Indiana University - Bloomington

Network Science

Measuring Facebook Influence Using Centrality Measures and Community Detection

Authors
Daniel Hinders
Nhi Tran

Professor
Yong-Yeol (YY) Ahn

May 4, 2020

Measuring Facebook Influence Using Centrality Measures and Community Detection

Daniel Hinders and Nhi Tran

May 4, 2020

Abstract

Within the past couple decades, social media has become a large part of everybody's life and our society. With the aid of social media, social influence can take place without physical human interaction and allow people to influence each other just by having social media accounts and the ability to access the Internet. After social media became popular, researchers and scientists have been studying the impact of social media using different methods such as surveys, trials, scientific experiments, and so on. Analyzing social media network graphs is another way of studying social media impact that this paper will cover. Given a public dataset of Page to Page network from Facebook in November 2017, multiple network graph measurements and a community detection method will help identify influential Fabebook pages within a network. By understanding the differences between measurements and what information they provide, it will be beneficial to apply similar measurements and analysis methods on other social media network to identify influential nodes.

1 Introduction

With individuals having the ability to influence others anytime from any location anonymously, it is important to be able to identify their power and influential level within a social network to carefully access the impact. Without the ability to identify influential nodes, a network can be too large to study and time-consuming for anyone to get useful information.

1.1 Background

Social media impact has drastically increased its effect on life in the last 10-15 years, the diverse and far reaching impacts cannot be overstated. Russian influence in the 2016 election, the Obama campaign utilizing social media to achieve a decisive victory in 2008 and social media fueling societal change and upheaval in the middle east and Hong Kong are some of the examples.

In another realm, corporations spend a massive 11.6 percent [1] average of their marketing through social media indicating vast impact on societies choices and buying habits through social media.

In the midst of these large organizations and vast movements in social media, there is a fasci-

nating and powerful impact propagated through lone individuals on social media. A single individual operating as a social media influencer is an existence that has become increasingly common in our societal vernacular. For many, a full time job as a social influencer has become a viable career.

Understanding all of these social media forces and impacts is important on many levels. Devastating negative ramifications: a nation state subverting democracy by impacting other countries' elections, hate groups spreading and inciting violence, and webs of misinformation leading to widespread distrust are all reasons alone to understand social media impact and influence.

1.2 Motivation and Objectives

The ultimate purpose of this paper is to analyze a public Facebook Page-to-Page network to identify important nodes and different communities.

The first objective is to answer the question that will observe the potential differences between multiple centrality measurements that help identify important nodes within a network - Can we simply use degree to identify powerful nodes in a network?

The second objective is to answer the question that will identify the different between a predetermined classification of a network and how likely would the nodes form a community within their own classification - Can we use page type to detect communities?

1.3 Existing Work

Social media is increasingly prevalent in commerce, politics, and all societal interaction. Unsurprising, research seeking to measure and understand social media influence and impact is prevalent.

There are several studies that are related to our project objectives. Several use similar centrality measures and some community detection techniques.

A Comparative Study of Modularity-Based Community Detection Methods for Online Social Networks [2] Karatas and Sahin also focus on communities in social networks using modularity. The study assesses datasets from multiple social media platforms using different static community detection algorithms and specifically focus on "modularity values, running time and accuracy" and compare and contrast these different algorithms for community detection assessing their strengths and weaknesses. The performance of the different algorithms were analyzed by measuring their modularity value, F1-score, and running time. Five different algorithms were used and the Louvain algorithm performed the best overall. The study suggests the following as problems with the current available community detection algorithms: stability, scalability, refinement on computational complexity, dynamicity, and prediction. These ideas prompt and encourage future research on community network structure and development of new community detection algorithms.

Social Circles in Facebook communities [3] The Social Circles in Facebook Communities study used the egonet dataset to assess specific subsets of the Facebook network which they termed social circles. Social circles are smaller groups within an individual's social network such as friends from highschool or the workplace. The study utilized the egonet data set to construct

various models of social circles and calculate Min Edit Distance and was purposed to: "predict what communities would form or what circles can be created using the given friendships" [3]. Initially Clique Percolation to model the social circle data, but later in the study, the researchers found that Spectral Clustering was more effective. "Spectral Clustering is a way to cluster based on 'Affinity' and connectivity" [3]. Our paper perhaps is a logical partner and expansion on this study. In the future perhaps integrating these two concepts would benificatial. The incorporation influence assessment might be a beneficial step towards identifying how communities would form around influential nodes.

Network Analysis of Page Likes from Facebook User Profiles [4] This study used the Python library NetworkX to analyze network data from Facebook in graphic form. Using Facebook ego network which structures data along a central vertex, the exercise started by analyzing betweenness centrality to determine the most connected individuals present in the dataset. After calculating betweenness centrality, the highest scoring 10 in the dataset were placed as central hubs within the graph. Next different communities were identified using community detection algorithms focusing on the modularity of the network or the fraction of edges.

2 Process and Approaches

Figure 1 identifies the steps that are necessary to complete the objectives of this report. Data are collected from Stanford Network Analysis Project. Once the data are collected, the next would be to clean up the network graph, perform high level analysis, calculations, detection community method. Then, all findings will be analyzed, visualized, and compared to get final results.

Figure 1: High Level Process

2.1 Collecting Data

Dataset [5] The dataset is November 2017 Facebook Large Page-Page Network from Stanford Network Analysis Project [5]. The dataset is a compressed zip file that contains the following files :

- `musae_facebook_edges.csv` – File that contains all edges between pages
- `musae_facebook_features.json` – JSON file that contains features of the nodes - this file will not be used.
- `musae_facebook_target.csv` - File that contains list of node ids, page names, and page types.

Each edge represents a mutual like between pages and each node is a Facebook page. There are four categories classifies for the Facebook pages: governmental organizations, politicians, television shows and companies. The raw data contains 22,470 nodes and 171,002 edges.

2.2 Data Cleaning and High Level Information

Tools The following software and tools were used to perform the analysis:

- **Python and Python Packages** - Python packages include Networkx, Pandas, and Matplotlib
- **Jupyter Notebook**
- **Gephi** - visualization and analysis tool for network graph

Cleanup After using Pandas Python package to read and combine the edges list and pages categories information csv files, the next step would be converting the data into graph object and removing self-loops, edges that connect a node to itself. After removing self-loops, the total edge count reduced from 171,002 to 170,823.

High Level Information The average degree is 15.2045, which means that on average, each node connects to about 15 other nodes.

Figure 2 provides the composition of page category within the network.

The same composition can be shown and visualized using Gephi as shown in **Figure 3**. One advantage of using Gephi over pie chart is the visual of clustered areas within a network. There is a highly connected area and some small clustered areas within governmental organization pages. Politicians and tv show pages also have more clustered area toward the outer part of the graph and Company pages don't have any obvious cluster.

However, graphing the whole population makes it difficult to get more information out of the network. One way to workaround that

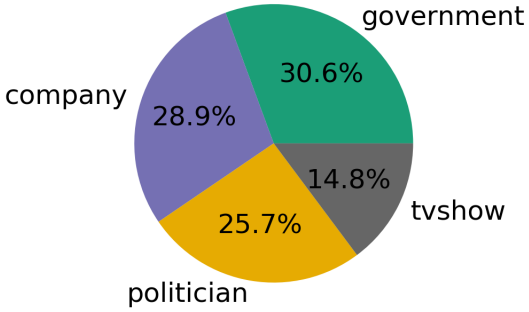


Figure 2: Percentage of Page Category

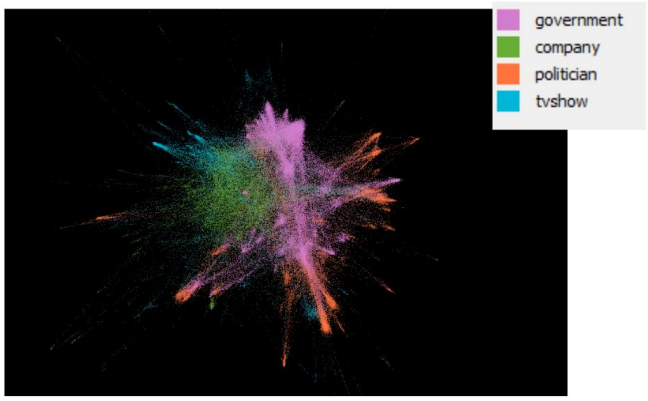


Figure 3: Whole Network Visualization by Page Category

is to filter the network into a smaller network of high degree nodes for more observations. By filtering the whole network to only select giant components with higher than 100 degree nodes, it resulted in a smaller network that only contains 314 nodes with 5,686 edges.

Figure 4 shows that governmental organizations and politicians pages made of the majority

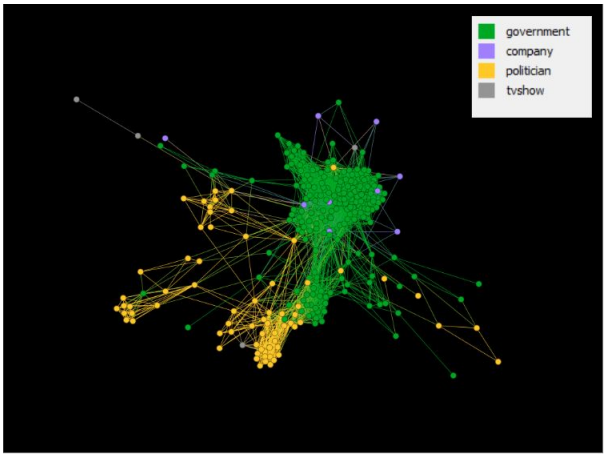


Figure 4: Visualization of Network with Nodes Greater Than 100 Degree

page_name	degree	page_type
U.S. Army	709	government
The White House	678	government
The Obama White House	659	government
U.S. Army Chaplain Corps	650	government
Honolulu District, U.S. Army Corps of Engineers	504	government
U.S. Department of State	468	government
FEMA Federal Emergency Management Agency	448	government
European Parliament	417	government
Army Training Network (ATN)	408	government
Defense Commissary Agency	387	government

Figure 5: Top Ten Highest Degree Pages

of the population in a network of higher than 100 degree nodes. It is also quite easy to identify the cluster areas (about 5 areas).

Another high level measurement that is important and perhaps the most obvious measure of influence was a measurement of the degree of every node in the network. Degree centrality provides a simple yet telling quantitative data point for the number of edges or connections a given node has [6].

Figure 5 contains the list of the top ten nodes that have the highest degree centrality within the whole network. One observation is that all of them are governmental organization pages; it aligns with the high composition of governmental organization pages shown in **Figure 4** earlier. It is interesting to see that the top five highest nodes are United State's related pages. Is it because the majority of Facebook users are in the United States or were the data skewed? Those are the questions that should be answered before making business or important decisions using this dataset.

page_name	degree	page_type
Facebook	380	company
NASA - National Aeronautics and Space Administ...	328	company
CNN	167	company
Walmart	124	company
Whole Foods Market	120	company
Microsoft	112	company
Army Knowledge Online/Defense Knowledge Online	112	company
Starbucks	106	company
Qantas	102	company
KLM Royal Dutch Airlines	97	company

Figure 6: Top Ten Highest Degree Company Pages

Additional filter of page type was added to get the top ten highest nodes of company pages, television shows, and politicians shown in **Figure 6**, **Figure 7**, and **Figure 8** respectively.

Both of the top highest nodes under company and politician pages are about half of the degree amount comparing to the highest node in governmental organization. Television show pages seems to have the least influential power within

page_name	degree	page_type
Today Show	141	tvshow
Home & Family	137	tvshow
tagesschau	119	tvshow
The Simpsons	110	tvshow
Glee	101	tvshow
So You Think You Can Dance	99	tvshow
Family Guy	91	tvshow
Dancing with the Stars	90	tvshow
MasterChef	90	tvshow
New Girl	90	tvshow

Figure 7: Top Ten Highest Degree TV Show Pages

this network.

2.3 Calculate Measurements

There are other centrality measurements that can be used to identify influential nodes within a network. It is important to understand the similarities and differences between these measurements and degree centrality to help with future network analysis. If degree centrality yields similar results to the rest of the other measurements, it might save time for researchers and scientists from having to calculate and perform deep-dive analysis.

2.3.1 Betweenness Centrality

One of the useful centrality measurements to analyze a network is betweenness centrality. Betweenness centrality finds the path between every node in the network, then finds the shortest path and the nodes that fall into the most of these shortest paths will have the highest betweenness score [6].

page_name	degree	page_type
Barack Obama	341	politician
Manfred Weber	326	politician
Joachim Herrmann	320	politician
Martin Schulz	236	politician
Arno Klare MdB	226	politician
Katarina Barley	224	politician
Katja Mast	222	politician
Angela Merkel	217	politician
Niels Annen	199	politician
Sir Peter Bottomley MP	174	politician

Figure 8: Top Ten Highest Degree Politician Pages

Figure 9 contains the top ten highests betweenness centrality nodes. The top ten nodes as calculated by their betweenness scores provides six governmental organization nodes, two politicians and two companies. The range of the scores is between 0.0155 for the 10th highest, and 0.116 for the 1st highest score.

2.3.2 Closeness Centrality

The next centrality measurement to analyze is closeness centrality. Closeness centrality provides the average path length between one node to every other vertex [6].

Figure 10 provided two companies, one

page_name	betweenness	page_type
Facebook	0.115790	company
Barack Obama	0.089628	politician
The Obama White House	0.039820	government
The White House	0.039805	government
European Parliament	0.025954	government
CNN	0.022697	company
NATO	0.019557	government
Joachim Herrmann	0.019308	politician
U.S. Embassy Ottawa	0.017641	government
U.S. Department of State	0.015456	government

Figure 9: Top Ten Highest Betweenness Centrality

page_name	closeness	page_type
Facebook	0.324158	company
The Obama White House	0.317475	government
The White House	0.317413	government
Barack Obama	0.316438	politician
U.S. Embassy Ottawa	0.303557	government
CNN	0.302080	company
U.S. Department of State	0.302027	government
U.S. Army	0.298901	government
U.S. Embassy in Mozambique	0.297753	government
NATO	0.297469	government

Figure 10: Top Ten Highest Closeness Centrality

politician, and seven governmental organizations in the top ten overall nodes with the highest Closeness centrality scores. The 10th highest had 0.0297 and the highest had 0.324.

2.3.3 Eigenvector Centrality

Lastly, Eigenvector centrality need to be measured for this network. Eigenvector centrality measures the connections a node has, not just the quantity, but also the quality or the influence of those connected nodes. A high Eigenvector score would indicate that a node is connected to many other highly influential, high scoring nodes [6].

page_name	eigenvector	page_type
U.S. Army	0.177841	government
U.S. Army Chaplain Corps	0.160628	government
Honolulu District, U.S. Army Corps of Engineers	0.136369	government
Army Training Network (ATN)	0.121057	government
Defense Commissary Agency	0.120850	government
The White House	0.117068	government
The Obama White House	0.115851	government
U.S. Army Materiel Command	0.113773	government
United States Air Force	0.108163	government
U.S. Army Garrison Red Cloud	0.106441	government

Figure 11: Top Ten Highest Eigenvector Centrality

Similar to degree, the top ten Eigenvector nodes are government related page types. The range of the top ten eigenvector nodes ranges from 0.106 to 0.178 shown in **Figure 11**.

2.4 Community Detection - Modularity

The second objective evolves around the idea of community within a network and method to detect those communities. According to a research article called "Defining and Identifying Communities in Networks", a community is a group of nodes that have more dense connections between themselves comparing to the rest

of the network [7]. One of the popular methods to detect community is by using Modularity measurement. Modularity is "the the portion of the edge connections within the same cluster minus the expected portion if the connections were distributed randomly" [8].

Gephi can calculate and divide a network into each Modularity Class by providing a resolution number. According to Gephi's instruction, the higher the resolution value, the fewer the modularity classes there are. By passing a resolution value of 10, Gephi detected 8 modularity classes. Half of the modularity classes made up 98.37% of the entire population; therefore, they can be the main four communities to compare to the four categories of Facebook pages (see **Figure 12**).

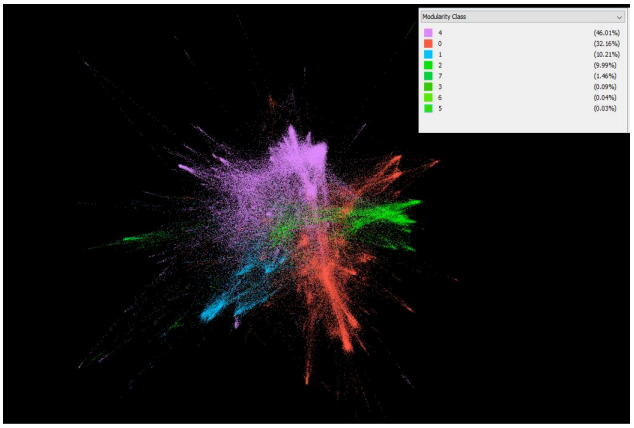


Figure 12: Modularity Classes Graph

References

[1] WebFX, "How much does social media marketing cost in 2020?." `https://www.webfx.com/Social-Media-Pricing.html`. Access Online on 2020-07-03.

[2] A. Karataş and S. Şahin, "A comparative study of modularity-based community detection methods for online social networks," `http://ceur-ws.org/Vol-2201/UYMS_2018_paper_68.pdf`, 2018.

[3] M. Nandi, "Social network analysis with NetworkX." `https://blog.dominodatalab.com/social-network-analysis-with-networkx`, July 2015. Access Online on 2020-07-03.

[4] K. Brauner, A. Kocheturov, and P. M. Pardalos, "Network analysis of page likes from facebook user profiles," ŭhttps://ufdc.ufl.edu/UF00091523/00851.

[5] S. N. A. Project, "Facebook Large Page-Page Network." `https://snap.stanford.edu/data/facebook-large-page-page-network.html`. Accessed Online on 2020-07-03.

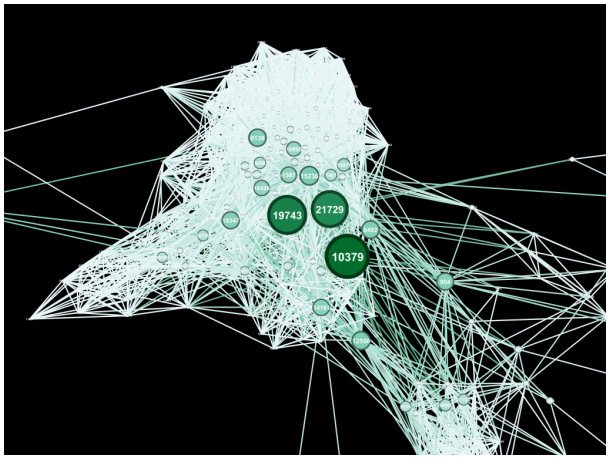
[6] M. E. Newman, "The mathematics of networks," *The new palgrave encyclopedia of economics*, vol. 2, no. 2008, pp. 1–12, 2008.

[7] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. 9, pp. 2658–2663, 2004.

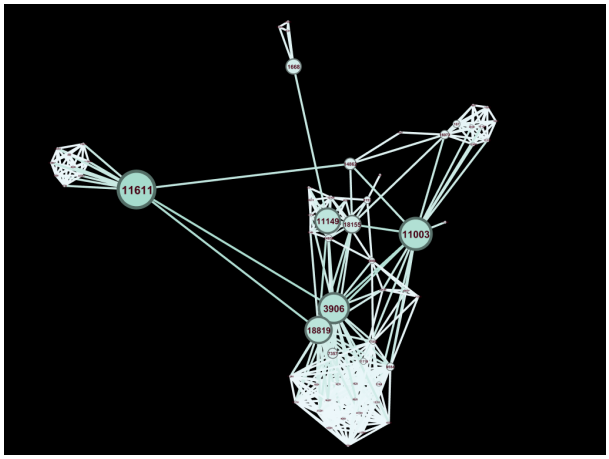
[8] W. Li and D. Schuurmans, "Modular community detection in networks." `https://www.ijcai.org/Proceedings/11/Papers/231.pdf`. Accessed Online on 2020-05-03.

Appendices

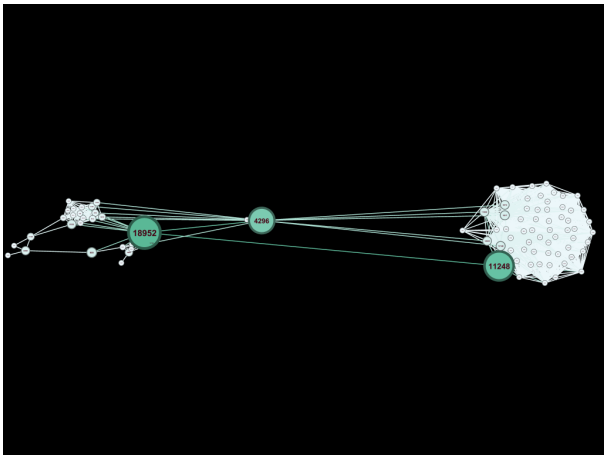
Appendix A



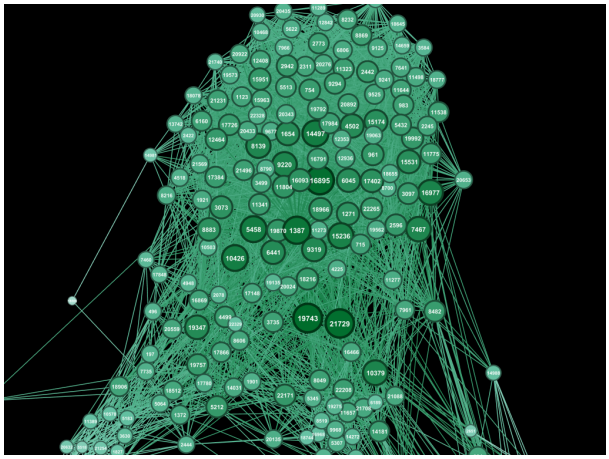
Appendix B



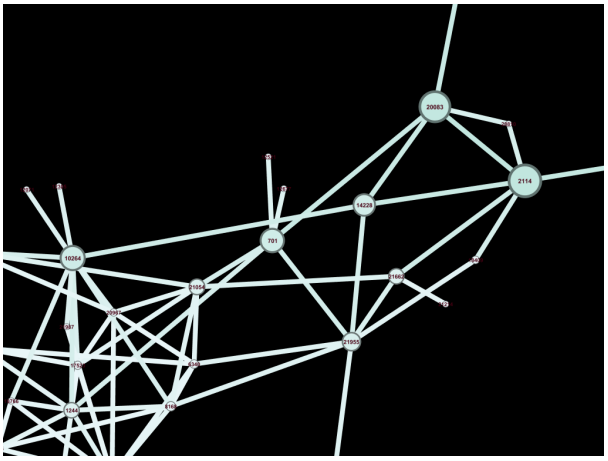
Appendix C



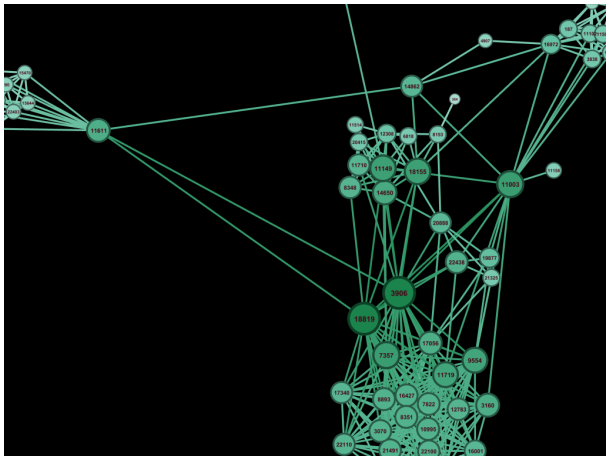
Appendix D



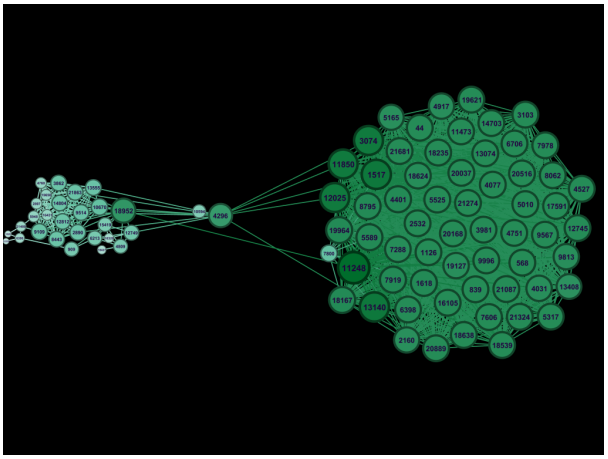
Appendix F



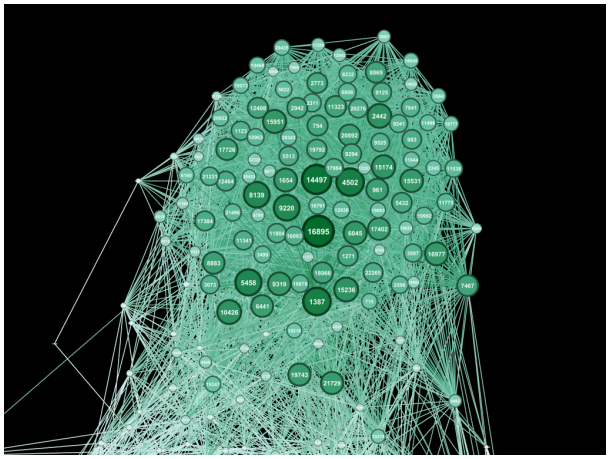
Appendix E



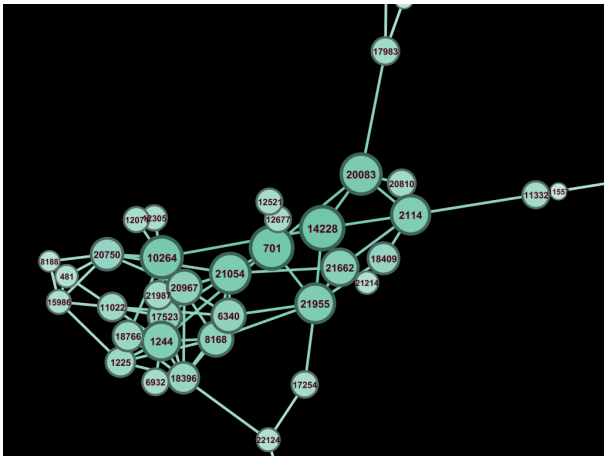
Appendix G



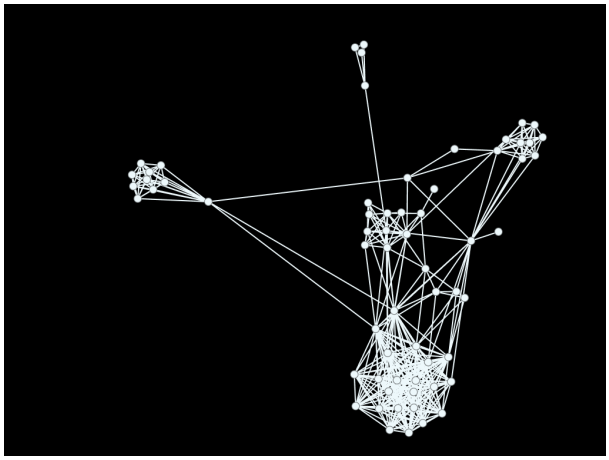
Appendix H



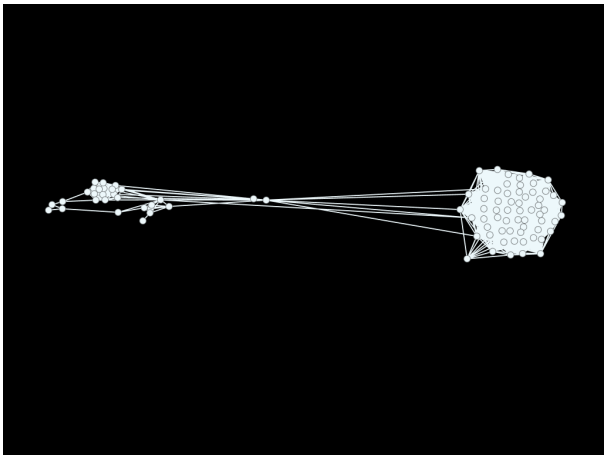
Appendix J



Appendix I



Appendix K



Appendix L