Social Networks 11 (1989) 39-63
North-Holland

CONNECTIVITY IN A CITATION NETWORK: THE DEVELOPMENT OF DNA THEORY *

Norman P. HUMMON and Patrick DOREIAN

University of Pittsburgh **

The study of citation networks for both articles and journals is routine. In general, these analyses proceed by considering the similarity of articles or journals and submitting the set of similarity measures to some clustering or scaling procedure. Two common methods are found in bibliometric coupling, where two citing articles are similar to the extent they cite the same literature, and co-citation analysis where cited articles are similar to the extent they are cited by the same citing articles. Methods based on structural and regular equivalence also seek to partition the article based on their positional location. Such methods have in common focus on the articles and partitions of them. We propose a quite different approach where the connective threads through a network are preserved and the focus is on the links in the network rather than on the nodes. Variants of the depth first search algorithm are used to detect and represent the mainstream of the literature of a clearly delineated area of scientific research. The specific citation network is one that consists of ties among the key events and papers that lead to the discovery and modeling of DNA together with the final experimental confirmation of its representation.

Introduction

In this paper, we develop new methods for analyzing the connectivity in directed networks. These methods are based on search algorithms, primarily *depth first search,* and an important variant, *exhaustive search.* Using these methods to analyze a citation network describing the development of DNA theory, we identify a set of papers that played a central role in the development of that theory. These papers are identified through their structural connectivity in the network. Our approach to the analysis of connectivity is to focus on sequences of links and nodes, called search paths. The properties of search paths are used to quantify various dimensions of connectivity.

* Revised version of a paper presented at the Eighth Annual Sunbelt Conference for Social Network Analysis, San Diego, California.

** Department of Sociology, University of Pittsburgh, Pittsburg, PA 15260, U.S.A.

0378-8733/89/$3.50

Analyzing the connectivity of the DNA research literature

Since the pioneering work of Garfield (e.g., 1955) and its strong reinforcement by Price (1965) one decade later, the idea of analyzing networks among scientific events has become common place. Most often, the events are scientific productions linked by citation so that citation analyses, of one sort or another, are legion. "(T)he citation is a precise, unambiguous representation of a subject that requires no interpretation and is immune to change in terminology" (Garfield 1979: 3). Citations are explicit linkages between papers that have some important content in common. Indeed, the idea of papers being linked by citation forms the foundation on which the construction of the *Science Citation Index* rests. While there are problems with taking citations at face value (e.g., self-citation, negative citation, window dressing and politically motivated flattery) there is a strong correlation between citation rates and peer judgments that holds for many disciplines (Garfield 1979: 63). There are many benefits that stem from analyzing the citation links between articles and, in aggregation, between journals. It is possible to map the intellectual content of field and demarcate their (porous) boundaries. Interaction between fields can be studied and an historical account of development of scientific thought can be constructed.

The majority of citation studies, for our purposes, can be grouped into two broad categories: those measuring the prominence, or importance, of publications and journals (within networks); and those analyzing the structure of citation networks. Citation counts provide some indication of the utility of a scientific production (and the journals containing these productions). The emphasis on measures attached to papers extends to journals and the *Journal Citation Reports* give immediacy indices, half lives, and impact factors for journals. More sophisticated and complicated measures based on the eigen structure of a normalized citation matrix have been constructed (Pinski and Narin 1979; Doreian 1987), although Noma (1987) suggests that these indices add little to the information contained in raw citation counts. For all of these measures, the structure of the network from which they are constructed remains implicit or secondary. The primary goal is a set of measures for *nodes.*

Structural analyses of citation networks emerge when the patterns of specific network relations are considered. The obvious relations are "cites" and its converse "cited" and various graphs depicting these

relationships can be constructed for a relatively small set of scientific events. [1] Garfield (1979: 81-97) reports on a set of these historiographs which includes the DNA citation network we analyze in this paper. [2]

Analysis of the structure of the network regardless of whether it is the relation "cites", or "cited by", or bibliometric coupling (Kessler 1963), or co-citation (Small 1973) focus on the *clustering of nodes* (be they articles, journals, or scientists) on the basis of ties connecting them. Our purpose here is to present and use a set of methods that focus on the *links* of the network rather than on the nodes.

The DNA citation network

The data for our connectivity analysis are taken from Garfield, Sher and Torpie (1964) and partially summarized by Garfield (1979). For

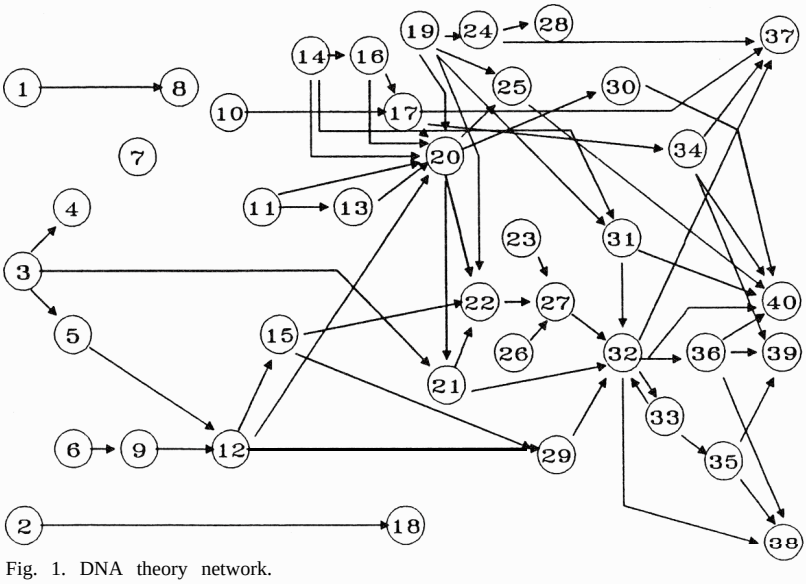


Fig. 1. DNA theory network.

[1] In linked-list form very large networks can be sorted and analyzed but their graphs are too large to draw and interpret.

[2] Of course, a primary, and crucial objective is to ensure. that the historiograph is accurate. Too much error or noise in the historiograph, or any network for that matter. compromises any analysis of its structure.

Table 1
The 40 milestone events of DNA research from Garfield *et* al. (1964) after Asimov (1963)

Event	Date	Author(s)	Discovery
1	1820	Bracconot	Isolation of specific amino acids from protein.
2	1860s	Mendel	Predictability of dominant and recessive traits in plants.
3	1869	Miescher	Isolation of nucleic acid.
4	1880	Flemming	Described replication of paired chromosomes within the cell nucleus.
5	1886	Kossel	Study of purine and pyrimidine content of nucleic acid.
6	1891	Fischer & Piloty	Isolation and synthesization of ribose as a freely occurring sugar.
7	1900	DeVries	Concept that spontaneous alteration of the chromosome can lead to mutation.
8	1900–10	Fischer	Demonstration of the peptide chemical linkage of amino acids forming protein.
9	1909	Levene	Identified the 5 carbon sugar ribose as a component of nucleic acid.
10	1926	Muller	Produced altered genes and mutants with X-rays.
11	1928	Griffith	Production of living capsulated bacteria from dead capsulated pneumococci.
12	1929	Levene	Discovery that certain nucleic acids contain deoxyribose (DNA).
13	1931	Alloway	Proof that genetic material from a dead strain influences characteristics of a live strain.
14	1935	Stanley	Isolated crystals of tobacco-mosaic virus.
15	1935	Levene	Proposed formulae assigning linkages between the nucleotides.
16	1936	Bawden & Pirie	Discovered the virus (cf. 14) was also a nucleoprotein.
17	1938	Caspersson & Schulz	RNA concentration is highest in cells where the rate of protein synthesis is highest.
18	1941	Beadle & Talum	Via X-rays produced mutant molds requiring precise amino acid supplementation.

19	1944	Martin & Synge	Development method of paper chromatographic separation of amino acids.
20	1944	Avery et al.	Discovered DNA carried genetic information that can change a strain into another.
21	1947	Chargaff	Purines and pyrimidines present in unequal quantities within nucleic acids.
22	1950	Chargaff	Different nucleotides in the chain are in random order.
23	1951	Pauling & Corey	Concept of polypeptides chains in a helical configuration.
24	1953	Sanger	Determined the amino acid sequence of insulin.
25	1952	Hershey & Chase	Nucleic acid portion of bacteriophage virus enters cell – not the protein shell.
26	1953	Wilkins	Developed X-ray diffraction methods for studies of nucleic acid.
27	1953	Watson & Crick	Constructed model of spatial molecular configuration of DNA (via method of 26).
28	1953	DuVigneaud	Extended 24 to determine amino acid sequence of oxytocin and vasopressin.
29	1955	Todd	Confirmed Levene's (15) formulae through chemical synthesis.
30	1953	Pallade	Discovered smaller particles associated with the microsomal fraction.
31	1955	Fraenlal-Conrat	Separated nucleic acid and protein shell of tobacco–mosaic virus.
32	1955	Ochoa	Isolated a bacterial enzyme producing polynucleotide strands of RNA.
33	1956	Kornberg	Produced synthetic polynucleotides of RNA from an enzyme.
34	1957–8	Hoagland	Demonstration of transfer RNA as a triplet code.
35	1960	Jacob & Monod	Discovered existence of second (Messenger) RNA.
36	1961	Hurwitz	Manufactured Messenger RNA in test tube (from DNA, nucleotides, enzymes).
37	1961	Dintzis	Demonstrated concept of protein construction (in 34) was accurate.
38	1961	Norvelli	Extended 36 via DNA nucleotides, ribosomes and amino acids.
39	1962	Mirsky & Allbrey	Messenger RNA isolated from mammalian cells.
40	1961	Nirenberg & Maltaei	Ultimate verification of triplet code (using method of 32).

Garfield *et* al. (1964: iii) "the history of science is regarded as a chronological sequence of events in which each new discovery is *dependent* upon earlier discoveries" (emphasis added). Successful research is seen as a sequence of important events having a time ordered sequence where later work is critically dependent on earlier work and important scientific goals are achieved. Garfield *et* al. constructed two historiographs stemming from Asimov's (1963) account of the history of DNA work in *The Genetic Code.* A carefully identified set of 65 specific research productions, cited by Asimov in his historical account, were examinated. These productions were grouped into 40 milestone events and the constructed historiographs represent these events and the ties that link them. One is reconstructed from Asimov's narrative while the other was obtained by carefully examining the citations made in each bibliography. [3] We analyze the second historiograph. Figure 1 contains the citation-based historiograph constructed by Garfield *et al.* (1964: transparencies) and reviewed by Garfield (1979: 88). The relation depicted is "cites". For each linked pair, the event on the right cites the event to the left. [4] Table 1 lists the 40 milestone events between 1820 and 1962.

Each event was categorized and coded into broad subject categories: nucleic acid chemistry (NC), protein chemistry (PC), genetics (G), and microbiology (M). The coding for each milestone event is shown in Table 2. [5]

Our purpose is to examine the links in this network with a view to finding the main stream of research through that network.

[3] If events A and B are among the milestone events, then if A cites B there is a direct connection between the events. If C is not a nodal event, but it is the production of an author of a nodal event, then if A cites C and C cites B there is a less direct link. Finally, if C is written by an author not represented in any of the milestone events and the pattern is A cites C and C cites B it is an even less direct link. In our analysis, we do not distinguish between direct and less direct citations.

[4] There is a 2-cycle between 32 (Ochea) and 33 (Komberg). The arrow in Figure 1 represents the flow of useful information rather than cites.

[5] Garfield et al. note (1964: 15) the need for postgraduate level training in the field being analyzed and, as they had project members with this training, we assume that reasonable decisions were made throughout their analysis and that we can rest our work on their painstaking sifting of the relevant citations.

Connectivity procedures

1. Weakly connected subgraphs

The first task is to examine the citation network to determine whether it has distinct subgraphs. Simple or weak connectivity in a directed graph can be determined by symmetrizing the network, and performing a depth first search. The DNA network contains four weakly connected subgraphs, as can be easily seen in Fig. 1.

Subgraph 1
1 8
Subgraph 2
2 18
Subgraph 3
3 4 5 12 9 5 15 22 19 24
28 37 17 10 16 14 20 11 13 21
32 27 23 26 29 31 40 25 30 34
39 35 33 38 36
Subgraph 4
7

It is clear that the main root for this network starts with node 3, the paper by Meischer written in 1871. The other subgraphs are small, containing only one or two nodes. [6]

2. Strongly connected subgraphs (cycles), and sorts

Depth first search can also find strongly connected subgraphs, or cycles. If a directed graph is also a DAG, it can be sorted using the depth first search algorithm. This sort is a topological sort and orders the nodes so that no node is before a node that points to it. Topological sorts of DAGs are, in general, not unique.

The DNA network contains only one cycle involving nodes 32 and 33. These two papers overlapped in time, as both have 1956 publication dates. Not surprisingly, citation networks are very nearly directed acyclic graphs, or DAGs.

Removal of one link (the citation of event 32 by event 33) from the DNA network, transforms it into a DAG. As the Watson and Crick

[6] While in this rather simple network, these weakly connected subgraphs can be picked out visually from Figure 1, we doubt that eyeballing techniques are useful for much larger networks.

research is cited by Ochoa (who cites two other events) and Komberg cites only Ochoa, this is a reasonable choice. [7] Given DAG, we can analyze either "cited-by" or "cites" (or both). Here, the network is transposed so the links are is-cited-by relations as this form of the network is directed forward through time, and represents the influence patterns and dependency relations in a citation network. The following topological sort order was found for the DNA network.

Sort Order DNA Network

26		
23		
19	24	28
14	16	31
11	13	
10	17	34
7		
6	9	
3	5	12
25		
21		
15	29	
22	27	32
36	40	
33	35	39
38		
4		
2	18	
	8	

The depth first search procedure outputs nodes in the listed order starting with 26, 23, 19, 24, etc. and ending with 4, 2, 18, 1, and 8. We have broken the sorted list into sub-sequences of connected paths so that the structure of the network is easier to see. Recall that the relation used to sort the network is one of dependency. Node B follows node A if it is dependent on node A. Figure 2 presents the dependency structure of the core of the network starting from the root node 3.

The core of the network begins with node 3 and ends with node 38. Nodes 26, 23, 6, 9, etc. are all prior to root node 3 because they have no dependency relation with node 3. Nodes 4, 2 and 18, and 1 and 8 are the separate subgraphs identified 'above. Note that the order of nodes 37, 36, 40 and 38 could be interchanged in the sort list changing the dependency relations. It is in this sense that the sort order of the

[7] **An alternative is to collapse events 32 and 33 into a single node.**

Table 2
The milestone events coded by research area

Event	Protein chemistry	Genetics	Nucleic chemistry	Microbiology
1	√	–		
2		√	–	–
3	–	–	√	
4	–	–	–	–
5		–	√	–
6	–		√	–
7			–	–
8	√	–	–	
9		–	√	–
10	–	√	–	–
11	–	√	–	–
12	–	–	√	–
13	–	√	–	–
14		–	–	√
15	–	–	√	–
16	–	–	√	√
17	√		√	–
18			–	–
19	√	–	–	–
20	–		√	–
21	–	–	√	–
22	–	–	√	–
23	√	–	–	–
24	√		–	–
25	–	–	√	√
26	–	–	√	
27	–	–	√	–
28	√	–	–	
29	–	–	√	–
30	√	–	√	
31		–		√
32	–	√	√	–
33		√	√	–
34	√			–
35		√		–
36	–		√	–
37	√		√	
38	√	–	√	
39			√	
40				

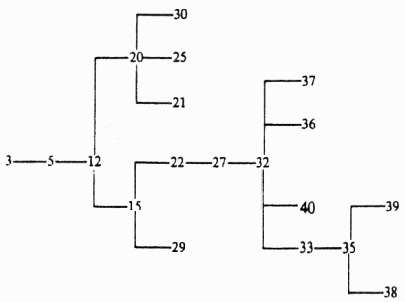


Fig. 2. Dependency structure of core of network (based on topological sort).

network is not unique; a very large number of alternative lists could be constructed using such changes.

3. *Connectivity and path lengths*

Another way to examine connectivity in these networks is to compute the path distances between node pairs. The distance of most interest in analyzing connectivity is the *longest path* or detour between each pair of nodes. This computation is accomplished with an important variant on the depth first search algorithm, the exhaustive search algorithm. Briefly, the exhaustive search algorithm finds all possible search paths through the network. To compute the maximum distance, it is only necessary to record the maximum distance for each node pair across all possible search paths emanating from the start node.

The maximum path distance for the DNA network is 10 links. Six out of 676 search paths are of length 10, including two search paths emanating from the root node. The longest paths from the root node to all other nodes are given in Table 3.

Table 3
DNA Network

Maximum path distance from node 3							
Node	Distance	Node	Distance	Node	Distance	Node	Distance
4	1	5	1	12	2	15	3
20	3	21	4	22	5	25	4
27	6	29	4	30	4	32	7
33	8	35	9	36	8	37	8
38	10	39	10	40	9		

Table 4
Main path sort order compared with longest path distance

Node	Order	Node	Order	Node	Order	Node	Order
3	0	5	1	12	2	20	3
30	4	25	4	21	4	15	3
29	4	22	5	27	6	32	7
37	8	36	8	40	9	33	8
35	9	39	10	38	10		

The longest path vector corresponds very closely to DAG sort order reported above. Repeating that portion of the sort order represented in Figure 2, we see that only the longest path distances for nodes 15 and 33 depart from the expected order, and both these are at distances only one link different from their implied sort order values in Table 4.

Figure 2 shows that both nodes 15 and 33 could be interchanged with other nodes to change their order to the expected value. However, such changes would move other nodes out of their expected order.

4. Network connectivity and search paths

We propose a new index of link connectivity based on the measure of traversal counts in search paths through the network. The construct this index, we propose three related operationalizations.

First, suppose we extract the subgraph from the network in Figure 1 that represented all possible paths from node 3 to node 22. It would look like Figure 3.

If there are N nodes in the subgraph, there exist N(N − 1) possible subgraphs connecting all directed node pairs in the network. For the graph in Figure 3, only 19 of the possible 42 node pairs have connect-

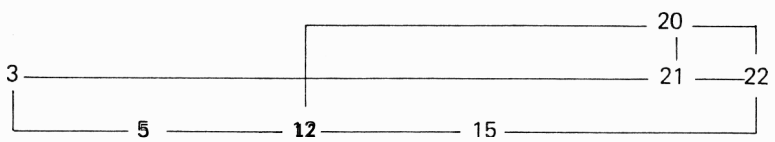


Fig. 3. Graph of all links from node 3 to node 22.

ing links and paths. Thus, we can construct subgraphs that connect the node pairs:

3-5	3-12	3-15	3-20	3-21	3-22
5-12	5-15	5-20	5-21	5-22	
12-15	12-20	12-21	12-22		
15-22					
20-21	20-22				
21-22					

To compute the traversal counts for each link, we construct the adjacency matrices for all the subgraphs connecting these node pairs. These matrixes can be "stacked" by corresponding row and column nodes. The traversal counts of interest are the projected counts of all links connecting node pairs projected onto a base matrix. [8] The resulting projection matrix contains counts of the number of times each link was involved in connecting all node pairs using all subgraphs derived from the network. We call this the ***node pair projection count*** (NPPC) method of generating traversal counts. The network is presented as a graph valued by traversal counts. For Figure 3 this is:

```
 3:  5 (6) 21 {2}
 5: 12 (10)
12: 15 (6) 20 {9}
15: 22 (4)
20: 21 {8} 22 (4)
21: 22 (5)
```

The link with the highest traversal count of 10 is 5-12. This means that this link was a member of 10 of the 19 subgraphs that connect all node pairs. Links that bypass several nodes, such as 3-21 generate low traversal counts. Traversal counts reflect the connectivity that both precedes and follows a link in a search path.

We propose two other ways of computing link traversal counts, and both are based on the exhaustive search algorithm. As noted above, this algorithm generates all possible search paths through the network emanating from an origin node. The count of the number of times a link is traversed by all possible search paths is a simple way measure of the importance of that link. We label this the ***search path link count***

[8] Cf. projecting back along the time axis to the phase space for a dynamic system.

(SPLC) method of computing traversal counts. The traversal counts computed by the SPLC method for the simple network in Figure 3 are:

```
 3:   5 (3)  21 {1}
 5:  12 (6)
12:  15 (3)  20 (6)
15:  22 (4)
20:  21 (4)  22 (4)
21:  22 (6)
22:
```

The third method is also based on the set of all search paths emanating from a start node, but instead of simple link counts, it accounts for all connected node pairs along *the search paths*. Thus a link in the middle of a search path will receive a higher traversal count than links at the ends of the search path because "inner" links are involved in connecting more node pairs than links at the beginning or the end of a search path. We label this the *search path node pair* (SPNP) method of computing traversal counts. The traversal counts using the SPNP method for the network in Figure 3 are:

```
 3:   5 (13) 21 (2)
 5: 12 (20)
12:  15 (6) 20 (15)
15:  22 {4}
20:  21 (8) 22 (4)
21:  22 (6)
22:
```

These traversal counts are analogous to counts of the number of geodesics that run through a node in Freeman's (1978) centrality measure. However, we are concerned with the connectivity of links rather than the centrality of nodes. There is an obvious duality between the centrality of nodes and the connectivity of links.

For the DNA network, the maximum traversal counts and links for the three methods are: search path link count for link 27-32 is 328; search path node pair for link 22-27 is 1178; and node pair projection method for link 27-32 is 152. All three methods identify the same pair of links with the two highest counts in the network. The 22-27 link connects the Chargaff paper to the Watson and Crick paper, and the 27-32 link connects the Watson and Crick to the Ochoa paper.

These traversal counts can be used in another important way: they define the *main path* through the citation network. We can use the traversal counts to determine a search path through the network that

reflects the greatest connectivity in the network. At any node, we choose the next link in the path as the outgoing link with highest traversal count. By repeatedly applying this choice rule, we define a path through the network that follows a structurally determined ***most used*** path. This link selection technique is an example of a priority ***first search*** algorithm, where the priority is set by the traversal counts. It is our intuition that the main path, selected on the basis of the most used path will identify the main stream of a literature.

For the DNA network, a traversal count priority first search identifies the same search path for all three methods of generating traversal counts. Thus the DNA network main path contains the following

Node	Paper
3	Miescher, 1871
5	Kossel, 1886
12	Levene with Mori and London, 1929
20	Avery, MacLeod and McCarty, 1944
21	Chargaff, 1947
22	Chargaff, 1950
27	Watson and Crick, 1953
32	Ochoa, 1955-1956
36	Hurwitz, 1960
40	Nirenberg and Matthaei, 1961-1962

Figure 4 shows the main path superimposed over the network.

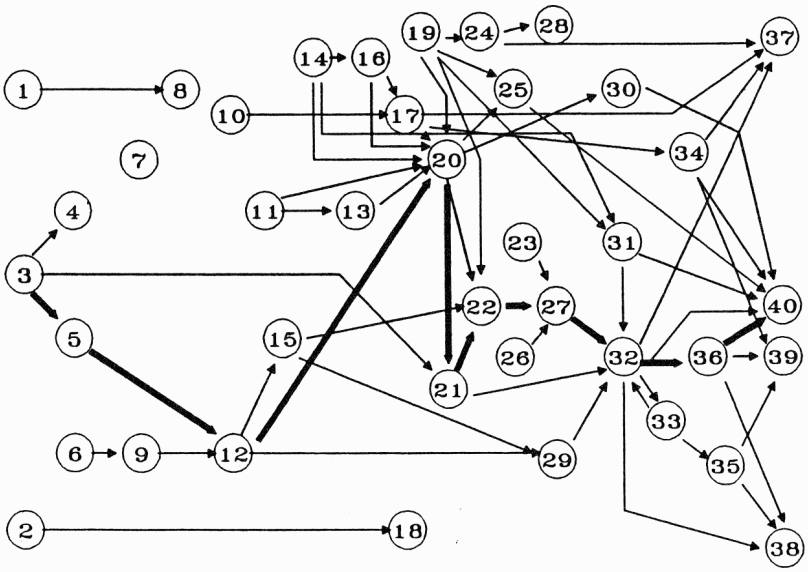


Fig. 4. DNA main stream.

5'. The network of main paths

The analysis reported above reports the main path starting with node 3. There are sound reasons for this choice, both from other connectivity analyses, and other findings reported below. However, what happens if this assumption is relaxed, and all nodes in the network are selected as start nodes of a main path analysis? Table 5 reports the results of such an analysis.

Table 5
The set of all main paths in the DNA network (based on SPNP traversal counts)

Start node	Main path
1	8
2	18
3	51220212227323640
5	12 20 2122 27 32 36 *40*
6	91220212227 32 3640
9	1220212227 32 3640
10	1720212227 323640
11	1320212227 32 3640
12	20212227 323640
13	20212221323640
14	1617 2021222732 3640
1s	2227 32 3640
16	1720212227 323640
17	20212221323640
19	2227323640
20	212221323640
21	2227 323640
22	27 32 3640
23	27 32 3640
24	28
25	*40*
26	27 32 3640
27	323640
29	32 3640
30	*40*
31	32 36 *40*
32	36 40
33	35 38
34	*40*
35	38
36	*40*

Thirty-one of the forty nodes have outgoing links, and therefore can be used as start nodes for the analysis. Table 5 shows that, when considering the node specific main paths, only four nodes result in main paths that do not join the main path that emanates from node three. Of these four, two are part of subgraphs that are not even weakly connected to the main network: these are nodes 1 and 2. The remaining nodes, 24 and 35, are directly connected to terminal nodes 28 and 38 respectively, and none of the main path nodes are reachable from these nodes. Thus, virtually all start nodes have main paths that converge to the main path from node 3, and for the exceptions it is impossible to reach the main path. Connectivity in this citation network converges to the main stream of this literature.

We now discuss the significance of this sequence of scientific events from other perspectives.

Corroborative evidence

Citation importance

DNA was discovered in 1869 by Miescher (node 3) so that this node has to be non the main path: it is.

Asimov (1963) identified event 20 (the discovery by Avery *et al.* that dexyribonucleic acid (DNA) carries genetic information that was capable of transforming one bacteria strain into another from which the DNA was extracted) as truly critical. Garfield *et al.* (1964) constructed a weighting scheme. so that events can be measured in terms of the types of ties (direct, strong indirect, and weak indirect with each successive type of tie being weighted less) incident to them. According to their index, event 20 has the greatest nodal weighting. It is reasonable to expect that such a key event be on the main path: it is. We know that Watson and Crick (event 27), together with Wilkins, shared a Nobel Prize for their work and that Ochoa (event 32) shared a Nobel Prize with Komberg (event 33). Such prize winning events should be on the main path: they are. [9] Garfield *et al.* (1964: Appendix II) give the

[9] Events 33 and 32 have a reciprocal link between them. As Ochoa's work was drawn on explicitly by event 40, our main path has only event 32. Wilkins (node 26) shared the Nobel Prize with Watson and Crick (node 27). His development of X-ray diffraction was critically important and the "main path" from node 26 leads immediately to the main path of the network. *Methodological* breakthroughs and new conceptual developments (node 23) may have this ancillary standing of being just off the main path in citation networks.

nodal weighting value for each nodal event. Starting with the highest weighting value and stopping when one of the terminal nodes is encountered, the set of nodes with the highest weighting values are, in descending order of the weighting index, event 20, event 32, event 22, event 21, event 36, event 27, and event 40. A strong criterion of adequacy is that all of these nodes with the highest nodal index be on the main path: all of them are.

The dominant field via Q-analysis

The data in Table 2 contain the coding of the milestone events by research area. Let A be matrix obtained by having a value of 1 where there are check-marks and a value of 0 where there are dashes. Atkin (1977) presents a systematic methodology whereby the structure of such a matrix can be explored. Each research area becomes simply the set of events containing its content while each event is represented by the research areas it contains. Each research area becomes a simplex and together they form a family of simplexes. Similarly, each event is a simplex made up of its own content areas and together they form a family of simplexes. These, together with their faces, form dual simplicial complexes. We consider the one concerning the family of research areas.

With A the data matrix, its transpose is A'. The product A'A gives a 4 **x** 4 matrix expressing the extent to which the research areas overlap. The overlap is the number of events containing both research areas and is referred to as a (common) face. A (modified) [10] Q-analysis consists of an exploration of the connective structure of common faces. Table 6 gives the relevant details.

The top panel of Table 6 shows the shared face matrix (where the entries express the size of the overlap rather than the dimension of the face). Nucleic chemistry is found in 26 of the 40 milestone events. Similarly, protein chemistry is found in 12 events, genetics in 11 events, and microbiology in 4 events. The off diagonal elements give the number of events where pairs of research areas are both present. Thus, of the 26 nucleic chemistry events, six share protein chemistry content, while five share genetic content. Similarly, nucleic chemistry shares

[10] If U is a 4 x 4 matrix of ones, a Q-analysis uses the matrix $\Lambda'\Lambda - U$. We ignore U, but continue to use q as a notation.

Table 6
Q-analysis of DNA content complex

(a) *Shared face matrix*

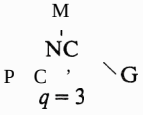
	PC	G	NC	M
Protein chemistry PC	12			
Genetics G	1	11		
Nucleic chemistry NC	6	5	26	
Microbiology M	0	0	3	4

(b) *Q-analysis*

Values of q	Complexes	Qq
26-13	{NC}	1
12	{NC}{PC}	2
11-7	{NC}{PC}{G}	3
6	{NC,PC}{G}	2
5	{NC,PC,G}	1
4	{NC,PC,G}{M}	2
3	{NC,PC,G,M}	1

(c) Graphs of equivalence classes

PC-NC-G
$q = 5$.



three events with microbiology. A Q-analysis proceeds by considering progressively smaller values in the shared face matrix. For a given value of q the complexes present in the Q-analysis form equivalence classes. For values of q = 26 through 13 there is a single equivalence class made up of nucleic chemistry. For q = 12 there are two separate equivalence classes made up of nucleic chemistry and protein chemistry. These equivalence classes are shown in the middle panel of Table 6 and the right hand column simply counts the number of equivalence classes present. For q = 11 through 7 there are three equivalence classes each containing a singleton: nucleic chemistry, protein chemistry and genetics. For q = 6 nucleic chemistry and protein chemistry join (as there are 6 events sharing their content) while genetics remains in a separate equivalence class. For q = 5 all three join in a single equivalence class. For q = 4 the area of microbiology joins the set of complexes but it is disjoint. Finally, for a q = 3 all areas are linked together in a single equivalence class. The third panel of Table 6 gives pictorial represen-

tations of the single equivalence classes found for $q = 5$ and for $q = 3$. For $q = 5$, nucleic chemistry is the subject area linked to both protein chemistry and genetics. For $q = 3$ nucleic chemistry is again the core having separate links to protein chemistry, genetics and microbiology. None of the other areas are connected directly. It is very clear that nucleic chemistry forms the core of the DNA content complex. Given its importance, it ought to be present on the main path. Indeed, *all* of the events of the main path have nucleic chemistry as all or part of their content.

Both the weights assigned by Garfield *et al.* and the central core identified via a (modified) Q-analysis provide corroborative evidence for the identification of the main path. The seven nodes with the highest nodal weighting are all on the main path, and with nucleic chemistry being identified as the core of the research specialty.

A comparison with equivalence approaches

We have shown that other paths, beyond the main path, can be picked out in the network. Event 2 could be taken as a start point although the path would reach only event 18. Each of events 10 and 11 can be taken as a start point and paths can be traced that reach one or more of the terminal events. The emphasis on connectivity is on the strands that *connect* the research productions that cumulate in a clearly identified research specialty. The objectives of clustering citation networks is quite different and can be seen as a complementary analysis.

There are two broad approaches to the clustering of citation networks: one is found in citation analysis; the other stems from the idea of equivalence found in the social network literature. In the citation analysis literature, the citation relation has been clustered directly (for example, Narin *et al.* 1972; Carpenter and Narin 1973).

Two further network relations have been constructed from these citation patterns. Kessler (1963) suggested the idea of bibliometric coupling whereby two articles are similar to the extent that they share citation to common sources. As defined, a bibliometric coupling relation is entirely static and the similarity is defined in terms of the authors' bibliographies. Yet scientific fields are dynamic and Small (1973) defined a new form of document coupling in the form of co-citation to, in part, reflect this. Two articles are similar to the extent

they are jointly cited by subsequent authors. "Co-citation patterns change as the interests and intellectual patterns of the field change" (Small 1973: 265). Moreover, they "can be used to map out in great deal the relationship between key ideas" (Small 1973: 266). It is primarily through the creative use of co-citation analysis that the content of fields can be mapped and the interactions between them studied. Longitudinal study of co-citation analysis permits a dynamic picture depicting the emergence of fields and specialties.

Once co-citation has been defined and measured for pairs of articles in a network, the major data analytic tool is cluster analysis. Moreover, by virtue of the gigantic size of the networks usually considered, computational constraints have limited the analyses primarily to single link clustering, even though it is prone to producing string-like clusters. Nevertheless, Small and Griffith (1974) and Griffith et al. (1974) employed this tool to map out the structure of science in terms of its diverse fields and specialties. Sets of coherent clusters were established, solely on the basis of the pattern of co-citation, that correspond to clearly defined scientific areas. The mosaic of science is depicted in terms of clusters, each representing a research specialty or field, and a network linking the clusters. Small's (1977) longitudinal study of the collagen research specialty showed the emergence and change of a specific field. Many other co-citation analyses have been conducted and the idea of co-citation has been extended to co-word analysis (e.g. Callon et al. 1983) whereby scientific productions are similar to the extent they share the same key words, or words in their titles.

In the social network analysis formulation of equivalence considerable attention has been focussed on representing and homomorphically reducing networks. The initial formulation is found in Lorrain and White (1971) where, loosely, two objects are structurally equivalent if they are connected to exactly the same other objects. If we consider the relation cited by, then structural equivalence maps directly into co-citation: two articles are structurally equivalent if they are cited by the same other articles. [11] With regard to citing, structural equivalence appears to map into bibliometric coupling. Both the relations, cites and cited, can be analyzed simultaneously, in which case articles are structurally equivalent if they cite the same sources and are cited by the same

[11] Indeed, Small's measure of the extend to co-citation could be a useful measure of structural equivalence.

subsequent articles. Block modeling tools have been used in the sociology of science (for example, Breiger 1976 and Lenoir 1979) suggests their use in co-citation analysis. This being the case, algorithms such as CONCOR (Breiger *et* al. 1975) or STRUCTURE (Burt 1976, 1987) may be useful. In the context of journal networks, Doreian and Fararo (1985) and Doreian (1985) have used Burt's algorithm to cluster journals into coherent clusters. A generalization of structural equivalence is regular equivalence where, again loosely, two objects are regularly equivalent if they are connected in equivalent ways to equivalent objects (White and Reitz 1983). Such a generalization may be particularly useful for studying multidisciplinary networks.

Given the time ordering of the citations and that, with the exception of a single two-cycle, the network is a directed acyclic graph with multiple start points, regular equivalence is the most relevant definition. The expectation is that such an analysis will pick out intellectual generations in the citation network. Figure 5 gives the regular equivalence dendrogram for the citation network of Figure 1.

Event 7 is picked out as the isolated node in the network. Events 1, 2, 3, 6, 10, 11, 14, 23, and 26 belong to a cluster that can be viewed as the first generation events for the research area. All are events that can be viewed as bringing genuinely new information to the research area. It includes all of the early events such as the Bracconot event in 1820. It also includes event 23 where Pauling and Corey propose the concept of a helical configuration for polypeptides chains and Wilkins' development of X-ray diffraction methods for the study of nucleic acid. Both are events that Watson and Crick build upon in constructing their critically important spatial model of DNA.

A second cluster is made up of events for, 8, 18, 28, 37, 38, 39, and 40. All can be viewed as terminal events and they appear to fall into two categories: capstones and deadends. It appears that the research of event 8 lead nowhere according to the citation network, as did research contained in events 4, 18, and 28 which can all be viewed as deadends. However, the research contained in events 37, 38, 39, and 40 can all be seen as capstones on which future research can be built.

A third cluster is made up of the events 24, 25, 32, 33, 34, 35, and 36. With one exception (event 33) these form the last but one generation as they are one step away from terminal nodes. The further cluster of events can be viewed, in the main, as the second generation of events one step away from the first generation events. In the middle portion of

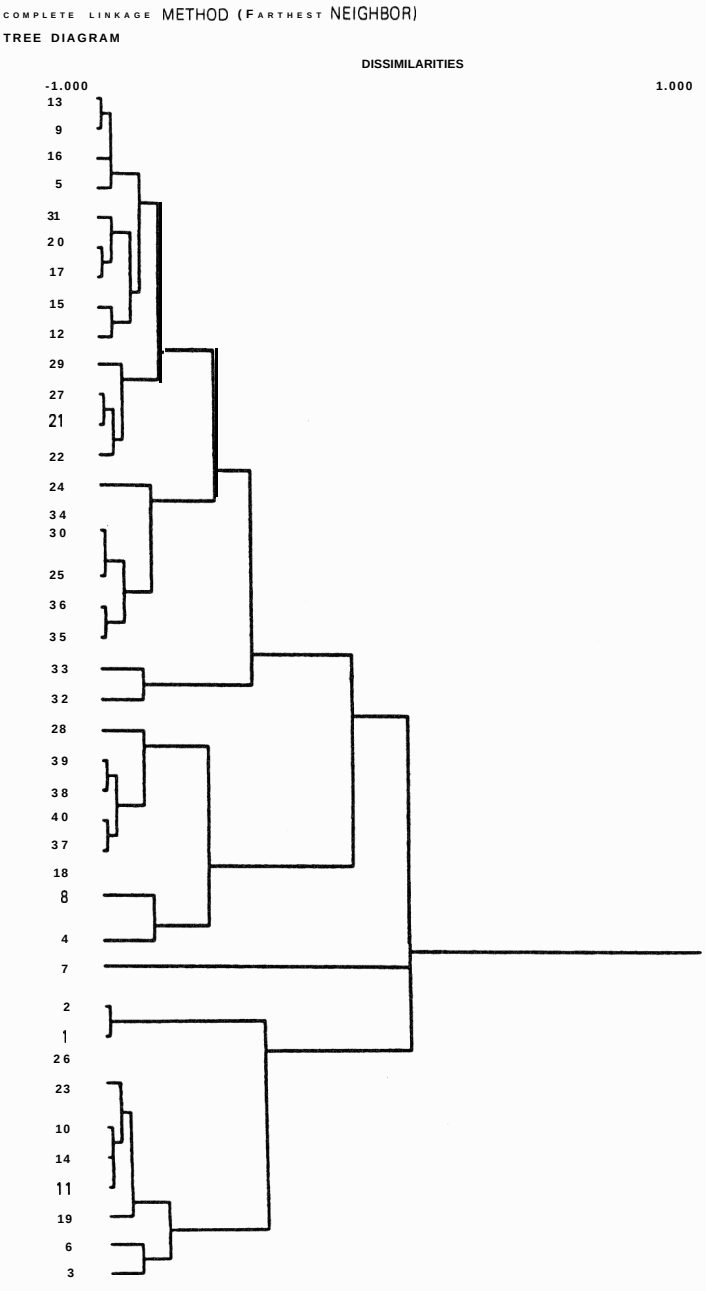


Fig. 5. Dendrogram for regular equivalence clustering.

the network there is some confusion as different events are at different distances from the initial and terminal events. Regular equivalence does pick out the intellectual generations, ignoring the time dimension – so that Wilkins' work of 1953 is regularly equivalent to Miescher of 1869 – in a fairly coherent fashion. It is also clear that partitioning the network in terms of intellectual generations collapse together the paths that can be traced from the initial work to the terminal work.

Conclusion

The citation network that describes the important events in the development of DNA theory has been subjected to a variety of analyses. First, the network was sorted, and an ordered set of nodes were identified by their dependency relations. This set begins with node 3 and includes nodes to all the other nodes on the main path. Next, the longest path analysis was reported, and the same basic set of nodes was identified. Third, the connectivity measure of traversal counts was introduced. The highest traversal counts fall along this same connected set of nodes. To formalize this observation, a priority first search was used to trace the main path from node 3. Finally, when all possible main paths were generated, there is a tendency for connectivity to converge the main path, and then follow it.

Using a completely different methodology, other researchers -had already identified the events that make up the main path as the most significant in the development of DNA theory. Thus our formal connectivity analysis employing network search techniques is completely consistent with the analysis of other researchers.

In summary, three widely different methodologies come to the same general conclusion about the structure of the DNA network, and, more importantly, about the social process by which this exciting field developed.

References

Asimov, I.
1963 The *Genetic Code.* New York: New American Library
Atkin, R.H.
1977 ***Combinatorial Connectivities in Social*** Systems. Basel: Birkhauser.

Breiger, R.
1976 "Career attributes and network structure: A block model study of a biomedical research specialty." *American Sociological Review 41:* 118-135.

Breiger, R.L., S.A. Boorman and P. Arabie
1975 "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling." *Journal of Mathematical Psychology 12:* 328-383.

Burt, R.S.
1976 "Position in networks." *Social Forces 55: 93-122.*

Burt, R.S.
1987 "Structure: Sociometric indices, cliques, structural equivalence, density tables, contagion, structural autonomy and equilibrium in multiple network systems," *Technical Report 3.2,* Research Program in Structural Analysis, Columbia University.

Callon, M., J.P. Courtial, W.A. Turner and S. Bauin
1983 "From translations to problematic networks: An introduction in co-word analysis." *Social Science Information 22:* 191-235.

Carpenter, M. and F. Narin
1973 "Clustering of scientific journals." *Journal of the American Society for Information Science 24: 425-436.*

Doreian, P.
1985 "Structural equivalence in a psychology journal network." *Journal of the American Society for Information Science 36:* 411-417.

Doreian, P.
1987 "A revised measure of standing for journals in journal networks." *Scientometrics II:* 71-80.

Doreian, P. and T.J. Fararo
1985 "Structural equivalence in a journal network." *Journal of the American Society for Information Sciences 36: 28-37.*

Freeman, L.C.
1978 "Centrality in social networks: I Conceptual clarification." *Social Networks 1: 215-239.*

Garfield, E.
1955 "Citation indexes for sciences." *Science 122 (3159):* 108–111.

Garfield, E.
1979 Citation Indexing – Its Theory and Application in Science, Technology, and Humanities. Philadelphia: Institute for Scientific Information Press.

Garfield, E., I. Sher and R.J. Torpie
1964 The Use of Citation Data in Writing the History of Science. Philadelphia: Institute for Scientific Information.

Griffith, B.C., H.G. Small, J.A. Stonehill and S. Dey
1974 "The structure of scientific literature, II: Toward a macro- and micro-structure for science." *Science Studies 4: 339-365.*

Kessler, M.M.
1963 "Bibliographic coupling between scientific papers." *American Documentation 14:* 10–25.

Lenoir, T.
1979 "Quantitative foundations for the sociology of science: On linking block modeling with co-citation analysis." *Social Studies of Science 9: 455-480.*

Lorrain, F. and H.C. White
1971 "Structural equivalence of individuals in social networks." *Journal of Mathematical Sociology I:* 49–80.

Narin, F., M. Carpenter and N.C. Berlt
1972 "Interrelationships of scientific journals." *Journal of the American Society for Information Science 23: 323-331.*

Noma, E.
1987 "Using influence weights to evaluate the scientific importance of journals" (preprint). Booz, Allen and Hamilton, Philadelphia.

Pinski, G. and F. Narin
1979 "Citation influence for journal aggregates of scientific publications: Theory with application to literature of physics." *Information Processing and Management 12:* 297-312.

Price, D.J.D.
1965 "Networks of scientific papers." *Science 149:* 510–515.

Small, H.G.
1973 "Co-citation in the scientific literature: A new measure of the relationship between two documents." *Journal of the American Society for Information Science 24: 265-269.*

Small, H.G.
1977 "A Co-citation model of a scientific specialty: A longitudinal study of collagen research." *Social Studies of Science 7:* 139-166.

Small, H.G. and B.C. Griffith
1974 "The structure of scientific literature, I: Identifying and graphing specialties." *Science Studies 4:* 17–40.

White, D.R. and K.P. Reitz
1983 "Grasp and semi-group homomorphisms on networks of relations." *Social Networks 5:* 193-234.