John Zhao
7 May 2023
DS210

Project Summary:

The following project seeks to determine the degrees of separation in several social network datasets. The datasets that I will be working with are the combined Facebook dataset of active users found at Stanford SNAP, and the Twitch dataset of active users found at the same locations. The datasets are also included in the github repository and can be downloaded at the following links:
Facebook Network: https://snap.stanford.edu/data/ego-Facebook.html
Instructions: download facebook_combined.txt.gz -> open and extract, drag file into src folder.
Twitch Network: https://snap.stanford.edu/data/twitch-social-networks.html
Instructions:
Download twitch.zip -> open file -> open ENGB -> download musae_ENGB_edges.csv -> drag file into the src folder.
Note: the musae_edges.csv has had the top (header) columns removed in the project! Please edit or use the included csv in the src folder.

The definition of the nodes and edges in both datasets is that the nodes represent profiles of users, with identifying information anonymized. The datasets both contain undirected edges which represent the users that are within the same 'circles' or are friends. The Facebook dataset contains a total of over 88,000 edges and over 4000 vertices. The Twitch dataset contains over 30,000 edges and 7000 vertices. The project seeks to determine a metric for the degrees of separation between vertices using a BFS search algorithm. This was chosen because the O(V+E) complexity makes it efficient for medium sized graphs. Distances between vertices were computed using the BFS algorithm to determine the distance between each node and every other connected node in the graph. The average of all the distances is taken as the measure of the degree of separation of the social network.

After implementing the BFS search, I determined that the average degree of separation in the social networks is slightly different but roughly equivalent. The average degree of separation is 3.6925068 in the Facebook network, and 3.6776154 for the English-speaking Twitch networks. A noteworthy observation was that the degree of separation is higher for the Facebook network than the Twitch dataset. This makes sense as the average Twitch user shares the same interest of gaming with other users, whereas the Facebook network includes a more diverse population.

In addition to evaluating the degree of separation of each network, I investigated the importance of the most travelled nodes (or most frequent connectors within a network) influence the degree of separation. I implemented a BFS search to check each neighbor of a node and updating the distances, then repeating for each node reachable from the starting node, then sorting the nodes into a descending order list of highest distances. I then investigated the impact of removing the top 1, 3, and 5 most travelled nodes from the graph. Due to runtime limitations, the project code specifically compares the average separation in the facebook dataset to that when the top 5 most travelled nodes are removed. The result of

removing the most travelled nodes is a minor decrease in the degree of network separation. This makes sense because the most travelled nodes or people connecting a network increase the connectivity of a network and add layers of separation by making a farther network connection (such as a friend of a friend of a friend, etc.) possible.

Nodes Removed	Degree of Network Separation (people)
1	3.6915
3	3.6896
5	3.6877