

## Mapping world-wide science at the paper level

Richard Klavans, SciTech Strategies, Inc., Berwyn, PA, 19312, USA  
email: rklavans@mapofscience.com

Kevin W. Boyack, Sandia National Laboratories, P.O. Box 5800, MS-0310, Albuquerque, NM 87185, USA email:kboyack@sandia.gov

**Abstract:** This article describes recent improvements in mapping a highly representative set of the world-wide scientific literature. The process described in this article extends existing work in this area in three major ways. First, we argue that a separate structural analysis of current literature vs. reference literature is required for R&D planning. Second, visualization software is used to improve coverage of the literature while maintaining structural integrity. Third, quantitative techniques for measuring the structural integrity of a map are introduced. Maps with high structural integrity, covering far more of the available literature, are presented.

### Introduction

The first major attempt to map the world-wide scientific literature was done at the Institute for Scientific Information (ISI) in the 1970's (Small & Griffith, 1974) using a method that relied on the analysis of co-citations (the co-occurrence of references in a set of scientific documents). This method was based on (a) selecting highly cited references, (b) using co-citation frequency to estimate relatedness between these reference papers, (c) clustering these reference papers, (d) assigning current papers to the document clusters, and (e) developing an additional measure of the relatedness between document clusters. These documents clusters were called specialties. Relationships between specialties were estimated and used to generate maps of science (Small, 1999; Small & Sweeney, 1985; Small, Sweeney, & Greenlee, 1985).

The second major attempt to map the world-wide scientific literature was done by the Center for Research Planning (CRP) in the early 1980's. Unlike ISI, where the research was more exploratory, CRP had a specific development goal in mind. Their interest was in a war map of science, where executives could plan their next manoeuvre (Price, 1986, see discussion in Small, 2003, p. 395). As an example, an executive at SmithKline Beecham described how the science-mapping techniques from CRP resulted in a major re-allocation of R&D dollars among their therapeutic areas, and was critical to their decision to be the first major pharmaceutical firm to invest in genomics (Norling, Herring, Rosenkrans Jr., Stellpflug, & Kaufman, 2000). CRP had a different conception than ISI of how these maps would be used, which led to a different development trajectory.

There were many methodological problems faced by ISI and CRP in these early years. They were limited by high computer costs – their work required the use of a mainframe computer. They used relatively small samples to represent the entire domain (ISI only sampled 1% and CRP sampled 2.6% of the available reference documents). Others were concerned that their analytical procedure (single-linked clustering) would provide incorrect results in the form of large heterogeneous clusters. It was unclear how to determine if the maps were accurate or not. Many of these concerns were addressed quite openly in an excellent review of the two techniques, but were left unresolved (Franklin & Johnston, 1988).

Since the 1980's, there has been little published on advances in the state of the art in analyzing and then visualizing a large number of documents (over 100,000) to create 'war maps of science'. The lack of progress is surprising given related scientific and technological development that could help create more comprehensive, more accurate and more affordable maps of science. Low cost computing is now widely available. Alternative measures of

relatedness have been proposed. Alternative clustering and visualization programs are available (Börner, Chen, & Boyack, 2003). The maps of science presented in this paper build upon these developments and introduce new methods for quantitatively assessing the structural integrity of a map of science. The explicit goal of the research reported in this paper is to generate objective, accurate maps of science, on off-the-shelf stand-alone PCs, that can be used by policy-makers to improve R&D planning and evaluation.

### **Research Communities from Two Perspectives**

There are two major ways to generate document clusters from a file of current documents. The first method (used by ISI and CRP) does not cluster the current documents per se. Co-citation analysis clusters a highly select set of reference papers. The reference papers are interpreted as exemplars (Small, 2003). Current papers are then assigned to these clusters of reference papers, and are sometimes referred to as a ‘research front’ (Price, 1986).

The second major method is to cluster the current documents based on common references, common words, or common phrases (for a review, see Börner et al., 2003). The resulting cluster of current papers is not the same as a research front. Far more current papers could be included – including current papers that are unlikely to have any impact in the future (e.g. they don’t have any bibliographies).

We suggest that there are important theoretical differences in a structural analysis of science based on current papers vs. reference papers. We use the concept of a research community (Kuhn, 1970) to highlight these theoretical differences. Clusters of current papers reflect current research communities, and represent how current research is organized around similar topics or research questions. Current research communities reflect a shared cognitive framework about research topics.

Clusters of reference papers represent how past research is currently being utilized. We label these as ‘base research communities’. Base research communities represent the building blocks of science that are being used by current researchers. Base research communities reflect the shared cognitive framework about the research methods used to address current topics.

We suggest that the distinction between current research communities and base research communities has important theoretical and practical implications. Relationships between these two structural representations need to be explored. For example, one might expect far more diversity in topics than methods (more current research communities than base research communities). Individuals, organizations or nations may have greater strengths in certain topics but relative weaknesses in underlying methods, or visa versa. Plans to improve one’s strengths are quite different if one wants to strengthen activity in a topic or capability in a method. These two types of communities have significantly different implications for R&D evaluation and planning.

This study develops two different maps using these two different perspectives. We use bibliographic coupling (Kessler, 1963) to identify current research communities and co-citation analysis to identify base research communities. In bibliographic coupling, two current papers are related if they cite one or more of the same references. One can use bibliographic coupling to cluster current papers in a manner that is parallel to the way that co-citation was used by ISI (Morris, Yen, Wu, & Asnake, 2003). One can then assign reference papers to these clusters in a similar fashion to the way this is done by ISI. The results, in both cases, are

clusters of documents where each cluster consists of a set of current papers and a set of reference papers.

### Measuring Relatedness at the Paper Level of Analysis

ISI and CRP used the raw co-occurrence frequency as the measure of relatedness between two papers. Two papers were considered more related if they had higher co-citation counts, and less related if they had lower co-citation counts.

Alternative measures of relatedness have been proposed in the literature (cf. Jones & Furnas, 1987) to overcome the biases associated with using raw frequency. We used two measures in this study. First, we used the raw frequency measures as a basis for comparison. Second, we used a modified cosine measure (a cosine adjusted for expected co-occurrence frequencies) because of previous work that suggested that this is the most accurate measure for co-citation analysis (Klavans & Boyack, in press).

The modified cosine measure (K50) was developed in order to adjust for expected co-citation levels before scaling for size. The expected value in each cell of the  $n \times n$  matrix, where  $n$  is the number of papers that are being clustered, depends on the total values in the corresponding rows and columns (with an adjustment for the fact that the diagonal in the matrix is zero). The equation used was:

$$K50_{i,j} = K50_{j,i} = \max \left[ \frac{(F_{i,j} - E_{i,j})}{\sqrt{S_i S_j}}, \frac{(F_{j,i} - E_{j,i})}{\sqrt{S_i S_j}} \right],$$

where  $F_{ij}$  is the frequency of co-occurrences of paper  $i$  and paper  $j$  in the  $n \times n$  matrix,

$$\bar{F}_i = \frac{1}{n} \sum_{k=1}^n F_{i,k}, \quad k \neq i,$$

$$S_i = \sum_{j=1}^n F_{i,j}, \quad j \neq i,$$

$$SS = \sum_{i=1}^n S_i.$$

Klavans & Boyack only tested the accuracy of the modified cosine measure for journal-journal relationships. This study looks further to see if there are comparable differences in modified cosine vs. raw frequency measures at the paper level of analysis.

### Improving Coverage

ISI and CRP reported that their initial analyses were based on an initial set of 5,000,000 reference papers. Both were faced with the same trade-off between coverage and structural integrity (in this context, structural integrity refers to the likelihood that papers in a cluster will be on the same subject). They only focused on papers with very high relatedness (the higher the relatedness, the more likely that the papers were on the same subject). They advocated for different clustering algorithms that would maintain structural integrity. ISI had a more stringent threshold for relatedness – using only 51,000 reference papers and subsequently clustering these documents into 9,420 groups (called specialties). CRP used a stratified sample to increase the coverage to 128,000 papers and used a different clustering method (generating 28,000 specialties). CRP claimed greater validity due to higher coverage, and ISI counter-claimed that structural integrity may have been sacrificed (Franklin & Johnston, 1988, p. 369).

This study uses visualization software to overcome some of the problems mentioned above. One advantage of visualization software is that one can significantly increase coverage (one can map the higher related and lower related data points) and then let the map determine the number and composition of clusters. Ideally, the highly inter-related papers would gravitate into clusters and the papers that have low relatedness will be isolated (i.e., there won't be a closely neighboring paper). One can then rely on nearest neighbor data to determine thresholds (whether points should be included in a cluster or not), and one can refer back to the map to determine if clusters of papers have been artificially split or over-aggregated.

The effect of the visualization software on coverage of the reference literature is dramatic. We have used the combined 2002 SCIE and SSCI databases to generate an initial set of 10,911,939 unique reference papers – twice the number used by ISI and CRP in the early 1980's. We excluded any reference paper that had a maximum co-citation frequency of three or less. In addition, we excluded any reference paper that had only one co-citation partner. There were 718,964 reference papers that met these thresholds – six times the number used by CRP and fourteen times the number used by ISI.

The visualization software also allows for a significant coverage of the current literature. The combined 2002 SCIE and SSCI databases generated an initial set of 1,069,764 current papers. We excluded all current papers with a maximum co-occurrence (bibliographic coupling) frequency of one, or that only co-occurred with one other current paper. Application of these thresholds left 731,289 papers to use in generating a map of current literature.

It is unclear, however, whether an increase in coverage using the method described above provides better insights into the structure of science. Increasing coverage may only increase the number of papers in a cluster (if the correct number of research communities were initially defined). The visualization method may result in structures that don't 'make sense' because of extreme dimensional reduction (from over 700,000 dimensions to 2 dimensions). It is this last issue (structural integrity) that will be addressed in the next section.

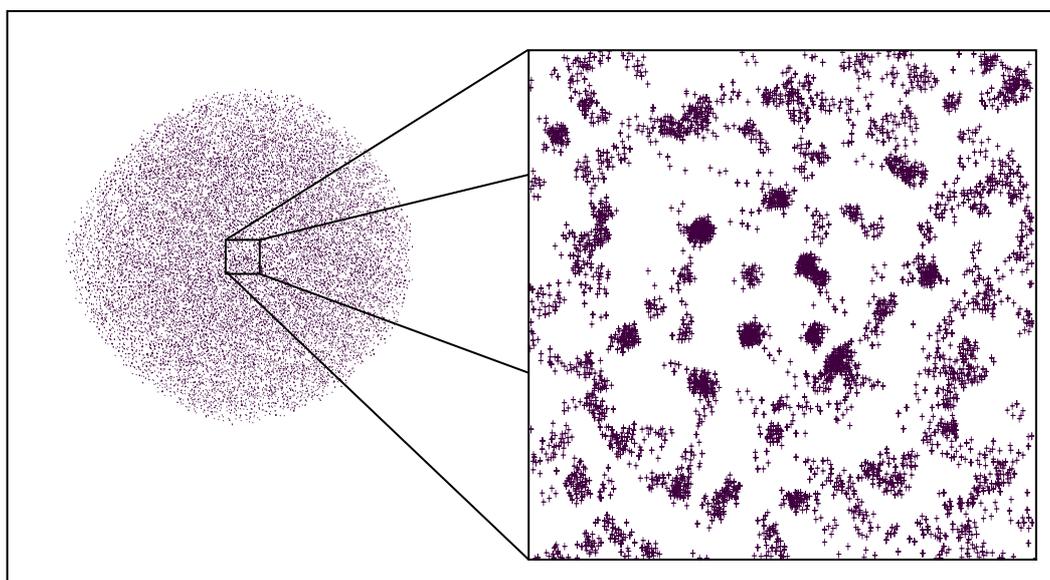
### **Structural Integrity of a Map of Science**

ISI and CRP used different algorithms to cluster their data, and then utilized expert judgment to determine if individual clusters 'made sense'. This means that experts in particular disciplines looked at the clusters of documents to determine if there was a coherency to the group. In general, the algorithms generated definable clusters and each clusters was judged to be on a particular topic (Small & Griffith, 1974). The clusters thus had structural integrity.

Since it has already been established by ISI and CRP that clusters of reference papers based on raw co-citation frequency counts and single-link clustering generate coherent clusters, experts were not used to judge whether the clusters generated in this study had structural integrity. There is little reason to suspect that the clusters of reference papers generated in this study, from the raw frequency measure and from a modified cosine measure that has been shown to be more accurate than raw frequency (Boyack, Klavans, & Börner, in press; Klavans & Boyack, in press), would have any less coherence. Nor was there any reason to suspect that clusters of current papers using these methods would not be coherent.

Quantitative methods were used to determine if one measurement approach generates maps that have greater (or lesser) structural integrity. We focus on two units of analysis: disciplines and research communities. For disciplines, we look at the tendency for papers in the same discipline to be located in the same area of the map, and for research communities we look at

the tendency for papers in a cluster to be assigned to the same discipline. Both aspects of structural integrity are defined and empirically assessed using the visualizations in Figures 1 as a guide.



**Figure 1: Maps of current science from ISI 2002 data using raw frequency (left). The panel on right is an 8x enlargement of the central section of the map.**

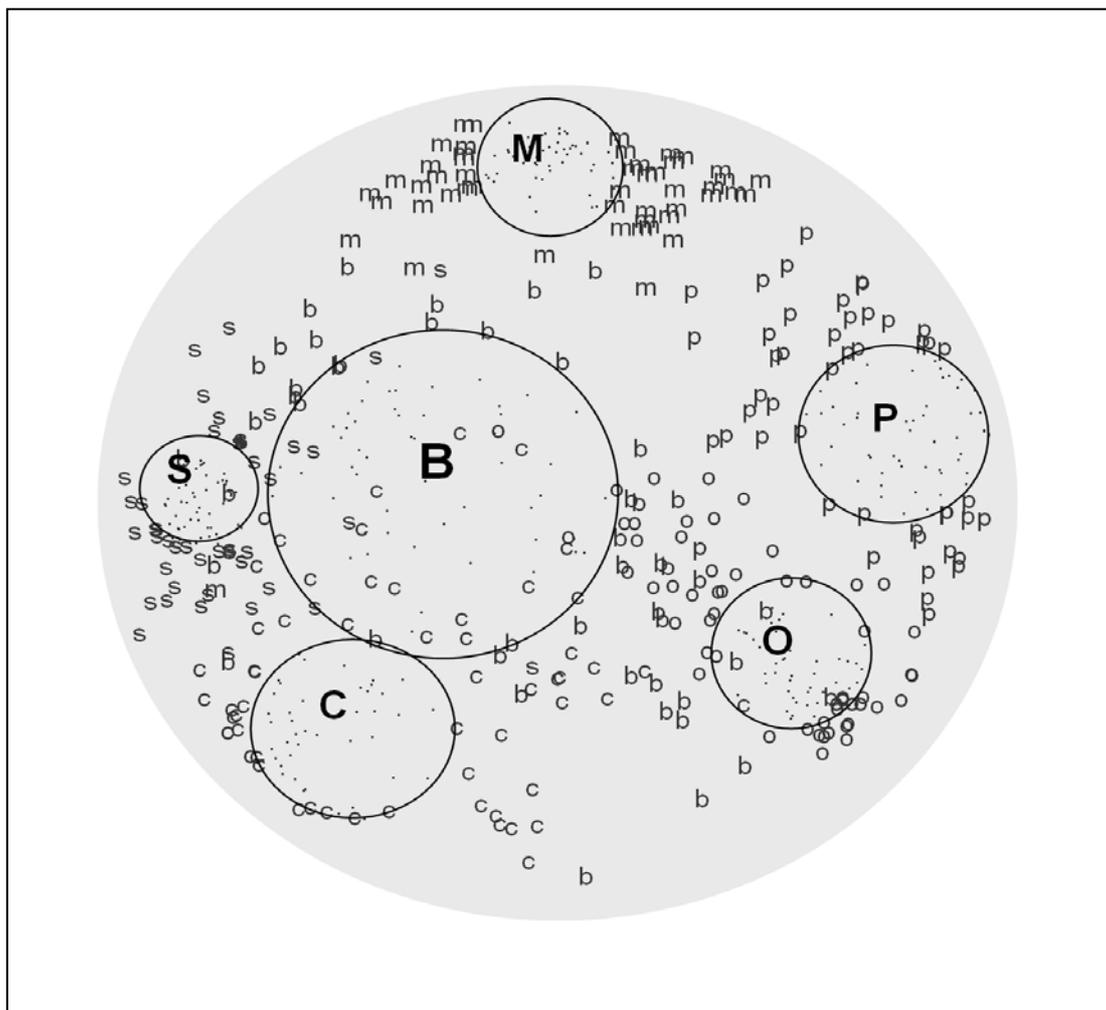
The map in Figure 1 is based on the raw frequency measure of relatedness between 731,289 current papers. The papers fill in a large circular space with highly granulated clusters of papers (left pane). A drill-down of one small section of the map (right pane) shows that the papers form natural clusters of different size. The clusters are of reasonable size; the average cluster is approximately twice as large as those reported by ISI and CRP and there are no super clusters as reported by ISI and CRP. Maps generated from reference papers, and from a modified cosine measure are qualitatively similar to those shown in Figure 1.

#### *Structural Integrity of Disciplines*

One way to assess structural integrity is to determine if papers in a discipline tend to be grouped in the same general area. Each of the papers was assigned to ISI's (227 categories) disciplinary classification schema. We could then determine if (a) the papers in a discipline tend to cluster together and (b) whether the relative position of these disciplinary clusters provide additional insights about the structure of science.

Figure 2 represents the spatial location of a sample of papers in six disciplines. The map from Figure 1 was used. The background in Figure 2 represents the total number of papers available and corresponds to the left side of Figure 1 (the granulated circle). The circles in the foreground of Figure 2 represent six disciplines (out of 226 potential disciplines) that were selected to illustrate how structural integrity can be assessed.

A random sample of 100 papers from each discipline was plotted in Figure 2. The actual sizes of these six disciplines range from a low of 1400 papers (Social Psychology) to a high of 20099 papers (Condensed Matter Physics). These papers were represented by a lower case letter if they were outside the 50% domain of their discipline, and were represented by a very small dot if they were within the 50% domain of their discipline.



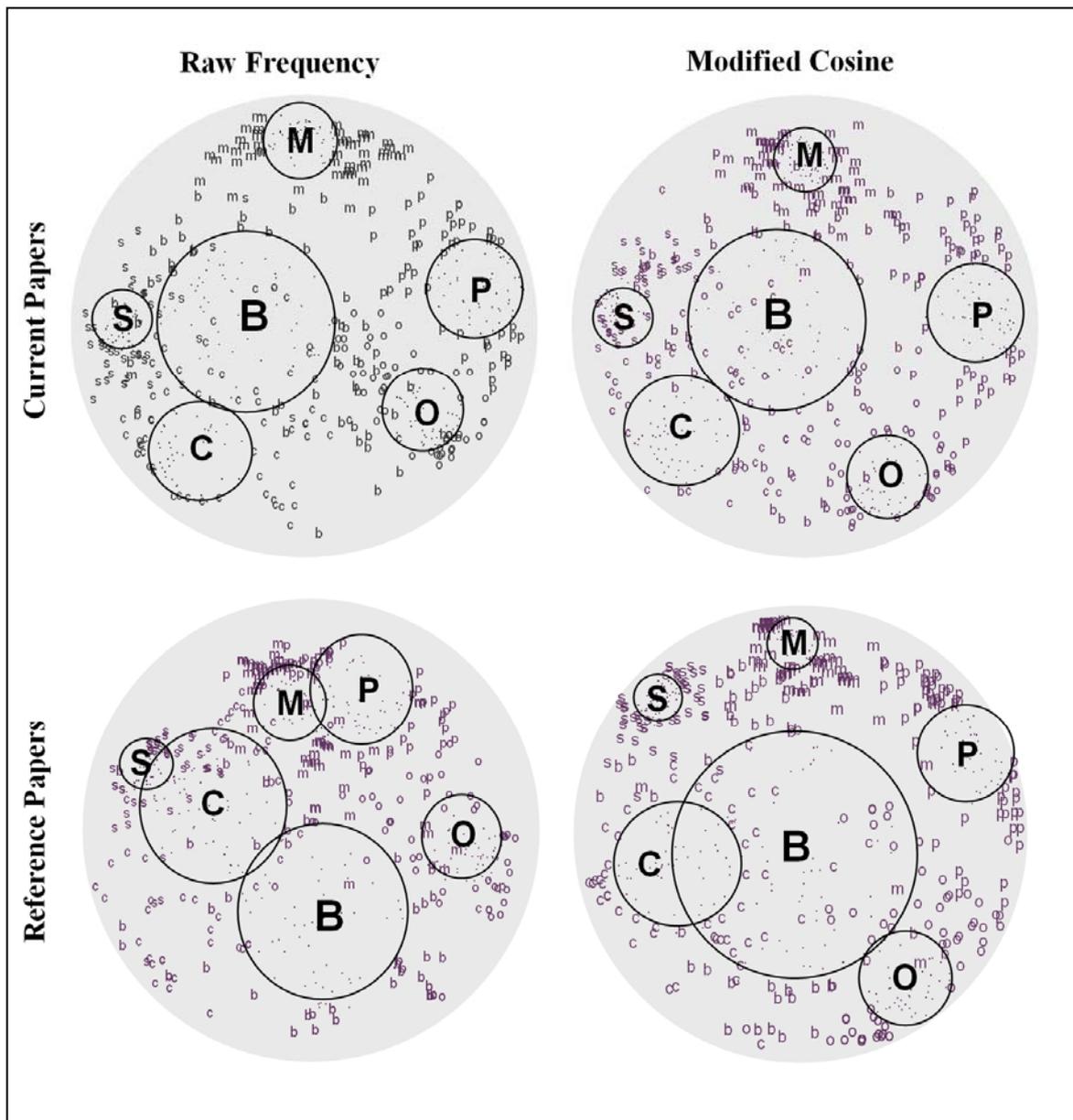
**Figure 2: Maps of current science from ISI 2002 data using raw frequency. Circles represent the 50% domain of six disciplines: Mathematics (M); Condensed Matter Physics (P); Organic Chemistry (O); Cardiovascular Research; (C); Social Psychology (S) and Biology (B).**

The location of these six disciplines reflects an underlying structure of science that appears in all of the maps we will explore. We rotated and flipped the maps so that mathematics (M) would be at the top of the circle and condensed matter physics (P) would be to the right of mathematics. Note that the papers in mathematics (the dots within the circle) tend to form a tighter circle than the papers in physics, and that there are no papers from the other five disciplines that are within these circle. The disciplines that are between mathematics (not shown) include many of the engineering and material science disciplines as well as other physics-related disciplines and some chemistry-related disciplines.

As one moves clockwise in Figure 2, the next disciplines to appear are organic chemistry (O), cardiovascular research (C) and social psychology (S). Biochemical disciplines (not shown) are generally between organic chemistry and cardiovascular research. Most of the biomedical research disciplines are in the third quadrant of this circle (between 6 o'clock and 9 o'clock). The social sciences are in the fourth quadrant of the circle (between 9 o'clock and 12 o'clock). One will find neurology, psychiatry and psychology at the border between the

biomedical sciences and the social sciences. One will find operations research and computer sciences at the boundary between the social sciences and math.

The sixth discipline, biology (B), is more diffuse and has a more central location on the map. This is almost tautological. A discipline whose papers are spread across the map will have an average position closer to the center. The circle representing 50% of the papers in the discipline would be larger. At this point, it is not possible to determine if the relatively large circles (in disciplines such as biology) are a reflection of reality (the papers are more likely to be linked to papers in many disciplines) or a sign of low structural integrity in this particular map (the circle should be smaller because the papers are more likely to link to other papers in that discipline).

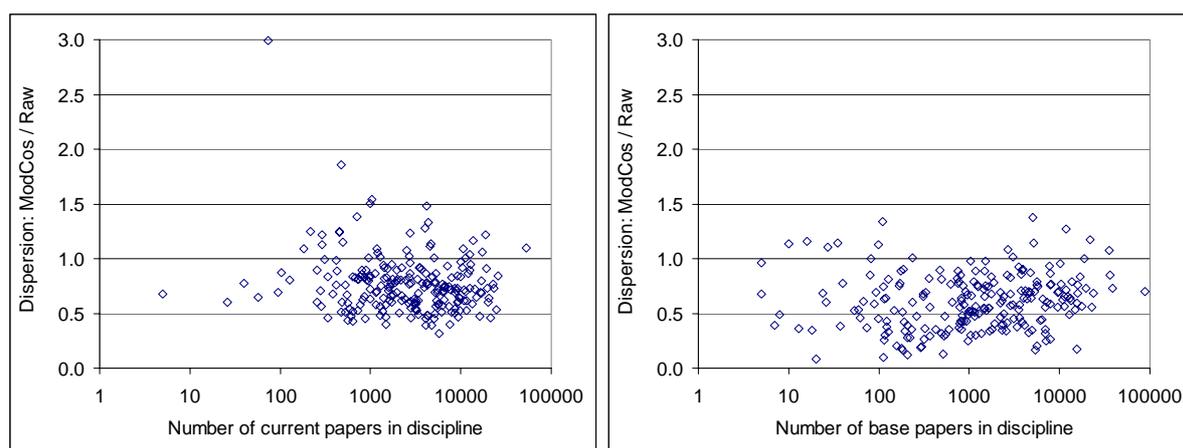


**Figure 3: Maps of current science and base science from ISI 2002 data using raw frequency and modified cosine. Circles represent the 50% domain of six disciplines: Mathematics (M); Condensed Matter Physics (P); Organic Chemistry (O); Cardiovascular Research; (C); Social Psychology (S) and Biology (B).**

Figure 3 compares the relative size and location of the six disciplines in four maps: those based on current papers vs. reference papers and those using the raw frequency vs. modified cosine measures of relatedness. Three of the maps provide similar structural pictures. There is reasonable spacing between math, physics, organic chemistry and social psychology. Biology is larger and in the middle. The major exception is the base science map using the raw frequency measure. In this map, math and condensed matter physics are relatively close together. Cardiovascular research is closer to social psychology and mathematics. Biology is further down and to the right.

Figure 3 also suggests that the circles (the 50% domains of the disciplines) are of relatively constant size when one compares both maps of current science (the top two maps), but quite different when one looks at the maps of base science (the bottom two maps). The base map using raw frequency has a smaller circle for biology but a larger circle for math. We looked therefore looked at the dispersion of papers assigned to 227 disciplines (using ISI's journal-discipline classification schema). The map that does a better job of grouping the papers (by discipline) is considered superior in terms of structural integrity. This allows us to test whether the structural integrity of disciplines is greater in the modified cosine map or raw frequency map.

Dispersion was used as a measure of the structural integrity of each discipline. We define dispersion as the average squared distance of each paper in a discipline from the centroid of the discipline. Ratios of dispersions for the current and base maps by discipline are shown in Figure 4. For current papers, the majority of disciplines (86%) have lower dispersion (and thus are more concentrated) in the modified cosine map than in the raw frequency map. The same is true for base papers, where 93% of disciplines are more concentrated in the base modified cosine map than in the base raw frequency map. The larger percentage of tighter disciplines in the modified cosine maps suggests that this measure generates maps with greater structural integrity. This is not surprising, given the recent finding that this is a more accurate measure of relatedness.

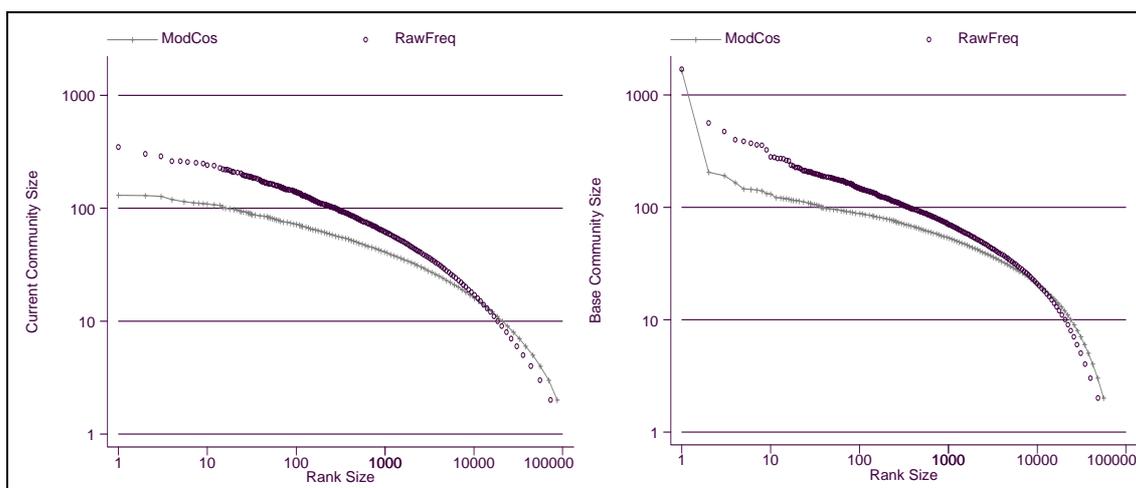


**Figure 4: Ratios (ModCos / Raw) of dispersions for (a) current and (b) base disciplines.**

#### *Structural Integrity of Research Communities*

A second way to measure structural integrity is to focus on clusters of papers within the map. Clusters, or research communities, were identified for each map using a modified single-link clustering algorithm. The maps of current papers had 96,500 and 84,619 communities, respectively, for the modified cosine and raw frequency maps, while the maps of base papers

had 60,773 and 54,115 communities, respectively, for the modified cosine and raw frequency maps. Distributions by size of community are shown in Figure 5.



**Figure 5: Research community size distributions for (a) current and (b) base maps.**

Of particular note is the problem of very large clusters. CRP and ISI felt that very large clusters were problematic. ISI dealt with this problem by arbitrarily saying that any cluster over a certain size should be broken up. CRP dealt with this problem by using a stratified threshold and clustering method.

Extremely large clusters are indicative of problems of structural integrity. This suggests that a map with no (or few) extremely large clusters should have better structural integrity than a map with many. For the 100 largest clusters, the modified cosine maps (both current and base) have smaller clusters than the raw frequency maps (see Figure 5). This is one indication that the modified cosine maps may have better structural integrity.

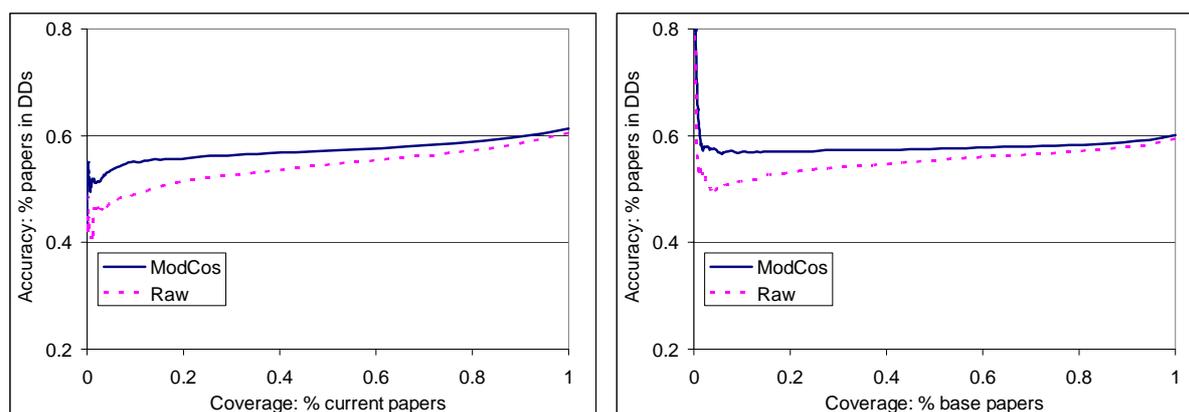
The existence of extremely large clusters only appears in the base maps. The modified cosine map only has one extremely large cluster (1645 reference papers), and the next largest cluster has 206 reference papers. The raw frequency map also has an extremely large cluster (1701 reference papers) and twenty seven clusters that are larger than 206 (the second largest cluster in the modified cosine base map).

Overall, the modified cosine map has fewer extremely large clusters. But it is important to remember that one has to drill down and see whether the content within these clusters of papers 'makes sense'. To measure this, we make the assumption that a community should be dominated by a single discipline. The map that does a better job of having the larger communities dominated by a discipline is considered superior.

The number of papers in a research community belonging to the dominant discipline (DD) for that community was thus used as a measure of structural integrity at the research community level of analysis. The DD for each research community was identified along with the number of papers belonging to that DD. Contributions from the DDs were summed, and the cumulative number of papers associated with DDs divided by the total number of papers in the associated research communities was used as a proposed measure of structural integrity of research communities. Summing for this measure went from the largest to the smallest research communities in order, since we are more concerned with large communities than

extremely small ones. Communities with only 2 or 3 papers have less meaning than those with more papers.

Figure 6 shows the relative structural integrities of the two measures for the current and base maps at the research community level. Although the final values (at 100% coverage) are close, the modified cosine measure maintains disciplines within research communities better than the raw frequency measure over the entire range of community sizes for both the current and base maps.



**Figure 6: Cumulative fractions of papers in dominant disciplines (DD) for (a) current and (b) base maps.**

While Figure 6 shows cumulative fractions, Table 1 gives fractions for different groups of cluster sizes. This shows that while the modified cosine measure is more accurate at large community sizes, the raw frequency is actually more accurate than the cosine for smaller communities. Thus the narrowing in the distances between the modified cosine and raw frequency curves with greater coverage in Figure 6.

**Table 1: Structural Integrity of Research Communities**

Cluster size	Fraction of papers in dominant disciplines (DD)					
	100+	51-100	26-50	11-25	5-10	2-4
Base- Raw Freq	0.504	0.550	0.574	0.599	0.650	0.745
Base- ModCos	0.569	0.562	0.576	0.590	0.631	0.731
Current- Raw Freq	0.484	0.524	0.554	0.591	0.642	0.743
Current- ModCos	0.531	0.530	0.564	0.582	0.627	0.722

### Summary

Table 2 summarizes the improvements reported in this paper. By utilizing hardware, software and intellectual development during the past 20 years, it is now possible to (a) separate two phenomena (base research communities vs. current research communities), (b) overcome some of the methodological problems implicit in identifying communities within a network, and (c) illustrate how the structural integrity of a map can be assessed.

Table 2 lists the fundamental differences between maps based on two perspectives. The Base Map was generated from clustering the reference papers and then assigning current papers. The Current Map was generated from clustering the current papers and then assigning

reference papers. Both approaches take one year of data, and identify clusters that include both current papers and reference papers. The difference is in how the clusters are generated. We suggest that current research communities capture the themes that researchers are working on. We suggest that base research communities focus on the tools/techniques that researchers are building on. They represent a different way to understand the structure of research, and may have different policy implications. This distinction is reflected in the fact that there are fewer base research communities than current research communities, despite the fact that the input files are quite similar (718,964 papers for the base map; 731,289 papers for the current map).

**Table 2. Improvements in Mapping Science**

	<b>ISI 1983</b>	<b>CRP 1984</b>	<b>This Paper 2002 Mod Cosine</b>
Visualization (at paper level)	No	No	Yes
<b>BASE MODEL/MAP</b>			
# Reference papers	5,239,536	5,150,772	10,911,939
# Reference papers in model	50,994	128,238	718,964
% Reference papers in model	0.97%	2.49%	6.65%
# Base Research Communities	9,420	28,128	60,773
Average #papers/community	5.4	4.6	11.8
# Current papers assigned	303,225	315,567	691,673
% Current papers assigned	44.0%	46.6%	64.7%
<b>CURRENT MODEL/MAP</b>			
# Current papers	688,678	677,011	1,069,764
# Current papers in model	No current model	No current model	731,289
% Current papers in model			68.4%
# Current Research Communities			96,500
Average #papers/community			8.2
# Reference papers assigned			10,080,472
% Reference papers assigned			92.4%

One of the methodological problems we have addressed in this paper is coverage – the ability to cover more of the literature using advancements in measurement and clustering. The base maps generated by this method cover a much higher absolute and relative number of available documents. For instance, we were able to cluster 6.65% of the reference papers. This is significantly higher than the percentages from co-citation methods developed by ISI and CRP (Franklin & Johnston, 1988, p. 370), and could be easily doubled simply by relaxing one of our thresholds.

The third contribution of this paper deals with how one can assess whether a map is more accurate. This is critical to technological progress in this field. We need methods for comparing different maps – not just in terms of what insights they may have – but whether one is truly more accurate than the other.

We find that maps based on a modified cosine measure have greater structural integrity than those generated from a raw frequency measure at both the discipline and research community

levels of analysis. The perceived problem of extremely large heterogeneous clusters that plagued ISI and CRP was far less of a problem using the modified cosine measure. There was only one 'super-group' of papers in the modified cosine base map, and the larger clusters were dominated by a single discipline.

Another method for measuring structural integrity, not reported in this paper, is to focus on the individual papers- and determine if the neighboring papers are on the same topic or not. We are presently investigating this possibility by using keywords. But whether one focuses on structural integrity from a disciplinary, community, or paper level of analysis, we strongly recommend that future work focus more heavily on quantitative techniques for measuring the accuracy of any map of science.

## References

- Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37, 179-255.
- Boyack, K. W., Klavans, R., & Börner, K. (in press). Mapping the backbone of science. *Scientometrics*.
- Franklin, J. J., & Johnston, R. (1988). Co-citation bibliometric modeling as a tool for S&T policy and R&D management: Issues, applications, and developments. In A. F. J. van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 325-389). North-Holland: Elsevier Science Publishers, B.V.
- Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6), 420-442.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10-25.
- Klavans, R., & Boyack, K. W. (in press). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*.
- Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (2nd ed.). Chicago: University of Chicago Press.
- Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology*, 54(5), 413-422.
- Norling, P. M., Herring, J. P., Rosenkrans Jr., W. A., Stellpflug, M., & Kaufman, S. B. (2000). Putting competitive technology intelligence to work. *Research-Technology Management*, 43(5), 23-28.
- Price, D. D. (1986). *Little Science, Big Science and Beyond*. New York: Columbia University Press.
- Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9), 799-813.
- Small, H. (2003). Paradigms, citations, and maps of science: A personal history. *Journal of the American Society for Information Science and Technology*, 54(5), 394-399.
- Small, H., & Griffith, B. C. (1974). The structure of scientific literatures, I: Identifying and graphing specialties. *Social Studies of Science*, 4, 17-40.
- Small, H., & Sweeney, E. (1985). Clustering the Science Citation Index using co-citations. I. A comparison of methods. *Scientometrics*, 7, 321-340.
- Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the Science Citation Index using co-citations. II. Mapping science. *Scientometrics*, 8, 321-340.