Final Project Report

DS 210

Leonardo Perez-Loynd

I chose a data set from SNAP representing an email network from a large European research institution. I was interested in doing the 6-degrees of separation prompt to find the average number of neighbors it took for any given node to be connected to everyone else in the network.

Most of the interesting discoveries I made were with my bfs_all_nodes function which makes a hashmap storing the node as the key and the average distance of each node to all the other nodes. In my main I printed these values from the hashmap and I was surprised to see that some nodes had either an average distance of 1 degree or Nan. I discovered by looking through the data set that the emails with an average distance of 1 were emails that only had one edge. By thinking about this as an email network you can deduce that these were possibly external emails that were emailing only one other person who did not then go and email other people. Again by looking at the data set I deduced that nodes with an average distance of Nan were emails that had only sent an email to themselves and no other emails.

The other functions I implemented were a read file function to read the text file line by line ignoring the comments at the top that begin with # and returning a vector of tuples representing the edges. The adjacency list function took the vector of tuples and made a hashmap where the keys are the nodes and the values are the vectors of outgoing node connections. The BFS function conducts a breadth-first search with an input of a starting node and the adjacency list and returns a hashmap with the

connecting node and the distance between them. This function is used to work the bfs_for_all function. Lastly, the average degrees of separation function uses the adjacency list and bfs_for_all functions to calculate the average distance of all nodes to other nodes in the set.