Visualizing Evolving Networks: Minimum Spanning Trees versus Pathfinder Networks

Chaomei Chen
College of Information Science and Technology
Drexel University
chaomei.chen@cis.drexel.edu

Steven Morris
Electrical and Computer Engineering
Oklahoma State University
samorri@okstate.edu

Abstract

Network evolution is a ubiquitous phenomenon in a wide variety of complex systems. There is an increasing interest in statistically modeling the evolution of complex networks such as small-world networks and scale-free networks. In this article, we address a practical issue concerning the visualization of network evolution. We compare the visualizations of co-citation networks of scientific publications derived by two widely known link reduction algorithms, namely minimum spanning trees (MSTs) and Pathfinder networks (PFNETs). Our primarily goal is to identify the strengths and weaknesses of the two methods in fulfilling the need for visualizing evolving networks. Two criteria are derived for assessing visualizations of evolving networks in terms of topological properties and dynamical properties. We examine the animated visualization models of the evolution of botulinum toxin research in terms of its co-citation structure across a 58-year span (1945-2002). The results suggest that although high-degree nodes dominate the structure of MST models, such structures can be inadequate in depicting the essence of how the network evolves because MST removes potentially significant links from high-order shortest paths. In contrast, PFNET models clearly demonstrate their superiority in maintaining the cohesiveness of some of the most pivotal paths, which in turn make the growth animation more predictable and interpretable. We suggest that the design of visualization and modeling tools for network evolution should take the cohesiveness of critical paths into account.

CR Categories: I.3.6 [Methodology and Techniques]; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism – Virtual Reality; E.1 [DATA STRUCTURES] -- Graphs and networks.

Keywords: Network evolution, network visualization, co-citation networks, Pathfinder networks, minimum spanning trees.

1 Introduction

The significance of understanding the evolution of a complex network is widely recognized. For example, recent research in complex network theory has focused on statistical mechanisms that govern the growth of small-world networks [Watts and Strogatz 1998] and scale-free networks [Barabási et al. 2000]. Scale-free networks are characterized by a power law degree distribution. A major concern is how to simulate the evolution of a network that demonstrates such special topological properties so that one can improve the understanding of real-world networks. Few empirical studies have examined changes in the topological properties of a network over time.

Visualizing fundamental changes in scientific networks is one of the toughest challenges for research in information technology. The shortage of comprehensive examinations of the evolution of citation networks is due to various reasons, including the lack of an overarching framework that accommodates underlying theories and system functionalities across relevant disciplines, the lack of integrated network analysis and visualization tools, the lack of widely accessible longitudinal citation network data, and the lack of tools that specifically facilitate the analysis of network evolution.

Network visualization has a long history in information visualization, such as, SemNet [Fairchild et al. 1988], ConeTree [Robertson et al. 1991], NicheWorks [Wills 1999], and Hyperbolic Browser [Lamping and Rao 1996]. Researchers are increasingly interested in visualizing emerging patterns in association with evolving information structures, using tools such as Disk Trees and Time Tubes [Chi et al. 1998] and Botanical trees [Kleiberg et al. 2001].

A common problem with visualizing a complex network is that a large number of links may prevent users from recognizing salient structural patterns. A practical strategy is to reduce the number of links shown. There are several link reduction algorithms. The question is which one preserves the underlying topological properties best. Furthermore, as far as an evolving network is concerned, the resultant network should also preserve dynamical properties that characterize the evolution.

In this article, we study the role of two link reduction algorithms in visualizing the evolution of networks. A minimum spanning tree (MST) is widely known and commonly used in information visualization. On the other hand, Pathfinder network scaling is a procedural modeling algorithm originally developed by cognitive psychologists to capture salient relationships between concepts [Schvaneveldt 1990]. The strengths of such relationships are typically measured by human experts' subjective ratings of how similar those concepts are. Prior studies exclusively used Pathfinder networks to represent interrelations between concepts or keywords. Our earlier work has extended the use of Pathfinder networks to a much richer range of applications, especially co-citation networks [Chen 1998; Chen and Paul 2001]. In fact, an MST is a special case of a Pathfinder network because a Pathfinder network is the set union of all the possible MSTs derived from a network [Schvaneveldt 1990].

IEEE Symposium on Information Visualization 2003,
October 19-21, 2003, Seattle, Washington, USA
0-7803-8154-8/03/$17.00 ©2003 IEEE

Pathfinder networks have demonstrated various useful features in co-citation studies [Chen 2002; White 2003]. However, the Pathfinder network-scaling algorithm has its limitations. In order to achieve a network of high clarity and legibility, it is necessary to impose the so-called triangular inequality throughout the network. While this requirement leads to the simplest representation of the essence of an underlying proximity network, this is at a considerable computational cost. Additionally, as the size of the original network increases, the algorithm requires a considerable amount of memory to run. Therefore, it would be desirable if either an equivalent but more efficient algorithm can be developed, or a hybrid approach can be used to achieve cost-effectiveness. In contrast, MST algorithms such as Kruskal's algorithm and Prim's algorithm can be efficiently implemented, but may not capture local structures as accurate as Pathfinder. Now the question is how these properties influence the visualized network evolution. To our knowledge, this issue has not been specifically addressed.

In this article, we aim to address a number of issues concerning visualizing the evolution of a network with special reference to the use of MST and PFNET. 1) What should be a preferable topological structure of a visualized network? 2) What are the additional criteria for visualizing the evolution of a network? 3) To what extend can MST and PFNET be expected to meet such criteria? 4) What are the implications of our finding on visualizing the evolution of a network in general? The rest of the article is organized as follows. Related work is outlined first. Criteria are derived in terms topological properties and dynamical properties. Then we examine these criteria in MST and PFNET versions of animated visualizations of co-citation networks in botulinum toxin research between 1945 and 2002. The results are analyzed and their implications for further research are discussed.

2 Network Visualization

Graphically representing nodes and links is the most commonly used approach to network visualization. Much of the attention in graph drawing has been given to the efficiency of algorithms and the clarity of end results.

2.1 Link Reduction

The most widely known graph drawing techniques include force-directed graph drawing algorithms and spring-embedder algorithms [Eades 1984]. The primary goal of these algorithms is to optimize the arrangement of nodes of a network algorithmically, such that nodes connected by strong links in a graph-theoretical model appear close to each other in the final geometric representation, and weakly connected nodes appear far apart. Force-directed algorithms often lead to node placements that are aesthetically appealing. These algorithms, however, face some challenges in terms of efficiency, especially in terms of scalability, which is closely related to the clarity of a visualized network.

Cluttered network visualizations should be avoided whenever possible. An excessive number of links in a display may severely obscure the discovery of essential patterns. A commonly used strategy to reduce clutter is to reduce the number of links. There are several ways to achieve this goal. Three popular ones are analyzed below.

The first option is imposing a link weight threshold and only include links with weights above the threshold [Zizi and Beaudouin-Lafon 1994]. This approach is straightforward and easy to implement. However, it does not take the intrinsic structure of the underlying network into account, so the transformed network may not preserve the essence of the original network.

The second option is extracting a minimum spanning tree (MST) from a network of N vertices and reducing the number of links to N – 1. This approach guarantees the number of links in the transformed network is always N – 1, whereas option 3 may not have such upper bounds. For instance, we know that a Pathfinder network is the set union of all possible MSTs of the original network, but the number of distinct MSTs depends on the weight distribution of individual links. Therefore, the number of extra links varies not only from network to network, but also from measurement to measurement. For instance, Noel, Chu, and Raghavan [2002] showed that using document co-citation counts normalized as cosine coefficients or Pearson correlation coefficients can lead to MSTs of different topological properties, and that the former resulted in more favorable structures, i.e. the presence of highly connected nodes with a fixed number of links, although the size of their MST is relatively small, less than 200 nodes.

The third option is imposing constraints on paths and excluding links that do not satisfy the constraints, for instance, as in Pathfinder network scaling [Schvaneveldt 1990]. Pathfinder network scaling is a typical example of this approach. The topology of a PFNET is determined by two parameters q and r and the corresponding network is denoted as PFNET(r, q). The q-parameter specifies the maximum length of a path subject to the triangular inequality test. The r-parameter is the Minkowski metric used to compute the distance of a path. The most concise PFNET for visualization is PFNET ($q = N–1$, $r = \infty$) [Chen 2002; Chen and Paul 2001; Schvaneveldt 1990]. In an author co-citation analysis (ACA), White [2003] demonstrated that a 120-node PFNET derived from author co-citation counts was predominated by a number of high-degree nodes. In contrast, if author co-citation links were weighted by Pearson correlation coefficients, the resultant PFNET did not have this pattern. He concluded that using raw counts in ACA would be a preferred method. As a side note, the use of Pearson correlation coefficients is studied in [Ahlgren et al. 2003], where an example is constructed to show that Pearson correlation coefficients could lead to counter-intuitive results in author co-citation analysis.

2.2 Network Evolution

The latest advances in statistical mechanics of complex networks have attracted much attention [Albert and Barabási 2002]. Small-world network properties as well as power-law degree distributions are found in scientific collaboration networks [Newman 2001a; Newman 2001b]. The growth of scale-free networks has increasingly become the focus of the attention. Most network growth models draw upon the rich-get-richer notion and cumulative advantage. As a result, if the degree of a node indicates its "richness," a node with a higher degree will have a better chance to receive the next new link than a node with lower degree. In a citation network, this means that a highly cited article is more likely to be cited again than a less frequently cited article. This type of growing mechanism is known as *preferential attachment*.

Newman [2001a] studied the evolution of scientific collaboration networks in physics and biology and found that the more collaborators a scientist has, the more likely that he or she will work with even more collaborators. Barabási and his colleagues [Barabási et al. 2002] found that preferential attachment mechanisms could statistically reproduce the topological properties of the co-authorship networks of mathematicians and neuroscientists.

One of the underlying assumptions is that the study of networks scientific papers can reveal insights into the dynamics of scientific frontiers. Price suggested that it is possible to identify objectively defined subjects in citation networks and particularly emphasized the significance of understanding such moving frontiers in depicting the topography of current scientific literature [Price 1965].

Small and Griffith [1974] pioneered the method of mapping the structure of scientific literatures, especially through analyses of co-citation networks. Small [1977] subsequently demonstrated the occurrence of rapid changes of research focus using the example of collagen research. Documents clustered by their co-citation links can represent leading specialties. The abrupt disappearance and emergence of such document clusters indicate rapid shifts in research focus. By tracing key events through a citation network, Hummon and Doreian [1989] successfully re-constructed the most significant citation chain in the development of DNA theory. Their study has great impact on subsequent studies of citation networks in the graph drawing community [Batagelj and Mrvar 2001; Brandes and Willhalm 2002].

An interesting study Powell et al. [2002] analyzed the evolution of the biotechnology industry through a study of a network of contractual collaborations in the field between 1988 and 1999. The nodes in the networks are organizations and the links are collaborative ties. Various stages of the network were visualized. No link reduction or pruning was made. It appears to be particularly problematic to identify significant topological and dynamical patterns in such visualization models because of the high density of the underlying network.

An et al. [2001] suggested that the evolution of citation networks could be useful in predicting research trends and in studying a scientific community's life span. Few studies in the literature visualized the growth of an evolving network. Chen and Carr [1999] represent the evolution of the field of hypertext by visualizing its author co-citation networks over consecutive periods of time. The evolution of discourse is visualized in a recent example [Brandes and Willhalm 2002].

3 Criteria on Preferable Network Visualization

Two criteria are derived based on the above analysis for qualitatively evaluating network visualization.

3.1 Criterion I: Topological Properties

The most recognizable patterns in a network are stars, rings, and spikes [Rosch et al. 1976]. The first criterion for selecting a preferable topological structure of a visualized network is the presence of hubs, or stars, in derived networks. The notion of reference points is proposed in [Krumhansl 1978], referring to conceptually or visually salient or distinctive points in a geometric model. Such reference points play the role of a reference context to which other points are seen "in relation to." For instance, a star in a network is a node which is the only node many nodes connect to. The "starness" of a pattern is also studied by Rosch et al. [1976]. A star pattern indicates the star node carries the most information, processes the highest cue validity and the most differentiated from one another. It has been demonstrated in [Chen and Davis 1999] that star patterns emerged in a hybrid PFNET of documents and users' profiles and profiles are in the center, connecting to documents. The preference of star-like patterns is also implicit in Salton's model of an effective indexing space for information retrieval [Salton 1989]. In such indexing spaces, similar documents should be easily separable from the rest of documents so that as one is retrieving a relevant document, it is possible to scoop many other relevant ones in its vicinity and to reject documents located remotely.

Existing studies appear to suggest that co-citation counts are likely to form such star patterns in both MST and PFNET. In terms of small-world networks, star-rich networks have relatively high clustering coefficients; we will return to this subject later in the article. The first part of our study is to identify the boundary conditions of this claim so that one can select the most appropriate method for a given network.

3.2 Criterion II: Dynamical Properties

Our second criterion focuses on the need for visualizing the evolution of a network. What makes a good visualization of an evolving network? The second criterion imposes additional constraints on the visualization of network evolution. Criterion I emphasizes the topological properties of preferable network visualization. Criterion II requires that the changes of topological properties over time must preserve the integrity of emergent trends or patterns. Visualizing network evolution should not merely inform users of changes of individual nodes and links; rather, it is essential to inform users how an intrinsically cohesive structure changes locally and globally in organically. A fragmented growth picture cannot be considered as an adequate visual representation. For instance, Branigan and Cheswick [1999] use their Internet Mapping techniques to show how the Internet in Yugoslavia was affected by the war. The focus is no longer on an individual connection; instead, it is now on the connectivity of a subset of nodes. It also follows that Criterion II implies a level of predictability; a good visualization should give the user various clues of where a new node is likely to appear and where a new path is likely to emerge.

4 MSTs versus PFNETs

Based on the available evidence in recent studies reviewed in earlier sections, both MST and PFNET appear to be capable of meeting the first criterion when conditions on the proximity measurements are satisfied. For instance, MSTs of similarity measures normalized by cosine coefficients tend to have several hubs or star nodes, whereas PFNETs of author co-citation counts with no normalization at all were found to have similar clustering patterns. MST is a common choice in information visualization. Clusters in MST appear to reflect the concepts of hubs and authorities. We also know that MST algorithms are more efficient than PFNET algorithms. Therefore, a number of theoretically and practically important questions now need to be addressed. Will MSTs be a generally better choice? As far as co-citation networks are concerned, will MSTs in general meet the second criterion? To what extent will the topological properties of highly clustered PFNETs be preserved by the use of raw author co-citation counts? Will PFNETs stand up the second criterion for visualizing the evolution of document co-citation networks?

In this study, we construct animated visualization models of the evolution of a research field from 1945 through 2002 in both MST and PFNET. This is essentially an empirical study. We compare the resultant models against the two criteria derived earlier in this article. The goal is to identify examples that can identify the boundary conditions in association with the selection of MST or PFNET. The evolution of the underlying research field is represented by the evolution of its co-citation network over its 58-year span. The nature of components of the co-citation network is identified in both MST and PFNET models using an independent method – accumulative co-citation clustering.

4.1 Botulinum Toxin Research (1945-2002)

The chosen research field for our empirical study is botulinum toxin research between 1945 and 2002. Botulinum toxin is a poison produced by the anaerobic bacteria *Clostridium botulinum* [Jankovic and Brin 1997]. The toxin is one of the most potent poisons known, as little a .1 to 1 µg of toxin can be fatal to humans. It attacks the synapses used by the nervous system to activate muscle movement, preventing the production of the neurotransmitters, thereby causing muscle paralysis. Death can occur if the toxin paralyzes the respiratory muscles. There are seven forms of the neurotoxin, designated A through G. Additionally, *C. botulinum* produces a two-part cytotoxin designated C2 and an exoenzyme, designated C3.

Botulism, the medical condition caused by botulinum toxin, was first systematically studied by J. Kerner, a German medical officer, in the 1820's. The bacteria *C. botulinum* itself was first isolated and its toxin identified by Ermengem in 1897. Most of the different toxin types were identified in the first half of the twentieth century. Modern toxin research started with a seminal paper by Burgen, et al, in 1949, which revealed that the toxin attacked the neuromuscular junction.

4.2 Co-Citation Networks of Botulinum Toxin

Co-citation networks of botulinum toxin research were derived from a citation dataset, containing citation records from 1945 to 2002. Figure 1 shows a power law model of the relationship between the number of nodes and the number of links in co-citation networks at various citation thresholds. For instance, at the thresholds of 5, 10, and 25 citations, the size and the density of the co-citation networks are: 1,250 nodes and 91,483 links, 516 nodes and 19,631 links, and 104 nodes and 2,677 links.

In the rest of the article, we primarily focus on the 516-node co-citation network. In addition, we briefly discuss two PFNETs without any normalization on the co-citation counts: one is a 407-node author co-citation network for authors who have more than 15 citations; the other is a 380-node document co-citation network for articles with more than 12 citations. These two networks are analyzed in order to identify the extent to which a PFNET can keep the number of links close to N.

The weight of a link in the network was calculated in two ways: first, weight links by direct co-citation counts; secondly, weight links by normalized co-citation coefficients. The following normalization (I) is used in this study:

$$sim(d_i, d_j) = \frac{cc(d_i, d_j)}{\sqrt{c(d_i) \bullet c(d_j)}} \qquad \text{(I)}$$

where $cc(d_i, d_j)$ is the number of times document *i* and document *j* are cited together, and $c(d_i)$ and $c(d_j)$ are the number of times document *i* and document *j* are cited respectively. Alternatively, one may choose to use the following normalization (II), but a detailed comparison between the two is beyond the scope of this article:

$$sim(d_i, d_j) = \frac{cc(d_i, d_j)}{c(d_i) + c(d_j) - cc(d_i, d_j)} \qquad \text{(II)}$$

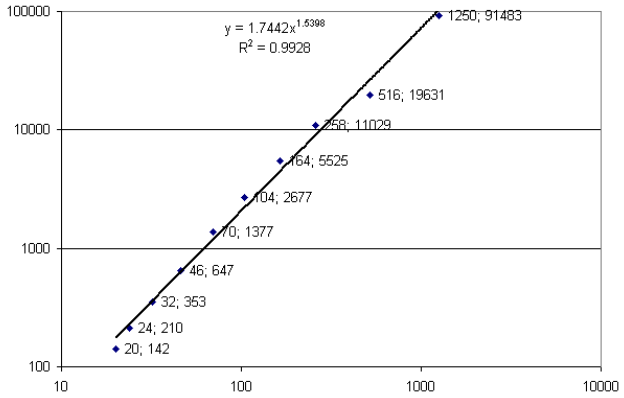

Figure 1. Log-log plot of the size of co-citation network at various citation thresholds, from 5 through 50 increased by 5. X axis is the logarithm of the number of nodes. Y axis is logarithm of the number of links.

MSTs were extracted using Prim's algorithm. PFNETs were extracted using the algorithm described in [Schvaneveldt 1990]. Both types of network models were examined against the first criterion given in Section 3. In order to examine the compliance to the second criterion, animated visualizations were generated as a sequence of annual snapshots of the evolving network throughout the 58-year period. The animated visualization revealed two types of state transitions as originally specified in [Chen and Kuljis 2003]. The connectivity of the underlying co-citation network was represented by three node states and three link states. The three node states (NS) of an article are:

NS1. Pre-publication state.
NS2. Published but not yet cited.
NS3. First citation detected.

Similarly, a co-citation link connecting two articles has three states (LS) as well. Suppose article A_i was published earlier than article A_j.

LS1. Both A_j and A_j in NS1.
LS2. Both A_j and A_j in NS2 or NS3.
LS3. First co-citation detected.

The method used to label and explore research topics in the network models is outlined as follows. For this purpose, research fronts are considered as collections of papers on specific research problems in a field [Morris et al. 2003]. Base reference clusters are groups of references that represent the foundational knowledge used by workers when investigating research problems. Research fronts can be found by clustering documents that tend to cite the same references, using bibliographic coupling [Kessler 1963] as the basis for measuring similarity between pairs of papers. Base reference clusters can be formed by clustering references that tend to be cited together, using co-citation [Small 1997] as the basis for measuring similarity between pairs of references.

In this study, research fronts were identified by agglomerative clustering using only papers that had at least five bibliographic coupling counts with some other paper in the dataset. Similarity calculation was based on Salton's cosine coefficient [Salton 1989] applied to bibliographic coupling counts. The titles for each research front were derived manually by exploring titles of papers within each research front for common themes. Base reference clusters were formed by agglomerative clustering using only references that had been cited 10 or more times. Similarity calculation was based on Salton's cosine coefficient applied to co-citation counts. For each base reference cluster, labels were found by using the label of the research front that contained the most citations to references in the cluster. A map of the references in the pathfinder network was produced identifying each reference by its base reference cluster membership, which allowed labeling of sections of the pathfinder network based on base cluster labels.

5 Results

The MST model indeed contained many clusters. Many articles did not connect to any other articles in their cluster apart from the cluster center. Figure 2 shows the 516-node MST based on the normalized co-citation counts. A three-dimensional visualization with the citation counts depicted in the third dimension also confirms that the cluster centers tend to have higher citation counts than non-center members of clusters. The MST model in this particular case evidently met the first criterion and it would be reasonable to hypothesize that MSTs can meet the criterion in a broader range of networks.

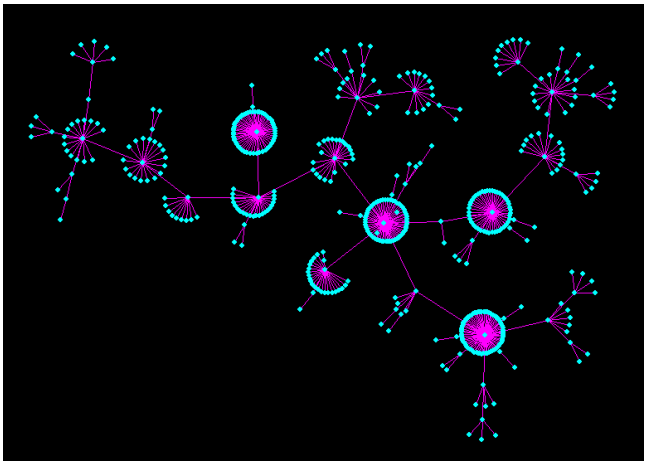

Figure 2. The MST visualization of the 516-node co-citation network on botulinum toxin (1945-2002) is predominated by star nodes. Co-citation counts are normalized (I).

However, an examination of the animated visualization over the 58-year span indicates that the MST model did not meet the second criterion, which requires the visualized network to convey the evolution of globally and locally cohesive structures. A key question is how the relationship between the center of a cluster and other non-center members in the cluster was depicted over the course of evolution. In general, due to the arbitrary choice inherited from the MST algorithms, one cannot guarantee the uniqueness of an MST. As a result, an MST may not preserve all the necessary links for representing the growth of a co-citation network. If this is the case, then important diffusion patterns may be distorted or inadequately represented by the extracted MST model. Users will probably find it hard to understand the way new nodes and new links emerge. The nature of the problem will become clear shortly when we contrast the growth animation of the PFNET and MST models.

The 516-node PFNET ($q = N - 1$, $r = \infty$) is shown in Figure 3. The two parameters q and r were chosen to ensure that the extracted PFNET has the least number of links. The network in this case contains 525 links, which gives the node-link ratio of 0.98. We have developed a number of visualization methods to identify the nature of local structures of a PFNET, including node color mapping based on principle component analysis (PCA) on co-citations normalized as cosine coefficients, chronologically synchronized animated visualizations of state transitions for both nodes and links, and base reference cluster memberships based on the clustering algorithm outlined at the end of Section 4, where clusters are formed independently from algorithms used in modeling the network. In Figure 3, each node is depicted as its cluster number. The PFNET and the clustering methods appear to have a nearly perfect match between each other.

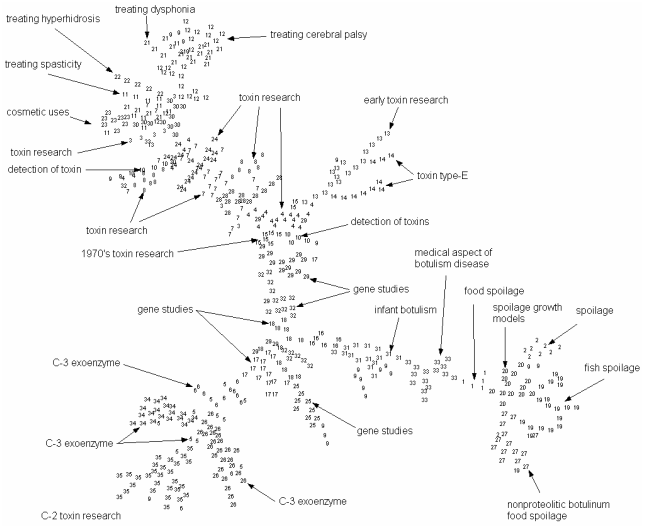

Figure 3. The PFNET visualization of the 516-node co-citation network ($q = N - 1$, $r = \infty$), containing 525 links.

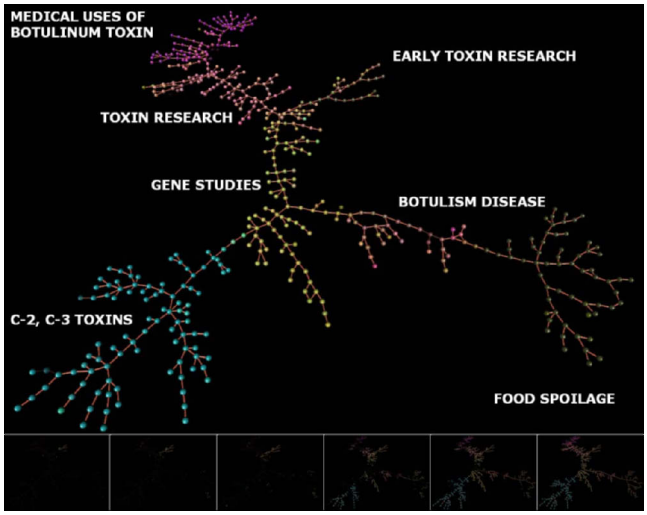

Figure 4. The 516-node PFNET consistently partitioned by base reference clusters and PCA factors. The PFNET is predominated by strongest paths.

Several distinct research fronts emerged in the 1980's. Gene sequencing research on *C. botulinum* started in the early 1990's. Toxin research base references are located in the areas slightly above the center of the map. Furthermore, research fronts have opened up on C2 cytotoxin and C3 exoenzyme recently. Additional research in botulinum toxin is using the C3 exoenzyme to study Rho proteins. C3 exoenzyme is also being studied as a possible neurotrophic drug, used for encouraging nerve growth. Base references related to C-2 and C-3 toxins are located in the South-Western region of the map in Figure 3.

Other research fronts in botulinum toxin research have focused on the mechanisms of food spoilage (located in the South-Eastern region of the PFNET in Figure 3), the medical aspects of botulism (along the branch stretching towards East in Figure 3), and infant botulism (also in the same branch), which is often attributed to cases of sudden infant death syndrome. There is also research being performed to develop methods of detecting and assaying the botulinum toxin in the environment. Botulinum toxin can be used as a chemical/biological warfare agent and possible bioterror weapon, making the search for a cheap and efficient detection method an important area for research. There does not appear to be a consistent set of base references for botulinum toxin bioterrorism and biological warfare as there is for the case of anthrax [Morris et al. 2003].

The medical uses of botulinum toxin have received a great deal of public attention. Scott et al. [1973] described experiments on monkeys to treat eye alignment disorders. In the 1980's it was noticed that many patients being treated for blepharospasm (a disorder of clamping of eyelids) using botulinum toxin exhibited reduced facial wrinkles and improved cosmetic appearance. Based on this effect, Carruthers [1992] reported the use of botulinum toxin for cosmetic purposes. The toxin has gained wide use for this purpose and as a result, the toxin is the current focus of much public attention.

In the 1990's botulinum toxin has been studied for the treatment of, spasticity, dysphonia (clamping muscles), achalasia (a disorder of clamping throat muscles that interferes with swallowing and food ingestion), cerebral palsy, hyperhidrosis (excessive sweating), anal fissures, and more. Jankovic [1991] presented an important review on medical uses of botulinum toxin. Most medical uses of botulinum toxin are based on toxin type A, which is manufactured under the commercial name of Botox.

The PFNET in Figure 5 was colored by factor loading from PCA based on cosine coefficients of co-citation frequencies. It is clear from Figure 3 and Figure 4 that base reference clusters and PCA factors are consistent with each other. For instance, C-2 and C-3 toxins base references identified by timeline visualization correspond to the area in light blue, which is consistently identified by PFNET and PCA.

Unlike the MST model, the PFNET model was not predominated by high-degree nodes. If we use the first criterion alone, MSTs would be a more preferable choice than PFNETs. In addition, PFNETs derived from raw co-citation counts appear to form more interpretable structures than normalized versions, cf. [White 2003]. However, our further study of the second criterion indicates that this may not be the case if we take the temporal factor into account.

The examination of the second criterion was based on animated visualizations of the 58-year growth history of the field. Figure 5 shows some of the frames in the animation sequence. The enlarged frame shows the diffusion process of how several base reference clusters emerge and spread. State transitions were shown by changing the transparency level of nodes and links in question. The four smaller frames in the figure were selected from the animation sequence to show the emergence of early toxin research in late 1940s and the research front of gene studies formed at the center more recently.

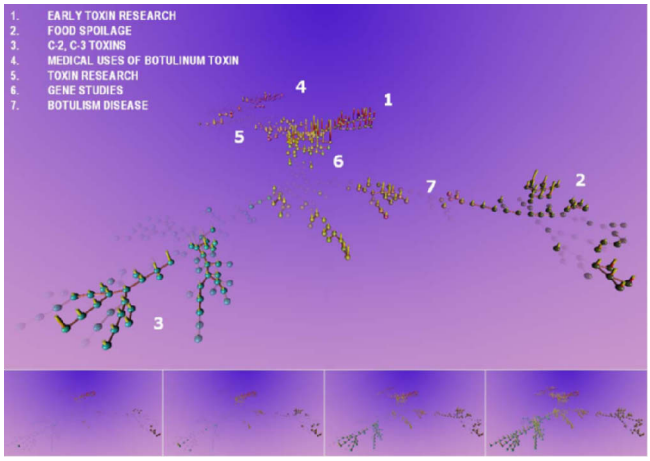

Figure 5. The animated PFNET sequence shows the evolution of the field as the PFNET network becomes populated.

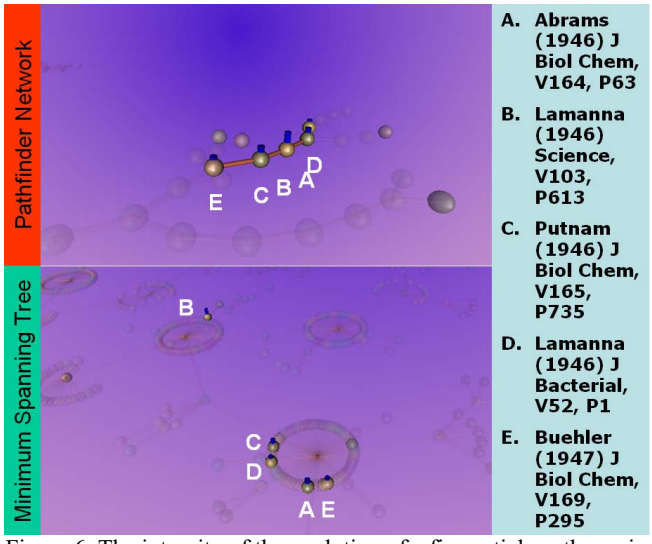

Figure 6. The integrity of the evolution of a five-article pathway is well preserved in PFNET (top), whereas the same pathway is fragmented in MST (bottom).

The animated PFNET visualization model demonstrated that nodes with similar colors often emerged simultaneously and formed local structures. And these local structures were reinforced by the timely emergence of salient co-citation links. The growth process can be represented by the dynamics shown in such local structures. Features such as continuity, predictability, and local cohesiveness in the PFNET indicated that the second criterion was met. More significantly, it was found that these properties were missing from the MST model (See Figure 6). In PFNET, five pioneering toxin research articles formed a distinct thread, or a pathway. One can follow the development of the thread visually as new nodes and new links extend the pathway. In contrast, in the MST model, four of the five articles were in the same cluster and one article was found in a different cluster.

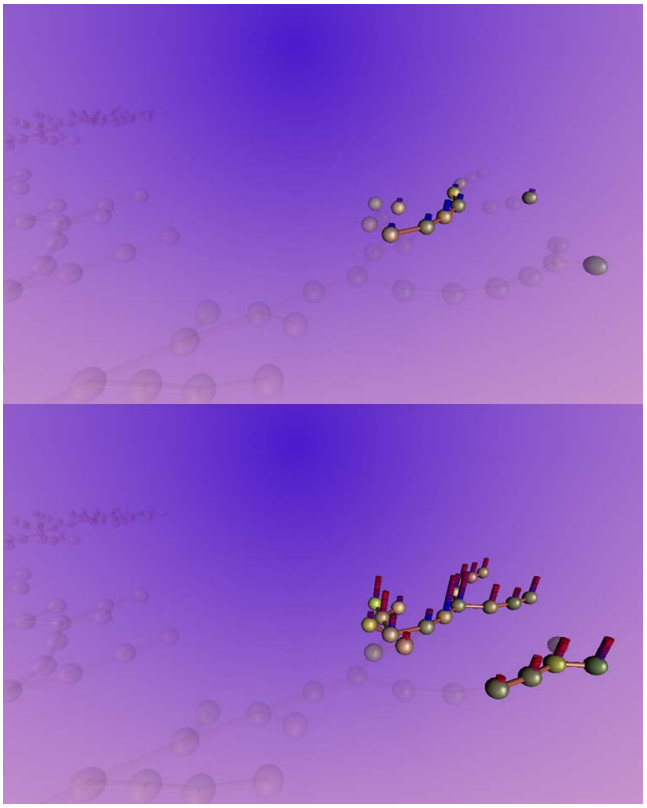

Figure 7. Two snapshots of PFNET taken subsequently clearly show the organic growth of the co-citation pathways. Such pathways were destroyed in MST.

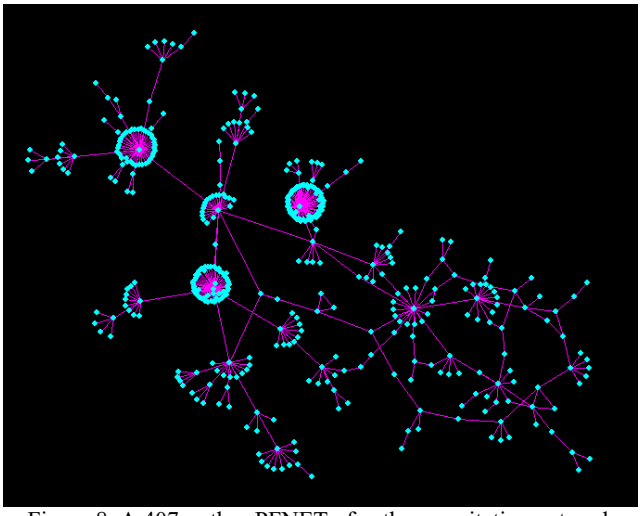

Figure 8. A 407-author PFNET of author co-citation network, containing 428 links (node/link ratio 0.95). The entry threshold is 15 or more citations for each author.

The animation showed that the pathway clearly captured by PFNET was simply not in the MST model. None of the four articles was the center of the cluster. Each article in the cluster connected to the center through a single co-citation link. However, such links in MST were not necessarily the earliest or the most salient co-citation links. MST may have excluded some vital links in the crucial pathways in the course of evolution.

A distinct advantage of the PFNET is evident in Figure 7, which clearly shows the evolution of the co-citation network, starting with the short pathway at first, and then continuing the growth by the emergence of the second pathway alongside. The two frames were separated by a few years. Users can easily recognize the nature of the newly added nodes.

Finally, we briefly discuss the potential weakness of using un-normalized link weights. Figure 8 is a 407-node PFNET of author co-citation network, containing 428 Pathfinder links. The node/link ratio is 0.95. Figure 9 is a 380-node PFNET of document co-citation network, containing 523 links. The node/link ratio is as low as 0.73. Instead of normalizing link weights, the two PFNETs used direct co-citation counts as link weights. A low node/link ratio means that the PFNET has too many links and it can quickly lose the advantage of a Pathfinder network as the PFNET drifts too far from a tree structure.

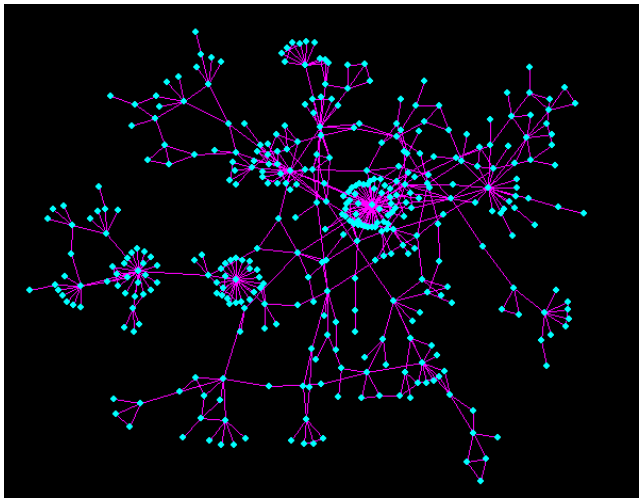

Figure 9. A 380-article PFNET of a document co-citation network, containing 523 links (node/link ratio 0.73). The entry threshold is 12 or more citations for each article.

6 Conclusion

In conclusion, the topological and dynamical criteria have enabled us to distinguish visual-spatial features of MST and PFNET in the context of network evolution. PFNETs tend to better preserve the evolution of networks. On the other hand, MST algorithms can be implemented more efficiently. Integrative strategies are likely to be a fruitful approach.

Given the findings of this study, several lines of research are worth considering: 1) visualizing the evolution of co-citation networks in other subject domains, 2) visualizing scientific networks other than co-citation networks, such as citation networks, bibliographic coupling networks, co-authorship networks, and social networks, 3) visualizing a wider range of networks, for example, small-world networks and scale-free networks, and 4) visualizing the evolution of complex systems that are not necessarily represented as networks.

We expect that visualizing the evolution of networks will stimulate further development and refinement of visualization techniques in general and fruitful collaborations between information visualization and other scientific communities.

Acknowledgements

The citation analysis research award from the Institute for Scientific Information (ISI) and the American Society for Information Science and Technology (ASIST) is acknowledged.

References

AHLGREN, P., JARNEVING, B., AND ROUSSEAU, R., 2003. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficent, *Journal of the American Society for Information Science and Technology*, 54, 6, 550-560.

ALBERT, R. AND BARABASI, A., 2002. Statistical mechanics of complex networks, *Reviews of Modern Physics*, 74, 1, 47-97.

AN, Y., JANSSEN, J., AND MILIOS, E., 2001. Characterizing and mining the citation graph of the computer science literature, Dalhousie University, Halifax, Nova Scotia, Canada CS-2001-02, September 26, 2001.

BARABÁSI, A.-L., ALBERT, R., AND JEONG, H., 2000. Scale-free characteristics of random networks: The topology of the world-wide web, *Physica A*, 281, 69-77.

BARABÁSI, A. L., JEONG, H., NÉDA, Z., RAVASZ, E., SCHUBERT, A., AND VICSEK, T., 2002. Evolution of the social network of scientific collaborations, *Physica A*, 311, 590-614.

BATAGELJ, V. AND MRVAR, A., 2001. Layouts for GD01 graph-drawing competition.

BRANDES, U. AND CORMAN, S. R., 2002. Visual unrolling of network evolution and the analysis of dynamic discourse. In *Proceedings of IEEE Symp. Information Visualization (InfoVis '02)*, Boston, MA., 145-151.

BRANDES, U. AND WILLHALM, T., 2002. Visualization of bibliographic networks with a reshaped landscape metaphor. In *Proceedings of Proc. 4th Joint Eurographics - IEEE TVCG Symp. Visualization (VisSym '02)*, 159-164.

BRANIGAN, S. AND CHESWICK, B., 1999. The effects of war on the Yugoslavian network, vol. 2003: Lumeta.

CARRUTHERS, J. D. A., 1992. Treatment of glabellar frown lines with c-botulinum-a exotoxin, *J Dermatol Surg Onc*, 18, 17.

CHEN, C., 1998. Generalised Similarity Analysis and Pathfinder Network Scaling, *Interacting with Computers*, 10, 2, 107-128.

CHEN, C., 2002. *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. London: Springer-Verlag.

CHEN, C. AND CARR, L., 1999. Visualizing the evolution of a subject domain: a case study. In *Proceedings of Proceedings of the IEEE Visualization'99 Conference*, 449-452.

CHEN, C. AND DAVIS, J., 1999. Integrating spatial, semantic, and social structures for knowledge management. In *Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS '32)*, Hawaii.

CHEN, C. AND KULJIS, J., 2003. The rising landscape: A visual exploration of superstring revolutions in physics, *Journal of the American Society for Information Science and Technology*, 54, 5, 435-446.

CHEN, C. AND PAUL, R. J., 2001. Visualizing a knowledge domain's intellectual structure, *Computer*, 34, 3, 65-71.

CHI, E., PITKOW, J., MACKINLAY, J., PIROLLI, P., GOSSWEILER, R., AND CARD, S., 1998. Visualizing the evolution of web ecologies. In *Proceedings of Proceedings of CHI'98*, Los Angeles, 400-407.

EADES, P., 1984. A heuristic for graph drawing, *Congressus Numerantium*, 42, 149-160.

FAIRCHILD, K., POLTROCK, S., AND FURNAS, G., 1988, SemNet: Three-dimensional graphic representations of large knowledge bases, in *Cognitive Science and its Applications for Human-Computer Interaction*, R. Guidon, Ed.: Lawrence Erlbaum Associates, pp. 201-233.

HUMMON, N. P. AND DOREIAN, P., 1989. Connectivity in a citation network: The development of DNA theory, *Social Networks*, 11, 39–63.

JANKOVIC, J., 1991. Therapeutic uses of botulinum toxin, *New England Journal of Medicine*, 324, 1186.

JANKOVIC, J. AND BRIN, M. F., 1997. Botulinum toxin: historical perspective and potential new indications, *Muscle Nerve Suppl*, 6, S, 129-145.

KESSLER, M. M., 1963. Bibliographic coupling between scientific papers, *American Documentation*, 14, 10-25.

Kleiberg, E., van de Wetering, H., van Wijk, J. J. 2001. Botanical visualization of huge hierarchies. In Proceedings of IEEE Symposium on Information Visualization 2001 (InfoVis'01). Oct 22-23, 2001. San Diego, CA. 87-94.

KRUMHANSL, C. L., 1978. Concerning the applicability of geometric models to similar data: The interrelationship between similarity and spatial density, *Psychological Review*, 85, 5, 445-463.

LAMPING, J. AND RAO, R., 1996. The hyperbolic browser: A focus plus context technique for visualizing large hierarchies, *Journal of Visual Languages and Computing*, 7, 1, 33-55.

MORRIS, S. A., YEN, G., WU, Z., AND ASNAKE, B., 2003. Timeline visualization of research fronts, *Journal of the American Society for Information Science and Technology*, 55, 5, 413-422.

NEWMAN, M. E. J., 2001a. Clustering and preferential attachment in growing networks, vol. 2003: arXia:cond-mat/0104209.

NEWMAN, M. E. J., 2001b. The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci. USA*, 98, 404-409.

NOEL, S., CHU, C. H., AND RAGHAVAN, V., 2002. Visualization of document co-citation counts. In *Proceedings of Proceedings of the 6th International Conference on Information Visualisation*, London, England, 691-696.

POWELL, W. W., WHITE, D. R., KOPUT, K. W., AND OWEN-SMITH, J., 2002. The evolution of a science-based industry: Dynamic analyses and network visualization of biotechnology, in *http://www.fek.umu.se/dpcc/powell.pdf*.

PRICE, D. D., 1965. Networks of scientific papers, *Science*, 149, 510-515.

ROBERTSON, G. G., MACKINLAY, J. D., AND CARD, S. K., 1991. Cone trees: Animated 3D visualizations of hierarchical information. In *Proceedings of CHI '91*, New Orleans, LA, 189-194.

ROSCH, E., MERVIS, C. B., GRAY, W., JOHNSON, D., AND BOYES-BRAEM, P., 1976. Basic objects in natural categories, *Cognitive Psychology*, 8, 336-356.

SALTON, G., 1989. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, Mass.: Addison-Wesley.

SCHVANEVELDT, R. W., 1990. Pathfinder Associative Networks: Studies in Knowledge Organization, in *Ablex Series in Computational Sciences*, D. Partridge, Ed. Norwood, New Jersey: Ablex Publishing Corporations.

SCOTT, A. B., ROSENBAUM, A., AND COLLINS, C. C., 1973. Pharmocologic weakening of extraocluar muscles, *Invest Ophthalmol*, 12, 924.

SMALL, H., 1997. Update on science mapping: Creating large document spaces, *Scientometrics*, 38, 2, 275-293.

SMALL, H. G., 1977. A co-citation model of a scientific specialty: A longitudinal study of collagen research, *Scoial Studies of Science*, 7, 139-166.

SMALL, H. G. AND GRIFFITH, B. C., 1974. The structure of scientific literatures I: Identifying and graphing specialties, *Science Studies*, 4, 17-40.

WATTS, D. J. AND STROGATZ, S. J., 1998. Collective dynamics of 'small-world' networks, *Nature*, *393*, 440-442.

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, 5, 423-434.

WILLS, G. J., 1999. NicheWorks: Interactive visualization of very large graphs, *Journal of Computational and Graphical Statistics*, 8, 2, 190-212.

ZIZI, M. AND BEAUDOUIN-LAFON, M., 1994. Accessing hyperdocuments through interactive dynamic maps. In *Proceedings of ECHT '94*, Edinburgh, Scotland, 1994 126-135.