Katie Chen
Cindy Chiu
Keri Mallari
CS&SS 569

# Visualizing Network Data

**Background on Social Network Analysis**

Social Network Analysis (SNA) has its theoretical origins in the work of early sociologists Georg Simmel and Emile Durkheim, where they emphasized the formal properties of social relations and the configurations of social relationships. Lewin and Moreno scoped this work by investigating specifically the 'space' of social relations and its characteristics as a 'network' [1].

Social Network Analysis is useful for *investigations* of kinship patterns, community structure, interlocking directorships, and more [2]. The principal types of data are 'attribute data' and 'relational data.' Attribute data relates to the attitudes, opinions and behaviors of agents. For example, data collected through surveys and interviews are often referred to as attributes of individuals that can be quantified for statistical analyses. On the other hand, relational data are the contacts, ties, and connections between agents such as group attachments and meetings. These relationships cannot be simplified to an individual agent. While there are prescribed analyses for these types of data, they are often collected alongside one another as parts of the same study.

Relational data is central to the principal concerns of sociology, where their primary concern is on the structure of social actions. Structures are built from relations, and structural concerns are investigated through collection and analysis of relational data. Social network analysis developed from Radcliffe-Brown's concept of 'social structure' and began investigating the 'fabric' and 'web' of social life. From these metaphors of understanding 'interweaving' and 'interlocking' relations in which social actions were organized, social networks emerged and researchers began to investigate the 'density' and 'texture' of social networks.

**Applications of Social Network Analysis**

Social Network Analysis is used extensively in our data rich society across a wide range of applications and disciplines.

Social Media Analysis is one of the most popular applications of SNA to understand behavior between users and organizations on websites such as twitter, facebook, and yahoo. SNA is also used to explore security risks such as leakage of data, mistreatment

of user data, and access limitations of users by tracking the suspected and genuine users in a social media network.
Education and Health are also domains where SNA is used such as by tracking social relations among students and teachers, and social relations between patient and doctor. Studies have also utilized SNA to analyze the spread of diseases.
Large textual analysis also benefits from the SNA with the introduction of Quantitative Analysis, where nodes are formed by noun phrases and links by verbs, directly expressing the action of one node upon the other [4].

**Statistical Models for Social Network Data**
There are several metrics that can be used to describe features of network data. Latent Space Model and Exponential Graph Model are two statistical models that commonly used to analyze data about social and other networks.

1. Latent Space Model

One of the statistical models that models social network data is latent space models developed by Hoff, Raftery and Handcock. The idea of this method is to measure the probability of forming a relational tie (edge) between two individuals (nodes) depending on their distance in the latent space. In Hoff's article, the latent space, also called social space, is defined as "a space of unobserved latent characteristics that represent potential transitive tendencies in network relations." [8] With the assumption of the presence of a relational tie between two individuals is independent of other ties, the conditional probability of adjacency matrix $Y$ is

$$P(Y|Z,X,\theta) = \prod_{i \neq j} P(y_{ij}|z_i, z_j, x_{ij}, \theta)$$

where $N$ is the set of nodes, $Y$ is the edges with $y_{ij}$ represents the edge value between node $i$ and $j$, $Z = (z_i)_{i \in N}$ is the unobserved latent positions, $X = (x_{ij})_{i,j \in N}$ is the observed pair-specific covariates, $\theta$ is regression parameters [9].
There are two models to measure $P(Y|Z,X,\theta)$, which is distance model and projection model. In general, individuals that share similar characteristics usually get closer in the social space, and therefore the probability of having a tie between them is higher. Distance model uses the latent position between nodes to measure the probability. The probability of forming a tie is also higher if the characteristics of two nodes have the same direction (angle between them is small), and less likely to form a tie if the angle between two nodes is large. Distance model uses the size of angle between two nodes to calculate the probability [8].
In R, there is a package called *latentnet* for estimating latent space models. It is used to estimate nodes' latent positions in the social space.

2. Exponential Random Graph Model (ERGM)

Another statistical model for network data is exponential random graph model, first proposed by Holland and Leinhardt in the early 1980s. It is commonly used in social network analysis.

When constructing the exponential random graph model, the choice of parameters is based on the interest of the observed network. For example, if we are interested in the reciprocate relationship, then reciprocity can be one of the parameters in the model. When fitting this model to the network data, graphs with more reciprocate relationship would get higher probability than graphs with little reciprocity [10].

The general form of exponential random graph model for predicting the presence of tie between nodes is

$$P(Y|\theta) = \frac{1}{c(\theta)} exp(\theta^T s(y))$$

where $Y$ is the adjacency matrix with $y_{ij}$ represents the edge value between node $i$ and $j$, $s(y)$ is the graph statistics which can be chosen by interested property of network graph (i.e., density, reciprocity, attributes), $c(\theta)$ is the normalising constant, and $\theta$ is the model parameters indicate whether there is a small or large amount of statistics. $\theta$ can be estimated by MLE, which makes the observed network most likely.

In R, there is a package called *ergm* that is used to analyze network data based on exponential random graph models.

**Challenges and Improvement of Network Visualization**

In order to include network visualizations in scientific paper, the graphs are restricted to two dimensions. However, for large and complicated networks, the conventional hairball diagrams are not presenting well in two dimensions. The audience can not gain insight from the complicated and overlapping lines. The following section discusses a few methods to improve the graph and provide better interpretation to the audience.

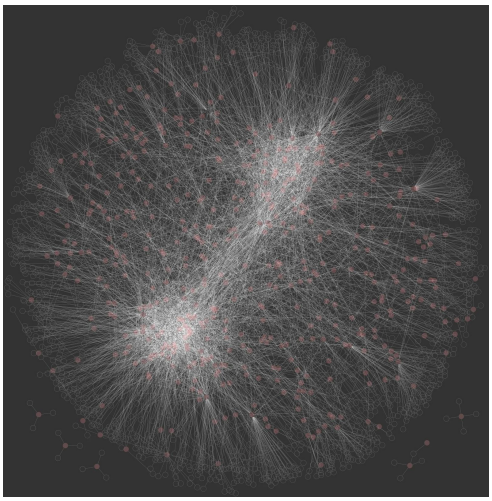

**Figure 1.** Conventional Hairball Network Diagram

1. Dimension Reduction

The first method is Dimension Reduction. It is taking the high-dimension data and projecting them to a lower dimension. The diagram creators can use Principal Component Analysis to determine which dimensions capture the most information and compose the diagram to those dimensions. Dimension reduction can help capture as much information as possible and fit them within two dimensions. It can also reduce the complexity of the diagram.

2. Featured Based Layout

Another possible solution is to focus on certain features and emphasize the network visualization using the certain feature. When the nodes have spatial attributes, the most common method is to overlay a layer of geometric image, such as map, to the network visualization [5]. This method emphasizes geometric relations between each node. However, this method might hide other interesting findings and relationships other than geometry for the nodes and edges.

Krzywinski et al. have created hive plots in attempt to address the shortcomings of the traditional network plots. The hive plot is a rational visualization method for drawing networks. In hive plots, nodes are constrained to linear axes and edges are drawn as curved links. The node-to-axis assignment and position are determined by the creator based on the properties they are interested in [6]. This plot can show the most important feature the creators are interested in. Also, hive plots uncover structure in the graph that could not be previously seen with other network diagrams for large datasets. Hive plots also work well in both small size of the data and easily scalable even with complex data set. However, it will lose some dimension features, such as geometry. Hive plots are also not aesthetically awe-inspiring compared to other network visualizations.

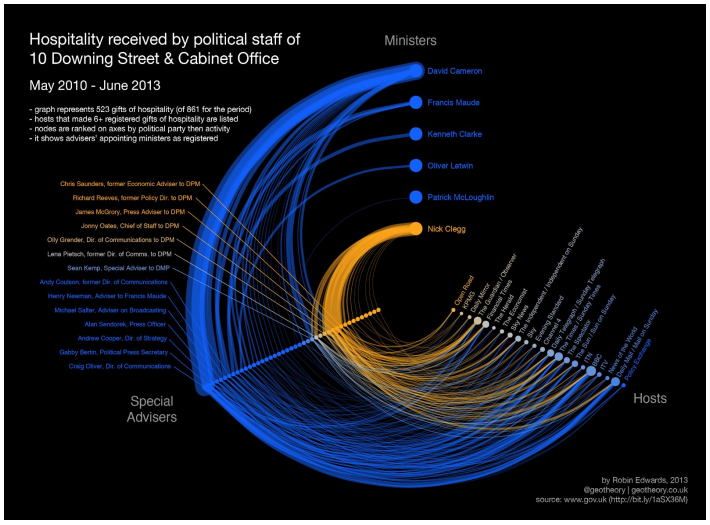


**Figure 2.** Hive Plot

3. Matrix Display

Another method to depict the relationships between discrete categorical values is matrix display. Each node is assigned to a one row and column in matrix format. The (i,j) and ( j,i) matrix elements associated with the j-to-i and i-to-j links. Like link maps, matrix display also concentrates on the link of the network [7]. The advantage of matrix display is that it resolves the problem of observing overlapping lines in conventional network diagrams. It transforms the thickness of the links into the saturation of the color blocks. It can also show the clusters by ordering the nodes with similar behavior in sequential orders. However, there are some limitations with matrix display. For instance, the matrix display gives up the geography of the nodes. The creator can arrange the nodes in approximate geographical order but the audience can not distinguish the geography at a glance.

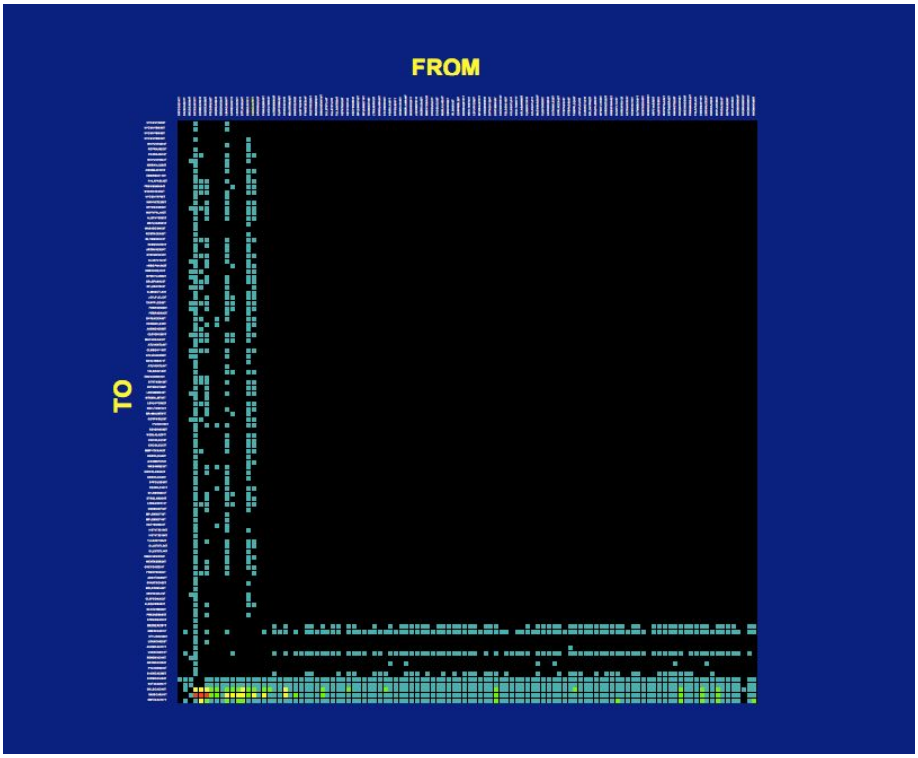


**Figure 3.** Matrix Display

If the data visualization is not limited to two dimensions, some challenges in two dimensions will be less of an issue in three dimensional graphs. For instance, edge crossing in complex networks will be less of a problem. Instead of trying to fit multiple dimensions to two dimensions, the creator can set up interactive tuning parameters that enable the user to filter on the dimensions they are interested in [6].

References

[1] Scott, J., & Carrington, P. J. (2011). *The SAGE handbook of social network analysis*. SAGE publications.
[2] Scott, J. (1988). Social network analysis. *Sociology*, *22*(1), 109-127.
[3] Golbeck, J. (2013). *Analyzing the social web*. Newnes.
[4] Sudhahar, S., Veltri, G. A., & Cristianini, N. (2015). Automated analysis of the US presidential elections using Big Data and network analysis. *Big Data & Society*, *2*(1), 2053951715572916.
[5] Gibson, H., Faith, J., & Vickers, P. (2013). A survey of two-dimensional graph layout techniques for information visualisation. *Information visualization*, *12*(3-4), 324-357.
[6] Becker, R. A., Eick, S. G., & Wilks, A. R. (1995). Visualizing Network Data. *IEEE Transactions on Visualization And Computer Graphics,* 16-21.
[7] Krzywinski M., Birol I., Jones S., Marra M. (2011). Hive Plots — Rational Approach to Visualizing Networks. *Briefings in Bioinformatics*
[8] Hoff, Peter D; Raftery, Adrian E; Handcock, Mark S. “Latent Space Approaches to Social Network Analysis.” *Journal of the American Statistical Association*, 01 December 2002, Vol.97(460), pp.1090-1098.
[9] Kim, Bomin, et al. “A Review of Dynamic Network Models with Latent Variables.” *Statistics Surveys*, vol. 12, 30 May 2018, pp.105-135., doi:10.1214/18-ss121.
[10] Robins, Garry, et al. “An introduction to exponential random graph (p*) models for social networks.” *Social Networks*, May 2007, Vol 29, pp.173-191.