CS579

Project 1:
Social Media Data Analysis

Authors:
Brandon Bennitt
Kemen Goïcoechea

Supervisor:
Kai Shu

Spring 2022

Contents

1 Introduction

In this project, we will visualize and analyze our Twitter network. Our objective is to familiarize ourselves with a Social Network API and a network visualization software. In the end, we wish to calculate meaningful statistics to provide insight into the network we visualized.

2 Data Collection

In this project, we decided to create a friendship network, which means that we analyzed the common links between someone's followers. Our starting point was Brandon's twitter account as he has around 280 followers. We decided that looking at his followers followers/friends would give us insight into how his followers are connected to one another.

2.1 API Account Creation

Our first step was to create a Twitter API account. Creating a Twitter API account was free, quick, and easy. Then, we needed to upgrade the account from "Essential" to "Elevated" to be able to gather information about Brandon's followers.

2.2 Twitter Data Gathering

Once the developer account was granted approval, we had access to an API Key and Access token. Thus, we tried to use it with *tweepy* by following *this tutorial*. Then, we found this *twitter-graph library*, created by Edouard Leurent. This library automates the conversion of the raw twitter data to a *.csv* file. Therefore, we used it to gather the followers' data from Brandon's account. One obstacle we faced was that only 15 followers' accounts could be scraped every 15 minutes. To automate the process until completion, we wrote a python script that calls a function in the library every 15 minutes.

When we first started scraping the data, we came across a follower that consistently caused the twitter-graph library to seemingly do nothing. After a few runs of seeing this follower stall out the script, we manually inspected the follower on twitter. We discovered the reason that he was stalling out the data scraping process was because he had over 1 million people he was following and 250 thousand followers. Since this seemed to be a random person following Brandon that he nor any of his followers knew, Brandon removed him as a follower on his twitter account. Upon successful removal of the follower, the script was able to continue on as expected.

Finally, with a few hours of run, we obtained a dataset of 281 nodes and 8994 edges.

3 Data Visualisation

3.1 Gephi configuration

As suggested in the project description, we used *Gephi* to visualize the graph. It is an intuitive software that has a graphical interface. We followed the tutorial provided by *twitter-graph* to set up Gephi with the resulting *.csv* files.

We tried many configurations of parameters to visualize the graph. The best layouts that we found were $ForceAtlas$ and $ForceAtlas2$. In particular, these two layouts allow users to control the attraction between the nodes and if the nodes overlap. Then, we were able to change the color, the size and the thickness of the different components to make the graph clearer.

Below is a screenshot of the overall network created from Brandon's twitter followers. In the next section, we will dive further into the clusters that were naturally occurring after the $ForceAtlas2$ layout was ran until equilibrium.

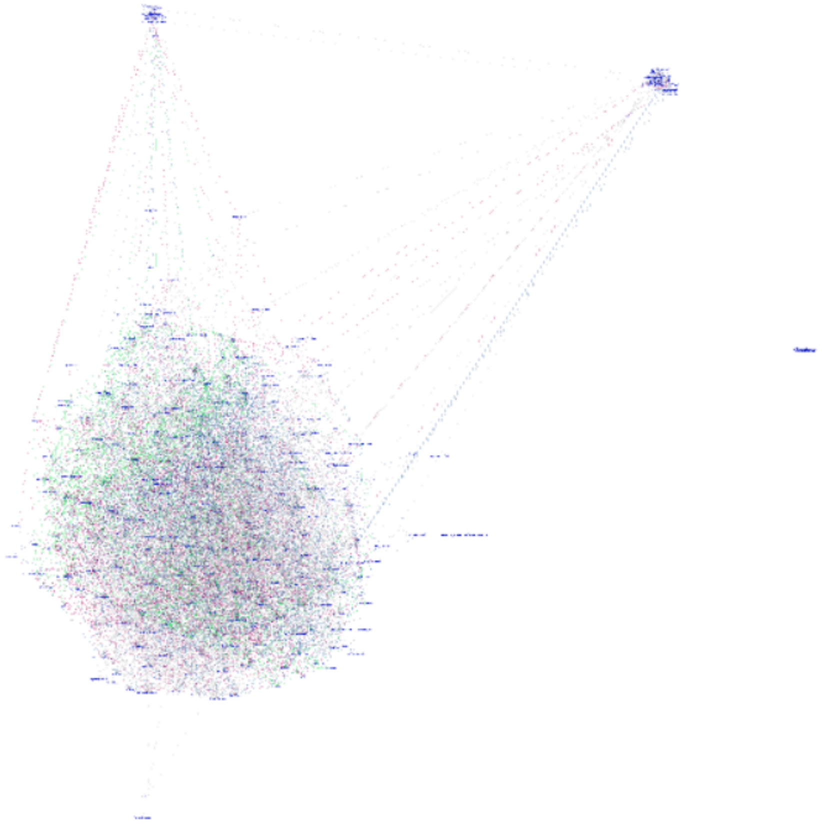

Figure 1: Full Twitter Followers Network

3.2 Analysis

At a first look, one can notice that there are multiple different clusters that are inherently separated in the graph. By looking a little closer into the nodes that are clustered together in this network, Brandon was able to provide some insight into why these clusters exist based on his knowledge of his followers.

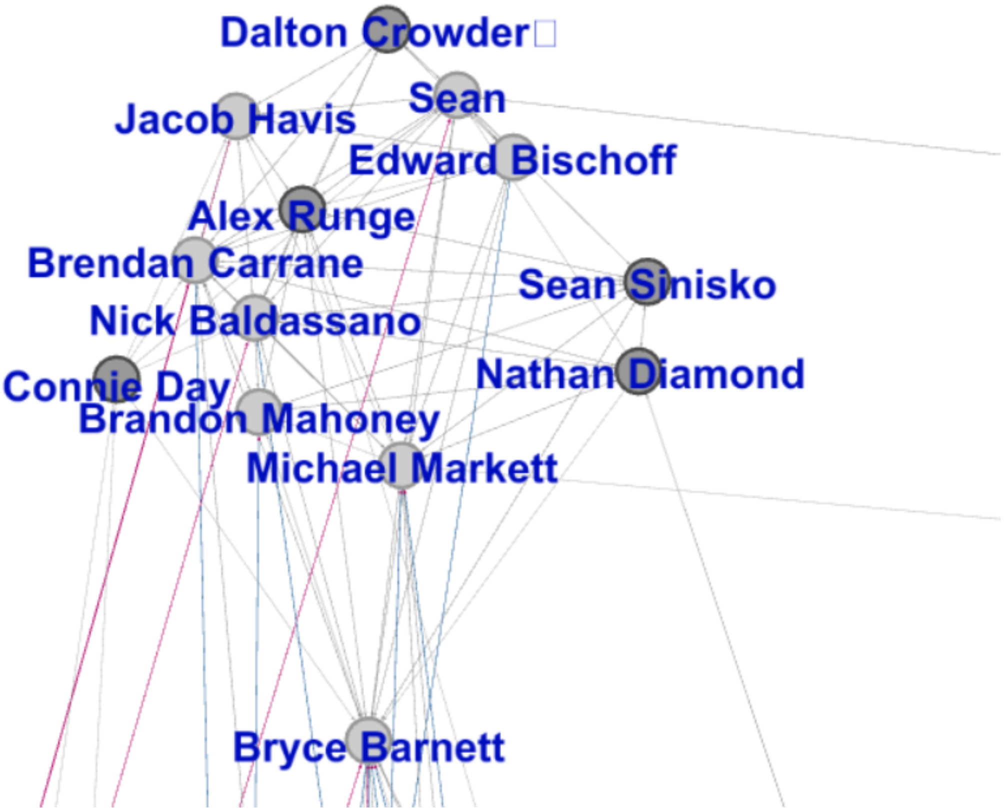

Figure 2: Travel Baseball Team

Looking at this small cluster that is in the top left of the full twitter graph, we can see Brandon's travel baseball team from middle school. Since all of these people know each other but don't know many other people who follow Brandon, they are grouped in their own cluster in the top left of the graph. What is interesting to point out is Bryce Barnett played on the travel baseball team, but also lived closer to Brandon growing up than the rest of the kids on the travel baseball team. Therefore, it seems that since Bryce may know more people from the rest of Brandon's network, he is being slightly pulled out of the travel baseball group and towards the the rest of the network.

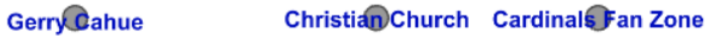

Figure 3: Seemingly Random Nodes From Top Center

Looking at these three nodes, we can see that they are not connected, but they were grouped together relatively tightly. One theory for this may be that they have similar followers or follow the same people, but none of them are in Brandon's twitter network. Otherwise, we were not able to draw any conclusion on why they were grouped close together, except the fact that it is fairly close to a lot of baseball players in Brandon's network. This makes sense since Cardinals Fan Zone is a baseball fan page. The other two nodes do not seem to fit this category though.

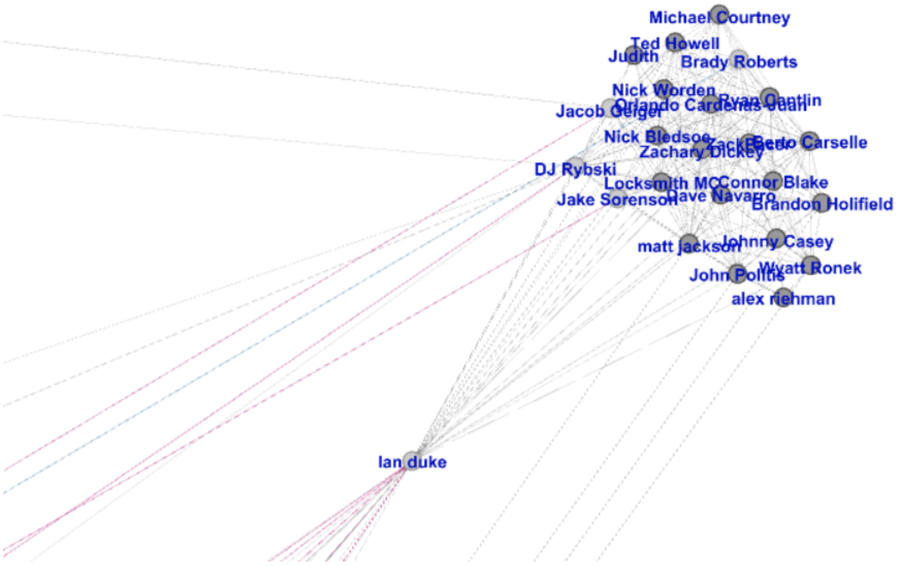

Figure 4: IIT College Baseball

This cluster is pictured in the top right of the full twitter graph. All of the people pictured in this cluster are players or coaches on the IIT baseball team with Brandon. The obvious outlier is Ian Duke. Ian was a high school friend of Brandon before going to IIT, so he knows a lot of the same people that Brandon does from high school. Since he was a year older than Brandon, he does not know all of the same people from high school. Therefore, since they know more people in common from the IIT baseball team, it makes sense he is in between clusters but pulled closer to the IIT baseball team.

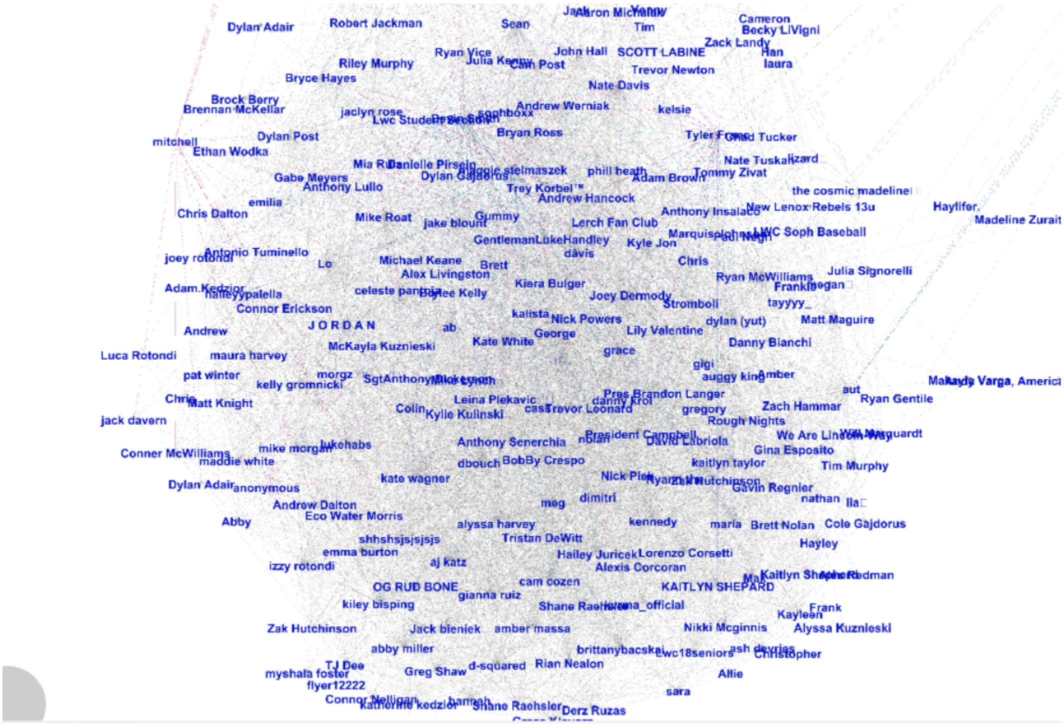

Figure 5: High School

This cluster is pictured in the bottom left of the full twitter graph. These are all of the people that Brandon is connected with from high school. A few interesting nodes that need to pointed out are all of the nodes at the top of the high school cluster (also pictured in the next figure below), as well as the node 'aut'. The nodes at the top of the high school cluster were on the high school baseball team with Brandon. Since these people all play baseball, it makes sense they are closer to the other two baseball clusters on the full twitter graph. The node labeled 'aut' is Brandon's girlfriend from high school. Since they went to the same high school and are still dating through college, she is connected with people from the IIT baseball team as well Brandon's high school. Therefore, it makes sense she is on the right side of the high school cluster.

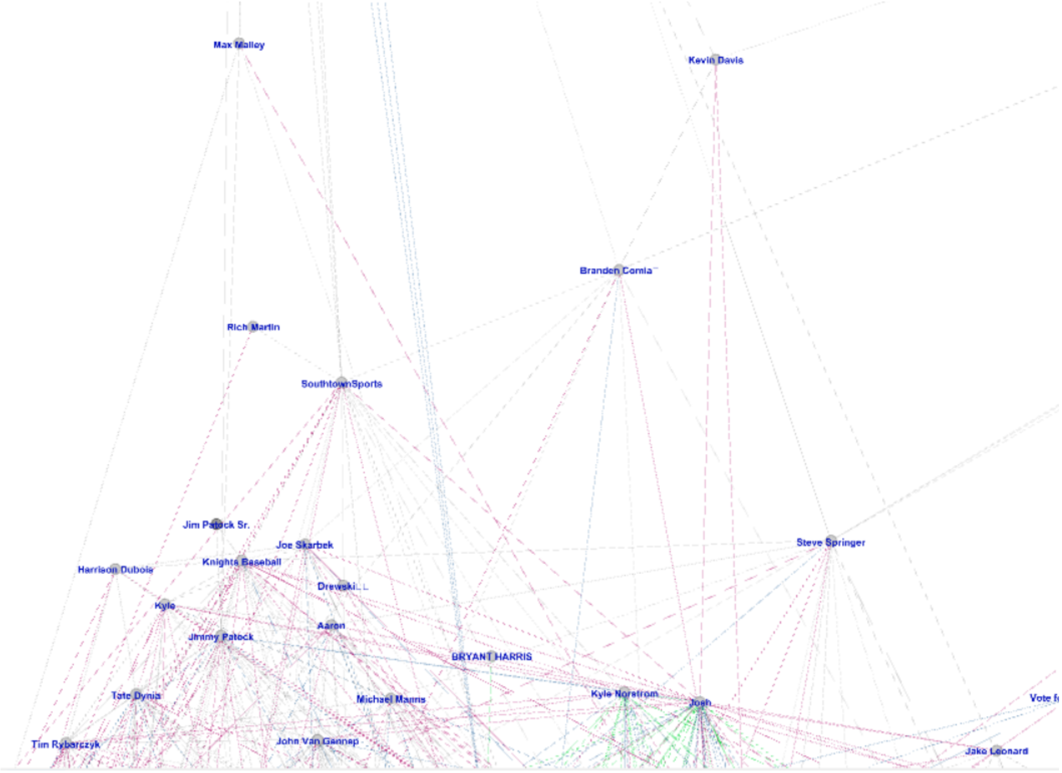

Figure 6: High School Baseball

Pictured above is the top of the high school cluster described earlier. The people that are somewhat in between the high school baseball cluster and the travel baseball cluster were opponents of Brandon when he played baseball in high school. Since the community of baseball players in Brandon's county were close and knew each other, it makes sense they were placed in between the travel baseball players and the high school baseball players.

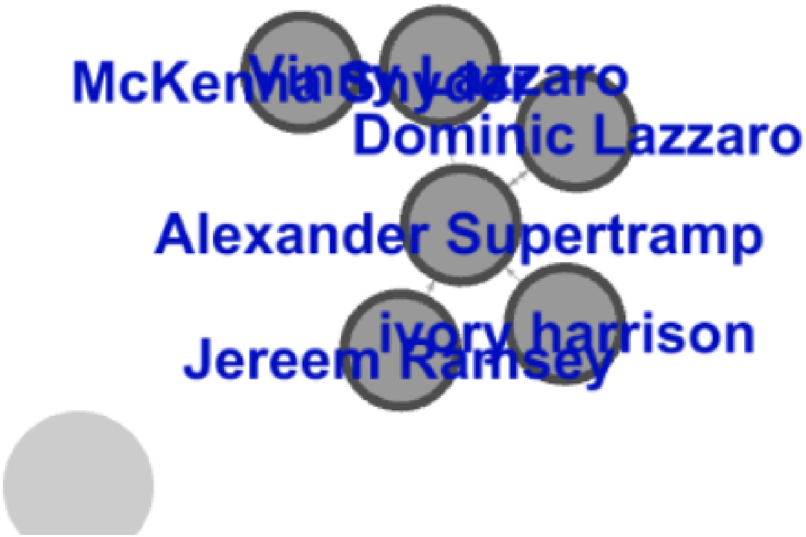

Figure 7: Cousins

This cluster is pictured on the right side of the full twitter graph. It contains nodes that are all connected, but have no edges to any other nodes in the graph. These people are all relatives of Brandon. Since Brandon's family and friends don't have too much crossover, it is reasonable they are connected yet isolated from the rest of the nodes.

To summarize, each cluster and node's location could be logically explained from Brandon's knowledge of his followers and their relationship to one another, even if they are all not fully connected by edges.

4 Network Measures

In the study of graphs and networks, the degree of a node in a network is the number of connections it has to other nodes and the degree distribution is the probability distribution of these degrees over the whole network.

Below are the distribution graphs for the total degree distribution, the in-degree distribution, and the out-degree distribution. We can see that the general trend in each of the graphs follows Zipf's law, which states the rank-frequency distribution is an inverse relation. This means that we should expect few nodes to have a large degree and many nodes to have a small degree. Below, we see the law holds.

The average degree was calculated to be 32.

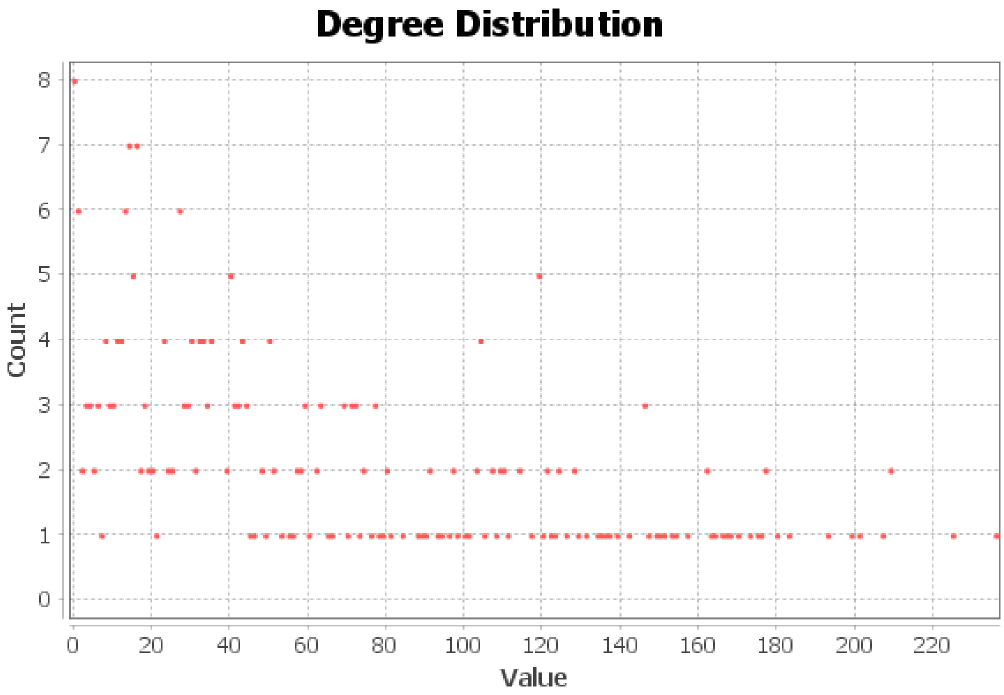

Figure 8: Overall Degree Distribution

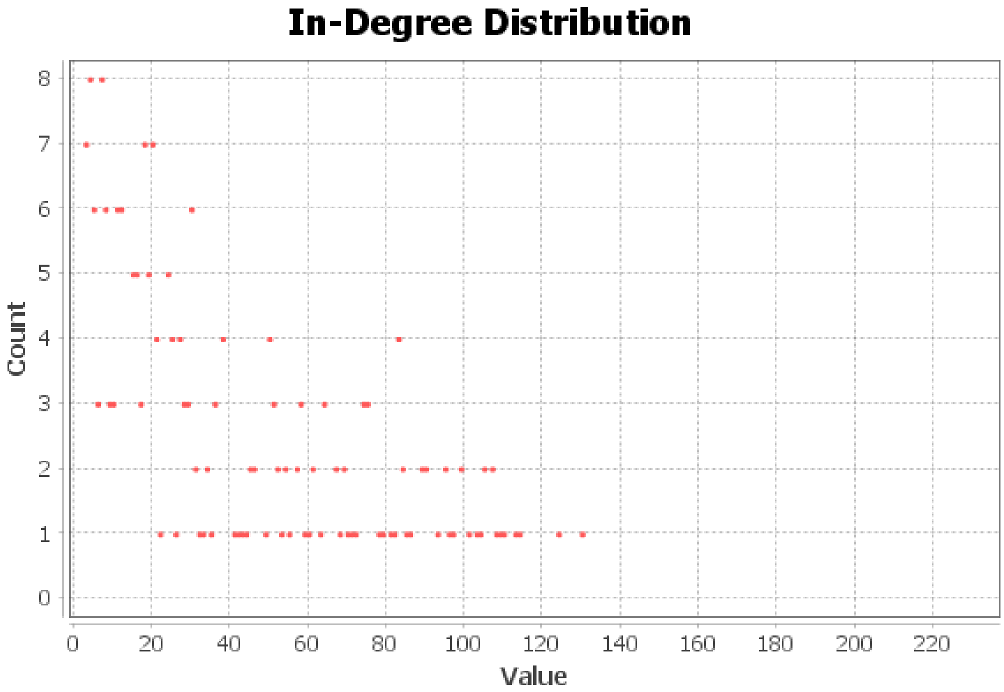

Figure 9: In-Degree Distribution

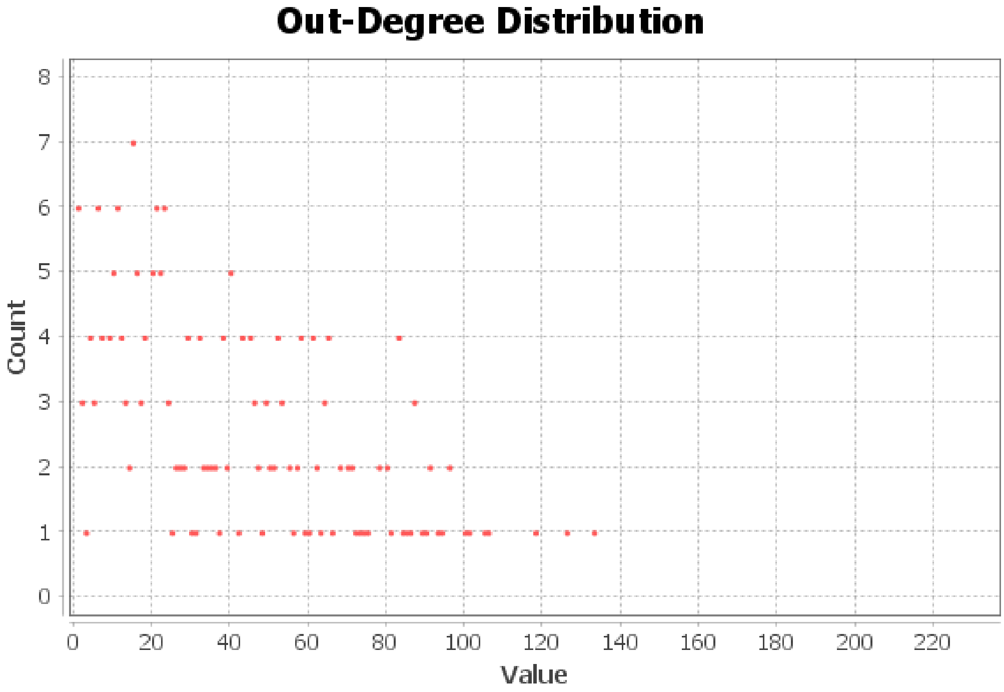

Figure 10: Out-Degree Distribution

In this project, we also decided to analyze the network diameter, betweenness centrality, and closeness centrality. Network Diameter is the longest graph distance between any two nodes in the network (i.e. How far apart are the two most distant nodes). Betweenness centrality measures how often a node appears on shortest paths between nodes in the network. Closeness centrality is the average distance from a given starting node to all other nodes in the network. The network diameter was measured to be 6. Below, we can see the distributions for betweenness centrality and closeness centrality.

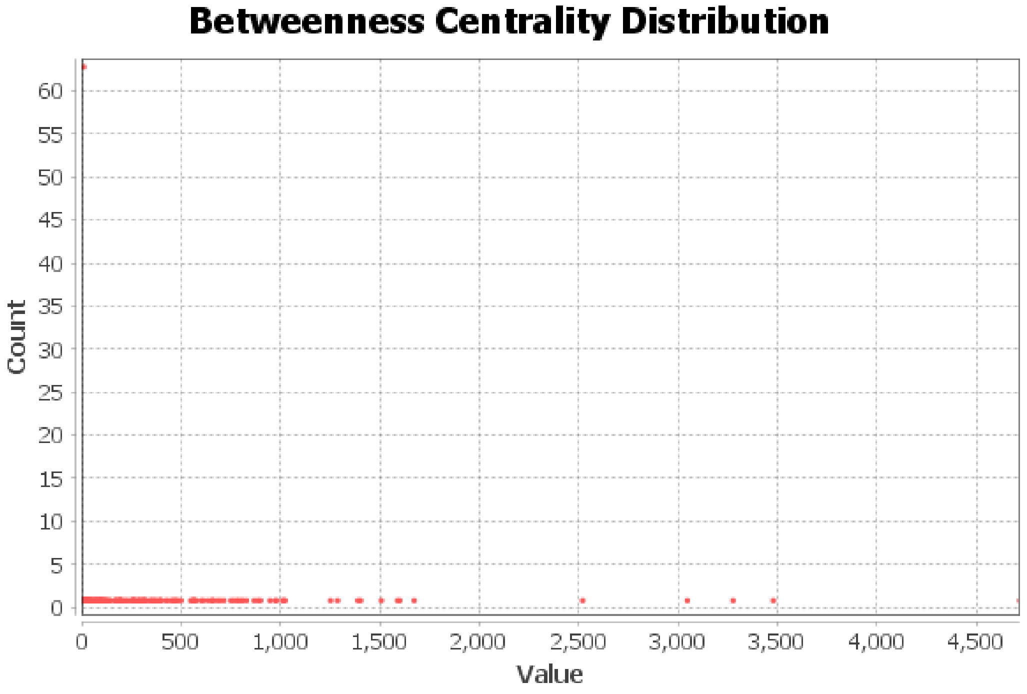

Figure 11: Betweenness Centrality Distribution

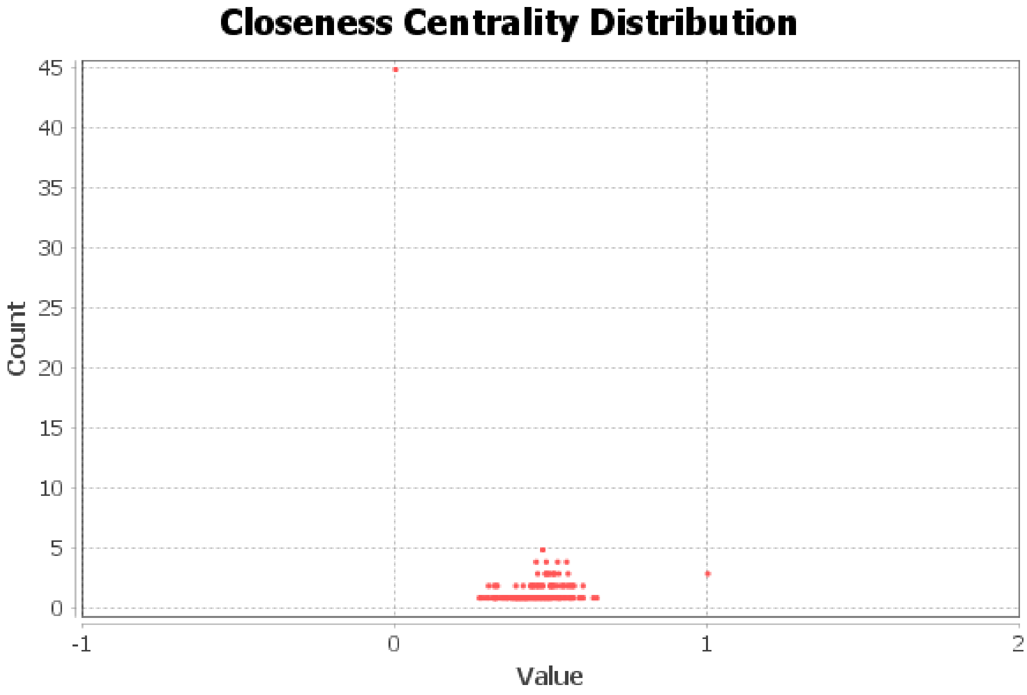

Figure 12: Closeness Centrality Distribution

We can see that the betweenness centrality distribution is heavily weighted towards a value of 0, and is sparse for higher values. The closeness centrality distribution sits right around .5.

5 Conclusion

To sum up, we were able to generate and visualize a Twitter account followers' network. This project allowed us to create and use a Twitter's API account which can be very useful in the future. The gathering and visualization of the data was extremely surprising as we did not expect to be able to collect as much data freely. The accessibility of that much information is almost scaring.

Overall, we are proud of the resulting graph obtained. A lot of information can be extracted from its structure, especially when combined with prior knowledge of the nodes included.

6 Group Participation

The work load has been well divided between the members of the group. On the one hand, Kemen took care of the *Data Collection* part with the creation of a Twitter API's account, the understanding of the libraries and the automation of the gathering of data. On the other hand, Brandon worked on the *Data Visualization* and the *Network Measure* parts. This includes the understanding of the software *Gephi*, the parameter tuning for the visualization and the measures extraction.

Then, the two members explained to each other the skills learned. Finally, the report has been equally written.

References

[1] Tweepy library, *tweepy*

[2] Steve Hedden, *How to download and visualize your Twitter network*

[3] Edouard Leurent, *twitter-graph*

[4] Gephi Layouts, *Gephi Tutorial Layouts*

[5] Gephi Visualization, *Gephi Tutorial Visualization*

[6] Degree Distribution Definition, *Degree Distribution Wikipedia*