Final Project Report

Dataset link: roadNet-CA.txt.gz

For this final dataset, I picked a dataset that consisted of California's road network. To explain it in terms of graphs, intersections and the endpoints are considered nodes while the roads connecting these endpoints or intersections are considered edges, which were undirected. The actual dataset had two columns, with the first being the "from" node and the second being the "to" node, in which there were 1965206 nodes and 2766607 edges. I used modules and split up the entire code into four files, with three representing my main code with information I wanted to find, and the fourth representing test functions and the place to run the code that were in the three other files. The main goal was to find degrees of separation and how connected the road network was. The first was dedicated to finding the number of triangles and the fraction of closed triangles in the dataset, with some important features for both pieces of information being the adjacent list and the binary search that I used. I know that California is an extremely large state with a vast number of roads and numerous urban and suburban areas where these roads intersect and end. Therefore, the vast number of triangles is not surprising and more significantly, the fraction of closed triangles being around twenty six percent is not surprising and very telling as well, as its definition is the ratio of the number of triangles to the number of possible triangles. Adding on this idea, another file is specified towards the average clustering coefficient, which is how nodes tend to tie together. Given the vast size of the state, and not reliant on the high-density population of the state, this amount is around four percent, where initially I thought it would be higher. Lastly, I wanted to find out the usual distance between pairs of nodes in which I only tested up to 100,000 and specifically used breadth-first search. The distance was around 105, which was expected given the large cities and highly rural areas. I generally did not find any

piece of information too surprising, but it did lend its insights into how different communities and California's geographical nature affected its road network and the growth of suburban areas in these cities are likely to create more triangles, higher fraction of closed triangles and average clustering coefficient along with a lowered usual distance between pairs of nodes.