann not only directly cites the Glashow paper, he also discusses its high similarity to his own work: "Many authors have taken steps in the same direction, and in particular we should mention the work of Glashow and Weinberg as being the *most closely related to ours*" (*9*) (Author's italics). Furthermore, these papers display a strong bibliographic coupling, citing five common previous papers. Based on this example, one might be tempted to hypothesize that co-citation, like bibliographic coupling, measures subject similarity. However, the same agreement between bibliographic coupling and co-citation strength was not observed in other cases.

Table 2 displays the absence of any clear relationship between bibliographic coupling strength and co-citation frequency. This is especially evident in the case of Lovelace '68 and Veneziano '68, the strongest co-citation linkages on the diagram. These papers had a bibliographic coupling strength of only one. Once again, a reading of the Lovelace paper reveals why it came to be so closely associated with the Veneziano paper, and, further, why it was linked to the Gell-Mann work. In the first paragraph of his paper Lovelace stated: "Recently Veneziano gave a simple formula which exhibits the Regge pole=resonance duality. We show here how this new development in Regge theory can be applied in a rather unexpected direction, . . . There is also reason, from at least seven parallel predictions, to believe [the Veneziano formula] connected with chiral symmetry." (*10*) Lovelace has applied the Veneziano formula in a new way which succeeds in establishing a link with the Gell-Mann work on chiral symmetry. Thus, two papers which were strongly linked by co-citation were only weakly tied by bibliographic coupling, although they were clearly related in content.

Of the 36 possible couplings among the 9 papers in Table 2, there are 8 instances of high co-citation between papers (six or more times) in which there was no bibliographic coupling at all. In 11 cases of very strong co-citation (15 or more), eight had bibliographic coupling strengths of either one or zero (four had couplings of one, and four had zero). These results suggest that bibliographic coupling is a less reliable indication of subject

TABLE 2. Co-citation, bibliographic coupling, and direct citation between pairs of papers

	Gasiorowicz '59	Gell-Mann '62	Gell-Mann '64	Gell-Mann '68	Glashow '68	Lovelace '68	Veneziano '68	Weinberg '66	Wilson '69	
Bjorken '66	1 6 no	9 0 yes	7 1 yes	13 3 no	5 1 no	2 0 no	2 0 no	2 2 no	11 4 yes	co-citation bibl. coupling direct citation
Gasiorowicz '59		6 9 yes	9 2 yes	11 5 yes	10 8 yes	1 3 no	1 0 no	9 6 yes	1 17 no	
Gell-Mann '62			19 4 yes	32 0 yes	18 1 no	3 0 no	0 0 no	6 3 no	7 1 yes	
Gell-Mann '64				20 0 yes	15 0 yes	4 0 no	0 0 no	9 1 yes	6 1 no	
Gell-Mann '68					50 5 yes	7 0 no	7 0 no	19 1 no	18 4 yes	
Glashow '68						1 1 no	0 0 no	10 2 no	7 0 yes	
Lovelace '68							68 1 yes	21 1 yes	1 4 no	
Veneziano '68								17 0 no	1 0 no	
Weinberg '66									0 1 no	

Each paper in the table is indicated by the first author's last name and the year the paper was published. There are three entries for each pair of papers. The top entry in each box is the co-citation frequency (determined from the 1971 *SCI*), the second entry is the bibliographic coupling strength, and the third is a "yes" if the later paper cited the earlier one, and "no" if it did not.

similarity than co-citation, although co-citation may, in addition, reflect "semantic" relations among cited papers, analogous to those observed in patterns of co-occurrence. Bibliographic coupling has been used to assemble groups of papers on particular subjects (*11*). Co-citation could also be used to form these groups after the papers have been cited, but the present data suggest that the results of these two grouping procedures would differ quite significantly.

A much better predictor of strong co-citation linkages between papers is provided by the direct citation patterns, that is, the citing of one paper by another. Figure 1 shows that the strongest co-citation links are between papers connected by direct citation. Of the 11 cases of strong co-citation (15 or more), eight are direct citation connections. In fact, all cases of direct citation are represented by co-citation strengths of at least 5, and many are much higher. The average co-citation frequency for papers connected by direct citation was 17.8, while the average for papers not connected by direct citation was 4.9. On the other hand, there are nine cases of co-citation strength of 6 or more which are *not* connected by direct citation. Thus, direct citation does not predict all strong co-citation linkages between papers in a field, although it more closely parallels such linkages than does bibliographic coupling.

• Conclusions

It apepars that an interpretation of the significance of strong co-citation links must rely both on the notion of subject similarity and on the association or co-occurrence of ideas. We have established, at any rate, that this measure in many cases reflects the existence of direct citation links and corresponds to significant intellectual connections within the field. Two applications in information retrieval are immediately apparent. A secondary index based on highly co-cited papers would allow sequential searches through a citation index, retrieving a list of new documents at each co-cited entry point. Second, co-citation could be used to established a cluster or "core" of earlier literature for a particular specialty. This 'core" could serve as a "profile" for that specialty and, thus, the basis of an SDI system.

Another area for the application of co-citation is in the study of the specialty structure of science (*12*). The pattern of linkages among key papers establishes a structure or map for the specialty which may then be observed to change through time. Through the study of these changing structures, co-citation provides a tool for monitoring the development of scientific fields, and for assessing the degree of interrelationship among specialties.

• **Acknowledgement**

I would like to thank Professor Belver Griffith of Drexel University and Drs. Eugene Garfield and Morton Malin, Institute for Scientific Information, for discussions and valuable suggestions.

References

1. Kessler, M. M., "Bibliographic Coupling Between Scientific Papers," *American Documentation*, 14 (No. 1): 10–25 (1963).
2. Garfield, E., "Citation Indexes for Science," *Science*, 122: 108–111 (1955).
3. Salton, G., "Associative Document Retrieval Techniques using Bibliographic Information," *Journal of the ACM*, 10 (No. 4): 440–457 (1963).
4. Price, N. and S. Schiminovich, "A Clustering Experiment: First Step Towards a Computer-Generated Classification Scheme," *Information Storage and Retrieval*, 4: 271–280 (1968).
5. Stevens, M. E., V. E. Giuliano and L. B. Heilprin, Eds., *Statistical Association Methods for Mechanized Documentation* (National Bureau of Standards Miscellaneous Publications 269, 1965).
6. We can also give a more formal definition of co-citation in terms of set theory notation. If A is the set of papers which cites document *a* and B is the set which cites *b*, then A∩B is the set which cites both *a* and *b*. The number of elements in A∩B, that is n(A∩B), is the co-citation frequency. The relative co-citation frequency could be defined as n(A∩B) ÷ n(A∪B).
7. Garfield, E., "Citation Indexing, Historio-Bibliography, and the Sociology of Science," *Proceedings of the Third International Congress of Medical Librarianship*, Amsterdam, 5–9 May 1969, (Excerpta Medica 1970) pp. 187–204.
8. Cawkell, A. E., "Search Strategies using the Science Citation Index," in *Computer Based Information Retrieval Systems*, B. Hougton, Ed. (London, 1968) pp. 27–44.
9. Gell-Mann, M., R. J. Oakes and B. Renner, "Behavior of Current Divergencies under $SU_3 \times SU_3$," *The Physical Review*, 175 (No. 5): 2195–2199 (1968).
10. Lovelace, C., "A Novel Application of Regge Trajectories," *Physics Letters*, 28B (No. 4): 264–268 (1968).
11. Schiminovich, S., "Automatic Classification and Retrieval of Documents by Means of a Bibliographic Pattern Recognition Discovery Algorithm," *Information Storage and Retrieval*, 6: 417–435 (1971).
12. Programs are currently being developed for automatically determining co-citation and using these data as input to a cluster generating program. The results of some preliminary experiments using this system to examine scientific specialties will be reported in a forthcoming paper.