Analyzing Social Network using Gephi

Khaing Phyo Wai, Ei Ei Thu
University of Computer Studies, Mandalay
khaingphyowai.mm@gmail.com, eieithuet@gmail.com

Abstract

Social network analysis (SNA) is the study of social networks to understand their structure and behavior. The study of the social networks is the possibility of collecting web log data. The data used for analyzing social networks is relational data, which can be obtained from different resources including content available on web pages, user interaction logs and social interaction information provided by users. This paper used social interaction information for determining the relationship between users in the social networks using Gephi. An application called netvizz was used in order to get the web log from social network. This analysis also examines how SNA methods and tool can be used to evaluate degree distribution and relationships between friends in a social network.

***Keywords**: Social Network Analysis, Gephi*

1. Introduction

A social network is a social structure of people, related directly or indirectly to each other through a common relation or interest [7]. Social networks are increasingly gaining importance in the day-to-day living of Internet users [6]. In recent years, SNA has increasingly been used as an approach in scientific disciplines and has gained attention by the public. SNA is based on an assumption of the importance of relationships among interacting units or nodes [1]. These relationships are defined by linkages among units or nodes and are a fundamental component of SNA [2]. Social networks can be mathematically modeled as graphs and, thus, graph theory has become inextricably related to social network analysis with a long history of research [10].

Relational data can be visualized in matrix form or in graphic form. A network is a set of nodes connected by a set of ties. The nodes can be anything including individuals, teams, organizations and groups with common ideas [3]. Ties connected pairs of nodes can be directed or undirected and can be dichotomous or weighted. When network analysts collect data on ties from a set of nodes, they call it relational data.

Social network analysis tools can be used to analyze a social network. The analysis targets are mainly focused on resources from the web, such as its content, structures and the user behaviors. Application of data mining techniques to the World Wide Web, referred to as Web mining, can also be used for the analysis of social networks.

2. Literature Review

SNA is a unique methodology with its own version of data collection, statistical analysis, and presentation of the results [11]. The most important tenet is that it enables researchers, practitioners, and educators to see how actors are located or embedded in the overall network [5]. This way of thinking creates an advantage of multilevel analysis [12]. Its methodology enables the analysis of relationships between individuals, groups, teams, cliques, agencies, and organizations. Thus, most of the time, a network analyst would be concerned with how an actor is located in the network and how that very structure is created by the relationships among those actors [15]. With the increase of network-related research and its applications, several analysis tools were developed to examine social network dynamics [9], [17].

3. Types of Social Networks

Social networks that exist today come in many varieties, but they share a common goal, they are all built to empower relationship building. Some exist to help you discover people and content, some are strictly focused on communication, and some are focused on showcasing creativity each of which convey different aspects of your personality.

3.1. Personal Networks

This social network exists to help you stay connected with existing relationships by sharing important moments with friends. These networks thrive on personal content. Whether sharing a milestone, an interesting article, or a location based check-in with a friend, this is the place to do it.

3.2. Content sharing Networks

A content sharing network is often used to have larger scale conversations. Although these networks help nature both new and existing relationships, they tend to be the best place to start to build new relationships and grow your audience organically. This networks use a combination of personal and professional content to show more dimensions of your personality and different aspects of your life. It can be intimidating to enter the blurry territory that sits between personal and professional but a lack of social media presence can harm you more than help you. Content sharing networks can be used as tool to further your relationships with professional connections.

3.3. Shared Interest Communities

Shared interest networks are very community-oriented, informative by nature, and driven by both professional and personal interests. People often use these communities to learn about a skill, showcase things they have learned, or stay connected with professional contacts. Shared interest communities are all about learning. Asking questions is a great way to show your commitment to learning about a skill as well as a great conversation starter. Social media is a great way to discover interesting people, but the real magic happens when you meet someone in person.

4. Measurements in Social Network Analysis

A primary use of graph theory in social network analysis is to identify the importance of actors. Centrality and prestige concepts seek to quantify graph theoretic ideas about an individual actor's prominence within a network by summarizing structural relations among the nodes [13]. They are closely related to hyperlink analysis and search on the Web.

An actor's prominence reflects its greater visibility to the other network actors. An actor's prominent location takes account of the direct sociometric choices made and choices received (outdegrees and indegrees), as well as the indirect ties with other actors. Both centrality and prestige are measures of degree of prominence of an actor in a social network [8].

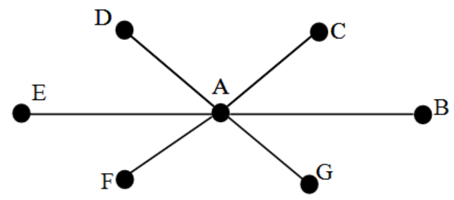

Figure 1. An example social network

4.1. Centrality

Measuring the network location is finding the centrality of a node. These measures give us insight into the various roles and groupings in a network who are the connectors, mavens, leaders, bridges, isolates, where are the clusters and who is in them, who is in the core of the network, and who is on the periphery. A person with extensive contacts (links) or communications with many other people in the organization is considered more important than a person with relatively fewer contacts. The links can also be called ties. A central actor is one involved in many ties. There are different types of links or involvements between actors. Thus, several types of centrality are defined on undirected and directed graphs.

4.1.1. Degree Centrality

Social network researchers measure network activity for a node by using the concept of degrees which is the number of direct connections a node has. In an undirected graph, the degree centrality of an actor is simply the node degree which is the number of edges of the actor node. In a directed graph, the degree centrality is defined based on only the out-degree which is the number of out-links or edges. In Figure 1, Actor A has degree six, all other actors have degree one. This logic underlies measures of centrality and power based on actor degree. Actors who have more ties have greater opportunities because they have choices. This autonomy makes them less dependent on any specific other actor, and hence more powerful.

4.1.2. Closeness Centrality

In closeness centrality, the centrality is based on the closeness or distance. The basic idea is that an actor is central if it can easily interact with all other actors. In Figure 1, actor A is more powerful than the other actors in the network is that actor A is closer to more actors than any other actor. Actors who are able to reach other actors at shorter path lengths, or who are more reachable by other actors at shorter path lengths have favored positions.

4.1.3. Betweenness Centrality

Actor A is advantaged in the network is because actor A lies between each other pairs of actors, and no other actors lie between A and other actors. If A wants to contact F, A may simply do so. If F wants to contact B, they must do so by way of A. This gives actor A the capacity to broker contacts among other actors to extract service charges and to isolate actors or prevent contacts. This aspect of a structurally advantaged position then is in being between other actors.

4.1.4. Eigenvector Centrality

Eigenvector centrality is a measure of the importance of a node in a network [14]. It assigns relative scores to all nodes in the network based on the principle that connections to nodes having a high score contribute more to the score of the node in question.

4.1.5. Centralization

Centralization is the difference between the numbers of links for each node divided by maximum possible sum of differences. A centralized network will have many of its links dispersed around one or a few nodes, while a decentralized network is one in which there is little variation between the numbers of links each node possesses.

4.2. Prestige

Prestige measures of prominence apply only to directed graphs, taking into account the differences between sending and receiving relations. A prestigious actor is defined as one who is object of extensive ties as a recipient. Hence, the prestige cannot be computed unless the relation is directional or the graph is directed.

4.2.1. Degree Prestige

A prestigious actor enjoys high popularity, shown by receiving many ties from others. Hence, it measure actor-level degree prestige as in-degree. To standardize the degree prestige index, divide outdegree by network size (g-1). Dividing by (g–1) standardizes the prestige value to the range from 0 and 1. The maximum prestige value is 1 when every other actor links to.

4.2.2. Proximity Prestige

Proximity prestige is computed as the mean length of all the shortest paths connecting a given node to the nodes within its influence domain. This analog to closeness centrality considers the proximity of actor to other actors in its influence domain, the set of all network actors that can reach actor, directly and indirectly. Because this measure ignores actors unable to reach actor, proximity prestige can be calculated for unconnected graphs, but the values depend on the network's size.

4.2.3. Status or Rank Prestige

Rank prestige is, by far, the most commonly used prestige measure and there exist a number of well-known methods to compute one or another flavor of such a measure. It is mutually reinforcing and, hence, it requires a series of iterations over the whole network. It takes into account the prestige of nodes linking (direct or indirectly) to a given node that is why it requires iterative algorithms and, in some sense, it describes how well connected is a node to other well connected node. Of course, the prestige ranks of all actors in the network are determined simultaneously in this manner.

5. Social Network Analysis Software

Social network analysis software is used to identify, represent, analyze, visualize, or simulate nodes and edges from various types of input data (relational and non-relational), including mathematical models of social networks [16]. The output data can be saved in external files. Various input and output file formats exist. Visualization is often used as an additional or standalone data analysis method. With respect to visualization, network analysis tools are used to change the layout, colors, size and other properties of the network representation. Social network analysis software generally consists of either packages based on graphical user interfaces (GUIs), or packages built for scripting or programming languages. GUI packages are easier to learn, while scripting tools are more powerful and extensible. Widely used and well-documented GUI packages include NetMiner, UCINet, Pajek (freeware), GUESS, ORA, and Cytoscape. Among them, I used Gephi in this analysis to measure social network because it can support various types of input and visualize network in real time.

5.1. Gephi

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs [4]. It is a tool for people that have to explore and understand graphs. The user interacts with the representation, manipulate the structures, shapes and colors to reveal hidden properties. It uses a 3D render engine to display large networks in real-time and to speed up the exploration. A flexible and multi-task architecture brings new possibilities to work with complex data sets and produce valuable visual results.

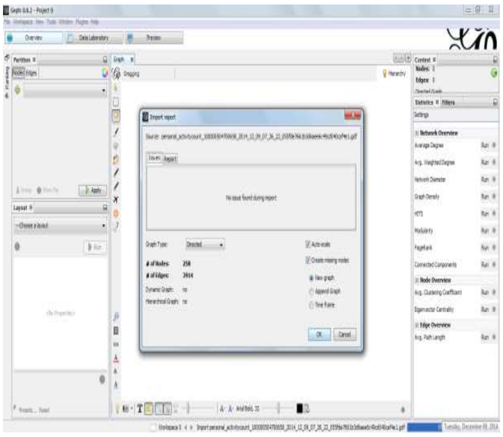

Figure 2. Importing web log to Gephi

5.2. Analyzing Social Network using Gephi

The objective of this paper is to analyze my Facebook friend network using Gephi. The analyzed results can reveal information about the groups of highly clustered people, the most influential person in terms of connections and the connecting persons between different clusters of people in my friend network. The graph shown in Figure 3 is a representation of my friend network taken from Facebook, and rendered using Gephi. There are 258 nodes and 3914 edges in this graph. In order to get the web logs from Facebook, I used an application called netvizz, which can extract data from different parts of the social network.

Modularity is a method that allows us to fast unfolding of communities in large networks. It is community detection algorithm. In order to make the sub groups more distinguishable, statistics are available in Gephi.

In the modularity statistical report it is stated that my network has 19 communities, but 13 of them have only 1 member in it. So in practical terms, there are only 6 communities, ranging from 5 to 90 members.

The average graph-distance between all pairs of nodes is called distance. Connected nodes have graph-distance 1.The diameter is the longest distance between any two nodes in the network. Betweenness centrality measure how often a node appears on shortest paths between nodes in the network.

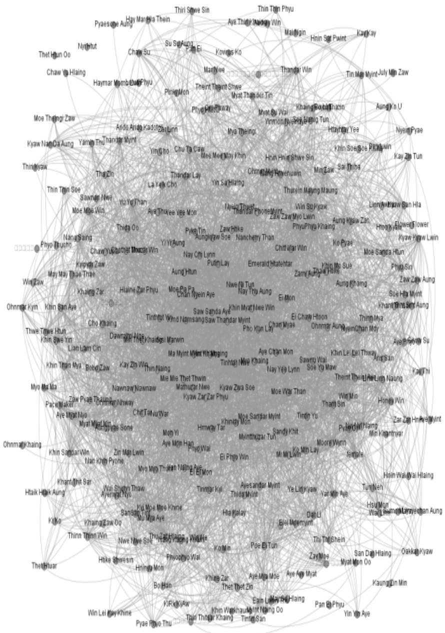

Figure 3. A graph of friend social network

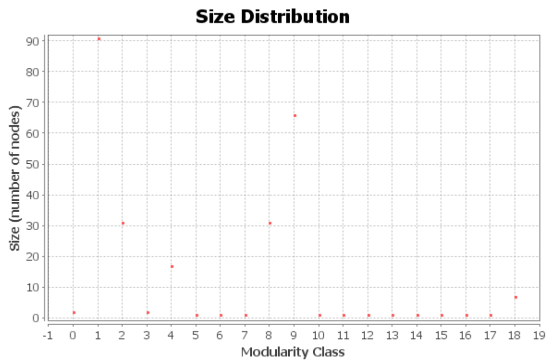

Figure 4. Modularity report

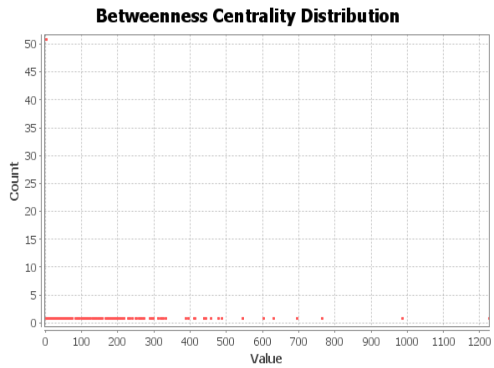

Figure 5. Betweenness centrality report

The graph ranked by betweenness centrality is shown in the Figure 5. To get that measure, I ran statistics on graph distance, from what I learnt that the diameter of my network is 7, and the average path length is almost 3.The number of shortest path in my graph is 21697.

Another single-node question that I wanted to ask to my data is about connectivity. Who among my friends know more of my friends? In order to do that, I ran the Average Degree statistics, and rank the nodes' sizes by its degree, so the most connected nodes appear bigger. As an overall network statistic, the average degree is 15.171, but with a high dispersion, ranging from 0 to more than 110, as shown in the Figure 6.

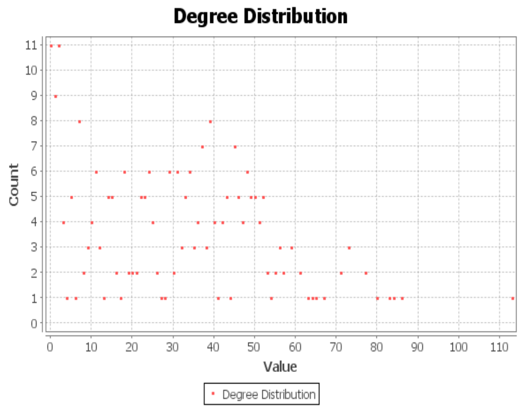

Figure 6. Degree distribution report

Gephi allowed to easily navigating through the layouts, partitioning, ranking and filtering of my network. It was very nice to be able to hover or click on a node, and see its connections. That served as a hard filter, and it helped me to perform separated analysis for the different groups. Also, it is able to manage large datasets. It can play with a dataset several times larger than my Friends.

6. Conclusion and Future Work

Analyzing social network will help to determine the browsing interest of the website users and the relationships between users. This paper focused on the use of social network analysis tool to analyze the relation of a user's friend social networks and interactions. Social network analysis tool called Gephi was used to evaluate the various measures of centrality such as degree, closeness, and betweenness and to identify patterns of interaction among actors between networks. Visual representations of social networks are important to understand network data and convey the analyzed result. As Gephi is an open source platform, in the future I have to extend Gephi by adding some advanced function to detect the source of distributing hate speech that can lead to be some unwanted problems via social networks.

References

[1] A. Goyal, F. Bonchi, and L.V.S. Lakshmanan, *Learning Influence Probabilities in Social Networks*, Proceedings of the Third ACM International Conference on Web Search and Data Mining, , 2010, pp. 241 -250.

[2] Borko, F, *Handbook of Social Network Technologies and Applications*, Springer publication, 2010.

[3] C.K. Dagli and W.M. Campbell, *Individual and Group Dynamics in the Reality Mining Corpus*, IEEE/ASE International Conference on Social Computing (SocialCom), 2012.

[4] Clément Levallois, Gephi – *advanced functions Filters, metrics, plugins and workspaces*, 2014.

[5] Enhua Tan and Lei Guo, *SPAM Behavior Analysis and Detection in User Generated Content on Social Network*, IEEE, 2009, pp298-305.

[6] Erlin and Norazah Yusof, *Analyzing Online Asynchronous discussion using Content and Social Network Analysis*, IEEE , 2008, pp-872-877.

[7] Ghosh, R., and Lerhman, K., *Predicting Influential Users in Online Social Networks*, Proceedings of KDD Workshop on Social Network Analysis (SNA-KDD), Washington DC, USA, 2010.

[8] Grabisch, M., Rusinowska, A., *A Model of Influence in a Social Network. Theory and Decision*, 2010, pp-69-96.

[9] J. Chen, J. Fagnan, R. Goebel, R. Rabbany, F. Sangi, M. Takaffoli, E. Verbeek, and O. Zaiane. *Meerkat: Community mining with dynamic social networks*, Data Mining Workshops (ICDMW), IEEE International Conference, 2010, pp- pages 1377 –1380.

[10] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, *The anatomy of the Facebook social graph. Structure*, 2011.

[11] K.R Sunnetha and R.krishnamoorthi, *Identifying User Behavior by Analyzing Web Server Access Log File*, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[12] Marcelo maia and Jussara Almeida, *Identifying User Behavior In Online Social Networks*, Socialnet'08, ACM, pp-199-207.

[13] Mislove and Alan, *Measurement and analysis of online social networks*, Proceedings of the 7th ACM IGCOMM conference on Internet measurement", 2007.

[14] Newman, M. E. J., *Finding community structure in networks using the eigenvectors of matrices*, Physical Review E, 74, 2006.

[15] Rupam Some, *A Survey on Social Network Analysis and its Future Trends*, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6, June 2013.

[16] Takaffoli, M., Sangi, F., Fagnan, J., and Za¨ıane, O., *A Framework for Analyzing Dynamic Social Networks*, Applications of Social network Analysis (ASNA), Zurich, Switzerland, September 2010.

[17] Zafra, A., Romero, C., Ventura, S. DRAL, *A Tool for Discovering Relevant e-Activities for Learners,* Knowledge and Information System, 2013.