Searching for intellectual turning points: Progressive knowledge domain visualization

Chaomei Chen*

College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104-2875

This article introduces a previously undescribed method progressively visualizing the evolution of a knowledge domain's cocitation network. The method first derives a sequence of cocitation networks from a series of equal-length time interval slices. These time-registered networks are merged and visualized in a panoramic view in such a way that intellectually significant articles can be identified based on their visually salient features. The method is applied to a cocitation study of the superstring field in theoretical physics. The study focuses on the search of articles that triggered two superstring revolutions. Visually salient nodes in the panoramic view are identified, and the nature of their intellectual contributions is validated by leading scientists in the field. The analysis has demonstrated that a search for intellectual turning points can be narrowed down to visually salient nodes in the visualized network. The method provides a promising way to simplify otherwise cognitively demanding tasks to a search for landmarks, pivots, and hubs.

The primary goal of knowledge domain visualization (KDViz) is to detect and monitor the evolution of a knowledge domain (1). Progressive knowledge domain visualization is specifically concerned with techniques that can be used to identify temporal patterns associated with significant contributions as a domain advances.

Many aspects of a scientific field can be represented in the form of a scientific network, such as scientific collaboration networks (2), social networks of coauthorship (3), citation networks (4), and cocitation networks (5). Scientific networks constantly change over time. Some changes are relatively moderate; some can be dramatic. Understanding the implications of such changes is essential to everyone in a scientific field.

Researchers have been persistently searching for underlying mechanisms that may explain various changes and patterns in scientific networks. On the other hand, this is an ambitious and challenging quest because of the scale, diversity, and dynamic nature of scientific networks that one has to deal with. In this article, we introduce a previously undescribed method designed to reduce some of the complexities associated with identifying key changes in a knowledge domain. We focus on cocitation networks, although we expect that the method is applicable to a wider range of networks.

The key elements of the method draw their strength from a divide-and-conquer strategy. A time interval is divided into a number of slices, and an individual cocitation network is derived from each time slice. The time series of networks are merged. Major changes between adjacent slices are highlighted in a panoramic visualization of the merged network. The primary motivation of the work is to simplify the search for significant papers in a knowledge domain's literature so that one can search for visually salient features, such as landmark nodes, hub nodes, and pivot nodes, in a visualized network. The entire progressive visualization process is streamlined and implemented in a computer system of the author called CITESPACE.

The rest of this paper is organized as follows. We first review prior studies of the growth of a knowledge domain and then identify the key issues to be addressed by our method. The progressive visualization method is described and illustrated with an example in which we identify intellectual turning points in the field of superstring in theoretical physics. Identified articles associated with visually salient features are validated with the leading scientists in the field of superstring.

Related Work

Two strands of research are relevant. The focus of our research can be expressed in two key questions. How does a scientific field grow? What has been done for visualizing temporal patterns, especially in relation to network evolution?

Scientific Revolutions

The most widely known model of science is Thomas Kuhn's *Structure of Scientific Revolutions* (6), in which science is characterized by transitions from normal science to science in crisis and from crisis to a scientific revolution. Kuhn's theory suggests that scientific revolutions are a crucial part of science. The notion of paradigm shift is widely known in virtually all scientific disciplines. Kuhn's model has generated profound interest in detecting and monitoring paradigm shifts through the study of temporal patterns in cocitation networks. Small (7) identified and monitored the changes of research focus in collagen research in terms of how clusters of most cocited articles change over consecutive years. Small's study predated many modern visualization techniques. However, the representations of cocitation clusters were isolated from one year to another; significant temporal patterns or transitions may go unnoticed if they fall between the clusters from different years.

In our earlier work (8), we used animated visualization techniques to reconstruct citation and cocitation events in their chronological order so that one can examine the growth history of a domain in a broader context in a similar way to how we play a video in a fast-forward mode. The animated visualization enabled us to identify paradigm-like clusters of cocited articles corresponding to significant changes in the field of superstring, the same topic we will revisit with our method, but the visual features of some of the groundbreaking articles were not distinct enough to lend themselves to a simple visual search. Our earlier methodology did not include time slicing, multiple thresholding, and merging. One of our main objectives is therefore to improve visualization techniques so that groundbreaking articles can be characterized by distinguishable visual features.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviation: KDViz, knowledge domain visualization.

*E-mail: chaomei.chen@cis.drexel.edu.

The concept of a research front is also relevant to how science grows. A research front consists of transient clusters of most recently cited works in the literature of a scientific field (9). A research front represents the state of the art of a field, and research fronts move along with the underlying scientific field as new articles replace existing articles.

The recent interest in complex network analysis is a potentially fruitful route to improve our understanding of scientific networks as well as general networks (10). Studies in complex network analysis, especially in relation to small-world and scale-free networks, focus on two broad issues, namely topological properties and generative mechanisms of networks. Various growth models such as preferential attachment (10–12) have been developed in the study of network evolution. However, much of the work has concentrated on abstract network representations rather than on concrete networks and their practical implications. We emphasize the integral role of the semantics of such networks in understanding the profound dynamics of network evolution. We expect that the progressive method described in this article can provide a useful instrument for examining the evolution of a scientific network, and that the concrete example of network evolution can lead to insights into a broader range of networks.

Visualizing Temporal Patterns

A good example of visualizing thematic changes in a collection of text documents is THEMERIVER (13), which uses the metaphor of a thematic river to depict temporal changes of word frequencies. An intensified theme can be identified if one can detect increasingly widened word frequency streams. This is a relatively straightforward task, given that one needs only to tell how much the width of a stream changes over time. In contrast, it is often much more complicated to detect temporal patterns in higher-dimensional data or higher-order relationships.

A number of methods such as INDSCAL (14), PROCRUSTES analyses (15), and thin-plate splines and deformation analysis (16) can be used to compare dimensional representations. PROCRUSTES analysis, for example, aligns two configurations by stretching and rotating operations so that the remaining differences are where the two configurations really differ. Similarly, thin-plate display renders the difference between two configurations as a deformed plate. The degree of the deformation indicates the extent to which the two configurations differ. Such methods are efficient in detecting local and short-range discrepancies between two almost identical configurations, but the performance degrades if the discrepancies are long range in nature or a substantial part of the configurations is involved. In KDViz, we need to consider both short- and long-range changes between two adjacent snapshots of a domain, although few empirical studies have examined these techniques in the context of KDViz.

A network can change over time in various ways and can change its topology by adding new nodes and links as well as removing existing nodes and existing links. A network can also change the intrinsic attributes of its nodes and links; for example, citations of articles in a cocitation network tend to increase over time.

Much of the existing approaches to visualizing the evolution of a network falls into one of two categories: the slide-show and panorama approaches. Just like in a slide show, the former aims to highlight the changes as the viewer moves forward, sometime back and forth, in a time series of snapshots. The latter aims to pack synthesized temporal changes into a single image.

The slide-show approach has several advantages, including being easy to implement and flexible to use. This approach often provides additional visual aids to help viewers identify changes between adjacent snapshots. Recent examples include the visualization of how a discourse evolves as a network of words (17) and the visualization of semantic structures across different time planes (18). However, research in perceptual cognition has shown that comparing two images back and forth can be cognitively vey demanding and prone to error.

The panorama approach aims to depict temporal as well as spatial changes in such a way that viewers can detect a trend or a pattern by studying a single image. This approach could minimize the disturbance to the viewer's mental model (19). Related work in this area includes incremental graph drawing (20) and the timed network display function in PAJEK (21). Our earlier work on using animated visualization techniques to depict temporal changes in a cocitation network also belongs to this category (8).

Progressive Visualization Issues

A progressive visualization method aims to visualize the evolution of a network over time. The following three issues need to be addressed for visualizing time-sliced networks: (*i*) Improving the clarity of individual networks; (*ii*) highlighting transitions between adjacent networks; and (*iii*) identifying potentially important nodes.

The first issue is concerned with the clarity of individual networks' representations. One of the major aesthetic criteria established by research in graph drawing is that link crossings should be avoided whenever possible. A network visualization with the least number of edge crossings is regarded as not only aesthetically pleasing but also more efficient to work with in terms of the performance of relevant perceptual tasks (22). The number of link crossings may be reduced by pruning various links in a network. Minimum spanning trees and Pathfinder network scaling are commonly used algorithms. The major advantages and disadvantages of these scaling techniques are further analyzed below.

The second issue is concerned with progressively merging two adjacent networks, so that one can identify which part of the earlier network is persistent in the new network, which part of the earlier network is no longer active in the new network, and which part of the new network is completely new. Much of the novelty of our method is associated with the way we address this issue.

The third issue is concerned with the role of visually salient features in simplifying search tasks for intellectual turning points. Visually salient nodes include landmark nodes, pivot nodes, and hub nodes.

Issue 1: Improving the Clarity of Networks

Cocitation networks often have a vast number of links, and displaying links indiscriminately is the primary cause of clutter. There are two general approaches to reduce the number of links in a display: threshold- and topology-based approaches. In the threshold-based approach, the elimination of a link is determined solely by whether the link's weight exceeds a threshold. In contrast, in a topology-based approach, the elimination of a link is determined by a more extensive consideration of intrinsic topological properties; therefore, such approaches tend to preserve certain topological intrinsic properties more reliably, although the computational complexity tends to be higher.

Pathfinder network scaling is originally developed by cognitive scientists to build procedural models based on subjective ratings (23–25). It uses a more sophisticated link-elimination mechanism compared to minimum spanning tree (MST) and can remove a large number of links and retain the most important ones. Given a network, one can derive a unique Pathfinder network that contains all of the alternative MSTs of the original network. MST is increasingly a strong candidate in a series of KDViz studies (8, 26–28).

The goal of Pathfinder network scaling, in essence, is to prune a dense network. The topology of a Pathfinder network is

determined by two parameters, r and q. The r parameter defines a metric space over a given network based on the Minkowski distance so that one can measure the length of a path connecting two nodes in the network. The Minkowski distance becomes the familiar Euclidean distance when $r = 2$. When $r = \infty$, the weight of a path is defined as the maximum weight of its component links, and the distance is known as the maximum value distance.

Given a metric space, a triangle inequality can be defined as follows,

$$w_{ij} \leq (\Sigma_k w^r n_k n_{k-1})^{1/r},$$

where w_{ij} is the weight of a direct path between i and j, $wn_k n_{k+1}$ is the weight of a path between n_k and n_{k+1}, for $k = 1, 2, \ldots, m$. In particular, $i = n_1$ and $j = n_k$. In other words, the alternative path between i and j may go all the way around through nodes $n_1, n_2, \ldots, n_k$, so long as each intermediate links belong to the network.

If w_{ij} is greater than the weight of alternative path, then the direct path between i and j violates the inequality condition. Consequently, the link *i-j* will be removed, because it is assumed that such links do not represent the most salient aspects of the association between the nodes i and j.

The q parameter specifies the maximum number of links that alternative paths can have for the triangle inequality test. The value of q can be set to any integer between 2 and $N - 1$, where N is the number of nodes in the network. If an alternative path has a lower cost than the direct path, the direct path will be removed. In this way, Pathfinder reduces the number of links from the original network, whereas all of the nodes remain untouched. The resultant network is also known as a minimum-cost network.

The strength of Pathfinder network scaling is its ability to derive more accurate local structures than other comparable algorithms, such as multidimensional scaling and minimum spanning tree. However, the Pathfinder algorithm is computationally expensive; the published algorithm is in the class of $O(N^4)$. KDViz approaches built on the Pathfinder network scaling algorithm have a potential bottleneck if one needs to deal with large networks. The maximum pruning power of Pathfinder is achievable with $q = N - 1$ and $r = \infty$; not surprisingly, this is also the most expensive one, because all of the possible paths must be examined for each link. In addition, the algorithm requires a large amount of memory to store the intermediate distance matrices. This is the first of the three issues our method is to deal with. The method follows a divide-and-conquer strategy.

Issue 2: Merging Heterogeneous Networks

The second issue identified above is concerned with progressively merging two temporally adjacent networks. Depending on the nature of a knowledge domain, networks to be merged could be heterogeneous as well as homogeneous in terms of intrinsic topological properties and additional attributes of nodes and links. For example, intellectual structures of a knowledge domain before and after a major conceptual revolution are likely to be fundamentally different as new theories and evidence become predominant. Cocitation networks of citation classics in a field are likely to differ from cocitation networks of newly published articles. The key question is, what is the most informative way to merge potentially diverse networks?

A merged network needs to capture important changes over time in a knowledge domain's cocitation structure. We need to find when and where the most influential changes took place so that the evolution of the domain can be characterized and visualized. Few studies in the literature investigated network merge from a domain-centric perspective. The central idea of our method is to visualize how different network representations of an underlying phenomenon can be informatively stitched together.

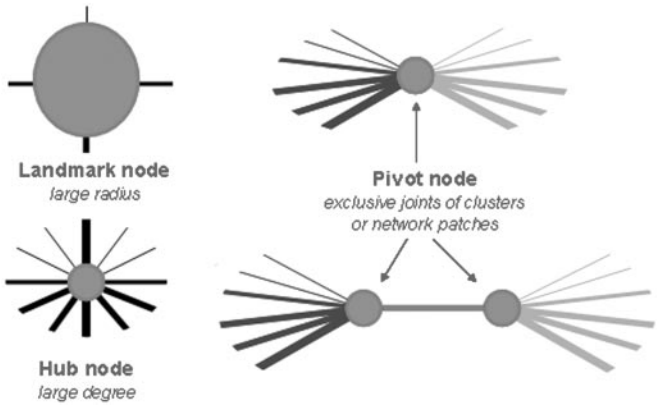

Fig. 1. Three types of visually salient nodes in a cocitation network.

Issue 3: Visually Salient Nodes in Merged Networks

The third issue addressed by our method is concerned with the identification of potentially important articles in a cocitation network. The importance of a node in a cocitation network can be quickly identified by the local topological structure of the node and by additional attributes of the node. We are particularly interested in three types of nodes: (*i*) landmark, (*ii*) hub, and (*iii*) pivot nodes (see Fig. 1).

A landmark node has extraordinary attribute values. For example, a highly cited article tends to provide an important landmark regardless of how it is cocited with other articles. Landmark nodes can be rendered by distinctive visual–spatial attributes such as size, height, or volume. A hub node has a relatively large node degree; a widely cocited article is a good candidate for significant intellectual contributions. A high-degree hub-like node is also easy to recognize in a visualized network. Both landmark and hub nodes are commonly used in network visualization. Although the concept of pivot nodes is available in various contexts, the way they are used in our method is previously undescribed. Pivot nodes are joints between different networks; they are either the common nodes shared by two networks or the gateway nodes that are connected by internetwork links. Pivot nodes have an essential role in our method.

Methods

The method includes the following procedural steps: time slicing, thresholding, modeling, pruning, merging, and mapping. Although pruning is not always necessary, it is a potentially valuable option when dealing with a dense network. All steps are implemented in CITESPACE.

Procedure. The input to CITESPACE is a set of bibliographic data files in the field-tagged Institute for Scientific Information[†] Export Format. The outputs of CITESPACE include visualized cocitation networks; each network is shown in a separate interactive window interface.

Time Slicing. The entire time interval can be sliced into equal-length segments. The length of each segment can be as short as a year or as long as the entire interval. If appropriate data

[†]These data are extracted from *Science Citation Index Expanded* [Institute for Scientific Information, Inc. (ISI), Philadelphia, PA; Copyright ISI]. All rights reserved. No portion of these data may be reproduced or transmitted in any form or by any means without the prior written permission of ISI.

become available, it is possible to slice the data thinner to make monthly or weekly segments. Currently, sliced segments are mutually exclusive, although overlapping segments could be an interesting alternative worth exploring.

Thresholding. Citation and cocitation analysis typically sample the most highly cited work, the cream of crop, with a single constant threshold. However, a single constant threshold is a crude sampling mechanism if the citation patterns over an extended period are being considered. By default, both citations and cocitations are calculated within each time slice, as opposed to across all time slices.

Time slicing provides the flexibility to tailor a threshold more closely to the characteristics of citation and cocitation activities in each individual time slice. This flexibility is expected to reduce the bias associated with a single one-size-fits-all threshold. One can even compare and merge two very different networks within this framework, for example, a network of articles from Nobel Prize-winning scientists and a network of technical reports. The key questions are: what is the common ground between two networks? How can one extract insights into the internetwork relationship from such common ground? A flexible threshold configuration can find a common ground more easily.

The cocitation network in a given time slice is determined by three thresholds: citation, cocitation, and cosine coefficient thresholds. In CITESPACE, the user needs to select desired thresholds for three specific time slices, namely the beginning, middle, and ending slices. CITESPACE automatically assigns interpolated thresholds to the remaining slices. In practice, the user starts with an arbitrary threshold configuration and then adjusts thresholds accordingly based on the reported statistics such as the citation population and the numbers of nodes and links in a network.

In the citation world, articles are not created equal. Some articles have much more than their fair share of citations, some have less, and some have none at all. Citations depend on many underlying factors. For example, success breeds success; a highly cited article is likely to receive more citations than a currently less frequently cited article. To detect intellectual turning points, we are particularly interested in articles that have rapidly growing citations. In the following superstring example, we use a simple model to normalize the citations of an article within each time slice by the logarithm of its publication age, the number of years elapsed since its publication year. The rationale is to highlight articles that increased most in the early years of publication. More sophisticated models can be derived based on citation distribution models of a given dataset and a model of the growth and decay of scientific citations (29). Building such models is significant and challenging in its own right.

Modeling. By default, cocitation counts are calculated within each time-sliced segment. Cocitation counts are normalized as cosine coefficients, $cc_{cosine}[i, j] = cc[i, j]/sqrt(c[i], c[j])$, where $cc[i, j]$ is the cocitation count between documents i and j, and $c[i]$ and $c[j]$ are their citation counts, respectively. The user can specify a selection threshold for cocitation coefficients; the default value is 0.15.

Alternative measures of cocitation strengths are available in the information science literature, such as Dice and Jaccard coefficients. In earlier studies, we used Pearson's correlation coefficients. Recently, researchers began to examine how Pearson's correlation coefficients transform the underlying structure of a cocitation network (30), but available evidence is still inconclusive (31, 32). Although the impact of various cocitation metrics on the resultant network visualizations is worth pursuing, the topic is beyond the scope of this article.

Pruning. Effective pruning can reduce link crossings and improve the clarity of the resultant network visualization. CITESPACE supports two common network-pruning algorithms, namely Pathfinder and minimum spanning tree. The user can select to prune individual networks only or the merged network only or to prune both. Pruning increases the complexity of the visualization process. In the following section, visualizations with local pruning and global pruning are presented.

In this article, we concentrate on Pathfinder-based pruning. To prune individual networks with Pathfinder, the parameters q and r were set to N_k - 1 and ∞, respectively, to ensure the most extensive pruning effect, where N_k is the size of the network in the kth time slice. For the merged network, the q parameter is (ΣN_k)- 1, for $k = 1, 2$.

Merging. The sequence of time-sliced networks is merged into a synthesized network, which contains the set union of all nodes ever to appear in any of the individual networks. Links from individual networks are merged based on either the earliest establishment rule or the latest reinforcement rule. The earliest establishment rule selects the link that has the earliest time stamp and drops subsequent links connecting the same pair of nodes, whereas the latest reinforcement rule retains the link that has the latest time stamp and eliminates earlier links.

By default, the earliest establishment rule applies. The rationale is to support the detection of the earliest moment when a connection was made in the literature. More precisely, such links mark the first time a connection becomes strong enough with respect to the chosen thresholds.

Mapping. The layout of each network, either individual time-sliced networks or the merged one, is produced by using Kamada and Kawai's algorithm (33). The size of a node is proportional to the normalized citation counts in the latest time interval. Landmark nodes can be identified by their large discs. The label size of each node is proportional to citations of the article, thus larger nodes also have larger-sized labels. The user can enlarge

Table 1. Time slicing and threshold settings for a set of small networks, where f_c is the citation frequency threshold and f_{cc} is the cocitation frequency threshold

Time slices	f_c	f_{cc}	Cite space size	Top cited	Sample, %	cc (cosine ≥0.15)
1985–1987	3	1	604	16	2.65	58
1988–1990	10	3	2,740	15	0.55	30
1991–1993	50	7	12,214	18	0.15	62
1994–1996	60	10	16,147	19	0.12	53
1997–1999	80	10	19,716	20	0.10	60
2000–2002	85	15	22,449	20	0.09	54
2003	25	10	9,594	13	0.14	34
Total (unique)			83,464	121 (82)	Mean 0.54	Total 351

cc, cocitation.

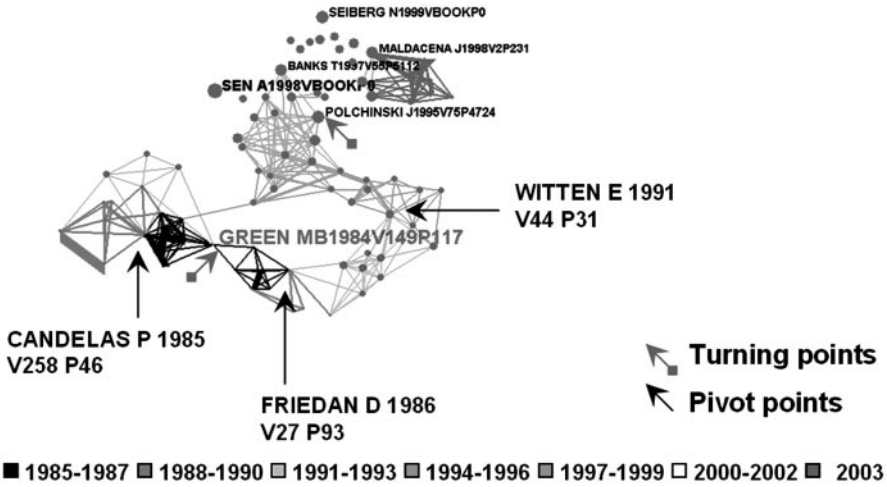

Fig. 2. An 82-node merged network without global pruning. See a color version at www.pages.drexel.edu/~cc345/citespace/Figure 2.png.

font sizes at will, and both the width and the length of a link are proportional to the corresponding cocitation coefficient. The color of a link indicates the earliest appearance time of the link with reference to chosen thresholds.

Visually salient nodes such as landmarks, hubs, and pivots are easy to detect by visual inspection. CITESPACE currently does not include any algorithms to detect such nodes computationally. Instead, the visual effect is a natural result of slicing and merging, although additional computational metrics may enhance the visual features even further. A useful computational metric should reflect the degree of a node, and it should also take into account the heterogeneity of the node's links. The more dissimilar links a node connects to others, the more likely the node has a pivotal role to play. In the following example, we consider only nodes that have a degree of 10 or higher for visual inspection.

Superstring. The method is applied to the visualization of how cocitation networks of superstring in theoretical physics evolved over time. Two superstring revolutions are documented over the last two decades: one in the mid-1980s and one in the mid-1990s (34). We reported animated visualizations of the superstring cocitation networks by using a single constant citation threshold, and Pearson's correlation coefficients were used to measure the strength of each cocitation link. The changes of the cocitation network were animated by the growing height of a citation bar and as a state transition process. Although key articles to the revolutions were identifiable in the resultant animations, they did not quite lend themselves to a simple visual inspection. We expect that the progressive visualization method can make it more easily to identify intellectual turning points by visual inspection. The superstring dataset in this study is updated to include citation data between 1985 and 2003.

Visualized networks were validated by the leading scientists in the field of superstring. We showed the merged map, without pruning, to John Schwarz (California Institute of Technology, Pasadena) and Edward Witten (Princeton University, Princeton). Schwarz is the coauthor of the article that triggered the first superstring revolution; Witten has written a number of highly cited articles on superstring and is also the top-ranked physicist in a list of the 1,000 physicists most cited between 1981 and 1997. The list was compiled by the Institute for Scientific Information, who were asked to explain the nature of intellectual contributions identified by pivot points and hubs in the networks.

The 19-year time interval was sliced into six 3-year segments, starting from 1985–1987 and ending with 2000–2002, plus a 1-year segment for 2003. Two sets of results were generated from two separate runs: one used relatively higher-threshold settings, which resulted in small networks (Table 1); the other used lower-threshold settings for larger networks (Table 2). Two versions of the larger network are shown: one without global pruning (Fig. 3) and the other with global pruning (Fig. 4). Links were color-coded by the earliest establishment rule. Darker colors indicate links from earlier time slices, whereas lighter colors indicate links from more recent slices. Networks in individual time slices are not shown due to page limitations.

Results

Table 1 shows the size of the cite space and details of individual networks and the merged network. The size of the cite space in a given time slice is the number of articles that have at least one citation within the given time slice; the size is generally increasing over time. The size for 2003 is smaller, because the 2003 data are still incomplete. The merged network contains 82 articles, and various pivot points are evident at a glance (Fig. 2).

Table 2 shows the threshold setting for a sequence of larger networks. The cocitation network in each time slice represents approximately the top 1% most cited articles. The merged network contains 647 unique articles, which collectively made 1,097 appearances in these time slices. In other words, 41% of articles appeared in more than one time slice. The locally pruned version of the merged network is shown in Fig. 3; the globally pruned version is shown in Fig. 4.

As shown in Fig. 3, color-coded links in effect partitioned the merged network into several major clusters of articles. Clusters of the same color represent cocitations made within the same time slice. More importantly, as we expected, within-cluster cocitation links are evidently more common than between-cluster links. A strongly clustered network also makes it easy to identify pivot nodes and between-cluster links. Six structurally strategic nodes are identified in Fig. 3, including the 1984 Green–Schwarz article, which triggered the first superstring revolution. However, the 1995 Polchinski article that triggered the second superstring revolution was not obvious in the dense visualization; Polchinski introduced the fundamental concept of D-branes in that article.

The 1984 Green–Schwarz article is a typical pivot node; it is the only contact point between two densely connected clusters in blue (1985–1987). It was this article that sparked the first superstring revolution, the famous 1984 Green–Schwarz anomaly cancellation paper. Friedan's 1986 article is a distinct pivot node connecting blue (1985–1987), pink (1988–1990), and green clusters (1991–1993). Witten's 1986 article is a pivot between a

Table 2. Time slicing and threshold settings for a set of larger networks, where f_c is the citation frequency threshold and f_{cc} is the cocitation frequency threshold

Time slices	f_c	f_{cc}	Cite space size	Top cited	Sample, %	cc (cosine ≥0.15)
1985–1987	2	1	604	39	6.46	229
1988–1990	4	3	2,740	114	4.16	283
1991–1993	15	7	12,214	200	1.64	1,263
1994–1996	20	10	16,147	229	1.42	895
1997–1999	25	10	19,716	223	1.13	956
2000–2002	30	15	22,449	180	0.80	486
2003	10	10	9,594	112	1.17	131
Total (unique)			83,464	1,097 (647)	Mean 2.4	Total 4,243

cc, cocitation.

blue cluster (1985–1987) and a yellow cluster (2000–2002). Fig. 3 also contains a couple of smaller clusters that are completely isolated from the main super cluster. Small clusters in red (2003) indicate the candidates for emerging clusters. We were able to find Polchinski's 1995 article in a smaller-sized merged network, but the article must be overwhelmed by the 4,000 strong links of the larger network. Nevertheless, the quality of the visualized network is promising: intellectually significant articles tend to have topologically unique features.

Articles by Maldacena, Witten, and Gubser–Klebanov–Polyako, located toward the top of the major network component, were all published in 1998. When we asked Witten to comment on an earlier version of the map, in which citation counts were not normalized by years since publication, he indicated that the Green–Scharwz article is more important to the field than the three top-cited ones, and that the earlier articles in the 1990s appeared to be underrepresented in the map. The apparent mismatch between citation frequencies of nodes and their importance judged by domain experts was partially corrected in the network shown in Fig. 2. Witten's comments raised an important question: is it possible that an intellectually significant article may not always be the most highly cited?

Fig. 4 shows the merged network pruned by Pathfinder; the pruned version contains fewer links than the version in Fig. 3. Much of the within-cluster links is reduced to links between cluster centers and other cluster members. Links between non-center members are essentially removed. The overall structure is simpler and easier to explore. In addition, the number of link colors attached to a node distinguishes a pivot node from a nonpivot node. If a node connects to other nodes through links in a single color, it is not regarded as a pivot node, because it does not imply intellectual transitions over time. In contrast, if a node joins several different-colored links, it is a good candidate for an intellectual turning point, because if paths connecting articles in different clusters must go through a pivot point, the pivot point is likely to have a unique position in the literature.

The Green–Schwarz article is located toward the center of the visualization; it joins links from four different time slices. The 1995 article by Candelas *et al.* is similar in terms of the link colors. According to Institute for Scientific Information's Sci-

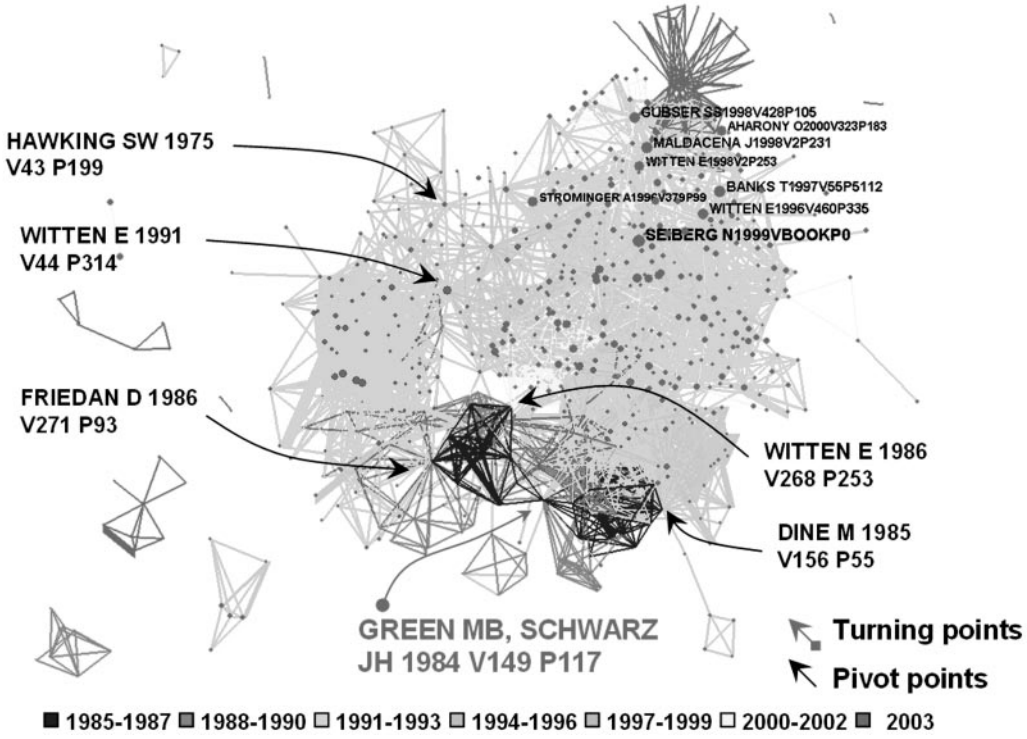

Fig. 3. A 624-node merged network without global pruning. See a color version at: www.pages.drexel.edu/~cc345/citespace/Figure3.png.

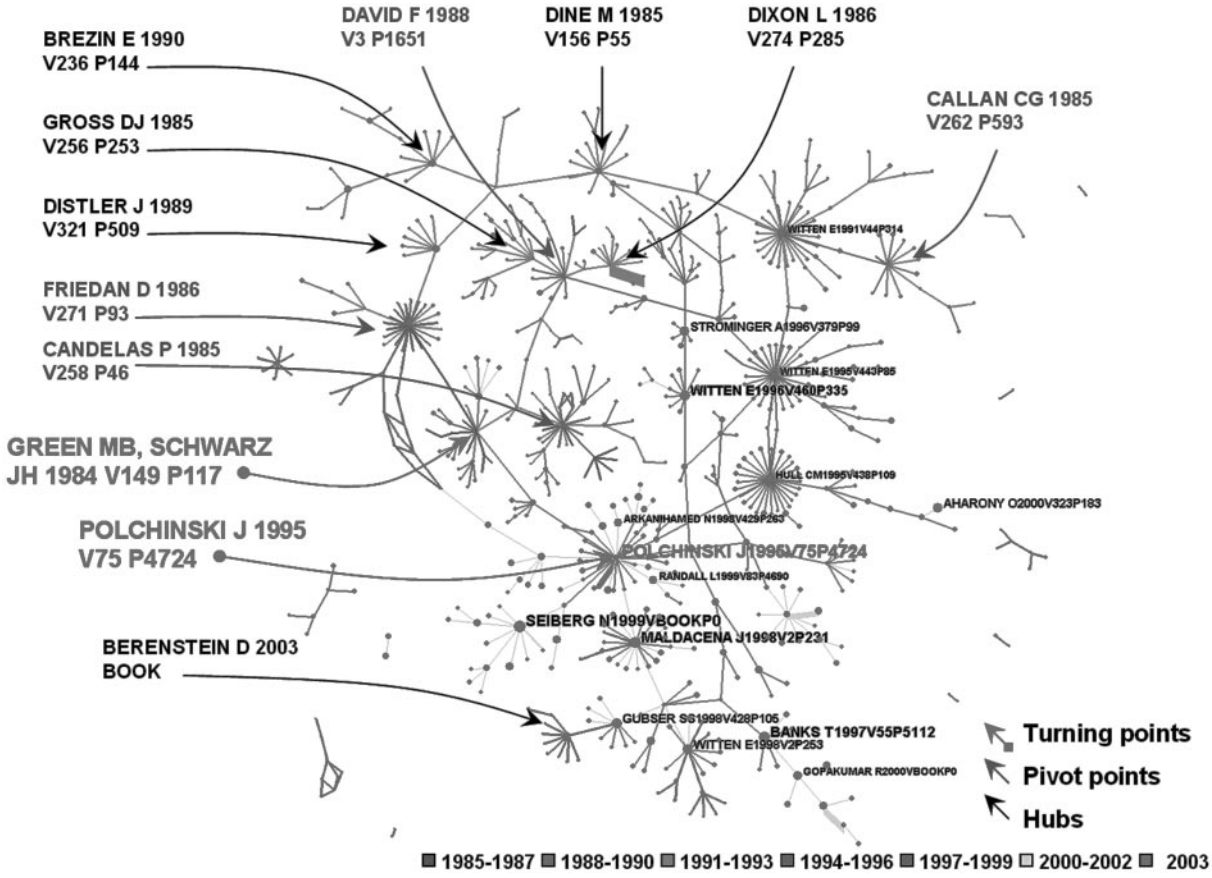

Fig. 4. A 624-node merged network with global pruning by using Pathfinder ($q = N - 1$, $r = \infty$). See a color version at www.pages.drexel.edu/~cc345/citespace/Figure4.png.

ence Citation Report (1981–1998), Candelas *et al.* 's article has a total of 1,538 citations during that period, and its average annual citation is 110. The 1995 Polchinski article can be easily found at the lower center of the map; according to Schwarz, it explained the concept of D-branes, a crucial ingredient in almost all modern string theory research. It appears to be more of a hub than a pivot node, connected by links of only two colors, brown (1997–1999) and yellow (2000–2002). To the left of Polchinski's article is a cluster centered by the 1998 Maldacena article. Schwarz noted that in this article, Maldacena made a major new discovery that in certain circumstances relates string theories to quantum field theories.

The comments from domain experts have confirmed that both versions of the merged network indeed highlight significant articles, and these articles tend to have unique topological properties that distinguish them from other articles. The globally pruned version is easier to explore than the local-pruning-only version.

Discussion

The results are particularly encouraging because the presence of pivot nodes enables us to narrow down the visual search quickly to a small number of good candidate nodes for intellectual turning points. An easy identification of such turning points is an important and necessary step toward effective detection of paradigmatic changes in a knowledge domain. The small network is particularly clear, containing both turning points. The larger network without pruning is cluttered, although it is still possible to identify several pivot points.

The interpretation and validation of the visualizations have greatly benefited from help from leading scientists in the knowledge domain. The work has also shown that using a variable threshold could be a potentially good practice for citation analysis in general.

In comparison to our earlier visualization of the superstring cocitation network (8), this method tends to produce more distinct visual features for key articles. More importantly, such visual features appear to be independent to the amount of citations of a node. In other words, a lower citation rate is not necessarily preventing a node from having salient visual features, suggesting that cocitations must have played a greater role. Pivot nodes can be identified even if they have relatively fewer citations. This could be a particularly useful feature for the detection of significant articles that could be easily overlooked by falling below a single high-citation threshold.

The 1984 Green–Schwarz article for the first revolution is a typical pivot node, whereas the 1995 Polchinski article for the second revolution is more of a hub than a pivot node. This finding suggests that before we have further evidence, it would be sensible to examine both types of visualizations, pruned and unpruned, in a study of intellectual turning points.

In comparison to other methods for detecting changes of networks over time, our approach simplifies cognitively demanding tasks of comparing a sequence of network snapshots. The progressive visualization method allows us to focus on much simpler tasks of locating pivot nodes and cluster centers. The color-coded links enables the user to trace temporal patterns through the network visualization.

The progressive visualization method introduced here has practical implications. It provides scientists with a roadmap of their own field. Witten commented, "It was fun to look at it." A longstanding challenge is to be able to visualize cocitation networks of a domain as quickly as new bibliographic data become available so that one can monitor the changes of a domain more closely on a monthly or even weekly basis. The approach provides a practical starting point. Users have the flexibility to slice a time interval into smaller as well as larger segments.

Using overlapped time slices could be a valuable alternative to explore in future studies. Currently, adjacent time slices are mutually exclusive to highlight the magnitude of a potentially important change, whereas overlapping slice segments may blur such changes and make them less obvious to detect.

An unsolved issue is concerned with the detection of abrupt changes in citations within a short period. We normalized the citations of an article by its publication age. Additional metrics of pivot nodes should augment the power of visual inspection even further. Knowledge discovery and data-mining techniques, such as Kleinberg's burst-detection technique (35), are expected to play a substantial role in identifying a paradigm shift.

Finally, the role of domain experts in KDViz needs to be further investigated. Experts in the fields are the best sources to seek validations and interpretations. On the other hand, one should also use domain visualizations with caution; and it should be made clear that algorithmically generated domain visualizations, however crafted, merely portray the complexity of an underlying domain from a limited perspective. If KDViz can stimulate scientists to look at their own field from a different perspective and pose new questions about the evolution of their domain, KDViz will ultimately become a practical tool to study science itself.

Conclusion

The progressive KDViz method simplifies the tasks of tracking significant changes of a knowledge domain's cocitation network over time. Cognitively demanding tasks of comparing complex networks back and forth are simplified to tasks of locating pivot points and cluster centers in visualized networks.

The divide-and-conquer strategy maximizes the strengths of algorithms and reduces the influence of their weaknesses. The cosine cocitation coefficients are effective enough to pick up the most intellectually significant articles, whereas the Pathfinder-enhanced version improved the quality even further.

CITESPACE provides an experiment platform to investigate new ideas and compare existing approaches. We plan to make a further refined version of CITESPACE available in the near future to researchers, practitioners, and educators in various disciplines and obtain their first-hand experience in capturing the changes of their own domains.

Further studies and in-depth case studies of progressive KDViz should be encouraged. For example, can this method detect the merge of two domains or the split of a single domain into a few new ones? Can this method detect scientific revolutions in other disciplines? Will it work with alternative representations of a knowledge domain, such as the preprint archives used by physicists and other sources? KDViz is a challenging route, but it is also potentially rewarding for scientists in so many different knowledge domains to have easy access to the big picture of their own fields.

We give special thanks to John Schwarz (California Institute of Technology) and Edward Witten (Princeton University) for help in interpreting the visualizations. The 2002 Institute for Scientific Information/American Society for Information Science and Technology Citation Analysis Research award is acknowledged.

1. Chen, C. (2003) *Mapping Scientific Frontiers: The Quest for Knowledge Visualization* (Springer, London).
2. Barabási, A.-L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A. & Vicsek, T. (2002) *Phys. A* **311,** 590–614.
3. Newman, M. E. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 404–409.
4. Garfield, E. & Small, H. (1989) in *Innovation: At the Crossroads Between Science and Technology*, eds. Kranzberg, M., Elkana, Y. & Tadmor, Z. (Neaman, Haifa), pp. 51–65.
5. Small, H. & Greenlee, E. (1989) *Commun. Res.* **16,** 642–666.
6. Kuhn, T. S. (1962) *The Structure of Scientific Revolutions* (Univ. of Chicago Press, Chicago).
7. Small, H. G. (1977) *Soc. Stud. Sci.* **7,** 139–166.
8. Chen, C. & Kuljis, J. (2003) *J Am. Soc. Inf. Sci. Technol.* **54,** 435–446.
9. Price, D. D. (1965) *Science* **149,** 510–515.
10. Albert, R. & Barabasi, A. (2002) *Rev. Mod. Phys.* **74,** 47–97.
11. Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. (2000) *Phys. Rev. Lett.* **85,** 4633–4636.
12. Newman, M. (2001) *Phys. Rev. E* **64**.
13. Havre, S., Hetzler, E., Whitney, P. & Nowell, L. (2002) *IEEE T. Vis. Comput. Graphics* **8,** 9–20.
14. Carroll, J. D. & Chang, J.-J. (1970) *Psychometrika* **35,** 283–319.
15. Gower, J. C. (1975) *Psychometrika* **40,** 33–51.
16. Bookstein, F. L. (1989) *IEEE T. Pattern Anal.* **11,** 567–585.
17. Brandes, U. & Corman, S. R. (2003) *Inf. Visual.* **2,** 40–50.
18. Erten, C., Harding, P. J., Kobourov, S. G., Wampler, K. & Yee, G. (2003) *Technical Report TR0304* (Univ. of Arizona, Tucson, AZ).
19. Misue, K., Eades, P., Lai, W. & Sugiyama, K. (1995) *J. Visual Lang. Comput.* **6,** 183–210.
20. North, S. C. (1995) in *Proceedings of Graph Drawing* (GD'95), ed. Brandenburg, F. J. (Springer, New York), pp. 409–418.
21. Batagelj, V. & Mrvar, A. (1998) *Connections* **21,** 47–57.
22. Ware, C., Purchase, H., Colpoys, L. & McGill, M. (2003) *Inf. Visual.* **1,** 103–110.
23. Schvaneveldt, R. W. (1990) *Pathfinder Associative Networks* (Ablex, Norwood, NJ).
24. Chen, C. (1999) *Inform. Process. Manag.* **35,** 401–420.
25. Chen, C. & Paul, R. J. (2001) *Computer* **34,** 65–71.
26. Chen, C., Cribbin, T., Macredie, R. & Morar, S. (2002) *J Am. Soc. Inf. Sci. Technol.* **53,** 678–689.
27. Chen, C., Kuljis, J. & Paul, R. J. (2001) *IEEE T. Syst. Man. Cy. C* **31,** 518–529.
28. White, H. (2003) *J Am. Soc. Inf. Sci. Technol.* **54,** 423–434.
29. van Raan, A. (2000) *Scientometrics* **47,** 347–362.
30. Ahlgren, P., Jarneving, B. & Rousseau, R. (2003) *J Am. Soc. Inf. Sci. Technol.* **54,** 550–560.
31. White, H. D. (2003) *J Am. Soc. Inf. Sci. Technol.* **54,** 1250–1259.
32. Chen, C. & Morris, S. (2003) in *IEEE Symposium on Information Visualization* (InfoVis'03) (IEEE Computer Society Press, Seattle), pp. 67–74.
33. Kamada, T. & Kawai, S. (1989) *Inform. Process. Lett.* **31,** 7–15.
34. Schwarz, J. H. (1996) arXiv:hep-th/9607067 (http://arxiv.org/PS_cache/hep-th/pdf/9607/9607067.pdf).
35. Kleinberg, J. (2002) in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp. 91–101.