# NETWORK ANALYSIS OF S&P 500 STOCKS

-IE532 Project

Submitted by –

Nimit Kapadia (nimithk2)

Prannoy Kathiresan (prannoy2)

Shreekant Gokhale (gokhale6)

## Abstract:

In this project, we have worked on the daily stock price data of around 500 companies in the S&P 500. We processed this data to form a stock correlation network and further examined it to find the most dominant sector in the market for the considered period. We also studied the work that has already been done in this field and analyzed the methods used by them to draw conclusions about the state of the market. We also studied the effect of threshold on the structure of the graph by varying the theta values over its range in the correlation matrix. Also, we explain the results that we are getting from the maximum clique on the correlation-based graph and MST on the distance-based graph.


## Introduction:

Many analytical problems can be seen as network problems. In various fields like economic production, efficient storage, and transportation, looking at the problem as a network has proven to be beneficial as it becomes easier to visualize and formulate the problem. In a similar way, financial data can be considered as a network for further analysis. In this case, networks are not formed from the directly available data, but it needs to be processed first and then converted to the network. In stock correlation networks each company is the node, and these nodes are connected with the edges if there is a correlation between these companies. Network formation is explained in detail in the following sections. Using these networks, the effect of the complex market can be analyzed in an instance from the plot. It is also possible to study the dynamic nature of the market as time changes by observing changes in the structures of the resulting networks. Finding dominant sectors in the market and independent sectors can be helpful for portfolio optimization and risk management. Also, specific patterns can be seen in these networks based on the states of the market like major crashes or a bubble (like the dot com case in the 1990s).

Analysis of the market in the network form is not very intuitive but can be really inciteful as it gives a perspective of the whole market for a particular time. This was the motivation to study the conversion of stock price data in the network for this project.

## Literature Survey:

Mostly correlation is the basis for the formation of a network. The graph formed based on the correlations is further filtered using various methods and only a few edges are maintained in the final graph. Minimum spanning tree, and threshold networks are the filtering techniques that we have used for this project. Filtering the graph helps us to extract the important information from a load of data we have for each company.

In the minimum spanning tree graphs, correlations are further converted into distances. This conversion helps with the MST formulation and visualization as well. The companies with higher correlations will be closer in the graph as compared to other companies with lower correlations. The minimum spanning tree helps to show us the intercorrelation between companies of different sectors.

From the positions of different sector companies, we can extract a pattern based on the state of the market. It is seen that, in case of the market crashing, there will not be any extreme correlation between the companies of a specific sector. Different sectors will have a moderate correlation resulting in the whole graph acting as a single cluster. In this case, we can see sectors will be evenly scattered in MST. On the contrary, when there is a bubble in the market, some specific sectors can be seen to be forming clusters in the MST which implies, companies in that sectors are highly correlated to each other but are comparatively independent of other sectors.

In the case of threshold networks, the edges with a correlation less than the chosen threshold are eliminated and the data is represented only above the threshold. In these kinds of networks, strongly correlated relations are filtered out from all the other relations. There is a possibility of a loss of information in this case, as for higher threshold values we can lose the data of some companies that are moderately correlated with others. This can be a problem especially when the independence of the companies is important. One such example is portfolio optimization where we want to choose companies from different sectors which are independent of each other. But in MST we are getting at least some information about every company stock, and we can see which sectors are closer to each other.

## Network Formation:

We have considered closing day prices for the stocks of 494 Companies from the S&P 500. We got this data from Kaggle which is updated daily. For this project we considered the post covid data starting from 1st April 2020. Firstly, we calculated the daily log return of the prices to make the data stationary. The stationary data basically means that the average mean value for the price of the stock would be constant over time. We need to make the data stationary before calculating the Pearson coefficient between different stocks. We created the correlation matrix of the dimension equal to the number of companies in the data. Then we filtered the correlation values based on the threshold chosen. The correlation values lower than the threshold are not considered.

In the case of the maximum clique study, we formed a network based on these correlation values. If the correlation between the two companies is greater than the threshold, then nodes corresponding to those companies are joined with the edge where the weight of the edge was equal to the correlation.

While creating the MST graphs, we first need to convert the correlation values to the corresponding distances. The conversion was done as follows:

$$d_{ij} = \sqrt{2\left(1 - cor_{ij}\right)}$$

After converting all the correlations into distances, we create a graph where edge weights are the distances. Based on that, we found out the minimum spanning tree. Note that in this case no threshold is used as we do not miss out on any data from the correlations.

## Correlation-based network: Maximum Clique

Maximum clique gives a subgraph where each node is connected to every other node in the subgraph. This implies that the nodes in the maximum clique will have higher intercorrelation with the other nodes from the subgraph than the nodes that are not in the subgraph.

We generated the correlation networks for various threshold values (-0.2, 0.25, 0.6, 0.7, 0.8, 0.9) and then found the maximum clique for each graph. We also found out the sectors to which the companies from the maximum clique belong and color-coded the nodes accordingly. The results of this study are shown in Fig 1-3.

The colors used for different sectors are given in the table below.

| Sector | Color |
|---|---|
| Basic Materials | Red |
| Communication Services | Blue |
| Consumer Cyclical | Yellow |
| Consumer Defensive | Green |
| Energy | Orange |
| Financial Services | Magenta |
| Healthcare | Cyan |
| Industrials | Black |
| Real Estate | Pink |
| Technology | Purple |
| Utilities | Gray |

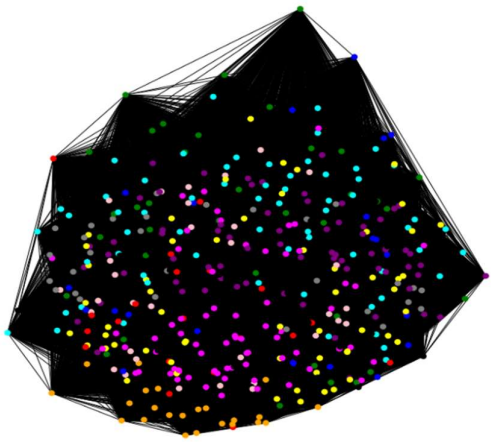
**Table 1: Color coding used for the results**

## Distance network: MST

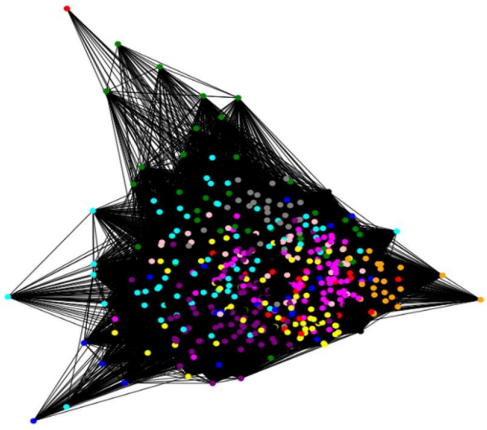
For formulating MST, we need to convert the correlation into distances. As mentioned above we created the distance matrix based on all the correlation values without using any threshold. Based on that we generated a graph and found out the lowest cost tree spanning all the companies from the data. We also color-coded the companies according to their sectors. The result for the spanning tree is shown in fig 4.

In the case of the minimum spanning tree, the companies with higher correlation would be closer to each in the graph. From the results, we can see that companies from the same sector are preferred over other connections in the minimum spanning tree.

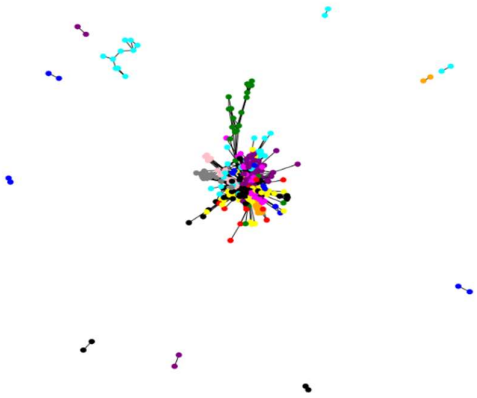
(θ = - 0.20)



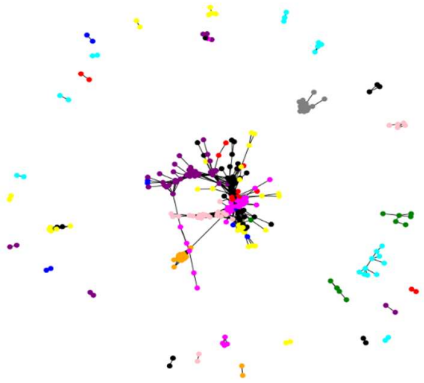
(θ =0.25)



(θ = 0.6)



(θ= 0.7)



(θ= 0.8)

(θ= 0.9)

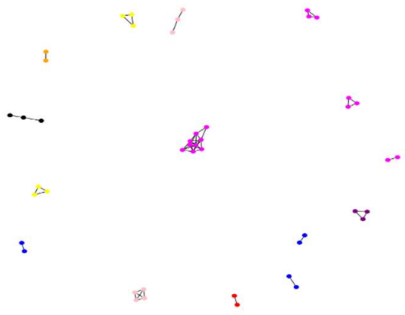
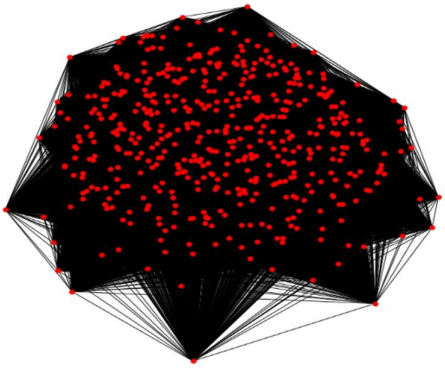
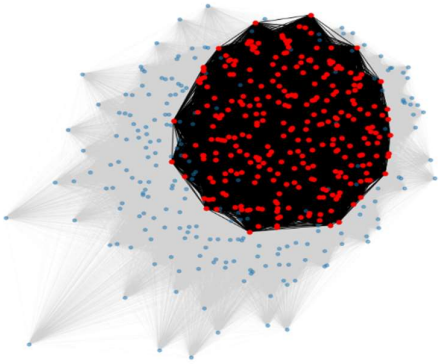


**Fig.1: Correlation-based threshold networks**

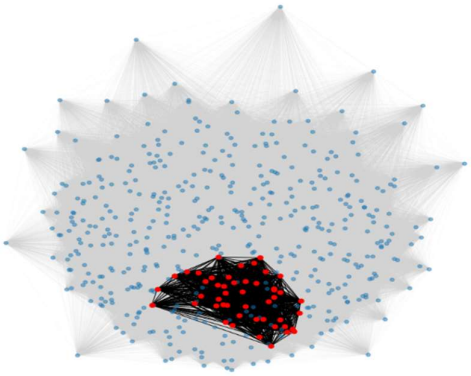
(θ = - 0.20)



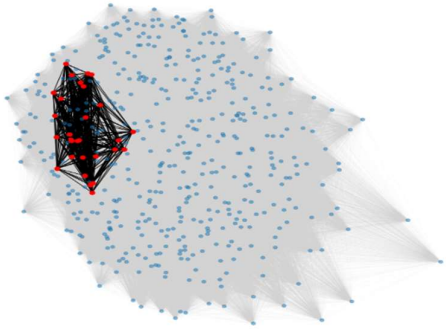
(θ=0.25)



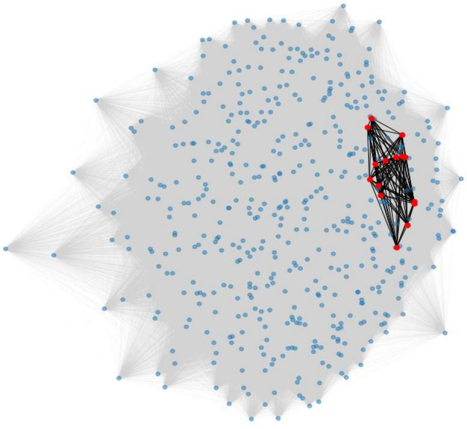
(θ= 0.6)



(θ= 0.7)



(θ= 0.8)



(θ= 0.9)

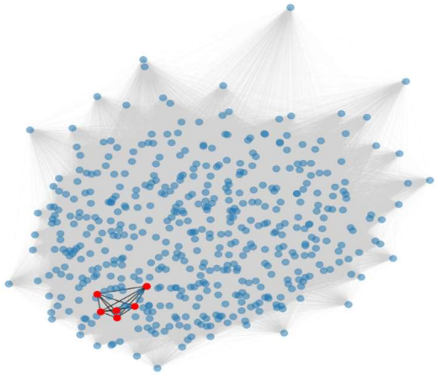
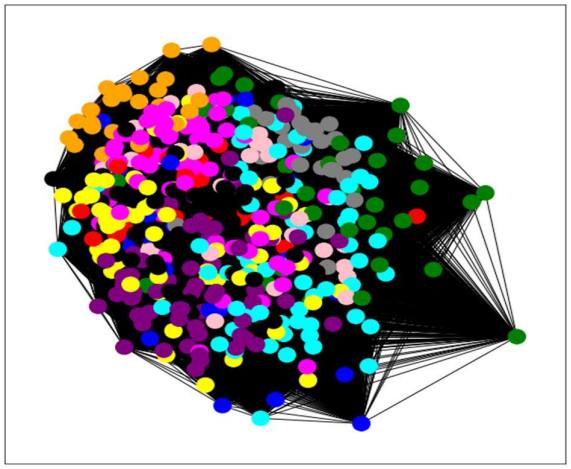
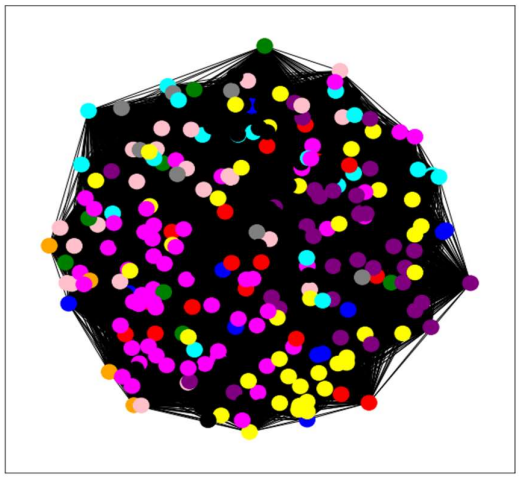


**Fig.2: Maximum Cliques in threshold networks**

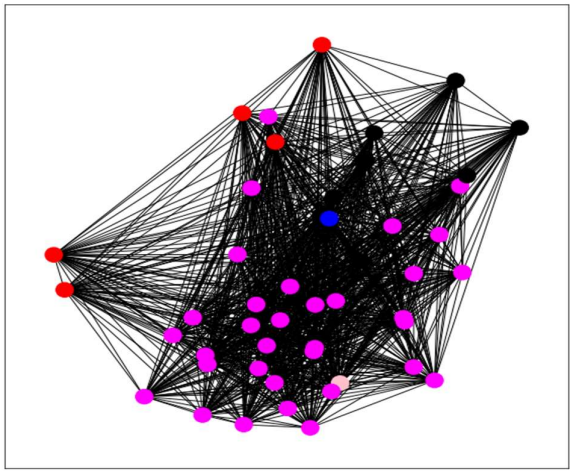
(θ = - 0.25)



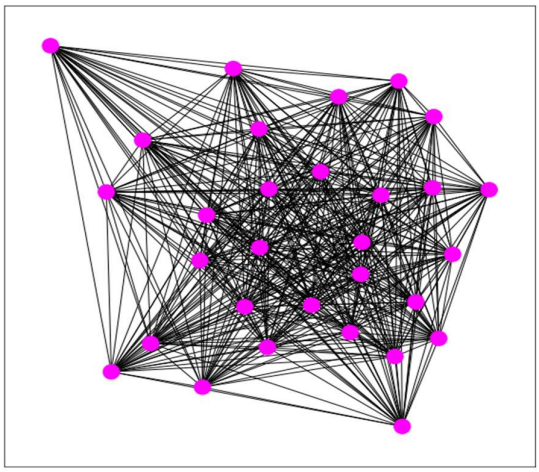
(θ=0.25)



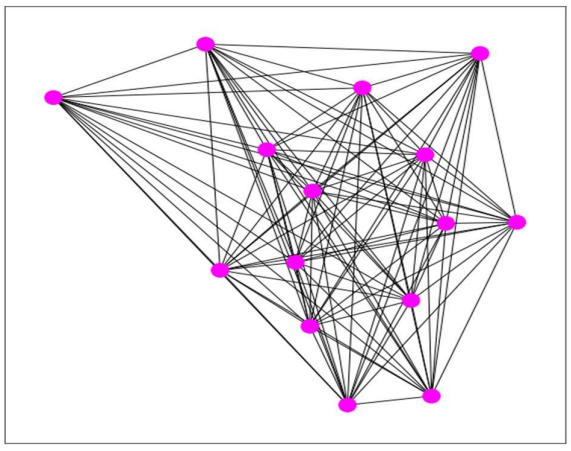
(θ=0.6)



(θ= 0.7)



(θ= 0.8)



(θ=0.9)

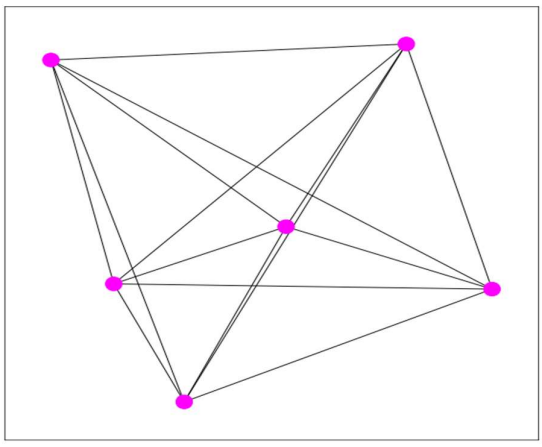


**Fig. 3: Sectors in maximum clique**

**Fig. 4: Distance-based Minimum spanning tree**

## Conclusion:

From the graphs in Fig 3 we can see that for higher threshold values, Financial Sector is found to be the most dominant sector for the chosen period. As we reduce the threshold value i.e., we allow lesser correlations to stay in the graph, we start seeing more sectors getting included in the maximum clique. We can see for lower threshold values; Basic materials, Communication Services, Industrials, and Financial Services were more dominant as compared to other sectors.

As we saw in MST, companies from the same sector preferred to be joined, which implies that intra-correlation for the sector is more significant than inter-correlation between the sectors. Even though MST is the lossless way of representing the data, we cannot infer the dominant sector. But from the patterns studied in the literature, we can predict the state of the market using MST.

Both the threshold network and MST representation of the stock correlation network is not robust. The structure of the network will change if there is any minor change in the considered data. To make the networks more robust, entropy measures are used in some literature works which study the structure of the graph based on the centrality measures.

## Future Scope:

As we studied the data limited to only one period, we could not include the dynamic nature of the market in the representation. We can study the data over multiple periods and based on the state of the market during the period, finding patterns in the MST is possible. Also, in the case where we want to find highly correlated stocks with one particular stock, we can replace the minimum spanning tree with the shortest path tree.

Additionally, from the independent sets and maximum cliques, we can find out the dominant and independent sectors in the market. This information can be utilized while planning the potential diversified investment. Also, it can be used to determine the volatility of certain portfolio.

## References:


1) https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks/code?resource=download&select=sp500_stocks.csv
2) Kukreti Vishwas, Pharasi Hirdesh, et. al., 'A perspective on correlation-based financial networks and entropy measures', Frontiers in Physics, 2020.
3) Kumar S., Deo N., 'Correlation and network analysis of global financial indices, Phys Rev E., 2012
4) Boginski Vladimir, Butenko Sergiy and Pardalos Panos,' Contribution for innovations in financial and economic networks', EE, 2003
5) Boganna G., Caldarelli G., et. al., 'Topology of correlation based minimal spanning trees in real and model markets, Phys Rev E., 2003
6) Boganna G., Caldarelli G., et. al., 'Networks of equities in financial markets', Eur Phys J B., 2004