

Collaborators/Credit: Kiko Xu

Dataset:

For my project, I've utilized a dataset that represents the road network of Pennsylvania, found at the following link: <https://snap.stanford.edu/data/roadNet-PA.html> (Stanford Large Network Dataset Collection). A node represents an intersection or endpoint in the road network, and an edge represents a road connecting two intersections or endpoints. The edges are undirected, meaning the road is bidirectional

Project Idea and Implementation:

In my project, I am evaluating the efficiency of a road network by examining the average path length. My program computes the average path length between nodes, providing insight into the networks' connectivity. I'm interested in understanding the level of connectivity in a specific region's road network and its distribution of roads with one, two, or no connections. To address these questions, I plan to calculate not only the average distances between nodes but also the summary statistics to gain a deeper understanding of Pennsylvania's road network. Additionally, I will determine the degree distribution of the network to further comprehend the nodes' connectivity. For my implementation, I have defined two structures, Node and Graph, in my code. The Graph structure, which contains a list of nodes and edges, is implemented to calculate the distance and average distances between nodes, the node degree distribution, and the graph's summary statistics. This structure forms the basis for computing the information required to answer my research questions.

Results:

From implementing a graph to calculate average path lengths, summary statistics, and degree distribution, I've learned that the road network is less dense than anticipated. Initially, I expected the average distance in the network to be around 1,000, but the results indicate an average distance of approximately 3,000. A notable observation is the standard deviation of the network, which is around 35,000. This high variability in path lengths across the road network suggests that there isn't a high concentration of paths of a specific length in the network. The degree distribution provides information on the number of nodes with connections. Many roads in the network are disconnected, as there are many nodes with no connections. There is also a comparable number of nodes with one connection, and a very small number of nodes with two connections (~30,000 compared to ~300). As a result,

many intersections in the sampled data are connected to only one other intersection, or none at all.