CSE 472 Project I - Social Media Data Analysis

Members : Tyler Black, Ujun Jeong

Task : Twitter Data Analysis

In this task, we analyze a real time twitter stream and visualize how actions such as mentions, replies, and retweets could be represented in graphs.

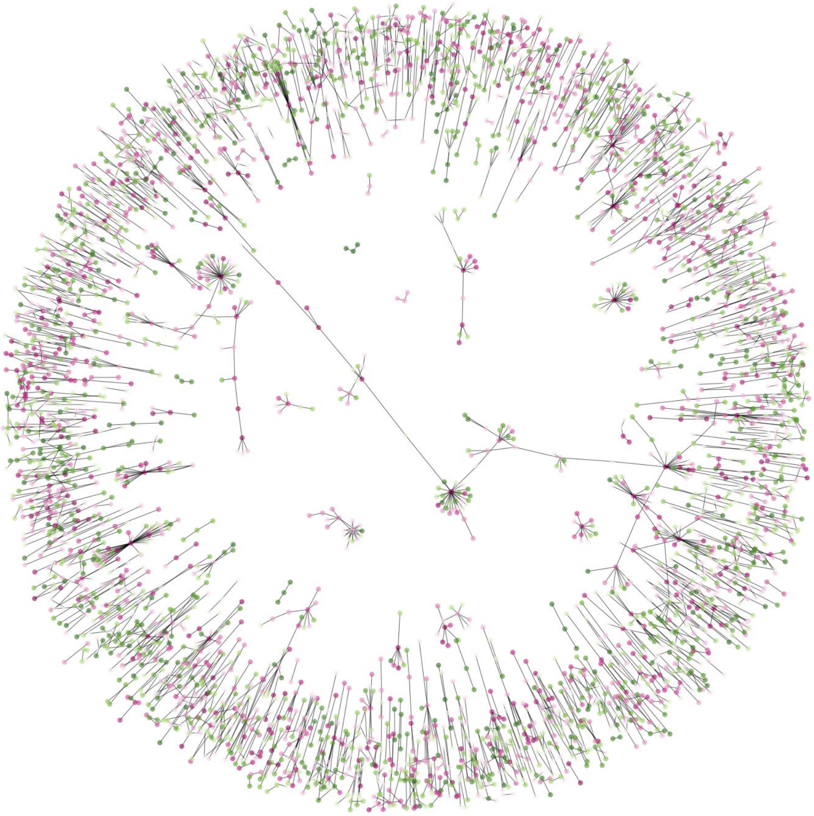
Step 1 : Data scraping

To crawl the data, we connected to Twitter through the API and the Tweepy package in python. We collected 3000 tweets which contain the keyword “Covid”, and analyzed how these tweets are published, received, and propagated through the network. We felt that 3000 tweets would represent a snapshot of activity at a given time in twitter, and could inform us of what the discourse would be. We chose the keyword Covid because it is a relevant current event that involves all sectors of the population for discourse.

The collected tweets are saved in json format and translated into a dataframe using the Pandas library. We represent users as nodes and edges as mentions, replies, and retweets. This forms a basic interaction network. In the end, these data are visualized in connected/unconnected subgraphs with undirected edges.

Step 2 : Data visualization

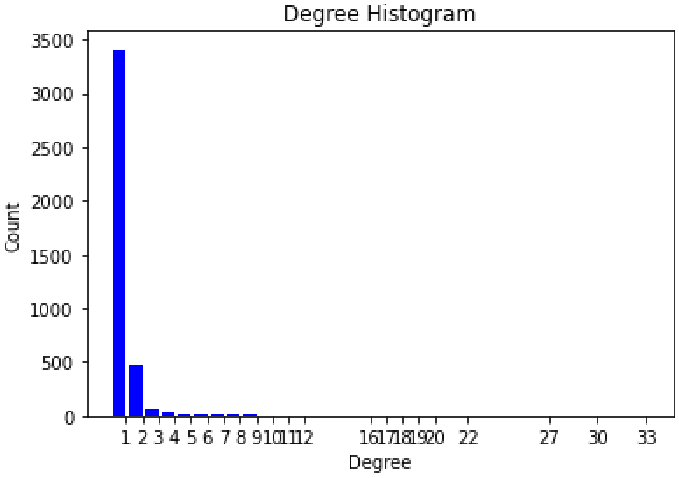
To visualize the data, we utilized the python library NetworkX to create an undirected graph of the 3000 tweets. Through visual analysis, we can see what the state of the discourse is like with respect to the network at the time of the tweet extraction. We see many different nodes, some connected, but without a major central connection of nodes. Some are larger than others. What this means, conceptually, is that there are users creating discourse containing the keyword “Covid” and other users are retweeting, quoting, and mentioning the original tweet. This graph indicates a fact we know, that the discourse on a topic is widespread and with original content from many users.



<Figure 2.1> Visualized result of tweet data based on “Covid” keyword.

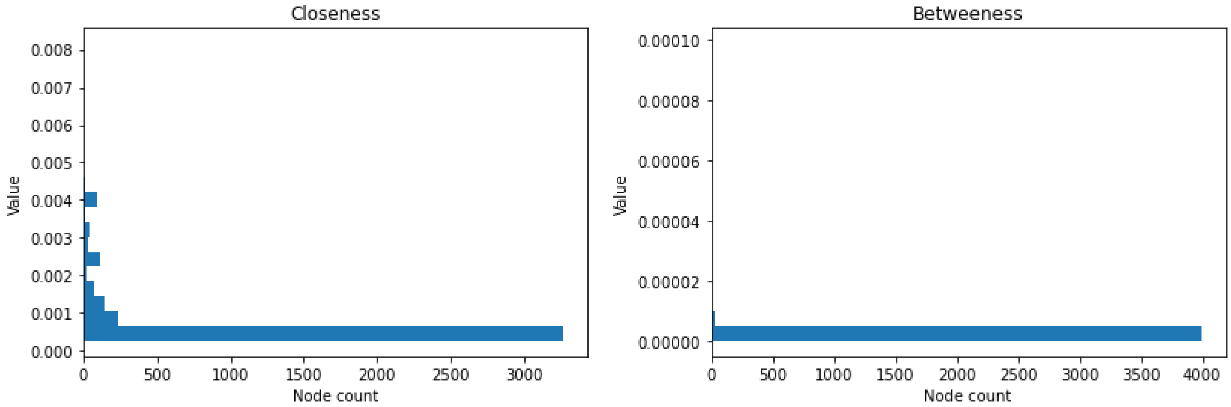
Step 3 : Network measures

First of all, we presented the degree distribution of the graph by histogram. Degree is the number of nodes adjacent to a node. In this graph, the x-axis represents the sequence of degrees which appear in the network, and the y-axis represents the count of each degree.



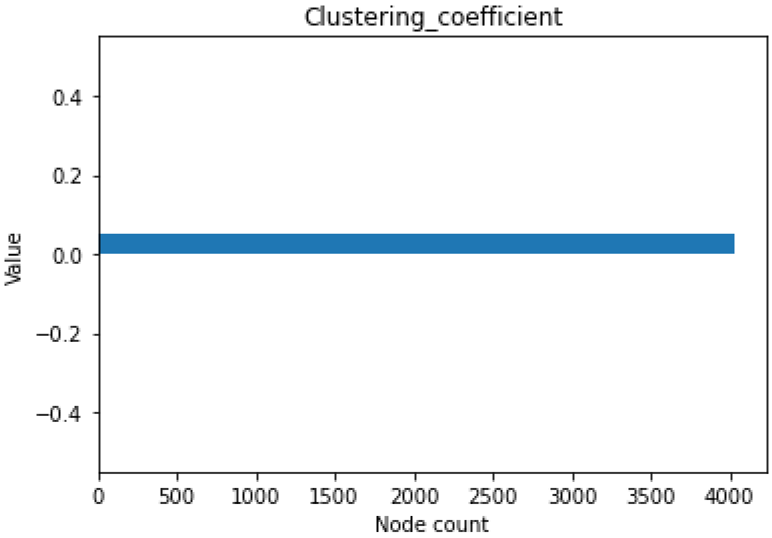
<Figure 3.1 Histogram of degree distribution>

Secondly, we calculated three types of measurements on the graph; such as clustering coefficient, closeness, and betweenness. Here, we used NetworkX again to plot measurements. From this experiment, we've found that our graph is mostly not connected to each subgraph, and it caused most of our measurements to result in zero. This is because most nodes in the subgraphs are propagated from one central node.



<Figure 3.2 Distribution of Closeness and Betweenness>

Lastly, we've proved the fact there are few neighbors in each node by measuring clustering coefficients. Most of clustering coefficients converged into zero in this graph, and it shows that there are almost no neighbors per each node. Through this experiment, we presented that characteristics of graphs are determined by what key word we are using, and how it is propagated among users.



<Figure 4.3 Distribution of Clustering coefficient>