SIGMA: A Statistical Interface for Graph Manipulation and Analysis

Greg Meyer, Yun Teng, Brynjar Gretarsson, Svetlin Bostandjiev, John O'Donovan, Tobias Höllerer

Department of Computer Science, University of California, Santa Barbara
meyer.greg.pro@gmail.com, {brynjar,alex,jod,holl}@cs.ucsb.edu

***Abstract*—In this paper, we present a statistical approach for gaining deep understanding of a graph visualization. The approach follows Shneiderman's vision that "visualizations simplify the statistical results, facilitating sense-making and discovery of features such as distributions, patterns, trends, gaps and outliers."[3]. Thus, the addition of statistical metrics within a graph visualization tool efficiently improves exploratory data analysis and allow analysts to discover new interesting relationships between entities. In addition, we believe that a statistical interface can play a role as a navigation control for a large graph visualization. This paper presents a discussion of design and implementation of the *SIGMA* statistics visualization module for WiGis [2], and example use cases showing how statistical views help to increase a user's understanding of graph visualizations.**

I. Introduction

With the advancement of rich internet technologies, and explosion of network data available on the web, interactive graphs are becoming more common as a means to explore, navigate and understand network data. However, scalability of these tools remains a weak point for most existing systems. For node-link graphs of more than a few hundred connected entities, most client-side graph visualization tools begin to slow down considerably. Usually network visualization tools use layout or clustering algorithms in order to clarify a chaotic visualization. However, this approach traditionally does not scale well. Now that the need for statistical graph representation has been motivated, we provide a brief discussion of related work and describe the design challenges and implementation choices used in the development of the *SIGMA* statistical interface.

A. Statistics in Graph Visualizations

Graph visualization tools abound, some popular examples include TouchGraph Navigator [5], Tom Sawyer [4] , GraphDice [1], and IBM's ManyEyes. These tools base their principal functionalities either on network visualization or statistical analysis but still do not support visualization of a broad scope of statistical functions in an interactive manner. Our novel statistical viewer and navigator, SIGMA, is implemented as a module for our existing graph visualization toolkit known as WiGis [2]. *SIGMA* is focused on the coupling of statistical views and node-link representations of data. WiGis was chosen as a supporting platform for this work because it has a modular design, making plug-in development easy. Addition of a statistics module supports novel research involving interactive analysis of large scale data, graph comparisons, classification and decomposition.

By enabling statistical analysis and control in a graph visualization tool such WiGis, we aim to provide a user with a rapid overview of data contained in a graph. Moreover, the goal is to highlight important components, nodes, edges or relationships, and anomalies that are not obvious in the traditional network view. Statistics allow for simplification/abstraction of a view, which can then support on-demand dynamic focus based on simple interactions, such as moving, deleting, clustering, zooming or running various algorithms to give the user a deeper understanding of the underlying data.

B. SIGMA Module

There is a huge variety of statistical analysis methods that can support graph analysis. In this initial work, we categorized candidate statistical methods into four groups to better organize the analytical process for the end user as they interact with a graph visualization. A *Global* statistics panel persists in every view, containing statistical metrics that apply to the entire graph. If two nodes are selected, a *Pairwise* panel appears containing metrics related to the selected node pair, such as Dijkstra's shortest path algorithm, for instance. Upon selection of a group of nodes, a *SubGraph* panel, gathering statistics for all selected nodes appears in the view. Similarly, when a single node is selected, a *Node* panel appears, giving statistical metrics for the current node. The following list describes the currently implemented statistics in the module :

1) Global Statistics:

- *Size*: Information about graph size in terms of nodes, edges and content.
- *Components*: Listing of distinct components.
- *Node Types*: List and number of nodes for each different type.
- *Degree Distribution*: Shows an interactive chart of degree distribution for the entire graph.
- *Average Path Length*: Average length over all paths.
- *Average Degree Centrality*: Average degree over all nodes.
- *Average In-Out Degree (only for directed graphs)*: Average number of in and out degree over all nodes.

2) Subgraph Statistics:

- *Average Path Length*: Average path length of the selected nodes.
- *Average Degree Centrality*: Average degree over all selected nodes.
- *Degree Distribution*: Presents a real-time degree distribution chart over all selected nodes.

3) Pairwise Statistics:

- *Shortest Path Distance*: Minimum hops number between two nodes (Shortest Path highlighted on the graph).

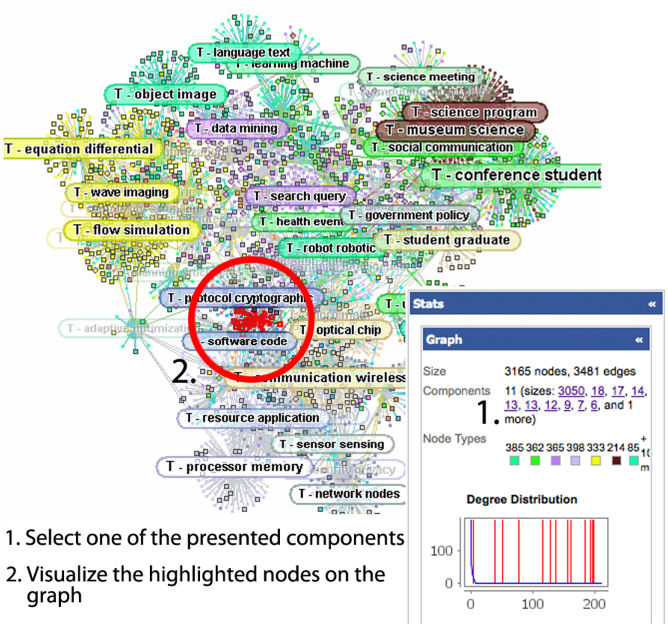
4) Node Statistics:

- *Degree*: Node's degree.
- *Neighbors*: List of all neighbors for a selected node.
- *In-Degree (only for directed nodes)*: Number of in-coming edges.
- *Out-Degree (only for directed nodes)*: Number of out-coming edges.

II. Use Cases

Since this work is in an early stage, a full live user evaluation is not yet available. To motivate our approach, two practical use cases from different domains are presented here. Each case describes a concrete example of the statistical viewer supporting discovery of previously hidden information, not easily detectable though the standard graph visualization interface.

A. NSF Dataset



Stats
Graph
Size 3165 nodes, 3481 edges
Components 11 (sizes: 3050, 18, 17, 14, 13, 13, 12, 9, 7, 6, and 1 more)
Node Types 385 362 365 398 333 214 85

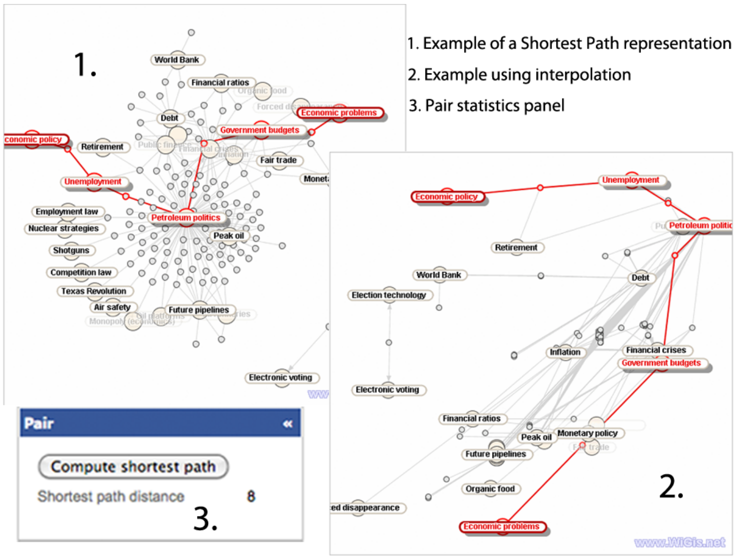
1. Select one of the presented components
2. Visualize the highlighted nodes on the graph

Fig. 1. Use of *Graph* statistics panel to discover isolated graph components

Figure 1 shows a visualization of a collection of awarded NSF grant proposals, showing the documents and topics that they relate to, based on Latent Dirichlet Analysis (LDA) modeling over their contents. The initial view shows the full graph. By clicking on elements in the *Graph* statistics panel, a user can quickly distinguish separate components that were previously hidden. With this feature, it is also possible to click on a disconnected component to highlight each of that node's neighbors, and reveal its graph position. When this occurs, statistical metrics such as the degree distribution chart shown in Figure 1 are automatically recomputed.

B. New York Times Dataset

Figure 2 shows a similar network of documents and topics from LDA analysis, in this case, a collection of New York Times news articles. A user selects a pair of nodes at random and the *Pairwise* panel appears, showing algorithms related to pairwise analysis. The user selects a shortest path algorithm, and in the main graph view, the set of shortest paths between the two selected nodes are highlighted in red, as illustrated in Figure 2. Figure 2 a.) and b.) shows a comparison between two different interaction methods for finding the shortest path between two nodes. In view a.), a user has selected the node pair and dragged them to the right of the screen. An *interpolation-based layout* moves all of the other nodes by a relative amount in the same direction, based on their graph distance from the moved node. The result highlights the shortest path on the right side of Figure 2 a.). Note that this view is specific to the two target nodes only, and all other nodes become clustered based on their graph distance from the selected node pair. Contrastingly, the view in 2 b.) shows a global *force-directed* layout of the graph, in which the shortest path is highlighted between two nodes. This view was arrived at simply by clicking on the button provided in the pairwise statistics panel that appeared on selection of the two nodes. Both views support shortest path analysis, but enable discovery of very different information about the graph.



Shortest path distance 8

1. Example of a Shortest Path representation
2. Example using interpolation
3. Pair statistics panel

Fig. 2. Two different representations of a shortest path algorithm a.) WiGis interpolation based view (pair-specific). b.) force-directed in a global view using SIGMA.

III. Conclusion

In this research abstract, we have discussed initial work on *SIGMA*, a statistical analysis tool for the WiGis visualization framework. Design details and two use cases on diverse data sets have been presented. Application of statistical analysis and navigation mechanisms to graph visualization tool such as *WiGis* improves a users understanding of the underlying graph. The *SIGMA* module allows a user to focus in on data that may have otherwise remained hidden in a traditional visualization such as force-directed node-link layout. In addition, the visualization tool improves and simplifies the comprehension of statistical metrics by allowing a user to see the results of a statistical method appear on the graph.

References

[1] Anastasia Bezerianos, Fanny Chevalier, Pierre Dragicevic, Niklas Elmqvist, and Jean-Daniel Fekete. Graphdice: A system for exploring multivariate social networks. *Computer Graphics Forum (Proc. EuroVis 2010)*, 29(3):863–872, 2010.

[2] Brynjar Gretarsson, Svetlin Bostandjiev, John ODonovan, and Tobias Höllerer. Wigis: A framework for scalable web-based interactive graph visualizations. In David Eppstein and Emden Gansner, editors, *Graph Drawing*, volume 5849 of *Lecture Notes in Computer Science*, pages 119–134. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-11805-0-13.

[3] Adam Perer and Ben Shneiderman. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 265–274, New York, NY, USA, 2008. ACM.

[4] Tom Sawyer Software. Tom Sawyer Visualization, 2009. Available at www.tomsawyer.com.

[5] Touchgraph. Touchgraph llc, http://www.touchgraph.com, 2004.