**ECE 232E - Project 5**

**Graph Algorithms**

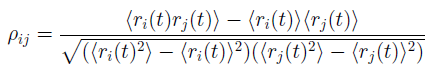
*Zhechen Xu (805030074), Yufei Hu (404944367), Qi Zeng (204946904), Liangkun Zhao (204947003)*

*June 15, 2018*

# **Part 1: Stock Market**

**Question 1:**

According to the definition:



Considering the best and worst situation, the upper bound of it is **1**, while the lower bound of it is **-1**.

After analyzing the formulations and investigating some references, there are five reasons we should use log-normalized return instead of regular one:

**A: Time-additivity**

**B: Log-normality:**

if we *assume* that prices are distributed [log normally](http://en.wikipedia.org/wiki/Log-normal_distribution) (which, in practice, may or may not be true for any given price series), then log(1 + r_i) is conveniently [normally distributed](http://en.wikipedia.org/wiki/Normal_distribution), because:

1 + r_i = \frac{p_i}{p_j} = \exp^{\log(\frac{p_i}{p_j})} 

**C: 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:

\log(1 + r) \approx r , r \ll 1 

**D: Mathematical ease:**

from [calculus](http://en.wikipedia.org/wiki/Calculus), we are reminded (ignoring the [constant of integration](http://en.wikipedia.org/wiki/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 upon [continuous time stochastic processes](http://en.wikipedia.org/wiki/Continuous-time_stochastic_process) which rely heavily upon integration and differentiation.

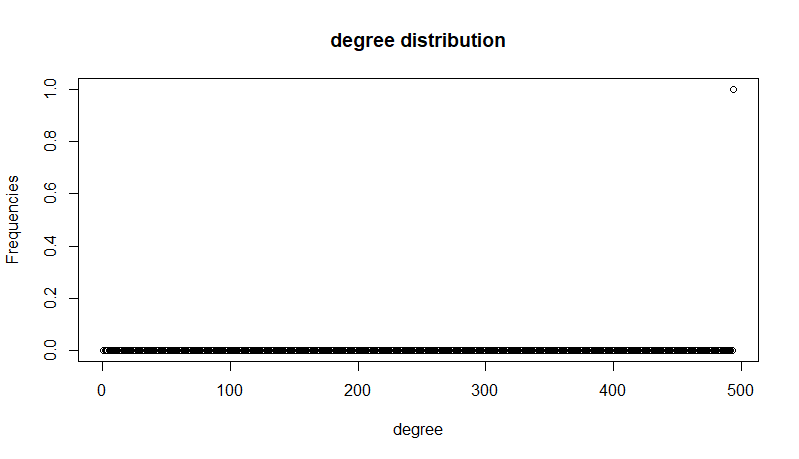
**E: Numerical stability:**

Addition of small numbers is numerically safe, while multiplying small numbers is not as it is subject to [arithmetic underflow](http://en.wikipedia.org/wiki/Arithmetic_underflow). For many interesting problems, this is a serious potential problem. To solve this, either the algorithm must be modified to be numerically robust or it can be transformed into a numerically safe summation via logs.

Reference: https://quantivity.wordpress.com/2011/02/21/why-log-returns/

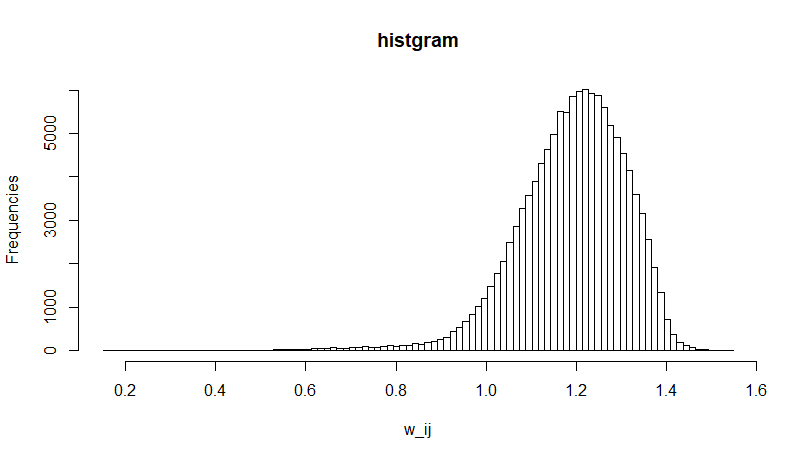
**Question 2:**

Degree distribution of the correlation graph:



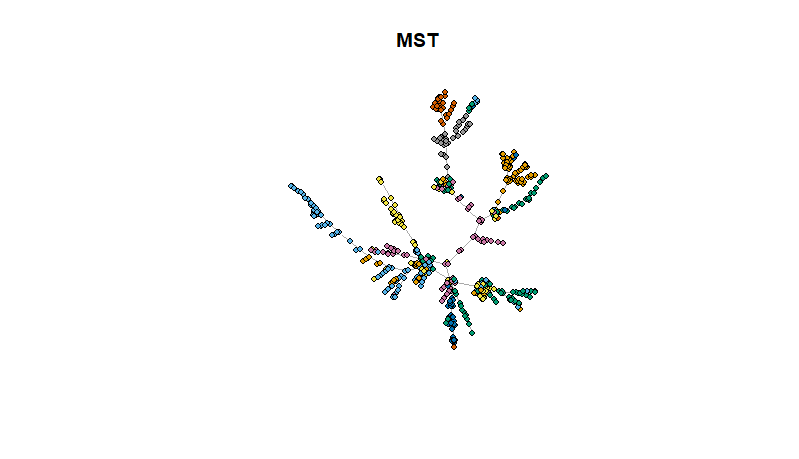
It’s a one bar plot.

The un-normalized distribution of edge weights:



**Question 3:**

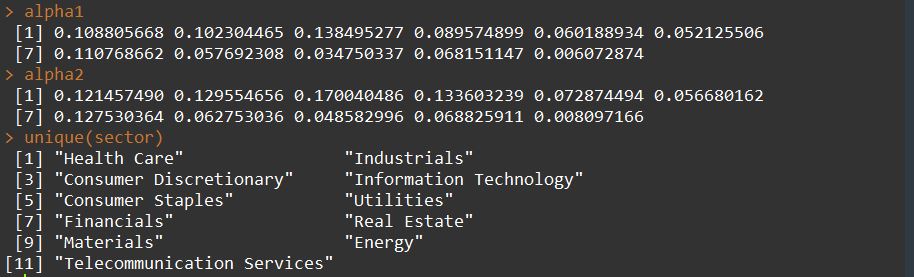
The MST and color-coded nodes:



As we can see in the plot, those nodes with the same color have more possibility to get together. This structure is called Vine clusters, which, in our problem, reflects the stocks in the same kind of sector.

**Question 4:**

The values of alpha for the two cases are summarized in the following table:



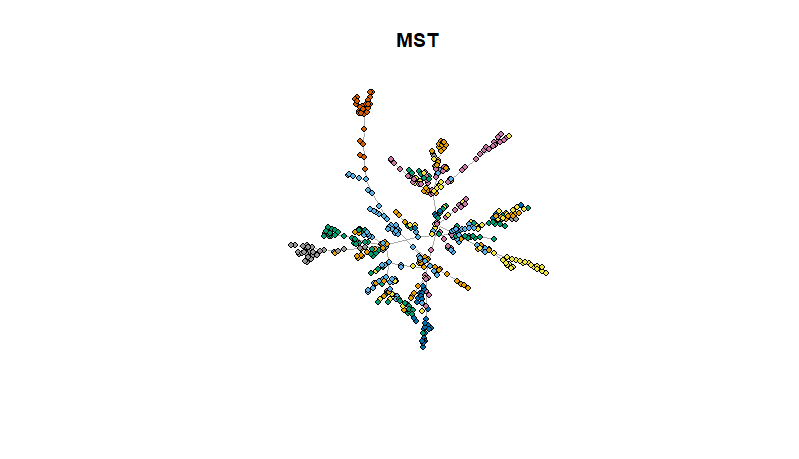
|  |  |  |
| --- | --- | --- |
| **Sectors** | **Alpha 1** | **Alpha 2** |
| Health Care | 0.108805668 | 0.121457490 |
| Industrials | 0.102304465 | 0.129554656 |
| Consumer Discretionary | 0.138495277 | 0.170040486 |
| Information Technology | 0.089574899 | 0.133603239 |
| Consumer Staples | 0.060188934 | 0.072874494 |
| Utilities | 0.052125506 | 0.056680162 |
| Financials | 0.110768662 | 0.127530364 |
| Real Estate | 0.057692308 | 0.062753036 |
| Materials | 0.034750337 | 0.048582996 |
| Energy | 0.068151147 | 0.068825911 |
| Telecommunication services | 0.006072874 | 0.008097166 |

The values are calculated from the cleaned data (494 stocks). We first conducted data preprocessing and discarded 11 stocks with different time length.

As its shown in the table, the values given by the second definition are always larger than the one given by the first definition. The reason behind that is because instead of analyzing one’s neighbors, we just defined the probability as the |Si|/|V|, which is clearly not a very good formulation when we have the real data. If we use the first definition, when calculating alpha, not all the stocks belong to one single sector have relationship with the one we investigated. And that’s why the second one is always larger than the first one.

**Question 5:**

The MST given by the weekly data:

****

Comparing with the plot in question 3, the stocks with same color do not stick together as a cluster. Rather, they are relatively distributed in the all graph. This is because when generating MST in this question, we used every Monday data, in other words, we discarded 4 day’s data when using every Monday data so that the result we got is very rough comparing with the accurate one computed using all the weekday’s data.