EE 232E Project 5
Graph Algorithms

Hengjie Yang, Sheng Chang, Wandi Cui, and Tianyi Liu

June 12, 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 $q_i(t)$.

According to the formula of ρ_{ij}, the upper bound should be 1 and the lower bound should be -1, where $r_i(t) = r_j(t)$ and $r_i(t) = -r_j(t)$ respectively.

Considering the reason why we use log-normalized return $r_i(t)$ instead of regular return $q_i(t)$, we have following analysis and explanations:

(a)log-normality: if we assume that prices are distributed log normally (which, in practice, depends on the given price series), then $r_i(t)$ is conveniently normally distributed, because:

$$r_i(t) = \log(1 + q_i(t)) = \log \frac{p_i(t)}{p_i(t-1)}$$

This is handy given much of classic statistics presumes normality.

(b)Approximate raw-log equality: when returns are very small (common for trades with short holding durations), the following approximation ensures they are close in value to raw returns:

$$r_i(t) = \log(1 + q_i(t)) \approx q_i(t), where q_i(t) \ll 1$$

(c)time-additivity: Consider a stock i with an ordered sequence of n trades. A statistic frequently calculated from this sequence is the compounding return, which is the running return of this sequence of trades over time:

$$q_i(1) \cdot q_i(2) \cdot q_i(3) \cdot ... \cdot q_i(n)$$

This formula seems really complicated since the product of normally-distributed variables is not normal. Instead, the sum of normally-distributed variables is normal. This can show another benefit of representing returns in log-normality:

$$r_i(1) + r_i(2) + r_i(3) + ... + r_i(n) = \log p_i(n) - \log p_i(0)$$

Thus, the compound return over n periods is merely the difference in *log* between initial and final periods. In terms of algorithmic complexity, this simplification reduces $O(n)$ multiplications to $O(1)$ additions.

(d)For Mathematical Ease: from calculus, we are reminded (ignoring the constant of integration):

$$e^x = \int e^x dx = \frac{d}{dx} e^x = e^x$$

This identity is tremendously useful, as much of financial mathematics is built on continuous time stochastic processes which rely heavily on integration and differentiation.

(e)Numerical stability: Sometimes the addition of small numbers is numerically safe, while multiplying small numbers is not as it is subject to arithmetic underflow.

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.

After constructing the correlation graph, the degree distribution we get is shown in Figure1 Not surprisingly, all vertices are connected with each other, so the degree of them are the same, equals to the number of vertices minus 1.

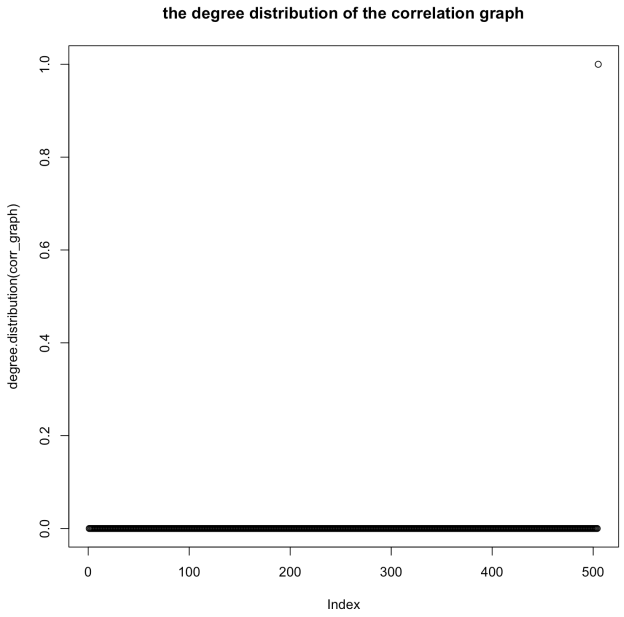

Figure 1: degree distribution of the correlation graph

The edge weights distribution is shown in Figure2

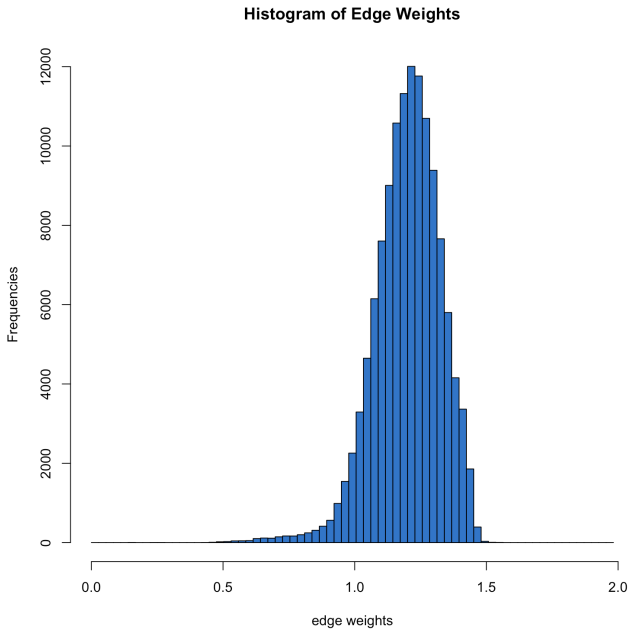

Figure 2: edge weights distribution of the correlation graph

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 in Figure3, and differnt color represents different sectors of the stock.

We found that there is an interesting pattern in MST: stocks in same sector tend to be clustered together, and the whole MST looks like a grape-shaped graph. In fact, this is understandable, since the weight of the edge represents the correlation of the stock. It is not difficult to imagine that stocks are highly correlated so the weight between them is likely to be comparablly smaller than the weight between stocks in different sectors. Since MST find the path containing all the vertices in the graph with minimum total cost, it is reasonable to traverse nearly all the vertices in one sector and than transfer to another. That is why the pattern of MST looks like several "sector clusters" hanging in the tree, which is called "vine clusters" structure.

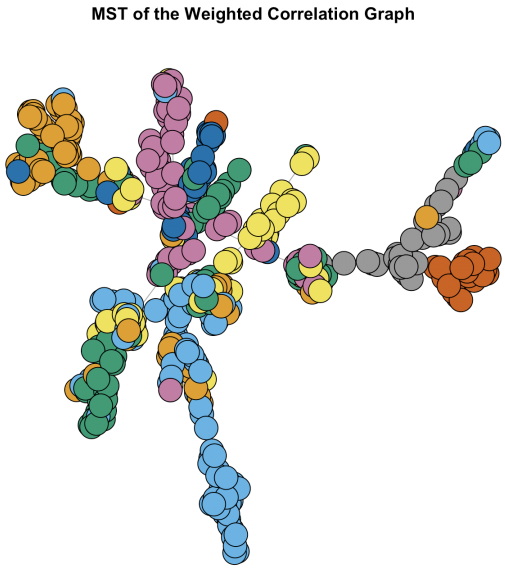

Figure 3: MST of the correlation graph

1.4 Sector Clustering in MST's

1.5 Correlation Graph for Weekly Data

2 Let's Help Santa!

2.1 Download the Data

2.2 Build Your Graph

2.3 Traveling Salesman Problem

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.

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

$$\rho = \frac{\text{Approximate TSP Cost}}{\text{Optimal TSP Cost}} \tag{1}$$

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

3 Analysing the Traffic Flow

3.1 Estimate the Roads

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

3.2 Calculate Road Traffic Flows

3.3 Calculate the Max Flow

3.4 Defoliate Your Graph