DS210 - Final Project

Aditya Chopra

Dataset: 'Circles' from Google+

Link:https://snap.stanford.edu/data/ego-Gplus.html

This project explores the data analysis of the six degrees of separation within the Google+ Social Network. Though the dataset is extremely simple, consisting of features (profiles), circles, and ego networks, it portrays the six degrees of separation theory very well. Using a graph, which consisted of 107614 nodes and 13,673,453 edges, we implemented the Breadth First Search (BFS) algorithm, which we integrated into our six degree separation analysis. Overall, the output of the program explored the degree of separation between two randomly selected nodes within the data and determined whether the six degree separation theory held true for the nodes and if it did it outputs its degree of separation.

The six degree separation theory states that any two people on earth, on average, are connected with six or less links apart. This is interesting to explore in terms of social networks, and in this specific dataset the 'circles' of each individual were shared and through our code we were able to 'mark' all the vertices that were connected. The project itself looks at a random dataset of individuals and therefore is perfect for exploring our theory of if the six degree analysis really holds true for this dataset.

The project is divided into 3 different files: graph.rs, bfs.rs, and main.rs. The graph.rs module creates the graph with the nodes and edges, parsing through the datafile. The bfs.rs module explores the BFS algorithm finding the shortest path between 2 nodes and looks at the connected components which are eventually stored in a HashMap. Finally the main.rs uses the other two modules to output the final six degree separation for the two randomly selected nodes.

Running the project is simple. Because it looks at two random data points, the output is different each time, but does not require any user input. It can be runned by using cargo run -- release and the program will randomly take two nodes and find the degree of separation between them.

The output of the dataset produces the amount of components, the random source and target ID's and the degree of separation between them. It was interesting to note that there is only one connected component, which shows us that all the components within the dataset are connected. To test whether all these connected components still follow the six degree separation theory, I wanted to create a function which loops through each node and finds if the degree of separation between them is greater than 6. However, doing this would take countless hours due to the large size of the dataset. Therefore, I decided to choose 5 more random pairs of nodes and test whether

they still follow the six degree separation theory. Though this does not cover the whole dataset, it allows us to randomly test whether the other nodes are also connected and with what degree.

In conclusion, the project analyzes the Google+ Social Network and demonstrates the use of the Six Degree Separation Theory by using randomly selected pairs of nodes and finding the shortest degree of separation with them. The output shows us that most users are connected within six or less degrees as shown by our multiple tests, verifying that for our dataset the theory applies true.