Social Networks 37 (2014) 56–64

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Visualization techniques for categorical analysis of social networks with multiple edge sets

Tarik Crnovrsanin[a,*], Chris W. Muelder[a], Robert Faris[b], Diane Felmlee[c], Kwan-Liu Ma[a]

[a] Department of Computer Science, 2063 Kemper Hall, University of California-Davis, One Shields Avenue, Davis, CA 95616, United States
[b] Department of Sociology, 2247 Social Science and Humanities, University of California-Davis, One Shields Avenue, Davis, CA 95616, United States
[c] Department of Sociology, Pennsylvania State University, Oswald Tower, University Park, PA 16802, United States

ARTICLE INFO

Keywords:
Social networks
Information visualization
Network analytics
Graph layouts

ABSTRACT

The growing popularity and diversity of social network applications present new opportunities as well as new challenges. The resulting social networks have high value to business intelligence, sociological studies, organizational studies, epidemical studies, etc. The ability to explore and extract information of interest from the networks is thus crucial. However, these networks are often large and composed of multi-categorical nodes and edges, making it difficult to visualize and reason with conventional methods. In this paper, we show how to combine statistical methods with visualization to address these challenges, and how to arrange layouts differently to better bring out different aspects of the networks. We applied our methods to several social networks to demonstrate their effectiveness in characterizing the networks and clarifying the structures of interest, leading to new findings.

Published by Elsevier B.V.

1. Introduction

Social network research is one of the fastest growing academic areas (Rivera et al., 2012) and it continues to expand within an array of social, physical, and biological sciences. One key element of this field of research is social network visualization, which refers to the use of "sociograms," or illustrative diagrams of the ties that connect actors in social networks. The use of graphical representations is one of the main defining properties of the field of social networks (Linton, 2004). While statistical metrics can more succinct, the right metric must be applied. It can be difficult to know a priori what metric will produce the right result, and it can be difficult to verify that the results are correct. Researchers use pictorial images of social networks to help successfully communicate and understand the content of the network and also to aid in uncovering novel, structural patterns within social networks, as well as to guide and confirm statistical metrics. Nevertheless, visual diagrams of social networks often suffer from a range of problems, the most common of which being the high density of edges and complex structures in large networks, yielding sociograms that often appear as indecipherable clouds of nodes and edges.

In the study of aggression networks (Faris and Felmlee, 2011), we identified visualization techniques that can address problems typical to social network visualization, and enhanced the techniques to improve clarity and highlight key structural elements of aggression network. In particular, we considered social networks composed of nodes that can be grouped categorically (i.e., students can be categorized by gender, grade, etc.). Similarly, the edges in a social network can often be divided according to categories (e.g. a friendship is different from an aggression relationship). We used the most common type of visualization, which directly represents relationships between actors as a node-link diagram. That is, the resulting sociograms represent actors with the use of points, or vertices, and the relationships between these actors with the use of lines, or edges, that connect these points. In this paper, we present several visualization techniques tailored to further analyze such social networks. We show how we incorporate statistical measures such as sensitivity analysis to filter nodes/edges from a node-link diagram leading to succinct visualizations, and how different layout designs help bring out structures of interest that would otherwise hidden. We demonstrate several enhanced visualization techniques that enable us to better understand and explain our empirical social network data, and also derive new findings.

2. Related work

Visual diagrams of network-related data have a lengthy history (Linton, 2004). The origins of their application in social network research began with the work of Moreno and Jennings in the early 1930s, in which the typical focus was on examining patterns of individuals' likes and dislikes (Moreno, 1953). Since those early

* Corresponding author. Tel.: +1 916 995 2025.
E-mail addresses: tecrnovr@ucdavis.edu, turokhunter@gmail.com (T. Crnovrsanin).

0378-8733/$ – see front matter. Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.socnet.2013.12.002

beginnings, the use of network visualization has developed into its own specialty field (Freeman, 2000; Brandes and Pich, 2006). This area of study currently receives substantial attention in a range of disciplines (Freeman et al., 1998; Brandes et al., 2001; Di Battista et al., 1994; Krempel et al., 2003; McGrath and Blythe, 2004). Considerable recent work focuses on dynamic visualizations of change in social networks over time (e.g. Moody et al., 2005; Demoll and Mcfarland, 2005; Brandes, 2011; Sallaberry et al., 2012). Other studies examine methods of producing images of web-based social networks (Heer and Boyd, 2005). Finally, researchers continue to develop novel software programs for the production of network images, each with its own unique characteristics, and several publications illustrate the implementation of novel software routines (Brandes, 2011; Heer and Boyd, 2005).

One key task in creating visual images of networks is to determine the appropriate geometrical layout of the nodes and edges. There are several well-defined criteria for assessing the accuracy and validity of a particular graph layout (Demoll and Mcfarland, 2005). Some common criteria (Brandes, 2011; Bertin, 1983) include:

1 edges of the same approximate length,
2 vertices distributed over the area, or
3 reduction of the number of edge crossings.

Nevertheless, optimization of such criteria can be intractable and often contradictory (Brandes, 2011). For surveys of many modern graph layout algorithms see Tollis et al. (1999) or Hachul and Jünger (2005).

The most traditional and commonly used layout algorithm for social network analysis are force-directed layouts (Kamada and Kawai, 1989), often referred to as "spring embedders" (Eades, 1984). In this well-known procedure, nodes in a network graph are positioned iteratively, where the edges connecting them are treated like springs that push and pull on them until the system converges to an equilibrium. By directly optimizing on these criteria, force-directed layouts aim both to distribute nodes widely in a two-dimensional space, and to simultaneously keep connected nodes relatively close to each other.

However, spring embedder techniques do not always scale nicely to large graphs (Brandes and Pich, 2009). Other approaches have been developed with the goal to improve network layout in terms of quality and algorithmic efficiency, especially for large graphs. One such technique (Brandes, 2011) is based on a variant of dimension-reduction methods, referred to as multidimensional scaling (Cox and Cox, 2001), in which the goal is to minimize stress. In this approach, the purpose of stress minimization is to determine positions for every node such that the Euclidean distances in the n-dimensional space resemble the given "dissimilarities" between the nodes, where dissimilarity is determined by graph-theoretic distances, such as the shortest paths (i.e. geodesics).

A fundamental problem that faces visualization of very large social networks, particularly those that use force-directed layouts, is that they often result in a tangled mess of incomprehensible lines; this is often referred to as the "hair-ball" problem. In this paper, we describe two analytic approaches to reduce clutter and produce cleaner network visualizations. First, in order to simplify the contents of a social network, we employ a type of sensitivity analysis that is based on commonly used, graph theoretic, network centrality measures. The findings from the sensitivity analysis (Correa et al., 2012) are then used in traditional graph layouts and node-link diagrams. Second, we employ a type of hierarchical clustering procedure called modularity clustering (Clauset et al., 2004) in order to create an abstraction of a network that is particularly useful in identifying higher level structures.

In this paper, we also show how to apply these analytic strategies in the application of three visual design techniques. The first technique is referred to as "edge bundling" (Danny, 2006). This technique routes similar edges together, which produces cleaner network displays. Next, we introduce a radial layout design that can effectively separate a graph into sub-groups, or communities, for an effective display of network sub-structure. Finally, we introduce the use of "n-partite network layouts" based on parallel coordinate diagrams, which we use to directly compare two or more distinct graphs or subgraphs, defined on the same set of nodes.

In the subsequent two sections, we introduce the techniques we chose to use and explain why and how we enhance them for achieve our goals. Then in the following section, we focus on the study of an aggression network dataset using these techniques. Here we investigate patterns of aggression and friendship among high school students and use visual sociograms to help address questions such as the following:

1 Which students are most likely to be the aggressors, and victims of aggression – those located on the periphery of the friendship network, or those located more centrally?
2 Are there differences by structural factors in patterns of aggression, such as grade level, gender, and race?
3 Do the bulk of aggressive ties occur among or between racial groups?

The techniques that we use are designed to visually reveal the answers to these types of questions.

3. Analysis techniques

To reduce clutter and produce cleaner network visualizations, we apply two analytic approaches. First, we show the use of centrality sensitivity analysis, which measures the importance of one node with respect to another. The aim of this technique is to simplify a network based on centrality metrics, which can then be represented using traditional graph layouts and node-link diagrams. Second, we utilize modularity based clustering, which separates nodes into groups based on the intra and inter group connections. This creates a hierarchical abstraction of a network that we can use to depict higher level structures more clearly.

3.1. Sensitivity analysis

There are four commonly used centrality metrics: Eigenvector (Brin and Page, 1998; Kleinberg, 1999), Markov (White and Smyth, 2003), Betweenness (Lister, 2008; Freeman, 1979), and Closeness (Jacob et al., 2005; Newman, 2003). Each of these measure vertices' overall importance with respect to the whole network. Sensitivity analysis measures a vertex's importance to the structure of the network relative to other vertices in the graph (Correa et al., 2012). This metric is essentially the derivative of centrality, and as such can be calculated similarly for any type of centrality. In this work, we used Eigenvector sensitivity. Eigenvector centrality is a measure of the importance of a node in a network, and is used by the PageRank (Brin and Page, 1998) and Hyperlink-Induced Topic Search (Kleinberg, 1999) algorithms. Rather than basing the importance of a node solely on how many connections it has, eigenvector centrality also takes into account the weights of connections to other nodes; a single connection to a highly important node can carry more weight than many connections to nodes of low importance. Eigenvector centrality sensitivity extends this notion to derive the importance of nodes relative to each other. While centrality gives one value per node, sensitivity gives a value for every possible pair of nodes in a network. To calculate a reference node's sensitivity to