Project Report:

Here is the link of the input file and please delete the first four lines of descriptions: https://snap.stanford.edu/data/roadNet-CA.html

This project provides an in-depth analysis of a graph data structure of the California road network, focusing on finding out the shortest distance between two random vertices, the distribution of shortest path lengths, testing if the distribution follows a power law, and identifying highly connected vertices that serve as hubs for the road network to improve the connectivity. The graph is constructed using vertex pairs read from a text file ("roadNet-CA.txt"), and various algorithms, such as breadth-first search (BFS), power-law fitting, and degree calculation, are employed to analyze the graph properties.

To run the project, compile and execute the primary function. The program initiates by reading vertex pairs from the "roadNet-CA.txt" file to create a graph object representing a road network. It then applies the BFS algorithm to calculate the shortest path lengths for two random vertices and computes the average shortest path length across all vertices in the graph.

Next, the project analyzes the distribution of shortest path lengths by creating a frequency map of the path lengths. The results are plotted in a separate PNG file named "distribution.png", which displays the frequency of different path lengths in the graph. From the graph, we can see clearly whether the distribution follows the power law or not. Here, the graph of my distribution looks almost like a normal distribution graph. Therefore, we can see that the California road network is not following the power law. Because road networks are more regular and spatially constrained due to the underlying geography and infrastructure planning, they tend to be a normal distribution.

Besides that, I employed a mathematical way to prove whether the distribution follows the power law instead of judging by intuition from the graph. Using a simple linear regression model, the program tests if the distribution of shortest path lengths follows a power law. This test is crucial, as many real-world networks, such as social networks and the Internet, exhibit power-law distributions. If the input distribution fits a power law, it could indicate that the graph exhibits some significant implications for further research or application.

Additionally, the project identifies highly connected vertices, or hubs, in the graph. These hubs are essential for maintaining the network's connectivity, as they serve as central points in the shortest paths between vertices. The top 10 hubs, ranked by their degrees, are displayed in the output, providing insights into the vertices with the most connections in the graph. However, from the previous distribution, we can see that the distribution of the shortest length does not follow the power law. It indicates California road network with some highly connected vertices,

which might lead to a certain level of local efficiency in reaching nearby intersections. However, this local efficiency does not necessarily translate to a global efficiency resulting in a power-law distribution of shortest path lengths. Some bottlenecks or inefficient routes might prevent it from exhibiting a power-law distribution for the shortest path lengths.

Finally, the project's output comprises several messages, such as the shortest path distances for the random vertices, the average shortest path length, the distribution of shortest path lengths, and whether the distribution fits a power law. Furthermore, it lists the top 10 highly connected vertices and their degrees.