The Role of Gephi and Key Algorithms in Network Visualization and Analysis

Abstract-- This paper will examine the role of tools such as Gephi in network visualization and will also take a closer look at several algorithms of Gephi that provide some important graph statistics. Several algorithms that we will examine in greater detail in this paper include PageRank, Betweenness Centrality, and Word Adjacencies.

***Keywords*—Algorithm, Betweeness Centrality, Gephi, Network Visualization, Neural Network, PageRank, Social Network, Word Adjacencies**

I. Introduction

Network visualization has become an increasingly important approach for how we view data in our increasingly connected world. Social networks, information networks, transportation networks, and a host nf other datasets can be brought to life through network maps. In recent years, the explosion of social media datasets has propelled network graphs into the visualization mainstream, resulting in a number of proprietary and open source tools that address the need to create and view networks. One of the leading tools of this genre is Gephi. The goal of Gephi is to make network visualizations accessible to all by providing a set of tools that handle the complex mathematics supporting the graphs. Gephi has a number of algorithms that help generate graph statistics. Some examples of these algorithms are PageRank, Modularity, Betweenness Centrality Distribution and more. Betweenness Centrality algorithm is a way of detecting the amount of influence a node has over the flow of information in a graph. It is often used to find nodes that serve as a bridge from one part of a graph to another. This algorithm, in particular, calculates the shortest path between every pair of nodes in a graph that is connected. Nodes that lie on the shortest paths have the higher betweenness centrality score. One key function of the betweenness centrality algorithm is helping microbloggers spread their reach on Twitter with future recommendations that target the Twitter users and enable them to interact with something new.

II. Related Works

The datasets used for our findings came from the GML repository at http://www-personal.umich.edu/~mejn/netdata/. The three datasets we used to create our figures and tables and to provide our statistics were the dolphin social network dataset, word adjacencies dataset, and the neural network dataset.

III. Background/Method

In this section, we take a closer look at the datasets and the algorithms we are using for our network visualization and analysis. The datasets we used are the dolphin social network dataset, the neural network dataset, and the word adjacencies dataset. The dolphin social network dataset describes an undirected social network of frequent associations between 62 bottlenose dolphins in a community living off Doubtful Sound, New Zealand. This dataset contains a list of all of links, where a link represents frequent associations between dolphins. A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. The neural network dataset is a directed, weighted network, representing the neural network of C. Elegans. The third dataset that we will be using in our findings is the word adjacencies network. This dataset describes an adjacency network of common adjectives and nouns in the novel *David Copperfield* by Charles Dickens.

The algorithms that we used as part of our network visualization and analysis include the PageRank algorithm and the Betweenness Centrality algorithm. The PageRank algorithm measures the transitive influence or connectivity of nodes. It can be computed by either iteratively distributing one node's rank (originally based on degree) over its neighbors or by randomly traversing the graph and counting the frequency of hitting each node during these walks. PageRank is named after Google co-founder Larry Page, and is used to rank websites in Google's search results. This algorithm counts the number and quality of links to a page, which gives an estimation of how important the page is. The underlying assumption is that pages of importance are more likely to receive a higher volume of links from other pages.

Betweenness centrality is a way of detecting the amount of influence a node has over the flow of information in a graph. It is often used to find nodes that serve as a bridge from one part of a graph to another. The Betweenness Centrality algorithm calculates the shortest (weighted) path between every pair of nodes in a connected graph, using the breadth-first search algorithm. Each node receives a score, based on the number of these shortest paths that pass through the node. Nodes that most frequently lie on these shortest paths will have a higher betweenness centrality score. The algorithm was given its first formal definition by Linton Freeman, in his 1971 paper "A Set of Measures of Centrality Based on Betweenness". The method Gephi uses for calculating Betweenness Centrality is detailed in "A faster algorithm for betweenness centrality".

IV. Results/Findings

A. Tables

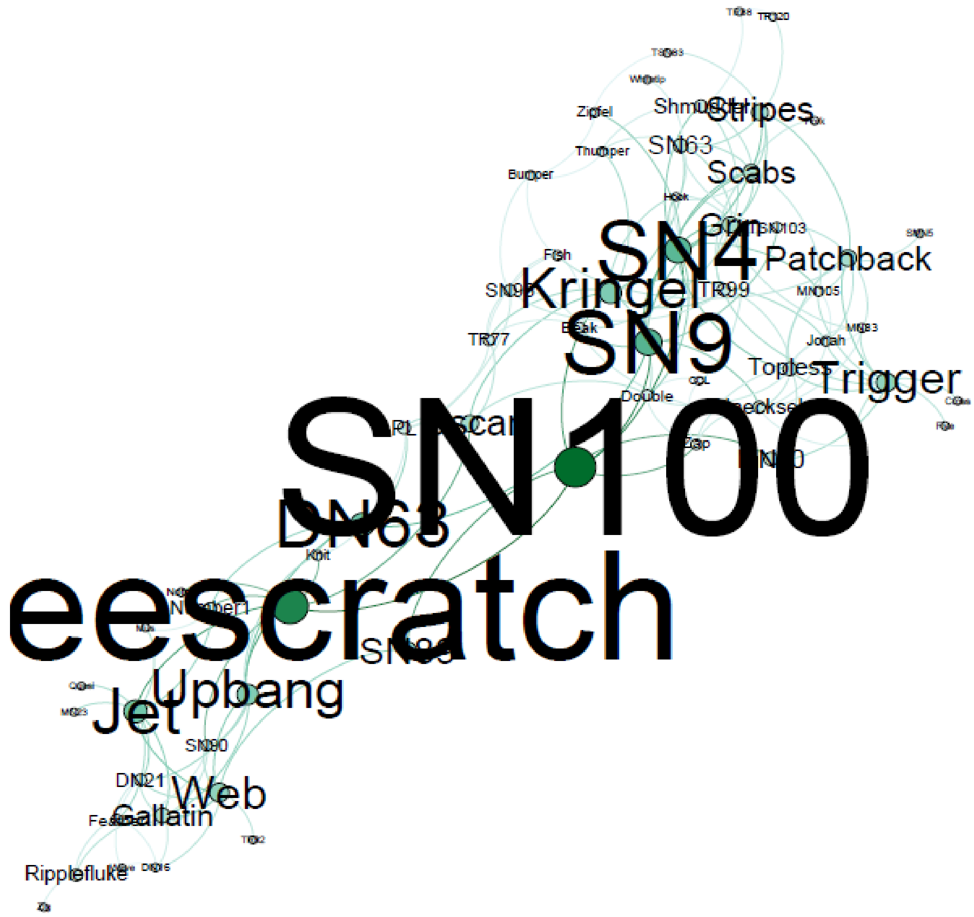
	Dolphin social network	Word Adjacencies	Neural Network
Average degree	5.129	3.795	7.896
Average weighted degree	5.129	3.795	29.694
Network diameter.	8	7	14
Graph density	.084	.034	.027

HITS		*	*
modularity	.517	.28	.484--- 0,1, 2, 3,4, or 5 with % from ~29 to 3%. This had fewer divisions than others partitions
PageRank		*	*
Connected Components	1 week	1 week/112 strong	1 week/57 strong

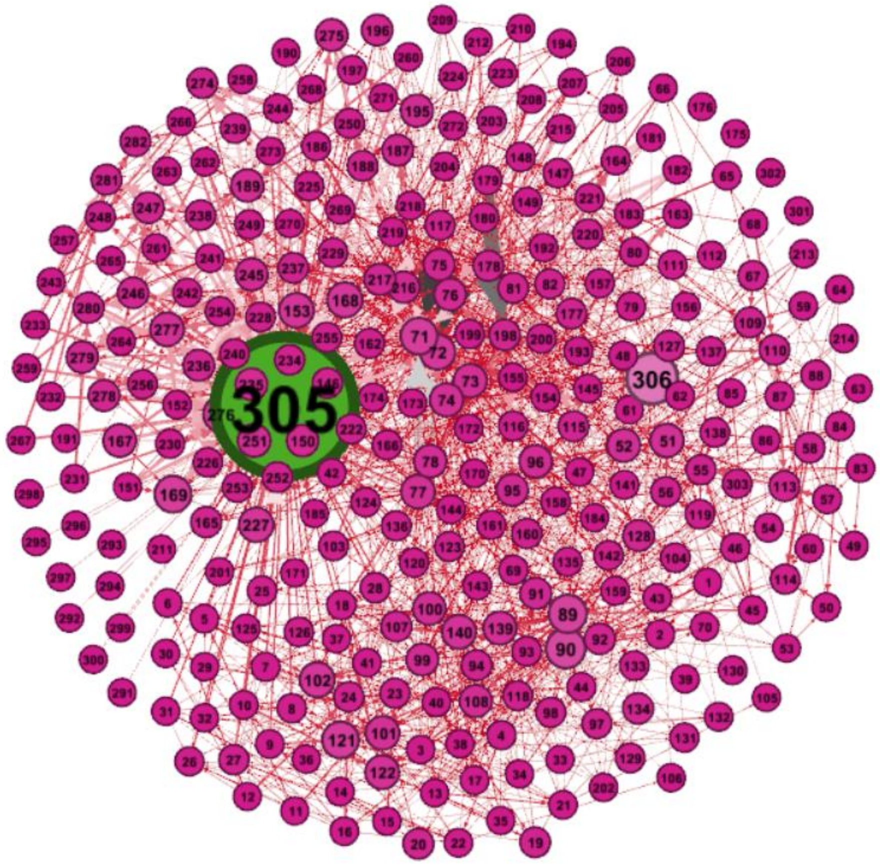
Table 1:

B. Figures

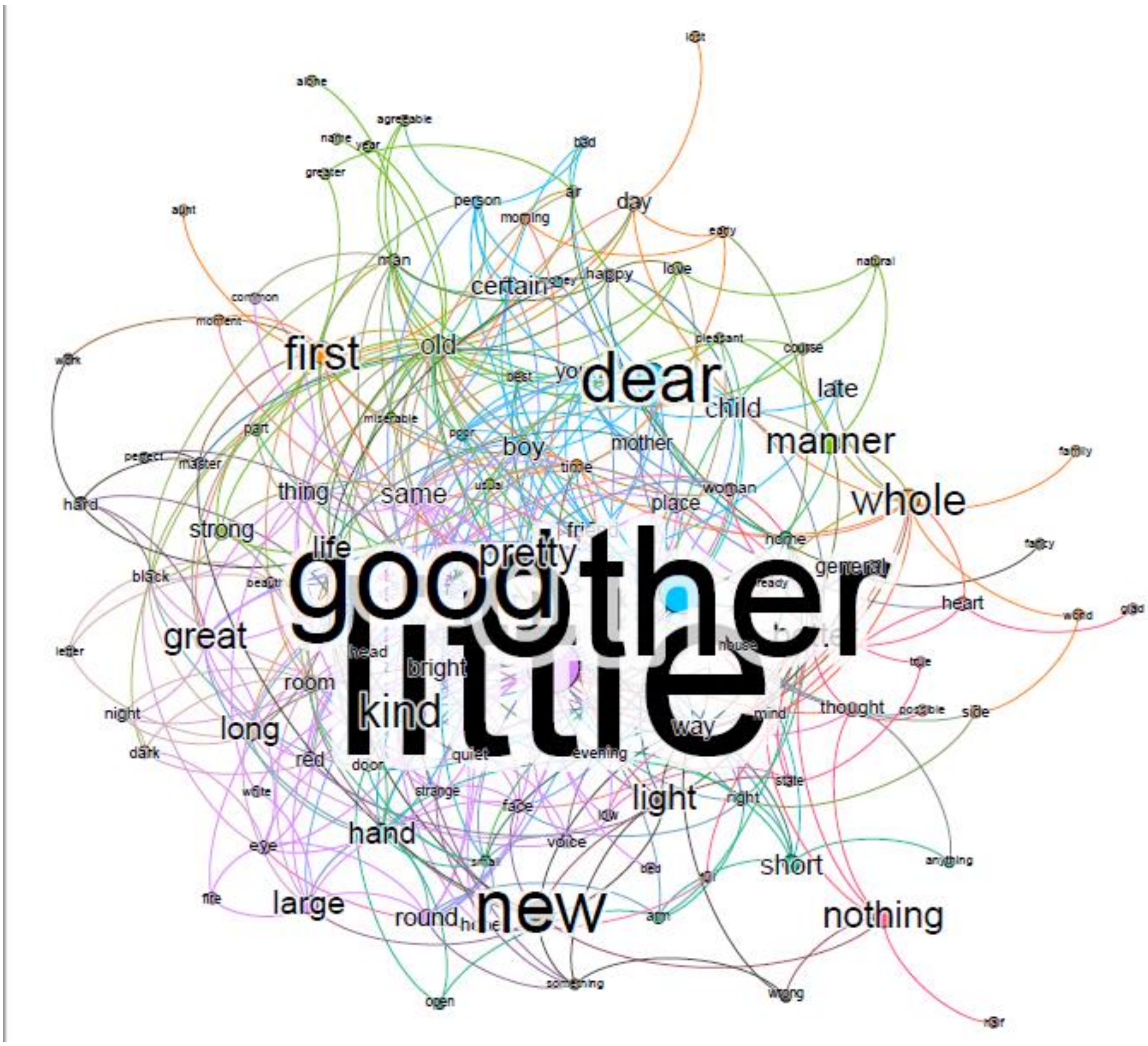
The first visualization we provided is the Dolphin data set which shows the relative communication between a group of 62 dolphins. For the dolphin data we used the Betweenness Centrality Algorithm. The Betweenness Centrality Algorithm works wonders in this kind of situation due to the way that natural social networks often have specific individuals who are bridges within select clicks, thus using an algorithm based around finding specific bridges would be perfect. Some other algorithms were attempted here as well to much more middling results, of note would be modularity which could make sense to show the possibility of specific social groupings but failed due an extreme lack in variance between node.



The second dataset used was the Neural Network Dataset. With this dataset the chosen algorithm was Page Rank. The idea here was since it is supposed to depict a neural network that using Page Rank it can hopefully dissect the most important and vital parts of the brain which it did quite spectacularly. In this situation Page Rank was the most prominent choice but some others were contending and interesting including Modularity, which pieced off specific parts of the brain, and Betweenness, which portioned off more centralized nodes, but even then the more centralization of Page Rank made the most sense.



The final dataset was the comparison between the adjectives and nouns in the book, David Copperfield. For this dataset once again the Betweenness Centrality Algorithm was used. This like the previous two makes the most sense, with how we are looking at the relationship between adjectives and the nouns in which they describe in a way that you are able to find the adjectives that most normally bridge to many different nouns within the book. There were others that were considered including one hypothesized as perfect being Page Rank, but this was not useful as there was a group of three major outliers that were used repeatedly with each other within the book.



V. CONCLUSION/DISCUSSION

For these three datasets were examined that gave quite a bit of information. Starting with the Dolphins we can see that Dolphins SN100 as well as eescratch appear to bridge the gap between what appears to be 2 distinct social circles within the small dolphin community, this raises questions about the social dynamics within dolphins and how they are able to create small sub groups within their large society. From this dolphin data you could push into examining not just the communication but the possibility for dolphins to develop hierarchy within their society as well as how they might react when presented with an outsider within their community.

For the Neural dataset there is not much that is truly unknown to humanity, but it is still very interesting. It is obvious that within the ranking sits a single very important node which was 305 (depicted by not only its size but also, its color being completely different). As there is no depiction for the exact representation of each node the guess from this end is that this node is the Medulla due to the medulla being the true center of most autonomic function within the human body and it being a most central part of the natural control over the entire brain and body. From this instead of focusing on the medulla which immediately draws eyes, it would be much more interesting to delve further into the rest of the brain and to see how specific neural networks are grouped together as well as how they are intertwined.

The final Dataset is the comparison between adjectives and the corresponding nouns within the book David Copperfield. This one is especially interesting as using betweenness you can see how much he may rely on certain adjectives, as you can see "little" is used distinctly more than any other with "good" and "other" following closely after. This kind of study would be most useful in the way you could then compare it to other similar books and compare the relative use of adjectives as well as then compare their favorability among readers to see how specific adjective noun combinations are viewed by readers.

VI. References

Dolphin social network: Please cite D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, *Behavioral Ecology and Sociobiology* **54**, 396-405 (2003).

Word adjacencies: adjacency network of common adjectives and nouns in the novel *David Copperfield* by Charles Dickens. Please cite M. E. J. Newman, *Phys. Rev. E* **74**, 036104 (2006).

Neural network: A directed, weighted network representing the neural network of C. Elegans. There are 297 nodes and 2345 edges. Data compiled by D. Watts and S. Strogatz and made available on the web here. Please cite D. J. Watts and S. H. Strogatz, *Nature* **393**, 440-442 (1998). Original experimental data taken from J. G. White, E. Southgate, J. N. Thompson, and S. Brenner, *Phil. Trans. R. Soc. London* **314**, 1-340 (1986)