Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 68, No. 3 (2006) 475–499

Quantitative evaluation of large maps of science

RICHARD KLAVANS,[a] KEVIN W. BOYACK[b]

[a] *SciTech Strategies, Inc., Berwyn, PA (USA)*
[b] *Sandia National Laboratories, Albuquerque, NM (USA)*

This article describes recent improvements in mapping the world-wide scientific literature. Existing research is extended in three ways. First, a method for generating maps directly from the data on the relationships between hundreds of thousands of documents is presented. Second, quantitative techniques for evaluating these large maps of science are introduced. Third, these techniques are applied to data in order to evaluate eight different maps. The analyses suggest that accuracy can be increased by using a modified cosine measure of relatedness. Disciplinary bias can be significantly reduced and accuracy can be further increased by using much lower threshold levels. In short, much larger samples of papers can and should be used to generate more accurate maps of science.

Introduction

Henry Small is the leader in generating macro-level maps of science from the scientific literature.[1–3] He and his colleagues have generated maps of science that are exemplars in the field. Each map is based on a very small sample of reference papers (the most highly cited references). These elite papers cover research in all areas of the natural sciences[1] or all areas in the natural and social sciences.[2,3] In all cases, maps were created by clustering these papers using co-citation analysis and then plotting the relationships between the clusters. These maps also represent methodological progress.

Received December 20, 2005

Address for correspondence:
RICHARD KLAVANS
SciTech Strategies, Inc., Berwyn, PA 19312, USA
E-mail: rklavans@mapofscience.com

0138–9130/US $ 20.00
Copyright © 2006 Akadémiai Kiadó, Budapest
All rights reserved

Better sampling techniques were used to reduce disciplinary bias. Clustering algorithms were modified in order to limit the size of a cluster of documents, and an iterative clustering approach was introduced so that existing visualization algorithms could be utilized.

There have not been any publications, however, that provide quantitative techniques for evaluating these maps of science. For example, there are no publications that show how a series of maps, using different methodological choices, could be generated and evaluated to determine which methodological choice results in a more accurate map. We have pursued this line of inquiry by focusing on journal-level maps (the relationship between scientific journals). Our first study looked at different measures of journal-journal relatedness to determine which measure generates the most accurate cluster solutions.[4] A second study focused on additional measures of journal-journal relatedness, developed a method for measuring local accuracy in a map (whether your neighbor is who you expect to see) and then used these methods to determine which measures of relatedness were better for different sample sizes.[5]

This paper builds on this avenue of inquiry, but shifts the level of analysis from scientific journals to scientific papers, expanding on our paper at ISSI 2005.[6] Paper-level maps are far more expensive and difficult to create. In the 1980's, one reviewer estimated that it cost $70,000US simply to generate cluster solutions from one year of data.[7] Days were required to generate these cluster solutions, weeks if one wanted to have a more representative sample of data. One can only be in awe of the resources that were invested by Small and his colleagues to generate the few maps that are available in the literature.

Things are different today. Many institutions now have electronic access to large amounts of bibliographic data as well as the computing resources to make sense of it. Software is now available to generate maps directly from a matrix of paper-paper relatedness measures. This combination enables us to generate large-scale maps of science more efficiently, and in a much more cost-effective manner. This capability allows us to generate maps with different measures and methods, which leads to the question posed in this paper – how do we determine which maps are better?

We present a relatively fast and inexpensive procedure for generating paper-level maps of science. This procedure uses a graph layout algorithm that can generate maps of science directly from a matrix of paper-paper relatedness measures. The procedure is used to generate eight paper-level maps of science using two different sampling strategies, two different measures of relatedness and two different approaches towards visualization.

This paper is organized into five sections. The first section presents two exemplar maps generated by Small and his colleagues (the earliest and the most recent). These maps illustrate how the field has progressed over the past 25 years, but also illustrate the problem addressed in this paper – how does one quantitatively evaluate and compare

maps of science? The second section describes our methodology for generating maps of science and the methodological choices that will result in eight different maps of science. The third section outlines how these maps will be evaluated with a primary focus on the tradeoffs between accuracy, coverage and disciplinary bias. The last two sections present the results of the evaluations, along with a discussion of the implications of this research.

Exemplar maps of world-wide science

This study focuses exclusively on maps of science that have been generated from a sample of documents covering many disciplinary fields. We will not attempt to review the many mapping efforts that focus on a specific field or area of science.[8–11] Nor will we review parallel efforts to map broad technical fields or the world-wide web. We limit our focus to the depiction of the structure among and between a diverse set of scientific papers in the attempt to map all of science.

Historically, there has been only one source capable of generating these maps – the combined science and social science databases provided by the Institute of Scientific Information (ISI), now a division of Thomson Scientific. Henry Small, chief scientist at ISI, has devoted a large portion of his career to generating and publishing macro-level maps of science from the ISI database. Two examples are provided here: the earliest map[1] and the latest map.[3] These maps illustrate many of the accomplishments made over the past 25 years and highlight some of the issues associated with generating large maps of science.

Figure 1 is a map of world wide science generated by Griffith et al.[1] They started out with 867,600 references (the number of unique references in ISI's natural sciences database during the first quarter of 1972). They limited their analyses to the 1,832 references that were most highly cited (reference papers had to have been cited 10 or more times in the first quarter of 1972 in order to be included). Three different thresholds were then used for clustering. The map shown in Figure 1 is based on a cluster threshold of 3 (a reference had to co-occur 3 or more times with another reference before it could be included in a cluster). This resulted in 115 clusters that covered 1,310 documents (0.15% of the available reference documents and 71.5% of the highly cited reference papers). The 41 largest clusters (clusters with at least 3 papers) were then selected and graphed so that nodes that are closer are more related.

The map in Figure 1 provided a first view of how science might be structured. The four largest nodes are biomedicine (the "hub" node on the left side of the map), chemistry (the "hub" node on the far right) and nuclear and particle physics (nodes at the upper middle of the map). Two of these nodes have the most number of links to other nodes (biomedicine on the left and chemistry on the right). In this map, physics does not have as many links to other areas of science as do the hub nodes.

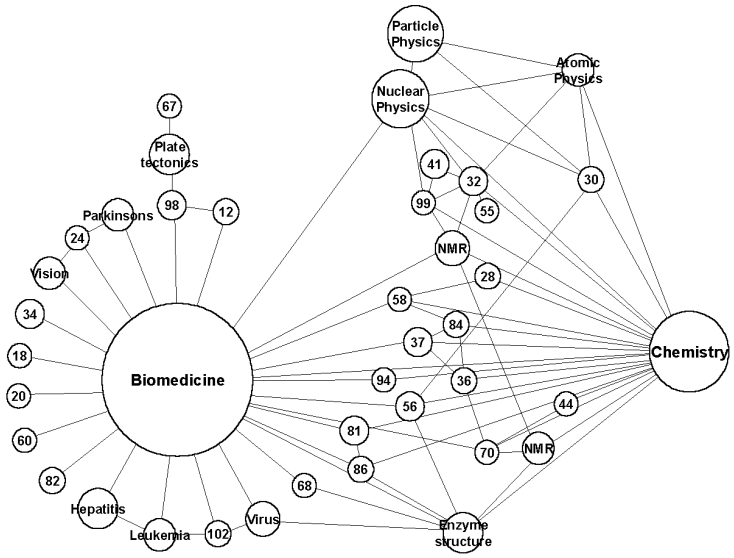


Figure 1. Redrawing of the map of science from Griffith et al.[1] Node sizes are scaled to the cubed root of the number of papers. Labels have been added for the largest nodes. Refer to the original paper for labeling for numbered nodes

The map in Figure 1 also provided some intriguing observations about the network relationships around major nodes. The nodes connected to biomedicine are like spokes around a wheel. They don't tend to be interconnected like the nodes connected to chemistry (on the right). But it's unclear whether these observations are a reflection of the structure of science or an artifact of the methodological choices (sampling threshold, measures of relatedness, clustering algorithm and visualization algorithm). Changes in any of these methodological choices may have generated a completely different map of science.

Small, Sweeney & Greenlee[2] revised their mapping technique in order to overcome three major problems: disciplinary bias, over-aggregation of documents, and limitations in visualization software. Disciplinary bias was a consequence of the sampling procedure. The sampling approach in the 1974 map used simple citation counts (selecting the references that were cited the most). References in biomedical papers were over-represented because biomedical papers tend to have more citations. References in mathematics were under-represented for the opposite reason (fewer citations per paper and relatively low citation counts). Small, Sweeney & Greenlee proposed using fractional citation counts as a way to reduce disciplinary bias. Fractional citation counts involved weighting each reference according to the fraction of references in the citing paper (references would each be weighted with the value of 1/4

if there were four references in a paper). The fractional values were then summed and reference papers with the highest counts were selected.

Over-aggregation (the existence of extremely large clusters of documents) was another problem addressed by Small, Sweeney & Greenlee.[2] There seemed to be general consensus that document clusters should not exceed around 50 to 60 papers.[7] The clustering algorithm that had been used (single link clustering) resulted in two nodes in Figure 1 that significantly exceeded this threshold (the biomedical node had 801 papers and the chemistry node had 92 papers). They proposed using a single link clustering algorithm that would impose a threshold of 50 papers on cluster size.

The third major problem that Small, Sweeney & Greenlee[2] addressed was limitations imposed by visualization software. The visualization software that was available in the 1970's and 1980's could not handle more than 100 nodes. But the methodological improvements mentioned above (lower disciplinary bias and limitations on cluster size) generated thousands of clusters. They proposed clustering the clusters in order to reduce the number of nodes. Four clustering iterations (clusters of clusters of clusters of clusters of documents) were needed in order to generate the number of nodes that could be visualized with existing software. These methodological improvements were used to generate a map in Small's 1985 and 1999 papers. Figure 2 represents the 1999 map using these new techniques.

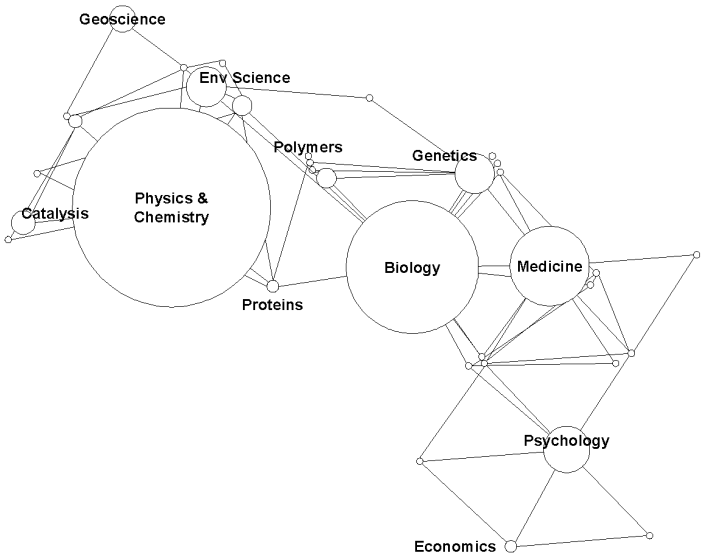


Figure 2. Redrawing of the map of science from Small[3] (Figure 3). Only the 35 fourth-level clusters are shown here. Third-level structure within clusters has been removed. The large clusters have been shrunk slightly to avoid covering smaller clusters

This map was based on an initial set of approximately 6,100,000 references (the number of unique references in all papers published in 1995, but limited to references that were published between 1981 and 1995). Their analysis focused on a sample of 164,612 references (based on the sequential application of full and fractional citation thresholds). Iterative clustering was used to generate 18,939 clusters (level 1), 2,402 clusters (level 2), 327 clusters (level 3) and finally the 35 clusters that are represented in Figure 2. Documents and clusters were dropped at each level. The first level of clustering dropped 75.3% of the sample documents. The final set of 35 clusters consisted of 36,720 documents (dropping an average of 36.3% of the documents over each iterative clustering step).

While the scope of this map is greater (it includes the social sciences), the overall number of nodes is slightly less than the number of nodes in the early map of Figure 1 (35 vs. 41). The most dramatic difference between these two maps, however, is the lack of similarity in structure. Figure 1 suggests that certain nodes (chemistry and medical science) are central to the communication patterns in their local area while physics is embedded in a more sparsely connected network. Figure 2 suggests that physics and chemistry are highly interconnected and medical science does not play a central role. Is this an artifact of the methodological choices, a more accurate reflection of how science is structured, or a reflection of change in structure? There is no way to answer the question based on what has been published.

The methodological improvements in Figure 2 had confounding effects on disciplinary bias. The use of fractional citation thresholds did have the desired result of reducing disciplinary bias (suppressing the number of biomedical references and increasing the number of physics references. But the use of iterative clustering (and the exclusion of documents during each clustering iteration) increased disciplinary bias.[3] Physics was now highly over-represented and clinical medicine was highly under-represented. Correspondingly, physics seemed to play a more central role and clinical medicine has a less central position in Figure 2.

These two maps help illustrate the major issues addressed in this paper. It's unclear how one might quantitatively evaluate the two maps presented above. Specifically, how might one measure the accuracy of a map? How much would accuracy deteriorate if one increased the number of documents (by lowering the sampling threshold and/or reducing the number of documents that are excluded during clustering)? Which map has lower disciplinary bias? What effect would disciplinary bias have on the accuracy of the map? What is the relationship between coverage (the number of documents), disciplinary bias and accuracy? Pursuing this line of research would help in the selection of measures and thresholds that generate more accurate maps of science. Pursuing this line of research would also generate more reliable insights into how science is structured and how it has evolved over time.

Generating maps of science

In generating our maps of science, we explore some of the "methodological space" inherent in that broad fundamental approach suggested by Small and his colleagues. First, we start from the same choice of database and focus on one year of data. We then use two different criteria for generating a sample of papers for further analysis. Two different document-document relatedness measures are then calculated for each sample. And finally, maps generated directly from paper-paper relatedness statistics are compared to maps based on document clusters.

Data

We have used the combined ISI databases from a single file year in this study (2002 SCIE and SSCI databases). There were 1,069,764 current records with 10,911,939 unique reference papers in the dataset. Of the over 1 million current records, we limit the number of "mappable" records to the 833,307 that are bibliographically coupled to another current record within the set. This excludes the majority of editorials, book reviews, and similar items that are indexed by ISI, and limits the dataset primarily to records containing publishable technical advances. This parallels the 10,911,039 reference papers used in the set in that, since they were referenced, they can be considered to be publishable technical advances that were worth referring to.

Two sampling approaches

Small[3] describes two major approaches to sampling a publication database. One could focus on the current papers or on the references. He and others have focused exclusively on the references in generating macro-level maps of science using co-citation techniques. We will generate two sets of maps: one based on an analysis of the relationship between current papers and another based on the analysis of the relationship between references.

We expect these maps to provide qualitatively different insights into the structure of science. An analysis of the relationship between current papers is expected to generate a thematic map. Current papers that are close to each other in a map are assumed to have the same theme. Reference papers that are close to each other in a map are assumed to deal with a related concept or exemplar. While these maps may be similar in structure, one can expect that the relationships between major scientific areas (math, physics, chemistry biology and medicine) may differ substantially.

Minimal thresholds were used to reduce the number of documents for further analysis. Different thresholds were used for current and reference papers. We excluded all current papers that had a maximum of 1 co-occurrence with any other documents in

their set, and those reference papers that had a maximum of 3 or fewer co-occurrences with any other documents in the set. Measures of paper-paper relatedness are relatively meaningless for the excluded sets of papers. This reduced the number of documents to 718,964 reference papers and 731,289 current papers.

Two measures of relatedness

We use co-occurrence frequencies to generate two different measures of paper-paper relatedness. Co-occurrence frequency between reference papers (e.g. co-citation analysis) is the approach first proposed by Small, and is the standard that was widely used in the field through the 1980s. Co-occurrence frequencies between current papers (e.g. bibliographic coupling) has also been proposed as a useful measure,[3,12] but has not been used. There has been no attempt to map the relationship between a large diverse set of current papers.

Alternative measures of relatedness have been proposed in the literature[13] to overcome the biases associated with using co-occurrence frequencies. In this study, we use K50, a modified cosine index (a cosine adjusted for expected co-occurrence frequencies). Previous work has suggested that this is the most accurate measure for co-citation analysis.[5] A raw frequency count and the modified cosine have been generated for each of the two data samples mentioned above.

Visualization approaches with and without document aggregation

Some of the newer visualization (or graph layout) algorithms are not limited to mere hundreds of nodes, but can handle hundreds of thousands, or even millions of objects.[14,15] Thus, the visualization routine no longer need be a bottleneck when handling very large document sets, and it is no longer required to cluster documents in order to visualize the relationship between large numbers of documents. In addition, these tools often have zoom and pan capabilities to allow exploration of very large data sets.

There are, however, theoretical and practical reasons to cluster the documents and then show the relationships between these clusters. Theoretically, some believe that the fundamental unit of analysis should be a research community rather than a document.[16] Since document clusters are one way to characterize these research communities, one should map the document clusters in order to gain insights into the fundamental structure of science. On the practical side, there is an upper limit on the number of objects that can be deciphered on the printed page. In addition, the visualization program we use, VxOrd, does not take into account groups of documents when it tries to locate documents in two-dimensional space. We were concerned that the

visualization algorithm might tend to optimize local conditions (improving the location of documents) at the potential expense of regional accuracy (improving the location of groups of documents).

We therefore visualize the relationship between documents in two ways. In both cases, we use a visualization algorithm, VxOrd,[15] that is capable of handling the relationship between hundreds of thousands of documents based on document-document relatedness measures. For the first approach we use VxOrd in native mode, which does not aggregate documents. For the second approach we change a parameter in VxOrd that enables natural aggregation of documents. The first approach tends to create a visually homogenous layout at local levels, while the second approach creates visually separated groups of documents. Subsequent clustering is performed on each map. A more detailed description of the visualization algorithm is provided in Appendix A. A description of the clustering algorithm is provided in Appendix B.

Evaluating maps of science

Figure 3 is an example of what a map of science with ~700,000 nodes looks like. Only the nodes are shown in this figure. VxOrd also provides information on the links between these nodes, but including this data in Figure 3 would hinder our ability to see the spatial relationship between nodes.

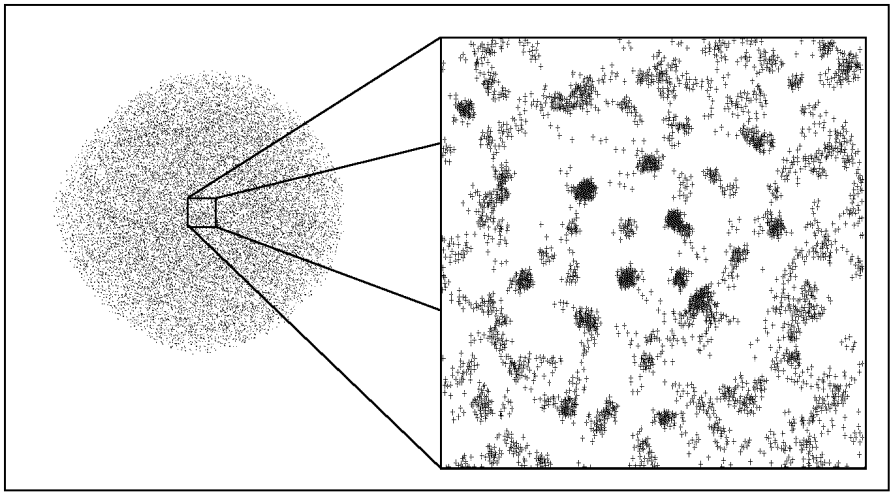


Figure 3. Maps of current papers using raw frequency co-occurrence data. VxOrd was run in "aggregation" mode to generate a map of over 700,000 papers (left). An 8x enlargement of the central section of the map (right) shows single papers more clearly

This figure illustrates one of the unique problems that one faces in evaluating and comparing large-scale maps. One cannot rely on simple visual inspections. Quantitative techniques are required to evaluate and compare these maps.

The following methods are proposed to evaluate each map of science. First, we evaluate the local accuracy of each measure of relatedness (the tendency for papers that are close to each other to be in the same discipline). Second, we evaluate the regional accuracy of each measure of relatedness (the tendency for papers in the same discipline to be located near each other on a map). Third, we evaluate disciplinary bias (the tendency for a sample of document to over/under represent different disciplines). And lastly, we evaluate the coherence of all clusters with 50 or more papers.

Local accuracy

Local accuracy is defined as the tendency for a pair of papers to belong to the same ISI disciplinary category. The measure of accuracy used here is based on Klavans & Boyack[5] in their analysis of the accuracy of different journal-journal relatedness measures. The following method is used to evaluate local accuracy of a map. First, the distance between linked papers in a graph is calculated. Second, each pair of papers is coded as "1" if both papers are in the same ISI category and "0" if not (the 24 ISI categories are used in this analysis). Journals that are coded in the multidisciplinary category are treated differently. The pair of papers is coded as ".5" if either paper is in the multidisciplinary category, because multidisciplinary papers can, by definition, link to so many other disciplines.

All pairs of linked papers are ordered by ascending distance. One expects that the first set of records (the most closely spaced pair) will be on the same topic, be assigned to the same ISI disciplinary category, and therefore be coded as a "1" (most accurate). The cumulative average is then calculated to determine the drop in accuracy as one includes all pairs of papers.

There are hundreds of millions of pairs of linked papers. We proceed through the list of pairs (from most related to least related) until 'n' unique papers are covered. For example, the first pair covers two papers. The next pair might cover two additional papers. We make note of the local accuracy (cumulative sum of the coding numbers divided by number of pairs) as we cover more and more of the papers until 100% of the papers are covered. This might require the review of only eight million pairs or as many as sixty million pairs. The results tell us the overall accuracy as one increases the number of papers from 2 (the initial case) to 100% of the sample (approximately 700,000 papers).

Accuracy vs. coverage curves are generated for each map (two samples, two measures of relatedness for each sample and two visualization layouts), where the x-axis shows cumulative coverage, and the y axis is the cumulative average accuracy.

Based on similar calculations with journals,[5] we expected that the curves would start high and slope down. Average accuracy should decline as one includes papers that have lower levels of relatedness to their nearest linked neighbors.

Regional accuracy

Regional accuracy is defined as the tendency for papers associated with a discipline to be close to each other on a map of science. Previous maps of science have tended to show that highly aggregated clusters have a strong disciplinary orientation.[4] Therefore, one would expect that papers in the same discipline should be located in the same general area on a map.

The same disciplinary classification system mentioned above is used. The journal affiliation of each paper allows one to assign ISI categories to each paper (multi assignments are allowed). The average position of each discipline is calculated. Dispersions (Euclidean distances between papers and the average position of the discipline) are used to measure regional accuracy. Low dispersions means that the map has higher regional accuracy.

Disciplinary bias

The following method is used to evaluate disciplinary bias. First, an ordered list is created (papers are ordered in terms of distance to nearest neighbor, in the fashion described earlier for local accuracy). Second, the disciplinary affiliation of the paper is coded. Fractional counts are generated if the paper is assigned to more than one discipline. Third, a cumulative disciplinary profile is calculated (the fraction of papers per discipline for all papers up to the nth related paper). Fourth, the squared difference between the disciplinary profile and the expected disciplinary profile is calculated. The expected disciplinary profile is based on the entire sample of current papers or references respectively. Disciplinary bias is 0 if the cumulative disciplinary profile (for a specific threshold level) equals the expected disciplinary profile.

Graphs showing disciplinary bias are then generated from all eight maps, with cumulative coverage on the x-axis and disciplinary bias on the y-axis. One would expect disciplinary bias to start relatively high and slope downward, dropping rapidly at first and then slowly as one covers all of the available literature.

Coherence of large clusters

As mentioned above, clustering (see Appendix B) is performed on each map. There is general consensus among those clustering the scientific literature that there is a maximum cluster size. Early work suggested a cluster size of around one hundred

papers.[7] Small imposes a more stringent size limit of fifty to sixty.[3] Clusters that were larger than this were considered problematic. We take a detailed look at all clusters with greater than fifty papers to examine their coherence.

Coherence is defined as the maximum percentage of papers in a cluster that are in the same discipline. A large cluster is more coherent if most of the papers are members of the same discipline. We expect that coherence is negatively correlated with cluster size (the larger clusters are, as assumed by prior researchers, to have lower coherence). We look at two criteria with regards to cluster sizes. First, the numbers of large clusters (> 50 documents) for each map is calculated. Second, we determine the average coherence of all large clusters.

Evaluation of eight maps of science

Local accuracy

The relative accuracies of the eight maps can be compared if one focuses on the trade-off between cumulative accuracy and coverage. "Accuracy vs. coverage" in this context is analogous to precision vs. recall in information science. The most accurate map is not necessarily the same for each level of coverage.

The curves in Figure 4 show that the modified cosine measure generates more accurate maps than the raw frequency measure. For current papers, the aggregated map based on the modified cosine measure (K50:aggregated) is superior at all levels of coverage. For reference papers, the K50:aggregated map has the highest accuracy from 40% to 87% coverage, and the K50:paper map has the highest accuracy from 87% to 100% coverage. The maps based on raw frequency measures are less accurate. We can also observe that the curves for the current paper maps are more spread out (greater variance). This suggests that the choice of measure is more important when generating maps of the current literature. The choice of measure has less of an impact on local accuracy when generating a map of reference papers.

The increase in accuracy over a broad range of sampling is counter to the expected pattern. The expected pattern (a drop in accuracy as one increases sample size) is implicit in early decisions to select a very small sample of papers and use very selective clustering thresholds. This curve suggests the opposite: one would have a more accurate map if the sample size is larger and the clustering threshold is more inclusive. This is especially true for generating maps of reference papers where local accuracy increases to 90% of the sample, and where the best map (K50:paper) could include 99% of the data with negligible impact on accuracy.

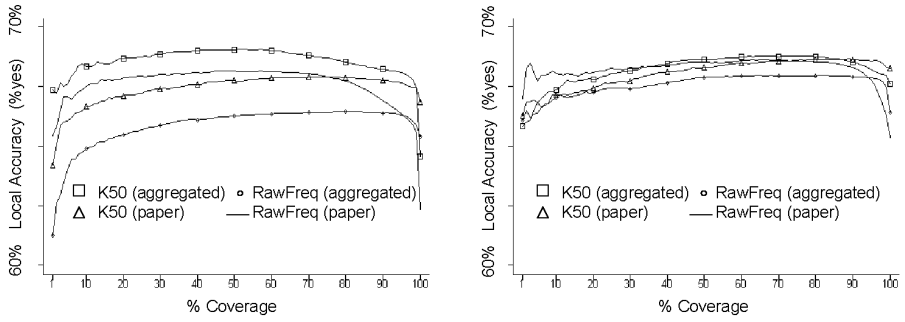


Figure 4. Accuracy vs. coverage for maps based on (a) current papers and (b) reference papers

Regional accuracy

The dispersion of papers by discipline is another way to evaluate the maps. Lower overall dispersions means that papers that are in the same discipline are closer to each other. The total dispersions for all four paper-level maps and the two best aggregated maps are listed in Table 1.

Table 1. Regional accuracy (cumulative dispersion by discipline) for different maps

	Current paper maps		Reference paper maps	
	Dispersion	Rank	Dispersion	Rank
K50:aggregated	555,197	2	469,212	1
K50:paper	516,412	1	550,976	2
RawFreq:paper	577,026	3	579,570	3

The maps based on the modified cosine measures have greater regional accuracy. As expected, the K50:aggregated map was better than the K50:paper map for reference papers (for the theoretical and practical reasons mentioned previously). But we were surprised that this relationship did not hold for the maps of current papers. The aggregated map of current papers has lower regional accuracy than the non-aggregated map of current papers. At this point, we can only guess at the reasons for this result. Perhaps themes (which correspond to clusters of current papers) are more multidisciplinary in nature than exemplars (which correspond to clusters of reference papers). Perhaps there are some disciplines that one would expect to be more dispersed as one shifts from a homogeneous document map to an aggregated document map.

In order to explore this further, we looked at the set of disciplines that had the unexpected increase in dispersion. Table 2 lists the 24 disciplines with their percent differences in dispersion (as a function of average dispersion) for the K50:aggregated map vs. the K50:paper map.

Table 2. Changes in dispersion for maps of current papers by discipline

	Decreases			Increases	
	Index *	ISI Discipline		Index *	ISI Discipline
1	–1.17	Space Science	10	0.06	Materials Science
2	–0.41	Economics & Business	11	0.07	Chemistry
3	–0.21	Psychology/Psychiatry	12	0.08	Education
4	–0.19	Computer Science	13	0.09	Clinical Medicine
5	–0.16	Mathematics	14	0.10	Multidisciplinary
6	–0.14	Physics	15	0.14	Biology & Biochemistry
7	–0.05	Engineering	16	0.15	Ecology & Environment
8	–0.03	Molecular Bio & Genetics	17	0.16	Immunology
9	–0.01	Geosciences	18	0.18	Neurosciences & Behavior
			19	0.18	Social Sciences, general
			20	0.28	Microbiology
			21	0.30	Pharmacology
			22	0.33	Plant & Animal Sciences
			23	0.36	Law
			24	0.61	Agricultural Sciences

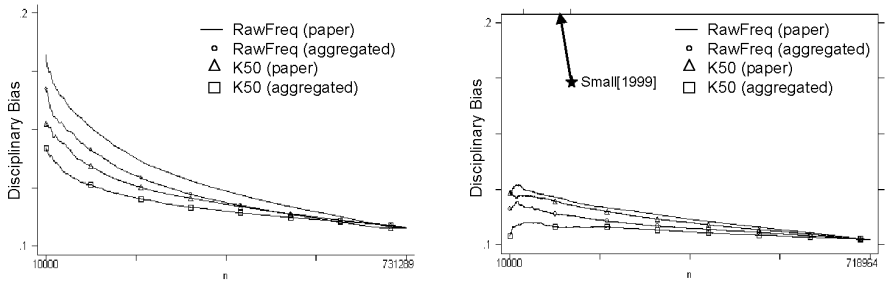
* (disp_k50:aggregated – disp_k50:paper/ [.5*(disp_k50:aggregated + disp_k50:paper)]

Only nine of the disciplines had the expected decrease in dispersion as one shifted from the K50:paper map to the K50:aggregated map. This included many of the engineering/math based disciplines (space, computer science, mathematics and physics) and a selected set of social sciences (economics & psychology). The unexpected increases in dispersions were in agricultural sciences, law, plant/animal sciences, pharmacology and microbiology. Many of these disciplines are associated with drug development. Development activities that build from engineering and math, which may be less multidisciplinary in nature, have the expected drop in dispersion.

Disciplinary bias

Disciplinary bias, or the tendency to over-represent some disciplines at the expense of others, may be another potential source of inaccuracy. We therefore measured the disciplinary bias as one increases document coverage. Figure 5 shows these disciplinary bias trends for the maps of current papers and reference papers. As expected, the disciplinary bias declines as one increases the sample size. In both cases, the

K50:aggregated maps are best (have the lowest disciplinary bias), even though they all converge at the right. Note that all curves converge at a value that is greater than zero because the expected disciplinary profiles were based on all papers, not just those included in the maps. For instance, the expected profile for current papers was based on the 833,307 papers that were bibliographically coupled, while only 731,289 are in the maps.



(a) (b)

Figure 5. Disciplinary bias for maps based on (a) current papers and (b) reference papers

It is important to note that most clustering algorithms, including our modified single-link algorithm, start at the left of the graph with the papers that have the highest relatedness. Therefore a map that has very high disciplinary bias at the beginning of the clustering process will be favoring the creation of clusters in certain disciplines. This may be more problematic in the generating of clusters of current papers where the disciplinary biases are much higher. This does not seem as serious of a problem in generating maps of the reference papers. In either case, the K50:aggregated map has the lowest disciplinary bias during this critical stage of cluster formation.

The disciplinary bias of Small,[3] calculated from values in his paper, is also shown in Figure 5b for comparison. His first level of clustering included 129,581 references, and had a disciplinary bias of 0.1745. The final level of clustering included 36,720 papers and had a disciplinary bias of 0.3899. Small's two sample sizes are indicated by hash marks at the top of Figure 5b. The scales used in this graph only allowed us to plot one of these points (indicated by a star). An arrow is used to indicate the direction of the other point. The disciplinary bias that resulted from clustering is quite severe. Dropping over 21.3% of the documents to create the first clusters, and then dropping 72% of the remaining documents by the time one reaches the fourth clustering iteration, results in an exceptionally biased set of papers.

Coherence of large clusters

Document clusters were identified for each of our eight maps using the modified single-link clustering algorithm detailed in Appendix B. The distributions of cluster size were plotted in order to evaluate the statistical characteristics of cluster size. Figure 6a shows the distributional characteristics for the current paper clusters. The non-aggregated maps (the two lines with steeper slopes) generate far more large clusters, with the largest clusters being quite large (4,403 and 10,000 papers for the RawFreq:paper and K50:paper maps respectively). The aggregated maps of current papers have much flatter curves. There are far fewer very large clusters, and the maximum cluster sizes are much lower (348 and 131 papers for the RawFreq:aggregated and K50:aggregated maps, respectively).

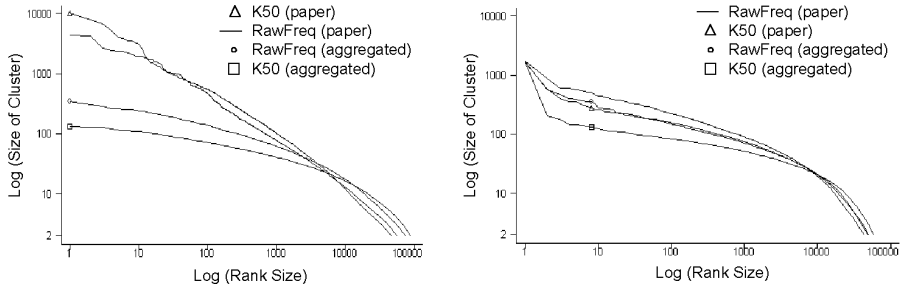


Figure 6. Cluster size distributions for maps based on (a) current papers and (b) reference papers

The difference between the lines in Figure 6a is mostly an artifact of the settings on the visualization algorithm that allow for edge-cutting. Many visualization programs maintain all of the edges and the corresponding maps do not show clearly defined cluster boundaries. By contrast, VxOrd allows cutting of edges after an initial solution is reached. Edge cutting has the effect of clustering the nodes. Minimum cutting (not used in this study) would generate a map with less agglomeration of nodes, and would correspondingly result in the identification of fewer, larger clusters. The default or native setting (used to generate the non-aggregated maps) results in some agglomeration (formation of strings of nodes and clusters of nodes). The maximum setting (used to generate the aggregated maps) will tend to break up strings of nodes into smaller clusters that have clearly defined boundaries. This effect is clearly seen in the distributions in Figure 6a.

The edge-cutting setting has a similar, but less dramatic, effect on the distribution of reference paper clusters (Figure 6b). The curve based on the K50:aggregated map is shallower than the curve based on the K50:paper map. The curve based on the RawFreq:aggregated map is shallower than the curve based on the RawFreq:paper map. But the curves all start with one large cluster that is almost the same size for all four maps (approximately 1,675 papers). This large cluster is dominated by papers in the Economics & Business discipline for all four maps. The K50:aggregated curve then drops down the fastest, and has the fewest large clusters.

The best curves (K50:aggregated) have large clusters that are within a reasonable range. Small suggested a maximum cluster size of 50 when he clustered about 130,000 reference papers.[3] We are clustering over five times more papers, and have a maximum cluster size of 131 (for the map of current papers) and only one cluster that exceeds that number for the map of reference papers.

The following analyses focus solely on those clusters with 50 or more papers. An analysis of variance was performed using the STATA statistical analysis package to evaluate the effect of different methodological choices (default vs. maximum cutting (i.e. aggregation); RawFreq vs. K50) on the percentage of papers that were in the dominant discipline. Two additional control variables were used: cluster size (extremely large clusters were expected to have lower percentages of papers attributable to a single discipline) and the discipline (large research communities in some disciplines, such as physics, may tend to be composed mostly of researchers from that disciplines). The ANOVA statistics are presented in Table 3.

The two ANOVA results in Table 3 have one strikingly similar result – the single most important independent variable in each case was the discipline. Large clusters associated with certain disciplines tended to have a very low (or very high) percentage of papers in the cluster. Surprisingly, cluster size only had an impact in one group of maps (reference paper maps) and did not have a sizable impact in the other group of maps. The least significant impact was associated with the methodological choices (using K50 vs. RawFreq or level of aggregation). Their effects were barely significant in two cases and had marginal impacts in explanatory value in all cases.

Table 3. ANOVA for % Dominant Discipline

Current Maps

Number of obs = 6492 R-squared = 0.3361
Root MSE = 0.180002 Adjusted R-squared = 0.3335

Source	Partial SS	df	MS	F	Prob > F
Model	106.059612	26	4.07921585	125.90	0.0000
k50	0.105449576	1	0.105449576	3.25	0.0713
max	0.281317031	1	0.281317031	8.68	0.0032
isi24	105.414756	23	4.58325026	141.46	0.0000
nrc	0.123902377	1	0.123902377	3.82	0.0506
Residual	209.470298	6465	0.032400665		
Total	315.52991	6491	0.04861037		

Reference Maps

Number of obs = 9076 R-squared = 0.3218
Root MSE = 0.169165 Adjusted R-squared = 0.3198

Source	Partial SS	df	MS	F	Prob > F
Model	122.845456	26	4.72482524	165.11	0.0000
k50	0.053313419	1	0.053313419	1.86	0.1723
max	0.128798077	1	0.128798077	4.50	0.0339
isi24	120.060839	23	5.22003647	182.41	0.0000
nrc	2.67210381	1	2.67210381	93.38	0.0000
Residual	258.952464	9049	0.028616694		
Total	381.79792	9075	0.042071396		

We therefore examined the average percentage of papers in a dominant discipline by discipline (Table 4). The analysis was limited to the large clusters associated with the K50:aggregated maps because they were the maps that had the fewest large clusters. Both maps (current papers and reference papers) suggest that large clusters dominated by physics, space science, and clinical medicine tend to be quite homogeneous. The large clusters that are dominated by pharmacology, biology & biochemistry, microbiology, molecular biology and genetics tend to be more multidisciplinary. This pattern is consistent with the observation made previously that the research communities associated with drug development activities are more multidisciplinary.

Table 4. Fraction dominant discipline (DomD) by discipline
(clusters from the K50:aggregated maps with 50+ papers)

Discipline	Current papers		Reference papers	
	%DomD	#clusters	%DomD	#clusters
Physics	83.8	57	78.6	65
Space Science	83.2	13	83.0	25
Clinical Medicine	70.4	194	71.6	314
Social Sciences, general	65.4	3	47.8	9
Plant & Animal Science	64.8	24	55.3	43
Chemistry	64.6	22	77.0	92
Geosciences	63.4	4	70.8	16
Agricultural Sciences	53.8	1		0
Ecology/Environment	52.8	5	48.2	15
Neurosciences & Behavior	52.0	30	60.8	95
Economics & Business	48.4	9	60.8	13
Immunology	47.8	16	55.0	56
Computer Science	47.5	10	64.8	1
Psychology/Psychiatry	46.2	10	62.8	41
Microbiology	45.9	9	53.2	44
Biology & Biochemistry	43.3	52	51.6	151
Molecular Biology & Genetics	43.3	24	53.4	130
Engineering	41.2	5	62.1	4
Pharmacology	38.9	5	46.7	20
Materials Science		0	49.2	1
Law		0	38.6	3
Mathematics		0	36.5	1
Multidisciplinary		0	31.5	11
Education		0		0

Discussion

We believe that large maps are better than small maps (e.g. maps based on a limited disciplinary scope) if one is interested in evaluating multidisciplinary research areas or anticipating the impact from cross-disciplinary research. The first large maps developed by Small were excellent illustrations of this principle. One could see how it would be inappropriate to only map physics, because of the important role of mathematics, chemistry and biology. This is even more true in the more detailed maps presented here. Many of the research communities are multidisciplinary (which means they would not have been properly specified if one only sampled a particular discipline). Many of the research communities in one discipline are related to research communities in another discipline (which means that one would have improperly specified the relationship between disciplines if one had a small map).

We have also shown that the large maps presented in this paper are a significant improvement over maps presented by Small. Table 5 summarizes the improvements in mapping worldwide science as reported in this paper. We make comparison only to reference paper maps, since no all-of-science current paper maps are available for comparison. By utilizing hardware, software and intellectual development during the past 20 years, it is now possible to generate a visual map at the paper level of analysis. These maps have greater coverage than the co-citation analyses previously reported in the literature. They cover a much higher absolute and relative number of available documents. For instance, we were able to cluster 6.6% of the reference papers. This is significantly higher than the 2.1% clustered by Small[3] in his first-level clusters, or the 0.6% retained in his fourth-level cluster map (reproduced here in Figure 2). In addition, disciplinary bias is significantly less in our maps, partially due to highly increased coverage, but also to the fact that our ordering was based on graph distances rather than citation thresholds.

Table 5. Comparison of reference paper map attributes

	Griffith et. al. (1974)[1]	Small (1999)[3]	This paper
Sample	1st Quarter 1972	1995	2002
# Reference papers	867,600	~6,100,000	10,911,939
# Papers in initial sample	1,832	164,612	718,894
# Papers in first level map	1,150	129,581	718,894
% Papers retained	0.13%	2.12%	6.59%
% Dropped during clustering	37.2%	34.3%	0%
Disciplinary bias	Unknown	0.175 – 0.39	0.102
# Clusters	41	18,938	63,369
# Papers per cluster	28.0	6.8	11.3

We identified over three times more documents clusters than Small.[3] Many of these additional communities had only one (or no) elite reference paper. We suggest that a high threshold level has resulted in underspecification of the number of research communities in science. It is quite possible that there are research communities with only one elite paper that forms the exemplar. And it is also easy to imagine that there are no elite references in some research communities, but rather just a general consensus on the literature that the community builds upon.

Cluster sizes are larger in this sample for the simple reason that we did not impose an arbitrary limit. Our investigations suggest that the coherence of clusters with more than 50 papers is mostly a disciplinary effect, not a size effect. We see no reason to impose this threshold on the clustering algorithm.

Using the results of this work, we make the following additional observations regarding the mapping of large-scale literature data sets. First, quantitative evaluation of maps on the scale of hundreds of thousands of papers is not only possible, but should be

done whenever such maps are generated. We have introduced several metrics for such an evaluation, and have applied them to our large scale maps generated from citation similarities. These metrics could also be applied to text based or other similarity types. Our analyses suggest that accuracy can be increased by using a modified cosine relatedness measure rather than raw frequencies. Disciplinary biases can be reduced and accuracy increased by increasing coverage of the scientific literature. Much larger samples of papers can and should be used to create the most accurate maps of science.

We note that our results are closely tied to the performance of the VxOrd graph layout algorithm, and thus we do not claim that other visualization methods will achieve similar results or findings. Nevertheless, VxOrd seems to have a number of unexpected side effects that are quite useful for large-scale graphs, including literature maps. VxOrd generates relatedness statistics that may have superior characteristics (in a separate study, we found that the re-estimated relatedness statistic between journals was more accurate than the raw relatedness statistics[5]). These re-estimated measures (e.g. distances between nodes) show that coverage can be increased without sacrificing accuracy. Fewer documents are dropped during clustering and disciplinary biases can be kept very low. The edge-cutting capability of VxOrd (see Appendix A) allows development of very good clustering algorithms. Edge-cutting seems to approximate the role of average-link clustering in that chaining is avoided because edges that would be associated with these chains are dropped during graph layout. Our results show that cluster size distributions are reasonable.

We did some spot checking of abstracts and titles in order to see if the research communities had the type of topical focus that we expected. The membership of a representative community is listed in Figure 7, and shows that the community does indeed have a strong topical focus. The only paper that appears not to fit is the sixth paper from the journal *Obes Surg*. Spot checking of another 50 communities gave similar results.

Journal	**Paper title**
Scientometrics	Reflections on **scientific collaboration**, (and its study): past, present, and future
Scientometrics	Continuity and discontinuity of **collaboration behaviour** since 1800
Scientometrics	Elite researchers in ophthalmology: Aspects of publishing strategies, **collaboration** and multi-disciplinarity
Scientometrics	The effect of team consolidation on **research collaboration** and performance of scientists. Case study of Spanish university researchers in Geology
Scientometrics	Recognition and **international collaboration**: the Brazilian case
Obes Surg	Progress of the International Federation for the Surgery of Obesity
Scientometrics	A. H. Zewail: **Research collaborator** par excellence
Scientometrics	Age structures of **scientific collaboration** in Chinese computer science
Philos Sci	The epistemic significance of **collaborative research**
Braz J Med Biol Res	The Brazilian **investment in science and technology**
Region Anesth Pain M	What do we measure by **co-authorships**?
Scientometrics	**Authorship patterns** in agricultural sciences in Egypt

Figure 7. Membership of community 73,336 from the 2002 current paper map of science. Common terms are highlighted to show the topical focus of the community

Using VxOrd for graph layout, the best overall maps are obtained using the modified cosine similarity measure and setting the cutting parameter to create "aggregated" maps. This combination gives the best combination of high accuracy at full coverage, and low disciplinary bias for both current paper and reference paper maps.

We invite others to test their algorithms on literature data sets of similar scale so that the results can be compared.

*

This work was supported by the Sandia National Laboratories Laboratory-Directed Research and Development Program. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

References

1. GRIFFITH, B. C., SMALL, H. G., STONEHILL, J. A., DEY, S., Structure of scientific literatures. 2. Toward a macrostructure and microstructure for science, *Science Studies*, 4 (1974) 339–365.
2. SMALL, H., SWEENEY, E., GREENLEE, E., Clustering the Science Citation Index using co-citations. II. Mapping science, *Scientometrics*, 8 (1985) 321–340.
3. SMALL, H., Visualizing science by citation mapping, *Journal of the American Society for Information Science*, 50 (1999) 799–813.
4. BOYACK, K. W., KLAVANS, R., BÖRNER, K., Mapping the backbone of science, *Scientometrics*, 64 (2005) 351–374.
5. KLAVANS, R., BOYACK, K. W., Identifying a better measure of relatedness for mapping science, *Journal of the American Society for Information Science and Technology* (in press).
6. KLAVANS, R., BOYACK, K. W., Mapping world-wide science at the paper level, Paper presented at the *10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, 2005.
7. FRANKLIN, J. J., JOHNSTON, R., Co-citation bibliometric modeling as a tool for S&T policy and R&D management: Issues, applications, and developments. In: VAN RAAN, A. F. J. (Ed.), *Handbook of Quantitative Studies of Science and Technology*. Elsevier Science Publishers, B.V., pp. 325–389.
8. CHEN, C., KULJIS, J., The rising landscape: A visual exploration of superstring revolutions in physics, *Journal of the American Society for Information Science and Technology*, 54 (2003) 446–453.
9. NOYONS, E. C. M., MOED, H. F., LUWEL, M., Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study, *Journal of the American Society for Information Science*, 50 (1999) 115–131.
10. BOYACK, K. W., Mapping knowledge domains: Characterizing PNAS, *Proceedings of the National Academy of Sciences*, 101 (2004) 5192–5199.
11. BÖRNER, K., CHEN, C., BOYACK, K. W., Visualizing knowledge domains, *Annual Review of Information Science and Technology*, 37 (2003) 179–255.
12. KESSLER, M. M., Bibliographic coupling between scientific papers, *American Documentation*, 14 (1963) 10–25.
13. JONES, W. P., FURNAS, G. W., Pictures of relevance: A geometric analysis of similarity measures, *Journal of the American Society for Information Science*, 38 (1987) 420–442.
14. BATAGELJ, V., MRVAR, A., Pajek - A program for large network analysis, *Connections*, 21 (1998) 47–57.
15. DAVIDSON, G. S., WYLIE, B. N., BOYACK, K. W., Cluster stability and the use of noise in interpretation of clustering, *Proceedings of IEEE Information Visualization 2001*, 2001, pp. 23–30.
16. SMALL, H., Paradigms, citations, and maps of science: A personal history, *Journal of the American Society for Information Science and Technology*, 54 (2003) 394–399.

Appendix A
Visualization algorithm

VxOrd[15] is a graph layout (visualization) algorithm that calculates the positions of data objects on a two-dimensional plane using similarities between the data objects. In the case of this paper, the similarities are the paper-paper relatedness matrices, and we use only the top 10 similarities per paper rather than the entire matrix as input. At the most basic level the VxOrd algorithm tries to place similar objects close together and dissimilar objects far apart. This process is achieved by moving the objects randomly around the solution space via a technique similar to 'simulated annealing'. The criteria for moving a node is the minimization of energy given by:

$$E_{i(x,y)} = \left[\sum_{j=0}^{n} \left(w_{i,j} \times l_{i,j}^{2} \right) \right] + D_{x,y}$$

where $E_{i(x,y)}$ is the energy of node i with n edges at a specific x y location, $w_{i,j}$ is the similarity between nodes i and j, $l_{i,j}$ is the Euclidean distance between nodes i and j, and $D_{x,y}$ is a density measure with respect to the area around point x,y. This density field is constructed as the sum of the energy footprints from each node, where the energy footprint is a function of r^{-2} from the node location.

The energy equation is gradually minimized in successive phases in an iterative fashion. The first phase (expansion) reduces the free energy in the system by expanding vertices toward the general area where they will ultimately belong. The following phase (cooldown) is similar to the 'quenching' step that occurs in simulated annealing algorithms, the nodes take smaller and smaller random jumps to minimize their energy equations. We have found that 800 iterations work well for complex graphs ranging from tens to hundreds of thousands of nodes.

VxOrd employs two additional features that lead to better graph layout: barrier jumping and edge cutting. Barrier jumping overcomes the situation where some nodes that belong near each other get "hung up" in a local minimum or behind an energy barrier. Barrier jumping is achieved by directly solving for the location that minimizes the attractive term in the energy equation for a single node. These calculations are done during the cooldown phase, and decline from 25% to 10% of all calculations over the course of this phase.

Edge cutting occurs during the expansion and cooldown phases when the edge weight to edge degree ratio goes below a threshold. In essence, edge cutting involves resetting the similarity value for a particular edge to zero for the balance of the calculation. VxOrd allows some control over the cutting parameter: default values can be used (called native mode in this paper), or the threshold can be reduced (to enable more cutting) or increased (for less cutting). In general, approximately 40% of edges are

removed using default cutting, while roughly 75% of the edges are removed when using maximum cutting (called aggregated mode in this paper). In practice, maximum cutting gives a result with very well defined and well separated clusters, with the majority of uncut edges remaining within clusters. With default cutting, the cluster boundaries are visually much less distinct.

Appendix B
Clustering algorithm

The clustering algorithm used in this paper was designed specifically to deal with extremely large two-dimensional graphs such as those generated by VxOrd. The input to the clustering algorithm is the distance between paired papers using the x,y coordinates from the graph. The original measures of relatedness (that were used to generate the graphs) are not used as inputs to the clustering algorithm because we found in an earlier study that they are generally less accurate than the re-estimated measures of relatedness (the distances between nodes on the graph).[5]

We started with single-link clustering because of computational simplicity (average-link clustering and other more sophisticated clustering algorithms cannot easily be run on graphs with 700,000 nodes). Single-link clustering starts by only focusing on paired relationships that are greater than a threshold level. These paired relationships are ordered (from the most related pair to the least related pair). In terms of assigning nodes to clusters, there are only three choices as one sequentially goes from the most related pair of nodes to the less related pair of nodes. Pairs can form a new cluster (if neither node has been previously assigned to a cluster); a node can be assigned to a cluster (if one of the nodes has been previously assigned); or clusters can be combined (if both nodes are assigned to different clusters).

It is the third choice that is the primary cause of the formation of very large clusters. Chaining (the combining of smaller clusters that just touch each other) is a common problem in single-link clustering. Small addresses this problem by requiring that clusters not be combined if the total size exceeds 50.[3] We suggest an alternative approach – generate a series of statistics that indicate the overlap between the two clusters that are candidates to be combined. We propose that clusters should not be combined if the statistic suggests that the clusters just barely touch. However, the clusters should be combined if the statistic suggests that the clusters overlap. Three statistics were used: the average location of the cluster, the number of documents in the cluster, and the average distance between documents within the cluster (as the documents were added sequentially). These statistics do not require additional

computational complexity since running averages can be easily calculated. The clustering algorithm runs extremely fast (less than $n\log n$ time) and is appropriate for clustering extremely large two-dimensional graphs.

The threshold used in the clustering algorithm is based on the distributions of relatedness statistics. Papers are ordered in terms of the distance to their nearest neighbor from the graphs. The cumulative number of papers is calculated (from closest distance to furthest distance). The resulting curve has the minimum distance to the nearest neighbor on the y axis and the cumulative number of papers on the x axis. Thresholds are located where the curve bends sharply. All of the maps generated in this study had a clearly defined bend in these curves.