Visualization of Large Citation Networks with Space-Efficient Multi-Layer Optimization

Mao Lin Huang and Quang Vinh Nguyen ***Member, IEEE***

Abstract — **This paper describes a technique for visualizing large citation networks (or bibliography networks) using a space-efficient multi-layer optimization visualization technique. Our technique first use a fast clustering algorithm to discover community structure in the bibliographic networks. The clustering process partitions an entire network into relevant abstract subgroups so that the visualization can provide a clearer and less density of display of global view of the complete graph of citations. We next use a new space-efficient visualization algorithm to archive the optimization of graph layout within the limited display space so that our technique can theoretically handle a very large bibliography network with several thousands of elements. Our technique also employs rich graphics to enhance the attributed property of the visualization including publication years and number of citations. Finally, the system provides an interaction technique in cooperating with the layout to allow users to navigate through the citation network. Animation is also implemented to preserve the users' mental maps during the interaction.**

Index Terms — **Citation Networks, Bibliographic Networks, Graph Visualization, Space-Efficient Visualization, Clustering.**

I. Introduction

SCIENTIFIC articles' analysis usually shows the coherence of literatures as well as their change in intelligible ways over time. One of the significant areas of the scientific analysis is citation and co-citation (or bibliography) networks. This area has been long studied in information science and other disciplines. Most of the recent techniques cooperate with an advanced visualization technique for assisting the analysis. Within the scope of this paper, we briefly review a few typical techniques in citation and co-citation analysis.

CiteSpace [3] and its upgrade version CiteSpace II [4] are the one of the most recent examples along this line of research. This system can detect, conceptualize and visualize emerging trends and transient patterns of citation and co-citation footprints in scientific literature. Although it provides a good visualization which shows both of cluster views and time-zone views, its 2D *spring-embedder* layout algorithm is slow and too much clustered for handling large scientific citation networks.

CiteWiz [8] is another framework for bibliographic visualization of scientific network. This technique can graphically present the hierarchies of articles with potentially very long citation chains. Besides the visualization of citation hierarchies, the authors also provide an interaction technique and an attributed property for enhancing the overview of important authors and articles which make it more useful for a wide range of scientific activity. Modjeska et al. [12] described a similar technique that can capture the user-required relationships between bibliographic entries or articles. This system also supports chronological structure with multi-articles overview, single-article view and spatial attribute-relevance view. Other visualization techniques for citation and co-citation analysis can be found at [1, 3, 11, 15, 17 and 19].

Graphs generated in real citation networks could be very large with thousands or perhaps hundred thousands of nodes and edges. As the result of rapid increasing of the size in networks, the large scale visualization of citation networks has been turned into one of the hottest topics in bibliography research. Therefore, the question about how to comprehensively display large graphs of citation networks on the screen becomes the key issue in the visualization. Large graphs can decrease significantly the performance of a visualization technique which normally performs well on small or medium datasets. Large graph visualization usually suffers from poor running time and the limitation of display space. In addition, the issue of "view-ability" and usability also arises because it will be impossible to discern between nodes and edges when a dataset of thousands of vertices and edges are displayed [9].

It seems that the classical graph model with a simple node-link diagram tends to be inadequate for large scale visualization with several thousands of items. The lack of formal hierarchical structures in the citation visualization applications could limit the conveying and perception of the complicated information. The limitations of classical graph visualization in citation networks generally are: 1) too many nodes to be displayed and the layout of such a large geometrical area could not be fitted in one single screen, and 2) the layout of the graph has inefficient utilization of display space.

Therefore, to address the first problem, a well established new graph model to accommodate with the visualization of large graphs is required. Clustered graph model [7] is one of excellent approaches to deal with such large graphs. They partition recursively the graph into the hierarchy of sub-graphs for clustered visualization, which simplifies the complex structures of large graphs through the global abstraction for

This work was supported in part by the Australia Research Council under Discovery Research Grant No. DP0665463.

Dr Mao Lin Huang is a senior lecturer at the Faculty of Information Technology, University of Technology, Sydney, Australia (e-mail: maolin@ it.uts.edu.au).

Dr Quang Vinh Nguyen is a post-doctoral fellow at the Faculty of Information Technology, University of Technology, Sydney, Australia (e-mail: quvnguye@ it.uts.edu.au).

easy interpretation, perception and navigation of large information spaces. To solve the second problem mentioned above, we need to optimize the layout algorithms that could maximize the utilization of display screen allowing more nodes to be displayed. The research from Ware [18] shows that more information can be displayed on very high-resolution and large screen, but it does not necessarily provide very much more information into the brain. Therefore, the investigation of new optimized space-efficient technique for visualizing large datasets could be more effective and economical than the use of expensive large display devices.

This paper presents a technique for visualizing large citation networks of several thousands of elements. Section 2 briefly describes the technical detail of our clustering algorithm. Section 3 presents the visualization including layout, navigation and interaction, and attributed display. Final is the conclusion and future work. Final section is our conclusion of the work.

II. CLUSTERING

The graph clustering method aims to quickly discover the community structure embedded in large citation networks and it divides the network into densely connected sub-networks. The proposed algorithm not only can run fast in time but also achieves a consistent partitioning result in which a connected graph is divided into a set of clusters of similar size. The balanced size of clusters could provide users with a clearer view of the clustered graph and thus it makes easier for users to visualize and navigate large networks. The combination between our clustering method and a space-efficient layout technique would enable the visualization of very large general citation networks with several thousands of elements.

To simplify the problem, our clustering algorithm is applied only to those nodes connected with large subgroups. The algorithm first groups none-connected nodes and nodes belonging to small subgroups of nodes into two large clusters respectively. The clustering process is hierarchically applied through the entire cluster graph from upper level clusters to downer levels until each cluster has a reasonable small number of nodes. Formally, we carry out the following steps in our clustering process:

1. Group all isolated or un-cited articles into one group.
2. Group all small connected subgroups of articles, i.e. a connected subgroup of articles with a few local citations, into clusters and merge these clusters into a large cluster.
3. For each large connected (or cited) subgroup of articles, apply clustering algorithm to partition it into small highly connected clusters.
4. Repeat step 3 for every cluster if a partitioned cluster is still large.

Figure 1 shows an example of our clustering technique performing on a citation network in terrorism research area. This figure shows that the network is divided into three subgroups. The top-right group is a collection of un-cited articles. The left group is a collection of small subgroups of small cited articles in which each subgroup contains small locally citing references. The bottom-right group is a collection of a large group of cited articles. This group is further partitioned into five highly connected clusters in which articles are more relevant within each clusters.

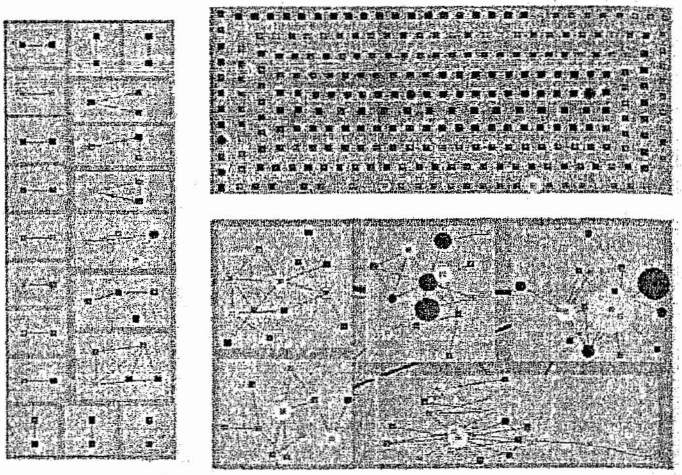


Fig. 1. An example of a visualization of an entire citation network in terrorism research area in which the network has grouped into sub-clusters.

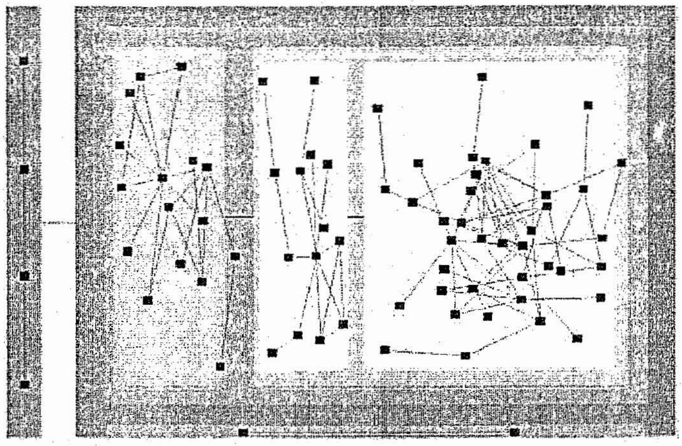


Fig.2. An example of the clustering result on a medium size graph, produced by Newman's method, in which sub-graphs are very unbalanced.

Technically, our clustering algorithm is a modification of Newman's algorithm [13] which considers the equality of size among clusters during the partitioning. Although Newman's algorithm is very fast and it performs well in some datasets, this technique could create unbalanced partitioning between clustered sub-graphs. This is because that this method does not consider the weight or the balance between merged subgroups. Figure 2 shows an example of a medium size citation graph produced by Newman's algorithm [13]. The figure clearly shows an unbalanced partitioning result among clusters in which the left cluster contains only 4 nodes while the right cluster has more than 40 nodes. The figure also shows that the clustering process needs to iterate 3 times running through the cluster tree to archive a reasonable good partitioning result for a clear view.

To overcome the problem of unbalanced clustering, we introduce a weight attribution into the calculation of quality increment ΔQ for defining the modularity value Q, defined by [13]. This weight variable ensures the balance between the connection strength and the node group size among clusters. Figure 3 shows another example of the clustering resulting from our algorithm on a large size graph in which clustered sub-graphs have quite similar size.

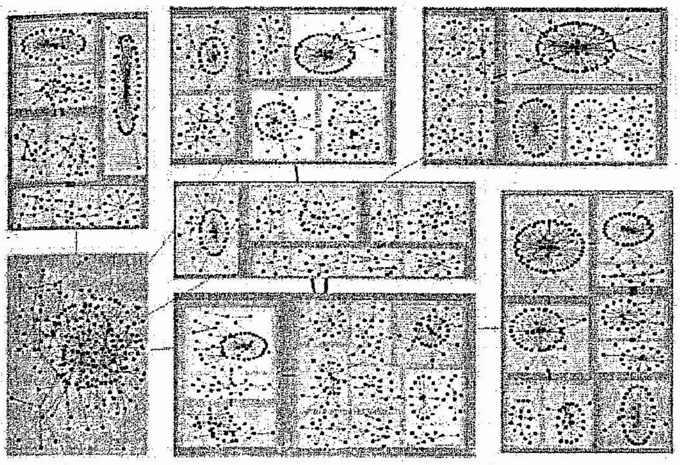


Fig. 3. An example of our clustering result on a large size graph.

III. VISUALIZATION

We use a new space-efficient visualization technique, similarly to *EncCon* [14], to optimize the geometrical space for visualizing large clustered citation graphs with several thousands of nodes and edges. The use of visualization techniques aims to provide a two-dimensional graphical visual interface for viewing the entire collection of citation networks and navigating to any particular sub-clustered group of articles.

In our visualization system, nodes are used to represent articles and their information while edges between nodes are representing citations among these articles. A node is associated with corresponding color and size to represent its attributed property including publication year and number of citations. The thickness of edges between clusters indicates the connection strength between those clusters.

Besides layout algorithms, we also use a multiple-views technique to provide both focus and context information during the navigation and interaction of graphs. In short, our visualization provides three views: a *main view*, a *full-context view* and a *current-context view* (see examples at Figure 4 and Figure 5). The *main view* is used for displaying focus information and it occupies almost the displaying area at the right-side. The *full-context view*, which is displayed as a small panel locating at the top-left side of the visualization, shows the entire context of the citation graph in high-level of abstraction with little detail. This view enables users to always maintain an overall mental map of the context information. The *current context view*, which is displayed as a small panel at the left side under the *full-context view*, also shows the immediate context during the navigation. Therefore, this view displays is a complement to the *full-context view* in which provides further detail of the context at a focus point during the navigation. The semantic zooming is employed in our navigation to enlarge the display of a particular part of the clustered graph based on user's interest at a time. We now describe the detail of our layout, navigation algorithms and attributed visualization.

A. Layout

Our layout algorithm is responsible for positioning of all nodes in a given clustered graph in a two-dimensional geometrical space. The layout of clustered graph is generated by using a combination of an extended new fast *enclosure* partitioning algorithm, called *Clenccon,* and a traditional *Spring Embedder* algorithm [6]. The *Clenccon* layout algorithm is only applied to those non-leaf sub-graphs in which the space utilization and computational cost issues are crucial. On the other hand, the *Spring Embedder* algorithm is applied to the calculation of the position for those leaf sub-graphs, which contain a small number of nodes and the space utilization issue becomes less important and, therefore, the aesthetic niceness and flexibility issues need to be more considered. The visualization also displays a high-level node-link diagram to present the overall clustering structure explicitly.

Our *Clenccon* layout algorithm inherits essentially the advantage of space-filling techniques [2, 10] that maximize the utilization of display space by using area division for the partitioning of sub-trees and nodes. Note that the issue of space utilization becomes significantly important when visualizing large citation networks with thousands or even hundred thousands of nodes and edges because of the limitation of screen pixels. It is similar to *EncCon* [14] that use a rectangular division method for recursively positioning the nodes hierarchically. This property aims to provide users with a more straightforward way to perceive the visualization and ensures the efficient use of display space. However, our new technique is capable of handling clustered graphs rather than simple tree structures; therefore, the algorithm takes the connectivity property between sibling nodes into its partitioning process.

Technically, we first assign the entire rectangular display area as the local region to the clustered graph. We then recursively partition the local regions for every sub-clusters until all the clusters are reached. In our enclosure partition technique, each cluster is bounded by a rectangular local region centered at super-node (or sub-root node) and the drawing of the corresponding sub-clustered-graph is restricted within the geometrical area of. Therefore, the local region of a cluster is the sum of the rectangular areas assigned to its children. The position of a non-leaf node is defined at the center of the rectangle defined by its sub-graph. The position of a leaf node is defined by the *Spring Embedder* algorithm [6].

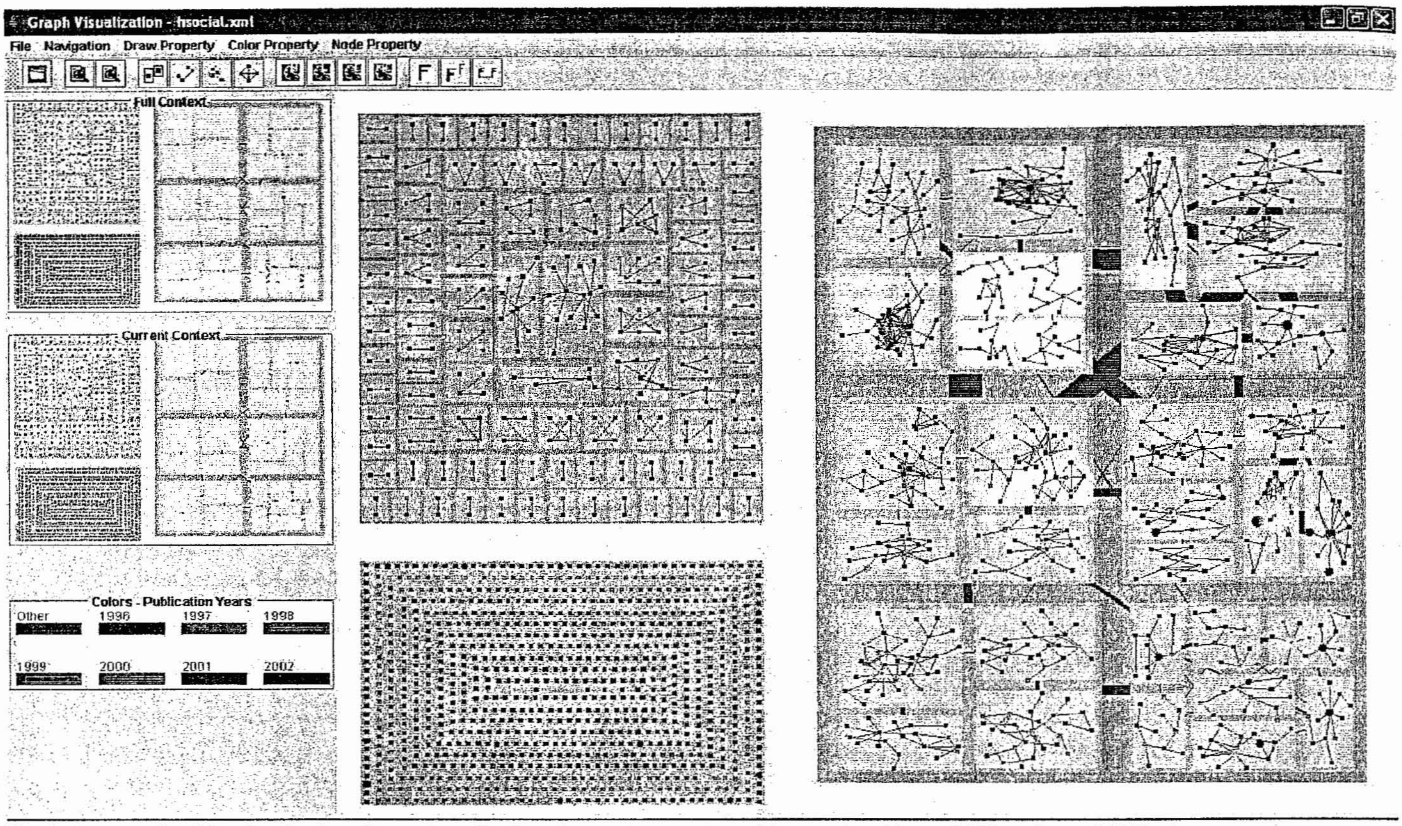

Graph Visualization - hsocial.xml
File Navigation Draw Property Color Property Node Property
Full Context
Current Context
Colors - Publication Years
Other
1996
1997
1998
1999
2000
2001
2002

Fig. 4. An example of the visualization of an entire collection of research papers in social networks with over 2000 vertices and 4200 edges.

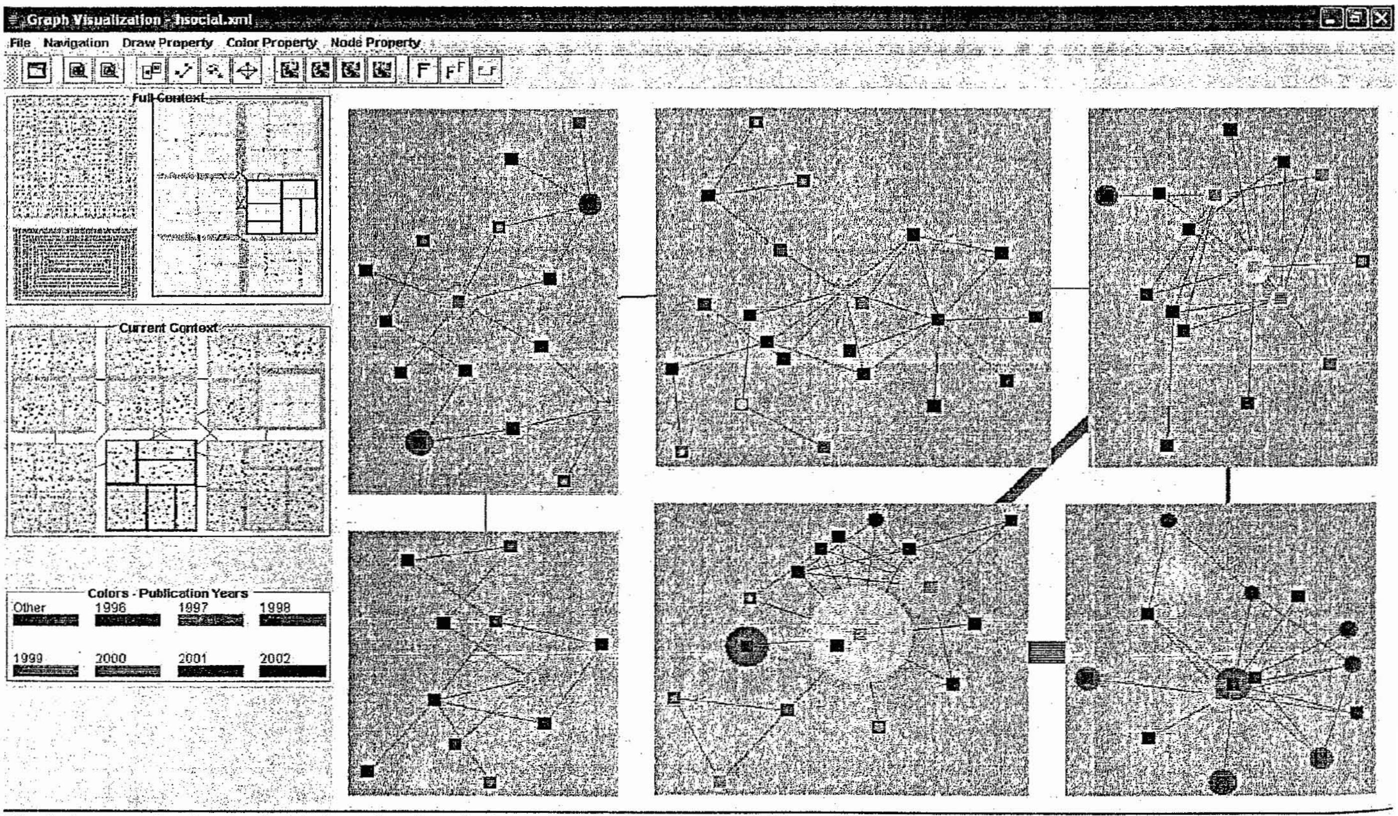

Graph Visualization - hsocial.xml
File Navigation Draw Property Color Property Node Property
Full Context
Current Context
Colors - Publication Years
Other
1996
1997
1998
1999
2000
2001
2002

Fig. 5. An example of navigational view of a selected sub-graph of same dataset used in Figure 4.

B. Navigation and Interaction

The main concerns of our visualization are not only for the geometrical positioning but also the navigation of the large citation network. This is because no layout algorithm can work alone to show detail of information for large dataset as a full functional interactive user interface. Therefore, the design of an associated navigation mechanism for interactively retrieving data items needs to be considered as part of the visualization. Besides layout algorithms, we also use a *multiple-views* technique [16] to provide both focus and context information during the navigation and interaction of graphs. In short, our visualization provides three views: a *main view* for focus information, a *full-context view* and a *current-context view* for displaying the entire context and current context information respectively (see Figure 4 and Figure 5).

Our navigation technique allows the exploration of data hierarchically to quickly focus on the interest part of data. The navigation works by zooming into an area of interest or enlarging a specific section of data while retaining simplified context views. The two *context views* updates their display automatically and highlights a focused sub-graph which to be viewed in details. This enables users to easily identify their focused region at the current context. The effectively uses of the semantic zooming to zoom in the focus articles while retaining context views which enable the interactive investigation of very large citation networks. Figure 5 shows an example of the navigation when a sub-cluster at the far-right of main cluster (at the right side) from Figure 4 is selected to enlarge. The figure also illustrates the highlight of the selected sub-cluster from the context views at the left-hand side.

C. Attributed Visualization

The *main-view* is displayed in a large area of the visualization. Rich graphical attributes are employed to in the display of clustered citation graphs to assist viewers to quickly identify the domain specific properties of data and the hierarchical structure. We first use background colors to assist viewers to quickly identify the hierarchical structure of the clustered graph. In our prototype, the local regions of nodes at different levels are painted with a same color but at different brightness (see those Figure 1, Figure 4 and Figure 5). This drawing property aims to provide a pleased view while retaining the clarity between sub-graphs.

Width of the edge is employed to represents the weight of the edge (or the number of connections between two articles or two groups of articles). For example, Figure 5 shows that the bottom-center sub-cluster has a strong connection to the top-right and bottom-right sub-clusters. This indicates a much stronger citing references among articles across these sub-clusters compared to other sub-clusters. Our system is also able to display all edges in their original connection from the abstract clustered display, i.e. citation from article to article. In addition, edges among leaf nodes are drawn with light-green color and edges among non-leaf nodes (or between two clusters) are drawn with light-gray color. This aims to provide an easy identification of the different type of edges.

Each node in our visualization is associated with a color corresponding to its publication year. Our program automatically matches colors for publication years based on an array of default colors. We first try to find a dominant period of publication years from the network. We then associate the corresponding colors for articles belonging to the dominant period. This method ensures the limited number of colors that we need to display at once in order to eliminate the confusing when too many colors are presented. Figure 4 to Figure 6 shows the display of citation network of research papers mainly from year 1995 to year 2002. In these examples, the colors corresponding to publication year of 1995 to 2002 are dark-red, chocolate, olive, green, dark-cyan, blue, and blue-violet respectively. Papers published at the other years are painted with gray color (see the color panel from Figure 4 and Figure 5).

Our visualization also presents the number of citations for articles using circles in which the area of a circle is proportional to the number of papers citing to the article. Therefore, the radius R of the circle is calculated by formula

$$R = \sqrt{\frac{n}{\pi}}$$

where n is the number of citations to the article. We also note that the area of the circle is not necessarily proportional to the number of connections (edges) in the visualization. This means that the paper could be cited by the papers from other fields and they are not included in our datasets for visualization.

The colors for the circles are same as the nodes but in light and semi-transparent condition. This drawing aims to emphasize the property of nodes but retaining the clarity of the overall display (see Figure 6 and Figure 7).

In addition, size of nodes is automatically adjusted based on the number of nodes and the screen resolution. This ensures the clarity of the visualization during the navigation (see Figure 4 to Figure 6).

Figure 6 and Figure 7 show example of our visualization of a small group of papers within a sub-cluster from social network and terrorism research domain, i.e. they are strongly connected or related. For example, Figure 6 illustrates a strongest influence from the publication of Watts, 1998 to the research community in the social network research community. Figure 7 also shows a strong concern of terrorism research community to Inglesby T.V on his publications in 2000 and 2001 titled "Anthrax as a biological weapon - Medical and public health management" and "Plague as a biological weapon - Medical and public health management" respectively. This could be because the risen of concern (after the catastrophic event of September 11th, 2001) about the homeland security and public health in the terrorism with biological weapon.

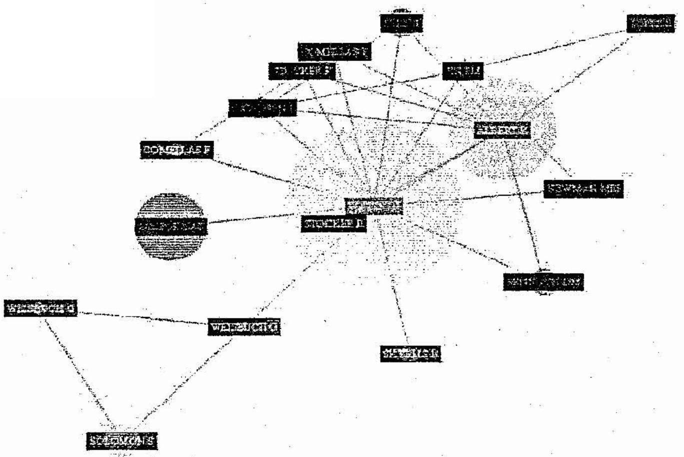
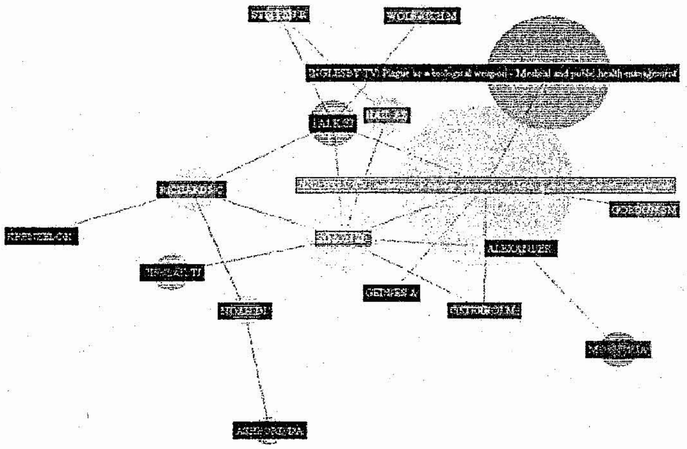


Fig. 6. The attributed visualization of a subgroup of articles Social Network research



INGLESBY TV: Plague as a biological weapon - Medical and public health management
ALEXANDER
GEDDES A

Fig. 7. . The attributed visualization of a subgroup of articles in Terrorism research

IV. Conclusion and Future Work

We have presented a technique for visualizing large citation networks using a space-efficient visualization technique. Our technique first partition a citation network into clusters of highly related articles. A new space-efficient visualization technique is then applied to archive the optimization of graph layout within the limited display space. Technically, the layout algorithm is a combination of the enclosure partitioning algorithm, *Clenccon,* and the *Spring-Embedder* algorithm which aims to handle very large citation networks. We also employ rich graphics to show the attributed property associated with articles and citing references. A *focus+context* interaction technique is also provided for the navigation through the network. Although, our technique is still at its early state and need further improvement, we believe that it is a valuable tool for researching in citation networks.

Acknowledgment

We would like to thank Prof. Chaomei Chen for providing the two datasets of citation in social network and terrorism research.

References

[1] U. Brandes, and T. Willhalm, "Visualization of bibliographic networks with a reshaped landscape metaphor", in *Joint Eurographics - IEEE TCVG Symposium on Visualization (VisSym '02).* ACM Press, 2002, pp. 159-163.

[2] M. Bruls, K. Huizing, and J. J. van Wijk, "Squarified Treemaps", in *Joint Eurographics and IEEE TCVG Symposium on Visualization.* Springer, 2000, pp. 33-42.

[3] C. Chen, "Searching for intellectual turning points: progressive knowledge domain visualization", in *Arthur M. Sackler Colloquium of the National Academy of Sciences.* The National Academy of Sciences of the USA, 2003, pp. 1-8.

[4] C. Chen. "CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature", to appear in *Journal of the American Society for Information Science and Technology,* 2006.

[5] C. Chen, C. (1999): Visualising semantic spaces and author co-citation networks in digital libraries, *Information Processing and Management,* vol 35, pp. 401-420, 1999.

[6] P. Eades, "A heuristic for graph drawing", *Congressus Numerantium,* vol 42, pp. 149-160, 1984.

[7] P. Eades and Q. Feng, "Multilevel visualization of clustered graphs", in *Graph Drawing (GD'96).* Springer, pp. 101-112, 1996.

[8] N. Elmqvist and P. Tsigas, "CiteWiz: A tool for the visualization of scientific citation networks". *Technical Report 2004-05,* Chalmers University of Technology and Göteborg University, Sweden, 2004.

[9] I. Herman, G. Melancon and M. S. Marshall, "Graph visualization in information visualization: a survey", *IEEE Transactions on Visualization and Computer Graphics,* vol 6 , pp. 24-44, 2000.

[10] B. Johnson and B. Shneiderman, "Tree-Maps: a space-filling approach to the visualization of hierarchical information structures", in *the 1991 IEEE Visualization.* IEEE, pp. 284-291, 1991.

[11] J. D. Mackinlay, R. Rao and S. K. Card, "An organic user interface for searching citation links", in *CHI 1995.* ACM Press, 1995, pp. 67-73.

[12] D. Modjeska, V. Tzerpos, P. Faloutsos and M. Faloutsos, "BIVTECI: a bibliographic visualization tool", in *the 1996 conference of the Centre for Advanced Studies on Collaborative Research (CASCON' 96).* IBM, 1996, pp. 28-37.

[13] M. E. J. Newman, "Fast algorithm for detecting community structure in networks", *Journal of Phys.* Rev. E 69, 066133, 2004.

[14] Q. V. Nguyen and M.. L. Huang, "EncCon: an approach to constructing interactive visualization of large hierarchical data", *Information Visualization Journal,* vol 4, 1, 2005, pp. 1-21.

[15] S. Noel, C. H. Chu and V. Raghavan, "Visualization of document co-citation counts", in *Sixth International Conference on Information Visualisation (IV'02).* IEEE, 2002, pp. 691-696.

[16] J. C. Robert, "On encouraging multiple views for visualization", in *International Conference on Information Visualisation (IV 1998).* IEEE, pp. 8-14, 1998.

[17] J. W. Schneider, "Naming clusters in visualization studies: parsing and filtering of noun phrases from citation contexts", in *10th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2005).* Karolinska University Press, pp. 406-416, 2005.

[18] C. Ware, *Information Visualization: Perception for Design.* Morgan Kaufmann, San Francisco, CA, 2004.

[19] H. D. White and K. W. McCain, "Visualizing a discipline: an author co-citation analysis of information science", *Journal of the American Society for Information Science,* vol 49, 4, 1998, pp. 327-355.

Proceedings

Fourth International Conference on Information Technology and Applications

ICITA 2007

15-18 January 2007

Harbin, China

Edited by

Dr. Dapeng Tien
Professor Shi Guangfan
Mr. Wang Guanran

Sponsored by

Heilongjiang University, Harbin, China
Shanghai Jiao Tong University, Shanghai, China
IEEE, NSW Section, Australia

Macquarie Scientific Publishing, Sydney, Australia