North Carolina School of Science and Mathematics

Modern Networks

Computational Analysis of the Wikipedia Network

Authors:

Jonathon K., Akshay P.,

Suhas R., Murali S.,

Sunwoo Y., Jennifer Z.

Instructor:

Dr. Daniel J. Teague

May 25, 2016

Contents

1 Summary

In order to design a comprehensive network, Wikipedia was used to create the overall structure. The nodes of the network represented pages, and the links between each node was found using the hyperlinks in each Wikipedia page. We created this network computationally by writing a Python program that started at a specific page and recursively created the network up to a certain depth.

We then ran this network through an analysis program, which used Gephi and Python to find different properties of the network such as the degree distribution, formation of clusters, degree centrality, Closeness centrality, and betweenness centrality. First it was found that the degree distribution graph closely followed the form of a power function, which was to be expected due to the nature of how the Wikipedia network was created. In addition, after generating a visualization of the network, it was found that the network formed many clusters of densely packed nodes. This behavior was explained by the method with which the network was created, as the program procedurally created the network using a specified depth. For degree centrality, the majority of the nodes were degree 1, which was expected as the nodes that were created at the end of the depth made up the vast majority of nodes and were all of degree 1. Interestingly, for closeness centrality, the nodes could be categorized into two groups, one with closeness centralities between 0.25 and 0.275 and the other between 0.325 and 0.35. This phenomenon demonstrated that Wikipedia articles tended to either cover a broad range of topics or refer to only the specific topics immediately around them. Finally, for the betweenness centrality, we found that the values were extremely low because the network analyzed had been created with only a depth of 2.

With more computational power and time, a network created with a greater depth could be analyzed. This would provide more insight into the nature of Wikipedia networks.

2 Introduction

Wikipedia is a free online encyclopedia and is a massive resource of information. Wikipedia features articles covering nearly the full spectrum of common human information from advanced physics concepts to famous Israeli celebrities to whatever else the mind desires.

Because of its unique dedication to covering all content, even knowledge outside the traditional academic areas of knowledge, Wikipedia sports an open policy towards content creation that is semi-regulated. This allows people all across the world who know niche topics to write and possibly publish their articles on the online resource and thus allows the knowledge contained within Wikipedia to be larger than any source had had prior. In fact, the English Wikipedia contains more than 5 million articles and Wikipedia as a whole, 38 million articles [1].

Part of the uniqueness of Wikipedia is that these articles are intraconnected, just like knowledge. Within a given article, another Wikipedia article can be hyperlinked to it allowing for the article author to give easy reference to concepts displayed within their article. This structure means that Wikipedia is a mathematical network where nodes are articles and edges are these hyperlinks. The characteristics of this network such as the average number of hyperlinks a page has (average degree), how interconnected the network is (cluster coefficient) and what are the most important articles of Wikipedia (centrality) were examined using quantitative methods and then the properties of unique characteristics of clusters were investigated within the network using qualitative analysis.

In this paper, instead of understanding Wikipedia as a whole, a knowledge domain was investigated: how much network science knowledge is contained within Wikipedia. This was done by only considering only “depth 2” or the edges originating from the nodes directly connected to the Network Science page, a grand total of 7,361 articles [2]. While ideally a depth 3 or depth 4 network would be investigated, time constraints kept our program from being able to run up to depth 3 efficiently (there would be over 100,000 articles) and there was no method of filtration. However .15% of the English Wikipedia was able to be mapped,

which is certainly no small task. Although relative to the size of Wikipedia, the data pool was small, the isolated data set helped to analyze interesting features of the Network Science Wikipedia network without having the rest of our professional lives consumed by researching this topic.

3 Methods

3.1 Computational Approach

For the analysis of the network structure of page link connections in the Wikipedia encyclopedia compilation, a Python program was created to collect and analyze the data. The program considered the input of an initial source page (for our analyses, we considered the source nodes of the "Networks Science" page) and the maximum number of iterations (corresponding to the maximum distance a neighbor can be from the source node, defined for convenience purposes as the depth of the network). From here, it systematically scanned through each Wikipedia page, collecting from each the page names and URLs of forwarding pointing hyperlinks that remain within the Wikipedia network of pages. These neighbors of the previously identified page (of a lower page depth) were then assigned a node number and stored in a .txt file with the corresponding node numbers of its direct neighbors for later analysis and visual output. This process was repeated for every stored page of depths less than or equal to the specified number of maximum iterations for the convenience of the user (to save both computational time and memory as both resources increased exponentially with the number of iterations run). For example, consideration of the network of pages of depth 2 originating from "Networks Science" would yield an output of the node numbers, page names, page URL, and direct neighbors of the source node, its direct neighbors, and its second neighbors, iteratively.

This output was then run through a Python program that then created a version of an ad-

jacency matrix, assigning 0s to each node for every node accessed in the analysis to which it was not connected.

This output was then inputted into the Gephi visualization platform to create a graphical representation of the network. Additionally, various measures of centrality (degree, betweenness, closeness) and clustering coefficients were calculated using NetworkX and Python to further analyze the behavior of the networks and their significant nodes.

4 Results

Unfortunately, due to limits in computational power and time, the maximum depth of network that could be feasibly analyzed was 2.

4.1 Visualization of the Network through Gephi Software

Below is the network that was created from the computational methods mentioned previously. The network began at depth of 2 starting from the "Networks Science" page. There are 7,361 nodes, each one representing a different page.

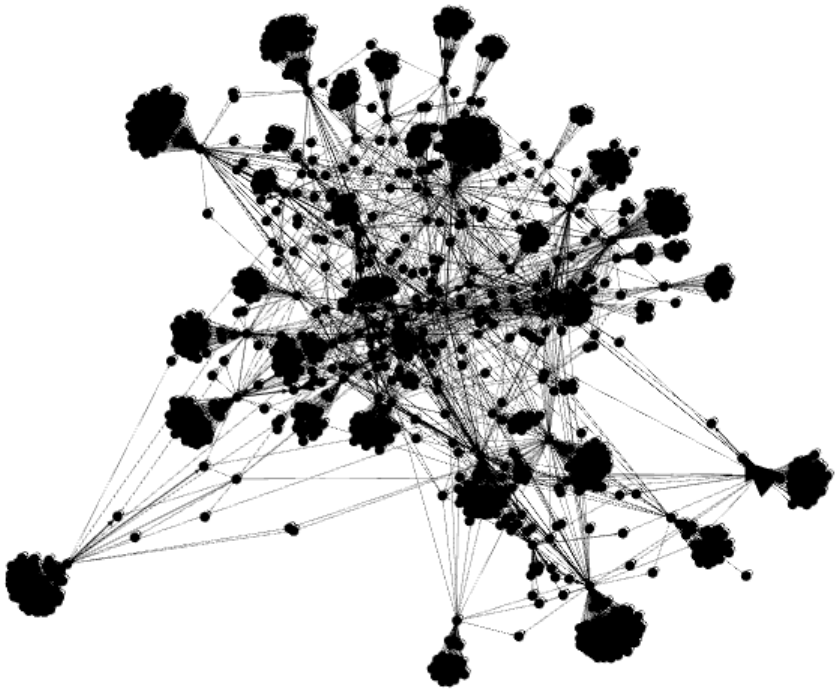

Figure 1: Partial Network of Wikipedia originating from the "Networks Science" Page

The average degree of the network was 5, however this is a misleading interpretation of the network created. There are many nodes that had a lower degree simply because they were at the end of the depth 2. Thus network exhibits a degree distribution that is similar to a Power Law Distribution starting from degree 5.

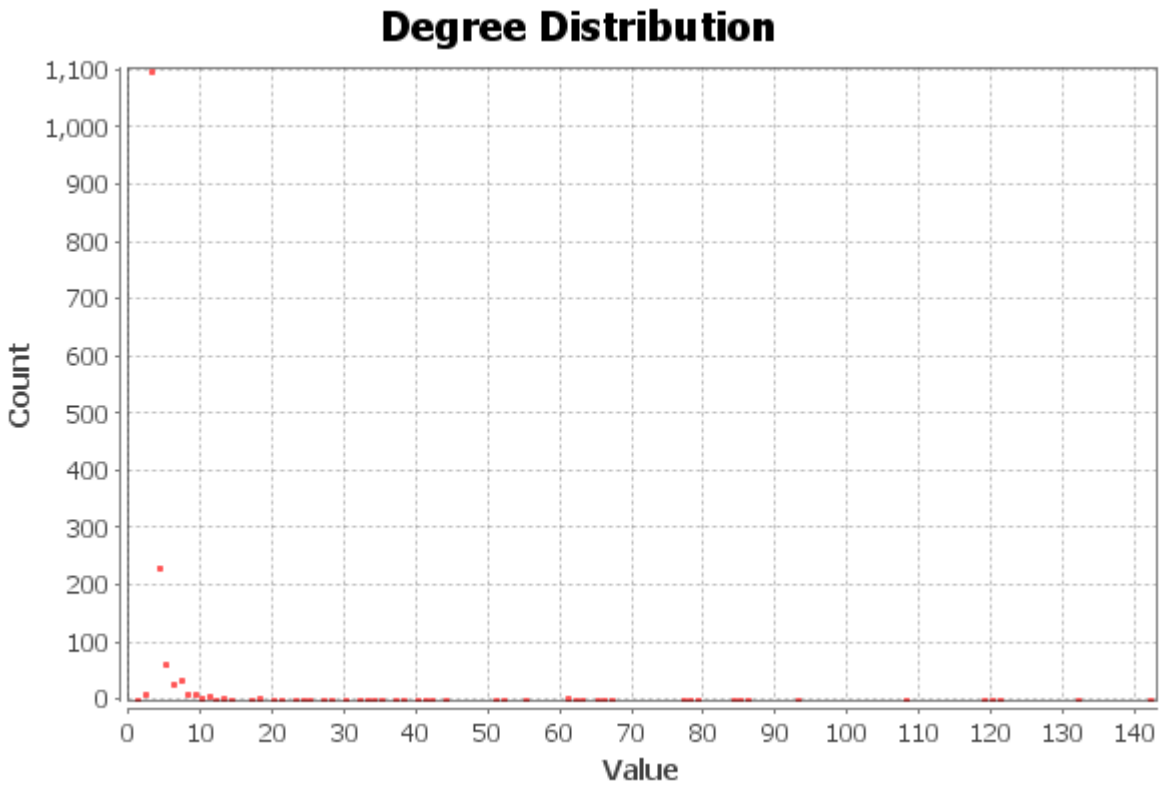

Figure 2: Degree Distribution of Partial Network above

Gephi provides an interesting approach of visually distinguishing nodes of the network by viewing their "close friends" or neighbors with which they share many other common neighbors forming a neighborhood or cluster of nodes. The Force Atlas graphical analysis technique link assumes nodes naturally repel each other but edges attract nodes. Thus a series of heavily interconnected nodes will form a cluster and the overall network is an aggregate of the individual neighborhoods. Then, the clusters could be qualitatively assessed by looking at which particular articles form these clusters. One of these clusters was considered [3]. The cluster, depicted in Figure 3, consists of 86 nodes and 90 edges. This cluster was dubbed the Internet Cluster due to it containing articles like Online Game, Video Hosting site and Ratemyprofessors.com (Dr. Teague would get a 5/5). While we postulate the existence of clusters in the full wikipedia network, unfortunately the clusters observed are simply outcomes of a node's hyperlinks outwards to their end nodes. These end nodes are only

attached to a source node which is in turn is attached to the Network Science node because the program only ran to "depth 2". Further exploration of other clusters unfortunately proves the theory correct, that the clusters are results of degree-1 nodes attached to a source node. Of course, it is an important observation to make that these degree-1 nodes are not truly degree-1 but rather any mathematical analysis treats them as such because the program was not able to process the hyperlinks of these end nodes. This, however, would be a problem, albeit a diminishing one, with any network considered except the full Wikipedia network of all 5,000,000 articles because for any given depth, there would be end nodes. Although there is no good method of analyzing clusters without a full network, clusters are still an interesting point of contention that should be investigated further but ultimately within the time constraint, it was not possible to obtain any conclusive results.

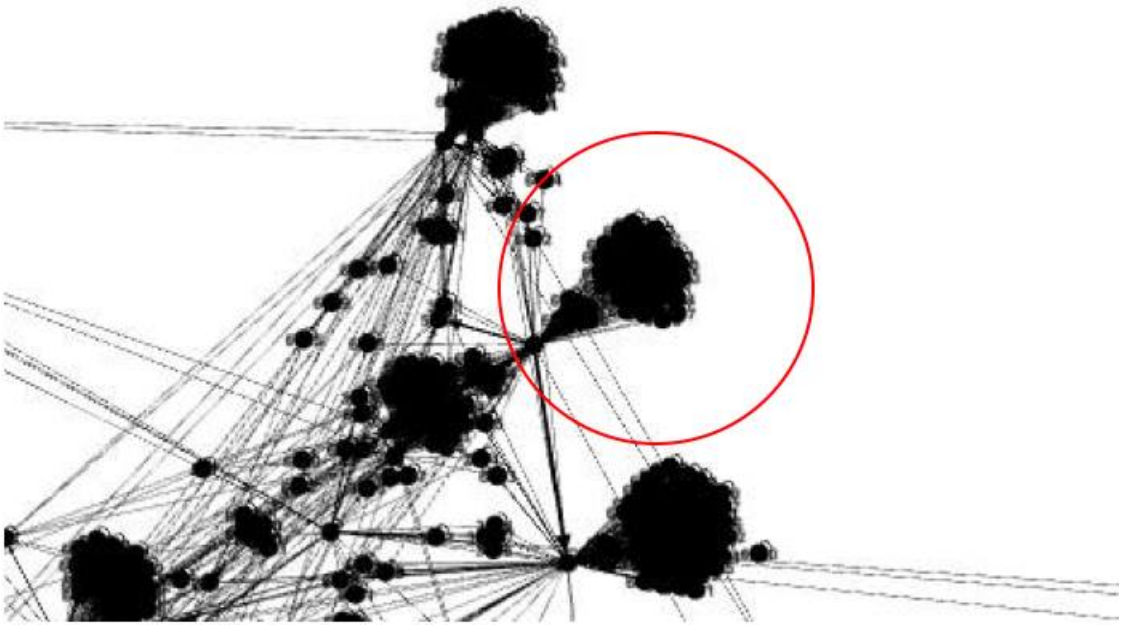

Figure 3: One of the clusters in the Partial Network

4.2 Analysis of the Network through NetworkX and Python

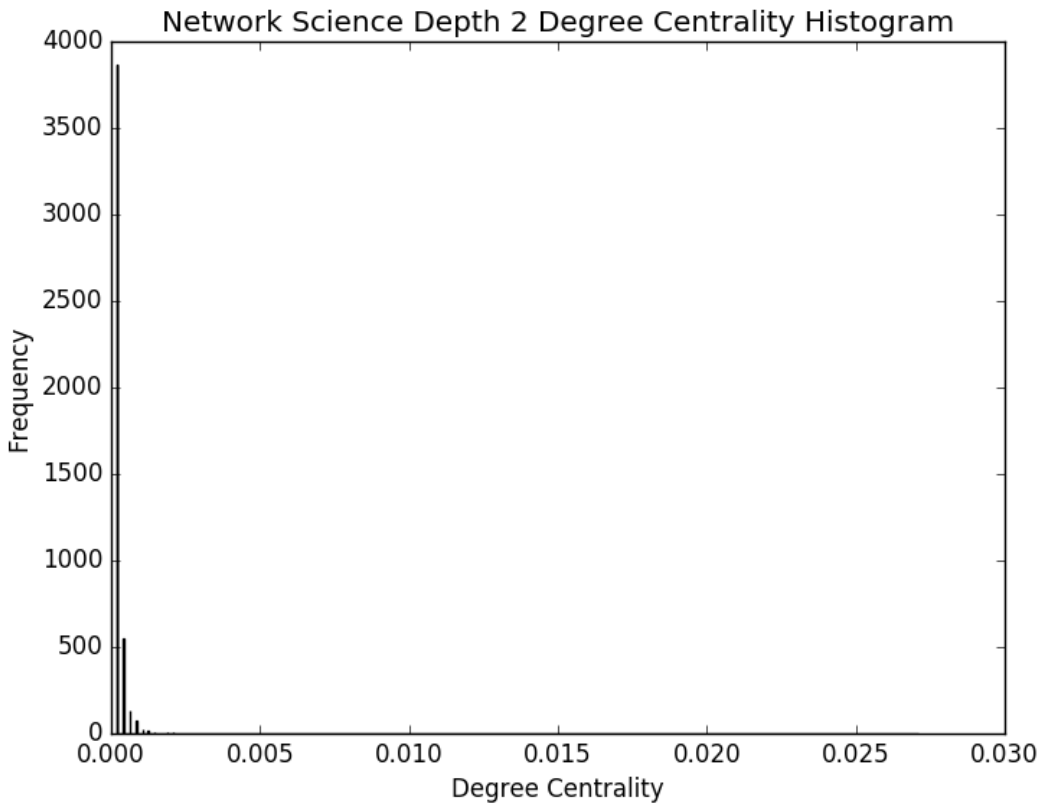

Figure 4: Histogram of the Degree Centralities of the Nodes

The degree centrality distribution shares the same shape as the regular degree distribution since it merely imposes a scalar (the total number of nodes other than the selected node that it could be connected to) on the degree of each node. While the distribution follows the expected power law distribution, this is heavily skewed by the fact that the analysis only follows each hyperlink through two levels, meaning that any page reached at the end of these two levels will have a degree that is only indicative of pages that have already been analyzed regardless of how many hyperlinks it may actually have. Since an exponentially greater number of pages will be reached by each level the analysis progresses, the majority of vertices will thus have degree 1 as long as the network is not fully complete.

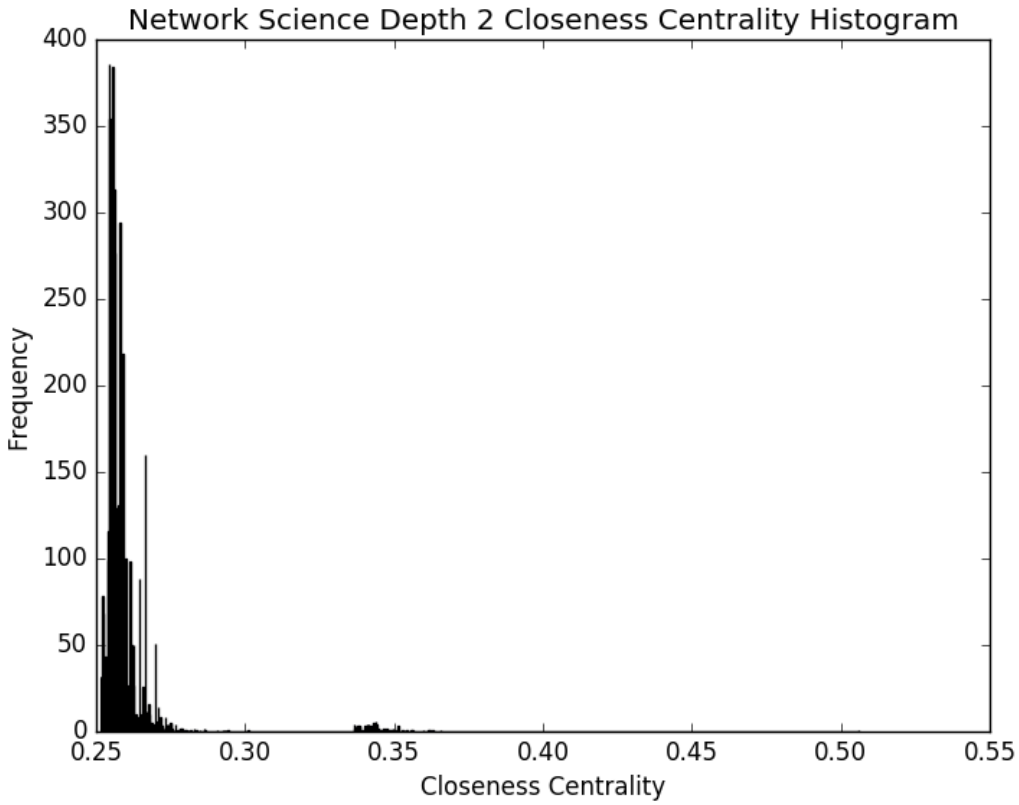

Figure 5: Histogram of the Closeness Centralities of the Nodes

The closeness centrality distribution demonstrates the presence of two separate types of nodes. One group of nodes has closeness centralities between .25 and .275, while the other has closeness centralities between .325 and .35. This is likely indicative of the fact that pages in Wikipedia typically are either overarching subjects that include a great number of hyperlinks for their underlying details or themselves one of these underlying details. This is exacerbated by the fact that these subject pages often have many similar topics redirected to them, preserving the separation between the two types of nodes.

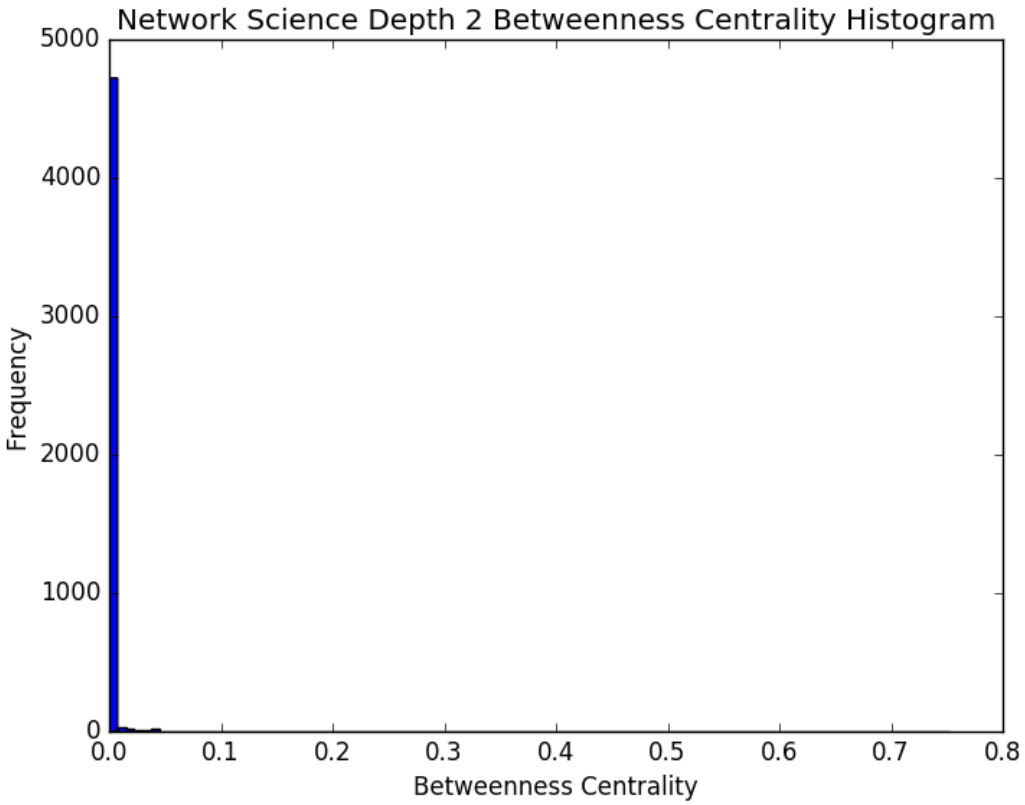

Figure 6: Histogram of the Betweenness Centralities of the Nodes

The betweenness centrality distribution demonstrates one of the detriments of only being able to analyze a Wikipedia network of depth 2. Since betweenness centrality relies entirely on the lengths of paths between nodes to distinguish between the centralities of various nodes, the fact that all possible paths between nodes are not reflected causes the betweenness centralities of all of the nodes to be inaccurate since the network appears much less connected than it is in reality. The shortest paths referenced in our analysis are in no way indicative of the actual shortest paths through nodes in Wikipedia.

Table 1: Table of Measurements Calculated Using NetworkX and Python

	Degree C.	**Closeness C.**	**Betweenness C.**
Max	Sociology: 0.0743	Network Science: 0.504	Network Science: 0.582
Min	Many Nodes: 0.000131	Many Nodes: 0.251	Many Nodes: 0
Average	0.000350	0.263	0.000369

Clustering Coefficient	0.04093793

5 Strengths

We tackled a barely considered problem that lacks extensive study. In fact, during background research, there was a startling lack of prior knowledge. This project is one of the few that takes a computational networks approach to understanding the Wikipedia encyclopedia. Moreover, we legitimately built a method that we could use to extensively network Wikiedia. With more computational power and time, it is even possible for this software to network all of the Wikipedia Database, along with providing valuable Networks statistics such as Closeness Centrality, Clustering Coefficient, and Degree Distribution.

6 Weaknesses

Due to lack of time and computational power, this project was only able to network about 7,000 of the approximately 5 million pages on the English Wikipedia. This means we were only able to analyze 0.15% of the total number of Wikipedia pages. However, considering the time frame given for this project, 7,000 nodes is still a massive data set with which we were able to perform meaningful calculations.

Since the program stops running at depth 2, this impacts the accuracies of some of our centrality calculations. Any page that comes at the end of the depth 2 will automatically have a lower degree than it should, because there are other multiple hyperlinks (edges) on that page itself. Due to this, the degree distribution and all measures of the centralities are skewed.

7 Conclusion

Although only approximately 0.15% of the total number of Wikipedia pages were analyzed in this project, there were still a vast amount of information to create the network from. The project was mainly limited by a lack of computational power that led to the network being

created with only a depth of 2. However, the majority of the results of the analysis were predictable and could be explained due to the nature of how the network was created. We found that the degree distribution and degree centrality graphs closely followed the shape of a power law function. Because the majority of the nodes in the network had a low degree, this was heavily represented in the degree centrality. Also, because the network had a depth of only 2, the visualization of the network showed the formation of many clusters, and the analysis of the betweenness centrality yielded very low values as expected. We also found that in the closeness centrality, the nodes could be categorized into two groups, which demonstrated interesting properties about the nature of the network.

The main limitation during this project the lack of resources to explore the network beyond depth 2, which limited the analysis of the network. Hopefully, given more time and computational power, the network can be greatly expanded, leading to more interesting conclusions about the network and about Wikipedia itself.

References

[1] "Wikipedia." Wikipedia. Wikimedia Foundation. Web. 25 May 2016. *https://en.wikipedia.org/wiki/Wikipedia*

[2] Korfiatis, Nikolaos Th, Marios Poulos, and George Bokos. "Evaluating authoritative sources using social networks: an insight from Wikipedia." Online Information Review 30.3 (2006): 252-262.

[3] Jacomy, Mathieu, et al. "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software." PloS one 9.6 (2014): e98679.