# Tracking modularity in citation networks

Yoshiyuki Takeda[1], Naoki Shibata[1], Yuya Kajikawa[1], Ichiro Sakata[1,2] and Katsumori Matsushima[1]

[1] *{ takeda, shibata, kaji, sakata}@biz-model.t.u-tokyo.ac.jp, matsushima@iijmio-mail.jp*
Innovation Policy Research Center, Institute of Engineering Innovations, School of Engineering, The University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-8656, Japan

[2] *isakata@pp.u-tokyo.ac.jp*
Todai Policy Alternative Research Institute, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

## Abstract

Clustering using cocitation and bibliographic coupling is an effective tool to analyze the structure of scientific research, but its arbitrariness in the setting of a clustering threshold is a problem. This study tracked modularity of citation networks in research domains, and we found that there are three stages in clustering of citation networks and it is universal across our case studies. In the first stage, core clusters in the domain are formed. In the second stage, peripheral clusters are formed, while core clusters continue to grow. In the third stage, core clusters grow again. By focusing on the elementary process in clustering, we can understand the structure of citation networks in each research domain and judge the location of each cluster and paper, which cannot be seen in the final clustering results.

## Introduction

Scientific activities are playing an increasingly important role in solving social problems and also as seeds of industrial innovations. In today's increasingly global and knowledge-based economy, competitiveness and growth depend on the ability of an economy to meet fast-changing market needs quickly and efficiently through the application of new science and technology. The capacity to assimilate and to apply new knowledge relies on scientific innovativeness. Therefore, for both R&D managers in companies or research institutions and policy makers, maintaining a comprehensive perspective on a knowledge domain and noticing emerging research domains has become a significant task.

Reflecting such importance, the amount of scientific knowledge measured by the number of academic publications has been rapidly increasing since the beginning of this century. Academic publications and patents provide the primary raw material, and their bibliometric information, i.e., authors, journals, titles, keywords, references, citations, and so on, constitute an adequate information source for the mapping or assessing fields or subfields of scientific and technological domains. Academic publications and patents constitute a generally accepted output indicator of scientific and technological activities. There are a number of reports and also many opportunities to create, disseminate, and apply knowledge to the real world in active research areas. But the speed and scope of development in such areas make it critical for researchers, engineers, and policy makers to notice information published across different research domains and different institutions. There is a commensurate increase in the need for scientific and technical intelligence to discover emerging research domains and the topics discussed there, even for unfamiliar domains (van Raan, 1996; Kostoff et al., 2001; Boyack & Börner, 2003).

Citation network analysis of scientific publications is a promising tool to overview scientific activities in a manner that individuals cannot handle. In his classical paper, de Solla Price (1965) originally introduced the concept of a research front. According to Price, there seems to be a tendency for scientists to cite the most recently published articles. The research front builds on recent work, and the network there becomes very tight. In a given field, a research front refers to the body of articles that scientists actively cite. Researchers have studied quantitative methods that can be used to identify and track the research front as it evolves over time.

Progress in computational speed enables us to treat a huge amount of data by bibliometric methods. In citation network analysis, cocitation (Small, 1973) and bibliographic coupling (Kessler, 1963) are frequently used to form a link between two documents. The created network data are used for visualization by using dimensional reduction techniques including multidimensional scaling (Small, 1977; Davidson et al., 1998), pathfinder networks (White, 2003; Chen et al., 2002; Chen, 2004), and force-directed placement (Davidson et al., 1998). A comprehensive review of knowledge domain visualization can be found in a recent article by Börner and others (Börner, Chen, & Boyack, 2003). Knowledge domain visualization is a useful approach for intuitively understanding the overall structure of research.

Clustering is also a useful technique for understanding the overall structure of research and to detect research fronts. Cluster analysis is based on the similarity measure that is calculated for each pair of nodes. There are a number of clustering algorithms such as single link, average link, and Ward’s method. Clustering results are also used in knowledge domain visualization (Small, 1999; Morris et al., 2003). For example, Morris et al. (2003) plotted each cluster along a horizontal track in the timeline, which is obtained after clustering by Ward's method using a cosine coefficient as a similarity measure on each link created by bibliographic coupling. The recursive clustering approach is a method to repeatedly perform clustering toward a network clustered once and is used to detect the hierarchical structure of a citation network. There are two methods for the recursive clustering approach, i.e., bottom-up (Small, 2006) and top-down (Takeda, Kajikawa, & Matsushima, 2007; Kajikawa et al., 2008) clustering is used to elucidate the hierarchical structure of a network. In bottom-up clustering, each cluster after clustering is treated as a node and then re-clustering is performed, while top-down clustering means that the initial nodes in each cluster are re-clustered and therefore subclusters are formed in an initial cluster.

Although clustering a citation network is a useful approach for elucidating and describing a network structure, there are some issues in the selection of similarity measures (Klavans & Boyack, 2006), clustering methodology, thresholds, and robustness of clustering (Hopcroft et al., 2004). In the selection of similarity measures, Klavans and Boyack (2006) compared the similarity of clustering results by intercitation with that by co-citation. They concluded that intercitation is more appropriate for clustering of similar documents. Intercitation also allows us to group papers that are only rarely cited, which is a significant portion of all papers (Hopcroft et al., 2004). Cocitation is less useful as a similarity measure, because it takes time to build up a cocitation record (Hopcroft et al., 2004). After defining the similarity measure, we select clustering methodologies having different characteristics and solutions, preventing unexpected results. For example, single-link cluster analysis gathers together links that share common papers. To prevent chaining, a maximum cluster size is set with a provision for re-clustering at a higher threshold (Small & Sweeney, 1985). And, to perform clustering, a threshold is sometimes set on the normalized co-citation coefficient of cosine similarity (Small, 2006). While a variance of threshold enables the user to freely navigate the knowledge domain from the macro to micro level, setting a threshold at a certain value is an arbitrary assumption, and we have less guidance for setting an adequate threshold. Therefore, the size of a research front obtained after clustering is dependent on the value of the threshold. In order to discuss the validity of such operations and to utilize the fruitful outcome of clustering results, understanding of the nature of citation networks is crucial but the existing literature is still insufficient.

The aim of this paper is to investigate the modularity change accompanying clustering. Recently, Newman (2004) defined modularity as the fraction of links in a network that connect nodes (within cluster links) minus the expected value of the same quantity in a network when random links are assumed between nodes with the same cluster divisions. Stopping clustering at the maximum modularity is a reasonable approach for dividing knowledge domains into appropriate units, but the user might require a different outcome on a different scale. Therefore, it is required that the system show different outcomes derived from different clustering thresholds and to navigate the user to overview the knowledge domain on a different scale, but in turn, there is no guidance for judging the relevance of such clustering results. Our motivation in this paper is to analyze the fundamental structure of citation networks by tracking modularity.

## Data and Methods

### *Data*

In this work, we analyzed intercitation networks of scientific publications in the above two domains. One is the energy domain, and the other is the materials science domain. Three sub-domains were constructed in energy domain, i.e., energy & fuels (E&F), the fuel cell (FC), and the solar cell (SC). For the materials science domain, two of its sub-domains, nanobiotechnology (NB) and chemical vapor deposition (CVD), were analyzed. The sustainability science (SS) was also analyzed as one of environmental science domains. These targets of analysis were set according to the specialty and interest of one of the authors (YK). We collected citation data of publications from the Science Citation Index (SCI) compiled by the Institute for Scientific Information (ISI). We used the Web of Science, which is a web-based user interface of ISI's citation databases.

The following queries are used to collect papers; fuel cell* for FC, solar cell* for SC, nano* and bio* for NB, chemical vapor deposition or chemical vapor deposition for CVD, and sustainab* for SS. The corpus for E&F was constructed by using journal categories of Journal Citation Reports (JCR). We collected papers published in the 68 journals classified under the category of energy & fuels in JCR. This is because when we use the simple query, energy, to collect energy-related papers, the results have a variety of non-energy-related topics and are noisy. A total of 152,514 records for E&F, 15,600 for FC, 16,199 for SC, 19,921 for NB, 46,525 for CVD, and 29,391 for SS were retrieved. We created citation networks using the above corpuses. In the construction of citation networks, we add cited papers by the above papers that were originally retrieved and collected using the above queries. Most of these cited papers were not included in the corpus retrieved by the queries. Then, we extract the maximum connected components in these networks. The number of papers in the components is 1,011,612 (E&F), 147,662 (FC), 150,607 (SC), 434,633 (NB), 467,292 (CVD), and 633,286 (SS) . These papers were then clustered using the algorithm described below, and modularity during the clustering was recorded.

### *Methods*

We focused on the topological clustering method in order to elucidate the structure of the above citation networks. Citation networks where each paper is connected by intercitation were divided into clusters. Of the many clustering methods and algorithms proposed, in this paper, we applied the method proposed by Newman (2004), because this could deal with large networks with small calculation time in the order of *$O((m+n)n)$, or $O(n^2)$* on a sparse network, with *$m$* edges and *$n$* nodes; therefore, this could be applied to large-scale networks (Newman, 2004; Newman & Girvan, 2004). The algorithm proposed was based on the idea of modularity. Modularity *$Q$* was defined as follows::

$$Q = \sum_{s} \left( e_{ss} - a_s{}^2 \right) = Tr(e) - \left\| e \right\|^2$$

where $e_{st}$ is the fraction of the edges in the network that connect nodes in cluster $s$ to those in cluster $t$, and $a_s = \sum_t e_{st}$. The first part of the equation, $Tr(e)$, represents the sum of density of links within each cluster. A high value of this parameter means that nodes are densely connected within each cluster. However, the maximum value of this is given if whole nodes are regarded as one cluster. The second part of the equation, $||e||^2$, represents the sum of density of links within each cluster when all edges are placed randomly. That is, $Q$ is the fraction of links that fall within clusters, minus the expected value of the same quantity if the links fall at random. A high $Q$ value represents a good community division where there are only dense links within clusters and sparse links between clusters. $Q$ = 0 means that a particular division gives no more within-cluster links than would be expected by random chance.

The algorithm to optimize $Q$ over all possible divisions to find the best structure of clusters is as follows. Starting with a state in which each node is the only member of one of $n$ clusters, we repeatedly join clusters together in pairs, choosing at each step the join that results in the greatest increase in $Q$. Usually, the clustering threshold by this clustering algorithm is set at the time that $Q$ reaches at its maximum, i.e., $Q_{max}$. Therefore, by stopping clustering at the maximum value of $Q$, $Q_{max}$, we can obtain the modularized network structure where there are only dense links within clusters and sparse links between clusters. It is reasonable to stop the clustering. But in this paper, instead seeking $Q_{max}$, we track $Q$ at each clustering step to elucidate the structure of citation networks. By doing so, we can use $Q$ as a clue to understanding the general structure of citation networks.

## Results

Figure 1 shows the result of our analysis of the modularity of citation networks during clustering. As shown in these figures, $Q$ increases as clustering proceeds, but after reaching $Q_{max}$, it suddenly decreases. But the increase of $Q$ is not linear throughout the entire process of clustering. At the early stage of clustering, $Q$ characteristically has some inflection points for all five domains. The number of inflection points is typically around four. There is a general tendency that in the early stage, some inflection points appear.

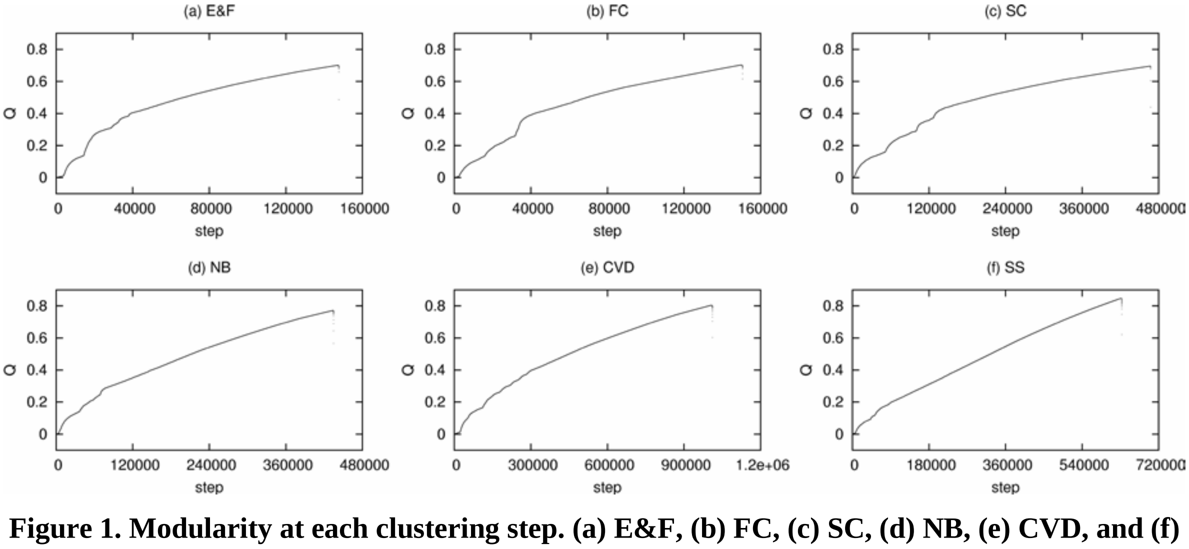

**Figure 1. Modularity at each clustering step. (a) E&F, (b) FC, (c) SC, (d) NB, (e) CVD, and (f) SS.**

In order to investigate the more detailed structure, we look at the number of nodes clustered in each step. In each clustering step, two clusters or nodes are gathered into one cluster. The size of clusters or nodes participating in clustering at each step is not always the same. Figure 2 shows the number of nodes in the bigger cluster, and Fig. 3 shows that in the smaller cluster.

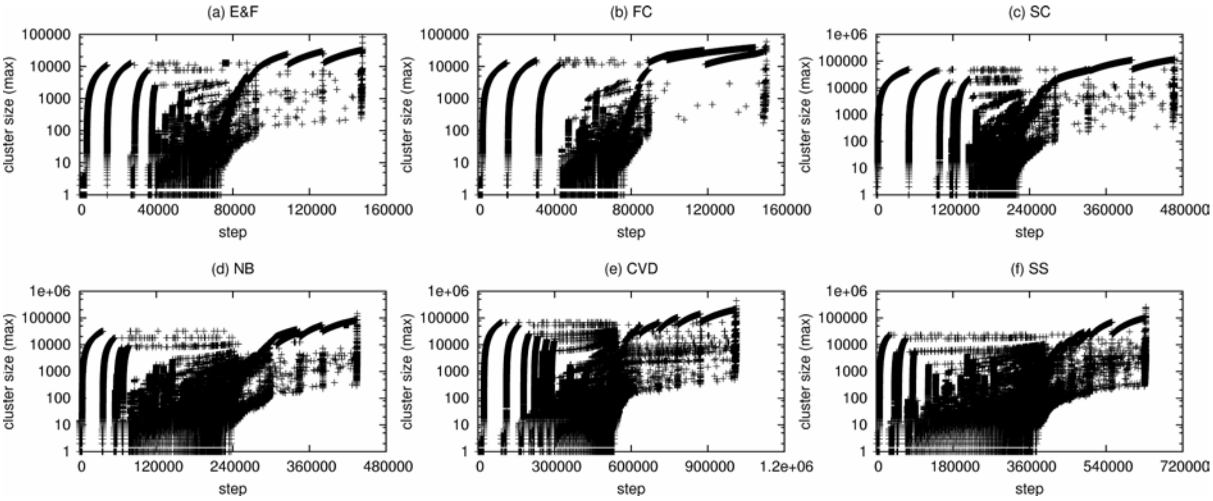

**Figure 2. Number of nodes of the larger cluster at each clustering step. (a) E&F, (b) FC, (c) SC, (d) NB, (e) CVD, and (f) SS.**

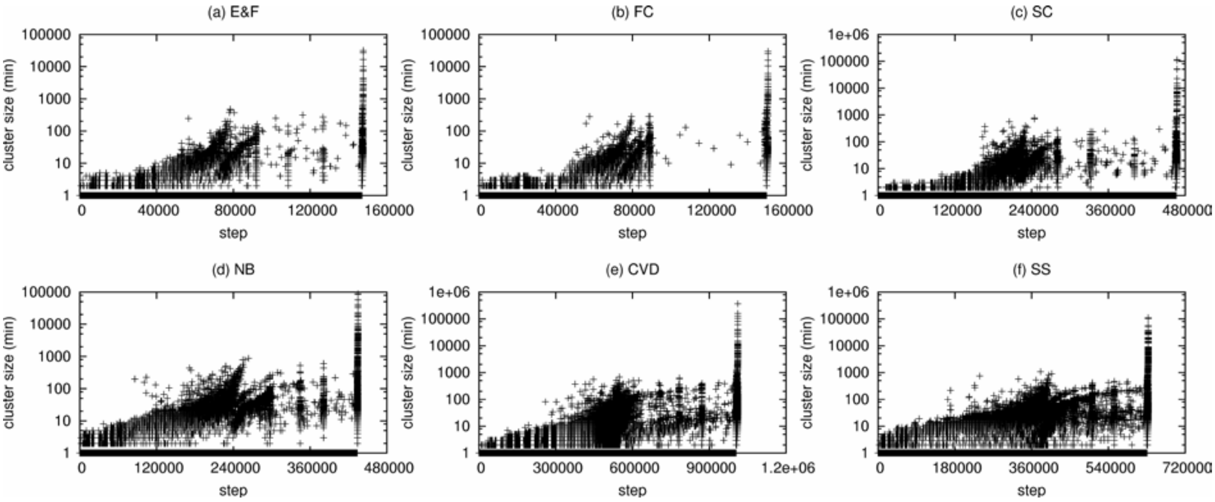

**Figure 3. Number of nodes of the smaller cluster at each clustering step. (a) E&F, (b) FC, (c) SC, (d) NB, (e) CVD, and (f) SS.**

When the number of node is one, it means that the cluster is a single node that has not participated in the clustering. At first glance, we can see three different stages in clustering. At the first stage, the plot is several continuous curves. At the second stage, the plot becomes more disturbed. At the third stage, it becomes several curves again.

For example, in the case of FC (Fig. 2(*)), until the clustering step around 40,000, four slopes are seen. This means that in this stage, each cluster grows exclusively, and when it reaches a certain size, the growing point shifts to the next cluster. The number of nodes in the cluster fused into the growing cluster is less than 10 (Fig. 3(*)). This exclusive growth continues until the clustering step around 40,000, which accords with the point at which nonlinear growth of *Q* appears as shown in Fig. 1(*). And the number of *Q* curves and inflection points in Fig. 1(*) seem to correspond with the number of continuous curves in Fig. 2(*). During the clustering step from 40,000 to 100,000, this exclusive growth is disturbed. And many medium-size clusters whose number of nodes is about 10 to 1,000 grow competitively (Fig.

2(*)). A feature of this stage is that the smaller cluster also has a variety of sizes (Fig. 3(*)). In this stage, these clusters coalesce into each other, but there is no dominant cluster that grows exclusively. When a pair of clusters coalesced, other pairs coalesced at the next step. After this competitive growth stage, only large cluster grows exclusively again after the clustering step of 100,000. Judging from Fig. 2(*), these clusters are the same as those generated in the first stage. The number of clusters fused into these large clusters is unity, and these are therefore single nodes. This growth of large clusters continues after $Q$ reaches $Q_{max}$. When we proceed with clustering after $Q$ reaches $Q_{\mathrm{max}}$, $Q$ suddenly decreases because large independent clusters coalesce and the clustered network is like a random structure.

## Discussion

In the above, we take FC as an example. But the remaining cases all exhibit similar trends, while there is a slight difference. But all domains studied in this paper have the three stages. At stage I, some clusters grow exclusively. At stage II, other medium-sized clusters grow competitively. At stage III, clusters generated at stage I grow exclusively again. This process is schematically shown in Fig. 4.

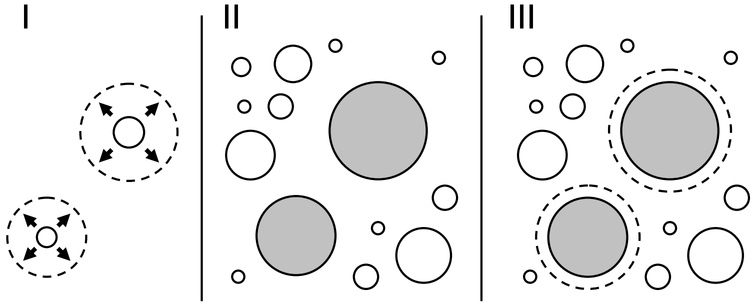

**Figure 4. Schematic illustration of the clustering process. I. Core cluster formation stage. II. Medium-sized cluster formation stage. III Core cluster growth stage.**

In stage I, several clusters emerge and these clusters grow collecting single nodes or small clusters whose sizes are less than 10. The growth is sequential, and a cluster grows exclusively and reaches a certain value. As determined by the clustering algorithm, the node that conveys larger $\Delta Q$ by clustering is preferentially aggregated. In the latter, we call clusters generated in stage I cluster A. The size of cluster A is larger than the other clusters formed in the later stage, and this type of cluster can therefore be regarded as the core cluster in each research domain.

In stage II, the exclusive growth of cluster A is disrupted, and the other clusters are formed and grow competitively. We call these clusters that form and grow in stage II cluster B. At the initial stage of the formation of cluster B, it grows by collecting single nodes or small clusters whose sizes less than 10 in a similar manner to cluster A. But after that, these clusters grow by coalescence of medium-sized clusters with sizes of around several tens or hundreds. The number of cluster B is larger that that of cluster A, but the size of each cluster is smaller in cluster B than that in cluster A.

In stage III, cluster A grows exclusively again by collecting mainly single nodes at each step. But these nodes are not included in stage I, and we can therefore regard these nodes as being in a more peripheral position in the network than nodes participating in the clustering of stage I.

Finally, when we proceed with clustering by exceeding $Q_{max}$, we can call this stage stage IV, and even cluster A coalesces within itself and $Q$ drastically decreases; at last, the network becomes a single cluster. This no longer gives an appropriate clustering result.

The above analysis gives us the following suggestions in the interpretation and utilization of the clustering results. One is to distinguish core domains (cluster A) and peripheral domains (cluster B). When cluster A grows, $\Delta Q$ is large, which means a denser citation network exists within these clusters than the other clusters generated at stage II. Therefore, we can regard these large clusters generated in stage I as core clusters in each research domain, while medium-sized clusters generated in stage II are peripheral. Therefore, it is reasonable to consider that cluster A includes main research topics in the research domain, while cluster B comprises rather minor topics although they should not be ignored. As shown in Table 1, network density defined by the number of links per node is smaller in cluster B than in cluster A. Therefore, we can regard cluster B as not being dense and thus not mature. But we must keep in mind that nodes included in cluster A can also be divided into two types. One is nodes clustered in stage I and other is nodes clustered in stage III. $\Delta Q$ in each clustering step is smaller for the latter than for the former, and therefore, nodes included in stage III are also peripheral and may be sometimes noisy in cluster A.

**Table 1. Density of networks.**

| | E&F | FC | SC | NB | CVD | SS |
|---|---|---|---|---|---|---|
| Cluster A | 1.90 | 2.67 | 2.45 | 2.09 | 2.77 | 1.50 |
| Cluster B | 1.15 | 1.11 | 1.09 | 1.14 | 1.14 | 1.08 |

As already mentioned above, for the purpose of clustering to obtain structures where there are only dense links within clusters and sparse links between clusters, it is desirable to set the clustering threshold at $Q_{max}$. But when we look at only the final clustering results, we cannot judge how clustering proceeds or what the characteristics of each node and each cluster are. By tracking the modularity during clustering, we can distinguish core and peripheral clusters and nodes in each cluster. One plausible direction for using these implications classified in this paper is to provide a mark to distinguish core and peripheral for users utilizing the clustering results such as policy makers and R&D managers.

## Conclusion

Clustering is an effective tool to analyze the structure of scientific research, but detailed structures obtained after and during clustering are not yet sufficiently elucidated. This study analyzed the structure of citation networks by tracking modularity during clustering. Five cases were analyzed, and these cases exhibit common characteristics. We found that there are three stages in clustering of citation networks and it is universal across our case studies. In the first stage, core clusters in the domain are formed and grow exclusively. In the second stage, peripheral clusters are formed and grow competitively. In the third stage, core clusters grow again by collecting peripheral nodes. By focusing on the elementary process in clustering, we can understand the structure of citation networks in each research domain and judge the location of each cluster and paper, which cannot be seen in the final clustering results.

## References

Börner, K., Chen, C., & Boyack, K.W. (2003). Visualizing knowledge domains. Annual Review of Information Science & Technology, 37, 179–255.

Boyack, K.W., Börner, K., (2003). Indicator-assisted evaluation and funding of research: visualizing the influence of grants on the number and citation counts of research papers. Journal of the American Society for Information Science and Technology, 54, 447-461.

Chen, C., Cribbin, T., Macredie, R., & Morar, S. (2002). Visualizing and tracking the growth of competing paradigms: two case studies. Journal of the American Society for Information Science and Technology, 53, 678-689.

Chen, C. M. (2004) Searching for intellectual turning points: Progressive knowledge domain visualization. Proceedings of the National Academy of Sciences of the United States of America, 101, 5303-5310.

Davidson, G. S., Hendrickson, B., Johnson, D.K., Meyers, C.E., & Wylie, B.N. (1998). Knowledge mining with VxInsight: Discovery through interaction. Journal of Intelligent Information Systems, 11, 259–285.

Hopcroft, J., Khan, O., Kulis, B., & Selman, B. (2004). Tracking evolving communities in large linked networks. Proceedings of the National Academy of Sciences of the United States of America, 101, 5249-5253.

Kajikawa, Y., Yoshikawa, J., Takeda, Y., and Matsushima, K. (2008). Tracking emerging technologies in energy research: toward a roadmap for sustainable energy. Technological Forecasting and Social Change, in press.

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

Klavans, R., & Boyack, K.W., (2006). Identifying a Better Measure of Relatedness for Mapping Science. Journal of the American Society for Information Science and Technology, 57, 251-263.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., & Saarela A. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11, 574–585.

Kostoff, R. N., del Río, J. A., Humenik, J. A., García, E. O., Ramírez, A. M., (2001). Citation mining: Integrating text mining and bibliometrics for research user profiling. Journal of the American Society for Information Science and Technology, 52, 1148-1156.

Morris, S.A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. Journal of the American Society for Information Science and Technology, 54, 413–422.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. Physical Review E, 69, 066133.

Newman, M.E.J., & Girvan M., 2004. Finding and evaluating community structure in networks. Physical Review E, 69, 026113.

Price, D. J. D. (1965). Networks of scientific papers. Science, 149, 510-515.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24, 265–269.

Small, H. G. (1977). A co-citation model of a scientific specialty: A longitudinal study of collagen research. Social Studies of Science, 7, 139–166.

Small, H., Sweeney, E. (1985). Clustering the science citation index using co-citations. I. A comparison of methods. Scientometrics, 7, 391–409.

Small, H. (1999). Visualizing science by citation mapping. Journal of the American Society for Information Science, 50, 799-813.

Small, H. (2006). Tracking and predicting growth areas in science. Scientometrics, 68, 595–610.

Takeda, Y. Kajikawa, Y. & Matsushima, K. (2007). Citation network of CVD research: research topics and journals. Chemical Vapor Deposition, 10, 523-525.

van Raan, A. F. J., (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics, 36, 397-420.

White, H.D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. Journal of the American Society for Information Science and Technology, 54, 423-434.