Emma Steel

DS210 Final Project Report

In my project I examined a data set that showed the connectivity of a subset of roads in California. The data was originally in a table with the first column being a list of roads and the second column being roads that intersect with the roads in the first column. I wanted to use this data set to better understand the connectivity between the roads and calculate centrality measures. The original data set was extremely large so I cut it down to include the top 10 most connected roads in the first column and then some of their intersection points in the second column, for a total of 100 points to be read into my program. In regards to the centrality measures, I chose to calculate the degree, closeness, page rank, and betweenness measures for the graph and the output of my code displays the best value for each.

For the degree of each node, I had to find how many neighbors each node had and which is equal to its degree. Then I output the one with the largest degree, which happened to be road 4. I have a key that matches the number of the road to its actual name given in the data set and I will attach it below. I was not surprised road 4 had the biggest degree, because it had the most intersection points. The second measure was closeness. For this measure I had to find the average amount of connections it takes to get to each node in my graph. For example, a node with a closeness value of 1 would mean each node in the graph is 1 node away from the original node. Again, road 4 came out to have the lowest closeness value (the lower the better) with a value of 1.25. This means that each node in the graph is approximately 1.25 nodes away from node 4, meaning it's highly connected. I then calculated the page rank because I was curious how this measure of centrality, which I often think of as more random, compared to the other modes that were not random and calculated. Of course with each run of the program, it's possible to get a new value for the page rank because it is a random computation but each time I found that the page rank value itself was very low. The last measure I computed was the betweenness, which was the most difficult for me. I had to figure out how many times each node in the graph interrupts the shortest path from every node to every node in the graph. The node with the highest betweenness was 4, which was not surprising based on my other results.

I found that overall, road 4 was the most connected road in my subset of data both locally and globally in the network of roads. This means, it has the most direct connections, or the most intersections with other nodes, and it is often in the shortest path between other nodes. One limitation to this project is I did not use the entire data set. I only used a small subset for two main reasons. The first being that I was interested in these centrality measures on a small scale, and did not need such a large subset of data to work on. The second reason is because the original data set contained over 50,000 points which would take a substantial amount of time to run on my computer, which was not feasible for this project.

Overall, I was very interested in this project and learning about how to calculate different measures of centrality. If I were to work on this project again, I would definitely work on the efficiency of my code and make sure I am finding the optimal solutions opposed to just code that works. Additionally, I think it would be very interesting to find other measures of centrality with this data set such as eigenvector centrality, which I did not have a chance to do on this project.

Road key:

0-01ST ST	6- 03RD TI ST	12- KEARNY ST	18 - 15TH ST	24- 11TH AVE
1-02ND AVE	7-04TH AVE	13- KANSAS ST	19- 09TH ST	25- 05TH ST
2-02ND ST	8-04TH ST	14 - JACKSON ST	20- 06TH ST	26 - 13TH ST
3-03RD AVE	9-HYDE ST	15 - IRVING ST	21- IDORA AVE	27 - 06TH AVE
4-03RD ST	10- HWY 101	16 - I-280 N OFF RAMP	22- 09TH AVE	28 - 14TH AVE
5-JEFFERSON ST	11- JUSTIN DR	17 -18TH AVE	23- INDUSTRIAL ST ON RAMP	