ECE232E Project4 Report

Luke Lim Angie Fan

May 2024

Question 1

To understand the range of ρ_{ij}, we consider two scenarios. First, when $r_i(t)$ and $r_j(t)$ are independent, the expected value of their product equals the product of their expected values, resulting in a numerator of ρ_{ij} that is zero, making ρ_{ij} equal to 0. In the second scenario, when $r_i(t)$ and $r_j(t)$ are correlated, we assume $r_i(t) = k \cdot r_j(t)$. This leads to the expected value of their product being k times the expected value of $r_i(t)^2$. Consequently, the value of ρ_{ij} becomes $\frac{k}{\sqrt{k^2}}$. When k is positive, ρ_{ij} equals 1, and when k is negative, ρ_{ij} equals -1. Therefore, the range of ρ_{ij} is between -1 and 1.

For standardizing weights, the correlation between stocks i and j should be based on normalized factors. When examining $q(t)$, their expectations are unequal, making it unsuitable for standardization across all i. However, for $r(t)$, the expectations can be shown to be zero (i.e., $E[r_i(t)] = \log(1) = 0$). Therefore, to provide a consistent standard, we use log-normalized data instead of regular return.

Question 2

The figure below depicts the histogram showing the un-normalized distribution of edge weights.

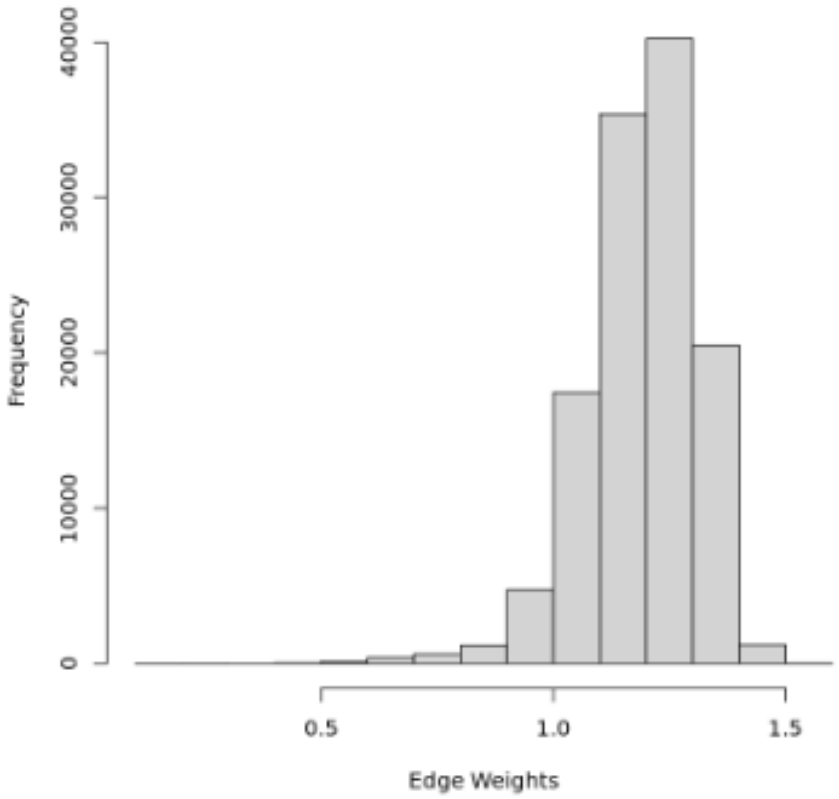

Figure 1: Histogram of Un-normalized Distribution of Edge Weights

Question 3

The figure depicts the MST extracted from the correlation graph, and each nodes are color-coded based on sectors.

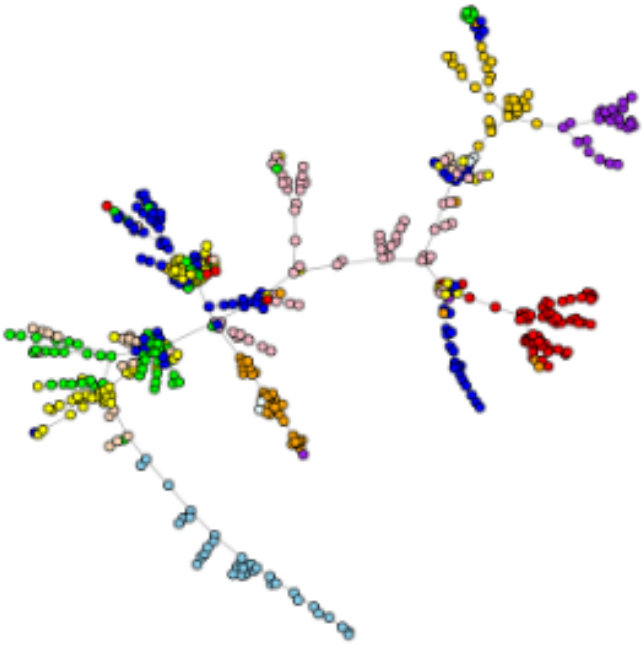

Figure 2: MST of correlation graph

From the figure above, we can observe several patterns. Stocks that are in the same sector, represented by the same color, usually group together in the MST. Conversely, stocks from different sectors, shown by different colors, are not connected. This means that stocks with high correlation are linked in the MST with the smallest possible edge weights. Conversely, stocks with low correlation have large edge weights.

Since the goal of the MST is to connect all nodes with the smallest total edge weights, it naturally groups highly correlated stocks. These groups form structures known as vine clusters, which look like grapes hanging from a main branch. Each cluster represents a different sector. Stocks within the same cluster tend to move in the same direction, implying similar investment strategies. Vine clusters give a long-term view of the stock market, using daily data.

Question 4

The figure below depicts the communities formed on the MST obtained from Question 3.

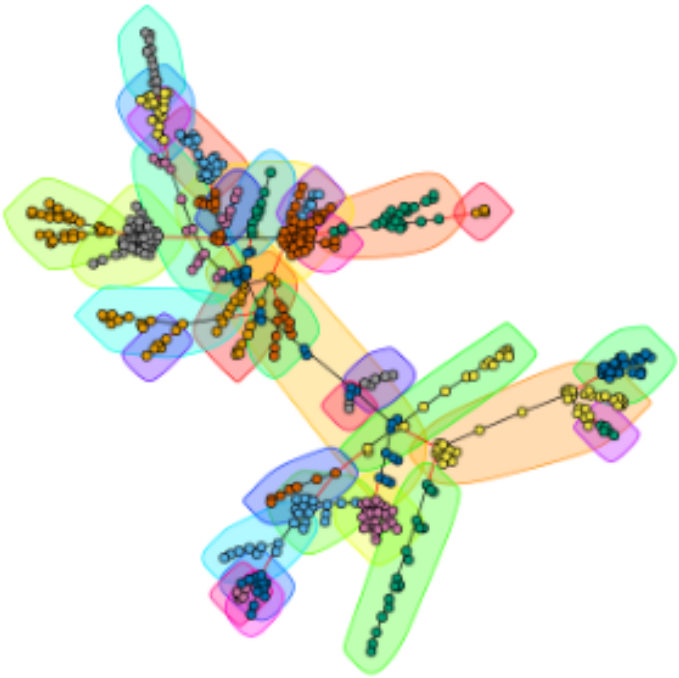

Figure 3: Communities Formed

The homogeneity score and completeness score were 0.682644 and 0.4792845 respectively.

Question 5

The value of α for case 1 was 0.828930 and case 2 was 0.11418. We notice that technique 1 yields a significantly higher value of α compared to technique 2. This outcome is expected because technique 1 utilizes the MST vine cluster structures of the correlation graph, focusing on highly correlated nodes that group together rather than all nodes. In other words, technique 1 leverages the local connectivity among neighboring nodes instead of the entire correlation graph to make decisions. On the contrary, technique 2 looks at all nodes within a sector and fails to capture local spatial connections or cluster formations because it considers the entire graph. Consequently, technique 2 only provides a broad probability estimate.

Question 6

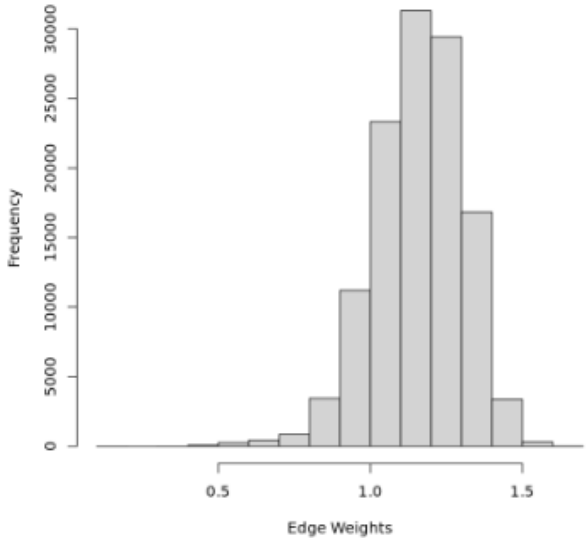

Figure 4: Histogram of Un-normalized Distribution of Edge Weights (Weekly Data)

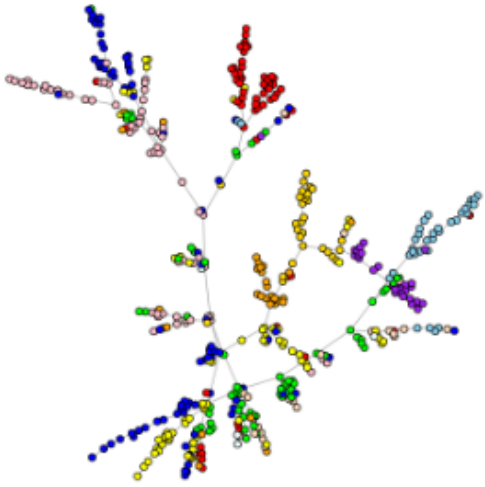

Figure 5: MST of correlation graph (Weekly Data)

Figure 6: Communities Formed (Weekly Data)

The homogeneity score and completeness score were 0.5811237 and 0.3900435 respectively.

The value of α for case 1 was 0.743957 and case 2 was 0.114309.

Question 7

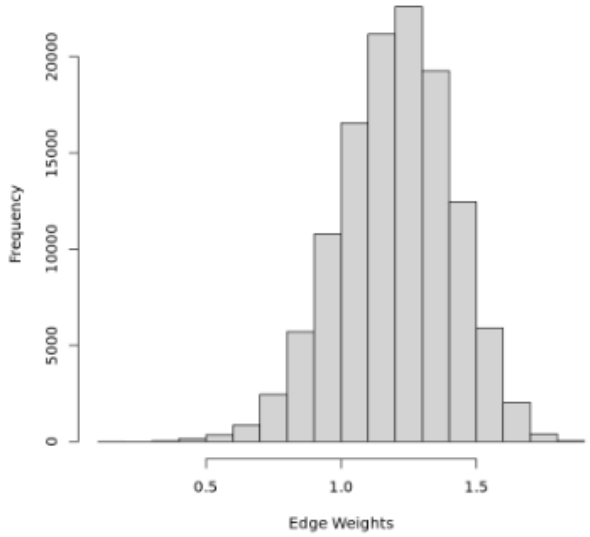

Figure 7: Histogram of Un-normalized Distribution of Edge Weights (Monthly Data)

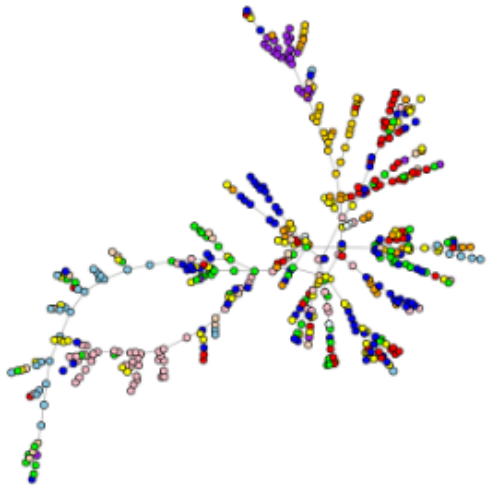

Figure 8: MST of correlation graph (Monthly Data)

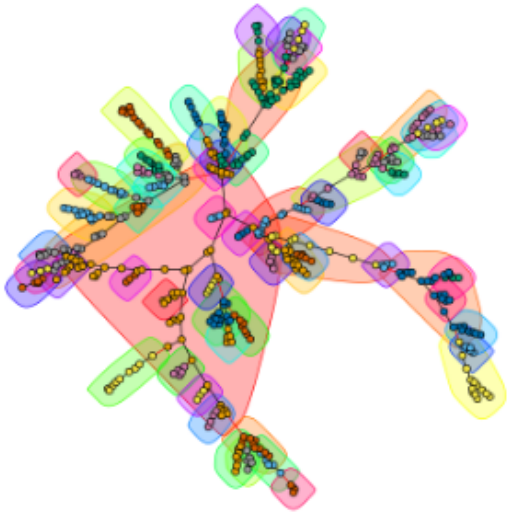

Figure 9: Communities Formed (Monthly Data)

The homogeneity score and completeness score were 0.4794473 and 0.2775512 respectively.

The value of α for case 1 was 0.484446 and case 2 was 0.114309.

Question 8

There are several differences in the results when performing Questions 2, 3, 4, and 5 with daily, weekly, and monthly data. Firstly, when observing the histogram of unnormalized distribution of edge weights, the daily data shows a left-skewed distribution, the weekly data shows a slightly left-skewed distribution, and the monthly data shows a bell-curved distribution. Secondly, for the MST, as we move from daily to weekly to monthly data, we observe that while some stocks still form vine clusters, a significant number of stocks from the same sector do not form clusters. The nodes do not form clearly separable regions in the MST graph, indicating that clustering is best with daily data and worst with monthly data. Similarly, for the communities formed, the communities are clearly defined with daily data, while the communities for weekly and monthly data are more disoriented. Additionally, the homogeneity score and completeness score decrease as we move from daily to weekly to monthly data. Lastly, the value of α decreases drastically from daily to monthly data; however, the α for case 2 does not decrease as much.

All these observations are possibly due to stocks from the same sector losing their correlation as the time scale increases, causing the edge weights in the correlation graph to increase among them even though they belong to the same node. Additionally, this means that it will be more difficult to assign a sector to an unknown stock if the data is sampled weekly or monthly.

Question 9

The number of nodes is 2649 and the number of edges is 1004955 in the graph G.

Question 10

The figure below depicts the Minimum Spanning Tree of the graph G.

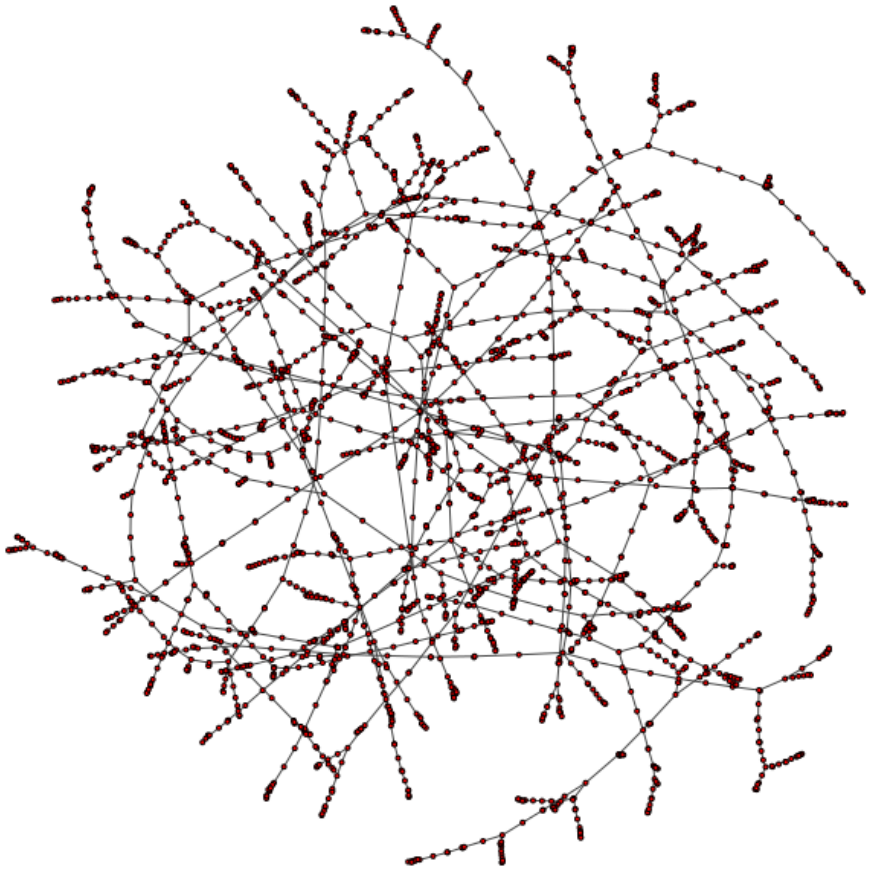

Figure 10: Minimum Spanning Tree of the graph G

The table below depicts the street address near the two endpoints of few edges.

ID	Census Tract (a)	Coordinates (a)	Street Address (a)	Census Tract (b)	Coordinates (b)	Street Address (b)
1	554001	[-118.133298 33.904119]	9421, Rosecrans Avenue, Bellflower, California, 90706, United States of America	554002	[-118.141448 33.896526]	9020, Somerset Boulevard, Bellflower, Los Angeles County, California, United States of America
2	461700	[-118.159325 34.153604]	500, Prospect Boulevard, Pasadena, California, 91103, United States of America	460800	[-118.172425 34.180286]	Brookside Golf Club, Arrovo Woods, Pasadena, CA 91109, United States of America
3	302201	[-118.251349 34.146334]	First United Methodist Church of Glendale, 134 North Kenwood Street, Glendale, CA 91206, United States of America	302202	[-118.248811 34.142665]	East Colorado Street, Glendale, CA 91205, United States of America
4	407101	[-117.967696 34.04101]	619 North Sunset Avenue, La Puente, CA 91744, United States of America	407002	[-117.980572 34.051745]	652 Puente Avenue, El Monte, CA 91746, United States of America
5	433401	[-118.043316 34.058606]	10468 Fern Street, South El Monte, CA 91733, United States of America	433402	[-118.038157 34.056957]	2433 Continental Avenue, El Monte, CA 91733, United States of America

Figure 11: Street Address Near the Two Endpoints for a Few Edges

The result makes sense because the goal of the Minimum Spanning Tree (MST) is to connect all addresses with the least total mean travel time. The simplest way to do this is by connecting nodes that are close together, as their travel times will be shorter. Prim's algorithm works by selecting the edges with the lowest weights (mean travel times), which means it connects nodes that are near each other. Our observations show that the average distance between these connected endpoints is about 0.7 miles.

Question 11

The triangle inequality states that in any triangle, the sum of any two sides must be greater than the third side:$a + b > c, \quad b + c > a, \quad a + c > b$. For our graph, this means that the direct mean travel time between two addresses is always shorter than the travel time when routing through a third address. By randomly sampling 1,000 triangles (each with 3 nodes), we discovered that 91.8% of them follow the triangle inequality.

Question 12

The upper bound on the empirical performance of the approximate algorithm, given by $\rho = \frac{\text{Approximate TSP Cost}}{\text{Optimal TSP Cost}}$, was 1.5663824728395292.

Question 13

The figure below depicts the trajectory the Santa has to travel:

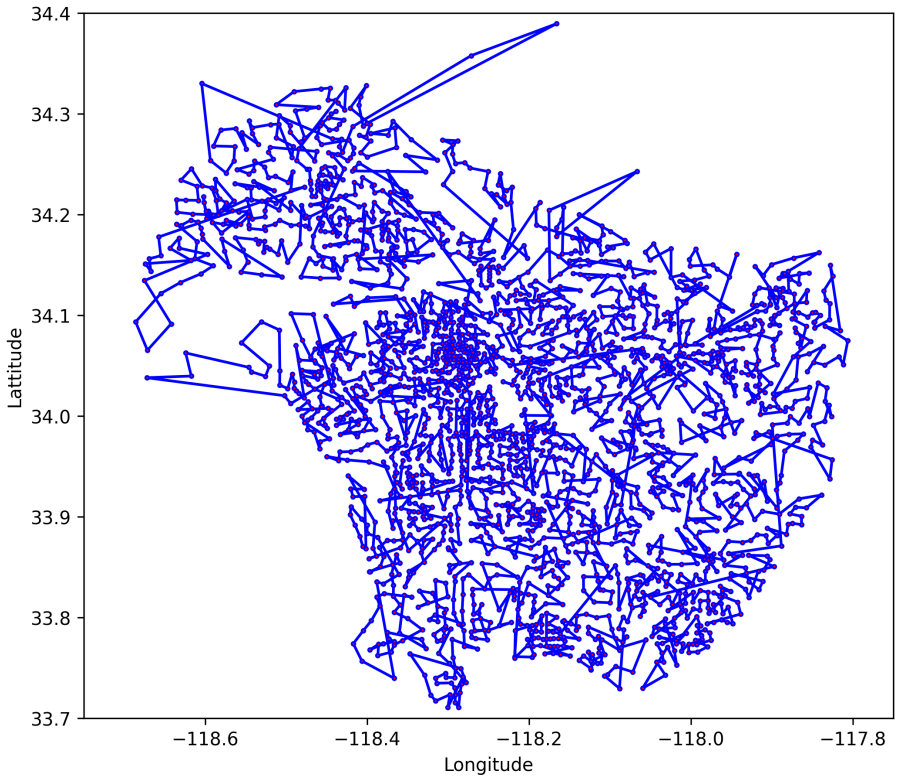

Figure 12: Trajectory Santa has to Travel

From the figure above, we see that the results make intuitive sense. Most of the edges do not overlap, but there are a few that do because the graph is not fully connected. As a result, we sometimes use Dijkstra's shortest path algorithm, which, unlike an Eulerian circuit, does not require each edge to be traversed only once. This leads to some edges overlapping, causing Santa to occasionally travel back to a previously visited node.

Question 14

The two figures below depicts road mesh of Los Angeles using Delaunay triangulation and graph G_Δ from the triangulation edges.

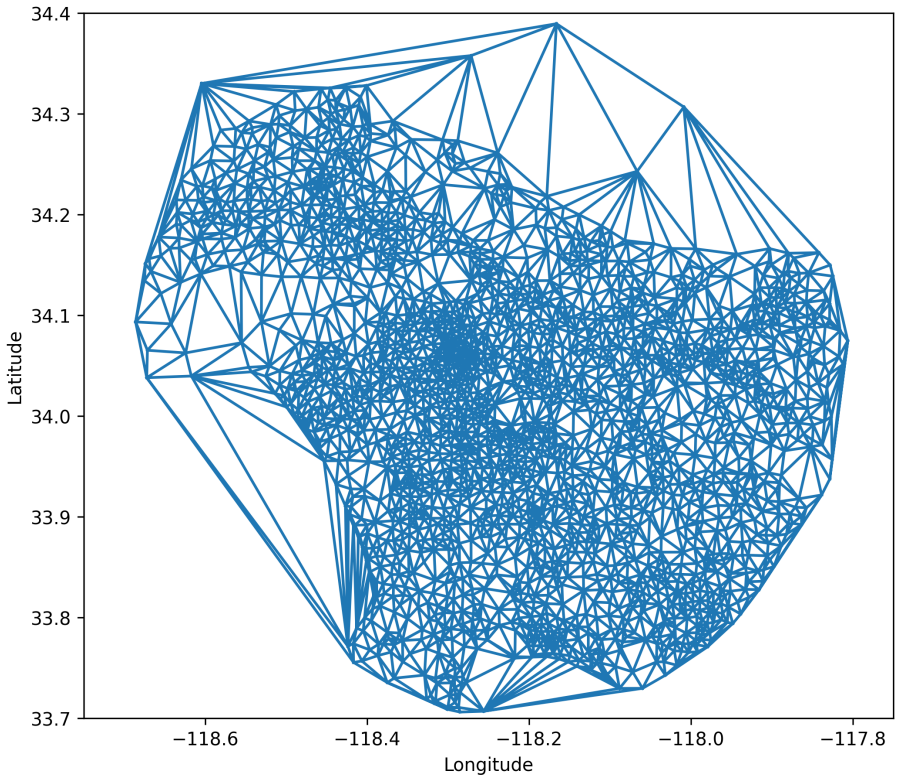

Figure 13: Road Mesh

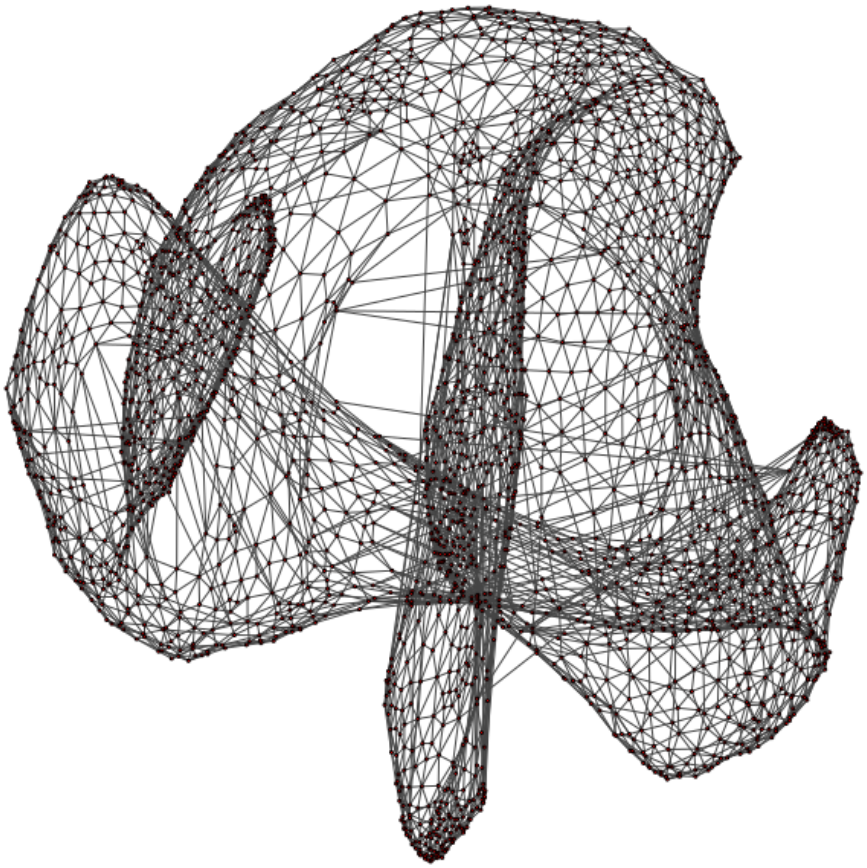

Figure 14: Graph G_Δ

Figure of the road mesh shows that the triangulation algorithm has largely captured the road network of Los Angeles, especially in the downtown area. However, it also generates some roads over oceans and mountains where there are no actual roads. This occurs because the Delaunay triangulation (DT) algorithm aims to avoid creating sliver or very narrow triangles by ensuring that no point lies within the circumcircle of any triangle. In other words, the DT algorithm fits each triangle inside a circumcircle. Figure 10 illustrates this concept, with nodes representing locations and edges resulting from triangulation, showing that most polygons lack extremely small acute angles.

Question 15

The derivation for traffic flow is as follows: the total distance is given by $\frac{\text{Velocity of car} \times \text{Mean Travel Time}}{60 \times 60}$, which converts data from seconds to hours. The gap between cars is $0.003 + \frac{2 \times \text{Velocity of car}}{60 \times 60}$ to account for the length of each car and the distance between cars. Thus, the total number of cars on the road is $\frac{2 \times \text{Total distance}}{\text{Gap}}$ to accommodate flow in both directions. Traffic Flow (cars/hour) is then calculated as $\frac{60 \times 60}{\text{Mean Travel Time}} \times$ Total number of cars on the road. Simplifying the derivation, we get: Traffic Flow (cars/hour) $= \frac{3600 \times \text{Velocity of car}}{5.4 + \text{Velocity of car}}$. The unit of velocity is in miles per hour. Using Pythagoras' Theorem, the velocity of a car between two coordinates is given by: Velocity of car $= \frac{69}{\text{Mean Travel Time}} \times \sqrt{(\text{Latitude}_{\text{point 1}} - \text{Latitude}_{\text{point 2}})^2 + (\text{Longitude}_{\text{point 2}} - \text{Longitude}_{\text{point 2}})^2}$.

Question 16

The number of edge-disjoint path is 5 and the maximum number of cars that can commute per hour from Malibu to Long Beach is 13095 cars/hour.

The two graph below depicts the road maps near Malibu and Long Beach.

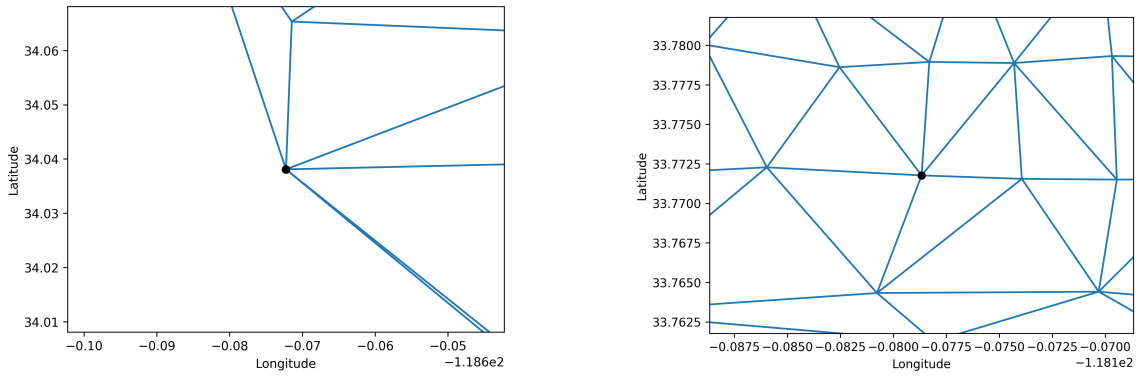

Figure 15: Road Maps near Malibu and Long Beach

Figure two plots above we see that both Malibu and Long Beach have 6 outgoing and incoming edges, respectively. Additionally, we can see that each node has the same degree. In a graph, the minimum number of outgoing and incoming edges for two nodes represents the number of edge-disjoint paths between them, which in this case is 6.

Question 17

The two graph below depict the road mesh and graph G_Δ after threshold method applied.

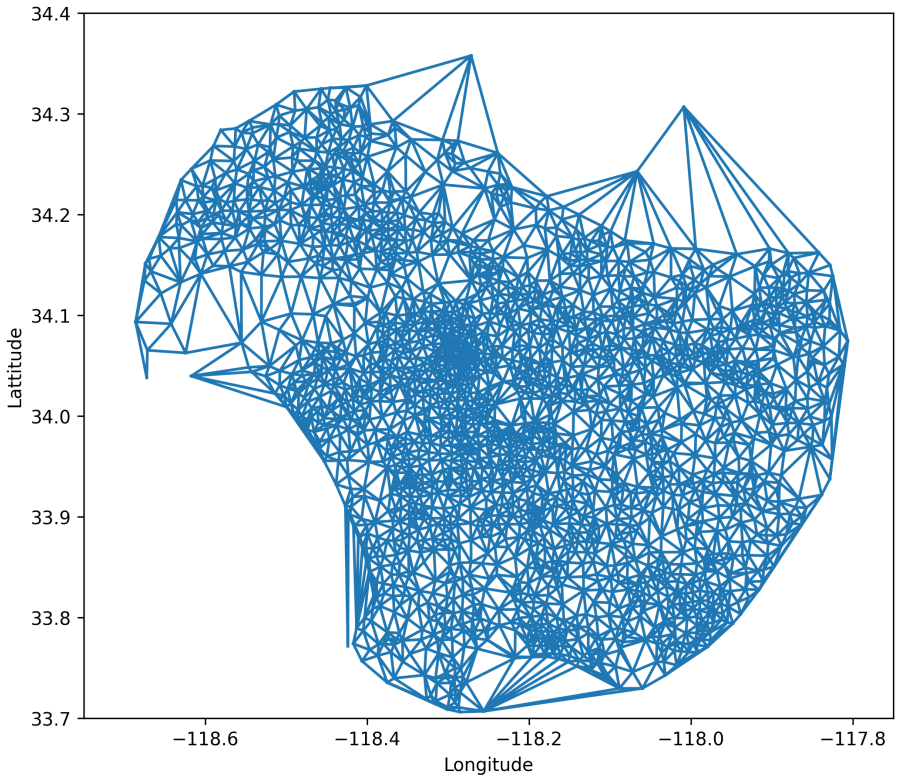

Figure 16: Road Mesh

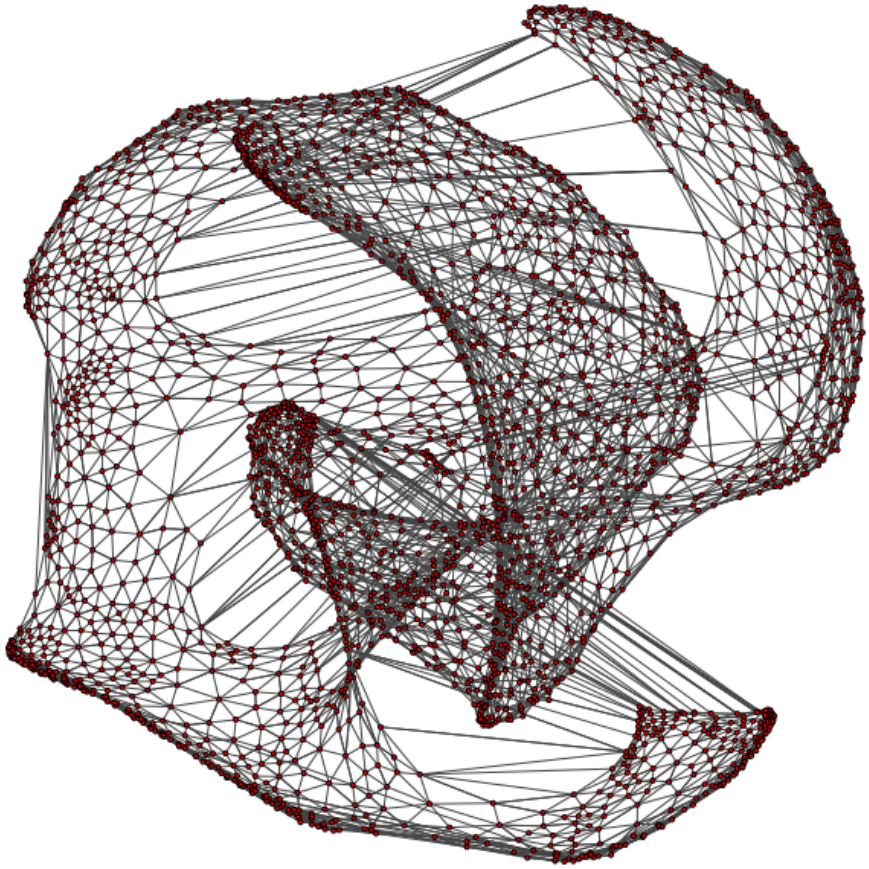

Figure 17: Graph G_{Δ}

From the road mesh, we can clearly observe changes before and after applying the thresholding method. After thresholding, some of the routes have disappeared, particularly those at the top and bottom left corner of the road mesh. This change indicates that the thresholding method was effective.

Question 18

The figures below depict road map near Malibu and Long Beach after thresholding.

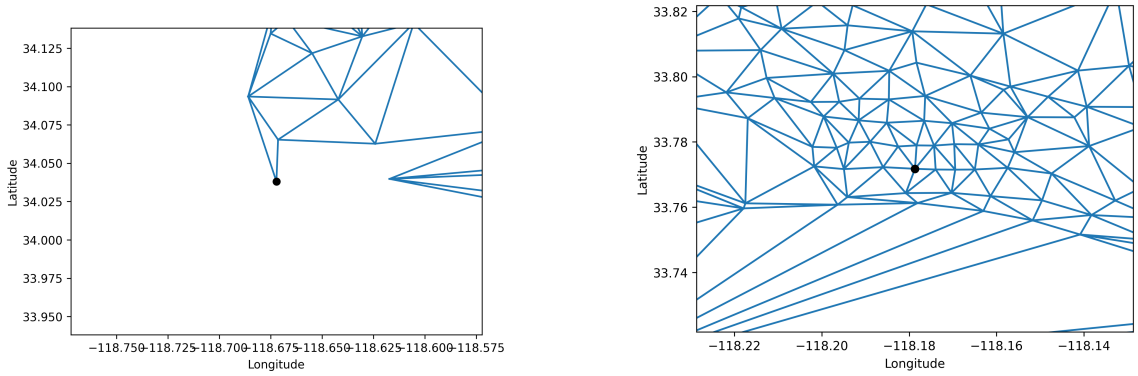

Figure 18: Road Maps near Malibu and Long Beach

From the figure above, we can see that there is no significant change to the road map near Long Beach. However, several paths in Malibu have been removed.