Rohan Anand
DS 210
5/10/21

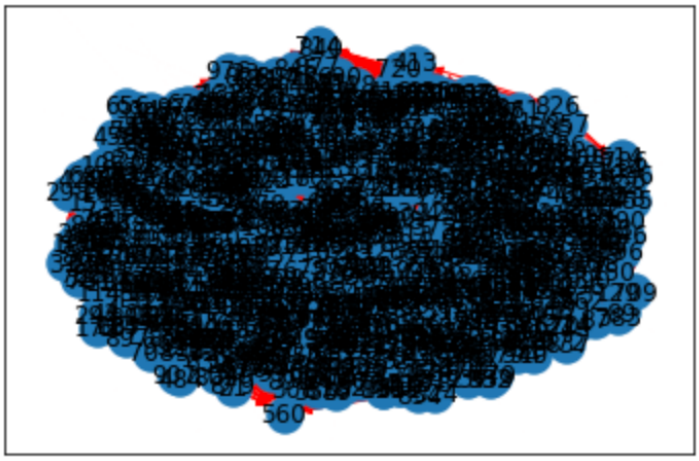
Introduction

For my final project, I choose a data set containing a subset of Twitter friends. The graph is directed and is presented in the form of an adjacency matrix.

Cleaning the Data

In order to turn the adjacency matrix into a text file I can read, I used pandas in python to first change the names of the users into numbers. Then, I iterated through the rows and columns of the data frame and used a dictionary to map which users were friends with which users. I found that the data was far too long and decided to sample edges by deleting some of them from the dictionary that I had created. I then wrote a text file in the same format as the previous homework, which was each node and what edges that they connected with. I also visualized the graph in python using networkx.

Visualizing the Graph



In the graph above, the blue represents the nodes and the red represents the edges. The large number of nodes crowd the graph, with the edges behind them. The graph is crowded due to the large number of nodes.

Six Degrees of Separation

The estimation for the average path length in a graph is given by ln N / ln E, where N is the number of nodes and E is the number of edges. Therefore, the estimation for the graph would be ln 999 / ln 250 which gives 1.25 rounded to two decimal places. To sample the average distance between nodes, I generated a random and used Dijkstra's algorithm to find the shortest distance from a starting node to each subsequent node in the graph. I sampled ten different nodes, finding the average distance each time, and finding the average distance of those ten nodes. The values fluctuated but centered around 1.74. For example, one sample run wielded an average of 1.75 with average distances being [1.64, 1.74, 1.76, 1.74, 1.75, 1.75, 1.75, 1.78, 1.73, 1.72]. The low number of 1.74 when compared the average in social media networks which is around 6, makes sense considering that this data set consists of friends who interact in a similar community.

Sample Run:

```
Average distance from node to node: 1.7452452452452452
Sample average for 10 different nodes: [1.7497497497497498, 1.7697697697697699, 1.7597597597597598, 1.7567567567567568, 1.7347347347347348, 1.7387387387387387, 1.7287287287287287, 1.7517517517517518, 1.74
27427427427427, 1.7197197197197198]
```

Friends have more friends than you

Friendship paradox claims that on average, your friends have more friends than you, which can be explained by the fact that friends share friend groups, and also have friends outside the friend group who are less likely to be antisocial than social. To approach this problem, I selected a node

and another node who are connected. I then counted the number of undirected edges that each node and calculated the average number of times that a friend had more friends than the original. The number fluctuates around 0.4, with more trials producing an average lower. As a result, the graph does not support the conclusion that a friend has less friends on average than their friends do. The reason for this could be due to lack of samples of friends, which can skew the average higher, since I only used one other friend to test the validity of the paradox.

Sample Run:

```
Average percentage of time your friends have more friends than you: 0.3
Sample of how many times your friends have more friends than you: [1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0]
```

Conclusion

I was able to use a dataset of Twitter friends to analyze degrees of separation, or the average number of steps to get from one node to another, using Dijkstra's algorithm. I found that the estimate for the average number of paths, 1.25, differed from the actual sampled value of 1.75. This is different than the usual 5-6 degrees of separation since the sample size is low compared to the 240 million users that computer scientists used to determine that number, and the data is not randomly sampled from twitter users, but are a community of friends. In addition, I tested the friendship paradox, which claims that your friend has more friends than you, by finding that the graph did not support that your friends have more friends than you on average. This is because the lack of testing with multiple friends leads to more variable results than testing with a large variety of friends which can skew the average upwards.