Scientometrics, Vol. 26. No. 1 (1993) 5-20

MACRO-LEVEL CHANGES IN THE STRUCTURE OF CO-CITATION CLUSTERS: 1983-1989

H. SMALL

Institute for Scientific Information (ISI) 3501 Market Street, Philadelphia, PA 19104 (USA)

(Received January 6, 1992)

At ISI we have used a consistent method for clustering the combined *Science Citation Index* and *Social Sciences Citation Index* for the last seven years (1983 to 1989). This method involves clustering highly cited documents by single-link clustering and then clustering the resultant clusters, a total of four times. This gives a hierarchical or nested structure of clusters four levels deep. Relationships among clusters at a given level can be depicted by multidimensional scaling, and by comparing successive year maps we can see how the relationships of major disciplines have changed from year to year. We focus mainly on the two highest levels of aggregation, C4 and C5, to make observations about structural changes in science involving the major disciplines. Distinction is made between changes which appear to be cyclic or oscillatory in nature, and those which appear to be more permanent or unidirectional.

Introduction

Of the various research endeavors in the field of bibliometrics, perhaps the most challenging is the attempt to map the structure of science. The mapping of science is based on a number of premises. First, that scientific knowledge can be represented as a network of concepts or ideas, and that these elementary entities can be aggregated to form macro-structures which bear some resemblance to the traditional branches of knowledge and disciplines of science. It is not important that this network resemble a geographic map, or cleanly separate individual topics, any more than a map of the brain's neuron connections would neatly organize human knowledge, but rather that the network is represented as truly and accurately as possible.

Second, it is assumed that each map is a snapshot at a distinct point in time of what is actually a changing and evolving structure of knowledge. It should be possible to follow this evolution either at the micro-level, where we deal with histories of individual scientific ideas and specialties, or at the macro-level, where change occurs in entire bodies of knowledge or their interactions with one another. This simultaneous change in multiple, interacting systems can be viewed as streams which

flow in parallel, sometimes converging to form broad rivers or diverging into smaller rivulets across time.

At ISI we have used the same method for clustering the combined *Science Citation Index®* and *Social Sciences Citation Index®* from 1983 to 1989. [1]This method involves linking highly cited documents by co-citation, applying single-link clustering and then clustering the resultant document clusters a total of four times, giving a hierarchical or nested structure of clusters four levels deep. Here I will be concerned with the highest, or most inclusive levels of the clustering, which come closest to showing the relationships between scientific fields or disciplines. My interest is in discerning change or continuity in such relationships over the seven year period.

The study of change at the field or disciplinary level raises difficult conceptual and methodological problems. This is because fields of science not only change internally as their knowledge bases change, but also externally in their relations to other fields. Some attempts have been made to examine disciplinary change using journal citation patterns,[2] author co-citation,[3] and document co-citation coupled with word similarity analysis.[4]

The use of a document clustering methodology, of course, makes no assumptions about the boundaries or interconnections of scientific fields. Rather it attempts to reconstruct science *a priori* from its elementary particles, the scientific papers. The laws which govern such an *Aufbau* or build-up of science, while constrained by the physical laws of Nature, are mainly sociological and psychological ones, in that they derive from the authors who write the papers and select what references to cite. A citation-based clustering method assumes, for example, that scientists in the same disciplines cite, by and large, the same pool of references, and also that the intensity of common referencing is an indicator of whether the entities are in the same discipline. Thus, we are concerned with group behavior, and whether it is purposeful or coordinated in some way.

Such an *a priori* approach of course, has its own problems. We do not know what principles should govern the *Aufbau* process. Specifically we must operationalize the meaning of "common referencing". For example, if we select single-link clustering, we obtain loose, weakly linked networks of research areas, whose constituents may only share references with their immediate neighbors. Complete-linkage clustering, on the other hand, yields only solidly linked and more isolated blocks of researchers, where each constituent must share references with every other. Sociological theory suggests that the method of linkage may vary with field.[5] We have used single-link clustering because of its simplicity of implementation for massive files.

Methodology

I will briefly review the methodology for producing high level clusters and maps to represent them. The single-link clusters of documents (about 9000 per annual file having two or more highly cited documents and containing a total of about 60,000 cited items) are called C1 clusters. Each of these clusters is collapsed to form a single super-document. All super-documents (C1 clusters) are subject to a second clustering (denoted C2) which yields single-link clusters of clusters, called C2 clusters. These number about 1000.

Continuing the process, each C2 cluster is taken as a super-document and clustered to form about 100 C3 clusters. By the fourth iteration, C4, the number of super-clusters has been reduced to about 10. Hence, with each level the number of entities is reduced by nearly a factor of ten. Of course, isolates are formed at each level along the way so that only about one-third of the original clusters are contained in the final C4 set, but these usually include the largest clusters.

The methodology for generating clusters at each level involves progressively raising the normalized co-citation threshold starting from some minimum value (e.g. zero) until single-link clusters are formed which do not contain more than a specified number of entities from the previous level. For all years and levels this maximum size has been set at 60. At the lowest possible threshold for a given level the majority of entities cluster together, and exceed the size limitation. For example, on Fig. 1 the systematic biology C4 cluster (#3) forms at 0.025 by breaking off from the main group. As the threshold continues to be incremented, cluster #1 (biomedical, physical, and social/behavioral sciences) and cluster #13 (earth science) disaggregate at 0.029. Each of these C4 clusters when formed contains by definition fewer than the maximum allowed number of C3 clusters.

We can think of the changing associations of disciplines as a process of competitive binding among fields, analogous to atoms competing for a binding site on a molecule. Competition comes about because we limit the number of entities which can bind by using a maximum cluster size. If more than this number of entities bind together, the linkage threshold is raised until enough fragments disengage that the resulting aggregates are within the size limitation. These fragments can then cluster together at the next higher level.

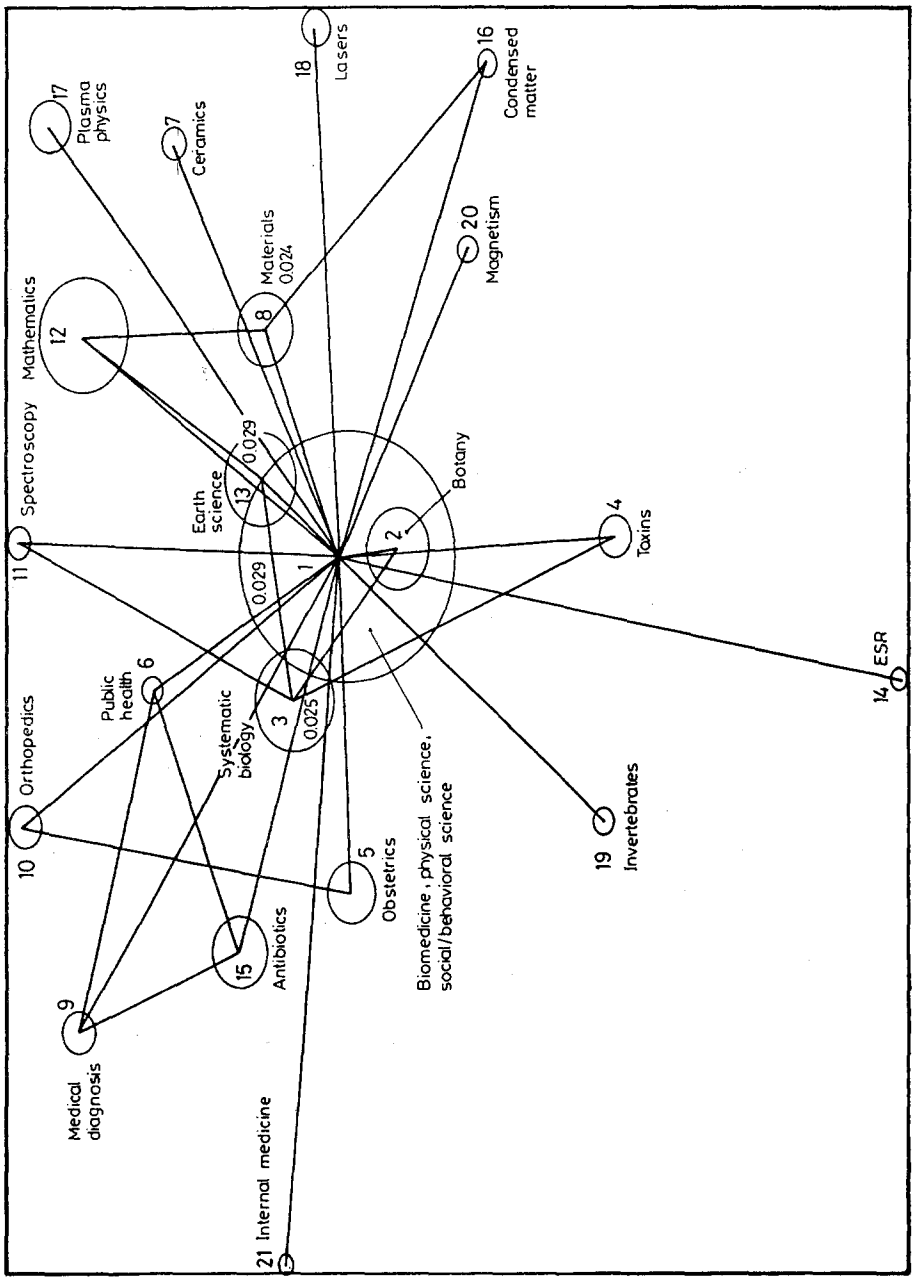

Fig. 1. 1984 SCI/SSCI C5 map

To give a hypothetical example, in one year field "A" might be strongly bound to field "B", but not so strongly bound to field "C". If the presence of field "C" pushes the cluster over the size limit, then the threshold will be raised until "C" forms a separate cluster. On the map for the next higher level we see the "A-B" aggregate linked to "C". If in the next year the link between field "C" and "A" has become stronger, then "C" may displace "B", and "B" will form a separate cluster. The map will then show the "A-C" aggregate linked to "B".

The results of a cluster analysis are traditionally shown as a tree structure. However, because clusters are constructed from linkages among objects it seems natural to display clusters as networks of connected nodes. The technique of non-metric multidimensional scaling[6] or other methods such as centroid scaling[7] can be used to display clusters by locating each of the objects at a point in space. Ideally, the location should represent the relation of that object to the other objects in the space. However, scaling is used here only to obtain an approximate representation of a network, and not to determine precise locations of objects. In this sense, the presence or absence of links is more significant than exact location.

When we compare maps from different years, we can see how fields change over time in their relations to each other, provided we know the correspondence between clusters across the years. The information is provided by cluster strings. The links between clusters across time are based on a normalized measure of the number of common highly cited documents in successive year clusters. A sequence of such continuing clusters is called a cluster string. Forming the string is itself a single-link clustering process. If inter-year linking is applied to higher level clusters, which are aggregates of lower level clusters, we can track the development of disciplines or specialty aggregates from one year to the next.

It is important to stress that each annual mapping is, in a methodological sense, independent of previous or later maps. We make no attempt to force a field appearing in one year to appear the next, for example by allowing an overlap in the citation data sets. Only citation and co-citation thresholds are held roughly constant from year to year. Also we have not attempted to rotate the scaling solutions to achieve maximum congruence between maps in different years. In this sense map orientation from year to year is arbitrary. However, in some cases it is not difficult to see how a rotation or reflection would bring two maps into better correspondence.

Two factors make continuity from year to year more difficult to achieve. First, since distributions of citation and co-citation scores obey the usual hyperbolic laws, items or links whose scores are close to the thresholds may be selected or not based

on very small changes. Second, the single-link clustering algorithm being a weak criterion for clustering has a tendency to form chained structures held together by single links. If critical links disappear below the threshold, the structure may be significantly altered. This instability is, of course, compounded by applying single-link clustering four times to create the higher level structures. Such instability to initial conditions is of course characteristic of fractal systems.[8]

Maps at C4 and C5

The maps shown in Figs 1 through 6 include the C5 maps for the years 1984, 1987, and 1988, and a map for the largest C4 cluster in the years 1984, 1985 and 1988. Two 1983 maps (a C5 and a C4) also relevant to this discussion were published previously, and may be referred to.[9] Copies of C5 and C4 maps for all years discussed here are available from the author. With these maps we can begin to examine in a qualitative way the changes in association of the major disciplines, which for the purpose of this analysis can be designated roughly as biomedicine, physics, chemistry, biology (including ecology), and social/behavioral science. Other easily distinguished fields such as mathematics, geoscience, and materials science will be discussed as well. In some cases the C5 maps have been labeled to show the co-citation thresholds at which the C4 clusters were formed to assist in understanding the changes from year to year.

The consistent feature of the C5 maps (Figs 1 through 3) is the presence of a large central super-cluster which always includes a large portion of biomedical science and usually, but not always, significant portions of the physical sciences. Surrounding this large central region are a more variable set of medium sized and small areas, all of which link to the central region. Many but not all of these outlying areas are more applied in nature than the centrally located areas.

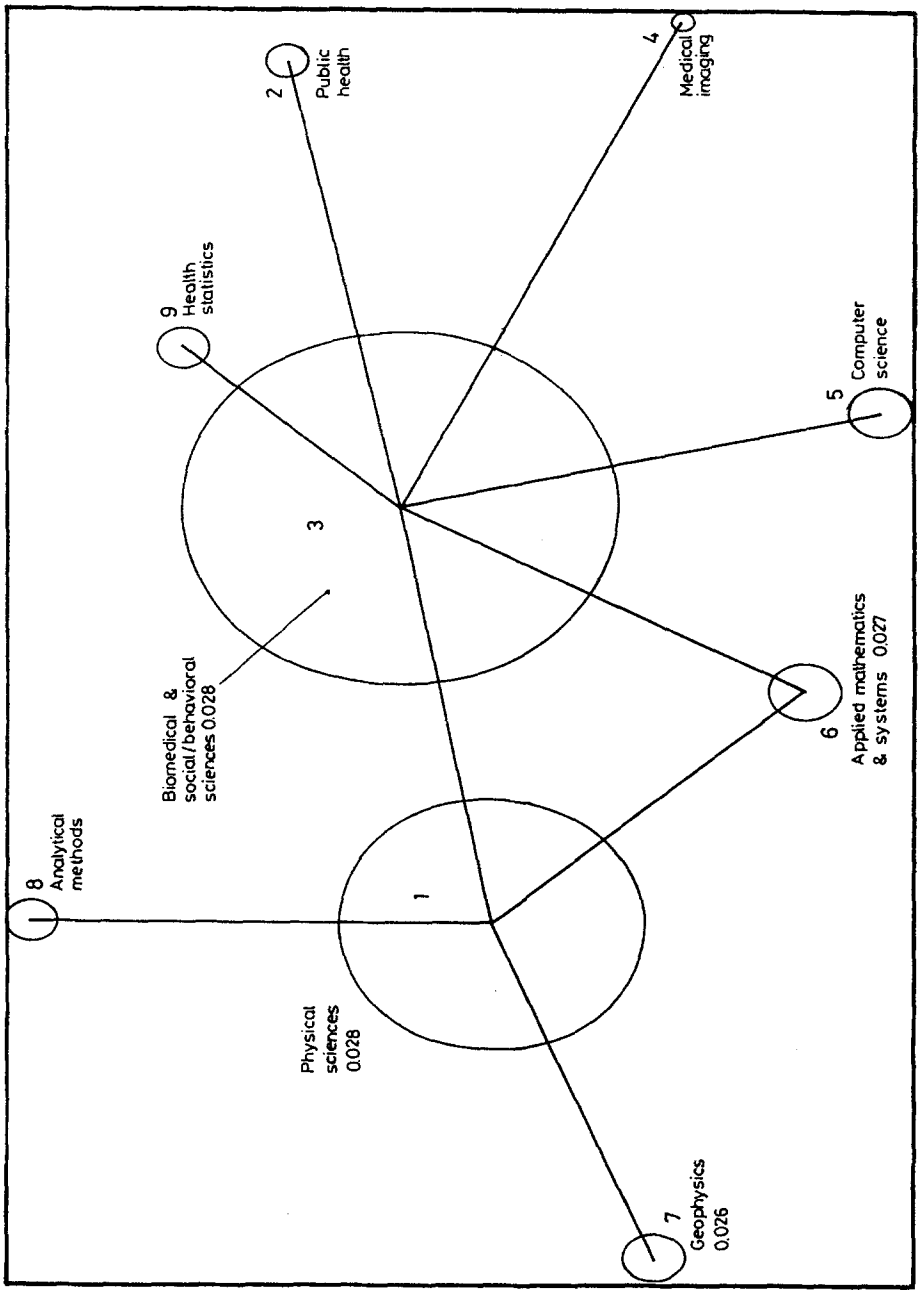

Fig. 2. Cluster map CS 0001 1987

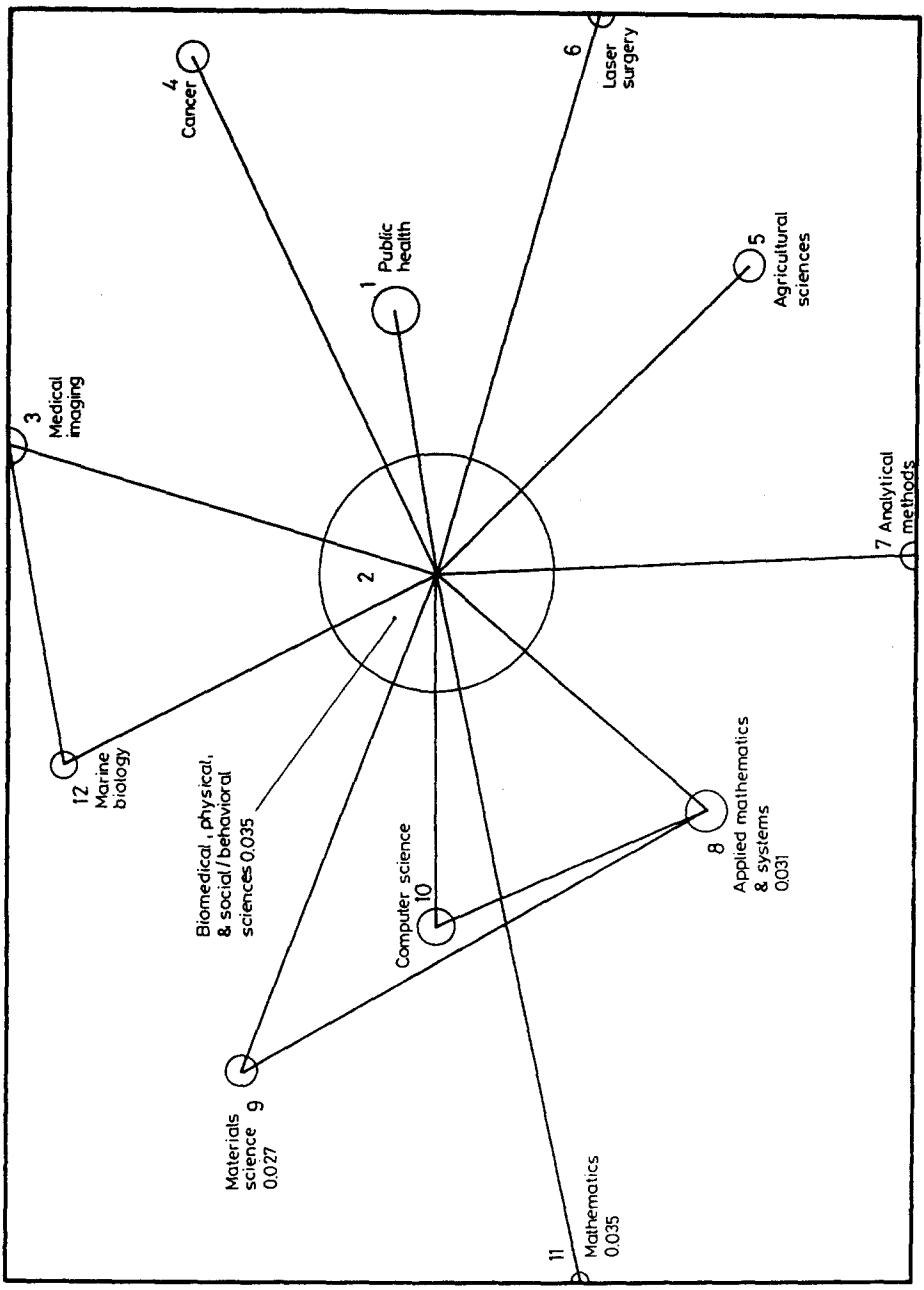

Fig. 3. 1988 global map

In most years the central C4 cluster dominates the C5 maps in terms of size, but in 1985 and 1987 (Fig. 2), the physical sciences form their own C4 cluster separate from the biomedical C4. This physical science C4 is the second largest area on these maps.

Turning to the maps for the largest C4 clusters (Figs 4 through 6), we find a large number of loosely linked C3 clusters, the largest of which is usually also biomedical. The fields most persistently linked at the C4 level are physics and chemistry. They have been linked in a C4 cluster in each of the seven years analyzed. Chemistry also often plays a mediating role by linking biomedicine to physics. In five of the seven years that the physical and biomedical sciences have co-existed in the same C4 cluster, the mediating field between chemistry and biomedicine has been organic or protein chemistry (see C4 maps for 1983, 1984, 1986, 1988 and 1989; e.g. Figs 4 and 6).

In these largest C4 clusters social/behavioral science, biomedicine, chemistry, and physics have aggregated in four of the seven years (C4 maps for 1984, 1986, 1988, and 1989; e.g. Figs 4 and 6). The largest C4 in 1983[10] contained biomedicine, chemistry and physics, but lacked the social/behavioral sciences which joined biological science in another C4 cluster.

With the exception of 1983, biomedicine is linked to the social/behavioral sciences in the largest C4 cluster. The point of attachment of social/behavioral science to biomedicine is usually neuroscience (see C4 maps for years 1984, 1985, 1987, 1988; e.g. Figs 4, 5, and 6).

The shift of social/behavioral science to biomedicine is brought about by a weakening of its link to biological science and a strengthening of its link to biomedicine. This is seen in the C4 cluster thresholds. In 1983 the behavioral-biological science C4 cluster was formed at a threshold of 0.027, while in 1984 biological science clustered separately from behavioral science at a lower threshold, namely 0.025 (Fig. 1). At the same time the aggregate of behavioral science and biomedicine was formed at a higher threshold, namely 0.029.[11]

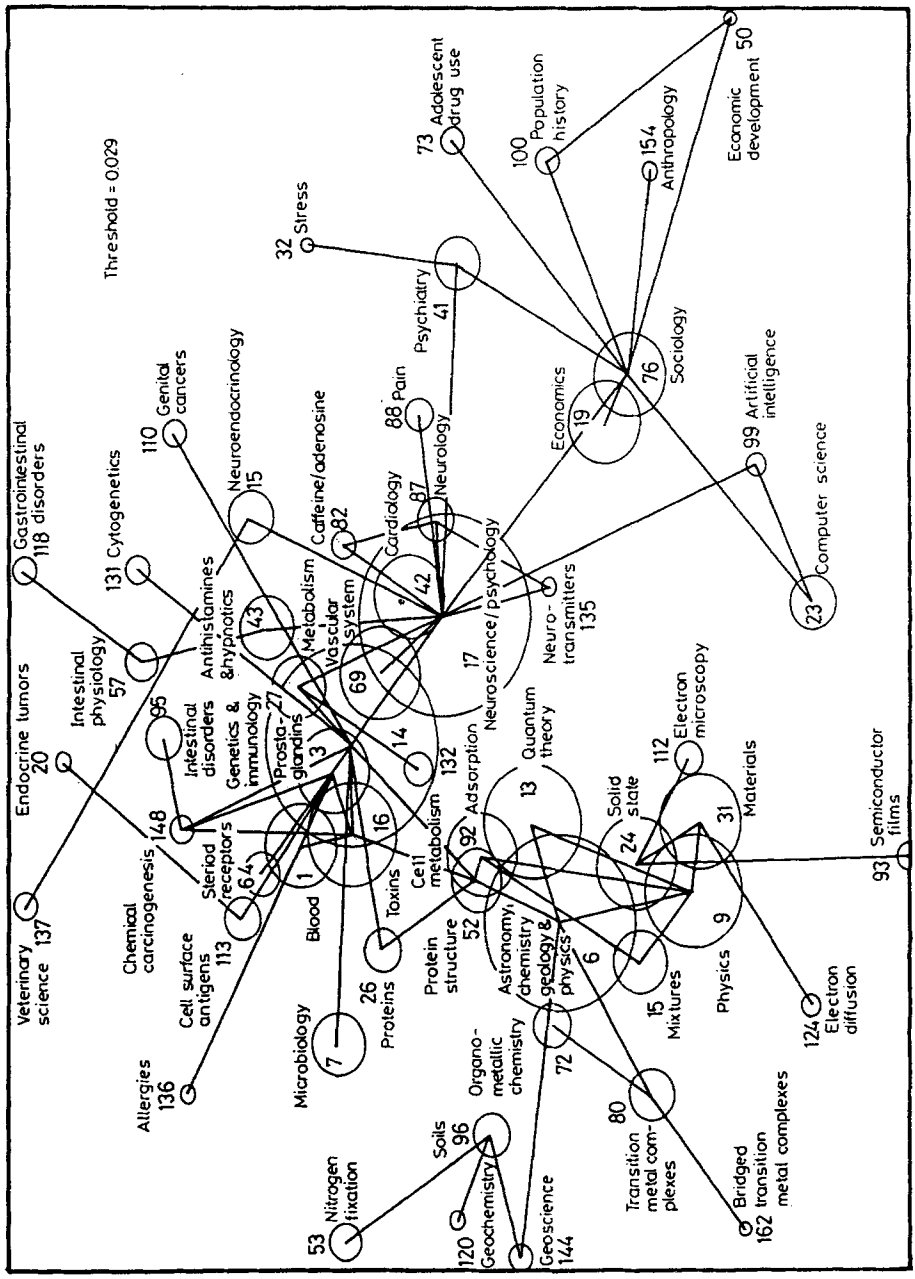

Fig. 4. 1984 C4 clusters, cluster num 00001

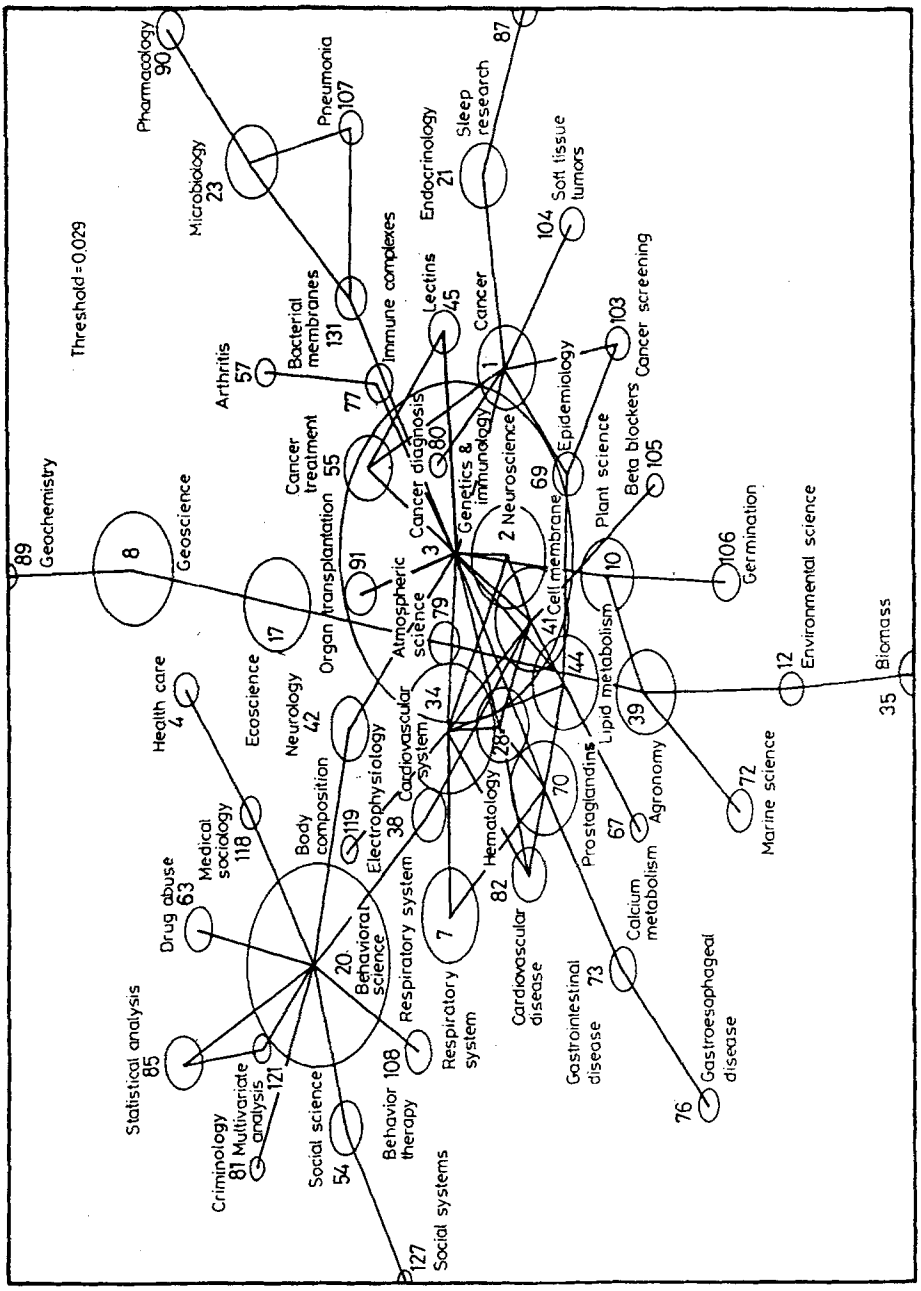

Fig. 5. 1985 C4 clusters, cluster num 00001

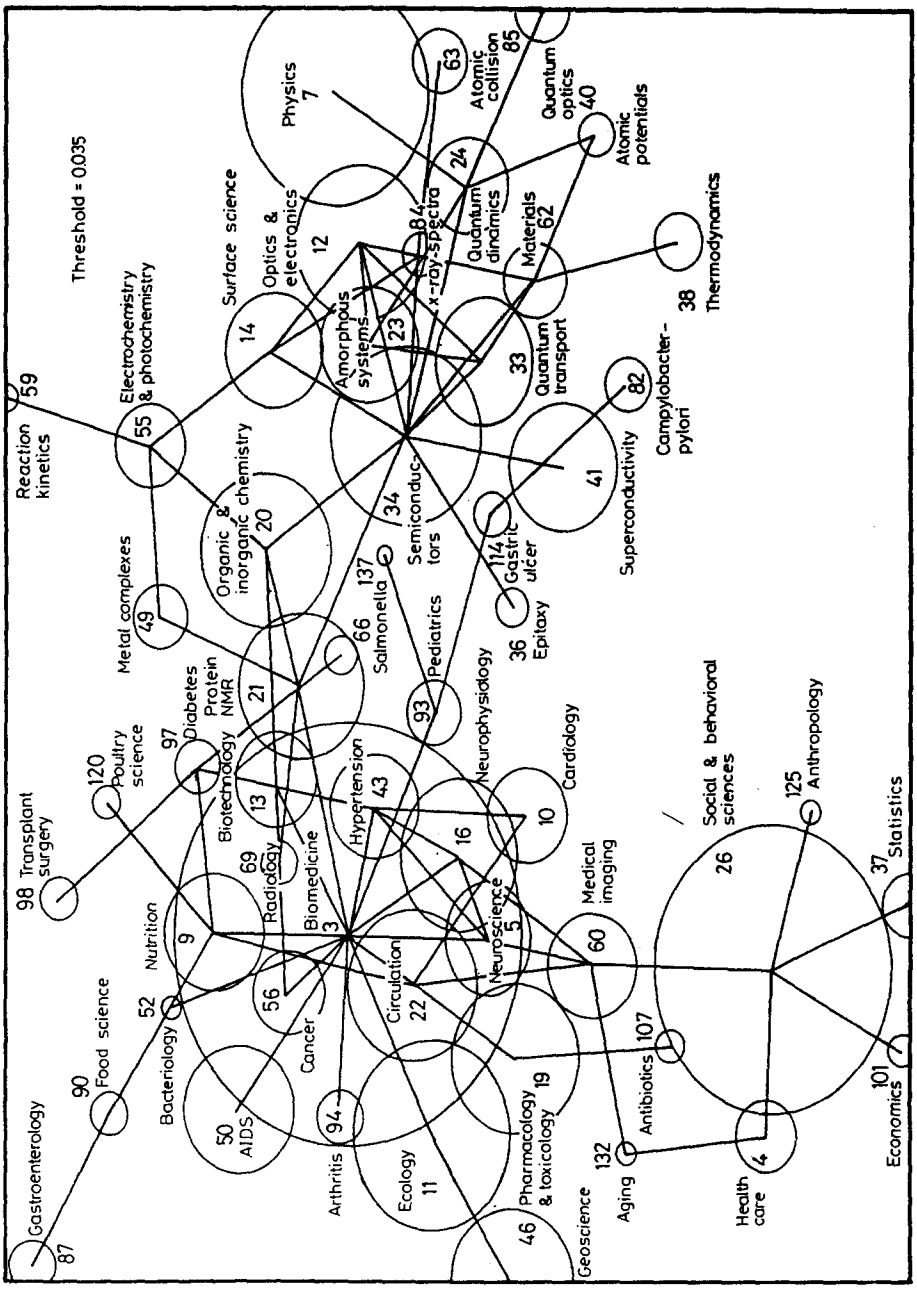

Fig. 6. 1988 C4 map: cluster #2

Only in the years 1985 and 1987 does the largest C4 cluster *not* contain the physical sciences. As noted above, in those two years chemistry and physics are contained in a separate large C4 cluster. This separation of physical science from biomedicine appears to be due to a strengthening of the link between biological science and biomedicine, and weakening of the link between chemistry and biomedicine. Biological science had formed a separate C4 cluster in 1984 at 0.025 (Fig. 1). In 1985 it aggregates with biomedicine at a higher threshold, namely 0.029. Similarly, chemistry and physics had aggregated with biomedicine in 1984 at 0.029, but in 1985 physics and chemistry form a separate entity at that level. Thus in 1985, biological science appears to have displaced physical science from linking with biomedicine. The point of attachment of biology to biomedicine in 1985 is plant biotechnology (Fig. 5).

Turning to 1986 we see the situation reversed. Physics and chemistry have rejoined biomedical science at an even higher threshold (0.032) than the one at which they were separately formed the previous year (0.029). At the same time, biological science separates from biomedical at a lower threshold (0.027), indicating a weakening of that connection.

This cycle repeats again in 1987 (Fig. 2). Chemistry and physics separate from biomedicine at a lower threshold (0.028) than the one at which they aggregated the previous year (0.032), and biological science again attaches to biomedicine at a higher threshold (0.028) than its previously separate existence (0.027). In 1988 (Fig. 3) chemistry and physics rejoin biomedicine at a higher threshold (0.035) than their prior separate threshold (0.028). This time, however, biological science does not disengage, but remains attached to biomedicine (Fig. 6). No displacement occurs, but there is an increase in binding strength (0.028 in 1987 to 0.035 in 1988).

Further insight into these shifts from year to year can be gained by examining the internal structure of the C4 clusters (Figs 4 through 6). In those years in which the biomedical and physical sciences co-exist in the same C4 cluster (1983, 1984, 1986, 1988, and 1989; e.g., Figs 4 and 6) the structure is essentially linear. For example, on the 1983 map,[12] we have the linear progression of biomedicine to chemistry to physics to mathematics. In 1984 (Fig. 4) the linear arrangement is social/behavioral sciences to biomedicine to chemistry to physics. This pattern is repeated in 1986, 1988, and 1989 (e.g. Fig. 6). In 1985 (Fig. 5) and 1987 the linear pattern is social/behavioral science to biomedicine to biological science. Such recurrent disciplinary connections suggest that these relationships are persistent. On the other

hand since they form a chain, a weakening of one link can bring about a rearrangement or reordering of a discipline, or its displacement by another.

The positions of other fields can also oscillate in other ways, namely between levels. For example, in 1983[13] and 1985 mathematics is attached to physics on the C4 map as a C3 cluster, while in 1984 (Fig. 1) and 1986 mathematics appears on the C5 map as a C4 cluster. This separate position on the C5 map continues in 1987, 1988 and 1989 (e.g. Figs 2 and 3), and mathematics does not cycle back again to physics.

Geoscience presents a similarly variable picture, but without a clear direction. It appears on the C5 map as a moderately large C4 cluster in 1983, 1984, 1986, and 1987 (e.g. Figs 1 and 2). But in 1985 and 1988 it is in the largest C4 cluster as a C3 cluster and is attached to biological science (Figs 5 and 6). Geoscience also appears in some years to be internally split, having one portion on the C5 map, and another on the C4 map attached to physical science (e.g. 1984, Figs 1 and 4).

Materials science is more clearly in a structural cycle. It appears as a distinct cluster on the C5 maps in 1983 and 1984 (e.g. Fig. 1), then submerges into the C4 map where it is attached to physics in 1985, disappears from view in 1986 and 1987, and then reappears on the C5 map in 1988 (Fig. 3) and 1989, thus coming full circle.

Discussion

One interpretation of this disciplinary cycling is that there is a structural oscillation between expansion and contraction in the disciplines. A similar pattern has been observed in case studies of cluster maps at the document, or C1 level.[14] These case studies have suggested a kind of pulsating model of specialty development, with alternating periods of discovery and consolidation. Discoveries appear as small densely linked groups of documents which are somewhat isolated. Following discovery there is a period of expansion when the field develops and ramifies the discoveries into a wider range of phenomena. Clusters representing this later stage are larger and more loosely structured than the discovery clusters.

This kind of alternation between periods of discovery and integration could also be occurring on the disciplinary scale. In one year a discipline might contract in order to build internally, and in a subsequent year reach out and link with other disciplines in order to apply its new findings in other fields. For this kind of coordinated collective behavior to exist there would have to be a feedback mechanism of some kind which effectively coordinated the research of the individual scientists in the discipline.

Another interpretation is that these structural oscillations are due to the method's sensitivity to initial conditions, rather than any changes in the relations of scientific fields, and that the true picture would show essentially static relations among disciplines, or only gradual changes. We do observe many consistent or recurrent patterns in these structures, such as the repeated linking of fields or linear sequences. The changes from year to year, in some instances, appear more sudden or discontinuous than warranted since the structure often returns to its former state the following year with little or no apparent progression. A middle ground is that the present methodology exaggerates or magnifies what are actually small changes. It remains to be seen, of course, how a change in the clustering algorithm, for example use of complete rather than single-linkage, would affect the results.

The goal of mapping science has clearly not been fully achieved. We have succeeded in building up a structure using a series of four iterations of clustering. We cannot claim that this is the only structure possible, or that other methods of aggregation would not lead to different structures which are more easily interpreted. Clearly the present methods are only a first step toward an accurate recording and rendering of the structural evolution of scientific knowledge, let alone providing a theoretical basis for understanding it. Nevertheless, I believe that further progress can be made, by sharpening both methods and data. The issue at stake is the existence of a collective mind for science.

References

1. H. SMALL, E. GARFIELD: The geography of science: Disciplinary and national mappings, *Journal of Information Science*, 11 (1985) 147–159.
2. P. DOREIAN: Testing structural-equivalence hypotheses in a network of geographical journals, *Journal of the American Society for Information Science*, 39 (1988) 79–85.
3. K.W. MCCAIN: Longitudinal author co-citation mapping: The changing structure of macroeconomics, *Journal of the American Society for Information Science*, 35 (1984) 351–359.
4. R.R. BRAAM, H.F. MOED, A.F.J. VAN RAAN: Mapping of science by combined cocitation and word analysis. Part 2. Dynamic aspects, *Journal of the American Society for Information Science*, 42(4) (1991) 252–266.
5. R.D. WHITLEY: *The Intellectual and Social Organization of the Sciences*, Oxford; Clarendon, 1984.
6. J.B. KRUSKAL: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrica*, 28 (1964) 1–27.
7. E. NOMA: The simultaneous scaling of cited and citing articles in a common space, *Scientometrics*, 4 (1982) 205–231.
8. A.F.J. VAN RAAN: Fractal dimension of co-citations, *Nature*, 347 (1990) 626.
9. See Ref. 1.
10. See Ref. 1.

11. A preliminary examination of 1990 clusters reveals that social/behavioral science has again separated from biomedicine and formed a separate C4 cluster.
12. See Ref. 1.
13. See Ref. 1.
14. H. SMALL, E. GREENLEE: Collagen research in the 1970s, *Scientometrics*, 10 (1986) 95 – 117.