Visualizing Science by Citation Mapping

Henry Small
Institute for Scientific Information, 3501 Market Street, Philadelphia, PA 19104. E-mail: hsmall@isinet.com

Science mapping is discussed in the general context of information visualization. Attempts to construct maps of science using citation data are reviewed, focusing on the use of co-citation clusters. New work is reported on a dataset of about 36,000 documents using simplified methods for ordination, and nesting maps hierarchically. An overall map of the dataset shows the multidisciplinary breadth of the document sample, and submaps allow drilling down to the document level. An effort to visualize these data using advanced virtual reality software is described, and the creation of document pathways through the map is seen as a realization of Bush's (1945) associative trails.

Rationale for Mapping

A map of science is a spatial representation of how disciplines, fields, specialties, and individual papers or authors are related to one another as shown by their physical proximity and relative locations, analogous to the way geographic maps show the relationships of political or physical features on the Earth. Lin (1997) has provided a useful typology of the various styles of representation, including hierarchical, network, scatter, and map displays. Of course there is nothing inherently two-, three-, or even N-dimensional about how scientific topics or papers relate to one another. Rather, it is a structure we impose on a collection of objects. Nevertheless, we find arranging information in space a natural and useful heuristic tool, perhaps because spatial relations play such an important role in everyday experience.

Here we can only speculate on the relation of spatial cognition and information. Clearly the human mind has the ability to organize experience in spatial terms and recall objects associated with physical locations. We have all had the experience of finding an object by remembering where we placed it. Books are arranged on library shelves by topic and with related topics in close proximity. Such an arrangement facilitates browsing and finding related items, and also returning to a title we found before simply by its location. The association of a physical object with a location has a nonspatial analogue: we often recall a fact from memory by associating it with something else. It seems reasonable that creating spatial environments with information items distributed in a stable and meaningful fashion has the potential of enhancing information usability and retrieval.

In the case of scientific literature, a spatial representation can facilitate our understanding of conceptual relationships and developments. A map of science can provide insight into a contemporaneous state of knowledge, which Holton (1996) lists as the first requirement of good history of science. It may also aid us in using information for making discoveries as envisioned by Swanson (1987). If we can visually spot literatures not directly related to one another but perhaps indirectly related, such topics may be good candidates for combination and knowledge synthesis.

The computer display of large-scale maps of science can now make use of virtual reality software and hardware capable of smoothly navigating large synthetic spaces (Steinberg, 1997). One such system is Sandia National Laboratory's EigenVR. A large-scale mapping could be defined as one involving the positioning of a few thousand documents, but a comprehensive, multidisciplinary mapping of science might involve tens or even hundreds of thousands of items which constitute the research front of modern science. The dataset reported on in this article consists of a two-dimensional mapping of a multidisciplinary sample of 36,720 documents.

Approaches to Mapping Science

The objective of most information visualization applications is the depiction of local structure. Either a set is retrieved and set members become the objects to be visualized (Lin, 1997), or a navigation or snowball process creates a set from some starting point (Small, 1994). The objective of the work reported here, however, is to create a global structure and overview of as large a sample of data as possible and then enable the user to explore the underlying fine structure. Following the terminology of Sneath and Sokal (1973), we will call the procedure for fixing or positioning objects in space "ordination."

Because the work reported uses citation data to generate the ordination, we first review some of the early efforts to

use citation data for mapping purposes. Citation data have an analogue and intellectual cousin in the world of hypertext links. There are a growing number of efforts to map web sites using such links (Hendley, Drew, Wood, & Beale, 1995; Savoy, 1996), and even some application of citation analysis techniques to web analyses (Pitkow & Pirolli, 1997; Larson, 1996). We will not attempt to review the many mapping efforts that utilize linguistic data, such as co-word, co-term, or co-classification analysis (Callon, Law, & Rip, 1986), two important new examples of which are the work of Wise et al. (1995) and Chalmers (1996). Another important avenue of research is the hybrid use of citation and word data to create maps (Braam, Moed, & van Raan, 1991a, 1991b).

The Citation Network

The citation network is a directed graph of great size and complexity, whose vertices can be chronologically ordered, and whose edges connect earlier with later vertices. The network embodies the communication patterns of millions of scholars, both living and dead. These patterns show how researchers go about embedding their work, both cooperatively and competitively, in the work of prior authors.

We know that the overall citation graph is extremely sparse, with regions of high linkage density. Such regions are found to correlate with subject areas or specialties (Small & Griffith, 1974). Despite the generally sparse linking as much as 98% of documents in the graph can be connected into a single component using data from a subject area such as analytical chemistry (Small, 1995). This seemingly contradictory weak connectedness and high localized density are perhaps due to a combination of factors: A focusing or concentration of effort in specific scientific areas, and the tendency on the part of each researcher to seek out a unique research niche.

Historical Review

Citation Networks

One of the earliest attempts to pictorially represent scientific development was Garfield's historiograph (Garfield, 1979). This is a diagram of citation patterns depicting the linking of papers forward and backward in time to trace the lineage of ideas over several generations. In a landmark study (Garfield, Sher, & Torpie, 1964), an historical account of the discovery of the genetic code was correlated with a citation network. Recently these data were reanalyzed using sociometric network analysis (Hummon & Doreian, 1989).

Price's article on citation networks (1965) represented citations between articles as filled cells in a citing/cited matrix. Regularities such as the periodic appearance of review articles, and the distinction between the archival and research front literatures, were seen as dense vertical or horizontal patterns on the matrix. Price's approach has recently been extended by Baldi and Hargens (1995).

The first computer visualization of citation networks was by Yermish (1975). This early system allowed forward and backward navigation of citation links, and displayed historiographs on a screen using a one-dimensional multidimensional scaling to fix the positions of documents along the horizontal axis, and publication year along the vertical axis. The most recent development in the graphic display of citation nets is the work of Mackinlay, Rao, and Card (1995) called the butterfly.

Two-Dimensional Maps: Co-Citations and Clustering

In 1973 Small and Marshakova independently proposed using highly cited papers and their frequency of co-citation as the building blocks for a mapping of science (Small, 1973; Marshakova, 1973). In 1974 Small and Griffith extended this approach to a large ISI© citation data file (Small & Griffith 1974; Griffith, Small, Stonehill, & Dey, 1974). Maps were constructed for both the microstructures of individual specialties, and macrostructures of broad fields, showing several scientific specialties in a common configuration. The technique of multidimensional scaling was used to display structure.

Eventually full annual files of ISI data were used, and up to four nested levels of clustering were performed, each level using the clusters obtained in the previous level as objects to cluster again (Small, Sweeney, & Greenlee, 1985). After about four iterations it was possible to create global maps which showed relationships between disciplines in physical and biological sciences (Small & Garfield, 1985). The advantages of this approach to mapping were, first, that co-citation provided a coefficient of similarity between documents, and a metric that could differentiate distances between objects. Second, clustering provided a "chunking" of the citation network, so that the complexity of document citation patterns could be hidden within a hierarchy of objects. In other words, regions of high intercitation could be collapsed, and represented by a simpler network of super-nodes (Fig. 1).

Unlike the historiograph approach, co-citation maps use two dimensions to depict subject relationships. Change over time is analyzed by comparing maps from successive time periods. The time variable is usually taken as the year of the citing articles. The patterns of co-citation in that year define the collective perceptions of citing authors and give rise to clusters of highly cited and co-cited works. Shifts in highly cited articles are then used to study the rate of intellectual change. A sudden shift in the cited article set of a specialty can signal a revolution in the field. Rapidly growing fields such as AIDS can be tracked from their birth, as they spawn multiple lines of research, and eventually emerge as major fields in their own right (Small & Greenlee, 1990).

The co-citation methodology was also extended to authors, using the primary author rather than the document as the unit of analysis. Here the analysis focuses on individuals whose collective citation patterns can be mapped with multidimensional scaling (White & Griffith, 1981).

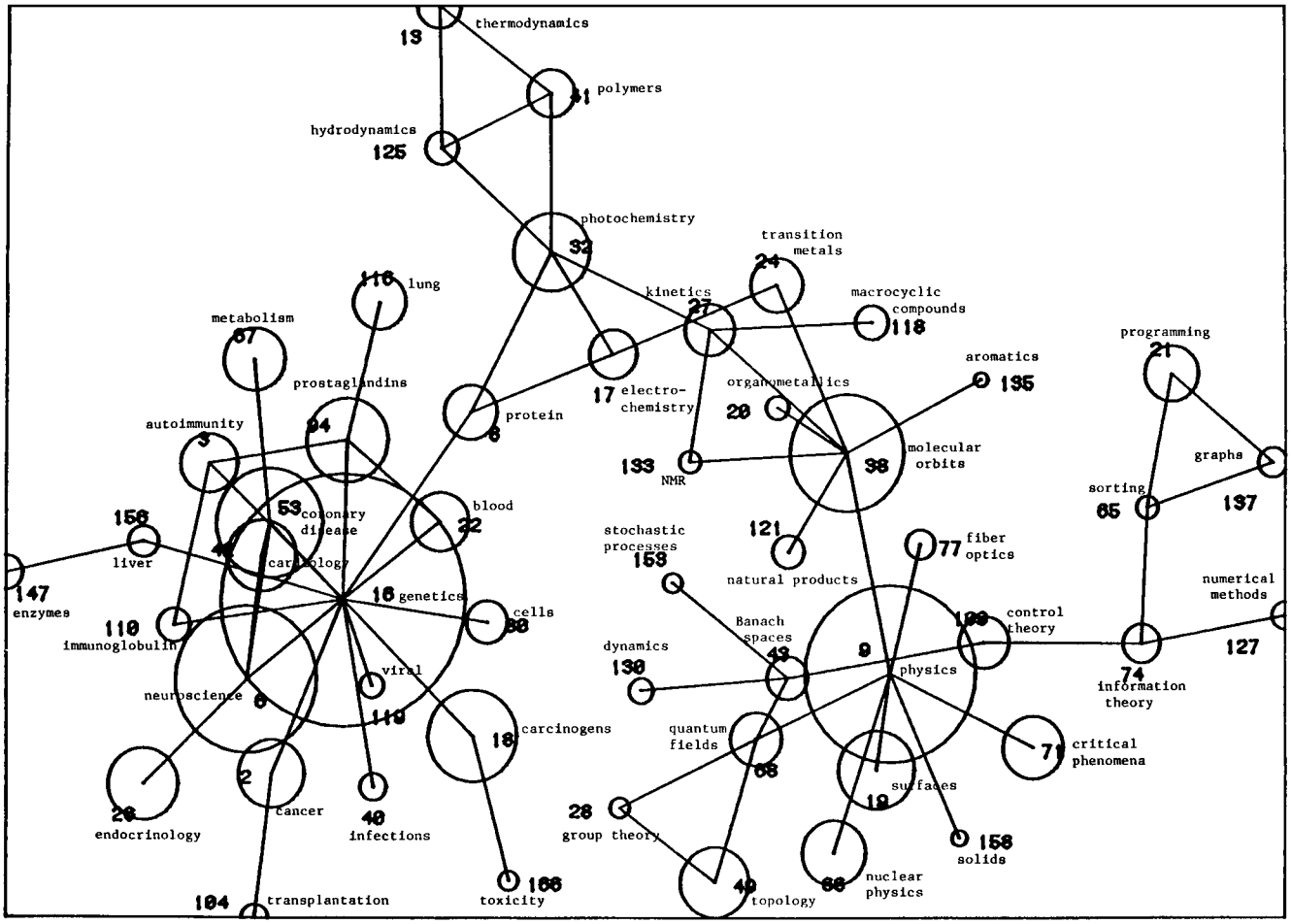

FIG. 1. An early co-citation map of science showing the major disciplines of the natural sciences: Biology, chemistry, and physics. The data are from a five-level co-citation analysis. Multidimensional scaling was used to position macro-clusters (from: Small & Garfield, 1985).

SCI-Map

In 1991 a PC-based science-mapping program called SCI-Map (Small & Rothman, 1994; Small, 1994) was introduced to allow the analyst to control and visualize the clustering process. Based on co-citation, the software allowed the user to select a seed document and create a cluster while viewing it on the screen. A number of different clustering options were available including single- and complete-linkage. Rather than attempt to build multidimensional scaling into the PC program, a method was developed for arranging documents in two dimensions by a geometric triangulation process (Fig. 2).

While SCI-Map allowed the exploration of the fine structure of an area, it was not well suited to mapping the macrostructure of science. SCI-Map shared with the earlier historiograph methodologies the stepwise building up of structures using individual documents. However, it proved impractical for users to build up large-scale maps of full fields or disciplines document by document. The earlier co-citation clustering approach, on the other hand, did provide a global view by iteratively clustering until macrostructures were formed. Therefore, what seemed called for was a combination of bottom-up and top-down approaches. One solution to providing both macro and micro views is described later in this article.

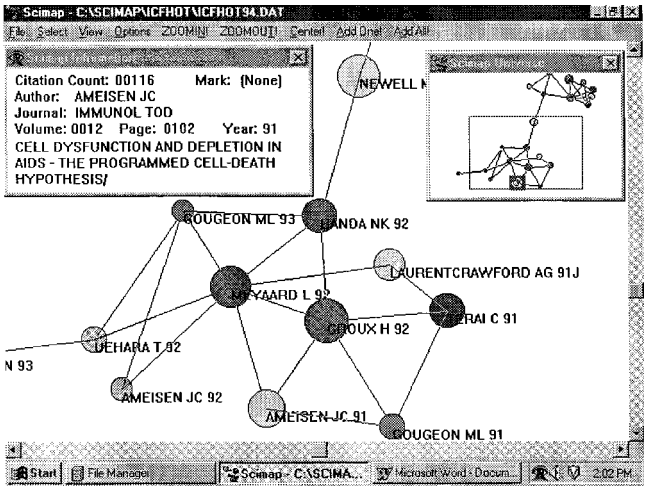

FIG. 2. A screen from SCI-Map showing a cluster on AIDS under construction. The universe window shows the network built up thus far, a portion of which is displayed on the full screen. An information window gives bibliographic information about the currently selected paper.

TABLE 1. Frequency of coupling forms in a citation network.

Indirect linkage	Couplings	Distinct links	Mean strength
CC	225,723 (50.6%)	201,737 (54.5%)	1.12
BC	120,199 (26.9%)	90,531 (24.5%)	1.33
LC	100,113 (22.5%)	77,583 (21.0%)	1.29
Total	446,035 (100%)	369,851 (100%)	1.20
Combined links	446,035	341,196	1.31

Measures of Document Similarity

The rationale for using indirect linkages, such as co-citation or bibliographic coupling (Kessler, 1963), is that they reinforce regions of dense direct citation and thereby facilitate the breaking up of the network into meaningful chunks. These procedures can be considered methods of citation amplification.

Considering the publication years of articles, there are three ways two articles can be connected by taking two steps on a citation network: 1) Bibliographic coupling (BC), which connects papers by one step back then one step forward; 2) co-citation (CC), which takes one step forward then one step back; and 3) a third form which connects older and younger papers by taking two steps in the same direction, either forward or backward. This third form has been called longitudinal coupling (LC), because it is capable of connecting articles across multiple years (Garofano, 1965; Small, 1995).

In order to gauge the prevalence of these three forms a 13-year dataset of articles in analytical chemistry was used consisting of 8,402 articles among which were 56,774 citation links. Each link connected one of the 8,402 articles with another item within the set. Hence, the links are a subset of the total citations (or references) made by the articles since links outside the set are not included. Table 1 summarizes the number of two-step citation paths in the dataset, expressed in terms of the total frequency of indirect linking (couplings) and the number of distinct coupled pairs of each kind. The reason for the numerical dominance of co-citations is not clear, but similar results have been obtained in other datasets. The last row, labeled "combined links," shows the results of collapsing the three forms of coupling into distinct links, which results in the same total couplings, but fewer distinct links than the sum of the three forms. The redundancy of the different forms is, however, only about 8%, and each form contributes significantly.

Clearly the role of the different coupling measures will vary with the length of the citing and cited time periods used. For example, CC and BC offer cross-sectional views given relatively narrow, one-year citing periods. LC becomes effective only when longer periods are used: At the end of a time period documents will be linked through their references to earlier items; at the beginning, through citations received; and at mid-period, through both references and citations.

Early on Amsler (1972) proposed combining co-citation and bibliographic coupling into a single measure. Recently, Small (1997) combined all indirect coupling modes and direct citations to form a composite measure. Whatever measures are used, singly or in combination, it seems appropriate to cast each as a coefficient that varies between zero and one. Such a normalization should take into account the total number of links, whether references or citations, incident on each node of the connected pair.

The choice of what coupling measure to use, of course, depends on the goals of the analysis. For a mapping of current papers the analyst might elect to use BC only. If the goal is to map older key papers from a current perspective, the best choice might be co-citation. Finally, if a mix of current and older papers is desired, then a combination of measures can be used. In the mapping exercise described below we use a one-year citing period for co-citations, namely, 1995, and a 15-year cited period, 1981–1995.

Methodology

Distance Transformations

For mapping or visualization, coefficients of similarity need to be converted into distances such that closely related objects are short distances apart and weakly related objects are further apart. In classical nonmetric multidimensional scaling or principal-components analysis, this transformation is achieved through a mathematical minimization method. In vector-space approaches, distances are a by-product of placing the objects in an N-dimensional coordinate space.

An alternative is what might be termed direct distance conversion, namely, a numerical transformation of a coefficient of association into a dissimilarity coefficient interpreted as a distance (Small, 1997). The simplest of these transformations subtracts the similarity coefficient from one to get a distance of zero for objects with a similarity of one, and a distance of one for zero similarity. Since the clustering procedure we use involves setting a threshold on the similarity, a further normalization is performed on the computed distance, dividing it by the maximum possible distance for the given threshold. Since the clustering threshold can vary from cluster to cluster, distance normalization has the effect of making the longest distance between linked objects within a cluster equal to one. This measure has the character of a basic unit of distance on our maps, and we call this unit the "Garfield."

Ordination by Triangulation

The classical methods of ordination mentioned above involve the solution of minimization problems to yield the final configuration of objects in coordinate space. For example, multidimensional scaling involves the calculation of a minimum value for a goodness-of-fit measure called stress. Hence these methods are computationally intensive.

An alternative to these methods is simple triangulation (Lee, Slagle, & Blum, 1977), used in the SCI-Map described above. Triangulation is a purely geometric proce-

dure yielding configurations that exactly represent small numbers of distances between objects, but lacks global optimization. The motivation for using triangulation is to see if a simpler and faster method is adequate for the visualization task at hand, thereby providing a computationally less demanding solution.

For the two-dimensional case, triangulation begins arbitrarily with one of the objects and places it at the origin of the coordinate system. Then the object closest to it is found and placed at the specified distance from the first object, oriented in an arbitrary manner. The location of the third object is fixed using distances from the first two objects, triangulating on them. This leaves a degree of freedom to pick one of the two possible quadratic solutions, above or below the line formed by the first two objects. From this point on we use the notion of repulsion from the center of gravity to select the quadratic solution furthest from the center. This causes the configuration to grow out from its center.

The disadvantage of triangulation is that it is dependent on the order in which objects are assigned positions. However, the speed with which the calculations can be made means that every object in the cluster can be tested as the seed, and the "best" solution selected. The definition of the "best" here is the solution having the highest sum of linkage values used in the ordination process. It should also be possible to select the solutions with the lowest "stress" value as in multidimensional scaling, although a minimum of this function would not be reached.

In triangulation the focus is on exact fits to the strongest links and shortest distances. The method generally does not attempt to fit weak or very long distances. Because there is a limit of one Garfield to the longest fitted distance, the triangulation process is similar to piecing together a set of objects using a collection of short sticks of variable length, reminiscent of Tinkertoy™ construction.

Sampling Papers With Fractional Citation Counts

Since our objective is to represent the structure of science using highly cited papers as markers for individual topics, we need to obtain a sample of papers which is as multidisciplinary as possible. Thus the sampling of highly cited papers needs to take into account variations of citation and reference intensity by field. Simply using an integer threshold for citations will skew the sampling to high referencing fields, such as biomedicine. A solution to this is to use fractional citation counting (Small, Sweeney, & Greenlee, 1985). A fractional citation count is one in which each citation is inversely weighted by the length of the reference list of the citing article.

We begin by using a low integer citation-count threshold below which we would not want to select a paper. This base citation rate was set at five citations per paper published in the 15-year period 1981 through 1995 using a single citing year, 1995. There were a total of 5,382,404 citations to 524,165 distinct items cited five or more times, which is 53% of all citations from 1995 to items in the 15-year period. Citing these selected documents were 527,114 distinct items from 1995. For each citing item the length of the restricted reference list was computed, that is, ignoring references to items cited less than five times. These counts were attached to each citing item. The fractional citation count was then computed by summing the reciprocals of these reference counts for a given cited item.

The final sample of cited documents was made by applying a threshold of 1.0 on the fractional citation scores. This yielded 164,612 cited documents, about 31% of all papers cited five or more times in the 15-year period.

TABLE 2. Number of clusters at five levels.

Level	No. clusters	No. objects	Mean size	Threshold
1	18,939	129,581	6.8	0.10
2	2,402	10,883	4.5	0.05
3	327	1,200	3.6	0.01
4	35	148	4.2	0.00
5	1	35	35.0	0.00

The Humpty-Dumpty Method

The procedure for visualizing multilevel hierarchies has been called the Humpty-Dumpty method because of its analogy to breaking a structure apart and then putting it back together: The database is first broken up into smaller pieces by clustering, and then the pieces are reassembled into an overall structure (Small, 1997).

There are three steps in the process: 1) Creation of a multilevel hierarchy of clusters or partitions starting with individual documents, 2) an ordination of objects within each cluster in the hierarchy, and 3) the integration of the local structures into a global structure or common coordinate space. This last step assembles the separate pieces into an overall structure and involves expansion and translation of clusters at all levels. It is important to note that this framework is independent of the type of data used for similarity calculations (word or citation based, etc.) as well as the specific clustering and ordination methods used.

Input to the process was the file created by fractional citation counting using co-citation as the linkage method. Clustering was carried out to create a hierarchy five levels deep. The clustering method was single-linkage with a maximum cluster size to limit chaining. The number of clusters consisting of two or more objects and the total number of objects contained in these clusters is given in Table 2 for each of the five levels. A maximum cluster size of 50 was used, and if the cluster exceeded this size, the normalized linkage threshold was incremented and the clustering repeated with the same seed document until the resulting cluster was within the size limit. Table 2 shows the number of clusters obtained at each level and the objects (documents at the first level) they contain. About 79% of the documents selected by the fractional citation threshold are contained in clusters of size two or greater. Only one cluster was obtained at the fifth level. This cluster contained 36,720 documents in what we call the main hierarchy. The remain-

TABLE 3. Clusters in the main hierarchy

Level	No. clusters	No. objects	Mean size
1	4,341	36,720	8.46
2	699	4,341	6.21
3	148	699	4.72
4	35	148	4.23
5	1	35	35.00

ing 71% of documents fall into clusters at lower levels that become separated from the main hierarchy at some point. For visualization we focus our attention on the 36,720 documents, but the same methods used to visualize the main hierarchy could be used on the separate subhierarchies. Table 3 shows the numbers of clusters and objects contained in the main hierarchy.

The second step is ordination, and in this experiment we have used the SCI-Map triangulation algorithm (Small & Rothman, 1994), which gives a two-dimensional configuration for each cluster centered at the origin.

The third step combines the information from the first two, expands, or scales up each of the individual cluster coordinate systems, working from the document level to the most aggregated level, until the root (in our case the fifth level) of the hierarchy is reached. Then working back through the levels, each coordinate system is translated so that its centroid is moved to the location of the parent object that contains it. The original application of this method also performed a rotation operation on configurations which was not used in this experiment (Small, 1997).

The critical step is the initial expansion because without it objects from each level would pile up on top of each other. The purpose of the expansion is to make room for underlying objects of varying size so that object overlap is minimized. The size of a cluster is determined by finding the circle whose origin is at the centroid and whose radius is just large enough to enclose the objects it contains. An expansion factor for the coordinate system is then determined satisfying the condition that adjacent circles along a minimal spanning tree path through the cluster will be exactly tangent, that is, not overlap.

Since this method will not completely eliminate object overlap, an overlap avoidance routine is also used. This procedure consists of traversing the objects in the cluster in a set sequence, and moving the higher member of each overlapping pair away from the centroid until all overlaps have been eliminated. Using overlap avoidance without expansion creates a close packing of objects. On the other hand, too much expansion creates too much space between documents making navigation more difficult. What seems best is a balance between the two extremes.

In the visualization, circles are used to represent documents as well as higher-level objects. As in SCI-Map the radius of a document is scaled to the cube root of its citation frequency which models the number of citations as the volume of a sphere. Thus, at the first level of aggregation where clusters consist of documents, the expansion is based on document circles whose radii are determined by their citation frequency. At the second and higher levels the size of an object is dependent on several factors including the number and sizes of the objects contained, their degree of expansion, and the shapes of their configurations, that is, whether they form tightly knit clumps or more loosely linked structures.

Table 4 gives the mean, minimum, and maximum radii of objects for each level of the hierarchy. Another way to characterize the final coordinate space is to find the dimensions of a rectangle that would enclose all objects. For the 36,720 documents this rectangle is found to be 144,681 by 148,876 Garfields.

Results

Description of the Overall Map

We begin with the overall map of the main hierarchy, consisting of 35 level-four clusters (Fig. 3). Each level-four object is represented as a circle whose size depends on the number and spread of lower-level objects it contains. The map shows four main topic regions: A large physical-science region on the left, a biology region to its right, medicine to the right of biology, and finally a behavioral/social science region below medicine. This fairly linear progression from physics and chemistry to social sciences is similar to patterns obtained in earlier co-citation mapping exercises (Small & Garfield, 1985), although it is unusual to see social science and physics on the same map. The convergence of these disciplines means that it is possible in principle to traverse the co-citation network of 36,720 documents starting in, say, sociology and eventually reaching the field of astrophysics. The map gives us a guide to navigating these links.

The large physics and chemistry circle on the left in Figure 3 is surrounded by a number of smaller, more specialized topics in chemistry and engineering, with geoscience and environmental science at the top. The environmental area consists of subclusters on climate and soil science, and is linked to the geoscience circle, which contains topics such as geologic evolution and earthquakes. An interesting outside or shortcut path links environmental science to aquatic science, which in turn links to plant/animal genetics on the right-hand, or biological, side of the map. Thus, aquatic science can be seen as a linking field.

The large physical-science circle contains a number of subclusters covering pure physics topics such as astrophysics, high-energy physics, superconductivity, and quantum

TABLE 4. Sizes of objects in the hierarchy (radii in Garfields)

Level	Mean radius	Min radius	Max radius
1	0.84	0.56	2.95
2	7.16	1.13	224.64
3	67.30	2.56	1,601.18
4	483.60	12.14	3,407.05
5	2,604.40	54.23	28,633.15

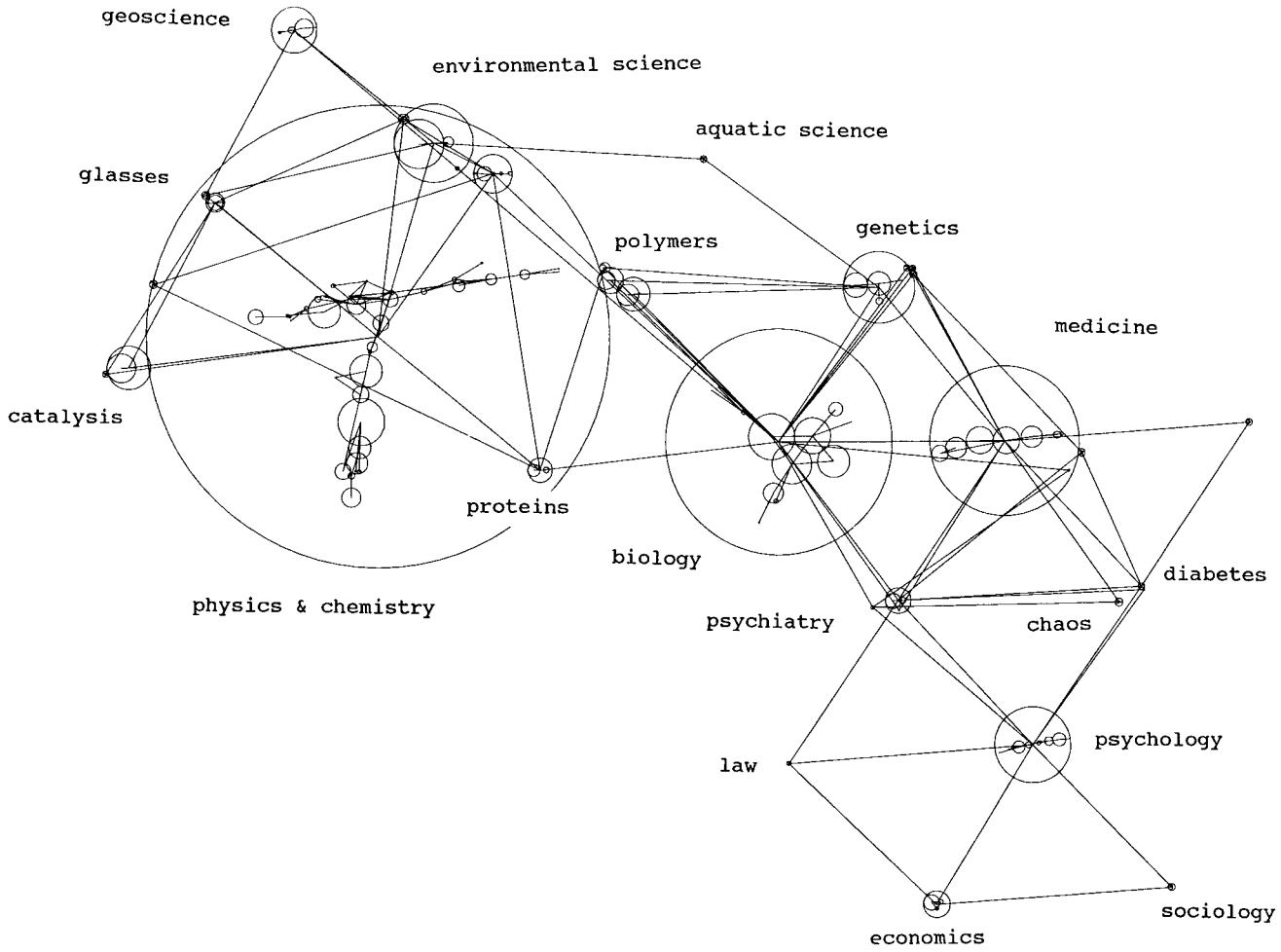

FIG. 3. The map is a two-dimensional computer visualization of 36,720 documents comprising a multidisciplinary sample selected by a fractional citation counting method, covering the years 1981–1995. The circles represent 35 fourth-level clusters, whose internal structure can be explored by zooming into the object. The configuration of third-level clusters is also visible within each fourth-level object. The sizes of the circles reflect the spread of lower level objects contained. Lines represent strong co-citation links among the clusters. Subject matter ranges from social science on the lower right to physical science on the left. Labels are based on a frequency analysis of article titles and journal category names.

physics. Linked to and branching off from these pure physics areas are topics in applied physics, materials science, and chemistry. These include laser fusion, nonlinear optics, surface science, transition metal complexes, and silicon polymers. The smaller chemistry areas which dot the perimeter of the physical-science circle cover topics such as electrochemistry, pyrolysis, polymers, catalysis, glasses, thin films, solid phase synthesis, and protein structure.

The biology region to the right of physics and chemistry includes large clusters on biology, and genetics. Within the biology circle are subclusters on immunology (including HIV), cancer genetics, cell adhesion, growth factors, and neuroscience.

The smaller plant and animal genetics circle above (Fig. 3) biology has subareas related to plant science with an emphasis on plant genetics and the study of plant DNA. This latter area is linked to a cluster on animal DNA evolution, human genetics, and heredity.

Medicine to the right of biology is closely intertwined with biology via numerous smaller interconnecting areas. Within the medicine circle are topics related to cardiology and the central nervous system. The unifying thread seems to be the various channels, receptors, and neurotransmitters that control these processes. Smaller medical clusters scattered around the biology region include radiology, transplantation, toxicology, human reproduction, epidemiology, ulcers, resistant bacteria, stroke, and sleep apnea.

Social and behavioral sciences occupy the lower right-hand region. Some of the main topics are economics, sociology, marketing, law, mathematical methods, social services, drug use, family violence, developmental psychology, racial identity, and feminism. There is also a close interlinking between medicine and behavioral areas through topics such as event-related potentials and psychological aspects of medical care.

Like aquatic science (noted above), a number of what might be called linking or interdisciplinary clusters emerge from this analysis because of their mix of topics or their unusual positions. The cluster on psychiatry, including adjustment disorder and depression, is a hybrid between psychology and clinical medicine and is in an intermediate

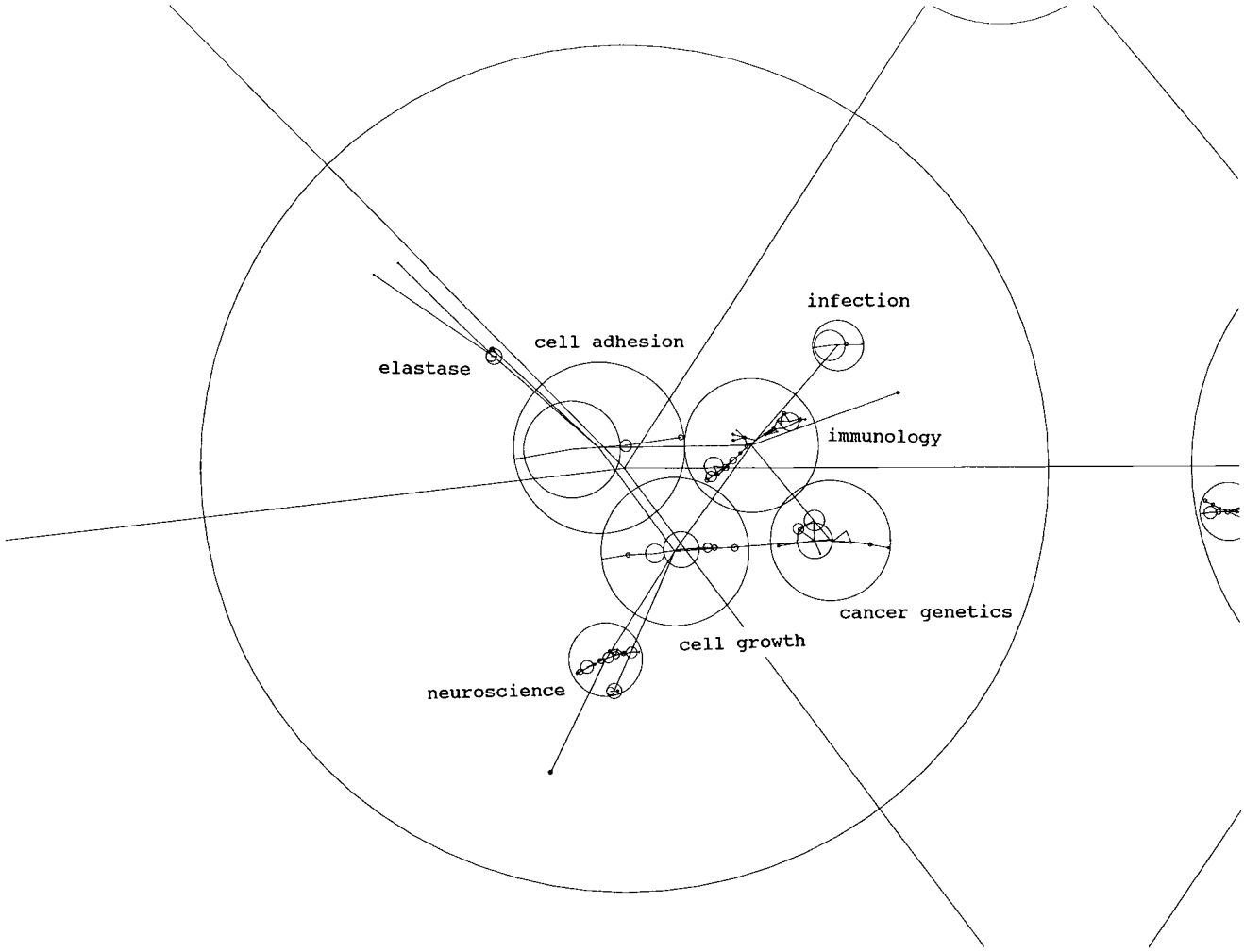

FIG. 4. A detailed view of the large biology circle at the center of Figure 3, showing the level-three clusters within it as well as the level-two objects within each level three cluster. Topics range from immunology to neuroscience.

position between the two. Similarly the cluster on diabetes situated between psychology and medicine deals with not only the treatment of diabetes, but also obesity and eating disorders, which have an important psychological component.

The cluster in this same vicinity labeled chaos contains subareas on applied physics and mathematics, and appears at first glance to be on the wrong side of the map. On closer inspection, however, the main focus of research is the application of chaos theory to biological and medical systems, such as microbial virulence, chaos in food chains, and population dynamics. Thus, in this case the links to the area of application are stronger than to the disciplinary source. A cluster just to the right of the physics and chemistry circle in Figure 3 labeled proteins involves an interdisciplinary mix of chemistry, physics, and biology. A closer examination of the subtopics reveals an emphasis on molecular recognition and molecular self-assembly, topics of interest to a number of fields, including pharmacology.

From this brief discussion of crossover fields we see that it is not uncommon for disciplinary boundaries to be breached for specific objects of investigation. The structure we see here seems to be partially determined by the requirements and opportunities for gaining new knowledge and partially by traditional disciplinary boundaries.

Volvox Display and Document Drill Down

The style of visualization used has been termed a volvox (McCain, 1996) because smaller, lower-level objects are represented as circles within larger, higher-level objects, which resembles a microorganism by that name. In Lin's typology (1997) it would be a combination of hierarchical and network displays. The advantage of this mode of representation is that it allows simultaneous viewing of relative location and hierarchical structure. Links between objects can also be drawn to clarify relationships that cannot be adequately captured by ordination. The volvox format can be easily extended to three dimensions by substituting spheres for circles.

The exploration of a subject area takes place by drilling down, showing each configuration as we progressively focus in, until we reach the document level. To illustrate this, we have identified a cluster that has a large number of recent

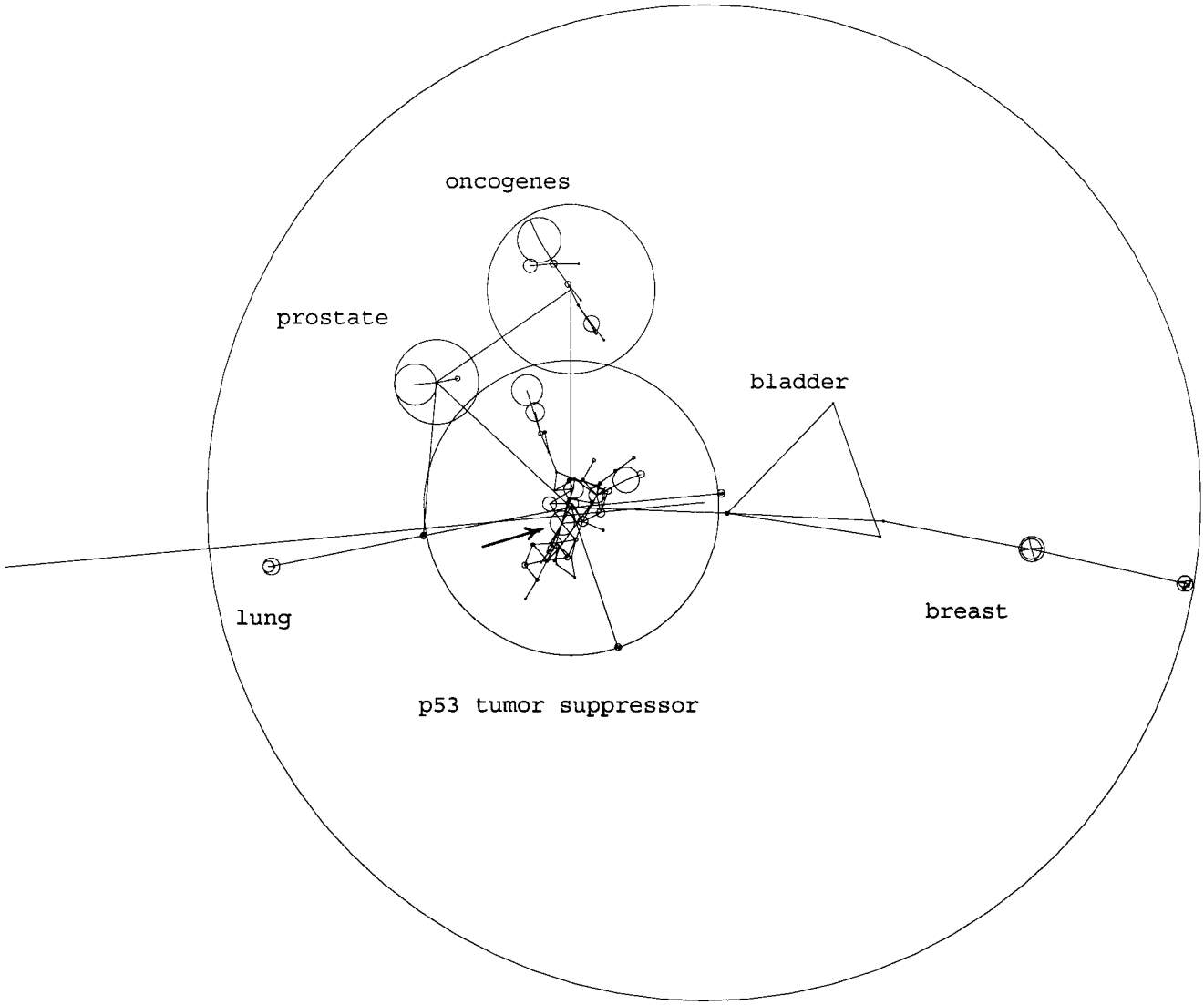

FIG. 5. A detailed view of the level-three cancer genetics cluster on Figure 4, showing the level-two objects contained as well as the level-one objects within them. Clusters dealing with various forms of cancer are arranged around the central cluster on the p53 tumor suppressor protein.

papers and thus has a high mean year of publication. Such objects are termed "hot fields." Across the document set in the main hierarchy, which draws on papers from 1981 through 1995, the mean year of publication is 1989.5. Our target cluster has a mean year of 1993.6, and is ranked within the top 2% of clusters by mean publication year. The topic of the cluster is p21, a protein implicated in the inhibition of cell division and more generally in the aging process (Cherfas, 1995, 1996).

We first drill down into the large molecular and cell biology circle, shown on the overall map, and view the objects contained (Fig. 4). We see subclusters on immunology, infection, cancer genetics, cell adhesion, cell growth factors, and neuroscience.

Zooming into the cancer genetics circle we find a few large groups and several smaller ones (Fig. 5). The large central circle is concerned with expression of the p53 tumor suppressor protein and apoptosis (programmed cell death) and its connection to cancer. Other smaller clusters are concerned with oncogenes and specific forms of cancer; namely, breast, prostate, and lung, and their genetic linkages. The hot field we are targeting is indicated by an arrow near the center of the p53 cluster.

Drilling into this region we obtain the view of Figure 6 showing the p21 cluster in the center containing 39 documents, surrounded by a number of smaller related areas, mainly on p53. The documents published in the most recent two-year period, 1994–1995, are unshaded and those prior to 1994 are shaded, clearly showing a tendency for items to segregate by age.

Representativeness of the Map

To broadly gauge the topic coverage of the map the ISI journal category for each article on the map was determined. Assignment to a category was made on the basis of the journal in which the article was published. Counts were made for the articles in the main hierarchy of clustering,

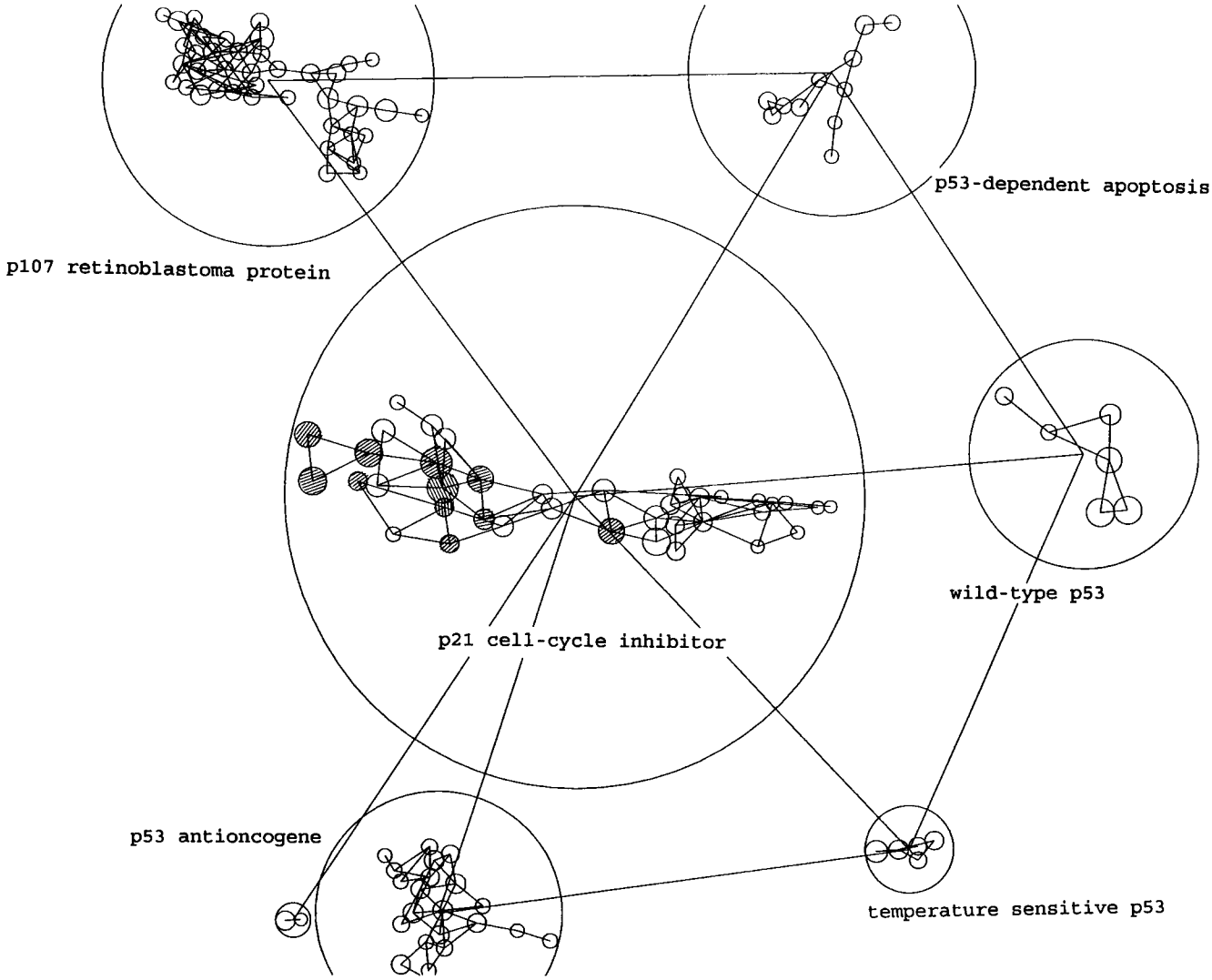

FIG. 6. A detailed view of a group of first-level clusters whose location on Figure 5 is indicated by an arrow within the p53 tumor-suppressor region. The largest level-one cluster is on the p21 protein containing 39 documents. Documents more than two years older than the base year, 1995, are shaded. Unshaded documents are from the two most recent years, 1994 and 1995, making this a "hot" cluster.

which the map in Figure 3 represents (36,720 items), and also for the full set of clustered items shown in Table 2 (129,581 items). The journal categories used were for the ISI product *Current Contents*,® and these categories were further aggregated into the 23 broad categories used in ISI's National Science Indicators (Institute for Scientific Information, 1997). For comparison purposes, baseline counts drawn from this indicator database were used for the period 1981 through 1995, which was the 15-year sampling frame for fractional citation counting.

Table 5 gives the percentage of articles for each of the 23 categories. From this we can gauge which topics are over- or under-represented on the map. For chemistry the sampling seems representative for both the main hierarchy and the full clustering. In physics there is an over-representation in the main hierarchy and a somewhat better representation in the full clustering. In clinical medicine there is an under-representation in the main hierarchy but a fairly accurate one in the full clustering. Social science is slightly over-represented in the main hierarchy. In general, biomedical fields seem somewhat under-represented and the physical-science areas over-represented, with representativeness improving for the full clustering regardless of field.

Since the length of the reference list is used in the fractional citation count procedure and biomedical papers in general have longer reference lists than physical-science papers, it appears that the fractional method has somewhat overcompensated for the expected biomedical bias. One way to correct for this is to increase the initial integer citation threshold and lower the fractional cutoff.

Linkages Between Clusters

Another way to approach the maps is to focus attention on the nature of the connections between areas, and attempt to understand why topics are linked together. Here we must reverse the process of map buildup and progressively unravel the interdocument connections.

TABLE 5. Percentage of articles by journal category.

Category	Main hierarchy	Complete clustering	15-year database
Agricultural sciences	0.58	1.27	2.59
Astrophysics	3.09	1.35	1.04
Biology and biochemistry	4.95	6.74	8.04
Chemistry	11.29	11.20	11.54
Clinical medicine	16.90	22.90	21.11
Computer science	0.34	0.88	1.06
Economics and business	2.99	1.49	1.37
Education	0.28	0.17	0.46
Engineering	3.29	5.22	7.28
Ecology and environment	1.22	1.90	1.93
Geosciences	4.04	2.42	2.33
Immunology	1.05	1.26	1.50
Law	0.58	0.28	0.32
Molecular biology and genetics	3.98	3.01	2.36
Microbiology	1.21	2.09	2.24
Materials science	2.92	2.44	2.88
Mathematics	0.45	1.19	1.60
Neuroscience	4.10	4.03	3.34
Physics	19.86	13.89	10.30
Plant and animal science	2.09	3.61	6.60
Pharmacology	0.81	1.59	2.45
Psychology/psychiatry	4.59	3.33	2.69
Social sciences, general	2.07	1.79	3.30

Links between clusters arise because not all co-cited documents are subsumed within a single cluster. The patterns of strong residual co-citation linkage in fact determine the macrostructures of the maps. At each level of clustering the co-citations are progressively collapsed to link larger and larger structural units, so that, for example, by the fourth level a single link might represent many hundreds of co-citation relationships connecting documents in separate macroclusters. By taking one such macrolink we can progressively break down the connection, first to a set of relationships between third-level clusters, then to relationships between second-level clusters, and so on until we arrive at the co-cited documents themselves. If at each stage of this decomposition the most strongly linked object pair is selected, we end up with the strongest interdocument linkage connecting the two macroclusters.

As an example, we start with the link between the level-four physics and chemistry (Fig. 3) and environmental science clusters on the left side of the overall map. Looking down one level we find the strongest link to be between level-three clusters on auroral plasma on the physics side and atmospheric carbon dioxide on the environmental side. At level two the strongest linked clusters are high-latitude ionosphere studies on the physics side and middle-atmosphere studies on the environmental side. Finally, at the document level we have a paper proposing a global model of thermosphere winds on the physics side co-cited with a paper on observations of sporadic sodium layer events in the atmosphere on the environmental side. In a general sense this means that atmospheric science provides the nexus between physics and environmental science.

Similarly an analysis of the link between the large biology and medicine clusters reveals that the most significant path connecting them involves the study of various receptors in the central nervous system, generally situated in the field of neuroscience. Between economics and psychology areas on the lower-right side of the map, the linking fields are management and organizational psychology, specifically the study of ownership and control of a business and collective rationality from the psychology side.

The absence of a drawn link on the map does not mean that no connection exists, only the absence of a strong one. As a case in point, the path between the large physics and chemistry and biology circles was found to be computer imaging research, that is, statistical image restoration methods on the biology side and dynamics of Monte-Carlo simulations on the physics side.

Discovering New Pathways Through Science

This glimpse into the nature of linking literatures is a prelude to a more comprehensive tour of the many pathways that connect the disciplines of science and which unify the physical, biological, and social sciences into a single structure. For example, we might want to find the shortest document path that touches all the level-four clusters, a kind of complete tour package. Or perhaps we would like to find a path that connects two areas not directly connected. The question in this latter case is whether such an indirect path might represent a chain of inference that could tie together the seemingly unrelated domains, in Swanson's sense (1987).

The steps involved in laying out a path are as follows: First we decide what level of objects will make up the path, that is, whether they are documents, level-one objects (clusters of documents), or higher-level objects. For our example we use level-one objects since they offer a slightly broader view than individual documents. Next we select a starting point and destination point. As the starting point, we select the biological-chaos cluster on Figure 3, and as the destination, the proteins cluster. The proteins area contains a subregion on molecular self-assembly. These innovative areas have no direct link with one another but finding a bridge between them may suggest interesting avenues of cross-fertilization.

Inspecting the map of Figure 3 helps us plan a route either following the drawn links or the proximity of objects. From the map we see that proteins has a link to biology but no direct link is shown from biology to chaos. Nevertheless, analysis of the linkage data reveals a weak link between them. As was done in the previous section, we find the strongest level-three object link for each level-four transition. In other words, as a boundary is crossed, we determine the most strongly linked objects at the next lower level. This is done both for the chaos to biology and proteins to biology connections. We find that chaos links into the neuroscience region within biology, and proteins links into the immunology region within biology at level three. Now we need to

TABLE 6. Path from chaos to molecular self-assembly.

	Level 1	Level 3	Level 4
1	Strange attractors; correlation dimension; dimension; attractor dimension; small data sets	Chaos	Physics
2	Make sense; brains make chaos; spatially chaotic dynamics; model; biological pattern-recognition	Chaos	Physics
3	Cat visual-cortex; oscillatory neuronal responses; visual-cortex; Cat; coherent oscillations	Chaos	Physics
4	Perception; monkey visual-cortex; functional logic; form color movement; depth anatomy physiology	Neuroscience	Biology
5	Macaque monkey; middle temporal visual area; middle temporal visual area (MT); relation; neurons	Neuroscience	Biology
6	Memory; pet; memory processing; retrieval; verbal episodic memory	Neuroscience	Biology
7	Schizophrenia; negative symptoms; schizophrenia review; schizophrenia differences; schizophrenia subtypes positive	Neuroscience	Biology
8	Risperidone; schizophrenia; treatment; new antipsychotic; ritanserin (R-55667)	Neuroscience	Biology
9	Central D2-dopamine receptor occupancy; D2-dopamine receptor occupancy; D(2) dopamine receptor occupancy; positron emission tomography	Neuroscience	Biology
10	Expression; D1; gene; cloning; dopamine-D2 receptor messenger-RNA	Neuroscience	Biology
11	Amphetamine; C-FOS; C-FOS gene; rat striatal neurons; neuroleptics increase C-FOS expression	Neuroscience	Biology
12	Expression; C-FOS; C-FOS expression; C-FOS protein; rat spinal-cord	Neuroscience	Biology
13	C-JUN; JUN; AP-1; FOS; different JUN	Neuroscience	Biology
14	Phosphorylation; map kinases; serine-73; serine-63; C-JUN	Immunology	Biology
15	RAF; activation; mitogen-activated protein-kinase kinase; RAS; map kinase	Immunology	Biology
16	SH2; SHC; GRB2; RAS signaling; RAS	Immunology	Biology
17	SH3 domains; SH3 domain; binding; identification; proline-rich peptides	Immunology	Biology
18	Larger proteins; N-15-labeled proteins; 3-dimensional heteronuclear NMR; 3-dimensional triple-resonance NMR; multidimensional heteronuclear NMR	Protein structure	Chemistry
19	Program; protein structures; macromolecular crystallography; molecular replacement; errors	Protein structure	Chemistry
20	Potentials; mean force; globular-proteins; calculation; protein models	Protein structure	Chemistry
21	Accessible surface-areas; proteins; proteins analytical equations; solvent accessible surface-area; solvation energy	Protein structure	Chemistry
22	Aqueous-solution; free-energies; electrostatic interactions; calculation; electrostatic potential	Protein structure	Chemistry
23	Hydrogen-bonded base-pairs; relative binding-affinity; nucleotide bases relative association constants; triply hydrogen-bonded complexes guanine-cytos	Organic chemistry	Chemistry
24	Molecular tweezers; convergent functional-groups molecular recognition; neutral substrates; microenvironment; complexation	Organic chemistry	Chemistry
25	Molecular recognition; diastereoselective molecular recognition; molecular recognition investigations; aqueous-media; chemistry	Organic chemistry	Chemistry
26	Synthesis; cyclodextrins; molecular self-assembly; molecular threads; self-assembly	Organic chemistry	Chemistry

find a path within the biology region that connects neuroscience and immunology.

In general, after drilling down from two sides into the same intermediate object, we then need to traverse a path within that object. We can use a map for this purpose. For example, the link from chaos at the third level to neuroscience turned out to be studies of brain activity using positron-emission tomography (PET) at level two. From the other side, the link from immunology to neuroscience was via the expression and cloning of neuronal receptors at level two. However, between the PET and neuronal receptors clusters were two additional intervening areas, one on schizophrenia and the other on the treatment of mental illness with clozapine. Hence, path formation involves alternatively navigating within and between objects.

The final pathway is shown in Table 6 as a sequence of 26 level-one clusters. This includes nine level-two clusters, five from level three, and three from level four. The table gives the "name" of the area in terms of the most frequently occurring word phrases in the titles of the articles making up the cluster. Shorter designations are given for higher-level groupings to show the movement from chaos, to biology, and finally to the chemistry areas of protein structure and molecular self-assembly.

To give a capsule summary of the journey, chaos clearly ties to brain activity in mental illness, which in turn ties to

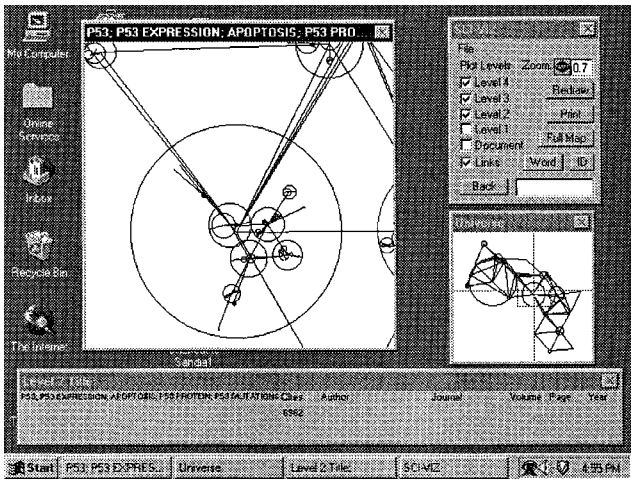

FIG. 7. A screen shot of the SCI-Viz PC interface for navigating nested maps. The Universe window at the lower right shows the overall map with a box and crosshairs indicating the current view in the main window. Information about the circle currently beneath the mouse pointer is given in the lower window.

brain chemistry. Brain chemistry links to signaling in the immune system, and the study of large proteins, which leads to the mechanisms for assembly of large molecules. It would be presumptuous to say because we have found this path through the literature, that somehow the mathematics of chaos has applicability to molecular self-assembly, or that the combination of chaos and molecular self-assembly would shed light on immune processes and the nervous system. The most we can say is that a literature trail exists circa 1995, and although it is not the only path that could be found, it is probably a significant one, due to the selection of strong linkages. A more comprehensive exploration of multidisciplinary pathways through scientific literature may in fact provide an empirical foundation for the unity of science (Holton, Chang, & Jurkowitz, 1996).

While maps at various levels are useful for visually navigating from topic to topic, this function can be automated by a shortest-path analysis. Such an algorithm, combined with the procedure for finding the strongest link when crossing cluster boundaries, provides what might be called a knowledge-connector system. This system automatically generates strongly linked document pathways for user specified starting and destination topics. For example, a pathway from economics to astrophysics was found consisting of 345 documents. The function of the map is then to display the path to the user and provide context for each step along the way.

User Interface Developments

The volvox style of mapping presents a number of challenges for computer visualization. The graphics primitives consist only of circles, lines, and text. However, there must be an efficient mechanism to zoom down into the structure to see the detail contained within objects nested several levels deep. The first system developed to display these data was a PC prototype implemented in Visual Basic called SCI-Viz, shown in Figure 7.

The user starts with an overview of the database similar to the overall map of Figure 3 showing level-four circles. The disciplines and topics represented by the various circles are discerned by reading labels activated by moving the mouse pointer over an object. The user can then elect to zoom into a given region. This is done by clicking in the area to be magnified using the left mouse button (zooming out with the right). By specifying which level objects to draw (one or more of five levels in the current dataset), the user can probe into the structure. Drawing speed is increased by storing the data on disk consistent with its geometric hierarchy, that is, having higher-level objects point to lower-level objects contained within them. Zooming in on one of the level-four circles the user can quickly draw the level-three objects within it, and so on until the document level is reached. The user can reposition the view by zooming out and then zooming in again at a different location.

The problem with this PC implementation is that the navigation is discontinuous, taking place in jumps by a sequence of mouse clicks. Clearly what is needed is the ability to continuously and smoothly navigate or pan across a map and fly into or out of a map to progressively reveal more or less detail. The systems best suited to this purpose are high-performance, virtual-reality systems, such as Sandia National Laboratory's EigenVR. This software allows the continuous navigation and zooming of an information landscape. Regions of high document density are represented by mountains or hills in the landscape, rather than by circles. The Sandia system also provides real-time peak labeling using a phrase analysis of article titles as zooming occurs. Searching by keyword or other descriptor is handled by an SQL pass-through to a relational database. Search hits light up as points on the map the user can navigate to. The EigenVR software can efficiently display the 36,000 document dataset described above on a Silicon Graphics O2 workstation with 256 MB of memory (Hendrickson et al., 1997). Future enhancements will likely expand this capacity significantly.

Conclusions

This article has described an approach to science mapping that utilizes citation data to provide a two-dimensional ordination of tens of thousands of documents across a multidisciplinary sample of scientific papers. Methods such as fractional citation counting, co-citation clustering, and triangulation have been combined with new methods which produce a unified ordination of a hierarchical arrangement of documents. The main novelty of the new approach is the ability to create a nested mapping, or volvox display, coordinating and arranging the details of lower level objects within higher level objects.

The structure when explored in terms of the distribution of topics covered by the map turns out to be one of the most multidisciplinary high-level maps yet to be generated using

co-citation data, ranging from astrophysics to sociology (Small, 1993). When compared with overall database statistics, the topic representation turns out to be slightly skewed to the physical sciences and away from biology, due to the initial fractional citation threshold. Adjustments in these thresholds should result in a more balanced representation of fields.

We have argued that a map of bibliographic data is a useful heuristic device by providing a visible organizing structure to information. One of our goals is to extend this to larger datasets. There are a number of strategies we can use to increase the comprehensiveness of the maps, such as employing lower thresholds and more hierarchical levels. In addition, combining other indirect citation link forms with co-citation can help extract maximum linking capability from citation data. Beyond this we can turn, as others have attempted (Braam, Moed, & van Raan, 1991a), to combining word and citation data. However, rather than aiming for total inclusiveness, the more limited objective of mapping the highly cited literature over a broad range of disciplines is a more realistic goal in the short run.

Many questions remain regarding the optimum methods to achieve this end. Clustering and particularly single-linkage has come under criticism (Burgin, 1995), but the method works well with citation linkage data, perhaps due to its high precision and inherent linearity. Triangulation as an ordination method is clearly inferior to the optimized configurations of multidimensional scaling, but may be adequate for showing patterns of strong linkage where considerable uncertainty may attend the weaker linkages. Another issue with bearing on computer visualization is a possible overexpansion of the coordinate space, and resultant reduction in document density. This places an additional burden on the computer display to efficiently and quickly move between areas containing relevant hits that may be separated by a great deal of empty document space. Tighter packing is always possible, but may lead to a loss of resolution between topics.

The most powerful and revealing types of analyses, however, exploit the network of linkages that propagate from document to document and eventually discipline to discipline. Interdisciplinary or crossover fields are frequently encountered, and the location of a field can occasionally defy its disciplinary origins. The patterns of linkage offer the possibility of exploring extended knowledge pathways, which in principle could traverse the entire map from social science to physics. The automatic generation of such pathways and their relationship to the potential for new knowledge and discovery in science remain to be explored. Perhaps for the first time, we will be able to use a map to purposefully navigate from one topic to another via a chain of key documents, and thereby creating a visualization of Vannevar Bush's associative trails (1945).

Acknowledgment

Support from Sandia National Laboratory, contract number AR-8321, is gratefully acknowledged.

References

Amsler, R.A. (1972). Applications of citation-based automatic classification. Unpublished report, University of Texas, Austin, TX.

Baldi, S. & Hargens, L.L. (1995). Reference network structure in turn-of-the-century physics: The case of N-rays. In M.E.D. Koenig & A. Bookstein (Eds.), Proceedings of the Fifth Biennial Conference of the International Society for Scientometrics and Informetrics (pp. 43–52), Medford, NJ: Learned Information.

Braam, R.R., Moed, H.F., & van Raan, A.F.J. (1991a). Mapping of science by combined co-citation and word analysis. I. Structural aspects. Journal of the American Society for Information Science, 42(4) 233–251.

Braam, R.R., Moed, H.F., & van Raan, A.F.J. (1991b). Mapping of science by combined co-citation and word analysis. II. Dynamical aspects. Journal of the American Society for Information Science, 42(4), 252–266.

Burgin, R. (1995). The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. Journal of the American Society for Information Science, 46(8), 562–572.

Bush, V. (1945). As we may think. Atlantic Monthly, 176, 101–108.

Callon, M., Law, J., & Rip, A. (1986). Qualitative scientometrics. In M. Callon, J. Law, & A. Rip (Eds.), Mapping the dynamics of science and technology (pp. 103–123). London: Macmillan.

Chalmers, M. (1996). A linear iteration time layout algorithm for visualising high-dimensional data. In R. Yagel & G. M. Nielson (Eds.), Proceedings: Visualization '96 (pp. 127–132), New York: Association for Computing Machinery.

Cherfas, J. (1996). Hot by any name: *sdi1* gene joins the roster. Science Watch, 7, 7.

Cherfas, J. (1995). Reports on cell-cycle regulators finally getting their due. Science Watch, 6, 5.

Garfield, E. (1979). Citation indexing—Its theory and application in science, technology, and humanities. New York: John Wiley.

Garfield, E., Sher, I.H., & Torpie, R.J. (1964). The use of citation data in writing the history of science. Philadelphia: Institute for Scientific Information.

Garofano, R. (1965). A graph theoretic analysis of citation index structures. Unpublished master's thesis, Drexel University, Philadelphia, PA.

Griffith, B.C., Small, H., Stonehill, J., & Dey, S. (1974). The structure of scientific literatures II: Toward a macrostructure and microstructure for science. Science Studies, 4, 339–365.

Hendley, R.J., Drew, N.S., Wood, A.M., & Beale, R. (1995). Narcissus: Visualising information. In N. Gershon & S. Eick (Eds.), Proceedings: Information Visualization '95 (pp. 90–96), Los Alamitos, CA: IEEE Computer Society Press.

Hendrickson, B., Wylie, B., Johnson, D., Davidson, G., Meyers, C., Small, H., & Pendlebury, D. (1997). Navigating science. (Abstract). AIPA97: Proceedings of the Symposium on Advanced Information Processing and Analysis, 25–27 March 1997 (p. 34), Fort Meade, MD: National Security Agency.

Holton, G. (1996). Einstein, history, and other passions (Chapt. 5). Reading, MA: Addison-Wesley.

Holton, G., Chang, H., & Jurkowitz, E. (1996). How a scientific discovery is made: A case history. American Scientist, 84, 364–375.

Hummon, N.P. & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. Social Networks, 11, 39–63.

Institute for Scientific Information. (1997). National Science Indicators on Diskette, 1981–1996: Documentation. Philadelphia: Author.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14, 10–25.

Larson, R.R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. In S. Hardin (Ed.), Proceedings of the 59th Annual Meeting of the American Society for Information Science: Global complexity: Information, chaos and control (pp. 71–78), Medford, NJ: Information Today.

Lee, R.C.T., Slagle, J.R., & Blum, H. (1977). A triangulation method for the sequential mapping of points from *N*-space to two-space. IEEE Transactions on Computers, 26, 288–292.

Lin, X. (1997). Map displays for information retrieval. Journal of the American Society for Information Science, 48, 40–54.
Mackinlay, J.D., Rao, R., & Card, S.K. (1995). An organic user interface for searching citation links. In I.R. Katz et al (Eds.), CHI '95 Conference Proceedings: Human factors in computing systems (pp. 67–73), New York: Association for Computing Machinery.
Marshakova, I.V. (1973). A system of document connections based on references. Scientific and Technical Information Serial of VINITI, 6, 3–8.
McCain, K. (1996). Personal communication.
Pitkow, J. & Pirolli, P. (1997). Life, death, and lawfulness on the electronic frontier. In S. Pemberton (Ed.), CHI '97 Conference Proceedings: Human factors in computing systems (pp. 383–390), New York: Association for Computing Machinery.
Price, D.J.D. (1965). Networks of scientific papers. Science, 149, 510–515.
Savoy, J. (1996). An extended vector-processing scheme for searching information in hypertext systems. Information Processing & Management, 32, 155–170.
Small, H. (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. (1993). Macro-level changes in the structure of co-citation clusters: 1983–1989. Scientometrics, 26, 5–20.
Small, H. (1994). A SCI-Map case study: Building a map of AIDS research. Scientometrics, 30, 229–241.
Small, H. (1995). Navigating the citation network. In T. Kinney (Ed.), Proceedings of the 58th Annual Meeting of the American Society for Information Science: Forging new partnerships in information (pp. 118–126), Medford, NJ: Information Today.
Small, H. (1997). Update on science mapping: Creating large document spaces. Scientometrics, 38, 275–293.
Small, H. & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. Journal of Information Science, 11, 147–159.
Small, H. & Greenlee, E. (1990). A co-citation study of AIDS research. In C.L. Borgman (Ed.), Scholarly communication and bibliometrics (pp. 166–193). Newbury Park, CA: Sage.
Small, H. & Griffith, B.C. (1974). The structure of scientific literatures. I: Identifying and graphing specialties. Science Studies, 4, 17–40.
Small, H. & Rothman, H. (1994). Investigations into the structure of science and social science using the SCI-Map system. In Identifying innovation in social science: Some bibliometric approaches (SPSG Review Paper No. 8), London: Science Policy Support Group.
Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the Science Citation Index® using co-citations. II. Mapping science. Scientometrics, 8, 321–340.
Sneath, P.H. & Sokal, R.R. (1973). Numerical taxonomy. San Francisco: W.H. Freeman.
Steinberg, S.G. (January 1997). Mapping science. Wired, p. 46.
Swanson, D.R. (1987). Two medical literatures that are logically but not bibliographically connected. Journal of the American Society for Information Science, 38, 228–233.
White, H.D. & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. Journal of the American Society for Information Science, 32, 163–171.
Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In N. Gershon & S. Eick (Eds.), Proceedings: Information Visualization '95 (pp. 51–58), Los Alamitos, CA: IEEE Computer Society Press.
Yermish, I. (1975). A citation based interactive associative information retrieval system. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, PA.