Applications of Network Theory in Finance

By

Jean Robin Raherisambatra (jraherisambatra@aims.edu.gh)

June 2020

AN ESSAY PRESENTED TO AIMS-GHANA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN MATHEMATICS

DECLARATION

This work was carried out at AIMS-Ghana in partial fulfilment of the requirements for a Master of Science Degree.

I hereby declare that except where due acknowledgement is made, this work has never been presented wholly or in part for the award of a degree at AIMS-Ghana or any other University.

Student: Jean Robin Raherisambatra

Supervisor: Philip A. Knight

ACKNOWLEDGEMENTS

My first and foremost gratitude goes to the almighty God for granting me the gift of life to undertake my studies at AIMS. I would also like to thank AIMS for giving me this wonderful opportunity to do my masters. Next I would like to thank my supervisor Dr. Philip A. Knight for the opportunity to work with him for my essay and to benefit from his guidance.

The completion of this essay and my studies would not have been possible without the support of the entire AIMS-Ghana family, including the student body and both the teaching and the non teaching staff. My special thanks goes to the wonderful and hardworking tutors at AIMS-Ghana for their enormous support and guidance, especially Fred Soglohu, my assigned tutor. I would also like to thank Tonny Rija Christian Ramarolahy and Jean Michel Menjanahary for their suggestions and fruitful discussions towards this essay.

My sincere thanks go to Mrs Beauty Beatrice Kwawu, the English & Communication Lecturer, for her support in learning and improving my English at AIMS, and for proofreading the essay.

I would finally like to thank my family for their support and for being my inspiration from the very beginning.

DEDICATION

I dedicate this work to my parents Mr. Dominique Alfred Ramangasoavina and Mrs. Edwige Razafindramavo, my siblings and my entire family.

Abstract

Financial networks based on stock prices can be constructed by using threshold correlation chosen from the correlation matrix. The importance of this approach is to select the relevant correlations from the correlation matrix. In order to study this, we collect N stock prices where N is a positive integer and with each stock price having n variables. We choose one variable of the same description for each stock price and form them to get a new data. This new data has N variables because we start with N stock prices. For instance, in this essay, the chosen variable is the closing price for each stock price and by using the *Pearson-correlation coefficient* we can form the *correlation matrix* between the stock prices. We then form an adjacency matrix A_c by setting an appropriate correlation threshold on the correlation matrix. The network associated to the matrix A_c has nodes corresponding to the stock prices and the edges corresponding to the link connecting a pair of stock prices. We analyze this network obtained by using appropriate tools in the network theory. We find that the financial network depends on the choice of the threshold value and the real-world network is more distinguished by the relationship between stock prices than the random network. We find that network theory helps us to gain understanding of several problems in quantitative finance such as the patterns of links between variation in stock prices and the evolution of trade between companies.

Contents

1. Introduction

Our everyday activities, interactions, and relationships as humans can all be given or represented by networks. This makes networks important. Examples of these networks are family and friends networks, transportation and street networks, the telecommunication network, the distribution network of shops, financial networks, and so on. In this essay, our primary focus is on financial networks. To do this, we introduce some appropriate tools from network theory.

We will focus on networks describing the correlation between financial entities. These entities may include banks, firms, markets, prices of good and services, amongst others.

The use of network theory in financial domain is relatively new. In particular, during the financial crisis of $2008 - 2009$, it stimulated a lot of new research in this area [1, 3].

Analysts have used network theory to gain an understanding of several problems in quantitative finance such as the patterns of links between movements in stock prices and the evolution of trade between nations. The concept and use network theory in financial systems have appeared in response to the information that modern finance displays a high degree of interconnection. Globalization has magnified the level of financial interconnection over many variety of institutions or organizations.

We focus on the topic of the financial network, which is a concept describing any collection of financial entities and the link between them. Financial networks are composed of financial nodes, where nodes represent financial agents, and edges represent a weighted relationships between nodes.

In this essay, we study a financial network where nodes correspond to stock price and the edges between them to weighted links obtained by using threshold correlation. We use the estimated correlation matrix to measure the relationship between stock prices. Mantegna was the first person who built networks based on stock price correlation, according to [6].

The goal of this project is to construct financial network based on stock prices by using appropriate data, correlation matrix and threshold correlation. After that we analyze this financial network obtained by using appropriate tools. And in the end of this essay, with this appropriate data, we give a comparison between existing real-world networks and random networks.

In the first chapter, we give some overview of network theory and we explain the steps to construct a network using threshold correlation. After that, we present some mathematical tools that are used in investigating and analyzing networks. In the last chapter, we provide some data description and analysis. In particular, we analyze the stock prices of commodities on the New York stock exchanges. Then finally, we can compare the data with both random networks and existing real-world networks.

2. Background and Terminologies

2.1 Notion of networks

2.1.1 Definition (Network). A *network* is a pair of the form (V, E), such that V be a nonempty finite set and $E \subseteq V \otimes V$, whose elements are not necessarily distinct. An element of V is called a *node* or *vertex* of (V, E) and an element of E is called an *edge* of the network (V, E).

A network G is also known as a *graph*.

2.1.2 Example. This below is an example of graph with six nodes labeled to $1, 2, 3, 4, 5, 6$ and with edges $E = \{(1,2), (1,3), (1,4), (1,5), (2,3), (2,5), (2,6), (3,4), (3,6), (4,6), (5,6)\}$.

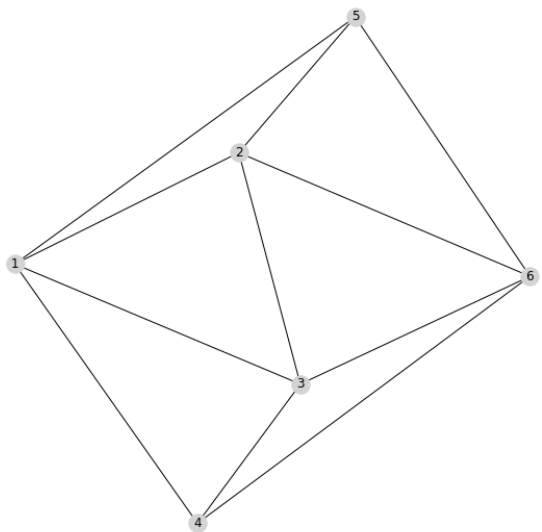


Figure 2.1: A graph with six nodes

Let G be a network. The set of edges E of G is called *reflexive* if $(v, v) \in E$ for all $v \in V$. It is called *anti-reflexive* if there is no edge of the form (v, v) (called *loop*). The network G is called *symmetric* or *undirected* if $(v_1, v_2) \in E$ for all $(v_2, v_1) \in E$. In case where G is both *symmetric* and *anti-reflexive*, and there is no duplicate edge, we say that it is *simple* network.

2.1.3 Definition. Let G be a network. A graph Γ is called *subgraph* if $V(\Gamma) \subseteq V$ and $E(\Gamma) \subseteq E$.

2.1.4 Definition. Let G be a symmetric network and $v \in V$. The *degree* of v is the number of edges connecting it to v.

2.1.5 Definition. A *path* of length p of the network G is a subgraph of G formed by $p+1$ vertices $v_0, v_1, \ldots, v_p$ of G, with edges $(v_0, v_1), (v_1, v_2), \ldots, (v_{p-1}, v_p)$. It is called a *cycle* of length p if $v_0 = v_p$.

The notion of connectivity is useful when we investigate the strength relations in networks.

2.1.6 Definition. A *network* is *connected* if any pair of nodes can be joined by a path.

2.1.7 Definition. A maximal connected subgraph of G is called *component*.

2.1.8 Definition. Suppose G is a simple network where $V = \{1, 2, 3, ..., n\}$.

For $1 \leq i, j \leq N$ define $a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases}$.

Then, the $n \times n$ square matrix $A = (a_{ij})$ is called the *adjacency matrix* of G.

2.1.9 Definition. A *distance* d on V is a mapping $d : V \times V \to \mathbb{R}$ that satisfies the following three properties:

1) For all $u, v \in V$, $d(u,v) \geq 0$ and $d(u,v) = 0 \iff u = v$.

2) For all $u, v \in V$, $d(u,v) = d(v,u)$ (symmetry).

3) For all $u, v, w \in V$, $d(u,v) \leq d(u,w) + d(w,v)$ (triangle inequality).

2.1.10 Definition. Let G be a network, with nodes labelled $1, 2, ..., n$ and A the adjacency matrix associated to G. We define a distance d as follows for all $u, v \in V$,

$d(u,v) = min\{k : (A^k)_{uv} \neq 0, k \geq 0\}$ where $(A^k)_{uv}$ is the (u,v)th entry of A^k and we let $d(u,v) = \infty$ if $\{k : (A^k)_{uv} \neq 0, k \geq 0\}$ is empty.

For each pair $u, v \in V$, one can easily show that $d(u,v)$ is the shortest path length between u and v in G.

The matrix $D = (d(u,v))_{u,v}$ is called the *distance matrix* of G whose (u,v)th entry is $d(u,v)$.

2.2 Construction of networks using threshold correlation

Correlation matrix:

Let N be a fixed positive integer and consider we have a system of data S composed of N different stock prices, each stock is represented by n variables (n is a positive integer) and each variable represent a time series of length T. In the language of networks, each element of S is represented by a node. We can find the correlation matrix associated with the system considered by using one variable that has the same description from each stock. We will use the correlation matrix to define correlation based-networks. In this case, the similarity between two nodes of the system is quantified by the linear correlation.

Now, define a matrix $C = (x_{ij})_{i,j}$ where x_{ij} is the linear (or Pearson) cross-correlation between element i and j, i.e:

$$x_{ij} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{\langle r_i^2 \rangle - \langle r_i \rangle^2}\sqrt{\langle r_j^2 \rangle - \langle r_j \rangle^2}} \tag{2.2.1}$$

where r_i and r_j are the investigated variables and for a vector r, the mean $\langle r \rangle$ is given by:

$$\langle r \rangle = \frac{1}{T}\sum_{k=1}^{T} r(k), \tag{2.2.2}$$

where $r(k)$ kth element of r. The correlation coefficient has values between -1 and $+1$, corresponding to perfectly anti-correlated and perfectly correlated variable, respectively. When, the coefficient is zero, the variables are not correlated. The matrix $C = (x_{ij})$ obtained is called the correlation matrix associated with the system of data.

2.2.1 Correlation threshold value. The correlation matrix C is a symmetric matrix (transpose $C^T = C$) then we have $\dfrac{N(N-1)}{2}$ distinct elements. Accordingly, we can filter out the most relevant coefficients by only dealing with values that exceed a threshold value c. This value is chosen empirically or analytically which indicates statistical significance of a correlation matrix. We use the threshold value to define an adjacency matrix and after that we get a network based on the threshold correlation.

The idea for choosing the threshold value is to filter out financial agents that are of particular interest.

We define the correlation network's adjacency matrix as:

$A_c = (a_{ij})$ where for $1 \le i, j \le N$ and

$$a_{ij} = \begin{cases} 1 & \text{if } |x_{ij}| \ge c \text{ for any } i \ne j \\ 0 & \text{otherwise} \end{cases} \tag{2.2.3}$$

where c is the threshold value choosen empirically in according to the entries of the correlation matrix C and x_{ij} the entries of C for all $1 \le i, j \le N$.

After we have built the adjacency matrix A_c from the threshold method, we also get the network G_c associated with A_c.

2.2.2 Definition. Let A_c be the adjacency matrix of a network G_c and $D = diag(A_c e)$ is the diagonal matrix whose entries are the degrees of each nodes. The matrix defined $L = D - A_c$ is called the *Laplacian matrix* of G_c.

In general, the network obtained from the threshold value is an undirected network without self loops but it may have multiple components.

3. Methodologies and Tools

In this Chapter, suppose our network G is a simple network and $V = \{1, 2, ..., n\}$ and E the edge set.

3.1 Clustering coefficient

All things in this section can be found at [5, p. 101-102].

In network analysis, a clustering coefficient quantifies the degree in which nodes be apt to cluster together.

3.1.1 The Watts-Strogatz clustering coefficient. The watts-Strogatz clustering coefficient specify how clustered a network is locally. The clustering coefficient for a node i in a network is given by:

$$C_i = \frac{number\ of\ transitive\ relation\ of\ node\ i}{total\ number\ of\ possible\ transitive\ relations\ of\ node\ i} \tag{3.1.1}$$

and denote by t_i the number of transitive relation of node i. In other words, denote by t_i the number of triangles attached to node i of degree k_i, then

$$C_i = \frac{t_i}{\frac{k_i(k_i-1)}{2}} = \frac{2t_i}{k_i(k_i-1)} \tag{3.1.2}$$

since $\frac{k_i(k_i-1)}{2}$ is the total number of possible transitive relation of node i .

So, the average clustering coefficient is given by:

$$\overline{C} = \frac{1}{N}\sum_i C_i \tag{3.1.3}$$

3.1.2 The Newman clustering coefficient. The presence of a relatively large number of triangles characterize many real networks. This characteristic feature of a network is a general consequence of high transitivity. Transitivity of a network is the global probability for the network to have neighboring nodes interrelated, thus inform the presence of firmly connected clusters.

The Newman clustering coefficient is also known as the transitivity index, it specify how clustered the network as a whole.

Let $|C_3|$ be the total number of triangles, and let $|P_2|$ be the number of paths of length two in the network. Then, the Newman clustering coefficient is given below:

$$C = \frac{3|C_3|}{|P_2|} \tag{3.1.4}$$

In this project, we need to compare a data with both random networks and existing real-word networks. To do this comparison, let us give a model for generating random networks: the Erdős-Réyni of random networks. We study some of the general properties of the networks generated by using this model such as expected number of node, clustering coefficient.

3.2 The Erdős-Rényi model of random networks

In this model,by Erdős-Rényi, we start with N isolated nodes. We then pick a couple of vertices with probability p and we add a link between them. In practice we fix a criterion value $p \in [0, 1]$ from which we generate the network. For each pair of nodes we generate a random number, r, evenly from $[0, 1]$ and if $p > r$ we add a link between them.

The Erdős-Rényi random graph is written as $G_{ER}(N, p)$ or $G_{ER}(N, m)$.

In the Erdős-Réyni random graph :

-the expected number of edges per nodes is $\overline{m} = \dfrac{N(N-1)p}{2}$;

-the average clustering coefficient is $\overline{C} = p$;

To have more understanding about this model, look at [5, p. 105-108] and [4, p. 233-256].

3.3 Fragments frequencies and specificities

Fragments are small sub-graphs with a particular topology in a larger network.

We are able to identify the small structural fragments of a system which are responsible for certain functional properties of the whole system. These kinds of structural fragments exist in financial network. We isolate these small fragments to understand how they work and gain understanding of their roles in the networks theory. For example, the triangles can indicate transitive relations in social networks. They also play a role in interactions between other entities in complex networks. In this section, we recall techniques to quantify some of the simplest but most important fragments or sub-graphs in networks.

We show how to determine whether the presence of these fragments in financial networks is just a manifestation of a random underlying process, or that they signify something more significant.

From now on, suppose that our network has N nodes and m edges.

3.3.1 Formulae to counting the most important small sub-graphs in networks. In this subsection, the theory giving the proof of the formulae to count small sub-graphs can be found in [5, p. 129-140], due to the limited length of this essay we just recall and give below the formulae from the notes to count the most relevant small sub-graphs.

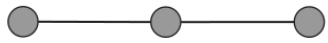
Denote k_i the degree of each node i in all formula and $\sigma(A) = \{\lambda : det(A - \lambda I) = 0\}$ the spectrum of the adjacency matrix A.

These below are particularly useful fragments for analyzing the financial networks:

$P_1, P_2, P_3, C_3, C_4, C_5, S_{1,3}, T_{3,1}$.

P_1 is an edge. The number of edges in a network can be obtained from the degrees of the nodes. See [5, p. 130]. P_2 represents a path of length 2. It is also a member of the family of star graphs ($S_{1,2}$).

(a) An edge



(b) Fragment of G denoted by P_2 or $S_{1,2}$.

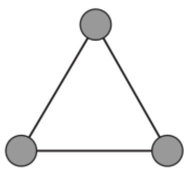
Each edge contribute to the degree of two nodes so:

$$|P_1| = m = \frac{1}{2}\sum_{i=1}^{N} k_i \tag{3.3.1}$$

The number of P_2 can be enumerated with the formula:

$$|P_2| = \sum_{i=1}^{N} \binom{k_i}{2} = \frac{1}{2}\sum_{i=1}^{N} k_i(k_i - 1) \tag{3.3.2}$$

Our next subgraphs are the cycle of length 3, C_3 and the path of length 3, P_3.



(a) Subgraph called triangle or cycle of length 3, denoted by C_3.

(b) A path of length 3, denoted by P_3.

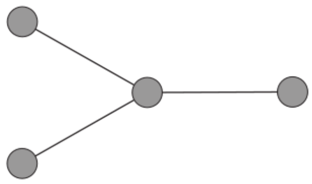
The number of triangles in the network is given below, where $\lambda_i \in \sigma(A)$ for each $i = 1, \ldots, N$ and A is the adjacency matrix associated of the network:

$$|C_3| = \frac{1}{6}\sum_{i=1}^{N} tr(A^3) = \frac{1}{6}\sum_{i=1}^{N} \lambda_i^3 \tag{3.3.3}$$

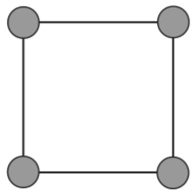
Then the number path of length 3 is given by:

$$|P_3| = \sum_{(i,j)\in E} (k_i - 1)(k_j - 1) - 3|C_3| \tag{3.3.4}$$

The next fragments are the star with four nodes $S_{1,3}$ and the cycle of length four C_4.



(a) A star sub-graph of the network, denoted by $S_{1,3}$.



(b) A cycle of length 4, denoted by C_4.

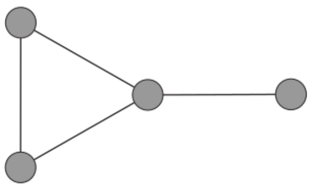
The number of $S_{1,3}$ star sub-graphs equals the number of times that the nodes attached to a given node can be combined in triples, namely,

$$|S_{1,3}| = \frac{1}{6}\sum_i k_i(k_i-1)(k_i-2) \tag{3.3.5}$$

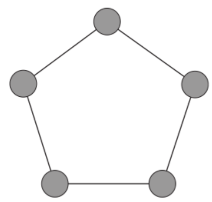
The number of C_4 is given below by:

$$|C_4| = \frac{1}{8}(tr(A^4) - 4|P_2| - 2m) \tag{3.3.6}$$

The next fragments are the tadpole $T_{3,1}$ and C_5, the cycle of length 5.



(a) The tadpole $T_{3,1}$ sub-graph



(b) A cycle of length 5, denoted by C_5.

Then, the number of tadpole $|T_{3,1}|$ is given below where t_i is the number of triangle associated for each node i:

$$|T_{3,1}| = \sum_{k_i>2} t_i(k_i-2) \tag{3.3.7}$$

The number of cycle of length five is given by:

$$|C_5| = \frac{1}{10}\left(tr(A^5) - 30|C_3| - 10|T_{3,1}|\right) \tag{3.3.8}$$

3.3.2 Network motifs. Network motif is one important local property of studying networks. They are sub-graphs that occur more frequently in a network than expected. Network motifs are small fragments that repeat themselves in a particular network or between many networks. Each of these small fragments, defined by a recurring pattern of interconnections between nodes, may reflect a framework in which specific functions are achieved efficiently.

Another way of phrasing this to note that network motifs are partial sub-graphs that are significantly over-represented in the data compared to a chosen random graph model that is assumed to fit well the data.

The relative abundance of a given fragment S can be estimated using the statistic:

$$\alpha_S = \frac{N_S^{real} - \langle N_S^{random} \rangle}{N_S^{real} + \langle N_S^{random} \rangle} \tag{3.3.9}$$

where N_S^{real} is the number of times the sub-graph S appears in the real-world network, and $\langle N_S^{random} \rangle$ is the average of the number of times that S appears in an ensemble of random network.

3.4 Centrality

In this section, we focus on some centrality measures, based on [5, p. 143-152] and [5, p. 158-160] for more information.

In many situations one may be interested in nodes which play an important role in a network. Centrality measure identifies which nodes are more significant central than others.

There are many different characterizations of the important node. An important node can communicate with other nodes, it closed to many other nodes, and it is indispensable to act as a communicator. Different centrality measure is obtained from these characteristics of the important node.

3.4.1 Degree centrality. The degree centrality is given by a nodes degree and measures the capability of a node to communicate directly with other nodes. By using the adjacency matrix, A, defined at (2.1.8), of G, the degree of a given node i is given by:

$$k_i = \sum_{j=1}^{N} a_{ij} = e_i^T A e = (Ae)_i \tag{3.4.1}$$

where a_{ij} is the (i, j)th entry of the matrix A and $e = (1, 1, \dots, 1)$ is the vector of ones of size N, all coordinates are one, e_i the vector with 0 coordinates except at the ith coordinate, which is 1 and e_i^T its transpose.

We say that the node i is more central than the node j if $k_i > k_j$.

3.4.2 Closeness Centrality. Closeness centrality is another measure of centrality of a node in a given network. We use closeness centrality to indicate the closeness of a node to all other nodes. Closeness centrality calculates as the average path length node by node for a centrality measure.

To compute the closeness centrality of each node, we need the following below:

1. the distance matrix D defined in (2.1.10);
2. the vector ones e, its entries are all one;
3. the vector e_i defined as all entries zero except the ith is one and e_i^T its transpose.

The closeness centrality of node i is defined as:

$$CC(i) = \frac{N-1}{e_i^T De} \tag{3.4.2}$$

The closeness centrality of a node i have the value between 0 and 1: $0 \leq CC(i) \leq 1$.

3.4.3 Remark. The more central nodes have higher closeness centrality and closeness centrality does not match degree centrality.

3.4.4 Betweenness Centrality. Betweenness centrality is another measure of centrality. It characterizes how important a node is in the communication between other pairs of nodes. Betweenness centrality accounts for the proportion of information that passes through node i on route from node j to node k and it assumes that the information travels along shortest path connecting one node to another.

The betweenness centrality of each node i in our network G_c is given by:

$$BC(i) = \sum_{i} \sum_{k \neq i,j} \frac{\rho(j,i,k)}{\rho(j,k)} \tag{3.4.3}$$

where $\rho(j,k)$ is the number of shortest paths from node j to node k, and $\rho(j,i,k)$ is the number of these shortest paths that passes over node i.

3.4.5 Katz centrality. The Katz centrality is another measure of centrality in a network. Let A be the adjacency matrix of a given network G and $e = (1,1,...,1)$ the vector of one, ie all entries of e are one.

The degree of node i, $k_i = (Ae)_i$, counts the number of walks of length one from node i to all other nodes. In 1953, Katz began to count not only the walks of length one but all walks of length started at node i. The intuition is that the closest neighbors have more influence over node i than more distant ones. The combination of the walks of all length makes us introduce an attenuation factor α such that more weight is given to shorter walks than to long ones. This is the reason that Katz did and the Katz centrality index is defined as:

$$K_i = \left[(\alpha^0 A^0 + \alpha^1 A^1 + \alpha^2 A^2 + ... + \alpha^k A^k + ...)e\right]_i = \left[\sum_{k=0}^{\infty}(\alpha^k A^k)e\right]_i \qquad (3.4.4)$$

(3.4.4) can be understood in terms of the resolvent matrix $(zI - A)^{-1}$ and this series converges so long as $\alpha < \rho(A) = max\{|\lambda| : \lambda \in \sigma(A)\}$, where $\rho(A) = \{\lambda : det(A - \lambda I) = 0\}$, and in this case $K_i = [(I - \alpha A)^{-1}e]_i$.

3.4.6 Eigenvector Centrality. Let λ_1 be the largest eigenvalue of the adjacency matrix A. So, the eigenvector q_1 associated with λ_1 is a centrality measure similar to Katz centrality coefficient. Furthermore, the i^{th} coordinate of q_1 is proportional to the number of walks that pass through i. The eigenvector q_1 is called the principal eigenvector of the matrix A, and their coordinates are called the eigeinvector centralities of the network.

All coordinates of q_1 are positive according to the Perron-Frobenius theorem.

3.5 Community detection

In network theory, communities are the topological organization to analyze or to investigate. For more understanding, see [5, p. 227-230] and [2, p. 358-396].

Nodes group together to form clusters, and clusters are also known as network communities. (Clusters may form for many reasons.)

3.5.1 Definition. *Communities* are groups of nodes more densely connected amongst themselves than with the rest of the nodes in the network.

3.5.2 Network density. Network density is given below:

$$\delta(G) = \frac{2m}{N(N-1)} \qquad (3.5.1)$$

where m is the number of edges in the network G and N the number of nodes in the network.

The properties of communities are more or less independent of individual nodes and of the network as a whole.

In this section let us give a method to detect communities.

3.5.3 Theorem (Fiedler vector)**.** *Look at [5, p. 84] for more information.*

Let G be a network. Assume that G is a connected network and L the graph Laplacian associated with G, see (2.2.2).

Suppose that the (second) smallest eigenvalue $\lambda_{N-1} > 0$ of L and denote by x be the eigenvector associated with $\lambda_{N-1} > 0$ and we associate the ith component of this vector x with the ith node.

Let $r \in \mathbb{R}$ such that $V = V_1 \cup V_2$ where $V_1 = \{v \in V | x_v \geq r\}$ and $V_2 = \{v \in V | x_v < r\}$.

Then at least one of the sub-networks of G induced by the sets V_1 and V_2 is connected. If $r = 0$ then both sets are connected.

The vector x is called the Fiedler vector.

3.5.4 Spectral partitioning. Spectral partitioning is a method to detect communities. An example of this spectral partitioning is the Fiedler vector. The Fiedler vector can be used to partition a connected network into two sub-networks. To develop this intuition, one can find that using spectral information parts a network into several fragments. The nodes are split according to whether x_i is smaller or bigger than r and choosing r to be the median value of the Fiedler vector x ensures that the network communities are evenly sized. The good popular choice of r is zero.

3.6 Assortativity

We give now the notion of assortativity coefficient based on [5, p. 181-184].

The correlation between the degrees of the nodes connected by links in a network may be useful to investigate. The idea of using these correlation is to classify that a network is assortative or disassortative. Respectively, assortative or disassortative indicates the tendency of high-degree nodes or lower-degree nodes to be connected to each other.

Measuring degree-degree correlation need to record the degrees, k_i and k_j, of nodes incident to every edges $(i, j) \in E$ in the network.

After measuring the degree-degree correlation, we can compute the statistics on this set of ordered pairs. Quantify the linear dependence between k_i and k_j by means of the Pearson correlation coefficient and we obtain a value $-1 \leq r \leq 1$. If $r > 0$ then the network is assortative and the network is disassortative if $r < 0$, and the network is neutral if $r = 0$.

$r > 0$ indicates that the degree-degree correlation is positive and $r < 0$ implies that the degree-degree correlation is negative.

The assortativity coefficient is defined as:

$$r = \frac{\frac{1}{m}\sum_{(i,j)\in E} k_i k_j - \left(\frac{1}{2m}\sum_{(i,j)\in E}(k_i + k_j)\right)^2}{\frac{1}{2m}\sum_{(i,j)\in E}(k_i^2 + k_j^2) - \left(\frac{1}{2m}\sum_{(i,j)\in E}(k_i + k_j)\right)^2} \tag{3.6.1}$$

where $m = |E|$ is the number of egdes in the network.

According to [5, p. 181-184], the assortativity coefficient can be written as:

$$r = \frac{|P_2(G)|\left(|P_{3/2}| + C - |P_{2/1}|\right)}{3|S_{1,3}| + |P_2(G)|\left(1 - |P_{2/1}|\right)} \tag{3.6.2}$$

where $C = \dfrac{3|C_3|}{|P_2(G)|}$ is the Newman clustering coefficient and $|P_{3/2}| = \dfrac{|P_3(G)|}{|P_2(G)|}$ and $|P_{2/1}| = \dfrac{|P_2(G)|}{|P_1(G)|}$ and $|P_1(G)| = m$, $|C_3|$ is the number of cycle of length three, $|P_2(G)|$ and $|P_3(G)|$ are the number of path of length two and path of length three in the network G, respectively.

4. Data Description and Analysis

To construct a financial network we need to use appropriate data from some stock prices. To do this, we need to collect N different data of stock prices in which each stock price is a function of n defined variables where N and n are positive integers.

4.1 Data Description

In this essay, we analyze two different financial networks denoted by G_1 with 20 nodes and by G_2 with hundred nodes. Both of these networks are generated from relationships between financial data related to the New York stock exchange. We use data from 2019-02-04 to 2020-01-17 for 20 items to generate G_1 and data from 2019-05-20 to 2020-05-19 for the 100 items to generate G_2. Each data set has $n = 6$ variables. These six variables are called *open, High, Low, Close, Adj Close, Volume*. The data in G_1 range from the stocks of companies, to currencies, commodities and indices, while the items in G_2, are only stocks of companies. To generate the networks, we create a new data by using the closing prices to generate a correlation matrix, one would expect that this matrix is symmetric with all elements on the diagonal equal to one since every stock price is related to itself. We then look at the correlation coefficient of the correlation matrix and choose a threshold correlation c to define an adajacency matrix A using the criterion in (2.2.3). This adds another index to our graphs. So the label of our graphs $G_{i,c}$ where $i = 1, 2$, $0 \leq c \leq 1$.

The data which we use is obtained from https://finance.yahoo.com/.

4.2 Data Analysis for $N = 20$.

Figure (4.1) illustrates a sample from one of the N=20 data series taken for G_1:

Date	Open	High	Low	Close	Adj Close	Volume
2019-02-04	167.410004	171.660004	167.279999	171.250000	168.315582	31495500
2019-02-05	172.860001	175.080002	172.350006	174.179993	171.195389	36101600
2019-02-06	174.649994	175.570007	172.850006	174.240005	171.254364	28239600
2019-02-07	172.399994	173.940002	170.339996	170.940002	168.010910	31741700
2019-02-08	168.990005	170.660004	168.419998	170.410004	168.208328	23820000
2019-02-11	171.050003	171.210007	169.250000	169.429993	167.240967	20993400

Figure 4.1: Data of the stock price of the Apple company

Figures (4.2) and (4.3) give some details of the data forming by the closing prices.

Date	AAPL	AMZN	BitcoinUSD	CBOE	Crude_Oil	DowJonesInd	FB	GoldJun	GOOG	IMMU
2019-02-04	168.315582	1633.310059	3459.154053	15.730000	54.560001	25239.369141	169.250000	1314.300049	1132.800049	15.15
2019-02-05	171.195389	1658.810059	3466.357422	15.570000	53.660000	25411.519531	171.160004	1314.199951	1145.989990	14.97
2019-02-06	171.254364	1640.260010	3413.767822	15.380000	54.009998	25390.300781	170.490005	1309.500000	1115.229980	14.78
2019-02-07	168.010910	1614.369995	3399.471680	16.370001	52.639999	25169.529297	166.380005	1309.400024	1098.709961	14.35
2019-02-08	168.208328	1588.219971	3666.780273	15.720000	52.720001	25106.330078	167.330002	1313.699951	1095.060059	14.04

Figure 4.2: Illustration of the part of the data

JNUG	Nasdaq	Netflix	Nikkei225	Russel2000	S&P500	TGTCorp	TSLA	USDJPY	WLL
55.171822	7347.540039	351.339996	20883.769531	1517.540039	2724.870117	69.806366	312.890015	109.438004	30.070000
56.613247	7402.080078	355.809998	20844.449219	1520.229980	2737.699951	70.124359	321.350006	109.960999	28.750000
53.133942	7375.279785	352.190002	20874.060547	1518.020020	2731.610107	69.873817	317.220001	109.941002	28.799999
50.698429	7288.350098	344.709991	20751.279297	1505.630005	2706.050049	69.247444	307.510010	109.973999	26.950001
54.078327	7298.200195	347.570007	20333.169922	1506.390015	2707.879883	68.303070	305.799988	109.755997	26.379999

Figure 4.3: Illustration of the part of the data

After we have constructed the appropiate data, we can compute the correlation matrix using the Pearson correlation coefficient defined in (2.2.1). Denote by $C = (x_{ij})$ the correlation matrix associated to the data have been constructed, see (2.2.1).

The correlation matrix coefficient associated with this data is given in Figure (4.4):

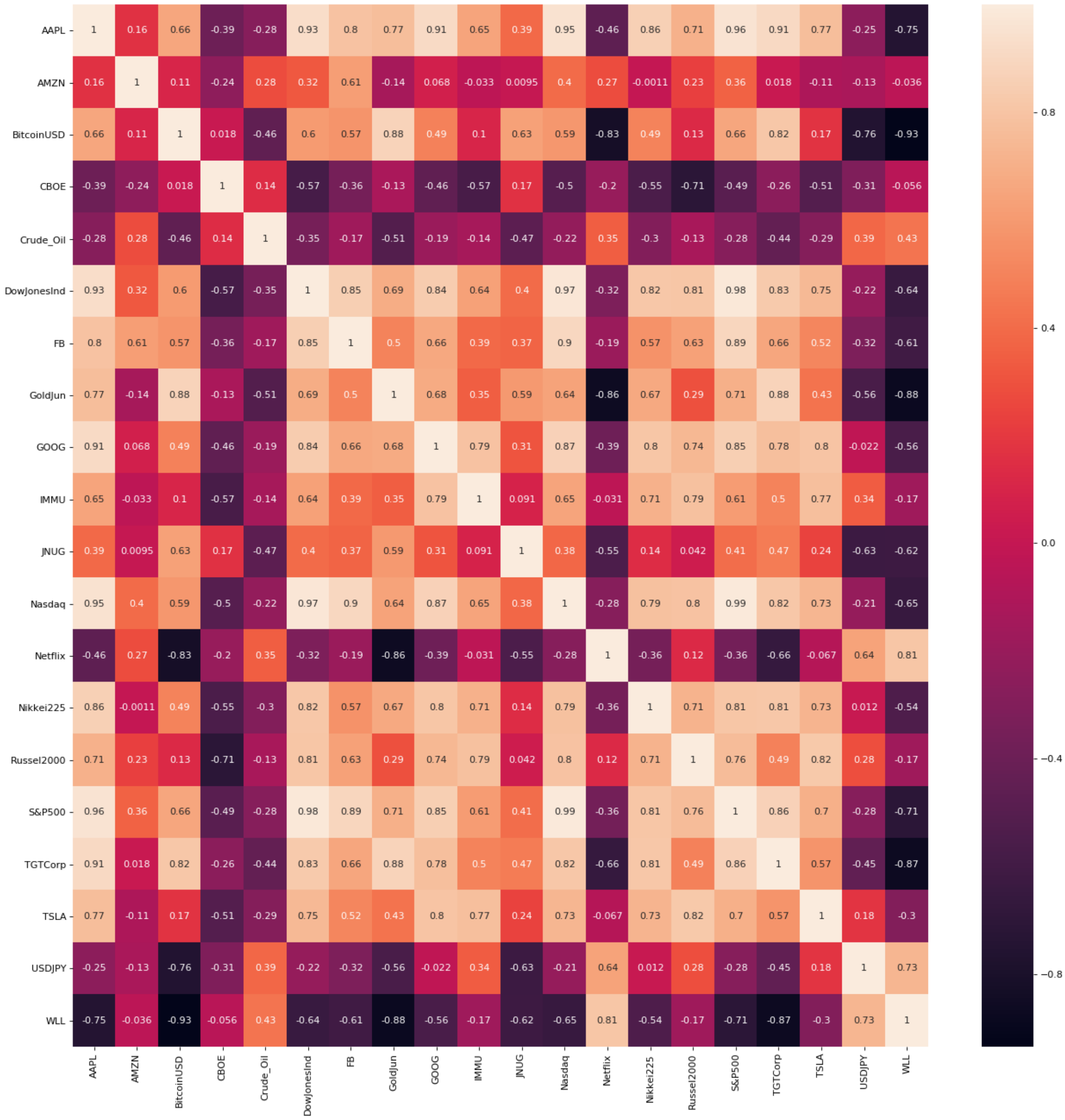


Figure 4.4: Correlation coefficient of the matrix C obtained by using Pearson correlation defined in (2.2.1).

To get the financial network, we need to form an adjacency matrix A obtained from choosing a threshold value c.

The threshold value is taken between 0 and 1 and in this work we pick threshold value in which there is a strong correlation between agents. In other words, we focused on the threshold value $c \in [0.6, 1]$.

For the three thresholds correlation $c = 0.6$, $c = 0.7$ and $c = 0.75$, we can see the connection between financial agents and the communities formed for using them.

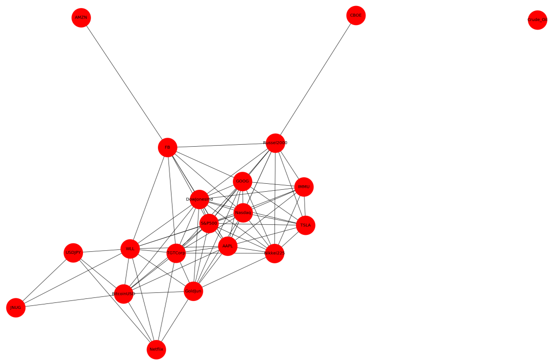
We use python code for all computations and for all networks obtained.

The adjacency matrix changes if we change the correlation threshold value. This means that the network we obtain depends on the threshold value.

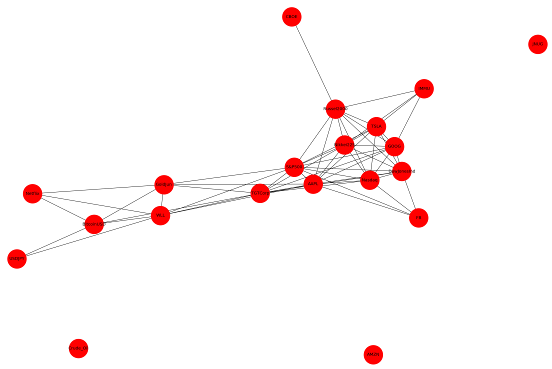
An example of the adjacency matrices corresponding to the threshold value $c = 0.6$ is given below:

$$A = \begin{pmatrix}
0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0
\end{pmatrix}$$

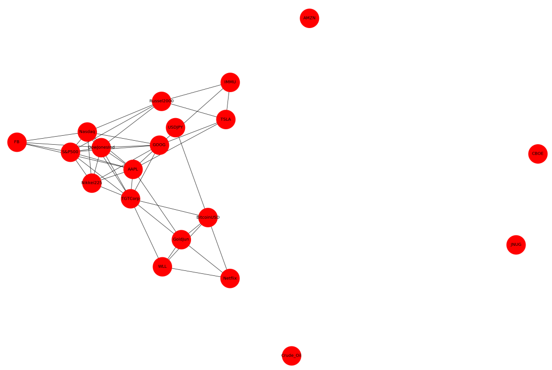
Below are the graphs for the threshold values $c = 0.6$, $c = 0.7$ and $c = 0.75$



(a) Network $G_{1,0.6}$ with the threshold value $c = 0.6$



(b) Network $G_{1,0.7}$ with the threshold value $c = 0.7$



(c) Network $G_{1,0.75}$ with the threshold value $c = 0.75$

Figure 4.5: Graphs with different threshold values for 20 nodes.

Every labeled node in the network represents the name of a financial instrument.

These figures above shows that different correlation thresholds give different networks. The networks represented in the figure(4.5a), figure(4.5b) and figure(4.5c) above, are all disconnected network.

4.2.1 Number of the most important fragments in the real-world networks. It is important to study the most important small fragments in the network. We apply the formulae given in (3.3.1) to compute the number of all most important small fragments. This helps us to understand the basic relationship structures between the stocks.

The table (4.1) below gives the number of basic fragments for some specific threshold:

$Fragments$	Number of fragments $for\ c = 0.6$	Number of fragments $for\ c = 0.7$	Number of fragments $for\ c = 0.75$
P_1	81	58	45
P_2	747	416	253
P_3	6411	2734	1299
C_3	187	99	55
C_4	1120	444	185
C_5	6616	1878	574
$T_{3,1}$	4961	2030	873
$S_{1,3}$	2228	964	456

Table 4.1: Number of the most relevant fragment in the network.

As expected, the threshold determines the number of edges in the graph and as we can see, as the correlation threshold increases the number of edges (P_1) reduces, and we can also see that the number of the other fragments reduces, this is also expected as the number of edges reduces. These are obvious trends, what are more important and interesting are the hidden trends. This result does not also seem to have any hidden trends. The only information we can deduce from this result is the basic relationship structure between the nodes. We can see that on the average, P_3 has a higher occurrence than the remaining fragments. We also see that the number of transitive relation decreases if we increase the threshold value. This means that the number of transitive relation of stock prices diminish if we increment the threshold correlation. This means again that the transitivity of the network decreases when we increase the threshold. We can also see high numbers of C_5, this means that the stock prices have deep cycle relations.

4.2.2 Centrality Measures. We use centrality measures to determine the nodes that are at the centre of our network, and by the centre, we don't mean geometric centre but centre with respect to the nodes with the largest number of relations.

Table (4.2) below summarizes different centrality measures and their corresponding values for $c = 0.6$, the results for $c = 0.7$ and $c = 0.75$ can be found in the appendix.

$Centrality$	$values$
$Degree\ centrality$ DC	[0.684 0.052 0.473 0.052 0.0 0.684 0.473 0.526 0.578 0.421 0.157 0.631 0.263 0.526 0.526 0.684 0.578 0.421 0.210 0.578]
$Betweenness$ $centrality:\ BC$	[0.056 0.0 0.065 0.0 0.0 0.056 0.109 0.028 0.017 0.0 0.0 0.033 0.005 0.009 0.104 0.056 0.041 0.0 0.001 0.126]
$Eigenvector$ $centrality:\ EC$	[0.321 0.021 0.184 0.023 4.183e-16 0.321 0.223 0.250 0.285 0.220 0.045 0.305 0.095 0.265 0.241 0.321 0.270 0.220 0.053 0.230]
$Closenness$ $centrality:\ CC$	[0.741 0.387 0.588 0.370 0.0 0.741 0.631 0.609 0.631 0.532 0.415 0.710 0.460 0.588 0.588 0.741 0.655 0.532 0.426 0.655]

Table 4.2: Different centrality measures for the network obtained by using $c = 0.6$ and $N = 20$ stocks.

According to Table (4.2), we notice that Dow Jones Ind. and S&P500 have the highest degree centrality, eigenvector centrality and closenness centrality while Facebook and Whiting Petroleum Coorporation have the highest betweeness centrality. Dow Jones Ind. and S&P500 are indices so we expect that they have more connections since their prices are directly affected by changes in the prices of their constituent stocks. On the other hand it is not very clear as to why Facebook and Whiting Petroleum will have the highest betweeness centrality. Given each pair on nodes, there is a shortest path that joins them, the betweeness centrality on a node is a measure of the number of such paths that go through it.

4.2.3 Network Motifs. Network motifs are fragments which appear more frequently in the real-world network than expected in the random one. So, to emphasise that which fragment can be considered as a motif in our real-world network, we need to investigate the abundance relative for each fragment. For more precisely, let us find the relative abundance for the most relevant fragments which were considered in the previous section but before that let us consider an Erdős-Rényi random network with $N = 20$ nodes and m the number of links which is the same as in the real-world network, for which the probability is given by $p = \dfrac{2m}{N(N-1)}$.

c	0.6	0.7	0.75
p	0.426	0.305	0.236

Table 4.3: Probability corresponding to each threshold $c = 0.6$, $c = 0.7$ and $c = 0.75$

We can record the number of $P_1, P_2, P_3, C_3, C_4, C_5, T_{3,1}$ and $S_{1,3}$ in the Erdős-Rényi graph. To determine the relative abundance of each fragment in the real-world network, we need to repeat the computation of the number of these fragments at least 50 times and we need to record the mean of each fragment.

After computation on python, the relative abundance of each fragment relatively to the real-world network based on the threshold value is given below:

Fragments	*Relative abundance for each fragment in* $G_{1,0.6}$	*Relative abundance for each fragment in* $G_{1,0.7}$	*Relative abundance for each fragment in* $G_{1.0.75}$
P_2	0.075	0.117	0.130
P_3	0.151	0.223	0.240
C_3	0.340	0.494	0.553
C_4	0.374	0.536	0.593
C_5	0.401	0.554	0.597
$T_{3,1}$	0.419	0.582	0.630
$S_{1,3}$	0.171	0.251	0.250

Table 4.4: Relative abundance for the most relevant fragments in the network.

where $G_{1,0.6}, G_{1,0.7}, G_{1,0.75}$ are the real-world network in the Figures (4.5a),(4.5b) and (4.5c).

We see in the Table (4.4) that for $c = 0.6$, $c = 0.7$ and $c = 0.75$, all the fragments are motifs with C_3, C_4, C_5 and $T_{3,1}$ having relatively high relative abundance values, and this means they are more relevant in our network. This tells us that transitivity of nodes and deep cyclic relations are relevant patterns in our network. The deep cycle relations tell us that a change in a stock price will create a cyclic ripple effect in the prices of the stocks depending on the length of the cycle. We can see that the $T_{3,1}$ is also a relevant fragment, this suggests that the deep cycle relations are linked.

4.2.4 Clustering Coefficient. Table (4.5) shows that the three real-world financial networks given in the Figures (4.5a), (4.5b) and (4.5c) are all assortative because the assortativity coefficient are all positive. We see that the average Watts-Strogatz clustering coefficient $\overline{C}_{G_c}$ of the whole network is significantly larger than that of a random network of the same size. Moreover, we see that both the average Watts-Strogatz clustering coefficient and the Newman clustering coefficient are very high in the real-world network for $c = 0.6$, $c = 0.7$ and $c = 0.75$. The high value of the Newman clustering coefficient indicates a large transitivity in the networks. The high value of the average Watts-Strogatz clustering coefficient implies a large number of transitive relation of each given node i. And we identify that the average Watts-Strogatz clustering coefficient and the Newman clustering coefficient decrease if we increase the threshold correlation.

The trend in percent of the network can see on the variation of the average clustering coefficient is 25% for the real world network and 19% for the random one. The Newman clustering coefficient decreases 10% if we increase the threshold correlation. These all mean that 25% of the transitive relation of a given node disappeared, 10% of transitivity disappeared if we increase the threshold correlation in the real world network. This happens due to the degree of the relationship between stock prices.

	network based on $c = 0.6$	*network based on* $c = 0.7$	*network based on* $c = 0.75$
Mean clustering of the network $G_{1,c}$: $\overline{C}_{G_{1,c}}$	0.67	0.62	0.52
Mean clustering of the random network : $\overline{C}_{G_{ER}(n,p)}$	0.44	0.27	0.25
Newman clustering coefficient C	0.751	0.713	0.652
Assortativity coefficient denoted by r	0.153	0.145	0.209

Table 4.5: Clustering coefficient and assortativity coefficient.

4.2.5 Communities. Communities are groups of nodes more densily connected than the rest of the nodes in the network. The network density is given in the table:

c	0.6	0.7	0.75
$\delta(G_{1,c})$	0.426	0.305	0.236

Table 4.6: Network density of the networks $G_{1,0.6}, G_{1,0.7}, G_{1,0.75}$.

To detect communities, we use the Fiedler vector method given in (3.5.3) and (3.5.4) to partition a network into a clusters (communities).

In this project, communities are financial agents frequently grouped together forming densely connected clusters which are poorly connected with other parts of the network.

For $c = 0.6$, we obtained two communities denoted by C_1 and C_2, and the other part is a member of an isolated node not in the communities. The Table (4.7) below described the elements of each communities and the elements of the member of an isolated nodes:

Elements in community C_1	*Elements in community* C_2	*Elements in neither communities*
$0, 3, 5, 7, 8, 9, 11, 13, 14, 15, 17$	$1, 2, 6, 10, 12, 16, 18, 19$	4

Table 4.7: Elements of the communities C_1 and C_2 of the network $G_{1,0.6}$.

For $c = 0.7$, we obtained two communities denoted by C_1 and C_2, and the other part is a member of an isolated nodes not part in the communities. The Table (4.8) described the elements of each communities and the elements of the member of an isolated nodes:

Elements in community C_1	*Elements in community* C_2	*Elements in neither communities*
$3, 5, 6, 8, 9, 11, 13, 14, 17$	$0, 2, 7, 12, 15, 16, 18, 19$	$1, 4, 10$

Table 4.8: Elements of the communities C_1 and C_2 of the network $G_{1,0.7}$.

For $c = 0.75$, we obtained two communities denoted by C_1 and C_2, and the other part is a member of an isolated nodes not part in the communities. The Table (4.9) described the elements of each communities and the elements of the member of an isolated nodes:

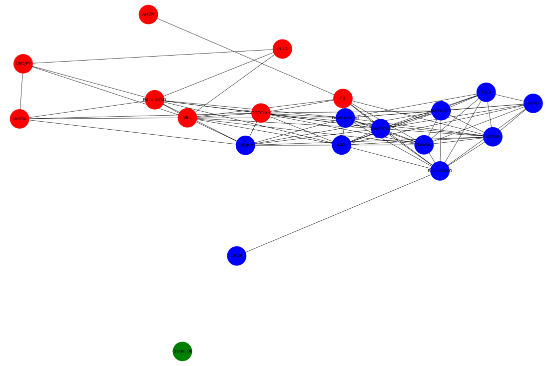
Elements in community C_1	*Elements in community* C_2	*Elements in neither communities*
$2, 7, 12, 18, 19$	$0, 5, 6, 8, 9, 11, 13, 14, 15, 16, 17$	$1, 3, 4, 10$

Table 4.9: Elements of the communities C_1 and C_2 of the network $G_{1,0.75}$.

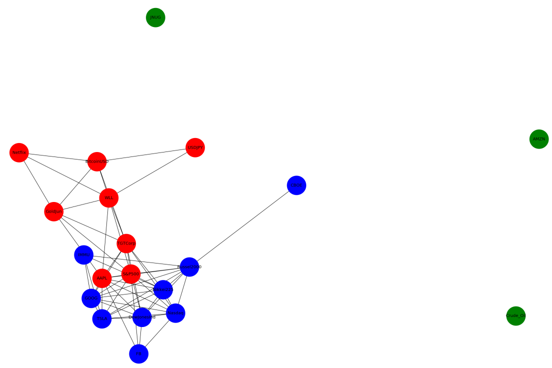
where node labeled 0=AAPL, 1=AMZN, 2=BitcoinUSD, 3=CBOE, 4=Crude-Oil, 5=DowjonesInd, 6=FB, 7=Goldjun, 8=GOOG, 9=IMMU, 10=JNUG, 11=Nasdaq, 12=Netflx, 13=Nikkei225 , 14=Russel2000, 15=SP500, 16=TGTcorp, 17=TSLA, 18=USDJPY, 19=WLL.

One would expect that nodes in a particular community will have a common property. In our results, the nodes are divided into two communities. A careful study of these communities does not reveal any clear basis on which the nodes are clustered. Each of the clusters have some few nodes that have a common property but this is not enough to make any claim. There might be a pattern but perhaps we are not noticing it due to the small number of nodes. Studying a financial network with a larger number of nodes will tend to give us a better picture of real world stock market network.

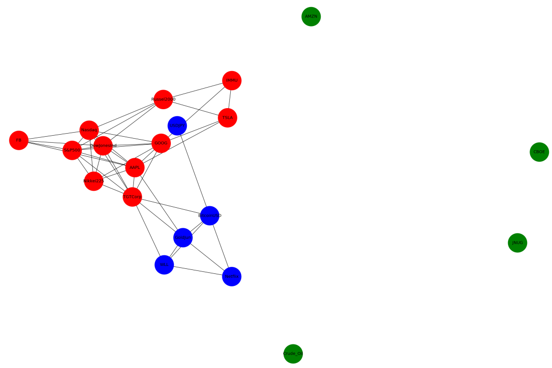
The Figures (4.6a), (4.6b) and (4.6c) below show these communities correspond to the network generated to $c = 0.6$, $c = 0.7$ and $c = 0.75$, respectively. We observed that the group of nodes colored in blue indicates one community and the group of nodes colored in red show the other community, while the nodes colored in green indicate that they are not in either community. We also observed that the number of the member of each community decreases if we increase the threshold correlation value. This happen due to the decreasing of the density of the community if we increase the threshold correlation.



(a) Communities in the network corresponding to the threshold $c = 0.6$



(b) Communities in the network corresponding to the threshold $c = 0.7$



(c) Communities in the network corresponding to the threshold $c = 0.75$

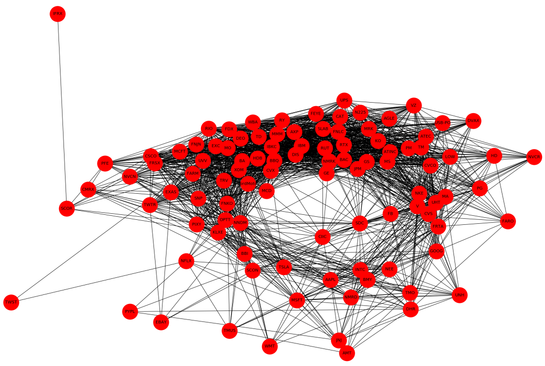
Figure 4.6: Communities with different threshold value $c = 0.6, c = 0.7$ and $c = 0.75$ for 20 nodes

4.3 Data Analysis for $N = 100$.

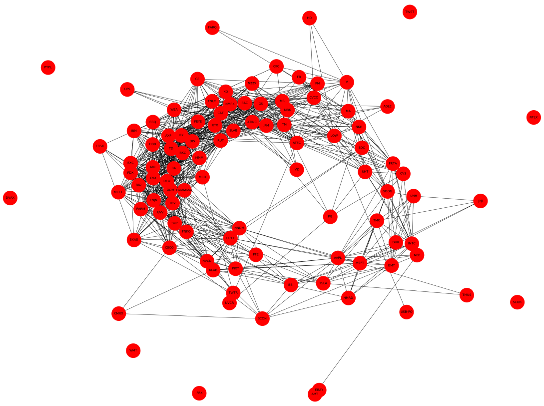
In all section and subsection, let us use two threshold correlation $c = 0.6$ and $c = 0.75$.

We use similar method and similar tools as in the first case to get our network based for the 100 stock prices.

The Figures (4.7a) and (4.7b) show the networks based on stock prices for $c = 0.6$ and for $c = 0.75$, respectively.



(a) Network $G_{2,0.6}$ with the threshold $c = 0.6$.



(b) Network $G_{2,0.75}$ with the threshold $c = 0.7$.

Figure 4.7: Networks with different threshold correlation for 100 nodes.

4.3.1 Number of the most relevant fragments considered in the network. . The Table (4.10) below show the number of the most relevant small-fragments in the network $G_{2,0.6}$ and

$G_{2,0.75}$:

$Fragments$	Number of fragments $for\ c = 0.6$	Number of fragments $for\ c = 0.75$
P_1	1846	886
P_2	77774	21936
P_3	3312753	557355
C_3	18025	4716
C_4	528420	81082
C_5	17017535	1529605
$T_{3,1}$	2405141	384363
$S_{1,3}$	1151536	200337

Table 4.10: Number of the most relevant fragments in the networks $G_{2,0.6}$ and $G_{2,0.75}$.

Similarly as the case of network with 20 nodes, we see in the Table (4.10) that the number of the fragments considered decrease if we increase the threshold correlation. This means that the structure of our network changes if we change the threshold value. In other words, for a given threshold value, there is a unique network $G_{2,c}$ correspond on it. So the interpretation of this result in the Table (4.10) is the same as on the interpretation of the result in the Table (4.1) of the network with 20 nodes.

4.3.2 Centrality measures. For $c = 0.6$, according to the degree centrality, the companies the Goldman Sachs Group Incorporation, Russel 2000, IBERIA BANK Corporation, Bank of America Corporation, the Walt Disney Company and Raytheon Technologies Corporation ordering respectively with highest degree of centrality, are the most central node in the network $G_{0.6}$. Betweenness centrality give us that the companies CIIG Merger Corporation, KLX Energy Services Holdings Incorporation, Nano Dimension Ltd, Facebook Incorporation, ComScore Incorporation and Wallgreens Boots Alliance Incorporation are most central companies in allowing communication between other pairs of companies and they simplify or constrain the communication or information between other companies in the network. Eigenvector centrality gives us that the companies IBERIA BANK Corporation, Russel 2000, the Walt Disney Company, the Toronto Dominion Bank, Royal Bank of Canada and Bank of America Corporation are the most central node and the closeness centrality give us that the companies Goldman Sachs Group Incorporation, Russel 2000, Raytheon Technologies Corporation, Bank of America Corporation, IBERIA BANK Corporation and the Walt Disney Company are the most central node in the network.

The intuition is analogous of the case $c = 0.6$ above for $c = 0.75$. For $c = 0.75$, the degree centrality give us that the companies ordered with the highest centrality respectively, Russel 2000, the Walt Disney Company, Royal Bank of Canada, IBERIA Bank Corporation, the Toronto Dominion Bank and Raytheon Technologies Corporation are the most central in the network. Betweenness centrality give us that the companies Russel 2000, LOWe's Companies, Alphat incorporation, Microsoft Corporation, Alphatec Holdings Incorporation and Ocean Power Technologies Incorporation are the most central companies in allowing communication between other pairs of companies. Eigenvector centrality give us that the companies Russel 2000, Royal Bank of Canada, the Walt Disney Company, IBERIA Bank Corporation, the Toronto Dominion Bank and American Express Company are the companies with highest centrality in the network. The closenness centrality give us that the companies Russel 2000, Raytheon Technologies Corporation, IBERIA Bank Corporation, the Toronto Dominion Bank, Royal Bank of Canada and Exxon Mobil Corporation have high centrality in the network.

4.3.3 Network motifs. We need to study the relative abundance to identify which of the fragments are considered as motifs. So, based on our real-world network with $N = 100$ nodes, let us consider an Erdős-Rényi graph with $N = 100$ nodes and m edges which is similar to our real-world network. The table below summarizes the probability of creating an edge between a pair of nodes:

c	0.6	0.75
p	0.37	0.18

Table 4.11: Probability for creating an edge between a pair of nodes in the Erdős-Rényi model for $c = 0.6$ and $c = 0.75$.

We can record the number of $P_1, P_2, P_3, C_3, C_4, C_5, T_{3,1}$ and $S_{1,3}$ in the Erdős-Rényi graph. To determine the relative abundance of each fragment in the real-world network, we need to repeat the computation of the number of these fragments at least 50 times and calculate the mean for each fragment.

The relative abundance of each fragment relative to the real-world network based on the threshold value is given below:

Fragments	*Relative abundance for each fragment in* $G_{2,0.6}$	*Relative abundance for each fragment in* $G_{2,0.75}$
P_2	0.068	0.175
P_3	0.147	0.354
C_3	0.360	0.674
C_4	0.393	0.745
C_5	0.440	0.808
$T_{3,1}$	0.445	0.780
$S_{1,3}$	0.167	0.387

Table 4.12: Relative abundance of each of the most relevant fragment in the network G_c.

where G_c is the real-world network based on threshold correlation, see Figure (4.7a) and (4.7b).

For $c = 0.6$ and $c = 0.75$, we see in the Table (4.12) that the fragments C_3, C_4, C_5 and $T_{3,1}$ have a larger relative abundance than the three others P_2, P_3 and $S_{1,3}$ in the network $G_{0.6}$ and $G_{0.75}$.

Then, we conclude that the number of the small fragments C_3, C_4, C_5 and $T_{3,1}$ in the network on Figures (4.7a) and (4.7b) are significantly larger than expected random, and they are considered as motifs.

4.3.4 Clustering coefficient and assortativity coefficient. The Table (4.13) summarize the average clustering coefficient both real-world network and Erdős-Rényi random graph, the Newman clustering coefficient and the assortativity coeeficient of the network $G_{2,0.6}$ and $G_{2,0.75}$:

	network based on $c = 0.6$	*network based on* $c = 0.75$
Mean clustering of the network $G_{2,c}$: $\overline{C}_{G_{2,c}}$	0.689	0.598
Mean clustering of the random network : $\overline{C}_{G_{ER}(n,p)}$	0.371	0.169
Newman clustering coefficient denoted by C	0.695	0.644
Assortativity coefficient denoted by r	0.352	0.355

Table 4.13: Clustering coefficient and assortativity coefficient of the network for $c = 0.6$ and $c = 0.75$.

We classify that the networks in Figures (4.7a) and (4.7b) are all assortative. For $c = 0.6$ and $c = 0.75$, we see in the Table (4.13) that the average Watts-Strogatz clustering coefficient and the Newman clustering coefficient of the whole real-world network are significantly larger than that of a random network of the same size. The high value of the Newman clustering coefficient indicates a large transitivity in the real-world network based on stock markets and the high value of the average clustering coefficient shows a large number of transitive relation for each given node i. Moreover, if we increase the threshold correlation, the average clustering coefficient and the Newman clustering coefficient decrease and they tend to zero if the threshold correlation tend to 1. The trend in percent of the network as seen on the variation of the average clustering coefficient is 10% for the real world network and 21% for the random one. The Newman clustering coefficient decreases 5% if we increase the threshold correlation. These all mean that 10% of the transitive relation of a given node disappeared, 5% of transitivity disappeared if we increase the threshold correlation in the real world network. This happen due to the degree of the relationship between stock prices.

4.3.5 Communities. We detect community by using Fiedler vector and the computation was done with python code.

For $c = 0.6$ and $c = 0.75$, the 100 stock markets of 100 companies divide in two communities

C1 and C2. The Tables (4.14) and (4.15) indicates the elements of C1 and C2:

$C1$	$SCOR, IFRX$
$C2$	$CAT, RY, FRSX, KO, GE, AMT, MMM, TWTR, NNDM, EBAY, PYPL, CMRX, MO, BAC, AGLE, WBA, PG, TWST, TMO, RTX, HD, FB, GS, MS, FNKO, UNH, RUT, NVCN, PFE, JNJ, CSCO NMRK, AAPL, FARM, MA, SCON, FordMotor, BMY, HDB, NMRD, DVAX, TMUS, GOOG, SDC, BBI, FDX, V, NVCR, FRTA, UPS, UHT, NEE, INTC, FNLC, MCD, FEYE, AXP, PM, IBM, DEO, NFLX, MSFT, N225, TSLA, RIO, FARO, SLAB, KLXE, JPM, TD, MCFT, SNP, CIIC, CVX, BA, FNJN, VZ, USB-PO, DIS, MRK, CVS, UVV, TRV, NKE, ATEC, LOW, XOM, BBQ, OPTT, IBKC, EXAS, PIXY, DHR, ATINC, EXC, WMT, TM, CVCO$

Table 4.14: Communities C1 and C2 and their elements for $c = 0.6$

$C1$	$AMT, TMO, UNH, JNJ, AAPL, MA, SCON, BMY, NMRD, TMUS, GOOG BBI, FRTA, UHT, NEE, INTC, MSFT, TSLA, CVS, DHR$
$C2$	$CAT, RY, FRSX, KO, GE, MMM, TWTR, NNDM, CMRX, MO, BAC, AGLE, WBA, PG, RTX, HD, FB, GS, MS, FNKO, RUT NVCN, PFE, CSCO, NMRK, FARM, FordMotor, HDB, SDC, FDX, V, NVCR, UPS, FNLC, MCD, FEYE, AXP, PM, IBM, DEO, N225, RIO, FARO, SLAB, KLXE, JPM, TD, MCFT, SNP, CIIC, CVX, BA, FNJN, VZ, USB-PO, DIS, MRK, UVV, TRV, NKE, ATEC, LOW, XOM, BBQ, OPTT, IBKC, EXAS, PIXY, ATINC, EXC, TM, CVCO$

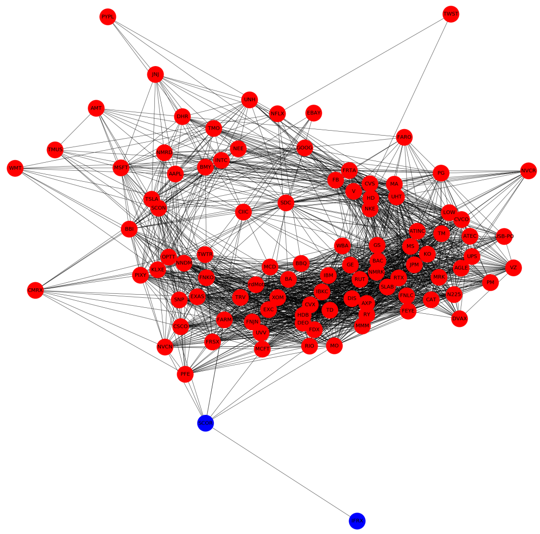
Table 4.15: Communities C1 and C2 and their elements for $c = 0.75$

The Figures (4.8a) and (4.8b) show these communities for $c = 0.6$ and for $c = 0.75$. The companies colored in blue indicates the member of the first communities and the companies colored in red the member of the second one. The companies colored in green is the companies in neither community. For $c = 0.75$, there are some companies in either communities are: eBay Incorporation, Twist Blosclence Corporation, PayPal Holdings Incorporation, ComScore Incorporation, Dynavax Technologies corporation, InflaRx N.V, Netflix Incorporation, Walmart Incorporation.

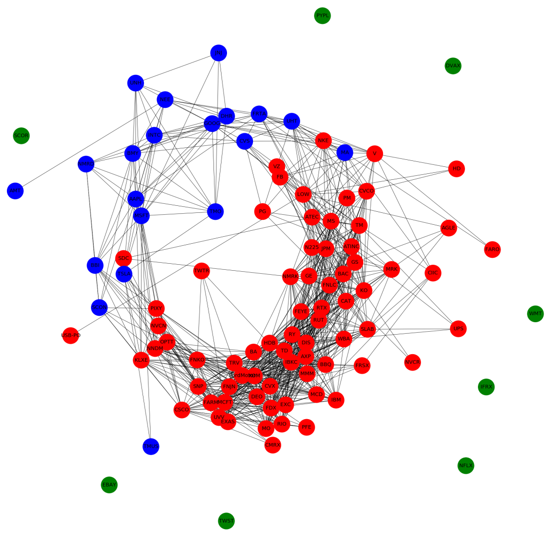
The network density for the networks $G_{2,0.6}$ and $G_{2,0.75}$ is summarized in the Table (4.16) below:

c	0.6	0.75
$\delta(G_{2,c})$	0.37	0.18

Table 4.16: Network density for each network $G_{2,c}$ corresponds for each threshold $c = 0.6$ and $c = 0.75$.



(a) Communities in the network corresponding to the threshold $c = 0.6$.



(b) Communities in the network corresponding to the threshold $c = 0.75$.

Figure 4.8: Communities with different threshold value for 100 nodes.

5. Summary and conclusions

In this essay, we have given a method for constructing financial network based on stock prices by using threshold correlation c, where $0 \leq c < 1$.

We have focused on the construction of financial network in which $c \in [0.6, 1[$. This means that we use strong threshold correlation if there is a connection between a pair of stock prices. Network analysis of New York stock exchange shows that a pair of stock prices are connected if the absolute value of the Pearson-correlation coefficient of these pair of stock prices is greater or equal to the threshold correlation.

We found that for each threshold value there is a unique network corresponding to it. So, the structures of our network depends on the threshold correlation. For instance, for $c = 0.6$, we get a disconnected financial network for 20 stock prices, while a connected financial network for 100 stock prices, this can be seen in Figures (4.5a) and (4.7a). Moreover, by varying the threshold correlation we get different structures of the network, this can also be seen in the Figures (4.5a), (4.5b), (4.5c), (4.7a) and (4.7b). Tables (4.1) and (4.10) indicate the number of the most relevant small sub-graphs in our network, the number of times that a small sub-graph appears in the network decreases if we increase the threshold correlation. This means that the number of connection between a pair of stock prices shrink if we increase the threshold correlation. So for instance, if we tend to one the threshold correlation the network associated to it tend to a null graph.

We found that centrality measure helps us to know which stock prices has an ability to communicate with other stock prices in the network, which stock prices are relatively adjacent to all other stock prices in the network, which stock prices is responsible for the communication between a pair of stock prices. Betweenness centrality identifies which stock prices constrain or simplify the information or communication between other stock prices in the network. The degree centrality and closeness centrality exhibit identical performance, with larger balance to systemic shape. Degree centrality and closeness centrality appear to be the most powerful measure, followed by eigenvector centrality.

We found that the triangles C_3, cycles of length four and length five, C_4, C_5, and $T_{3,1}$ appear more frequently than in the Erdős-Rényi random network. This can be seen from their relative abundance values in (4.4) and (4.12). Then, the number of triangles, number of cycles of length four and length five, and the number of tadpole are significantly larger than expected. So, we can consider the fragments C_3, C_4, C_5 and $T_{3,1}$ as network motifs in our real-world network of stock prices.

We found that the average Watts-Strogatz clustering coefficient and the Newman clustering coefficient decrease if we increase the threshold value. These mean that if we increase the threshold correlation, the number of transitive relation of a given node and the transitivity decrease. Moreover, there is a difference in the trends between the network with 20 nodes and the network with 100 nodes. To see this, for $c = 0.6$ and $c = 0.75$, in the Table (4.5), the network with 20 nodes disappeared 25% of the number of the triangle of a given node if in the Table (4.13) the network with 100 nodes 9% or 10% of its number of triangle associated of a given node. For $c = 0.6$ and

$c = 0.75$, we also found that 10% of the network transitivity disappeared for the network with 20 nodes, while 5% disappeared for the network with 100 nodes.

We also found that for $c = 0.6$, $c = 0.7$ and $c = 0.75$ we characterize that our real-world network is assortative because their assortativity coefficient are all positive. This assortative characteristic implies that low-degree nodes prefer to connect to other low-degree nodes, while high-degree stock prices are well related to other high-degree stock prices.

According to Figures (4.6a), (4.6b), (4.6c), (4.8a) and (4.8b), we saw that community detection gives us two different communities in the network, the first community colored in blue and the second in red, the stock prices in neither communities are in green. All members in a community receive the same information or communication like change in prices and the evolution of trade between companies. In other words, all members in a community have common properties. A stock is in a community if it is strongly connected with all other stocks in the community and they share the same feature like behavior and incomes. We also found that the structure of community depends on the threshold correlation. In fact, if we increase the threshold, the number of the members of the community decreases, this means that the network density decreases if we increase the threshold value. The number of member of a community shrink if we increase the threshold value for the network with 20 nodes, while the member of the community in the network with 100 nodes shrink and join on the other community if we increase the threshold correlation but if we tend to one the threshold value c we can find the same result for the financial networks with 20 nodes and with 100 nodes.

We conclude that the threshold correlation helps us to construct a network based on stock prices. Then complex network helps us to understand several problem in quantitative finance such as the pattern of links between movements in stock prices and the evolution of trade between companies. Using data from existing real-world network is more interesting than random network because we can find a larger number of connection between stock prices than in random network.

However, using threshold correlation and correlation matrix is just one method to construct network based on stock prices, there are more approaches that can be used to build a network based on financial entities as stock prices.

References

[1] J. Aswani. Impact of global financial crisis on network of asian stock markets. *Algorithmic Finance*, 6(3-4):79–91, 2017.

[2] A-L. Barabási et al. *Network science*. Cambridge university press, 2016.

[3] S. Battiston, J. B. Glattfelder, D. Garlaschelli, F. Lillo, and G. Caldarelli. The structure of financial networks. In *Network Science*, pages 131–163. Springer, 2010.

[4] E. Estrada. *The structure of complex networks: theory and applications*. Oxford University Press, 2012.

[5] E. Estrada and P. A Knight. *A first course in network theory*. Oxford University Press, USA, 2015.

[6] J-P. Onnela, K. Kaski, and J. Kertész. Clustering and information in correlation based financial networks. *The European Physical Journal B*, 38(2):353–362, 2004.

APPENDICES

Centrality measures for $c = 0.7$ and $c = 0.75$

Centrality	*values*
Degree centrality DC	[0.578 0.0 0.263 0.052 0.0 0.473 0.210 0.315 0.473 0.210 0.0 0.473 0.157 0.473 0.473 0.578 0.473 0.421 0.105 0.368]
Betweenness centrality : BC	[0.104 0.0 0.021 0.0 0.0 0.015 0.0 0.041 0.025 0.0 0.0 0.015 0.0 0.025 0.099 0.104 0.089 0.011 0.0 0.107]
Eigenvector centrality : EC	[0.351 8.06e-18 0.078 0.034 8.06e-18 0.318 0.159 0.150 0.144 0.317 8.06e-18 0.318 0.045 0.317 0.291 0.351 0.280 0.287 0.027 0.153]
Closenness centrality : CC	[0.641 0.0 0.408 0.328 0.0 0.538 0.408 0.481 0.538 0.374 0.0 0.538 0.336 0.538 0.518 0.641 0.561 0.499 0.328 0.499]

Table 5.1: Different centrality measures for $c = 0.7$

Centrality	*values*
Degree centrality DC	[0.473 0.0 0.263 0.0 0.0 0.421 0.210 0.263 0.421 0.157 0.0 0.421 0.157 0.315 0.263 0.421 0.473 0.210 0.052 0.210]
Betweenness centrality : BC	[0.098 0.0 0.093 0.0 0.0 0.027 0.0 0.058 0.069 0.001 0.0 0.027 0.0 0.0 0.016 0.027 0.205 0.010 0.0 0.011]
Eigenvector centrality : EC	[0.369 2.4e-18 0.091 2.4e-18 2.4e-18 0.358 0.208 0.136 0.338 0.097 2.4e-18 0.358 0.045 0.308 0.190 0.358 0.348 0.143 0.013 0.089]
Closenness centrality : CC	[0.538 0.0 0.408 0.0 0.0 0.493 0.370 0.438 0.493 0.338 0.0 0.493 0.311 0.455 0.370 0.493 0.563 0.382 0.275 0.394]

Table 5.2: Different centrality measures for $c = 0.75$

The link below shows the details of the Python code used during all computations:

https://drive.google.com/drive/folders/1eQU3F8fnB7meaenrM4-4qS0yBfvEr2xb?usp=sharing