ISSN 2278-7763

Survey of visualization techniques for very large network graphs using a case study*

[1]Monica Gokule, [2]Pravin Game, [3]Deepesh Patel

[1]Department of Computer Engg, PICT, Pune, India; [2]Department of Computer Engg, PICT, Pune, India; [3]SAS R&D Pvt.Ltd, Pune, India.
Email: [1]mgokule@yahoo.com, [2]psgame@pict.edu, [3]deepesh.patel@sas.com

ABSTRACT

Novel approaches have been proposed for large network visualization. However, visualizing large networks can be quite challenging if not impossible. This is due to limitations of the screen, complexity of layout algorithms and limitations of human visual perception. Some networks are too large to hold in the memory of most desktops let alone visualize it. Even if a network fits in memory, a single iteration of a good layout algorithm like a force-directed layout can take quadratic time. In this paper, we survey available techniques for visual analysis of very large graphs.
Keywords : Visualization; Graph visualization; Visualization techniques & methodologies

1 INTRODUCTION

The ever-increasing usage of internet in all walks of life is driving exponential growth in data traffic and a corresponding proliferation of very large networks. Examples of such networks include bibliographical collections, biological networks, market basket data, social networks and even World Wide Web. Visualization of such networks can lead to better understanding and clear presentation of patterns that can often be hidden.

Basically graph visualization consists of various steps such as graph preprocessing, mapping or layout, visual user interaction, graph analysis [1]. Graph preprocessing is carried out to reduce the graph size which consists of graph filtering and graph aggregation. There are many challenges in visualization of large graphs. Major problem in graph visualization is size of the graph which is increasing exponentially over the time period. One of the prominent examples of such networks is social networks having unimaginable sizes. Visualizing graph with millions of nodes and billions of edges has problem of discernibility wherein we are unable to recognize the details hidden. Such problem can be termed as hairballs. Hairballs add up to the complexity of visualization. They can pretend to communicate high information but just because they look complex does not mean they communicate complex information.

Even if the size of the display has increased, still it puts constraints on human eye which can wander over just 550 megapixels. Hence there is limitation of screen and human vision resolution [2]. Considering all the challenges there are many techniques that has been proposed to achieve better and clear visualization of large graphs.

2 VISUALIZATION TECHNIQUES

Visualization of graphs can be either static or dynamic.
For static visualization of graph, there are many techniques such as node link diagrams, matrix representation and other hybrid approaches while for dynamic visualization there are techniques such as animation, small multiples, traces and 3D approaches [17].

2.1 Node Link diagram

In node link diagrams, vertices represent nodes and an edge represents line segments, polylines shown in figure [1]. Majorly, node link diagrams contain fewer than 30 nodes and 40 links between them which is way shorter than millions of nodes and links contained in large graphs. Node link diagrams give good result for small graph visualization where number of nodes and links are limited. They also provide good readability but as the graph becomes denser, it faces occlusion and network hairball problems, where nodes and links overlap [3]. Some task consists of sequential search of graph elements, such as node finding by name which are increasingly difficult when the number of nodes becomes large. Also, path finding for such graphs becomes difficult.

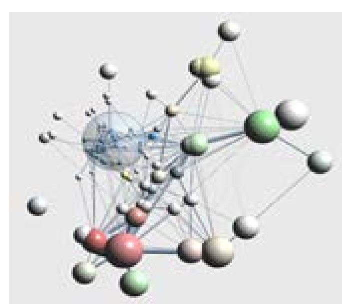

Fig.1

2.2 Matrix Representation

The graph can be represented in the form of Boolean valued connectivity matrix where rows and columns represent vertices of the graph [3].Matrix based representation offers intriguing alternative to the traditional node link diagrams. It has got two advantages, it exhibit no overlapping and is orderable. It has better readability than node link diagrams. Matrix view shows good result for large, denser graphs. Moreover, it has the ability to highlight missing connections. Matrix form of

ISSN 2278-7763

graph eliminate occlusion problem but as the number of node increases, size of the matrix also increases. Hence, path finding tasks may become difficult. For example, if two communities connected are tedious, it requires reading rows and columns back and forth alternately. Moreover matrices have significant area cost of (□(N 2)) so, it may not offer much advantage if the graph to be rendered is sparse.

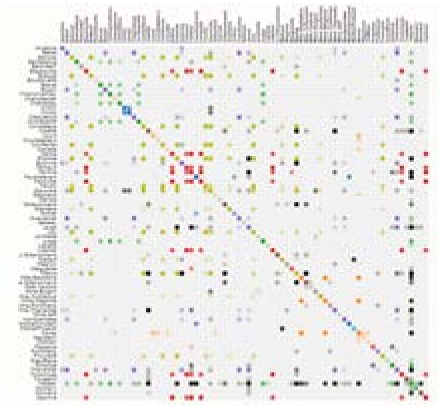

Fig. 2

2.3 ZAME

Zoomable Adjacency Matrix Explorer [4] is an exploration tool which is based on adjacency matrix. It allows analysts to dig into the details at many different levels. It mainly consists of three components. Hierarchical data structure which stores the graph attributes in aggregated form, automatic reordering wherein ordering the matrix to help visual analysis and accelerated rendering mechanism in which efficient display and caching of graph dataset is done.ZAME shown in figure 3.

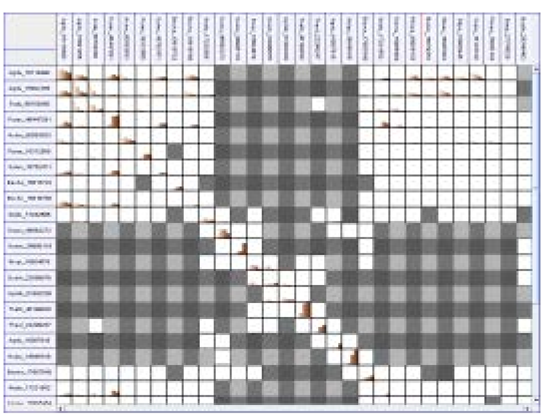

Fig. 3

2.4 NodeTrix

In Social networking graphs, vertices are actors (people) and edges are relationship. Graphs vary from very sparse to very dense. These networks face major challenge of acquiring a readable representation for both the overall sparse structure of a network and its dense communities.

NodeTrix[5] is a hybrid combination of node link diagrams and adjacency matrix representation. Here the communities represent as adjacency matrices. Intra-Community relationship uses adjacency matrices and inter-community relationship uses normal links. NodeTrix uses single level of aggregation.NodeTrix shown in figure 4.

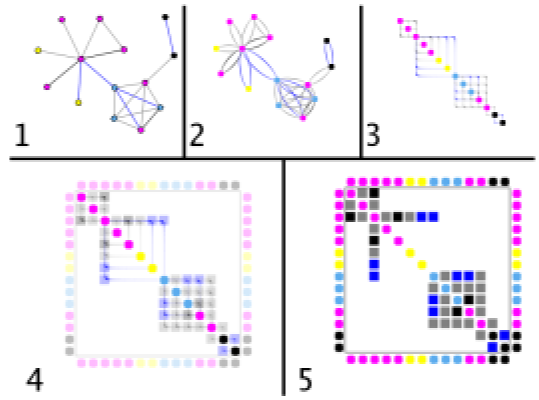

Fig.4

2.5 TreeMatrix

Clustered graph or compound graph can be thought as a rooted tree where the leaves can be considered as vertices of the graph & non leaves as cluster of nodes. Clusters are significant from user's perspective. Compound graph limits the amount of details the user can see at particular moment. If compound graph is large, then collapsing the cluster to drill down to the details can be appropriate than showing all the details at once. But collapsing the cluster can lead to loss of details. In large compound graph, edge crossing problem has to be there which makes the information indistinct.

TreeMatrix[6] is an extension to the NodeTrix. TreeMatrix addresses above problems. It makes use of adjacency matrices which eliminates edge crossing problem. Here, instead of collapsing the cluster each time, clusters are represented in the form of adjacency matrices to display the details inside. Basically, it is a hybrid representation of node link, adjacency matrix and arc diagrams.TreeMatrix shown in the figure 5 below.

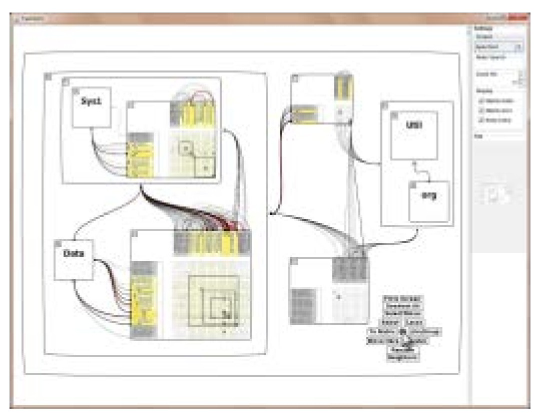

Fig.5

2.6 Hive Plot

It is a recent technique for large graph visualization. It is most powerful for dense and compact graphs when hidden patterns emerge. In hive plot [7][8], nodes are mapped onto the axis based on structural network properties. Communication links among the nodes are shown as arcs. The number of axes depends on the number of network properties we consider for visualization. The sub properties of networks can be shown using segmented axis. The large graphs containing millions of nodes and edges can be displayed over normalized axes.

We applied different dataset on hive plot using jhive[8] which is Java application for hive plot. The social networking dataset [15] consist of node IDs, their sources and communication links among the nodes. When this dataset applied on hive plot, the result shown in figure 6. The citation dataset [16] consists of paper ids and their publishing dates and also paper references. The resulting hive plot shown in figure 7. For example users have Facebook account and the same user can have twitter account. Consider these accounts as sources. The Facebook user can communicate to twitter user. The axes a1, a2, a3 shows sources of nodes and corresponding nodes are plotted on it. The tooltip for the last node on axis a3 shows node ids and number of nodes that has been clustered under one single node.

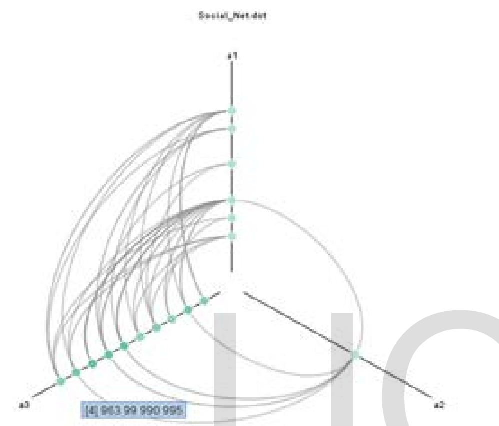

Fig.6

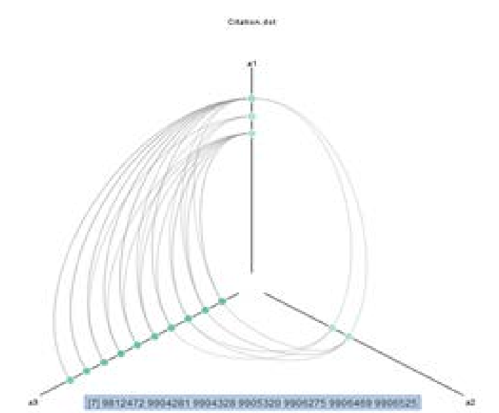

Fig.7

2.7 Arc Diagram

It is one of the techniques for visualization of large graphs. The main idea is to display the nodes along axis where edges are represented as arcs. The axis can be drawn either horizontally or vertically. Arc diagrams [9] are one dimensional layout and they may not provide overall structure of network as a two dimensional layout can.The large graphs containing millions of nodes and edges can be displayed over normalized axes.

Ordering the node is very important in arc diagrams. If the nodes are placed randomly, it may increase the arc length unnecessary. Moreover it may make visualization ambiguous. So to order the nodes there are many techniques, one of it is barycenter ordering [11]. It is an iterative technique where the average positions of the nodes are calculated and sorted according to their average positions. Nodes can either be sorted alphabetically, based on sizes, time stamp or any attribute.

There is enough room alongside nodes for long text labels or any small graphics such as pie diagrams, bar graph, and line graph to show multivariate data which is quite good advantage of using arc diagrams. But with the increase in density of arcs and intersections, the image becomes hard to read. Arc diagrams are sensitive to node ordering.

The paper IDs are plotted on horizontal axis based on their publication years. The arc represents the paper references. The color represents the cluster membership of nodes.

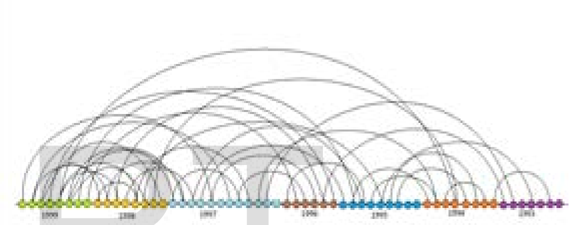

Fig.8

2.8 Circos

It is visualization tool to promote the analyzing relationships between genomes. It is circular layout. We applied social networking dataset [15] on Circos [12][13][14]. In figure 9, the outer labels indicate the sources of nodes while the color bands indicate the frequency of communication links among those sources.

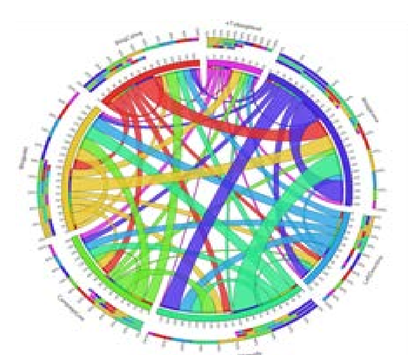

Fig.9

3 Discussion

Hive plot can directly compare two networks. It is powerful technique for large & dense graph visualization but it is directionally skeptic. In case of arc diagrams, it may not give overall structure of the network but if the nodes are ordered than it gives fair visualization of graph attributes. Moreover to show the multivariate data if any, there is abundant space to the right or below the node but if the communication links among the nodes increases then edge crossing problem may arise. Hive plot overcomes node occlusion and edge crossing problems. Arc diagrams do not facilitate displaying the frequency of communication links among the nodes while Circos can but comparison of two networks can't be done using Circos tool. Circos visualizes the data in circular layout. Circular layout is advantageous as compared to linear layout. Curved objects are always better to visually follow. Circular layouts easily elucidate the information to the user. In case of linear layouts if number of links crossings or interactions increases, it quickly leads to visual clutter.

The following table gives comparison among the techniques discussed above on which dataset is applied.

Comparision Table

Properties	**Techniques**		
	Hive Plot	***Arc Diagrams***	***Circos***
Size of graph	Powerful to handle dense & large graphs	Good for large graphs	Good for large graphs
Hairball/Visual clutter problem	No	No	No
Edge crossings problem	No edge crossing problem	Yes, if nodes are not properly ordered	No edge crossing problem
Layout	Linear	Linear	Circular
Direction property	Directionally agnostic	Arcs directed from left to right	Directed graphs can be represented
Frequency of communication links	It can show	It can't show	It can show

4 Conclusion

We discussed novel approaches towards the large graph visualization. These techniques provide fair visualization but there are major problems dealing with large graph data.

There is a need of generic technique which should be non-specific to any application and can render a large dataset without sacrificing the clarity of presentation and overcoming the problems in existing techniques. Circos exhibits many advantageous features but having some limitations. So the technique which will exploit these features and give the bird's eye view to the network structure is needed.

5 References

[1] T. von Landesberger and A. Kuijper," Visual Analysis of Large Graphs: State of the art and Future Research challanges",Computer Graphics Forum 30,2011

[2] Yifan Hu, "Current and Future Challenges in the Visualization of Large Networks", AT&T Labs –Research, unpublished.

[3] Mohammad Ghoniem, Jean-Daniel Fekete and Philippe Castagliola, "A Comparision of the Readability of Graphs using Node-Link and Matrix-Based Representations", Information Visualization,IEEE Symposium,pp.17-24,2004.

[4] Niklas Elmqvist and Thanh-Nghi Do, "ZAME: Interactive Large-Scale Graph visualization," INRIA, University Paris & University of Sydney, IEEE Pacific, Visualization Symposium, 2008, pp.215-222.

[5] Nathalie Henry, Jean-Daniel Fekete and Michael J.McGuffin, "NodeTrix: A Hybrid Visualization of Social Networks", Visualization and Computer Graphics, IEEE Transaction on vol. 13, pp. 1302–1309, December 2007.

[6] Sébastien Rufiange, Michael J. McGuffin and Christopher P. Fuhrman,"TreeMatrix: A Hybrid Visualization of Compound Graphs", Computer Graphics forum, vol. 21, pp.1-13, 2012.

[7] Martin Krzywinski, Inanc Birol, Steven JM Jones, and Marco A Marra,"Hive Plot-rational approach to visualizing networks", Oxford University, 2011.

[8] Hive plot.http://www.hiveplot.com, Referred date: Oct 9, 2013

[9] M.Wattenberg,"Arc diagrams: visualizing structure in strings", Information Visualization, IEEE, pp.110-116, 2002.

[10] Michael J.McGuffin, "Simple Algorithms for Network Visualization: A Tutorial", Tsinghua Science and Technology, vol.17, pp.1-16, August 2012.

[11] Circos.http://mkweb.bcgsc.ca/tableviewer/. Referred date: Nov 15, 2013

[12] Circos.http://circus.ca, Referred date: Nov 15, 2013

[13] M.Krzywinski, J.Schein, I.Birol, J.Connors,R. Gascoyne,"Circos: An information aesthetic for comparative genomics", Cold Spring Harbor Laboratory Press,June 2009

[14] Social Network dataset. http://networkdata.ics.uci.edu/data/polblogs/polblogs.gml. Referred date:26 Oct,2013

[15] Citation dataset. http://snap.stanford.edu/data/ Referred date:26 Oct,2013

[16] Daniel Archambault, Helen C and Bruno Pinaud, "Animation, Small Multiples, and the Effect of Mental Map Prservation in Dynamic Graphs", IEEE Trans Visualization Computer Graphics. May 2010.