ECE232 Project
Graph Algorithms

Juo-Wen Fu 805025223
Alex Hung 705035114
Te-Yuan Liu 805027255
Wesly Tzuo 704946416

June 22, 2018

1 Stock Market

1.1 Return correlation

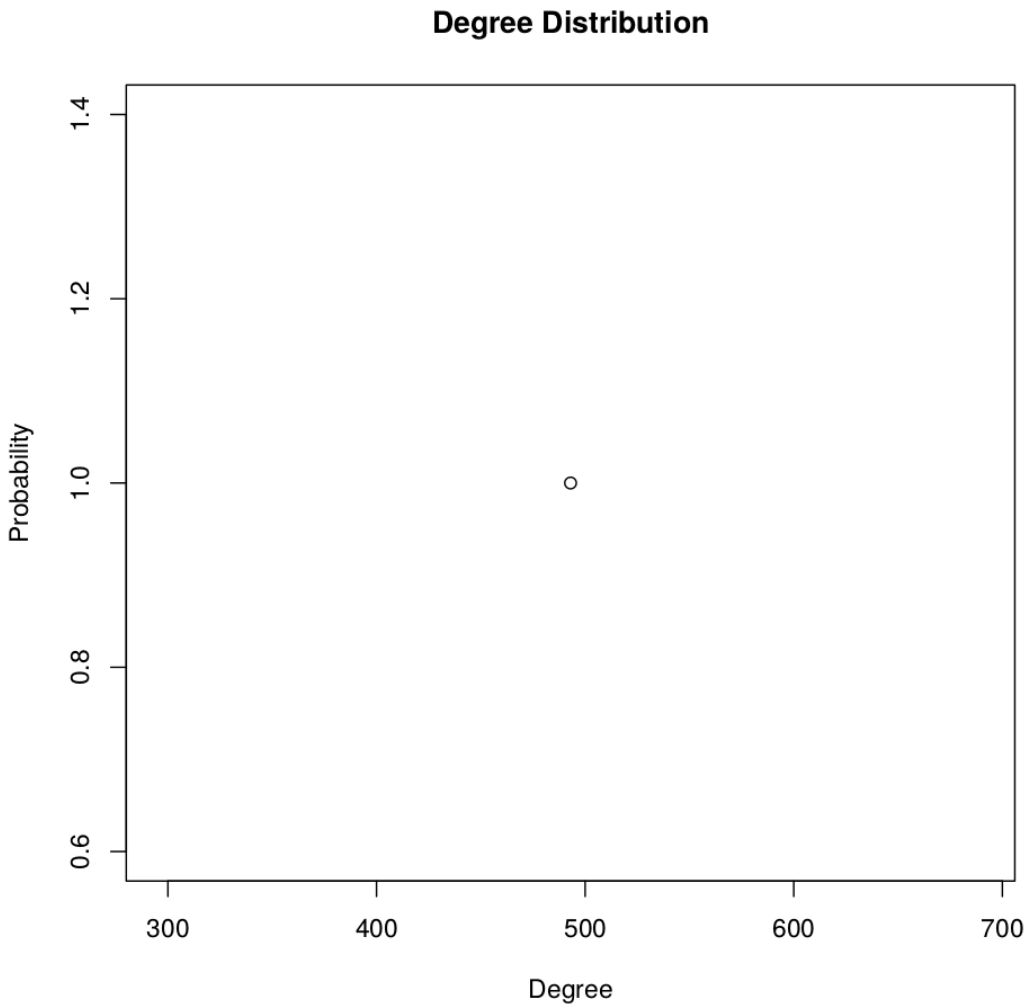
Question 1: Provide an upper and lower bound on ρ_{ij}. Also, provide a justification for using log-normalized return $(r_i(t))$ instead of regular return $(q_i(t))$.

The upper bound of ρ is 1 and the lower bound is -1. Log-normalized return is used instead of regular return because the data is lognormally distributed. In other words, the log of the data is normally distributed. This leads us to the idea of taking the log of the data and thus restore the symmetry to it.

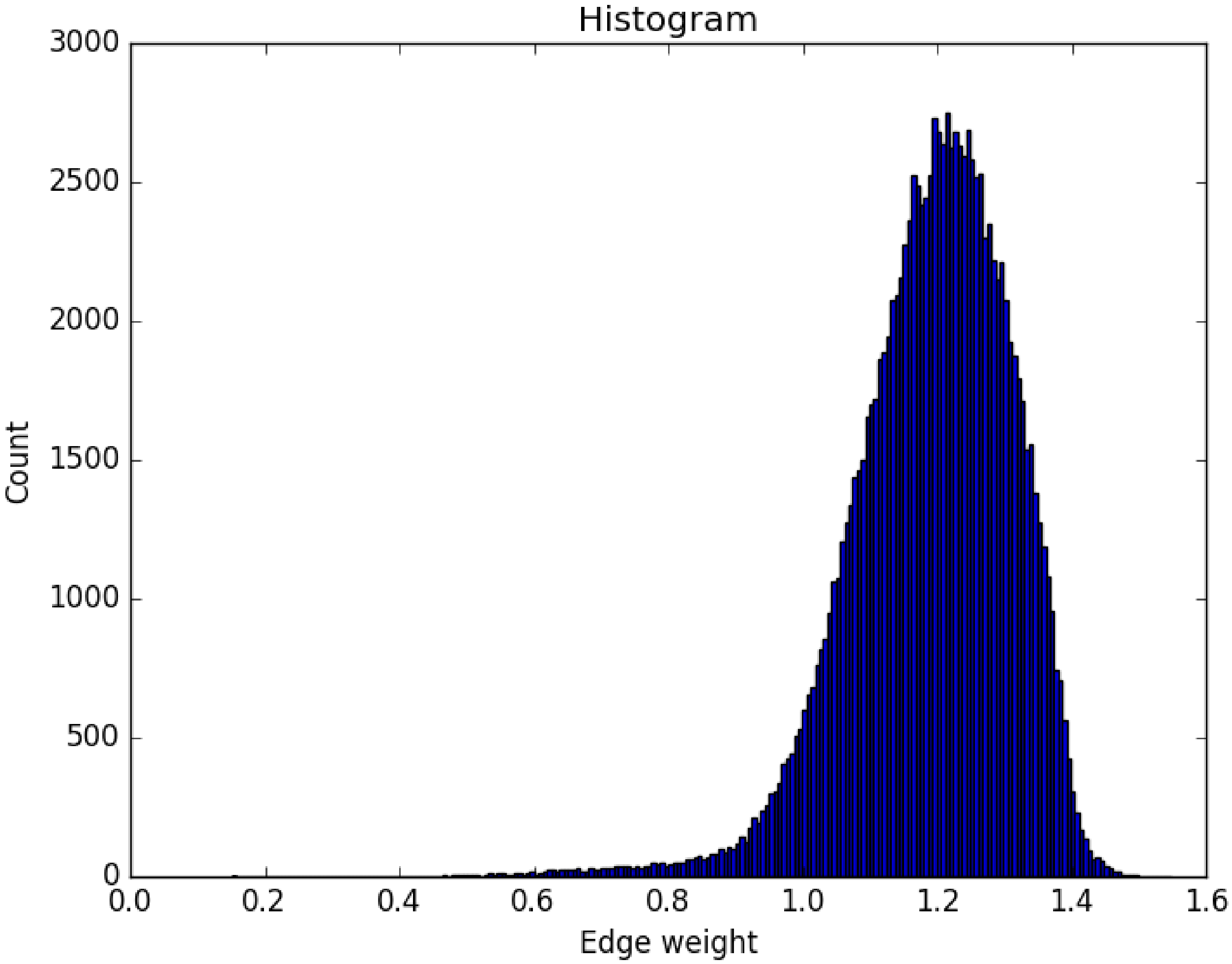
1.2 Constructing correlation graphs

Question 2: Plot the degree distribution of the correlation graph and a histogram showing the un-normalized distribution of edge weights.

The degree distribution graph is shown below.



From the above graph we know that every node has a degree of 494.
The edge weight histogram is shown below.

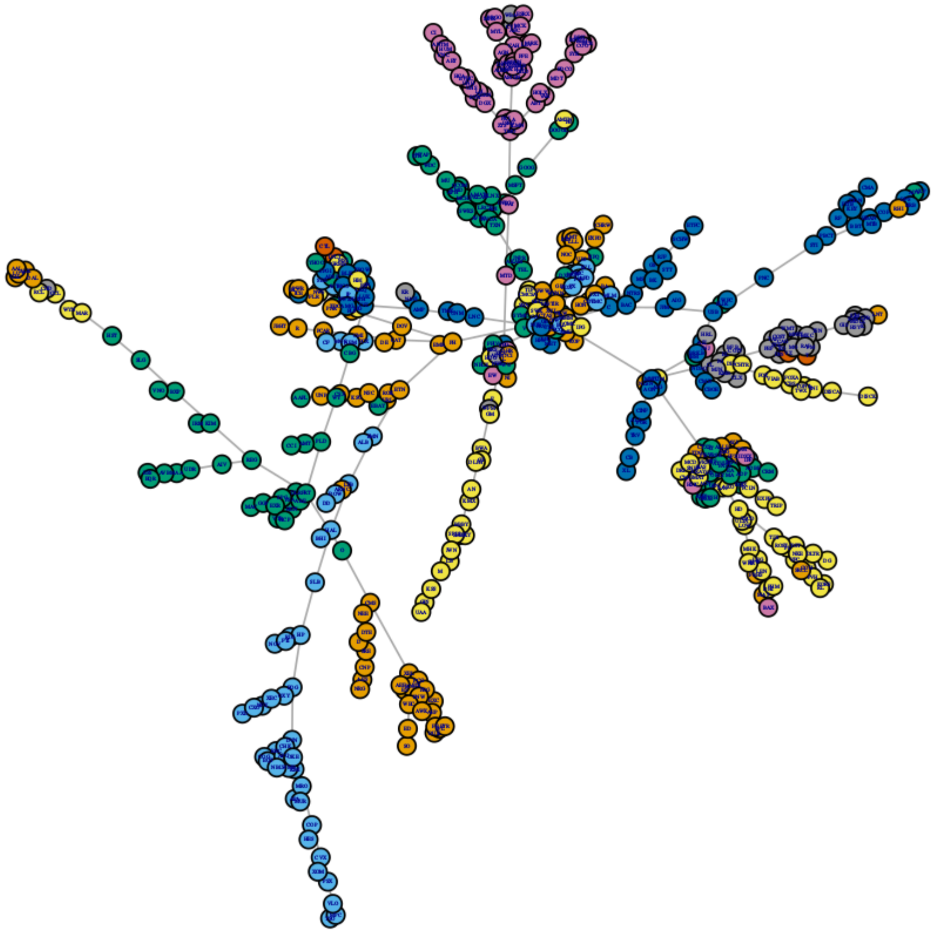


From the above graph we know that the edge weight histogram shows a normal distribution, which corresponds to the idea we mentioned in the previous question.

1.3 Minimum spanning tree (MST)

Question 3: Extract the MST of the correlation graph. Each stock can be categorized into a sector, which can be found in Name_sector.csv file. Plot the MST and color-code the nodes based on sectors. Do you see any pattern in the MST? The structures that you find in MST are called Vine clusters. Provide a detailed explanation about the pattern you observe.

The MST plot is shown below.



The Vine clusters pattern shown in the above graph reveals the fact that stocks in the same sector tend to have high correlation in price fluctuation and thus low edge weights between one and another. This observation verifies the idea that investors will have similar strategies of investment for stocks that are effected by the same economic factors, and thus stocks of closely connected industries will have similar price fluctuation and high correlation.

1.4 Sector clustering in MST's

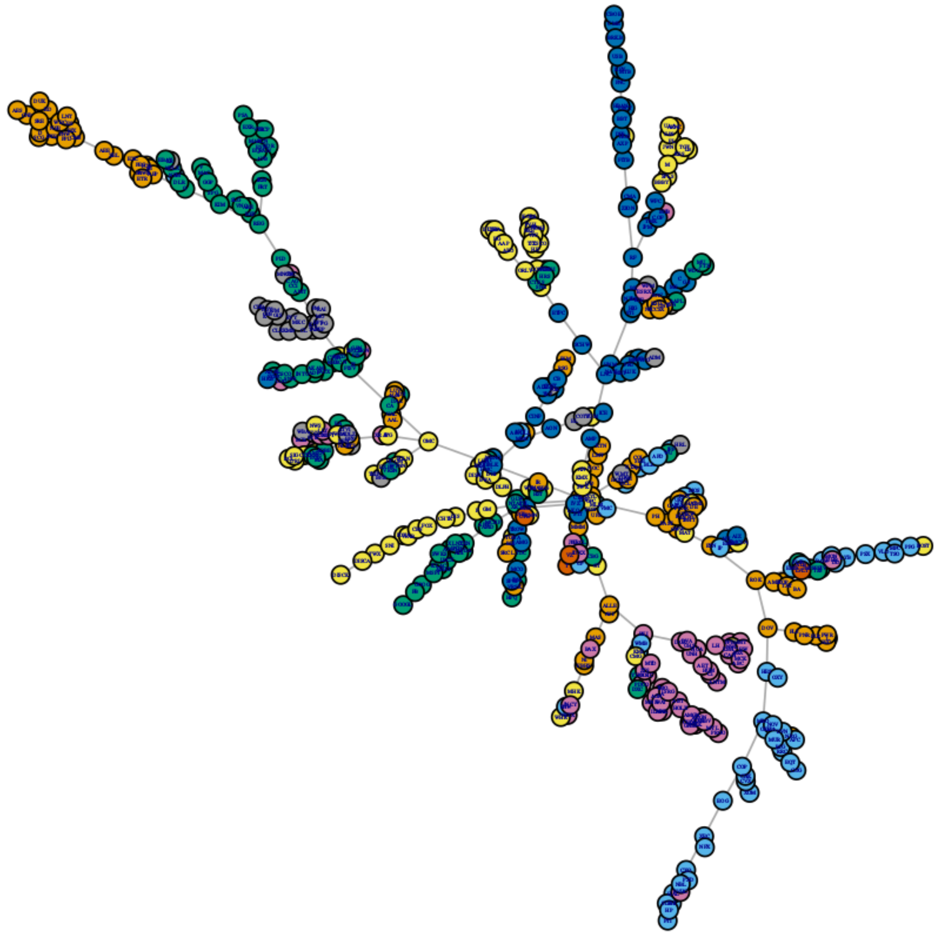
Question 4: Report the value of α for the above two cases and provide an interpretation for the difference.

The metric α using neighborhood structure ($\mathrm{P}(v_i \in S_i) = \mid Q_i \mid / \mid N_i \mid$) is 0.8289 and the other metric α using entire sector structure ($\mathrm{P}(v_i \in S_i) = \mid S_i \mid / \mid V \mid$) is 0.1142. The neighborhood based metric looks like the homogeneity measurement for clustering and it gives us the idea of how MST performs on finding the sector structures of the stocks. A high score on the neighborhood based metric indicates that most of the stocks grouped together are in the same sector and this is what we expect to see for a good clustering performance. The other metric seems to find the skew effect of the sector distribution and does not utilize the result of the MST. If the sectors distributed evenly, in other words, each sector has the same amount of stocks, then the metric score is low. However, if the sector distribution is skewed and one sector possesses most of the stocks, then the metric score is high.

1.5 Correlation graphs for weekly data

Question 5: Extract the MST from the correlation graph based on weekly data. Compare the pattern of this MST with the pattern of the MST found in question 3.

The MST plot is shown below.



The neighborhood based metric α for this network is 0.743, which is lower than the one using daily data. This is reasonable because as we sampled weekly data from daily data, we lost some correlation information. From the graph, we observe that the sector structure is more messy than the one using daily data.

2 Lets Help Santa!

Question 6: Report the number of nodes and edges in G.

After merging the duplicate edges by averaging their weights and reserving the giant connected component of the graph, we can get the following number of edges and nodes:

$$Edges : 311802$$
$$Nodes : 1880$$

Question 7: Build a minimum spanning tree (MST) of graph G. Report the street addresses of the two endpoints of a few edges. Are the results intuitive?

After building the MST of graph G, we select the first 5 edges and report their street addresses of the two endpoints in the following table:

E1	3300 Brodie Drive, South San Jose, San Jose	132.59
	4300 La Torre Avenue, South San Jose, San Jose	
E2	3300 Brodie Drive, South San Jose, San Jose	126.24
	3700 McLaughlin Avenue, South San Jose, San Jose	
E3	3300 Brodie Drive, South San Jose, San Jose	109.62
	400 Ginkgo Court, South San Jose, San Jose	
E4	1700 Coyote Point Drive, Shoreview, San Mateo	80.985
	1800 Helene Court, East San Mateo, San Mateo	
E5	1700 Coyote Point Drive, Shoreview, San Mateo	111.88
	600 Lexington Way, Oak Grove Manor, Burlingame	

As we can see from the table, the location of two end points of these edges are close to each other thus the weight of these edges are very small. This matches the intuition that the sum of edge weights in MST needs to be as small as possible.

Question 8: Determine what percentage of triangles in the graph (sets of 3 points on the map) satisfy the triangle inequality. You do not need to inspect all triangles, you can just estimate by random sampling of 1000 triangles.

In this part we randomly pick one node from the GCC and two of its neighbors. If they form an triangle, we will further check whether they satisfy the triangle inequality. After repeating this process 1000 times, we get 909 times that those points match triangle inequality.

Percentage of triangles that satisfy the triangle inequality: 91%

Question 9: Find the empirical performance of the approximate algorithm:

In general, the Approximate TSP Cost would be less or equals to two times of Optimal TSP Cost (Approximate TSP Cost <= 2×Optimal TSP Cost, $\rho <= 2$) if the graph satisfy the triangle inequality. However, only 91% of triangles satisfy the triangle inequality so the upper bound of ρ will not equal to 2.

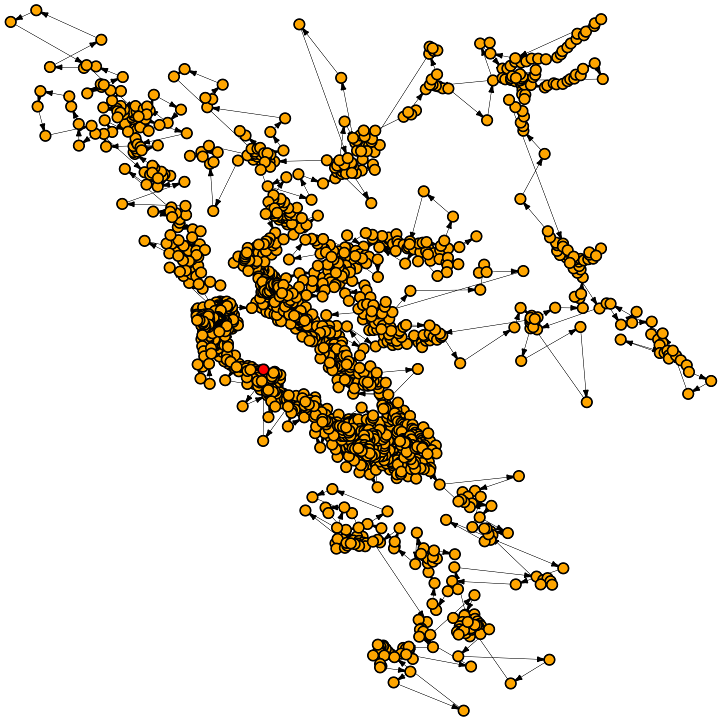
After using DFS which returns pre-order list of nodes, we can get order of nodes. However, it is not guaranteed that there is an edge between each nodes, thus sometime we need to estimate the edge weight by their distance. Here we sample 200 nodes and compute the ratio of their distance and the mean travel time. Finally we can get the Approximate TSP Cost:

$$\text{Approximate TSP Cost} = 428671.3$$

So we can estimate that the Optimal TSP Cost would be less than 210000.

Question 10: Plot the trajectory that Santa has to travel!

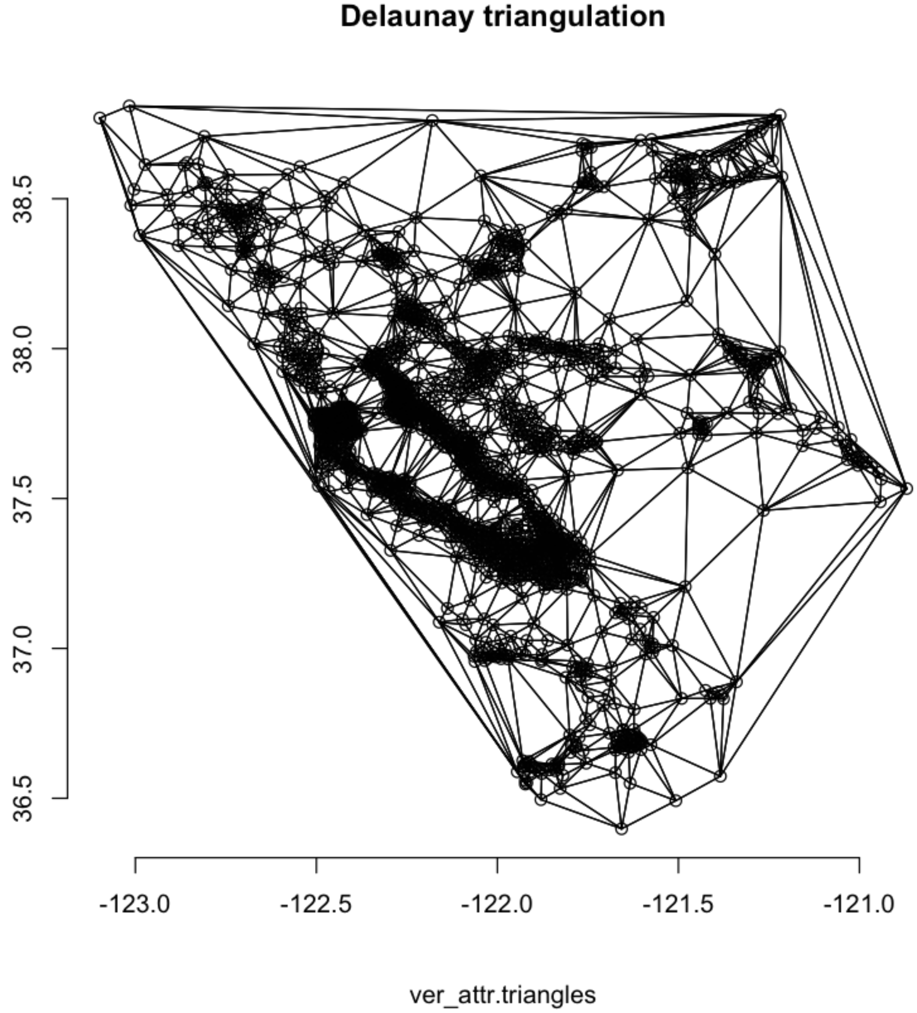
By adding edges between the nodes we get from Question 9, we can get the trajectory that Santa has to travel! Here we plot the nodes base on their coordinate and as you can see from the figure that the algorithm provides a reasonable trajectory.



3 Analyzing the Traffic Flow

Question 11: Plot the road mesh that you obtain and explain the result. Create a subgraph G induced by the edges produced by triangulation.

Delaunay triangulation is a max-min angle triangulation. It is a proximate method that satisfies the requirement that a circle drawn through the three nodes of a triangle will contain no other node. Since the triangles are created as equi-angular as possible, it is less likely to produce long skinny triangles. In this graph, we can see that most of the points are in the middle.



Question 12: Using simple math, calculate the traffic flow for each road in terms of cars/hour.

First, we calculate the distance between each pair of points. Next we calculate speed where $speed = \frac{distance}{time}$. Then we calculate the time for a car passes the end point plus the safety distance of 2 seconds to the next car, which is $\frac{0.003+2\times speed}{speed}$. An hour equals to 3600 seconds, thus $\frac{3600}{\frac{0.003+2\times speed}{speed}}$ is the traffic flow of a road. Since each road has 2 lanes, we multiply it by 2. The following equation is how we calculate the traffic flow for each road.

$$\text{traffic flow} = 2 \times \frac{3600}{\frac{0.003+2\times speed}{speed}} (\text{cars/hour})$$

Question 13: Calculate the maximum number of cars that can commute per hour from Stanford to UCSC. Also calculate the number of edge-disjoint paths between the two spots. Does the number of edge-disjoint paths match what you see on your road map?

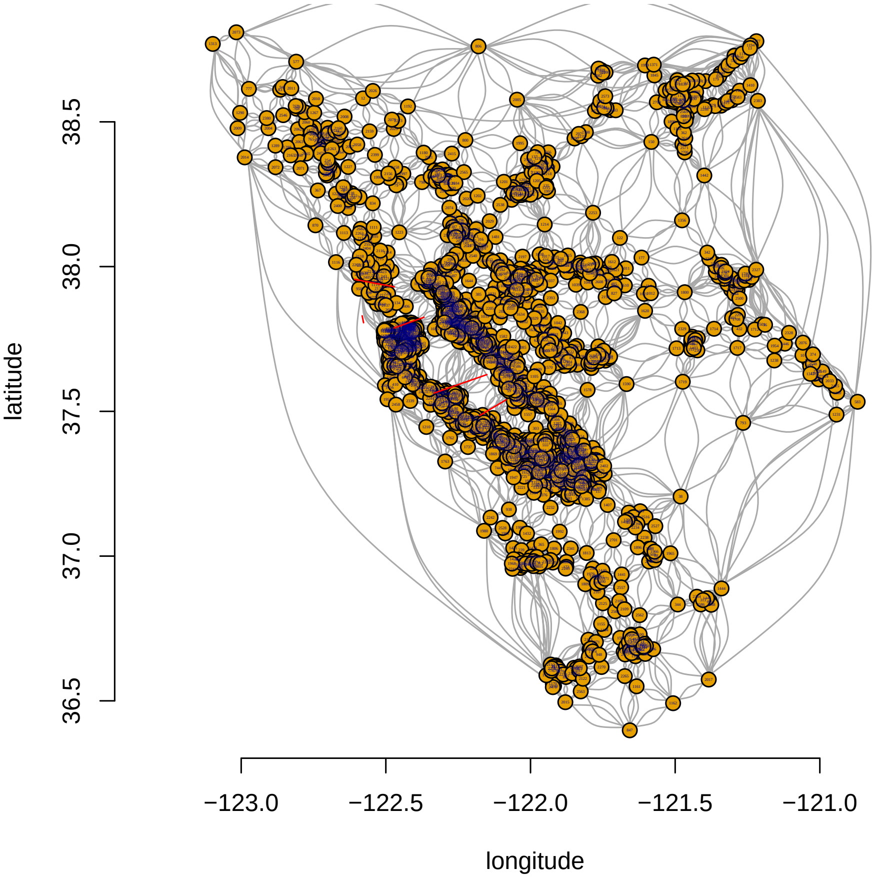
$$\text{max flow} = 15014.96$$

$$\text{number of edge-disjoint paths} = 5$$

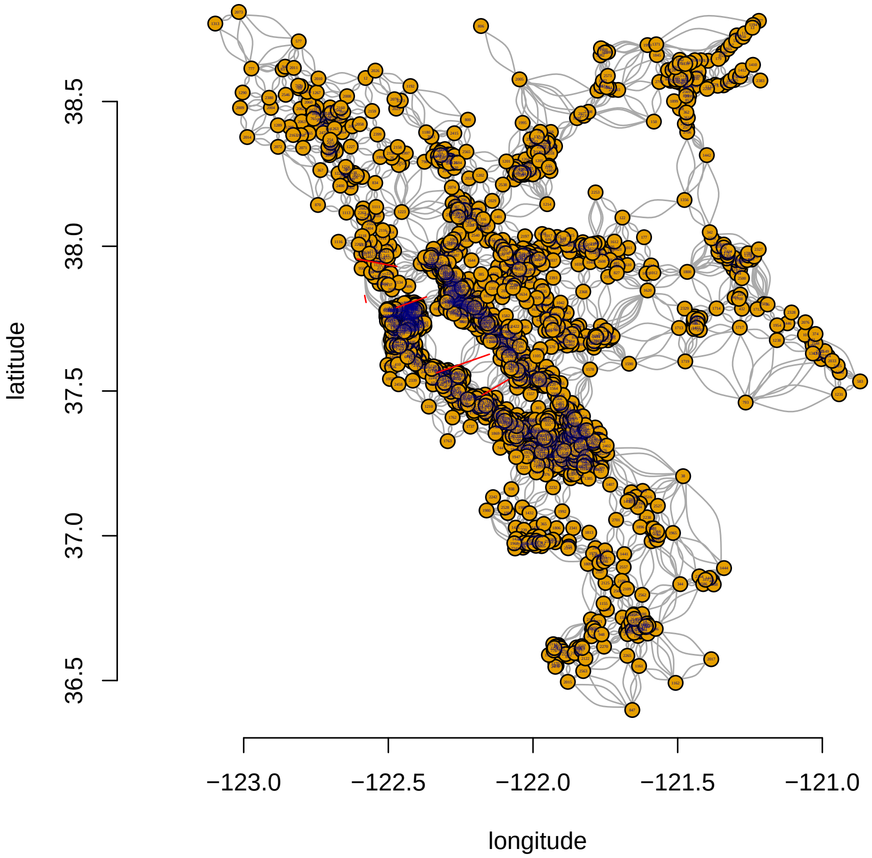
Yes. Since the number of out-degree of Stanford is 6 and the number of in-degree of UCSC is 5, the maximum number of edge-disjoint path we can get is 5.

Question 14: Plot $\tilde{G}_\Delta$ on real map coordinates. Are real bridges preserved?

For comparison purpose, we first plot G_Δ on real map coordinates. We use the average coordinates for each node. We also plot the five bridges in red color. Because the coordinates we use for each node are the average coordinates, they may not coincide exactly with the exact bridge coordinates.



It can be seen from the above plot of G_Δ that there are many unreal roads connecting nodes that are very far apart.
These unreal roads tends to have long travel time. We apply a maximum threshold of 1500 on the travel time of the roads to trim the fake edges and call the graph $\tilde{G}_\Delta$ and plotted in real map coordinates are shown below.



As we can see from the plot above, roads around the bridges are preserved. This is expected because these real bridges, as real roads, have short travel time and hence were not trimmed.
We also experiment with trimming the fake edges by applying a threshold on the speed of the vehicle on each edge. However, in this case many real edges were trim before the fake edges.

Question 15: Repeat question 13 for $\tilde{G}_\Delta$. Do you see any significant changes?

Using the new graph $\tilde{G}_\Delta$ we get

$$\text{max flow} = 15014.96$$

$$\text{number of edge-disjoint paths} = 5$$

We did not see any changes for the max flow. This is because the edges we trim have are outliers with very long travel time which do no contribute to the max flow calculation.