Michael Hyder
5/7/2023
DS 210
Final Project Write-up

For my final project, I measured the betweenness and the closeness of a subset of the intersections and roads of Pennsylvania. In my dataset, the nodes were intersections, and the edges were the roads that connected them. The betweenness centrality is a measure of how important an edge is to connect the whole graph together. Edges with high betweenness are often referred to as bridges because they connect two sides of a graph. If a vertex with high betweenness was to be removed from the graph, that would case the graph to be harder to span. The closeness centrality is a measure of how efficiently a node can spread information throughout a graph. A high closeness means the node is a part of many of the minimum spanning trees for this graph. In this project, I wanted to find the vertices that had both high betweenness and high closeness to find the most vital streets in the state. Streets with high closeness and betweenness are high traffic areas which may need maintenance more regularly.

My project starts by reading the data file. Originally, the file contained over 1,000,000 vertices and over 3,000,000 edges, but for the purpose of keeping the run time to a minimum, I conducted my tests on approximately 5,500 vertices. The program then turns the data into a adjacency matrix so the betweenness and closeness can be calculated. The betweenness calculation uses an algorithm that implements a breadth first search to iterate over all the nodes in the graph, and then calculates the betweenness based on how many shortest paths it is a part of. It then calculates the closeness in a similar way except using minimum spanning trees because we want to find how efficient the vertices are. Afterward it combines the measurements by multiply them together to find with vertices have the largest overall impact in the graph.

An interesting observation I found was that the nodes with high betweenness tend to have smaller closeness. This means that the intersections or highways that have the most people go through them are often not the most efficient. However, the vertices with high betweenness have more of an impact that those with high closeness. This makes sense because if a high traffic area were to shut down (removed from the graph), that would case a lot more issues than if a low traffic but high efficiency intersection closed.