

Henry Small and His Sciences Mapping

Xiaohua Zhou
Xiaohua.Zhou@drexel.edu

1. Introduction

Henry Small is Chief Scientist at the Institute of Scientific Information in Philadelphia. He is one of the foremost scholars in the area of developing and applying co-citation analysis. This important work has resulted in a better understanding of the structure, relationships, and evolution of the sciences. In 1973, he and Marshakova independently proposed using highly cited papers and their frequency of co-citation as the building blocks for a mapping of science, which spawned a great many new works on citation analysis and citation mapping of the scientific literature.

He has served on the JASIS editorial board since 1985 and is a Fellow of the American Association for the Advancement of Science (AAAS). In 1987, Dr. Small received the JASIS Best Paper Award and the Derek de Solla Price Medal from the journal "Scientometrics". In 1998, he received the Award of Merit from the American Society for Information Science & Technology (ASIST)—the highest honor the Association bestows on individuals.

Dr. Small received a joint Ph.D. in Chemistry and the History of Science from the University of Wisconsin. After a brief career as a historian of science at the American Institute of Physics' Center for History and Philosophy of Physics, he joined ISI in 1972. Prior to being named Chief Scientist, he held the position of Director of Contract Research.

Since he published the paper on co-citation in the scientific literature in the Journal of the American Society for information Science (JASIS) in 1973, Small has published around 25 papers in the area of co-citation analysis and sciences mapping. I pick five of them across over ten years and try to reflect his creative and exciting work in sciences mapping. They are in ascending chronological order: (1) *The Geography of Science: Disciplinary and National Mappings* (1985), (2) *Macro-level Changes in the Structure of Co-citation Clusters* (1993), (3) *A General Framework for Creating Large-scale Maps of Science in Two or Three Dimensions: The SciViz System* (1998), (4) *A Passage Through Science: Crossing Disciplinary Boundaries* (1999a), (5) *Visualizing Science by Mapping* (1999b).

It is interesting to track the author's continuous working on the improvement of the method of co-citation based science mapping within the selected document set. It is also exciting to capture his skillfully utilization of the result of science mapping. The two clues are always intertwined over time. The improvement in method enhances the utilization, and vice versa.

Small's continuous work can be further partitioned into the following goals: (1) To improve the performance of the method of global science mapping against the challenge of ever increasing documents in ISI database;(2) To optimize the clustering algorithm to reflect the "true picture" of the science structures; (3) To enhance the display of science maps by making best use of the techniques of virtual reality; (4) To deepen the analysis of the sciences evolution patterns.

The remainder of the paper is organized as follows. In section 2, we examine the role of fraction citation counts in sampling highly cited documents from ISI database. In section 3, we fully discuss the method of document co-citation clustering, including the co-citation analysis, hierarchical clustering and single linkage algorithm. In section 4, we look back at the author's idea in ordination and display for obtained clusters. In the next two sections, we examine the two utilizations of science mapping, pathway discovery and longitudinal co-citation analysis. In the last section, we summarize the whole paper and discuss the remaining work in this area.

2. Fractional Citation Counts

Citation analysis has potential limitations. One of them is that some abnormal citation behaviors will negatively affect the result of citation analysis. In order to remove noise as far as possible, the first step is to set a citation frequency threshold and select only the most cited documents for processing through a clustering algorithm.

Since citation rates differ across fields of science, applying a simple integer threshold biases the selection to high citing fields—biomedical research eliminates less citing fields such as mathematics. The method Small comes up with, while perhaps not ideal, gives quite satisfactory results. It is called fractional citation counting and amounts to assigning to each published paper or citing item one unit of strength to be divided equally among all its references. For instance, for a published paper with 10 references, each reference has 0.1 unit of strength. The fractional citation counting threshold in part eliminates the effect of uneven citing phenomenon across disciplines.

He applies this method through the selected document set. In 1985 and 1999b, he discusses the effectiveness measure of this method by matching the rough disciplinary distribution of highly cited documents selected by it to the distribution of source papers appearing in ISI index. If the percentages are comparable, then we are sampling in the field in proportion to its representation in the database. For example, we know that about 14 percent of the source items in the database are from social sciences. The number of highly cited documents in the social sciences selected by the fractional threshold is about 12 percent, indicating proportional coverage of that field.

Though the author has introduced fractional citation counts threshold, low integer cut off is still provided to prevent the introduction of random noise. For example, a document is cited by only 3 other documents whose reference lists lengths are all 3; hence the fractional count is

1.5. In case of fractional count threshold is set to 1.5, this document will be selected though it is obviously not a highly cited document.

In 1999b, Small fully discusses the choices of fractional cut-off and integer citation threshold in order to obtain a representative sampling from the database. According to the effectiveness measure in 1985, the highly cited documents may over- or under-represent a topic. The mapping result in 1999b shows that biomedical fields seem somewhat under-presented while physical sciences area over-represented. Since biomedical papers in general have longer reference lists than physical sciences, it appears that the fractional method has somewhat overcompensated for the expected biomedical bias. One way to correct for this is to increase the initial integer citation threshold and lower the fractional cut-off. However, the choices of integer citation threshold and fractional cut-off are kind of art.

3. Document Co-citation Clustering

Small presents a general framework for creating large-scale maps of science in two or three dimensions in 1998. To speak briefly, the strategy for achieving large-scale mapping is to break the database into smaller chunks by hierarchical clustering, and reassemble the pieces into an overall structure. Its advantage is that the computation time is much less dependent on the number of objects analyzed since only a small subset has to be dealt with at a given time. However, this more piecemeal approach requires that a combination of clustering and ordination steps be carried out. Small employed this framework and strategy to map science structure across the selected documents set.

3.1 Document Co-citation Linkage

According to the framework 1998, the first step in the process is to define links among objects. If the objects are documents, the linkage can be based on a direct relationship such as a citation linkage, or an indirect linkage such as the sharing of linguistic attributes, as in co-word analysis (Callon et al. 1985), or of citations in co-citation (Small 1995). Small, however, implements document co-citation as linkage in SciViz, a prototype of the framework.

Co-citation frequency is the number of times two documents are cited together by later papers. In order to obtain a fair representation of large and small areas, the association measure is normalized by dividing by the square root of the co-cited documents as below

. $Freq = AB / \sqrt{A * B}$ Where, AB denotes the co-citation frequency, A denotes the times cited of document A, B denote the times cited of document B. Small points out the advantage of this normalized document co-citation frequency as the linkage strength in 1999b. Since that two documents are cited by an author in the same paper to some extent accounts for the similarity of two documents in terms of intellectual content, the normalized co-citation frequency can be taken as a coefficient of similarity between two documents, and further a

metric that could differentiate distances between objects as below formula.

$$\text{Distance } A-B = (1 - \text{similarity}) / (1 - \text{similarity threshold})$$

Co-citation linkage takes another important responsibility in analysis of pathways through sciences as depicted in 1999a. A strong interdisciplinary link can occur when an author co-cites across the boundary of two disciplinary clusters. The interdisciplinary co-citation reaches out beyond the author's home cluster. This reaching out or stretching can import or export methods, ideas, models, or empirical results from the author's to the other field.

3.2 Hierarchical Clustering

In his framework 1998, Small thinks that hierarchical clustering solution is preferable because we can subsume smaller scale structures into larger scale structures. Given a set of N items to be clustered, and an N square distance (or similarity) matrix, the basic process of hierarchical clustering follows the steps as below:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

In Small or ISI version of hierarchical clustering, the process varies slightly. Firstly, the clustering is completed within four discrete iterations rather than N continues steps. Secondly, at each level, clusters in previous level are taken as the sole input for clustering, and newly formed clusters can't be merged to other same level clusters. C1 is the lowest level taking documents as input. C5 is the highest level and there is only a single super cluster at this level. The detail of the algorithm and its strength and weakness will be discussed in section 3.3. Now we switch to discuss the implications of the obtained five levels clusters.

The hierarchical clustering gives us the flexibility to expand or collapse the science structure. When you want to hide the complexity of co-citation patterns and focus on the macro tendency, you can move up to the high level clusters. When you want to observe the scientific ideas or specialities, and obtain concrete insight, you can move down to the low level clusters. This functionality is fully illustrated by the analysis of pathway through science in 1999a and the analysis of changes in science structures over time in 1993

The changes in science structures over time roughly can be classified into three patterns according to involved levels. First, it is micro-level evolution, where we deal with histories of individual scientific ideas and specialities, especially the emergence or disappearance of ideas or specialities at low level clusters. Second, it is macro-level change, where change occurs in the entire bodies of knowledge or their interactions with one another, especially the separation or aggregation of more than one fields or disciplines at high level clusters. Third, it is the

growth or recession between disciplines, topics or fields across levels. For example, in 1983 and 1985 mathematics is attached to physics on the C4 map as a C3 cluster, while in 1984 and 1985 mathematics appears on the C5 map as a C4 cluster.

Pathway discovery involves two kinds of path-form processes. One is to find a sequence of lower level objects within a higher level object. The other is to find the most strongly linked lower level objects within two adjacent higher level objects. The two modes of traversal might be termed stepping and jumping as shown in Figure 1. The “jumps”, of course, can involve traversing more weakly connected or distant objects and may entail larger shifts in subject matter while the “steps” are more strongly linked and closer in topic.

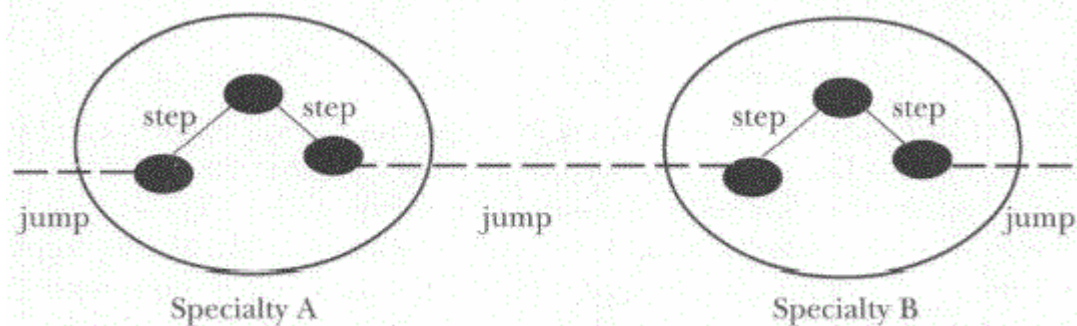


Figure 1. Path Formation: Steps and Jumps.

3.3 Single Linkage Method

Frequently used hierarchical clustering algorithms include single linkage, complete linkage, and average linkage. The distinctions of the above three algorithms are the methods they employ to recompute the similarities or distances in course of clustering as shown in Figure 2.

In single linkage clustering, distance between two clusters is defined as the minimum distance from any member of one cluster to any member of the other cluster. If similarity information is available, the similarity between two clusters is defined as the maximum similarity from any member of one cluster to any member of the other cluster. Similarly, distance in complete linkage clustering is defined as the maximum distance from any member of one cluster to any member of the other cluster while distance in average linkage clustering are defined as the average distance from any member of one cluster to any member of the other cluster.

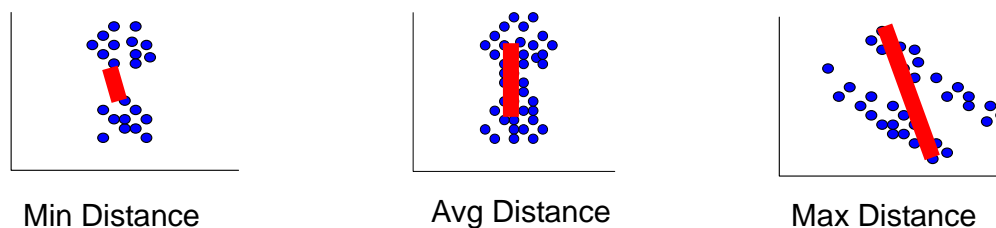


Figure 2. Comparisons among Single-link, Average-link and Complete-link Clustering.

Small adopts single linkage clustering through selected document set because of its simplicity

of implementation for massive files. However, its clusters sometimes are highly chained since single linkage is sort of loose and weak connection between objects. For this reason, he introduces a cut-off in cluster size and a similarity threshold to limit the amount of chaining. If the cluster size exceeds the cut-off, the similarity (normalized document co-citation frequency) threshold is incremented and the clustering repeated with the same seed document until the resulting cluster was within the size limit. Because of the similarity threshold, the iteration can't include all clusters obtained at given level or iteration in the macro-clusters obtained at the next higher level. For example, the global map at C5 includes only about one third of the C1 clusters reported in 1985, and similarly reported in 1999b, the remaining 71% of documents fall into clusters at lower levels that become separated from the main hierarchy at some point. Due to the mentioned deficiencies of single linkage, it is under criticism (Burgin, 1995).

Small further discussed the choices of hierarchical clustering method in 1993. Since the intensity of common referencing is an indicator of whether the entities are the same discipline, we must operationalize the meaning of "common referencing" using document methodology. For example, if we select single-link clustering, we obtain loose, weakly linked networks of research areas, whose constituents may only share references with their immediate neighbors. Complete-link clustering, on the other hand, yields only solidly linked and more isolated blocks of researchers, where each constituent must share references with every other. Sociological theory suggests that the method of linkage may vary with field.

4. Ordination and Display

After document clustering, the remaining first work is ordination that is the process used to position each object in space. There are numerous methods for ordination including factor analysis, principal components analysis, correspondence analysis, multidimensional scaling, and triangulation. In 1985 and 1993, Small adopts multidimensional scaling for ordination while he applies triangulation in framework 1997 and later works.

The second work is to combine individual ordinations for each object or cluster at specified levels into a common coordinate space which expresses their hierarchical relationships as geometric positions, and preserves their relative locations within and between levels.

4.1 Multidimensional Scaling

Multidimensional Scaling (MDS) is a set of techniques used to create visual displays—maps, from proximity matrices same as the input of clustering. The major output of MDS is a display of points, usually mapped in two or three dimensions based on their proximity in the original matrix. A major purpose of MDS is to capture as much of the original data as possible in only two or three dimensions. This simplification, while valuable, necessarily distorts the original data somewhat and can't account for all the variance in the proximity matrix. MDS programs summarize the distortion with a statistic called "stress", which is a criterion for

determining the “best fit” between the original input matrix “distances” and the estimated distance in the chosen low-dimensional solution.

Apart from distortion, MDS is also a computational intensive algorithm since it involves the calculation of a minimum value for a goodness-of-fit measure called stress. I think, that is the main reason Small turns to triangulation for ordination in the 3 most recent papers considering the ever-increasing number of scientific literatures in information age, while he applies MDS in 1993 and 1985.

4.2 Geometric Triangulation

Triangulation is a purely geometric procedure yielding configurations that exactly represent small numbers of distances between object, but lack global optimization. Small says in 1999b, the motivation for using triangulation is to see if a simpler and faster method is adequate for the visualization task at hand, thereby providing a computational less demanding solution.

Before applying triangulation, it is required to transform similarities into distances between objects. The simplest way subtracts the similarity coefficient from one. However, since different clusters may use different similarity threshold, it is necessary to normalize the distance. For this reason, the final transformation formula is defined as below:

$$\text{Distance } A-B = (1 - \text{similarity}) / (1 - \text{similarity threshold})$$

If similarity is equal to the similarity threshold, the distance is equal to one Garfield— the distance associated with the weakest link on the map.

In 1999b, Small describes the algorithm. The positioning of objects is accomplished by taking the two strongest links for each object to be added to the map. For the two dimensional case, it begins arbitrarily with one of the objects and places it at the origin of the coordinate systems. Then the object closest to it is found and placed at the specified distance from the first object, oriented in an arbitrary manner. The location of the third object is fixed using distances from the first two objects. This leaves a freedom to pick one of the two possible quadratic solutions, above or below the line formed by the first two objects. From this point on we use the notion of repulsion from the center of gravity to select the quadratic solution furthest from the center.

From the description of the algorithm, it is found that the algorithm depends on the order in which objects are assigned positions. However, the speed with which the calculations can be made means that every object in the cluster can be tested as the seed, and the best solution selected.

4.3 Global Coordinate Space

After positioning objects in individual clusters, it is time to put pieces into a common global coordinate space without overlap. Though both 1999b and 1998 fully discuss this topic, there is no remarkable progress made. Small mainly speaks to three points:

- (1) The strategy to scale up the individual clusters into the global space. It works from the document level to the most aggregated level. Then working back through the levels, each coordinate system is translated so that its centroid is moved to the location of the parent object that contains it.
- (2) Initial expansion to minimize the object overlap. The size of a cluster is determined by finding the circle whose origin is at the centroid and whose radius is just large enough to enclose the objects it contains. An expansion factor for the coordinate system is then determined satisfying the condition that adjacent circles along a minimal spanning tree path through the cluster will be exactly tangent, that is, not overlap.
- (3) The routine to avoid overlap. This procedure consists of traversing the objects in the cluster in a set sequence, and moving the higher member of each overlapping pair away from the centroid until all overlaps have been eliminated.

Another problem needed to be solved is the isolated objects or clusters at some level generated during iterative clustering. In 1998, he suggests considering them as separate worlds, or islands without connection to the mainland.

4.4 Volvox Display

The style of visualization used has been termed a Volvox (McCain, 1996) because smaller, lower-level objects are represented as circles within larger, higher-level objects, which resembles a microorganism by that name. In 1999b, Small mentions its advantage that it allows simultaneous viewing of relative location and hierarchical structure. Links between objects can also be drawn to clarify relationships that can't be adequately captured by ordination. The volvox format can be easily extended to three dimensions by substituting spheres for circles.

The other problem worth mentioning is the subject label of the clusters. Through the document set, Small bases the subject labels on a frequency analysis of articles and journal category names.

With the availability of powerful hardware and software systems, information visualization is rather promising. Small also sets new goals for PC-based display of science structures in 1999a. The goal set includes (1) continuously and smoothly navigation or zooming across a map and flying into or out of a map to progressively reveal more or less detail, (2) object (documents, ideas, specialties, fields or disciplines) probing on science map by text query, and (3) diverse representation for objects and their relationship like virtual reality.

5. Information Pathway

The basic requirement of a pathway through science is that the linked objects form a chain of significant connections. Ideally each connection represents a relationship whose logic can be determined by some form of content analysis. In an abstract sense, an information pathway

could be defined as a sequence or succession of information objects or events (document, descriptors, topics) such that each object along the path bears some kind of relationship to the objects that precede it.

There are several kinds of information pathways. One of them is complete path that tour the network with shortest document path. However, Small focuses on linear paths that connect arbitrarily selected beginning and ending documents or clusters, and presents a methodology for creating pathways through the scientific literature following strong co-citation links in 1999a. The method consists of processes: step and jump (see section 3.2, Figure 1) by which we can form a path connecting the beginning and the ending point.

As a matter of fact, it is not difficult to find such an information path by applying some simple algorithm in graph theory. In 1999b, he reports a knowledge connector system, which can automatically generates strongly linked document pathways for user specified starting and destination topics. The true challenge comes from the interpretation of each link in the pathway. A proper analysis of the nature of the topic transitions would require a content analysis of the co-citation passages for pairs of documents along the path. However, a first order approximation to understanding the nature of the transitions can be based on a close examination of the titles of the linked documents.

A lot of work still remains to deal with the interpretation of the pathways. However, this new specialty has great implications for retrieval, the unity of science, discovery, epistemology, and evaluation, which appeal to researchers in information science.

6. Macro Evolution of Sciences

Maps generated from annual ISI database can be linked year to year since the same pairs of papers and clusters are often co-cited in successive years. This provides a moving picture of frequently rapid evolution of knowledge. Using the annual science maps from 1983 to 1989 generated by the same method at ISI, Small did an amazing study on the evolution of science structures in 1993.

It is a quite challenging work because, first it is difficult to identify the same cluster across years, second it is difficult to measure or depict the change of structures over time. The author uses cluster strings to correspond between clusters across the years. The links between clusters across time are based on a normalized measure of the number of common highly cited documents in successive year clusters. A sequence of such continuing clusters is called a cluster string.

He thinks of the changing associations of disciplines as a process of competitive binding among fields, analogous to atoms competing for a binding site on a molecule. Competition comes about because of the limitation of cluster size in single linkage clustering. If more than this number of entities bind together, the similarity threshold is raised until enough fragments disengage so that the resulting cluster is within the size limit. These fragments can then

cluster together at the next higher level.

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

Applying this method, he discovers a “pulsating model” of knowledge evolution. That is, periods of discovery are indicated by relatively small groups of emerging clusters that isolated from the larger, established research disciplines. This is followed by periods of integration in which the new clusters become densely linked or even merged with other disciplines as their research is applied and extended in other fields.

Though Small made remarkable progress in longitudinal co-citation analysis, a lot of work remains to do. Firstly, we need other co-citation methods, e.g. author co-citation analysis, to verify his result. Secondly, since the mapping result depends on the cut-off of cluster size, it is kind of fractal systems. We need alternative clustering method or to fix the single linkage clustering algorithm we currently use. Thirdly, it is urgent to better represent the science evolution by visualization, otherwise only few bibliometric experts can discover and interpret the pattern of science evolution. Last, we can extend analysis of macro-level changes in science structures to that of micro-level. Actually, Small stated this problem in his 2003 personal review, which is not contained in the document set.

7. Conclusions

Small and ISI have formed a method of global science mapping since he proposed the method of document co-citation analysis in 1973. The mapping method can be summarized as following 3 steps: (1) Make a fairly representative sampling of highly cited document from ISI database by choosing appropriate integer citation cut-off and fractional citation counts threshold, (2) Employ four-iteration single linkage algorithm with cut off of cluster size and a variable similarity threshold for document clustering given normalized document co-citation frequency matrices, (3) Use multidimensional scaling or triangulation for individual cluster ordination, then put pieces into a common coordinate space resulting a volvox display.

Moreover, he contributes much importance to the longitudinal co-citation analysis which enables researchers to discover macro- and micro-level patterns of science evolution, and information pathway discovery and analysis. With the ever-increasing scientific literatures, he is still working on the improvement of mapping method performance and the enhancement of its utilization.

Without doubt, his amazing work in document co-citation analysis and science mapping has great implications for fields of bibliometrics and information visualization, for both

researcher and profession in information science, and scientific policy makers.

However, lots of work remains to deal with both existing problems and new challenges. While the single linkage method has advantages of simplicity of implementation for massive files, its clusters are sometimes highly chained because of loose and weakly connection. With the cut-off of cluster size, only one third of the lowest level clusters are subsumed in the super cluster. That the clustering result depends on the choice of cluster size threshold makes the result of longitudinal co-citation analysis instable. It is urgent to find an alternative or to fix the existing problem of single linkage algorithm.

It is still challengeable to better display the result of clustering for browsing although advanced computerized display hardware, software and techniques are available. Especially, there has been no insight to intuitively represent the science evolution and information pathway in multi-layers science maps thus far.

The work on longitudinal co-citation analysis and information pathway discovery is still at infant stage. For the former, it is urgent to define a series of micro- and macro-level quantitative indicators to measure the change of science structures over time based on which lots of work such as automation of structure change pattern discovery can be initiated. For the latter, it is of interest and challenge to make machine interpret the links in pathways and further incorporate this feature into information retrieval systems.

References

1. J.B. Kruskal (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 28, 1-27
2. Lee, R.C.T., Slagle, J.R., & Blum, H. (1977). A triangulation method for the sequential mapping of points from N-space to two-space. *IEEE Transactions on Computers*, 26, 288–292.
3. Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265–269.
4. Small, H. (1993). Macro-level changes in the structure of co-citation clusters: 1983–1989. *Scientometrics*, 26, 5–20.
5. Small, H. (1998). A general framework for creating large-scale maps of science in two or three dimensions: the SciViz system. *Scientometrics*, 41, 125-133
6. Small, H. (1999). A passage through science: crossing disciplinary boundaries. *Library Trends*, 48, 72-89
7. Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50, 799–813.
8. Small, H. & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. *Journal of Information Science*, 11, 147–159.