7-2022

A New Approach for Analyzing Financial Markets Using Correlation Networks and Population Analysis

Zahra Hatami
University of Nebraska at Omaha, zhatami@unomaha.edu

A New Approach for Analyzing Financial Markets Using Correlation Networks and Population Analysis

Dissertation

Presented to the

College of Information Science & Technology

and the

Faculty of Graduate College

University of Nebraska

In Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy in Information Technology

University of Nebraska at Omaha

by

Zahra Hatami

July 2022

Supervisory Committee

Dr. Hesham H. Ali

Dr. Peter Wolcott

Dr. David Volkman

Dr. Lotfollah Najjar

A New Approach for Analyzing Financial Markets Using Correlation Networks and Population Analysis

Zahra Hatami

University of Nebraska

Advisor: Dr. Hesham Ali, Ph.D.

With the current availability of massive data sets associated with stock markets, we now have opportunities to apply newly developed big data techniques and data-driven methodologies to analyze these complicated markets. As stock market data continues to grow, analyzing the behavior of companies listed on the market becomes a massive task, even for high-performance computing systems. Hence, new big data techniques like network models are very much needed. We conducted this study on data collected from CRSP during the years 2000-2021 inclusively.

In this study, we proposed a novel population analysis by constructing a correlation network model based on the monthly data of different companies' excess returns; additionally, we employed the Louvain clustering algorithm to generate individual clusters/communities. After constructing correlation networks from input data, hidden knowledge was extracted from the network by using community detection and measuring network centralities. The Louvain algorithm was applied to the network as a data analysis shortcut tool and grouped different companies with high correlations or similar financial behavior over the period of study. In each community, different centralities were measured. Centrality measurements came from Closeness, Betweenness, and Eigen centralities for this study. The empirical result of this study showed that the

meaning of centrality measurement in network analysis in the stock market has a different meaning compared to social network analysis. In most networks, high central entities are the most important entities; however, in this study, we learned that high centrality is not something that researchers should look for when developing and building a portfolio with low risk. What was discovered was that nodes in the network with lower degrees of centrality led to developing a diverse portfolio with lower risk, with acknowledgment of the Modern Portfolio Theory.

Since this new approach was applied on the years 2000-2021, this study revealed behavioral patterns from stock movements depending on different events such as the 9/11 attacks, 2008 economic crashes, and the Covid-19 pandemic. As a result of this study, we would like to suggest a system based on a weighted portfolio to make a proper decision in selecting portfolios that can outperform the benchmark during normal circumstances or crises.

This Dissertation is dedicated to my loving and supportive husband and daughter, my brilliant father, mother, brothers and, to the memory of my beloved grandmother.

ACKNOWLEDGEMENTS

I would like to extend my appreciation towards all the people who have supported me during this endeavor. I would like to express my deepest sense of gratitude to my advisor Dr. Hesham Ali for his continuous advice and encouragement from the initial to the final level of my research. I am also grateful to all my committee members: Dr. Peter Wolcott, Dr. David Volkman and Dr. Lotfollah Najjar for their perspectives and thoughtful feedbacks both in my research and my professional career. I want to express my appreciation my appreciation from my deep heart to my husband Siamak and my daughter Asal for their love and continuous support. My very sincere thanks to my mother, my father, and my brothers Afshin, and Saeid, who were my greatest supporters.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

The Financial Market is a market in which people trade financial securities. Securities include stocks, bonds, and precious metals. The stock market is an equity or share market, which is an aggregation of buyers and sellers of stocks. In other words, the stock market refers to the collection of markets and exchanges where the regular activities of buying, selling and the issuance of shares of publicly held companies take place. In the market, different companies from different economic sectors trade stocks, with the purpose of gaining profit. The market mechanism and market behavior are major concerns for long-term investors because they desire to make the most profit out of their investment. In fact, the way the stock market works is an age-old problem that researchers have analyzed using different tools and machine learning mechanisms. Investing in the stock market has continued to represent major challenges for many investors around the world.

The business world can be unpredictable, and even the most experienced investors may struggle with decisive actions. This unpredictability stems from factors such as institutional and political constraints, the specifics of economic processes in each country, the accessibility of information (and information dissemination), and so forth. While all these factors are crucial in business making decisions, the way in which people and consumers perceive and take in information is even more critical [1]. Predicting the trends of various stock entities has always been a challenge for both investors and

analysts aiming to study these trends. These analysts are interested in how various factors cause investors to survive in the market, and how they affect market performance.

Knowing the impact these factors have will better the understanding of the market's mechanism, and therefore help predict future movement in the market. Finding the trend of movements requires investors to obtain knowledge regarding fluctuations and the reasons behind them. Various research papers have aimed to find the reasons behind the stability and instability in financial markets.

The behavior of financial markets is, in majority, based on two theories: market equilibrium and behavioral theory. The market equilibrium stands for the availability of information for every investor in the market [2]. The behavioral theory seeks the influence of the psychological process in decision making while investors interpret the market based on their information on making investment decisions [3]. Therefore, even though the information is available for everyone, investors' decisions are influenced by the behavior of other investors. These theories tend to work well under normal circumstances, but it is unclear whether they hold during unexpected situations or disturbing events like economic crash or Covid-19. Many unknowns are associated with the emergence of unexpected crises and their economic impacts. Hence, this brings new challenges for policymakers in their process of making appropriate decisions.

The effect of each crisis depends on the size of the impacted economy and its degree of vulnerability [4]. For example, throughout the economic crash in 2007-2008, primarily caused by the housing bubble, the financial stocks shot up because of the huge amount of fiscal monetary stimulus packages and other governmental support. During the

Spanish Flu pandemic, sectors related to entertainment suffered twice as much compared to companies in the health sector [5].

The 2020 pandemic brought society to an unprecedent point where people were required to work from homes and use online stores for purchases. As a result, companies belonging to specific sectors, depending on the nature of their business, were likely to be affected more than others. Under normal circumstance, consumers spend more in consumer discretionary products. On the other hand, during economic hardships, consumers tend to focus on spending more on essential products. However, the recent data from stock markets showed that consumer behaviors did not follow their usual pattern. For example, during Covid-19, stocks belonging to companies like Amazon went up and such stocks seem to benefit from the situation. Such profit did not appear to be related to any economic boosting. The response of the market seems to be less predictable and depends on specific conditions that society and consumers are faced with.

Attempting to understand the key aspects of stock markets, like when to buy or sell stocks, or which stocks are performing well/poor, remains extremely difficult. New approaches need to be explored to extract knowledge from available data and allow investors to understand the behavior of the financial markets under different circumstances. Although many studies have been conducted to address these issues, they have primarily focused on using statistical models. Fundamental analysis and technical analysis [6] are two ways to analyze stock markets from a financial perspective. Fundamental analysis aims to calculate the intrinsic value of the stock with the information of the company and the analysis of the financial statements, and to compare it with the current value to make the right decision about buying and selling the stock of

the company; technical analysis is applied to find the right time to buy or sell a share. Researchers have analyzed the financial market from a big data analytics perspective using various computational and statistical methodologies, including Artificial Intelligence, Machine Learning, Artificial Neural Networks, Fuzzy Logic, and Support Vector Machines [7].

The goal of analyzing data from financial markets is to find hidden patterns that reflect how companies perform for a given period and what the primary factors that impact their performances are. Due to the high volume of data and the complexity of the relationships between companies, big data techniques like graph models and network science are very much needed. With the goal of conducting correlation network analysis in mind, we can structure the data in such a way that facilitate finding common patterns or similar behaviors among companies [8].

This study introduces population analysis and demonstrates three different applications of stock market network analysis. First, similarity/correlation network and community detection algorithms were used to group the stocks that have similarities in their behaviors across different communities. Second, population analysis utilizing network properties and enrichment analysis helped recognize the reasons behind the similarities and differences between one community compared to others. The last application employed in-sample and out-of-sample analysis to utilize the structural stock market network to develop the portfolio selection process. This analysis considered datasets throughout different periods of time. Therefore, the study proposed a population analysis that used correlation network and graph properties and compared the result of the analysis against well-known benchmarks, such as the S&P 500. The main reason for

conducting this research is to use population analysis to optimize portfolios. The traditional approach to optimizing portfolios uses statistical methods to calculate the variance of prices and returns on the selected securities of the portfolio. In this manner, portfolio selection and strategic planning are based on investors' goals, vision, timeframe, and the amount of risk that investors are willing to take. Our proposed approach used population analysis utilizing network analysis and its specific properties such as Eigen centralities, Closeness centralities, Betweenness centralities, and enrichment analysis to allow investors to build their portfolio based on a high-yield strategy with proportional risk.

In the following section, problem statements, identified gaps, and terminologies that were used in this study will be explained.

1.1 Problem Statement

This proposed research is to create a more robust correlation network model that will identify the significantly enriched input parameters and compare various groups of stocks with different performances using population analysis. Previous studies do not show how network and population analysis can help investors build their portfolios. Several gaps were identified from previous studies regarding stock market analysis.

1- Comprehensive studies have not been conducted to see what type of stocks are suitable for investors' portfolios for various sectors of the United States stock market. Up to now, studies have focused on different machine learning techniques to predict the future movements.

2- A comprehensive study is required to check companies' behavior in different sectors during different time periods to find patterns for all the condition (normal circumstances or crises) in the United States stock market. So far, studies focus on financial perspectives and statistical approaches and there is a lack of big data analysis in financial domains.
3- From previous studies, the significant parameters were not recognized based on overall stock performance and specific time series analysis.
4- There is a lack of robustness analysis on the correlation network models.
5- So far, there is no correlation network model that can recommend a portfolio based on normal and unexpected events that could enable investors or even ordinary people to identify which companies/stocks need to be selected in time based on their behaviors regarding their size and sectors.
6- A better and suitable/efficient clustering algorithm for correlation network models on financial markets have not been identified yet by comparing various clustering algorithms on finance databases.

To overcome the gaps above, we completed a comprehensive study of the US stock market from the CRSP dataset to analyze the groups/clusters of stocks using CNM and population analysis. At the end of this study, we conducted a recommendation system that will enable investors select their portfolios based on the results of this study.

All the gaps above could be overcome in different phases as below:

Goal 1. Construct a comprehensive study: For this research, the comprehensive literature review was conducted. The most relevant studies related to financial and statistical perspective were collected. Limitation and future works in those studies led to a more

advanced approach by applying the new methodologies. This study attempted to overcome the limitations from previous research and filled the gaps in the studies as much as possible.

Goal 2. Construct the Correlation Network (CNM): After conducting the comprehensive literature review and recognizing the gaps, a CNM was applied on the dataset collected from CRSP and Fama & French association. The outcome for the CNM was a list of the most correlated companies clustered into different groups. Stocks for potential portfolios were recognized based on different centrality measurements and portfolio selection formulas. Finally, a list of stocks in different sectors was compared with the S&P 500 benchmark for future decisions.

Goal 3. Construct Case Studies: Various case studies were conducted as part of this research. For example, temporal analysis was applied to the dataset to determine whether and how companies' behaviors change over time. From different case studies, the list of companies that outperformed in different time periods were recognized. The Louvain clustering algorithm was applied on the CNM to find the most correlated stocks in different communities. To have robustness analysis and avoid sample bias issues, different datasets were designed as in sample and out of sample and networks were constructed based on different correlation coefficients. The purpose of robustness analysis was to show that obtaining significant parameters is not dependent on the correlation coefficient. A three-step validation process was utilized in these phases. Step one was cross checking the result with previous studies; step two and step three was discussing the results with the experts in the financial domain to take their opinion and know the correct meaning of the produced results and new knowledge.

Goal 4. Construct the Pipeline: Creating the pipeline was one of the main goals of this research. It means this research tried to come up with a set of codes that could be used on different databases extracted from financial markets and produce the final outcomes in terms of significantly enriched parameters and performance plots for comparison of all the significant clusters. R programming language was used in this study to create this pipeline.

Goal 5. Construct Decision Support System: A recommended portfolio in terms of a decision support system was the outcome of this research. For this to happen, different packages and functions were used from R dashboard to evaluate the weight of stocks in the potential portfolio. The final evaluated stocks shaped the candidate portfolio that could outperform the benchmark and have a higher Sharpe ratio.

1.2 Business Terminology

Excess Return (ER): The return of investment in securities or portfolios that outperforms a benchmark or index with a similar level of risk is called excess return. These returns are widely used to measure the added value created by a portfolio or investment manager, or to measure management's ability to overcome the market. This value obtains with subtracting companies' return from market's risk free.

Portfolio: A portfolio is a set of different securities owned by an investor. The return on investments in a portfolio will be equal to the average return on that portfolio; however,

portfolio risk is often lower than the average stock risk in the portfolio and its amount depends on the interaction of events on the stocks in the portfolio[1].

Sharpe Ratio: Sharpe ratio is a measure of return on credit. This ratio was created by Nobel laureate William F. Sharpe (William F. Sharpe). Sharpe ratio is the return on excess of the exchange rate without any unit of volatility or total risk.

1.3 Network Terminology

Betweenness Centrality is the indicator of centrality in a graph based on the shortest path. Betweenness centrality is a way of detecting the amount of influence a node has over the flow of information in a graph [9].

Closeness centrality is the sum of the total distances from one node (v) to all other nodes in a network [10].

Eigen centrality measures a node's impact based on the number of links to other nodes in the network. The Eigen centrality reflects the importance of nodes connected to the current node because not all nodes are equivalent [11].

1.4 Organization of Dissertation

This dissertation is structured as follows. Chapter 2 will provide a literature review about correlation networks, works related to network analysis in the financial market and other domains, network analysis with emphasis on crises and network properties, and the

[1] https://www.lehnerinvestments.com/en/portfolio-risk-measure-manage-investment-portfolio/

theoretical framework behind in financial domains. Then in chapter 3, our proposed comprehensive conceptual methodology framework will be discussed. Chapter 4 will explain how we addressed research questions. Chapter 5-9 contains different case studies that were designed to answer research questions in different scenarios. Chapters 5-9 can be comprehensive chapters by themselves. Those chapters contain case studies and their building blocks including an introduction, datasets, methodology, and outcome. Chapters 5-9 paint a picture of how the model was developed overtime. Chapter 10 will address the robustness analysis for our research. The last chapter, chapter 11, will address the discussion of the results, the limitations involved with this work, and how said limitations can be addressed in future studies.

CHAPTER 2 LITERETURE REVIEW

Analysis of a network involves the recognition of which entities are connected to others in a network diagram. The entities and how they affect each other are important in finding the most influential entities in the network. A correlation network model (CNM) is built by groups of nodes (vertices) and edges in the graph model. Two nodes are connected by a directed/undirected edge if and only if their correlation coefficient is 0.75 or more. CNM is relevant, as the highly correlated nodes or nodes with dense connections would give us the information about the nodes with the same kind of behavior or characteristics. A regular network contains several components that interact with each other. The analysis of a network is the recognition of which entities are connected to others in a graph.

The strength of the relationships (i.e., the strength of the edges) can be measured with different kinds of correlation coefficients such as Pearson, Kendall, and Spearman. The coefficient of continuity is always between -1 and 1. A correlation coefficient above 0 means that there is a positive correlation; the closer the coefficient to +1, the stronger the positive correlation. A correlation coefficient below 0 means that there is a negative correlation between the two variables, and the closer the number is to -1, the stronger the negative correlation. Various correlation coefficients can be used as measurements depending on the nature of the data that creates the correlation network. For example, if the data are normally distributed, the Pearson correlation coefficient will be used to establish the edges' connections. A Pearson correlation-based network is an unweighted/undirected network that constructs and measures the network using Pearson correlation coefficients.

2.1 Different Types of Correlations Coefficients

Pearson r correlation: Pearson *r* correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables. For example, in the stock market, Pearson *r* correlation is used to measure the degree of relationship between the two stocks that are related to each other.

Kendall rank correlation**:** Kendall rank correlation is a non-parametric test that measures the strength of dependence between two variables.

Spearman rank correlation**:** Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. The Spearman rank correlation test does not carry any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal. With company size, then we can recognize other factors related to the companies for further analysis.

2.2 Correlation Networks in Various Disciplines

Different research domains such as social computing [12], biosciences, and civil engineering used correlation networks in their analysis [13]–[19]. The theoretical framework behind the correlation networks can address different problems in different domains. Another benefit of using the correlation network is extracting hidden information by using advanced visualization tools such as clustering [20]. In bioscience, using

clustering methods on the complex network created based on biological entities could help researchers recognize active genes associated with various stages of disease progression [15]. The outcome of the research could detect different early health conditions in different patients. They also allow researchers to utilize advanced visualization tools at different granularity levels [20], [21]. Recently, researchers used the correlation network in the civil engineering domain to assess the safety, deterioration, and performance of bridges and their infrastructures [13], [16], [19].

2.3 Analyzing Financial Markets: Different Perspectives

In a big data analysts' perspective, researchers have analyzed financial markets using various computational and statistical methodologies, including Artificial Intelligence, Machine Learning, Artificial Neural Networks, Fuzzy Logic and Support Vector Machines [22], minimum spanning trees (MSTs), and Planar Maximally Filtered Graphs, which are a topological generalization of the MSTs [23]–[25]. Additionally, large and complex daily and monthly reports produced by financial markets can benefit significantly from utilizing big data analytical tools such as network models and correlation analysis. A correlation network model (CNM) can be built by assuming that each company/stock is a node (a vertex) in the graph model, and two nodes are connected by a directed/undirected edge if and only if their correlation coefficient is above a particular threshold (such as 0.75 or above). The correlation network in the financial domain can be created based on different input variables such as prices or companies' returns. In [23], [26], based on the similarities in daily prices in different stocks, researchers built a

correlation network in which nodes represented equities in financial markets and the relationship between them as edges based on time series data. They found out that only a small number of stocks held influence over most stocks, and stocks in different economic sectors were dominated by the finance sector. In the network, companies are represented by nodes, and similarities between companies' prices are represented by edges. The result of the study revealed that finance sector[2] dominated the market compared to other economic sectors[3]. In other studies [26], [27], researchers established the network as a worldwide network based on a few entities such as price indices, companies, and stocks.

The correlation network analysis brings opportunities to the researchers to uncover hidden information from the network. For example, in [22], researchers could predict the next day's events using time series data in the correlation network using Hidden Markov Models to forecast the stock market. The result of different studies using big data techniques showed that governments could make use of the results of correlation networks while making economic decisions [28], [29].

In the big picture of acting correlation networks and how these networks are useful, [30] prove that the results of a correlation network can help governments to make wise economic decisions. There is a possibility of creating different sizes of network complexity. Researchers could build a small network constructed from just a few entities, such as price indices, companies, and stocks, or construct the network as a worldwide network [26], [27]. Analyzing the financial market with a correlation network and applying different machine learning techniques will release hidden information from the network.

[2] A finance sector is a group of companies that fit into one of the main categories such as banks and financial services

[3] Economic Sectors are large groups of the economy that categorized according to their place in the production chain

The behaviors in the financial market can be analyzed from a financial perspective through fundamental analysis and technical analysis [31]. Fundamental analysis is applied to find the intrinsic value of a share, and technical analysis is applied to find the right time to buy or sell a share.

There are also different statistical approaches to analyze the stock market (financial market behavior) [22], [32]. Recently, there has been a general inquiry as to whether a virus, similar in fashion to the novel coronavirus, has also affected financial markets in the past. Past researchers investigated the impact of the 2003 SARS epidemic on the economy and did not find any conclusive evidence showing that the epidemic had a large effect on the economy. They conducted a study using a non-parametric Mann-Whitney test and heteroscedastic t-test to examine the impact of the SARS epidemic on eight countries' stock market returns [32]. The analysis did not prove any effect from the SARS epidemic on the stock markets of those countries. Researchers even discovered that there was out-performance in stock market indices during the SARS period. However, their analysis proved that stock indices were affected negatively by Covid-19 in those representative countries [32]. The study applied the same statistical approach as the SARS pandemic investigation (Pooled OLS regression, conventional t-test, and MannWhitney test) on weekly panel data for Covid-19 to find the impact of the pandemic on the stock markets of sixteen countries. The test result showed that when human-to-human transmissibility had been confirmed, all the stock market indices and investors' behaviors negatively reacted to the news in both shorthand long events [32]. The study showed that when the number of Covid-19 weakly cases went up, stock markets return go down.

2.4 Network Analysis on Financial Markets with an Emphasis on Crises

The concept of network analysis is a growing topic of interest among many researchers across the globe. Through its devices, network analysis may be used to understand complex phenomena and systems in various fields of work. Such phenomena come in different forms, seeming to not even correlate to one another, displaying the wide range of applications available for network analysis. Disease virality, human information, infrastructure, and risk estimation in a financial context are all fields and phenomena where network analysis may come into play [33]. In biological studies, correlation networks come into play as an able tool, creating a way for vast sets of data to be input for analysis [20]. Resourceful insight can be extracted from network models displaying complex relationships through nodes and edges. A model reflecting the prior statement displays the interactions in a system, as well as complex relationships. For instance, microbial interactions in models of bacterial communities could highlight the relationship and dynamic between microbes [17]. Similarly, the functionality and ways of market dynamics, a focus of this literature, is also a prime example of the way information may be garnered from network models [33].

In the stock market, an established network can be used to observe and make predictions based on economic factors due to the cascading behavior of stocks [34]. In simpler terms, negative stock performances in communities can spread and impact a whole network [35]. Emphasis on researching relationships in the financial market is partly due to historical and even contemporary events, like crises, that have significantly impacted markets. Subsequently, the literature delves into the impact of the Great Depression and its

contribution to future market instability[36]. The collapse of 1929 highlighted the instability, as well as corruption, present in markets and proved how an economic crisis could have a major impact on social unrest and political discourse [36]. Additionally, periods of economic recession, like the 2007 crisis in the United States, lead to instability and consequences that hardly could be recovered from [37]. Such instability has made it so that even homeownership is not free of risk due to the mortgage and labor markets. Consequently, the increasing risk of homeownership contributes to economic and market instability [38]. The decline of housing prices in 2008 led to a weakening of the global economy, causing major losses to many financial institutions that were ruined or had to go through internal reconstruction. The crisis even called for some governments getting directly involved in response to the collapse of many financial institutions. Instability in one economy is bound to affect others' depending on common factors; this is shown as events occurring to the United States' economy are bound to impact India and the European Union's economy [39]. Similarly, in the Asian financial crisis 1997, the collapse of the Thai currency ended up having implications on Japan, South Korea, Indonesia, and other Asian countries [37]. Due to such demand, many researchers have turned their attention to calculating risk and recognizing the signs of an oncoming crisis [40]. Importantly, the onset availability of data provides great insight to investors competing with corporations in the business sector.

There is no doubt that major activities in the stock market have an impact on the economy. Researchers study the stock market from different perspectives; some studies have tried to analyze the market theoretically, and others look at the market empirically. Such studies could have been conducted with statistical analysis, big data overview,

financial perspectives, etc. The purpose of these studies is to deeply analyze the market to find the patterns behind the activities, predict the future movements, collect enough information in different periods, and to design proper strategies for any unexpected events to help the market recover as soon as possible. Therefore, researchers have tried to conduct different studies applying different methodologies to clarify the effect of different movements either in crises or booms. In [41], the reported study focused on analyzing the impact of the market theoretically and empirically. The result of the analysis showed that during the 1968-70 recession, there was 7% capital loss, 7% in nominal value, and 19% in real terms. Regarding the effect on economic sectors, corporate shareholdings declined by 14% [41].

The United States economic crash in 2008 was the result of a false housing boom, risky financial manipulations to grant high-risk mortgage loans, and lack of regulation of stock traders and speculators searching the capital market. The impact of this crisis brought major issues in different sectors: many lost their jobs in the housing sector, the financial market froze, and widespread economic recession occurred. Banks and financial institutions were forced to suspend lending, disrupting the economic cycle in many countries. Following these events, the world financial markets collapsed, the housing market stagnated, and the banking system's problems continued for years. The result of the study[42] showed that besides major economic sectors, such as the financial sector and the housing sector impacted by the crisis, the utility sector had been impacted on three fronts: financing, demand, and expansion. Even though it has been 13 years since the 2008 crisis, researchers are still trying to find similarities and differences in economic sectors' behavior with other crises such as the Great Depression, past pandemics (SARS), and recently the

Covid-19 pandemic. Another relevant study pointed out that studying simple mathematical models for economics is not enough to prepare for potential corrections to deal with future economic crises [43]. For example, the world was not well prepared to face the consequences of Covid-19 and its wide range of implications. Further, the study argues for the need of economists to work along with social scientists to better understand and face future economic challenges. The researchers in [44] compared various stock market crashes that happened in the 20th century and concluded that focusing on the stock market is not always a good indicator of economic health, nor is the success of the stock market is an indicator of financial stability.

Another study of the 2008 great crash was reported in [45]. The study concluded that it would be a geopolitical setback to the United States. Other studies highlighted that banks needed to quickly respond to market fluctuations [44], [46].

Since one of the significant ways to invest in the United States is the stock market [47]. Researchers have studied the behaviors in the stock market during unexpected events like the SARS pandemic and Covid-19 pandemic to gauge public reaction. The researchers applied a non-parametric Mann-Whitney test and heteroscedastic t-test to examine the impact of the SARS epidemic in 2003 on eight countries' stock market returns. The study showed that there was no significant effect on the economy due to SARS. However, the analysis of the stock market data during the 2020 pandemic demonstrated that stock indices were negatively affected in different countries [32]. Similar to the previous study on the SARS pandemic, researchers applied the same methodology on the weekly panel data to Covid-19 data to find any significant impact the pandemic had on the stock market in

sixteen countries. When the pandemic started and the time that different health departments verified human-to-human transmission, the stock market had both a short-term and long-term adverse reaction[32]. As of February, and March 2020, S&P 500 lost 1/3 of its value, and a group of researchers conducted a survey to measure respondents' expectations and households' financial prospects. The goal of the study was to find the impact of the 2020 pandemic on the households' decisions and how they adjusted their plan for future investments, savings, and spending [47]. The survey was conducted on the aggregate equity market and dividends to investigate how expectations change in a crisis, such as the pandemic. The result of the analysis showed that the investors' behaviors were equally likely to evolve during the crisis and Giglio et al. (2020b) documents that "while respondents on average downward revised their short-run expectations about stock returns and GDP growth during the crash, they remained optimistic about the long-run outlook, and that disagreement across investors increased over the crash" [47].

2.5 Community Detection

Correlation networks represent a powerful analytical tool to address complex problems due to their ability to integrate different data types and benefit from a wealth of theoretical concepts obtained from graph theory and network science. They also allow researchers to utilize advanced visualization tools at different granularity levels [20]. For example, in bioinformatics applications, correlation analysis has been successfully applied to measure changes in temporal relationships among biological elements, which leads to

the early detection of several health conditions. Along with the availability of more data, the clusters extracted from such networks can allow researchers to identify groups of active genes associated with various stages of disease progression [15].

Since the correlation network is identified as a powerful method to apply for complex datasets; accordingly, the researchers need a community detection method to extract the knowledge from the network. Different approaches such as community detection and applying network properties can be applied on the correlation network models to extract the information from it. In this regard, different clustering algorithms tested from researchers for different correlation networks structures and efficient algorithms identified for each domain. For example, the Markov Clustering Algorithm (MCL) given in [48] has been shown to be the most efficient compared to various graph clustering algorithms and Louvain algorithm was identified as another efficient data analysis shortcut in financial domain.

2.6 Network Property: Emphasizing on Centrality Measurements

In the case of network analysis, generally, stocks are represented by nodes, and edges or links represent returns. In hindsight of economic crises, stocks' behaviors are difficult to analyze due to complexities in networks. Through network analysis, researchers can determine network centrality, aiming to determine the significance of nodes in a network [49]. Adopting centrality measurements, as the literature states, could be key to determining the importance of stocks. Of course, grasping the concept of centrality and its means of measurement is crucial in building an understanding of nodes in a network of the

financial markets. As such, the literature roughly defines the means of measuring centrality as degree centrality (with focus on eigenvector centrality), Betweenness and Closeness centrality (central metrics).

Eigenvector centrality (degree centrality) stresses the significance of certain nodes in a network over other present nodes. In short, it determines that not all nodes are equally important in a network. Edges are crucial in assigning values to nodes in eigenvector centrality, meaning that the number of links a node has contributes substantially to its significance. However, a node with fewer edges could also have high centrality depending on the importance of said edges and which node they link to the analyzed node (central metrics). This is the reason that eigenvector centrality branches off degree centrality, which counts the edges connected to a node, therefore determining its significance in a network [50].

Closeness centrality creates a different approach to quantifying nodes. In closeness centrality, the importance of a node is determined based on its access to other nodes in a network; greater values signify greater importance as closeness centrality is determined by the reciprocal value of the average lengths of the shortest links between the given node and all other nodes (central metrics). Closeness centrality can determine how much a node can impact other nodes in a network since a node with high closeness centrality has more access to the other nodes in a network, and thus, more influence [50].

Betweenness centrality is the indicator of centrality in a graph based on the shortest path. Betweenness centrality is a way of detecting the amount of influence a node has over the flow of information in a graph [9]. It expresses the relative importance of a node by using the shortest path that node has created between other nodes.

Nodes can represent stocks and inside a network, and, depending on their centrality, they can be picked out to create an optimized portfolio depending on factors such as portfolio size. Therefore, it can be concluded that through network analysis of characteristics such as centrality, an optimized portfolio could be created to maximize gain in the stock market [35]. In related literature, the subject of study regularly makes use of data and networks with high centralities (i.e., [35]); however, other literature contrasts that statement and states that highly central nodes involve financially riskier investments [49]. Said literature reiterates that favorable portfolios involve heavily investing in low-central securities [26].

2.7 Theoretical Framework

2.7.1 General Theories

The evolution of unexpected crises and their economic impact is not clear. So, it brings challenges for policymakers in their process of making appropriate decisions. There are different theories behind the behaviors in the financial market. The common theories are the Efficient Market Hypothesis (EMH), the behavioral theory, traditional and modern portfolio theory (MPT). The EMH stands for the availability of information for every investor in the market [2]. The authors of [51] defines the information in the market into three different categories: weak, semi-strong and strong. Weak refers to information about historical prices, semi-strong is about adjusting prices with previous public information, and strong refers to monopolistic access to information from investors. Behavioral finance dissects the way in which an investor makes their decisions, which in turn affects the

market. While it is true that people have a sense of reason and rationality, behavioral finance assumes that this rationality acts within limits. Behavioral finance can be split into two different subcategories: micro behavioral finance and macro behavioral finance. Both assume that investors make decisions based on their own psychological biases, thus causing their decisions not to be the best always. Such decisions may cause anomalies, or disruptions, in the market, which could have destructive effects on many individuals' financial health, as well as the economy's financial health. As such, these anomalies must be stopped with the help of behavioral finance by understanding the psychology of the market and its investors as to make decisions to counteract the anomalies. Micro behavioral finance attempts to analyze behavior, and macro behavioral finance discloses and describes anomalies of the EMH that could be explained by models of people's behavior [1], [52]. The behavioral theory seeks the influence in the psychological process in decision making while investors interpret the market based on their information for making investment decisions [3]. Therefore, even though the information is available for everyone, people's decisions are influenced by other investors' behavior. Accordingly, this research tries to analyze the companies' behavior in financial markets considering behavioral theory and EMH.

2.7.2 Portfolio Theories

A portfolio, known as a portfolio of assets or an investment, is a combination of diversified assets that can include investing in housing, banks, stock exchanges, coins, currency, gold, and so on. So far, two theories to building a portfolio have been adopted: the traditional and modern theory (Curtis, 2004). The traditional approach implies that all

investors should have a personal portfolio that is unique and tailored to their needs [53]. This means that investors need to estimate the yield on the securities they intend to invest in before making their portfolios. Then, after estimating the yield, they select the securities that are expected to have the highest returns in the future for investment. American economist Harry Markowitz criticized the traditional portfolio theory [54]. He believed that it was almost impossible to find a share with the lowest risk and highest return, and that if people wanted to build a good portfolio, they need to find a balance between risk and expected return [55]. Markowitz believe that people should not only measure the risk but return of an asset for investing led to Modern Portfolio Theory. The theory was formed on the assumption that investors are inherently risk-averse, but their ultimate purpose is different [56]. In a 2004, study, Roy tried to provide a practical way to determine the best amount of interaction between risk and return. According to Roy, investors initially sought to preserve their original capital, then they would consider the minimum rate of return for their capital [57]. Therefore, they tried to avoid selecting the stocks/assets that had high deviations in their returns. Researchers in 1994 tried to evaluate investment funds' performance [58]. They used the word undesirable deviations for funds with a rate of return below the target rate. Their analysis of monthly data for the previous ten years in December 1992 for two mutual funds and six stock market indexes proved the usefulness of risk to be undesirable in evaluating the performance of capital funds [58]. Designing a framework of eighteen retirement funds for a performance assessment based on Sharpe ratio showed that undesirable risk in performance evaluation of assets is much better than focusing on the returns' standard deviation [59].

The question now is how do investors optimize their portfolios for the highest expected return with different levels of market risk? Modern Portfolio theory holds that there is no such thing as a full investment. What is important and should be considered is choosing a high-yield strategy, along with proportionate risk. Modern portfolio theory argues that individuals can design an ideal investment portfolio that maximizes returns by considering the optimal amount of risk. By investing in more than one share, the investor can gain the benefits of diversification while reducing their risk. To calculate portfolio risk, the variance of each asset along with the correlations between each asset pair can be calculated [60]. The correlation between assets, the percentage of investment in each asset, and the number of different stocks in which they are invested affect the total portfolio risk [60]. By building a diverse portfolio, the risk of high-risk assets will decrease by adding low-risk assets to the portfolio. In fact, by adding securities such as treasury securities and units of investment funds, the risk of the entire portfolio can be reduced.

According to this theory, the risk per share consists of two types. The first type is the systematic risk or market risk that cannot be eliminated (such as recession, changes in interest rates on bank deposits, etc.). The second type of risk is an unsystematic risk (risk per share that may result from poor management or sales) [61]. In fact, diverse portfolio is a model for the optimal allocation of an individual's wealth invented between risky assets. This model was focused on only the two factors of expected return and variance. In a 2002, study, Lien examined the relationship between risk and return in investment and creating portfolios. He pointed out the issue of investing. He mentioned that investing in financial institutions in the form of a portfolio needs a precise evaluation of the portfolio performance from several different indicators (i.e., return, share ratio). Moreover, with a

market focused largely on risk management due to unpredictability, equity sectors can make an effective addition to portfolios as they can increase investment rates and lower risks simultaneously. In accordance with MPT, assets that are not in direct competition with one another (meaning they do not correlate) help lower the risk in portfolios and can create opportunities for higher return rates. However, that does not mean it is a guaranteed fact as many factors come together to create an optimal portfolio [26], [62]. As the previously discussed literature states, lower centralities in networks (defined by nodes) could be the key to build an optimal portfolio to investors [49].

2.8 Shortcomings

From previous studies, we learned that there is a lack of big data analysis predicting the financial market movement. Current approaches are not as developed, and the quality of the results would not improve significantly with the addition of new data. To the best of our knowledge, there is a lack of comprehensive studies finding approaches to help investors build their portfolios. In this regard, constructing a portfolio based on computational analysis, such as correlation network and graph property, is needed. Analyzing the stock market is emerging to be a big data problem because it exhibits all the components of big data. There is a large amount of data with a variety of data types that need to be processed and stored. Therefore, there is a need for more data collection mechanisms.

In order to overcome the shortage in big data analysis in the financial domain, we designed a study using a combination of computational analytics techniques. The proposed method benefits from the availability of big financial data collected over many

years. This research is a combination of big data techniques since constructing correlation networks to find hidden information from networks and to find the market trend is a complicated process that cannot be achieved with traditional approaches. Moreover, revealing information from networks is a complex task. Clustering was one of the key tools we used to extract information from the network. Data clustering is an NP-hard problem that involves computationally intractable and heuristic problems. Since this study used clustering algorithms to extract a smaller group of companies with the highest degree of similarity, finding companies with similar excess returns turned into a heuristic process.

In this research, we aimed to conduct a comprehensive study of the US stock market using a novel population analysis approach. This proposed method aimed to study various events and see what sectors are resilient to said events, along with why and how they bounce back very quickly. Identifying these sectors could be an important source of guidance for investors to predict patterns in the stock market for financial gain, even in worst case scenarios. We want to enable investors to select their portfolios based on the results of this study and our recommendation system.

CHAPTER 3 METHODOLOGY

With the availability of massive data in the form of daily, weekly, and monthly in financial markets, researchers need to use powerful tools and techniques to get insight into data. Therefore, analyzing the behavior of companies listed on the market becomes a massive task, even for high performance computing systems. Hence, new big data techniques like network models are very much needed.

We proposed the use of population analysis and the employment of similarity network models as an objective measurement of the performance of each stock by showing the level of similarity among a population of stocks with regards to their ER over a period of time. The notion of population analysis employs correlation networks as a modeling tool and enrichment analysis as an analytical tool to reveal apparent and hidden patterns associated with the input data. The objective of population analysis in this study is to compare different communities extracted from the network in respect to a particular outcome measure, then identify which input features are enriched in this community. Because this study seeks to find the best portfolio as a diverse basket and the best match with famous indexes such as S&P 500, population analysis as the comparison tool is necessary.

3.1 Population Analysis

The term “population analysis” in the correlation networks analysis refers to comparing the group/cluster of nodes/companies with respect to various parameters.

Some parameters may be highly enriched in one cluster compared to another cluster. Applying a novel population analysis helped us compare individual data points with other data points in different clusters or populations regarding different performance levels. Therefore, population analysis allowed us to compare two or more clusters of companies with respect to one or more enriched parameters. The results of this analysis allowed us to discover the parameters that significantly affected the cluster. In other words, population analysis is the process of comparing and finding the various clusters to see how they are enriched with different parameters. In this type of analysis, depending on correlation network model (CNM) criteria and communities, different clusters were compared based on their properties to find the reason behind their enrichments. Further analysis such as portfolio selection process helped to identify the portfolios contain the stocks that can outperform the benchmark.

Accordingly, by applying population analysis on community detections, network properties, and portfolio selection process, different portfolios were identified as potential portfolios within the whole dataset. By utilizing all these processes, starting with population analysis, creating CNM, and ending with enrichment analysis, we were left with a small group of stocks from which to choose. Selective stocks could even be broken down further to find the stocks that are the strongest of the strong.

3.2 Network and Community Detections

Stock returns and their changes are among the most important factors in assessing the economic value of a company in the stock market, which reflects the investment decisions of individuals in the economic environment. Stock return changes are not

independent of each other. Hence, studying the correlation of stock behavior changes provides investors with a greater understanding of market performance. Stock market analysis based on networks provides a study of stock returns' correlations.

In the large data domain, researchers can use network analysis to create a correlation/similarity network to extract underlying information. The similarity networks were represented using graphs where stocks were the nodes. Two nodes were connected by an edge if the corresponding stocks' scores had a similar excess return pattern over the predefined period. We also utilized graph-theoretic mechanisms to zoom in and out of the population networks and extracted different types of information at various granularity levels. The use of network models also has the added feature that the quality of the analysis improves with the addition of new data, an important feature that traditional approaches, which focus on the assessment of each individual stock, do not have. This approach was particularly effective in analyzing temporal data or data collected for the same stocks over a period. The obtained networks can highlight consistencies as well as changes associated with certain groups in the financial market, which can be used in prediction analysis for planning purposes.

After constructing correlation networks from input data, hidden knowledge was extracted from the network by using community detection and measuring network centralities. Identifying communities containing highly correlated stocks provided information that could be used along with network properties, such as centrality measurements, to identify optimal portfolios. The Louvain algorithm was applied to the

network and grouped different companies with high correlations over the period of study [76] .

3.3 Enrichment Analysis

The notion of enrichment analysis as a useful technique under population analysis can provide valuable insight into the group of communities that contain different stocks. Enrichment analysis in the sense of knowing which specific parameter makes the community noteworthy among the rest of the communities. After constructing correlation networks and detecting the communities, to identify key stocks' parameters and find the significant parameters, enrichment analysis was employed on each community. In this study, the outcome parameter was the excess return (ER), and the input parameter was the sector and size of the companies. To find the significantly enriched sector/s and sizes in each community with respect to the remaining communities, we applied different correlation matrices between stocks in each community and compared the results.

3.4 Portfolio Calculation and Sharpe Ratio

Different graph properties could be applied to the network's data, and it would be helpful if the analysis relies on those well-established metrics. The notion of centralities is one of the well-defined measurements in network analysis. The centrality measurements are predefined for both nodes and edges. The meaning of the centralities is dependent on the structure of the network. For example, in social networks (friend networks), the higher centrality for a node represents how important that node is compared to others with low centrality. This study attempted to build a diverse portfolio containing the companies with similarities in their ER. Therefore, we counted and interpreted the centrality measurements

as connections to other companies. The notion of Eigen centrality measures a node's influence based on the number of links it has to other nodes in the network. In other words, Eigen centrality counts the number of links derived from a node while all links do not represent the same importance. Eigen centrality can be considered as an expansion of the degree centrality metric. The Eigen centrality reflects the importance of nodes connected to the current node, which means not all nodes are equivalent. Regarding the importance of nodes, in this study, the Eigen centrality displayed a value representing the connection strength among nodes; a higher value did not necessarily indicate a more critical node to be counted in portfolio selection, and a node with low Eigen centrality value could still be a good candidate in portfolio selection. Betweenness centrality is the indicator of the amount of influence a node has over the flow of information in a graph. Another notable measurement of centrality is closeness. Closeness centrality for one node can be calculated as the average distance of all distances from this node to all other nodes in the network [75]. Some scholars define closeness as the inverse of this average so that larger numbers would mean better performance. Closeness centrality scores each node based on their 'closeness' to all other nodes in the network. Therefore, if a node with high closeness centrality is affected, then the overall impact on the connectivity and distances on the network is severe. Like the Eigen centrality value, the highest closeness centrality does not mean the stock should necessarily be considered in the portfolio.

After measuring centrality scores, a specific algorithm was constructed based on the equal weightage of the centrality scores which came from Closeness, Betweenness and Eigen centrality for this study. As a result of this algorithm, a final centrality score was obtained to select the stocks in the portfolio that could outperform the benchmark.

Different potential portfolios were selected based on the high and low final centrality score. In the next step, the Sharpe ratio was calculated for all potential portfolios in the manner of low and high central scores. The Sharpe ratio measures risk-adjusted returns and has become the industry standard to examine stock/portfolio performance [77]. Modern Portfolio Theory states that adding assets to a diversified portfolio in which the assets are less than one when correlated with each other can reduce portfolio risk without sacrificing returns. Such diversification helps increase the Sharpe ratio of a portfolio. Although a portfolio can benefit from a higher return than its counterparts, it is only a good investment option if the additional risk does not accompany the higher return. The larger the Sharpe ratio of a portfolio, the better its adjusted performance is relative to risk. The performance of each potential portfolio was compared to the benchmark as well as the Sharpe ratio for low and high centrality score.

3.5 Weighted Portfolio

In this research, we seek to introduce a strategy that can be used to construct a portfolio. The aim of this study was to find the best strategy that investors can use from a certain investment perspective. In this study, many steps were taken to find the best strategies to build an optimal portfolio. First, different case studies were designed to understand stock movements and how they can behave in a portfolio. Then, based on the results of the aforementioned case studies, a practical strategy was suggested to determine the optimal approach to building a portfolio.

After grouping companies into low and high-central stocks, each low and high subcommunity’s excess return was compared against the benchmark. The comparative

results showed that low central stocks could outperform the benchmark. In other words, low central stock can predict the market movement; additionally, the diversity in groups of low central stocks was higher than high-central stocks in the terms of sector and size.

It is worth noting that stocks in the low and high central groups had the same weight. Meaning that, if there are 10 stocks in low-central stock, the weight (contribution) of all stocks is 1/10 of the share. The Sharpe ratio for this group of stock was smaller than those of high central stocks. This shows us that low central stocks can outperform the market, but at the same time, have smaller Sharpe ratios. As such, we moved to the final stage of our analysis: portfolio optimization. Importing candidate portfolios in R or Excel and using the Solver feature can change weight for stocks and get portfolios with higher Sharpe ratios. Therefore, we could optimize the weight of the stocks in low central subcommunities and get higher Sharpe ratios. This way of comparison is more realistic since the benchmark contains weighted portfolio.

3.6 Advantages of the Proposed Methodology

Networks play an important role in a wide range of economic phenomena because the economy can be considered as a network in working progress. In each network, each agent only interacts with its neighbors and hence, the network and the graph have the common features. Researchers use these two fields of study as synonym processes. In the economy, each entity (stock, financial sector etc.) is assumed as a vertex on the graph, and the relationship between entities which impact each other are considered as edges. Using the graph properties and the results of the graph theory can examine networks.

Since the behavior of the economy is not isolated from the behavior of individuals, the economy as a network works as a central identity in which many factors can be affected by its equilibrium and achievement. Correlation network modeling is one of the methods that researchers apply to the financial/stock market. A CNM is built by assuming that each company/stock is a node (a vertex) in the graph model, and two nodes are connected by a directed/undirected edge if and only if their correlation coefficient is more that predefined threshold.

Correlation network modeling is relevant, as the highly correlated nodes or nodes with dense connections would give us the information about the companies with the same kind of behavior or characteristics. From a correlation network, researchers can identify companies with similar patterns since they will be grouped in the same cluster. In other words, different clusters are formed from the highly correlated companies. Applying population analysis helps us to compare individual data points with other data points in different clusters or populations regarding different performance levels. Population analysis allows us to compare two or more clusters of companies with respect to one or more enrichment parameters. The results of this analysis allowed us to discover the parameters that significantly affect the cluster.

CHAPTER 4 HOW TO ANSWER RESEARCH QUESTIONS

The proposed methodology contained three main blocks: data procedure and filtering, population analysis which employs a correlation network, community detections and enrichment analysis, and finally, decision support system. The methodology in this study was gradually developed by designing the different case studies. Therefore, these three main blocks were progressively developed and applied to the datasets in the following case studies. In short, the methodology in this study was developed based on five case studies in order below (chapter 5-9).

The first case study was designed to examine the proposed approach to check the ability of the method to overcome the goals (chapter 5). The second and third case studies were designed to check the method during crises. Crises are unexpected events that affect the economy and society in different shapes. Depending on the economy, governments react to crises in different fashions. Some crises are expected, and some are unexpected. Expected crises are the result of previous economic decisions that did not produce the predicted results, such as the 2008 economic crash. For these types of crises, governments have set procedures to deal with their fallout. However, for unexpected events such as the Covid-19 pandemic, there was no way for governments to gauge the aftermath economically or socially. The following case studies (chapter 6 and 7) tried to test the model during the crises in 2002, 2008, and 2020. For those cases, we hypothesized about whether there is a relationship between the Covid-19 pandemic, the 9/11 attacks, and 2008 economic crash on the behavior of returns in different economic sectors. The fourth case study assessed the model for 20 years, from 2000 to 2021. The

result of the case study assessed whether the model could fabricate a potential portfolio that could outperform the benchmark. From 2000 to 2021, the socio-political state of the world experienced many fluctuations (chapter 8). The last case study was designed to show how the proposed model could select the stocks for the potential portfolio that can predict the market by having a good Sharpe ratio (chapter 9).

Each case study started with an introduction, followed by data procedure and filtering, methodology, and the analysis of results. Then, the limitation and future studies were explored; therefore, the subsequent case study was designed to overcome the limitations of previous cases. It is worth noting that each case study tests different correlation coefficients and clustering algorithms to ensure the analysis results do not depend on a specific correlation coefficient or clustering algorithm.

CHAPTER 5 A NEW APPROACH FOR ANALYZING FINANCIAL MARKETS USING CORRELATION NETWORKS AND POPULATION ANALYSIS

5.1 Introduction

Investing in stock markets has continued to present challenges for many investors around the world. Attempting to understand the key aspects of stock markets, like when to buy or sell stocks or whether stocks are performing well or poor, remains extremely difficult. Although many studies have addressed these issues, they primarily focused on using simple statistical models.

Fundamental analysis and technical analysis are two ways to analyze stock markets from a financial perspective [31]. Fundamental analysis calculates the stock's intrinsic value with the company's information, analyzes the financial statements, and compares it with the current value to make the right decision about buying and selling the stock of the company. Technical analysis is applied to find the right time to buy or sell a share by observing and analyzing stock price behavior in the past and using the patterns to predict the future price movements. Researchers have analyzed the financial markets from different perspectives using various computational and statistical methodologies, including Artificial Intelligence, Machine Learning, Artificial Neural Networks, Fuzzy Logic, and Support Vector Machines [22].The financial market brings the opportunity for investors to have access to the large and complex daily and monthly data. This data can be used by big data analytical tools, such as network models and correlation analysis. This study aimed to use a population analysis and correlation network model (CNM) approach to determine the specific sector that dominates in the stock market. The purpose

of applying population analysis was to compare different clusters, or groups, of stocks with respect to a particular parameter such as companies' returns, economic sectors, and companies' capitalization. The network models allow researchers to analyze each element in the network and look for networks' characteristics and features that are not possible under traditional approaches.

The general objective of this research was to demonstrate how network and community analysis could be applied to stock market data. We proposed employing CNM to create a correlation network of companies based on the time-series data of companies' excess returns (ER). After creating a correlation network, we applied the GLay clustering algorithm, which is a community detection algorithm [63], to obtain a set of companies that have similarities in their ER. Based on different communities from the network, we identified the top two communities produced by the clustering algorithm for further analysis. We then identified the economic sectors that dominated the market between the years 2000 through 2004 and found their relevant capitalization. In this study, we aimed to answer the following questions: What type of companies have the similar behavior during 2000-2004 based on their ER and amount of capitalization (Research Question 1: RQ1), and does any other factor affect their behavior outside of network analysis (Research Question 2: RQ2)?

5.2 Material and Methods

There were two sets of data involved to this case study, one set of data was collected from the Center for Research in Security Prices (CRSP). In addition, another data set was collected from the Fama-French (FF) data library. The CRSP dataset

contained a list of stocks/companies with their monthly information. After data cleaning and filtering, we looked at companies' Standard Industrial Classification (SIC), ER, and total capitalization (TCap) for all available stocks in the U.S. SIC is a code used to group companies with similar products or services at the end of the reporting period. SIC is used to identify companies' economic sectors. Economic sectors are groups that are categorized according to their place in the production chain. TCap results from the multiplication of price and number of shares (in 1000s) for each company. Based on the companies' capitalization, the companies' size will be divided into ten categories called deciles. In this study, companies that belong to decile 1-5 are called small-size, and companies that belong to decile 6-10 are called large-size companies. ER was obtained by subtracting the return value (a parameter in the CRSP data set) from the risk-free value (from the FF data set).

Under population analysis, correlation network analysis was applied in time-series data to find hidden information when it could not be found using traditional approaches [34], [64]. In this study, cluster analysis was applied on the network to group different companies whose degree of correlation between two companies was above a threshold.

After creating the communities/clusters, all the parameters from CRSP were added to the companies in each community for further analysis to do cluster enrichment for that specific community. As the communities were formed with high correlations among the nodes, we could infer that the overall behavior of the nodes within each community is the same.

5.3 Data Acquisition and Filtering

Two separate data sets were utilized in this research. The Center for Research in Security Prices (CRSP) is a variation of the research-quality stock database, which contains monthly data of all companies from 1926 to 2018. This data set includes five parameters/variables (Companies' ID (CUSIP), date, return, SIC, TCap). Another data set was collected from the Fama-French (FF) data library. Risk-free[4] was the only parameter that was used from FF data set. As a major parameter in this study, ER reflected the overall performance of companies based on different factors such as economic sectors and capitalization. ER values range between -1 and 2. A value of -1 means 100% loss, and a value of 2 means 200% gain. To study the features of the correlation network for companies, we first established the correlation network by extracting data from different data sets. A network, represented by a graph, is a collection of nodes and edges, N = (V, E), where each node represents a company, and an edge between two nodes reflects the relationship between the corresponding companies. We established the pilot study for the years 2000 to 2004 (inclusively). The reason for selecting these years is because the 9/11 attacks happened during this period, and we would have liked to study what sectors of companies are better to invest in case of having an unexpected event.

During the time-period 2000 to 2004, some companies ceased functioning, resulting in some missing data points for those companies. Therefore, to avoid any biased results for this pilot study, only companies with all data points for 12 months in these five

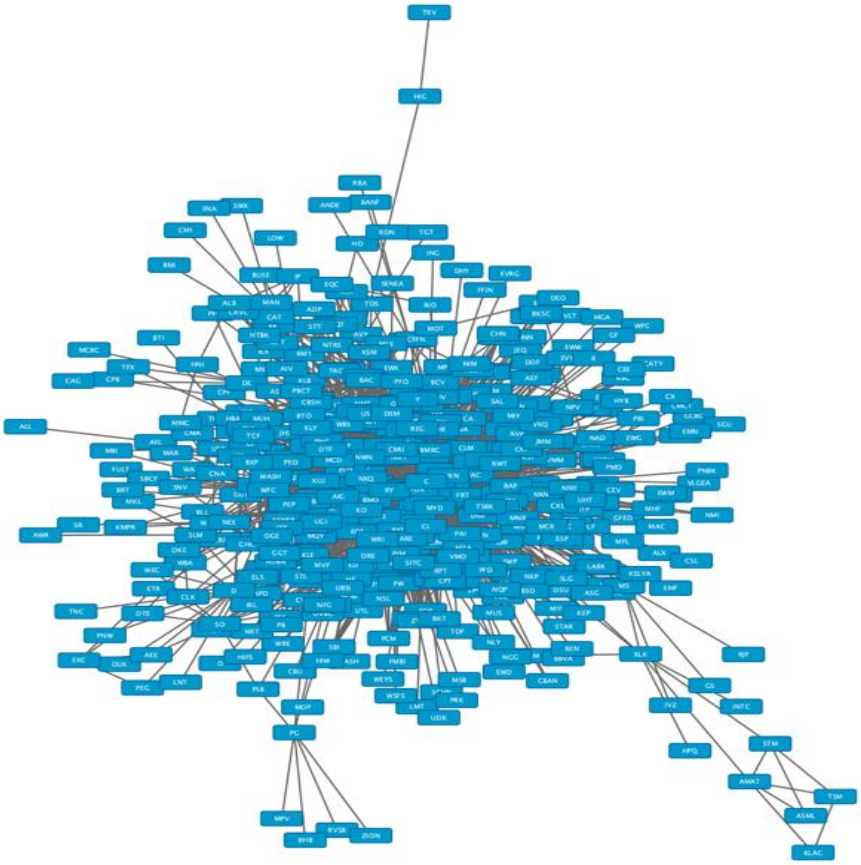
[4] risk-free is the rate of return of a hypothetical investment with no risk of financial loss over a given period of time

years were extracted. At the end of the data filtering, out of 5280 companies, 3427 companies were extracted for further analysis.

5.4 Correlation Networks and Community Detection

The correlation network was created based on the ER. The time-series data of ER for the 3427 companies was recorded as an input matrix. In the matrix, there were 60 data points (time-series data of 60 months) for each company. Since the data set is normally distributed, we applied the Pearson correlation coefficient to the ER matrix. In the constructed correlation network, each company as a node (vertex) is connected to another company with an undirected edge if their correlation coefficient is 0.75 or more and their significance of correlation is less than or equal to 0.05. This created a correlation network with companies as nodes along with highly correlated companies connected by edges as shown in figure 1.

Figure 1-Correlation network of companies' data between 2000 and 2004



In this study, due to the high similarity between companies' ER, cluster analysis was applied on the network as a data analysis shortcut tool to group different companies whose degree of correlation between two companies is above a threshold. Cluster analysis, or clustering, is a process by which a set of objects can be separated into groups. Each partition is called a cluster. The members of each cluster are very similar to each other in terms of their properties, and, in turn, the similarity between clusters is minimal. In this case, the purpose of clustering was to assign similar objects into one cluster and label with object's membership in the cluster. The financial market network is one of the most complex networks, bringing significant challenges to visualization. Creating

clusters from this complex network consumed considerable time and computational resources, and the results are not always useful. By using a specific topology, we had an opportunity to visualize the clusters from the community structure. Therefore, for this project, we applied GLay community structure detection algorithm (available in Cytoscape [21] that identified as an efficient layout for very large networks. GLay community clustering was applied to the network with all default parameters in Cytoscape [21] on the obtained correlation network to produce communities/clusters. GLay clustering was used since it has the ability for disintegration and could be used for large and complex networks that contain many nodes and edges [63].

In this study, from the 3427 companies, 2580 companies were involved in the network based on the above referenced network organization. Out of 2580 companies, 2477 nodes were placed in two communities (communities one and two). The subset of these two large communities is shown in Figure 2 and 3. Hence, communities one and two were considered for further analysis, and various experiments were conducted on these two communities.

Figure 2-Subset-Community one network with 1402 nodes and 156778 edges.

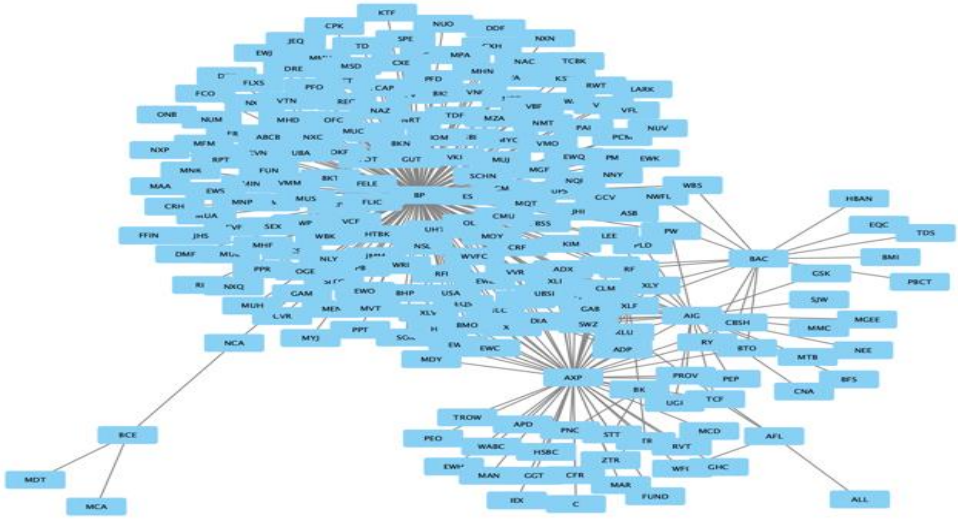
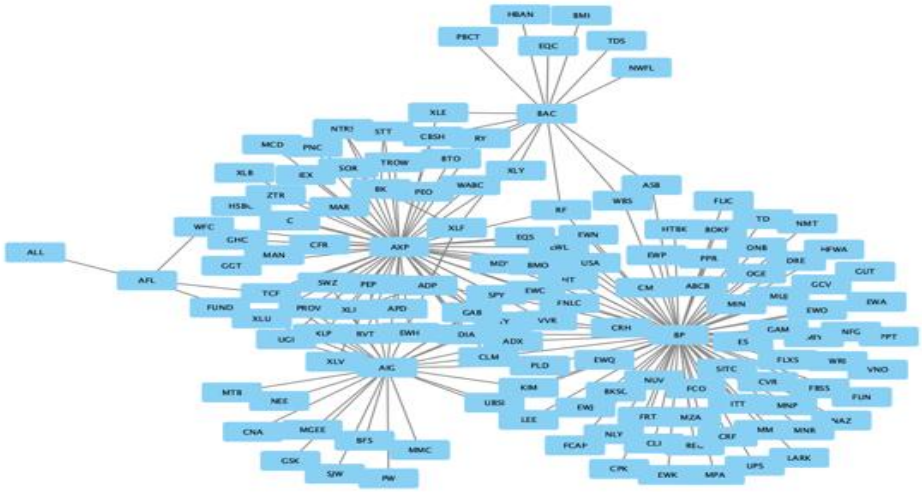


Figure 3-Subset-Community two network with 1075 nodes and 307087 edges.



5.5 Experimental Results

This section discusses various network properties and the application of the population analysis on the correlation network.

5.5.1 Network Properties of the Communities

The correlation network in Figure 1 represents 3427 companies (nodes) and 657890 relationships (edges). This network had 25 communities that were formed with a correlation of 0.75 or greater between edges.

This network was a scale-free network that followed a power-law node degree distribution. The power-law degree distribution means there are many nodes with fewer degrees and fewer number of nodes with more degrees. Two top communities (with respect to the number of nodes) were selected from this network, and the communities' statistics are shown in Table 1. Community 1 had 1402 nodes and 156778 edges, and Community 2 had 1075 nodes and 307087 edges. Some of the network statistics/properties of the top two clusters are shown in Table 1. The average degree of each cluster is the average number of edges of all nodes. The cluster density describes the potential number of edges present in the sub-network compared to the possible number of edges in the subnetwork. The higher the clustering coefficient, the higher the degree to which nodes in a graph are inclined to cluster together [65]. The higher values of the average clustering coefficient for each cluster/sub-network indicate that the nodes inside each cluster tend to be part of that cluster only. Table 1 shows that Community 2 had fewer nodes but had the highest clustering coefficient of 0.85. Once again, Community 2 had a higher density (0.532) compared to Community 1.

Table 1-Network statistics of top two communities.

Community	Avg_Degree	Density	Corr_Coef	#Node	#of Edges

1	223.689	0.16	0.71	1402	156778
2	571.325	0.532	0.85	1075	307087

5.5.2 Cross-Tabulation Properties

In this study, the data set obtained from communities was analyzed based on various parameters, such as economic sectors, companies' capitalization, and their degrees. Economic sectors are large groups of the economy, grouped according to their place in the production chain, by their kind of work (product or service) or ownership. There are 12 economic sectors in the economic era: consumer staples (NoDur), consumer discretionary (Durbl), industrial (Manuf), basic materials (Chems), energy (Energy), information technology (BusEq), communications (Telcm), utility (Util), real estate (Shops), health (Hlth), finance (Finance), and other (Other).

The degree indicates how many times each company is connected to other companies based on similarities in their ER. According to the degree range for the companies, we ranked the number of degrees into five categories. For example, rank one showed degrees between 1-200, and rank five showed degrees between 800-1000. The higher rank means more similarity with other companies within the cluster. Companies' capitalization results from the multiplication of price and number of shares (in 1000s) for each company. In the financial domain, a company's capitalization is divided into ten categories that are called deciles. A decile is a quantitative method of

splitting up a set of ranked data into ten equally large subsections[5]. The first category has the lowest, and the tenth category has the highest amount of capitalization.

5.5.3 Cross-Tabulation Results for Communities One and Two

Since Communities 1 and 2 had the highest number of companies in their community, in this section, we reported the result of the cross-tabulation analysis between degree-ranked, capitalization-ranked, and economic sectors for companies belonging to these communities. In 2000-2004, the finance sector dominated the market with 36%, and consumer discretionary with 2% had the lowest share in the market. The cross-tabulation analysis showed that, for ranked degrees one through five, the finance sector had the highest number of companies in the market followed by the information technology sector (Table 2 & Figure 4).

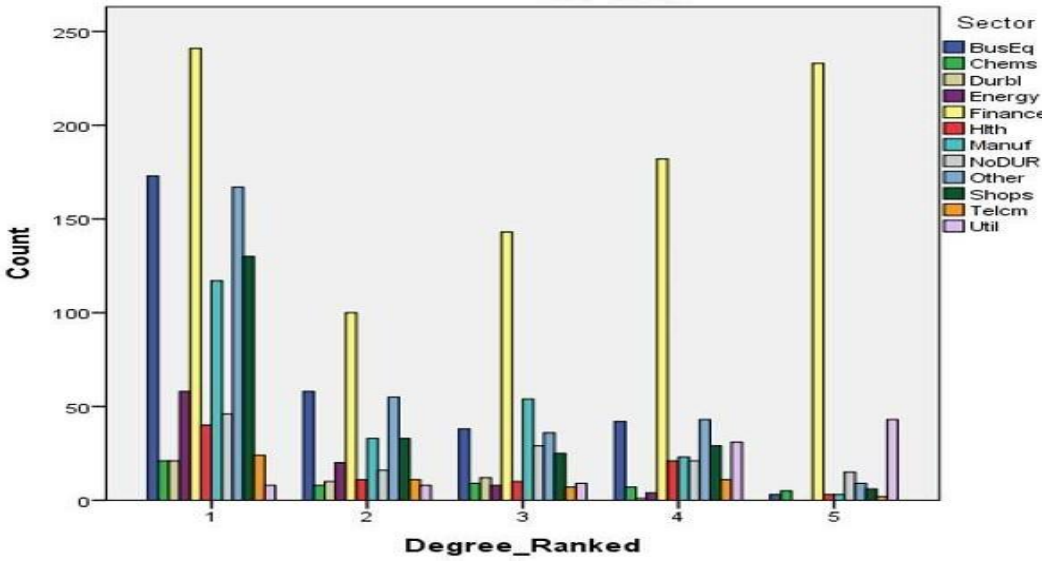
Table 2-Sector statistics for 2526 companies in 2000-2004

Sector	Frequency	percent
Consumer staples	127	5%
Industrial	230	9%
Energy	90	4%
Basic Material	50	2%
Information Technology	314	12%
Communication	55	2%
Utility	99	4%
Real State	223	9%

[5] https://cleartax.in/g/terms/decile

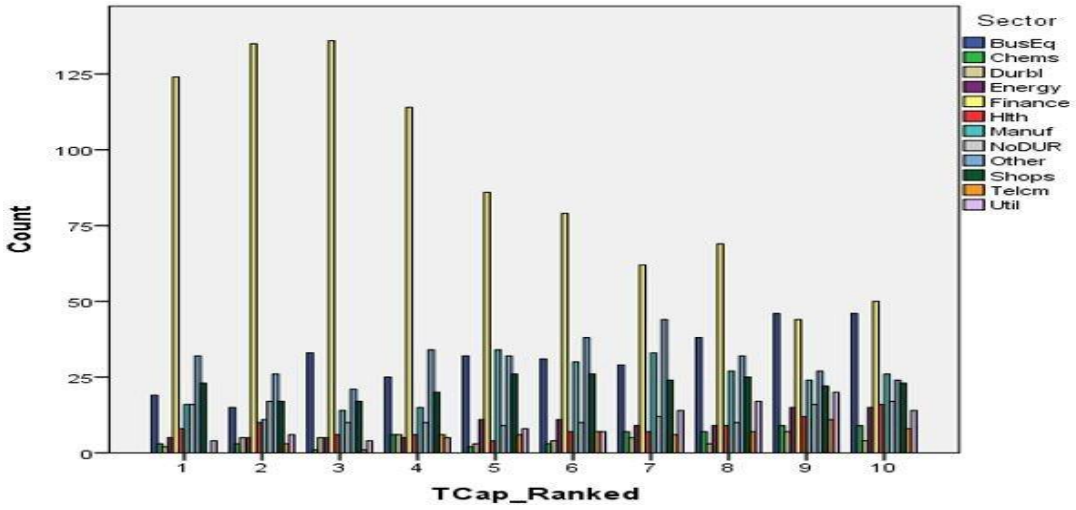
Health	85	3%
Finance	899	36%
Consumer Discretionary	44	2%
Other	310	12%
Total	2526	100%

Figure 4-Cross-tabulation between economic sectors and ranked degrees



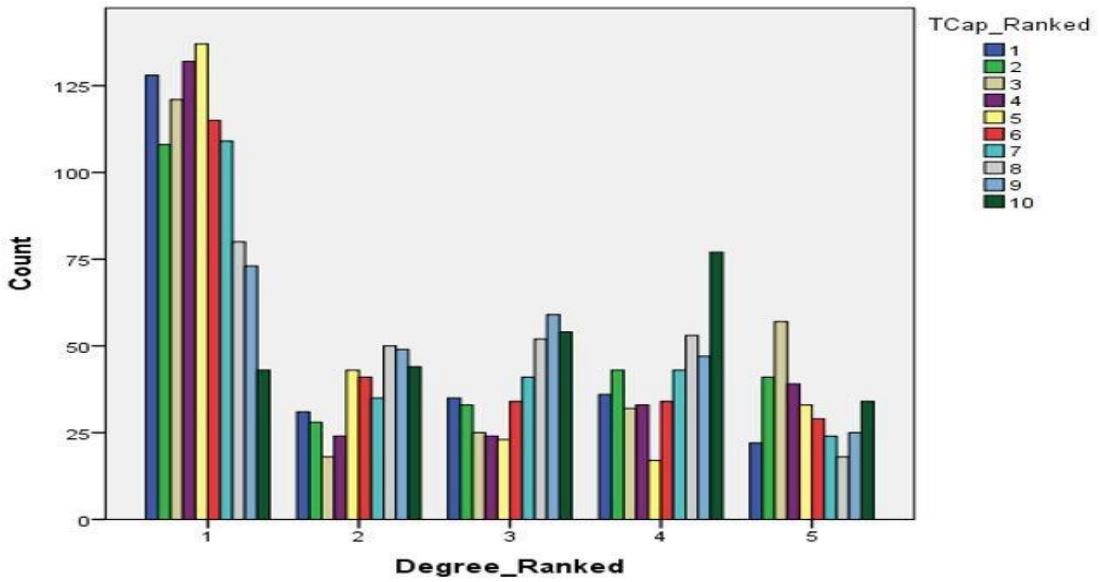
Applying cross-tabulation between economic sectors and capitalization (deciles) showed that companies belonging to the finance sector for all ten deciles had the highest share in the market during the time-period 2000-2004 (Figure 5).

Figure 5-Cross-tabulation between economic sectors and ranked capitalization in 2000-2004



Finally, a cross-tabulation analysis between degree-ranked and capitalization showed that the highest number of degrees belonged to companies that had the lowest amount of capitalization (ranked 1-4). In other words, companies that had a low amount of capitalization (deciles 1-4) had the most similarity in their ER with other companies. That means they had the same ER as other companies considered large companies (Figure 6).

Figure 6-Cross-tabulation between ranked degree and ranked capitalization in 2000-2004



As a result of this analysis, we can say that companies belonging to the finance sector from 2000 to 2004 had the highest degree of relationship (similarity in their ER) with other companies while they had the lowest amount of capitalization. Furthermore, based on this analysis, we can say that companies that have the most similarities in their ER movements compared to others in the community are those that even have the lowest amounts of capitalization.

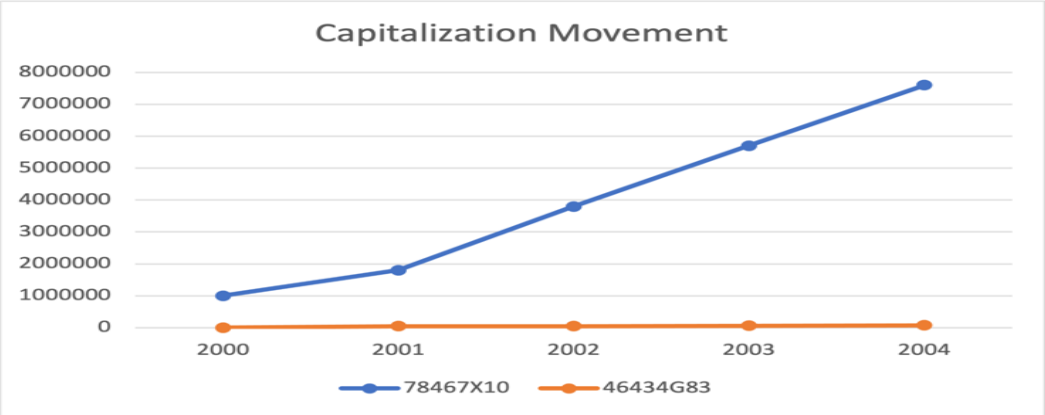
5.5.3.1 Analysis of 50 Randomly Picked Companies in Community One with Respect to Input Capitalization and Economic Sectors

For this part of analysis, we randomly picked the 50 companies in Community 1 as one of the largest communities in the correlation network for analyzing capitalization movements. In this community, 80% of companies belonged to the finance sector with

capitalization ranked 1-4, and the remaining 20% were companies belonging to other economic sectors.

As another comparison in Community 1, we compared the behavior of one high degree and one low degree company in the finance sector. The company ID ”46434G83” belonged to the highest-degree company and the company ID ”78467X10” belonged to the lowest-degree company. Tracking their capitalization showed that company 46434G83, belonged to the lowest-ranked capitalization decile (decile 2), while it had the most similarity in its ER compared to company 78467X10 as one of the largest size companies in the market (decile 9). Figure 7 shows the capitalization trend for these two companies in the finance sector in Community 1 during 2000-2004.

Figure 7-Comparison of average capitalization values of two companies from Community one in 2000-2004(Horizontal axis stands for” year” and vertical axis is “amount of capitalization”)

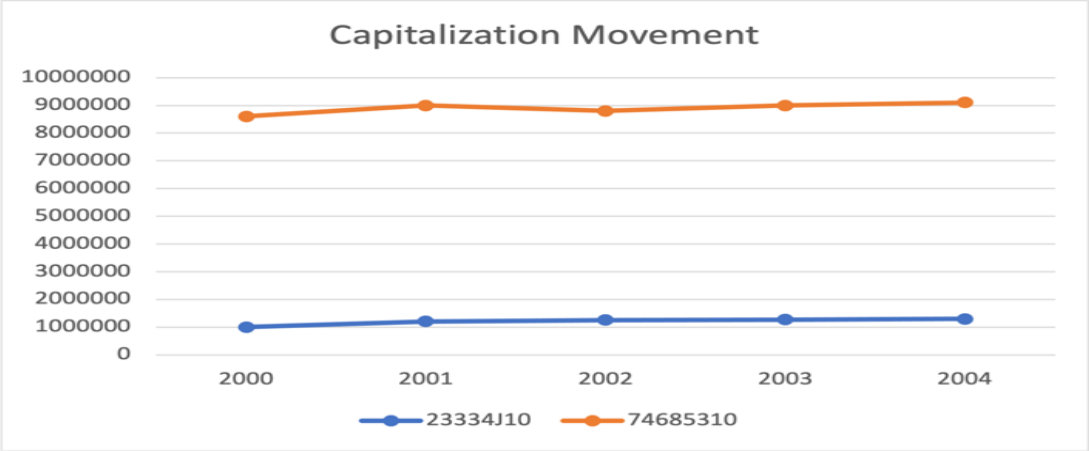


5.5.3.2 Analysis of 50 Randomly Picked Companies in Community Two with Respect to Input Capitalization and Economic Sectors

We applied the same analysis procedure on the next largest community (Community 2), and again, 50 companies were randomly picked. In this community, 96% of companies were in the finance sector, and from this 96%, 69% belonged to capitalization rank of 1-4.

As another comparison in Community 2, we compared the behavior of one high-degree and one low degree company in the finance sector. The company with CUSIP ”23334J10” and the company with CUSIP ”74685310” were selected. Tracking their capitalization showed that company 23334J10 belonged to the lowest-ranked capitalization decile (decile 3), while it had the most similarity in its ER compared to company 74685310 as one of the larger sized companies in the market (decile 7). Figure 8 shows the capitalization trend for these two companies in the finance sector in Community 2.

Figure 8-Comparison of average capitalization values of two companies from Community two in 2000-2004 (Horizontal axis stands for "year" and vertical axis is "amount of capitalization")



5.6 Conclusion

With the growing availability of financial data, new approaches are needed to take full advantage of such data and provide investors insightful knowledge about the behavior of companies in the financial markets. This research proposed a new approach for analyzing financial markets and extracting useful information from the large amount of financial data. Due to the complexity of financial market data, the proposed approach that utilized the concept of population analysis and correlation networks, along with associated enrichment analysis, allowing us to identify behavioral patterns of the financial market that are difficult to identify using traditional approaches.

To test our proposed approach, we conducted a case study on the stock market data of the United States for the years 2000-2004. We proposed a correlation network

model along with population analysis to conduct the study. We constructed a correlation network based on companies' ER. After creating the network and applying clustering algorithms, we extracted the top two communities and analyzed the associated companies. We collected various relevant information about the companies, such as the amount of capitalization and economic sectors. Based on the clustering analysis, we found that companies in finance sectors had the highest share in the market as compared to other sectors. We also showed that companies in the finance sector have similarities in their ER movements to that of large-size companies, even though they mostly had the lowest capitalization. Based on the obtained results, it can be concluded that investment in a small company with low capitalization in the finance sector, even during the crises, may yield a higher return than investment in large companies. From 2000 to 2004, companies in the finance sector kept their consistency with low capitalization and got the same ER as big companies with high capitalization (RQ1). Using population analysis, we did not find any parameters outside network characteristics that significantly affected the behavior of the companies under study (RQ2).

The proposed model and the reported results represent a starting point for a new direction in analyzing financial markets. The results show the viability of this new approach. However, additional studies with larger and more diverse data sets are necessary to make a case for utilizing the concept of population analysis in making important financial decisions. The limitation of this study is that we analyzed the market for a limited sample during the 2000-2004 period. To further validate the obtained results, we conducted a more comprehensive study using the proposed approach for different time periods and utilizing different types of data sets. We intend to apply the

concept of population analysis on different sets of data tied to independently established major crises to recognize the patterns that may be otherwise obfuscated. In addition to ER, future studies also include exploring other indicators such as different economic sectors and companies' sizes for comparing the behavior of companies in financial markets.

CHAPTER 6 A NOVEL POPULATION ANALYSIS APPROACH FOR ANALYZING FINANCIAL MARKETS UNDER CRISES – 2008 ECONOMIC CRASH AND Covid-19 PANDEMIC

6.1 Introduction

The economic cycle and its financial implications can be impacted by many parameters including various world events, internal developments within companies, inflation and interest rates, and releases of financial reports. Economic cycles can also be impacted by sudden or unexpected events, which may rise to the level of a local, regional, or work crisis. In such a case, there would be a significant decline in different economic activities that would continue for a few months. An economic crisis affects different economic sectors, rates of GDP, unemployment, and so forth. The effects of crises depend on their nature and the governments and industries' reactions [29]. An economic crisis can be categorized as a structural crisis such as the 2008 crash, or as a shock or a non-structural crisis such as the Covid-19 pandemic [36], [66], [67]. When an economic crisis happens, all developed and developing countries are affected by it. The economic crash in 2008 was categorized as a structural crisis since market players were the leading cause. The evidence of the financial crisis in 2008 displayed the weaknesses in the financial system. The problem originated with risky mortgage lending practices by financial institutions. Due to the structural nature of the economic crisis in 2008, the government and relevant financial institutions applied different strategies and pursued different policies that helped to some degree and provided a path for recovery [36], [66], [67].

The 2020 pandemic is widely categorized as a non-structural world crisis. As the virus spread from China, the initial reaction of many countries was to close their borders. As a result, different economic activities stopped progressing including major profit loss in transportation and travel industries as well as significant increase in unemployment rates. Unfortunately, there is no specific way to predict when a crisis will happen; however, different indicators could help anticipate a coming crisis. The probability of having a crisis increases when all or some of those indicators change. The most common indicators are the rate of GDP, unemployment rate, the devaluation of assets such as stocks. Lost profit or a low rate of profit in short/long-term deposits can be a sign of an approaching economic crisis [37], [40].

A significant change in stock market movements is one common indicator of a future crisis. In most situations, the decline in the stock market starts months before the crisis begins, and depending on the market, it will impact other markets around the world. As mentioned earlier, depending on the non-structural or structural nature of the crisis, the significant widespread effect of the crisis around the world could be different. For example, in the 2020 pandemic, the Wall Street stock market experienced its worst fall since the 2008 financial crisis, and different industries related to airlines, tourism, transportation, and energy were impacted. In this case study, we propose a new model to study the impact of crises on the financial market that is based on the concepts of population analysis and comparative evaluations. We attempt to compare the behavior of different economic sectors during the 2008 economic crash and the Covid-19 pandemic using this proposed model. We applied the correlation network and community/population-based analysis on data sets collected from these two events to find

similarities in companies' excess returns. Later, we grouped the companies based on their economic sectors to explain the reasoning behind their behaviors.

6.2 Methodology

As mentioned in methodology section, in chapter 3, the proposed approach in this study is based on the concept of population analysis, which is to assess an individual element as it compares to a group of peers. The first step in this approach is to build a similarity network based on the desired outcome—in this case, it is based on the similarity in the performance of stocks in a financial market based on their behavior for a certain period, followed by conducting the analysis based on the structural properties of the constructed network. Graphs are used to model the network, making it possible to take advantage of the numerous tools and algorithms available to extract relevant parameters and relevant information from the graph models. For example, a highly dense subgraph in the obtained network will represent a cluster of stocks with common characteristics as compared to other stocks under study. Such information would be particularly critical when attempting to assess the behavior of companies under a financial crisis.

When an economic crash or a worldwide pandemic happens, the immediate effects are always on the stock market, and especially on various sectors of companies in the stock market. However, there are always a few sectors that are resilient to any severe conditions in the stock market. This proposed method aims to study various event and see what sectors are resilient to these events, along with why and how they bounce back very

quickly. Identifying these sectors could be an important source of guidance for investors to predict patterns in the stock market for financial gain, even in worst case scenarios.

While studying an individual stock or a particular sector gives the information about a specific behavior, grouping/clustering the companies based on their historic behavior enables us to see the global picture. Once the companies are globally divided, based on their behavior, it is easy to study them at a high granularity level to understand how they are affected and resilient in various events. Comparing one group, with respect to the remaining groups, enables us to identify specific characteristics associated with each group, and what parameters are significantly impactful in this group compared to others. The very advantage of graph-based models, such as correlation network models in view of stock market related datasets, is that they allow us to establish similarities and differences among the companies based on their historical behavior and partition them into groups to assess the overall behavior of each group as related to the rest of the groups [13]–[15], [17]–[19].

The proposed population analysis method in this study allows us to establish valuable connections among stocks under study and conduct a comparative study of how various stocks or groups of stocks behave in general and under economic crises such as an economic crash or a pandemic. This section explains the series of steps performed right from dataset preparation until identifying various significant sectors using population analysis for both the economic crash 2008 and the Covid-19 pandemic. The following are the steps in this process (Figure 9):

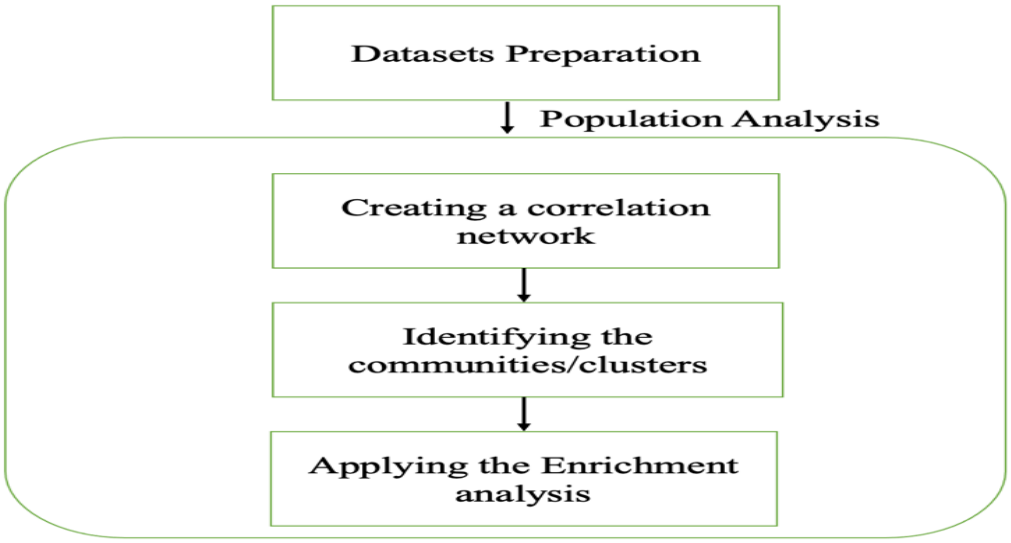
a) Datasets preparation for both the events

b) Creating a correlation network

c) Identifying the communities/clusters

d) Applying the enrichment analysis to identify the significant sectors.

Figure 9-Methodological steps



6.3 Datasets Preparation

The monthly stock market return data for the economic crash was collected from the Centre for Research in Security Prices (CRSP) database and the Fama-French (FF) databases. The pandemic data was collected from the Bloomberg database. Based on availability, the economic crash data over four years (48 months from 2006 t0 2009 inclusively) and pandemic data over eight months (from January 2020 to August 2020) were considered for analysis. The monthly return values from CRSP or Bloomberg databases were subtracted from the risk-free values of the FF database to obtain the

parameter Excess Return (ER). The ER values represent the overall behavior of the stock and provide the time-series data of the two datasets in terms of their ERs. The time-series data was be used as the input for constructing the correction networks in the next step. Please note that the 2008 economic crash dataset is referred to as CrashER, and the pandemic dataset as Covid-19ER. There were 1411 companies from 12 sectors in the CrashER dataset and 450 companies from 12 sectors in the Covid-19ER dataset.

6.4 Creating a Correlation Network

The correlation networks for the two individual datasets were created based on the correlations/similarity patterns of the time-series data of the ER values. We generated networks using different thresholds to create an edge between the nodes in the correlation network including 75%, 85%, and 95%. Different thresholds were tested for the robustness analysis. While different thresholds produced different networks, the overall analysis was not significantly impacted by the threshold variability. In this study, we are reporting the results obtained using networks generated with correlation parameters equal or above 0.95%. Figures 10 and 11 represent the correlation networks ($\rho \geq 0.95\%$) for the CrashER and the Covid-19ER datasets respectively.

Figure 10-Correlation network of CrashER-dataset

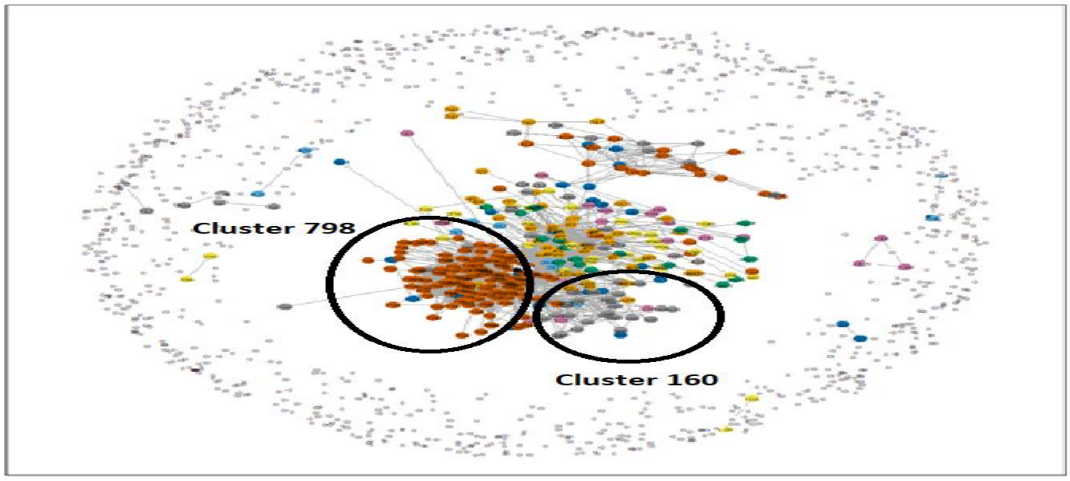
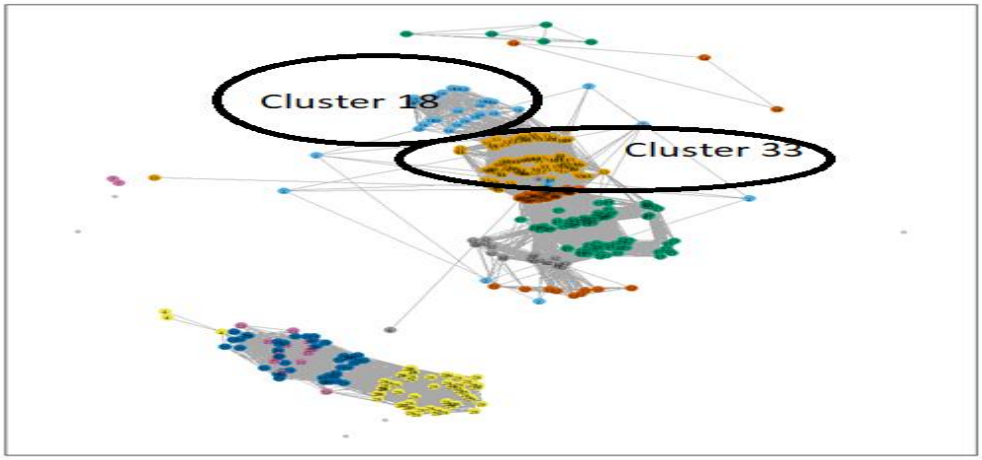


Figure 11-Correlation network of Covid-19ER-dataset



6.5 Identifying the Communities/Clusters

The main goal of building a network of stocks that reflects their behavior similarities is to identify groups of stocks that exhibit similar trends over a period time, in particular, during disrupting events or crises. To achieve this goal, we used the various

clustering algorithms to identify subnetworks of high density in the constructed networks. In this study, we are reporting the results obtained using the Markov Clustering Algorithm (MCL) since it has been shown to perform well on graph-based models [68]. MCL is a graph clustering algorithm based on random walks that can be tweaked using an input inflation parameter to influence the size of the clusters. The parameters range between 1 and 10 with large values producing small clusters. In our experiments, we tested different values of the inflation values on both datasets to identify the values that provide higher degrees of distinguishability among the clusters. In terms of quality of clustering, it is preferable to produce clusters with high homogeneity of elements in each cluster and higher separation among elements in different clusters. In this case study, we are reporting results obtained using Inflation values between 2 and 4 with 0.5 increments since they provide the best results in terms of homogeneity and separation. We used End Average Excess Return (EAER), which is an average excess return of the last month in the dataset, to assess the quality of the produced clustering, as shown in Table 3. For example, for the CrashER dataset, with the EAER values listed in the last column, the maximum difference between the lowest value of EAER (0.52 for cluster 49), and the highest EAER (5.11 for cluster 160) is 4.59, which was obtained with the inflation parameter value of 2.5. Similarly, different inflation values were tested for the Covid-19ER dataset, and again, 2.5 produced the best results. In Table 4, the maximum EAER difference is between cluster 18 (the lowest return value of -12.55) and cluster 30 (highest return value of 103.24) is 115.79. The minimum size of each cluster was calculated based on the average of all the cluster sizes produced. The clusters with at least

minimum size (mean size and above) were considered candidate clusters (CC) and they were considered for the population analysis in the next step.

Table 3-MCL clusters with the end average excess returns (EAER) (CrashER-dataset)

ROW	#Cluster	#Nodes	#Nodes (%)	EAER (%)
[1,]	49	29	2.055	0.52
[2,]	798	132	9.35	1.55
[3,]	160	25	1.77	5.11

Table 4-MCL clusters with the end average excess returns (EAER) (Covid-19ER-dataset)

ROW	#Cluster	#Nodes	#Nodes (%)	EAER (%)
[1,]	18	25	5.071	-12.55
[2,]	33	140	28.398	14.81
[3,]	29	47	9.533	16.09
[4,]	32	22		27.52
[5,]	27	93	18.864	29.02
[6,]	28	59	11.968	30.74
[7,]	30	43	8.722	103.24

6.6 Applying the Enrichment Analysis

In a process with so many non-deterministic parameters, such as the assessment and predictions of behavior of stocks under non-structural conditions, it is very difficult to measure performance in absolute values. Hence, the need for assessment by comparison. Stocks that perform better than most can be argued to be performing well as compared to those that perform worse than most. Hence, we employed the concept of population analysis, which was a process of assessing performance of one group/community in comparison with the rest of the groups/communities in the domain under consideration. The objective of population analysis in this study was to compare different communities with respect to a certain outcome measure, then identify which input features were enriched in those communities. In this case study, the outcome parameter was the excess return (ER) and the input parameter was the sector. We applied population analysis to see the significantly enriched sector(s) in each of the communities with respect to the remaining communities. To find the significantly enriched parameters by comparing various communities, we used hypergeometric distribution, as shown in Equation 1. Hypergeometric distribution is a method of sampling without replacement, used for gene enrichment analysis in earlier studies [69]. We used hypergeometric distribution with a False Discovery Rate (FDR) of 0.05 to identify the significantly enriched sectors in a certain community of stocks/companies with a common behavior pattern. To apply the hypergeometric distribution, we divided the companies into two sets; a target set and a background set. The background set consisted of the universal set; that was all companies used in the study. The target set was any community/cluster that we wanted to test if enriched by a particular sector.

Equation 1-Hypergeometric Distribution Formula

$$P(X \geq K) = \sum_{x=k}^{min(K,n)} \frac{\binom{k}{x}\binom{N-K}{n-x}}{\binom{N}{x}}$$

Where N is the total number of companies in the target set, K is the total number of companies in the same sector, n is the total number of companies in the given cluster, and x is the total number of companies in the given cluster that has the same sector name. P (X≥ K) is the probability of having at least k companies with the same sector name from the given cluster.

Table 5-Significantly enriched parameters for the given cluster 798 of CrashER-dataset. T.S: Target Set, R.S: Remaining Set, B.S: Background Set

Feature	T.S	P_VALUE	R.S	B.S
Basic Material	1	1	2	3
Utility	5	1	24	29
Finance	122	1.55e-10	25	147
Industrial	2	1	1	3
Real State	1	1	1	2
Other	2	1	1	3

6.7 Results

The two events, the economic crash 2008 and the Covid-19 pandemic, were compared in this study. As mentioned earlier in the methodology section, we used 48 months of data for the economic crash and eight months of data for the pandemic. Using the population analysis procedure, a correlation network was created based on the outcome parameter, and the resultant individual communities were analyzed for the commonly enriched input parameters. In this case study, the outcome parameter was the excess return (ER), and the input parameter was the sector. The correlation networks for the two individual datasets were created based on the correlations/similarity patterns of the time-series data of the ER values. Figures 10 and 11 represent the correlation networks ($\rho \geq 0.95\%$) for the CrashER and the Covid-19ER datasets, respectively. A clustering algorithm was applied to each of the correlation networks individually. Resultant communities/clusters were analyzed further to recognize the commonly enriched sector(s). The resulting networks are shown with different colors and numbers in Figures 10 and 11. The clusters with at least the minimum size (mean size and above) are the candidate clusters (CCs) considered for the population analysis.

The mean cluster sizes for the CrashER-dataset and the Covid-19ER dataset were 15 and 20, respectively. Three CCs were produced with equal and above the mean size of 15 for the CrashER-dataset as shown in Table 3, and seven CCs for the Covid-19ER-dataset as shown in Table 4. In addition, the cluster numbers, number of nodes (or companies), percentage of nodes in each cluster, and their EAER (%) are also shown in tables 3 and 4. Robustness analysis with different thresholds ($\rho \geq 0.75\%$, 0.85%, and

0.95%) was applied to create the correlation networks to identify what sectors are significant at different thresholds. Different thresholds were tested for the robustness analysis. While different thresholds produced different networks, the overall analysis was not significantly impacted by the threshold variability. These results were obtained using networks generated with correlation parameters ≥ 0.95%.

Hypergeometric distribution was applied on the resultant clusters to find the parameters' (in this case, sectors) enrichment. Table 3 shows that the clustering algorithm produces three communities/clusters for the CrashER-dataset with a mean size of 15 and above. The three communities were numbered 49, 160, and 798, with the number of nodes 29, 25, and 132, respectively. The total number of nodes was 187. In population analysis, each community was compared against the remaining communities as a target set and a background set using hypergeometric distribution to identify the significantly enriched sector(s). The background set was the set of all companies/nodes with the respective sector names associated with them, and the target set was the set of nodes/companies from the given community associated with the same sector name. For example, community 798 contained 132 nodes/companies from six different sectors, such as basic materials, utility, finance, industrial, real estate, and others, as shown in Table 5. To identify what sector was significantly enriched out of these six sectors, we applied hypergeometric distribution. For each sector, a target set and a background set was created. For example, there were three companies (Background Set) that were associated with the sector name "Basic Materials" in all three communities together. Out of these three companies, one company (Target Set) was present in community 798. Two companies were in the remaining communities (Remaining Set). The hypergeometric

distribution gave a p-value of 1 for the basic materials' sector. Hence, for any significance level (alpha value 0.01, 0.05, and 0.10), basic material was not a significant sector for community 798. For all the remaining sectors of community 798, a similar hypergeometric distribution was performed as shown in Table 6.

Table 6-Significantly enriched parameters for both the datasets

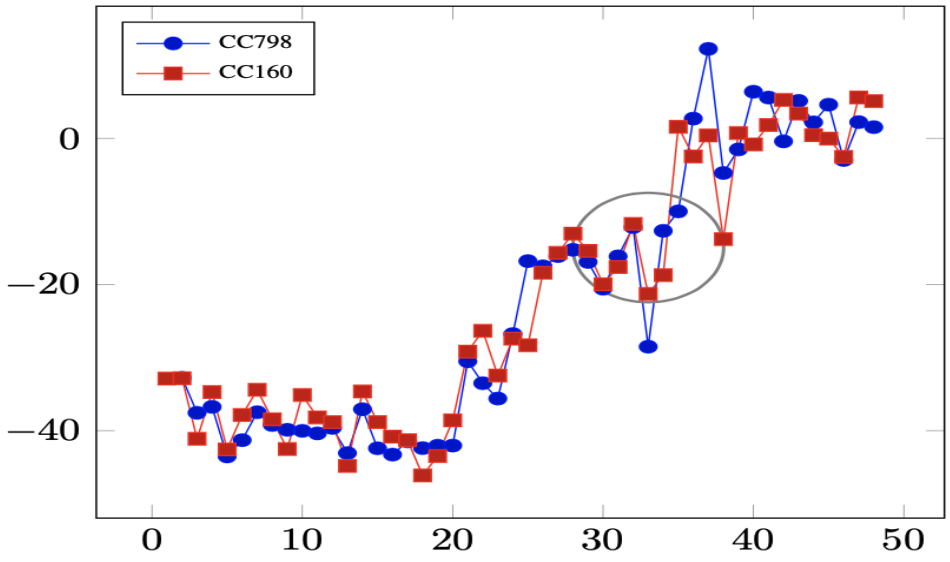
Dataset	Significant Parameters	Correlation Coefficient ($\mathcal{P}$)	#CC	P_Value
CrashER	Finance Utility	≥0.95	CC798 CC160	1.55e-10 5.38-21
Covd-19ER	Finance Energy	≥0.95	CC18 CC33	0.0007 0.0002

The finance sector was the only significant sector out of all the six sectors with a p-value of 1.55e-10. This process is applied for both remaining clusters with respect to all the sectors. The significantly enriched sectors for all the clusters and the datasets are shown in Table 6.

For the CrashER dataset, finance and utility sectors were highly significant parameters. Similarly, for the Covid-19ER dataset, finance and energy were the highly enriched sectors. We also observed that the finance sector was commonly enriched for both crises. The third column of Table 6 represents two CCs in each dataset with at least one significant sector. Only these CCs were encircled in the correlation network graphs

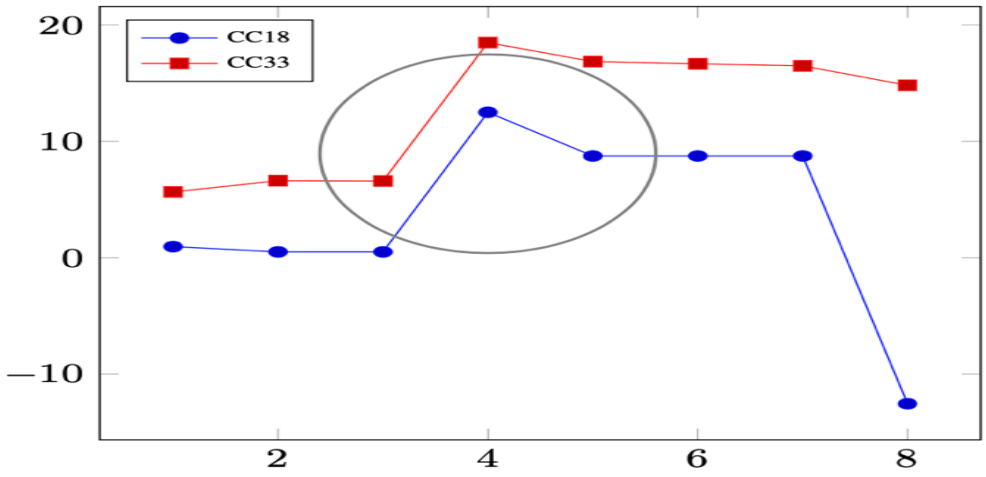
shown in Figures 12 and 13. For the same CCs, the average excess returns' line graphs (behavioral patterns) are plotted as shown in Figures 12 and 13.

Figure 12-Behavioral patterns of economic crash (x=months, y=average return)



From Figure 12, we see that there was a huge crash in September 2008 (the 33rd month and encircled with the event), and from Figure 13, we see that the Covid-19 lockdown started in March 2020 (the 3rd month in the graph encircled with the event). The population analysis results show that the finance and utility sectors were highly enriched for the CrashER dataset, whereas finance and energy sectors were highly enriched for the Covid-19ER dataset, as shown in Figures 12, 13, and Table 6 (see the encircled patterns), respectively. The reasons behind these sectors being significant to these crises are explained in the next section.

Figure 13-Behavioral patterns of Covid-19(x=months, y=average return)



6.8 Discussion

There were several key findings obtained from this similarity network model. Identifying the financial sector as a sector with common characteristics in both crises was one of the main outcomes of the proposed study, as shown in Table 6. The Federal Reserve and government interventions played critical roles in the crises. When the economic crash happened in the summer 2008, the government financially assisted Freddie Mac and Fannie Mae, American International Group, and as well as investment bank Merrill Lynch merging with Bank of America; in addition to Wachovia merging into Wells Fargo [42]. Similarly, the federal reserve provided substantial support to the financial sector during the Covid-19 pandemic. The availability of such assistance to financial institutions allowed them more leeway in granting loans to customers and improved their ability to face the challenges that resulted from the pandemic. As a result, banks were able to grant higher loans and decrease their operating costs. The funds provided by the Federal Reserve allowed banks to maintain a good deal of their financial

stability, even during economic crisis. All sectors were affected by the 2008 financial crisis; however, the utility sector was significantly affected/enriched in the CrashER dataset. Due to the nature of the crisis, the utility sector was affected by financing, demand, and expansion. Since there was decreased trust in the financial market, investors withdrew their investments; thus, this reaction caused more challenges in the market. Since maintaining stability in the market was difficult due to lack of trust, the utility deals decreased. This led to the common characteristics among utility companies that were revealed by our analysis. Similar results were mentioned in previous studies: "electric utilities are a major presence in the financial markets, particularly in terms of short-term borrowing. When the credit markets froze in mid-September 2008, several prominent utilities took proactive steps to secure access to funds by drawing from their bank credit lines" [42].

The energy sector was one of the sectors recognized as significant in the Covid-19ER dataset. Before the 2020 pandemic, the energy sector suffered from a price battle among key players in the market including Saudi Arabia and Russia. When the pandemic started and social restrictions got put in place, the previous problems in the energy sector accelerated. Based on previous experiences, the energy sector could make a profit if crude oil prices went above $50 per barrel. However, due to the restrictions in different economic activities, at the end of April 2020, the crude oil prices were $16 per barrel, which is the lowest price in the last 20 years[6]. Therefore, the low-price experience due to the current situation in the economy caused the energy sector to lose its profit. This led to

[6] https://www.statista.com/statistics/326017/weekly-crude-oil-prices

a significant correlation among companies in the energy sector as identified by the network model.

6.9 Conclusion

In this case study, a novel population analysis approach for analyzing the financial markets under crises was introduced and implemented with excess returns as the outcome parameter and the financial sector as the input enrichment parameter. The study showed that the proposed network analysis using graph modeling and community-based assessment provides a powerful method for dealing with complex problems utilizing high-volume datasets. Correlation networks were constructed using financial data associated with the behavior of stocks during two major events: the 2008 market crash and the 2019 global pandemic. For each network, graph clustering algorithms were used to identify groups of companies that behaved similarly in terms of their ER values. Each group or community was compared with all the remaining communities using enrichment analysis to find what sectors were significantly represented in the given community. For each dataset, two communities were identified as significant with at least one significant parameter. Results showed that finance and utility sectors were significant for the CrashER dataset, and the finance and energy sector were significant for the Covid-19ER dataset. It is particularly important to observe that the finance sector is a common sector in both crises. The reported analysis showed that no matter whether the crisis was due to a structural or non-structural event, the government always tried to intervene to protect the financial sector. Government support, such as low-interest loans or stimulus packages, was provided to stabilize the economy.

This study represents a starting point for a very promising line of research in analyzing the complex world of financial markets, under normal circumstances as well as during crises. Regarding the limitation of this work, we understand 48 months of data was available for the CrashER, while eight months of Covid-19ER was available to us at the time of testing the network model. Equal length time-series datasets would certainly improve the overall quality of the study, particularly to fully understand the aftermath of any major crisis in the stock market. Therefore, in the next case studies, we focus on tracking the real-time data to identify the signal/s at the beginning of a stock market crisis based on the comparison with the other crises using this method.

CHAPTER 7 A Novel Population Analysis Approach for Analyzing Financial Markets under crisis – A Focus on Excess Returns of the US Stocks Under 9/11 and Covid-19

7.1 Introduction

The business world can be unpredictable, and even the most experienced investors may struggle with decisive actions. This unpredictability stems from factors such as institutional and political constraints, the specifics of economic processes in each country, the accessibility of information (and information dissemination), and so forth. While all these factors are crucial in business making decisions, the way in which people and consumers perceive and take in information is even more critical [1]. The coronavirus pandemic, was an unforeseen event, starting in China and quickly spread across the globe. The Covid-19 pandemic affected all countries around the world in different ways. The effects of this pandemic on the financial market brought challenges for investors, including how to make proper decisions for their investments. The effect of disasters varied in different countries in contrasting ways. For example, the stock markets crashed after the 9/11 attacks with the top SP 500 companies dropping at least 14% [7], FTSE 100 in London lost 6%, while the DAX in Frankfurt lost about 8.5%. At the same time, the Swiss market and CAC in Paris lost 7% [4]. Now, regarding the novel coronavirus pandemic, the global financial systems were widely impacted due to the

[7] https://www.investopedia.com/financial-edge/0911/how-september-11-affected-the-u.s.-stock-market.aspx

unforeseen crises that have accompanied the virus. These crises caused the fastest drop in history in terms of the financial market.

Each crisis' effect depends on the size of the national economy and its degree of vulnerability [4]. In the 9/11 attacks, a study [70] found out that some sectors related to banking, insurance, and transportation were more vulnerable than other sectors. During the influenza pandemic in 1987, sectors related to entertainment suffered double what companies in the health sector suffered [5]. After the Coronavirus spread across the globe, market volatility rocketed upwards, and equities dropped. In the United States of America, market volatility levels during the pandemic were either comparable to or greater than market volatility levels in October 1987, and even greater than the market volatility during the economic turmoil of 1929 and 1930—the time of the Great Depression [71]. Covid-19, as a pandemic, brought society to the point in which people were required to work from their homes and buy their goods online. This means that companies belonging to a specific sector (specializing in a certain product) were affected more than the others. In normal circumstances, when the economy is strong, consumers spend more money on consumer discretionary products; in contrast, when the economy is weak, consumers spend more money on consumer staple products. However, the current evidence from the stock markets showed that consumers' behaviors have not followed their normal pattern during the pandemic. For example, during Covid-19, Amazon's stock price (Amazon is classified as "consumer discretionary"; in another economic definition it refers to consumer cyclicals) increased significantly. Amazon's profits did not relate to economic boosting, since similar stocks made money before the pandemic

and lost huge amounts of money during it [8]. This situation was a specific condition that the country and consumers were faced with. We expected that belonging to a specific economic sector played an important role in companies' behaviors during said time-period. The question was about what the similarities and differences are between companies' behaviors during the two crises—Covid-19 and the 9/11 attacks. This case study investigated the impact of the Covid-19 pandemic on the behavior of stock markets and the returns of different economic sectors and compared it with the impact of the 9/11 attacks. Based on the theoretical framework derived from the efficient market hypothesis (EMH) and behavioral theory, we hypothesized that there would be a relationship between the Covid-19 pandemic and the 9/11 attacks on the behavior of the returns in different economic sectors.

7.2 Methodology and Dataset's Procedures

Three separate datasets were utilized in this research. CRSP, FF and Bloomberg. From Bloomberg dataset, we used the data from 8 months, from January 2020 to August 2020. Finally, we had two datasets, and hence two correlation networks were created based on the two crises. One network was created based on the excess returns of the companies (call it 9/11 ER-dataset) for the years 2000-2002 inclusively, and another was for companies from the year 2020 (Covid-19 ER-dataset) to visualize the 9/11 attacks and the Covid-19 pandemic, respectively. Technically, we had 36 months' worth of data for 9/11 ER-dataset and 8 months' worth of data for Covid-19 ER-dataset. To identify the significantly enriched input parameters for various stock market sectors, we used the

[8] https://www.nasdaq.com/market-activity/stocks/amzn/advanced-charting

following methodological five-step process. We used the first four steps for each of the datasets, and the last step was used to compare both the datasets at a high-level.

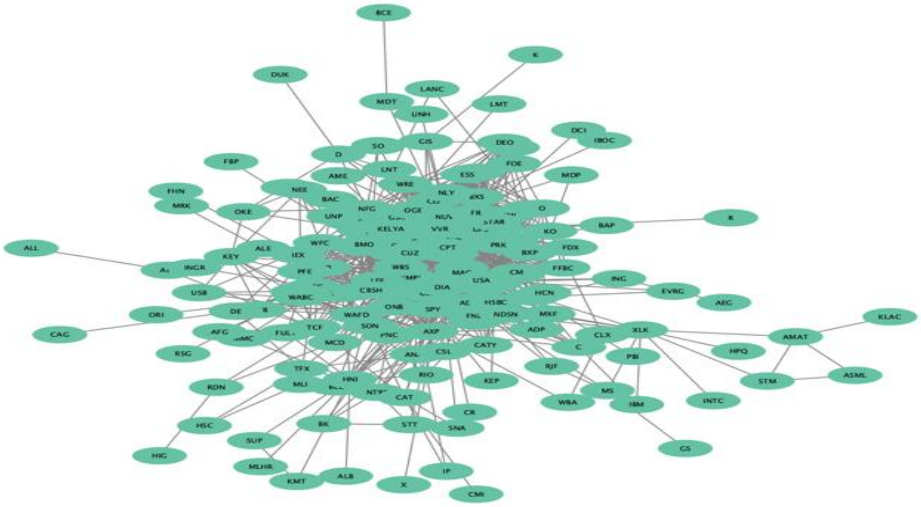
- Datasets preparation for both the events
- Create a correlation network based on excess returns and apply MCL clustering algorithm to generate individual clusters.
- Apply Hyper-geometric distribution-based enrichment analysis to find the significantly enriched input parameters for each cluster.
- Apply the population analysis to compare companies' clusters to see what parameters are significant for each cluster
- Comparing the significant parameters enriched for both the datasets.

In short, the overall process was explained as creating a correlation network based on some outcome parameter (in this case, it is excess return) and identifying the significantly enriched input parameters (in this case, the types of sectors) for each significant cluster. All the above five steps are explained in detail below.

7.3 Correlation Network Creation

The correlation network is a graph model, as shown in Figure 14, where each node represents a company in the stock market, and each edge represents the relationship between the stock returns of the companies.

Figure 14-Correlation network of 9/11 ER-dataset companies



The matrix of excess returns of the 9/11 ER-dataset, as shown in Table 7 for each company for 36 months, created the input matrix. Table 8 shows the corresponding correlation matrix created with the spearman-ranking correlations between the companies' excess returns given in Table 7. The correlation threshold considered was .90, which means that if there was a correlation coefficient ≥ 0.90 between any two companies, then the two companies were connected by an edge in the corresponding correlation network. Figure 14 is a correlation network created from the 9/11 ER-dataset.

Table 7-A sample input matrix of five companies for creating the correlation matrix

	$Month_1$	$Month_2$	$Month_3$	----	----	----	$Month_{35}$	$Motht_{36}$
AA00105510	-48.947	-58.647	-22.386	----	----	----	-10.456	-13.363
AA00195750	-37.157	-49.3981	-32.5038	----	----	----	-5.0578	-17.2143
AA00206R10	-52.5513	-54.2245	-36.3268	----	----	----	-0.9322	-15.8772

AA00282410	-50.8589	-54.2245	-40.3712	----	----	----	-7.4383	-19.6341
AA00915810	-52.7318	-52.5513	-35.8641	----	----	----	-11.9548	-13.8494

Table 8-A sample correlation matrix of five companies

	AA00105510	AA00195750	AA00206R10	AA00282410	AA00915810
AA00105510	1	0.693436	0.812098	0.749807	0.885199
AA00195750	0.693436	1	0.73333	0.715058	0.688546
AA00206R10	0.812098	0.73333	1	0.808752	0.739511
AA00282410	0.749807	0.715058	0.808752	1	0.7310171
AA00915810	0.885199	0.6885546	0.739511	0.731017	1

As mentioned in the previous sections, the clustering algorithm we used was the MCL. Again, similar to case study 3, inflation was a parameter (which is between 1 and 8) for the MCL that turned the subdivision of the bigger clusters to smaller clusters as it increased. An inflation value of 1 is too low, and 8 is too high. Therefore, we tested all the inflation values between 2 and 4 with 0.5 increments to identify clusters and find the optimal inflation value. The optimal inflation value was identified in such a way that the average of the last excess return difference between any two clusters was the maximum. An inflation value of 2.5 was identified as an optimal value that gave the ending highest

average excess returns difference among the clusters generated, as shown in Table 9. There are four columns in this table. The first column represents the cluster number; the second column represents the number of nodes in the cluster; the third column is the percentage of nodes in that cluster compared to the overall correlation of network nodes. The last column at the end is the average excess return (named as Avg. Return). From this table, we see that there were three clusters identified by the MCL at the inflation value of 2.5, with the Avg. Return difference of 5.645 between the clusters numbered 144 and 178.

Table 9-MCL clusters with the last average excess returns (9/11ER-dataset)

ROW	#Cluster	#Nodes	#Nodes (%)	EAER (%)
[1,]	178	21	5.556	-18.80
[2,]	144	56	14.815	-13.79
[3,]	91	25	6.614	-13.16

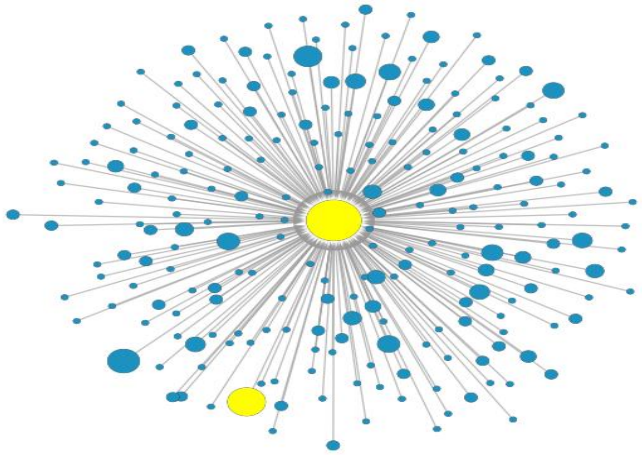
Similarly, from Table 10, we see that from Covid-19 ER-dataset, there were 7 clusters generated at the inflation value 3 with the maximum end average excess return difference being 115.8 between the clusters numbered 18 and 30. These two clusters are colored in Figure 15.

Table 10-MCL clusters with end average excess returns (Covid-19 ER-dataset)

ROW	#Cluster	#Nodes	#Nodes (%)	EAER (%)

[1,]	18	25	5.071	-12.55
[2,]	33	140	28.398	14.81
[3,]	29	47	9.533	16.09
[4,]	32	22		27.52
[5,]	27	93	18.864	29.02
[6,]	28	59	11.968	30.74
[7,]	30	43	8.722	103.24

Figure 15-Candidate clusters from correlation network of Covid-19 ER-dataset companies



7.4 Enrichment Analysis with Hyper-geometric Distribution

Enrichment analysis with hyper-geometric distribution given in [69], [72] was used to identify significantly overrepresented terms for a given gene set. Similarly, to identify the significantly enriched/overrepresented features (i.e., sectors in this case study), the enrichment analysis with hyper-geometric distribution applied for a given set

of stock market companies. As mentioned above, hyper-geometric distribution is a method of sampling without replacement and defined as follows.

$$P(X \geq K) = \sum_{x=k}^{min(K,n)} \frac{\binom{k}{x}\binom{N-K}{n-x}}{\binom{N}{x}}$$

Where P(X≥K) is the probability of examining at least k companies with a given finance feature value, N is the total number of companies in the background set (set of all companies), K is the total number of companies annotated with the given feature value, n is the total number of companies in the target set (companies in the given cluster), x is the total number of companies in the target set (given cluster) and annotated with the given feature value.

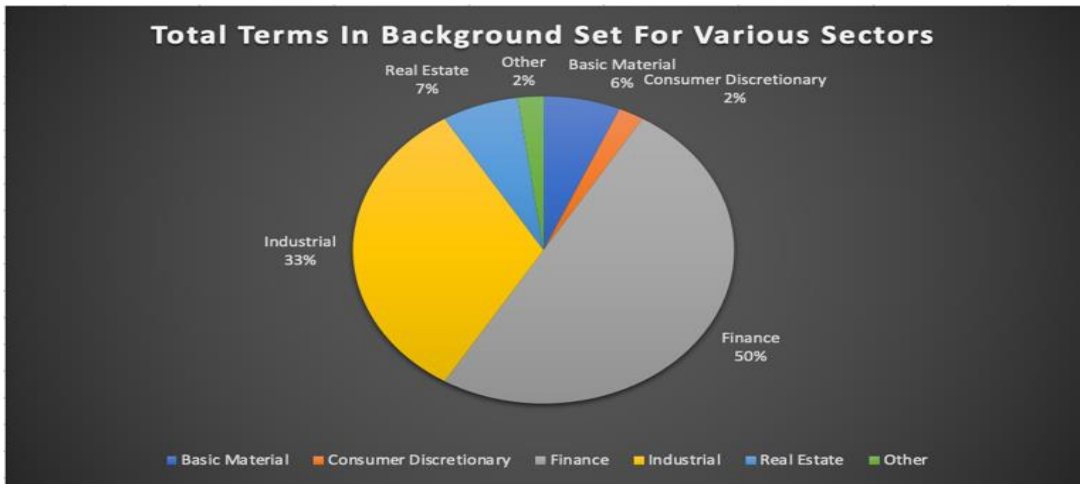
Table 11-Significantly enriched features/parameters for the given cluster 178 of 9/11ER-dataset compared to cluster 144

Feature	T.S	P_VALUE	R.S	B.S
Basic Materials	3	0.175231	0	3
Consumer Discretionary	1	0.547826	0	1
Finance	0	1	1	23
Industrial	13	0.000751	2	15
Real State	3	0.175231	0	3

Other	1	0.739511	0	1

Table 11 depicts the overrepresented sectors of companies of cluster 178. From this table, we see that there were six different sector types (or call them feature values, including the sector named “other”). Here, the sector is a feature, and each sector type is a feature value and a total of 46 companies (total count from Background Set column of Table 11 in the 9/11 ER-dataset) are distributed among these six sector types as shown in Figure 16.

Figure 16-Pie chart of companies of all sectors (9/11ER-dataset)



These 46 companies form a background set which were denoted by N. Out of 46 companies, 21 companies were part of the cluster under investigation (the total number of Cluster Terms column of Table 11) as shown in Figure 16. These 21 companies were considered as target sets, denoted by n. Then, we wanted to see if a particular sector type (feature value) was significantly enriched/overrepresented in the given cluster; we looked

at how many of the companies were labeled with a given sector type, and how many of them were part of the cluster under investigation. For example, from the pie chart shown in Figure 16, we see a total of 15 companies were of sector type “Industrial” (table 11 and figure 16). 13 companies out of 15 were significantly enriched with a p-value 0.000751 (see Table 11) with the sector type “Industrial”. The 15 companies formed the notation K (where K is the total number of companies annotated with the given feature value), and the 13 out of 15 companies form the notation x (where x is the total number of companies in the target set (given cluster) and annotated with the given feature value). For cluster 178, only the industrial sector was significantly enriched.

7.5 Analytical Results

Population analysis was used to compare the various clusters to see how they were enriched with different parameters (in this case, various sectors). As shown by Figures 14 and 15, we see that there were 3 clusters identified for the 9/11 ER-dataset, and 7 clusters identified for the Covid-19 ER-dataset. Nevertheless, from Table 12, we see that only two clusters were significantly enriched with at least one parameter/sector for each dataset. The CC (candidate cluster) column indicates this. From Table 6, we can also see that for the 9/11ER dataset, the finance sector and the industrial sector were significantly enriched compared to other sectors, with p-values 0.02397 and 0.000751, respectively. Similarly, for the Covid-19 ER-dataset the finance and energy sectors were significantly enriched, with the p-values 0.000756 and 0.000231, respectively.

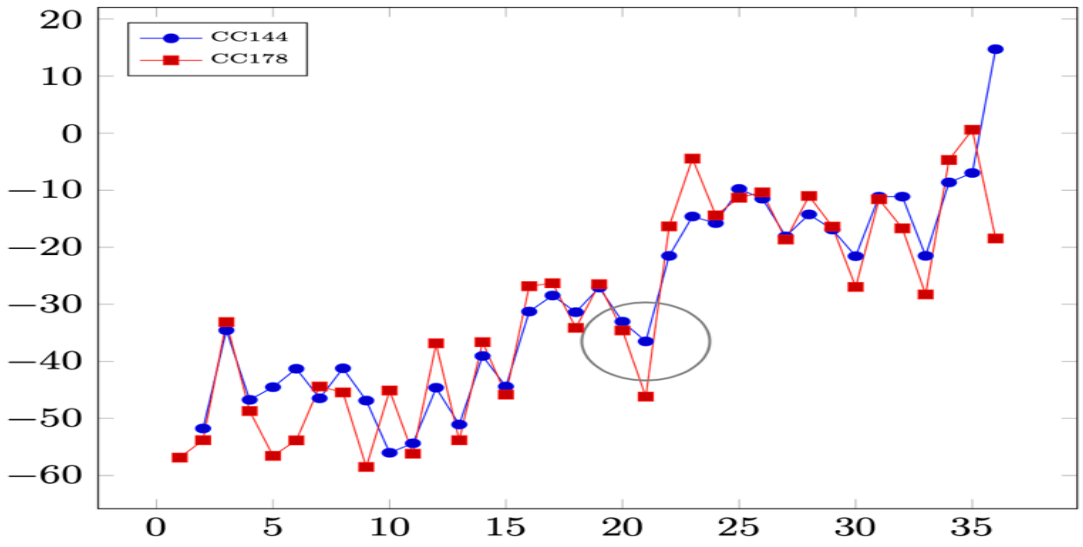
Table 12-Significantly enriched parameters for both the datasets

Dataset	Significant Parameters	Correlation Coefficient (ℑ)	#CC	P_Value
9/11ER	Finance Industrial	≥0.90	CC144 CC178	0.02397 0.000751
Covd-19ER	Finance Energy	≥0.90	CC18 CC33	0.000756 0.000231

7.5.1 Comparison of two disasters and their effect

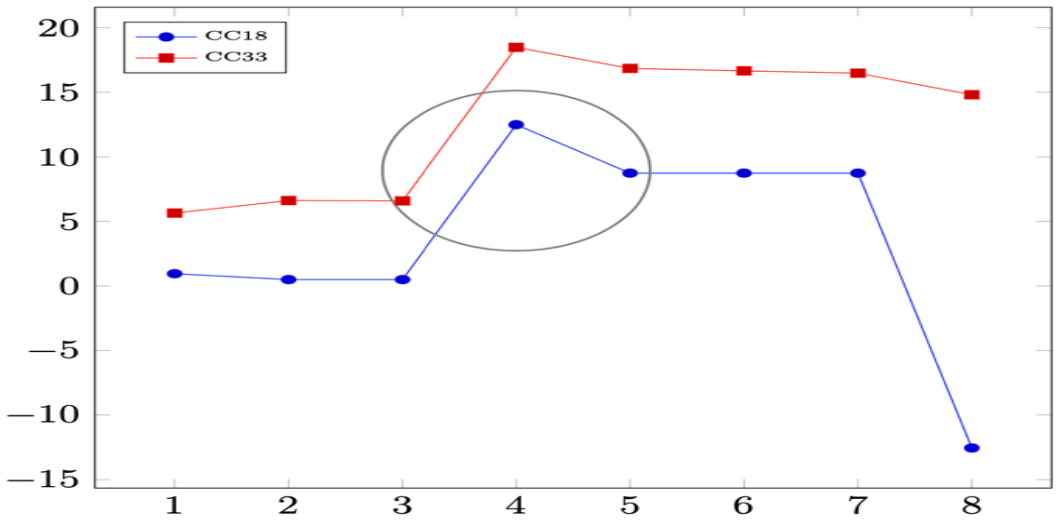
Figures 17 and 18 show the behavioral pattern of the excess returns for both datasets. Figure 17 is for the 9/11 ER-dataset. It shows the 36 months' patterns from January 2000 until December 2002. The attacks happened on the 21st month, September of 2001. Another thing to observe from this figure is that both the significantly enriched sectors follow the same behavioral excess return pattern after the attacks happened.

Figure 17-Behavioral patterns of cluster wise excess returns (9/11) (x=months, y=average return)



Similarly, from Figure 18, for the Covid-19 ER-dataset (it is an eight-month dataset, right from January 2020 until August 2020), we see that the first Covid-19 lockdown happened in late March (labeled in the figure), and again the excess return patterns increase their returns in the next month. Overall, both of the events' behavioral patterns were similar and companies' ER increased after the events took place.

Figure 18-Behavioral patterns of cluster wise excess returns (Covid-19) (x=months, y=average return)



7.6 Discussion

Several key findings were obtained from the similarity network model. Similar to the previous case study, the significance of the cluster that contained the finance sector was one of the main outcomes of the network. The reason behind this significant behavior in the financial sector is the same as case study 2. This finding was due to the significant influence of the Federal Reserve System on expected earnings of all members within the finance sector. Because of the shock to the US economy caused by the 9/11 attacks and the Covid-19 pandemic, the Federal Reserve significantly increased the supply of money within the economy. Year over year, increase in the US money supply, as measured by M2, was greater than 12% immediately after 9/11 and remained above 10% for the following four months [9]. The growth in the money supply after 9/11 was

[9] https://fred.stlouisfed.org/series/M2

significantly higher than the average annualized growth rate of 5.76% for all months prior to 9/11. This growth in available money provided an atmosphere of "easy money" for banks and financial institutions. The increase in funds from the Federal Reserve policy infused more loanable funds within financial institutions and decreased their cost of operations by lowering the cost of these loanable funds. Thereby, banks had the opportunity of increasing the amount of funds they loaned to their customers at a lower cost, increasing their profits. In addition to financial companies, 9/11 had a significant impact on the industrial sector. As the data shows, an unusually high correlation and clustering of returns occurred in the industrial sector after 9/11. Airlines, construction equipment, and industrial conglomerates' intrinsic values all dramatically declined immediately after 9/11 due to the uncertainty of future economic and global political shock. Note that all market trading was halted immediately after 9/11 and did not resume until one week later [10]. Immediately after trading resumed, industrial stocks faced punishment due to the uncertainty. However, in the months following 9/11, the uncertainty about the economy and the political environment subsided, resulting in a strong high correlated performance in market returns. Because the industrial sector was overly punished by investors on 9/11, the sector showed strong performance after the uncertainty subsided. Both the Financial and the Industrial sectors showed highly correlated returns through April of 2002. Like the economic and geopolitical shock of 9/11, the financial impact of the Covid-19 viral pandemic caused a high degree of uncertainty within the nation's economy. This uncertainty again created a high degree of return correlation within two sectors of the economy, the Financial and the Energy

[10] ttps://www.investopedia.com/financial-edge/0911/how-september-11-affected-the-u.s.-stock-market.aspx

Sectors. The Federal Reserve stepped in to address this shock with a major package of monetary stimulus [11]. Year over year, increase in the money supply, M2, increased by more than 20% immediately after recognizing the financial implications of Covid-19. Unlike previous monetary stimuli, this exceptional monetary growth continued for multiple months through the remainder of the year with an average annualized growth rate greater than 23.5%. This is the largest and longest monetary stimulus on record. The availability of "easy money" created by the Federal Reserve's increase in the money supply, along with a fiscal stimulus package, provided financial institutions with higher expected future earnings through cheaper and more accessible funds. Given the cheaper sources of funding and the incentive to create loans, the Finance sectors Net Interest Margin improved substantially. Thereby, equity prices for financial institutions moved in unison, causing a cluster effect from higher correlation among members of the sector. While financial institutions benefited from aggressive Federal Stimulus policy after the outbreak of Covid-19, the Energy sector did not see similar benefits because of the pandemic. The clustering of returns in the Energy sector was a result of higher correlations of negative price reactions within the Energy sector to the pandemic, and the tightening of social restrictions worldwide. Prior to the Covid-19 outbreak, the Energy sector was suffering from oversupply and a price war between Saudi Arabia and Russia. The Covid-19 economic shutdown and social restrictions exacerbated the problems within the Energy companies, which impacted operating margins. Crude oil prices slumped to a 20 year low to $16 per barrel at the end of April 2020[12]. At one time in the spring, crude oil prices traded at a negative $37 per barrel because of lack of storage

[11] https://www.brookings.edu/research/fed-response-to-covid19/.
[12] https://www.statista.com/statistics/326017/weekly-crude-oil-prices/

facilities and low demand. The low price of crude oil caused by the restrictions on economic activity affected profits across the energy sector, creating strong positive correlations and clustering of equity performance for the Energy sector. Based on the results obtained from the population analysis model, as shown in Figures 17 and 18, we concluded that the theoretical framework derived from the efficient market hypothesis and behavioral theory says that there was not a relationship between the Covid-19 pandemic and the 9/11 attacks on the behavior of the returns in different economic sectors because of the government support for this sector in both events.

7.7 Results

In this case study, we again tested our approach for analyzing the complex domain of financial markets. We showed that the proposed approach was very useful in providing accurate analysis of financial markets, particularly in dealing with the unpredictable factors often associated with global events. The constructed networks along with the associated enrichment analysis were shown to provide a valuable big data analytics tool to identify important patterns of the financial data, including those potentially hidden patterns. This approach makes it possible to identify those sectors impacted by the global events and how the government's decisions play a major role in the performance of such factors and their excess returns at the time of pandemics or other major events. We have shown that the behavioral patterns of the finance sector, specifically, during the two major events we studied were similar. As a result, we predict that finance sectors have good chances to recover from the impact of similar major events in the future. This concluded that our hypothesis that there was a relationship between two crises in terms of effecting the financial markets was true. We also showed that the

behavior of the financial markets under crises was not always consistent with the analysis provided by the behavioral theory and efficient market hypothesis frameworks.

Although the obtained results show that the proposed approach is very promising. We are well-aware of the limitations of the study. We used data associated with the U.S. market, even though the pandemic impacted almost every country in the world.

CHAPTER 8 PORTFOLIO SELECTION IN FINANCIAL MARKETS USING GRAPH MODELING AND POPULATION ANALYSIS

8.1 Introduction

In the competitive area that is the financial markets, data analysts strive to better interpret raw data to gain an edge on their like-minded competitors. Analysis of such data remains a challenge even among experienced data analysts as current methods of analysis each have advantages and disadvantages. In such a high-stakes area where it is easy to lose revenue, the demand for research on data analysis methods is ever-growing as researchers aim to solve common problems. A vast catalog of elements with little variation in the stock market exists to rile and cause issues for data analysts examining their subsequent difficult behavior. While stocks are individually monitored with the use of time-series data, analysis of the "why" factor in their behavior is hard to pinpoint due to the variety of factors influencing them. Therefore, the need to understand that the relationship among securities in the stock market has been increased by the availability of the data, and desire from investors to comprehend the function between stock movements. There are different approaches to analyzing the stock market from different disciplines. They come from financial overviews such as technical and fundamental analysis to statistical approaches [31], [73]. Network analysis is one of the popular approaches in the big data domain, especially in the social computing area, describing the characteristics or behaviors of variables in the complex network [74]. Network analyses

gives researchers the opportunity to reveal the information behind the relationship between elements. The use of network analysis by researchers seeks to find the underlying relationship between stocks in the financial markets. The motivation for using network analysis is that the performance of the stocks in the network is somehow correlated to different degrees. Their correlation can be either because of the nature of the market or the reflection of specific stock behaviors in different sectors. Community detection is an efficient way to extract information from the network. In community detection, stocks are grouped in different clusters depending on the structure of the network. Novel population analysis is one of the approaches that can be used to compare the behavior of a specific group/community of stocks with the rest of the groups. This case study introduces population analysis as an efficient and valuable approach utilizing graph theory properties to structure a well-organized portfolio in terms of diversity and comparability with popular indices. Networks are constructed from groups of nodes and edges, where edges represent the relationship between nodes. Constructing the edges between stocks could be considered implicit work since the correlation between the stocks stands on some well-defined attributes, such as returns, prices, and trade's volatilities over a specific period. In this regard, edges between stocks are created if the correlation is larger than a predefined threshold value such as 0.80. This case study presents three main analyses: population analysis using correlation and enrichment analysis, similarity network and community detection, and utilizing the structural stock market network using in-sample and out-of-sample analysis. The results of the analysis were compared against the benchmark.

8.2 Methodology

This case study examines the population analysis in a different time period containing normal and unexpected events. This case study tried to construct an algorithm utilizing equal weightage to the previously defined score of the two centralities. Hence, two centralities' scores were combined for each stock, and finally, the final centrality metric was created. Together, these centralities gave us a score that determined the overall position of stock on the network. Finally, using the centrality score from the algorithm, we further filtered the stocks by sector so that we could bring diversification to our portfolio.

8.3 Data Description and Preparation

Since this study tried to test and validate the proposed model in terms of in-sample and out-of-sample, there are four datasets and, accordingly, four correlation networks. The companies extracted from CRSP & FF from 2000-2004 (inclusively) and 2005-2009 (inclusively) were assigned for in-sample analysis, and companies from 2000-2009 and then 2010-2012 were assigned for out-of-sample analysis. At the end of filtering and data cleansing, there were 1411 companies extracted for each dataset. In each dataset, we looked at companies' Ticker, SIC, return, and TCap for all available stocks in the U.S.

The ticker is a one to four-letter code that provides information about a particular stock. Standard Industrial Classification[13] (SIC) codes are four-digit numerical codes that categorize companies' industries based on their business activities. It is worth noting that, in the datasets, 40% of companies belonged to the finance sector, and 60% belonged to

[13] https://www.investopedia.com/terms/s/sic_code.asp

the remaining sectors. The "returns" represent a monthly return for each stock. The excess return (ER) was calculated by subtracting the companies' returns from the risk-free value provided by FF. Risk-free[14] is a theoretical rate of return of an investment with zero risks. TCap is total capitalization for companies, and it was used to categorize companies into different deciles representing their size as small size and large size companies.

The table below represents the summary of the stock data from 2000-2012.

Table 13-stock information from 2000-2012)

#Stock	Excess Return	TCap	SIC
1411	Return-Risk Free	10 Categories (D1-D10)	12 Sectors

8.4 Network and Community Detection

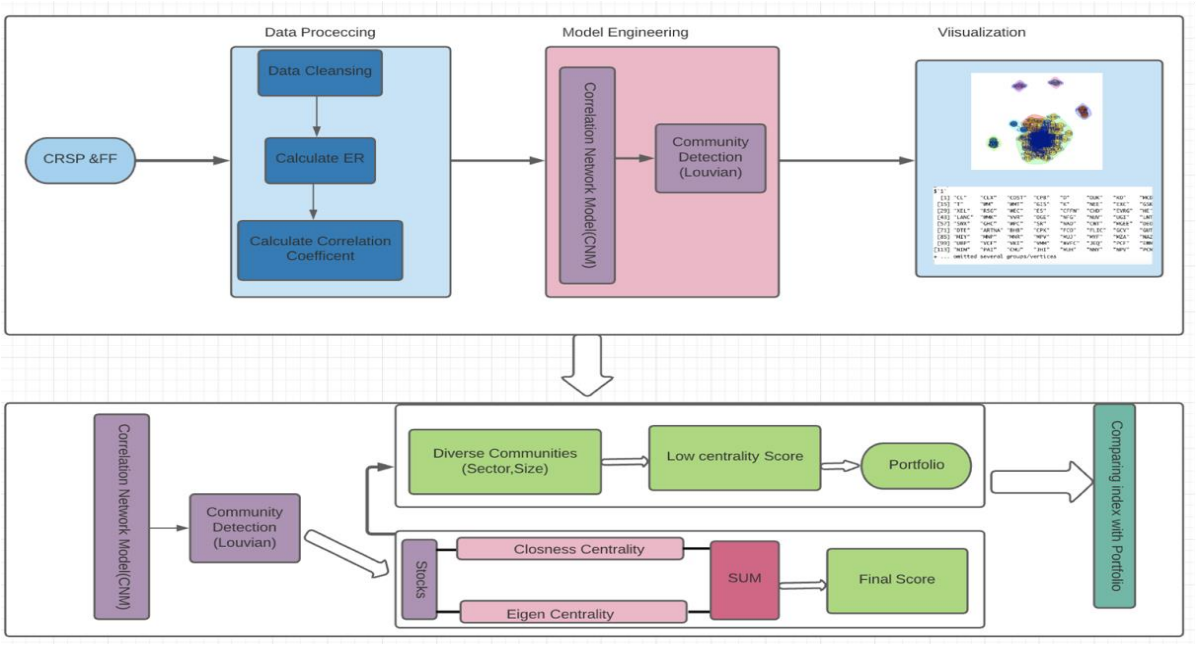
In modeling the correlation network and network analysis, stocks' tickers were represented as nodes, and edges between stocks represent the correlations of defined Excess return (ER) attributes over a selected time frame. Therefore, the correlation network created was based on the different correlation coefficients between stocks' ER. In this study, when the correlation was larger than the predefined threshold value such as 80%, 85%, and 90%, then the edges between nodes/stocks were created. As a result, the correlation network model manifested as an undirected weighted graph.

[14] ttps://www.investopedia.com/terms/r/risk-freerate.asp

Extracting information from the network is an important process in big data analysis. One way to find information from the network is community/cluster analysis. In this case study, communities were extracted from the networks based on the Louvain Algorithm. The Louvain algorithm [15] is a hierarchical clustering that recursively merges communities into a single node and executes the modularity clustering on the condensed graphs.

The figure below shows the modeling approach and architecture design from data preparation to building the portfolio.

Figure 19-Methodology overview



[15] https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/

8.5 Experimental Results

In this section, the analytical results will be presented in accordance with the order described in the methodology section above. This section begins with the way networks are constructed, followed by community detection, then enrichment analysis in in-sample and out-of-sample. The last section represents the portfolio selection and visualization of the analysis result.

8.5.1 Network Construction

Different correlation networks were created based on United States stock market companies from 2000-2012 (inclusively). Four correlation networks were constructed based on the merit of population analysis being applied to network analysis to reveal hidden information. Two of the networks were for in-sample analysis (data from 2000-2004 and 2005-2009), and two of them were for out-of-sample analysis (data during 2000-2009 and 2010-2012). Companies' tickers represent the nodes and correlation between the company's excess return (ER) as represented by edges. Excess return is a return on an investment minus returns on a risk-free investment. Correlation networks were constructed based on the highest correlation coefficient (90%) between stocks' ER.

8.5.2 Community Detection

To extract information from the complex networks, Louvain community detection techniques were applied to the networks. For further analysis, the most populated communities were selected from each network and its communities. Ten communities were extracted from the 2000-2004 network and nine communities were extracted from the 2005-2009 network. Three communities (1, 3, 4) containing the greatest number of

nodes were selected from the 2000-2004 network; and from network 2005-2009, communities 2,3, and 4 contained the greatest number of nodes selected for the enrichment analysis. The same community detection techniques were applied on the 2000-2009 and 2010-2012 networks. From the 2000-2009 network, out of seven communities, three communities (2, 3, 4) and out of nineteen communities from the network 2010-2012, two communities (15&17) were selected for enrichment analysis. Table 1 shows the communities' numbers and the number of nodes in different time-periods.

Table 14-Communities and Nodes (in-sample and out-of-sample)

Sample	Date	Communities	#Node
In-Sample	2000-2004	1	293
		3	198
		4	241
In-Sample	2005-2009	2	268
		3	171
		4	368
Out-of-Sample	2000-2009	2	180
		3	405
		4	333
Out-of-Sample	2010-2012	15	169

		17	163

8.5.3 Enrichment Analysis (in-sample)

The results of the enrichment analysis techniques used to find the significant parameters in each community showed that in communities 1,3, and 4, the finance sector had the highest number of shares compared to the rest of the economic sectors. Out of the three communities, community four was targeted as it had shares fairly distributed among the sectors, with 45% in the finance sector and 55% left for other sectors. In addition, community four was split halfway between large-size and small-size companies. In communities 1 and 3, the percentage split between large and small size companies was not even. In community 1, 90% of companies were grouped in small size, and only 10% belonged to large size; and in community 3, 40% belonged to small size and 60% belonged to large size companies.

Table 15-Enrichment Analysis (2000-2004)

#Community	Finance Sector	Large Size	Small Size
Community 1	83%	10%	90%
Community 3	39%	60%	40%
Community 4	45%	50%	50%

The result of enrichment analysis in the 2005-2009 communities showed that, in all three selected communities, the finance sector had the highest shares compared to other economic sectors. Meanwhile, in community 4, 50% of companies belonged to large size and 50% of companies belonged to small size categories; also, the share for the finance sector was fairly the same as the rest of sectors. In community 3, 60% were small size and 40% were large size and in community 2, 75% were small size and 25% were large size.

Table 16-Enrichment Analysis (2005-2009)

#Community	Finance Sector	Large Size	Small Size
Community 2	65%	25%	75%
Community 3	64%	60%	40%
Community 4	45%	50%	50%

8.5.4 Enrichment Analysis (out-of-sample)

The result of the enrichment analysis for out-of-sample in 2000-2009 showed that in communities 2,3, and 4, the finance sector had the highest number of shares compared to the rest of the economic sectors. Out of the three communities, community three was enriched with companies' size and sector. Community three had shares fairly distributed among the sectors, with 38% in the finance sector and 62% left for other sectors. In addition, community three was split halfway between large-size companies and small-size companies. In communities 2 and 4, the percentage was split between large size and

small size companies was not even. In community 2, 68% of companies were grouped in small size, and only 32% belonged to large size; and in community 4, 72% belonged to small size and 28% belonged to large size companies.

Table 17-Enrichment Analysis (2000-2009)

#Community	Finance Sector	Large Size	Small Size
Community 2	79%	32%	68%
Community 3	38%	50%	50%
Community 4	61%	28%	72%

The result of enrichment analysis in 2010-2012 communities showed that, in both selected communities, the finance sector had the highest shares compared to other economic sectors. Meanwhile, in community 15, 48% of companies belonged to small size, and 52% of companies belonged to large size categories; also, the share for the rest of sectors had a plurality of shares as compared to the finance sector. In community 17, 62% were small size, and 38% were large size.

Table 18-Enrichment Analysis (2010-2012)

#Community	Finance Sector	Large Size	Small Size
Community 15	18%	52%	48%
Community 17	53%	38%	62%

8.5.5 Portfolio Selection

From the portfolio calculation formula explained in the methodology section, all stocks in each community had one final score (equal weightage of Eigen and closeness centrality). From the enrichment analysis result, for in sample analysis in 2000-2004, stocks in communities 1 and 4 were selected, and from the 2005-2009 network, stocks in communities 2 and 4 were selected from portfolio calculation. These communities were selected since stocks belonging to those communities were the most diverse or least diverse stocks in the sense of economic sectors and the size of the companies. For example, within the in-sample 2000-2004, 50% of companies in community 4 were large size, and 50% were small size as compared to community 1 in which 10% were large size and 90% were small size. This condition was applied to selecting other communities in out-of-sample analysis. The table shows the communities selected for the portfolio selection process.

Table 19-Community Selection Criteria

Sample Type	# Community	Diversity
In-Sample 2000-2004	1 4	Least Most
In-Sample 2005-2009	2 4	Least Most

Out-Of-Sample 2000-2009	2 3	Least Most
Out-Of-Sample 2010-2012	15 17	Least Most

The result of our analysis showed that our model could outperform the S&P 500 in the communities with the most diversity in the sense of economic sectors and size of companies if companies had a low centrality score. For example, Figure 20 shows community 4 as a diverse community in 2000-2004 with high centrality scores, and Figure 21 shows this community when the centrality score decreased. Comparing Figures 20 and 21 shows that our model can outperform S&P 500 (Blue line= Our model, Red line=S&P 500—x-axis=year and y-axis=portfolio monthly return,) when stocks are selected from a diverse community with low centrality scores.

Figure 20-Community 4-High Centrality

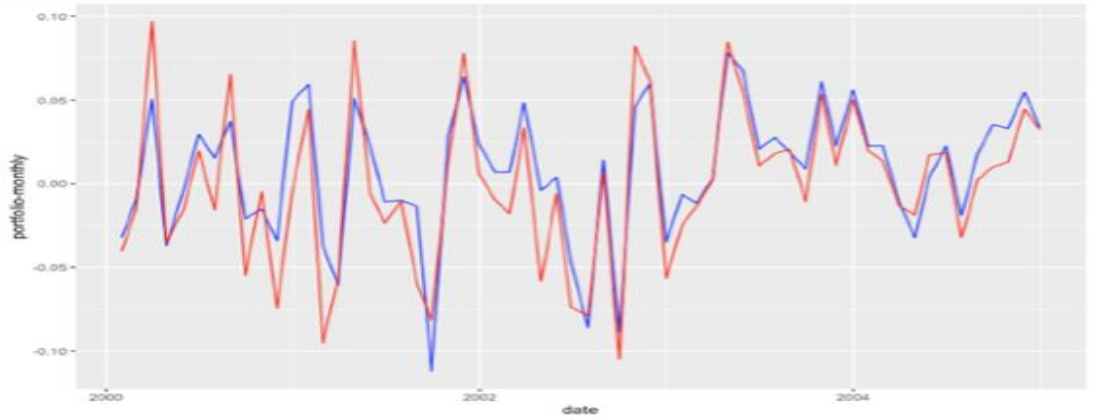
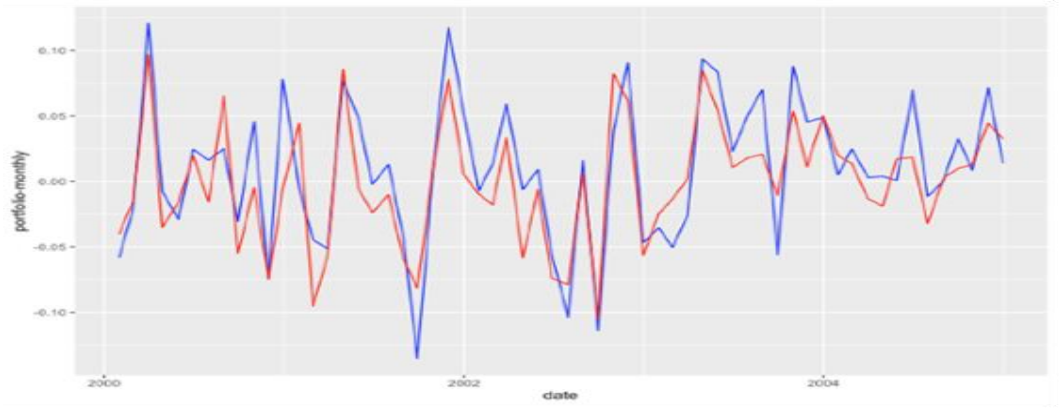


Figure 21-Community 4-Low Centrality



8.5.6 Visualization in Portfolio Selection in Different Targeted Communities

The figures below represent the side-by-side comparison between the least and most diverse candidate communities when their stocks had low centrality scores and high centrality scores. The results showed that constructing a portfolio by selecting stocks with low centrality scores from the most diverse communities could outperform the S&P 500 index. Meanwhile, constructing a portfolio by selecting the stocks with high centrality scores from the most diverse communities would not outperform the S&P 500 index. On the other hand, constructing a portfolio by selecting the stocks from the least diverse communities could not outperform the market, whether the stocks have low centrality scores or high centrality scores. However, the results in the least diverse communities proved that when centrality scores decrease, stocks performed better than high centrality scores' stocks.

Portfolio Selection- 2000-2004 (in-sample): Most Diverse Community

Figure 22-Community 4-High Centrality

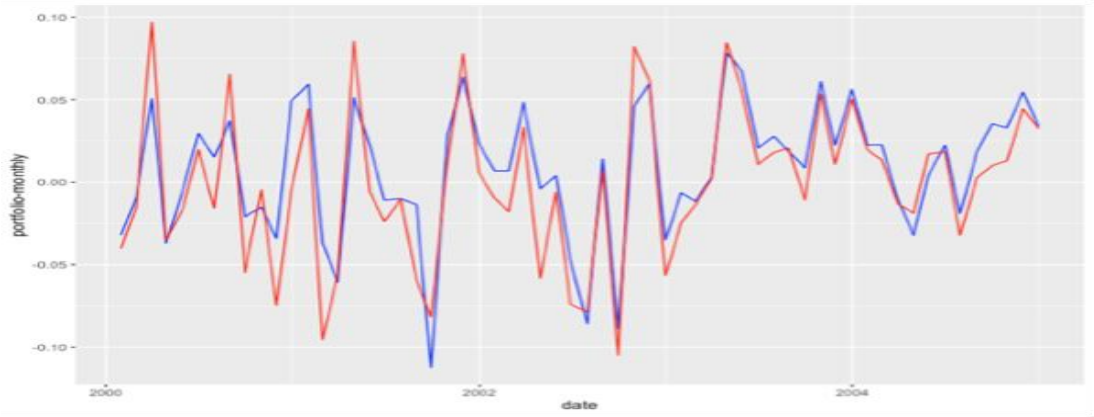
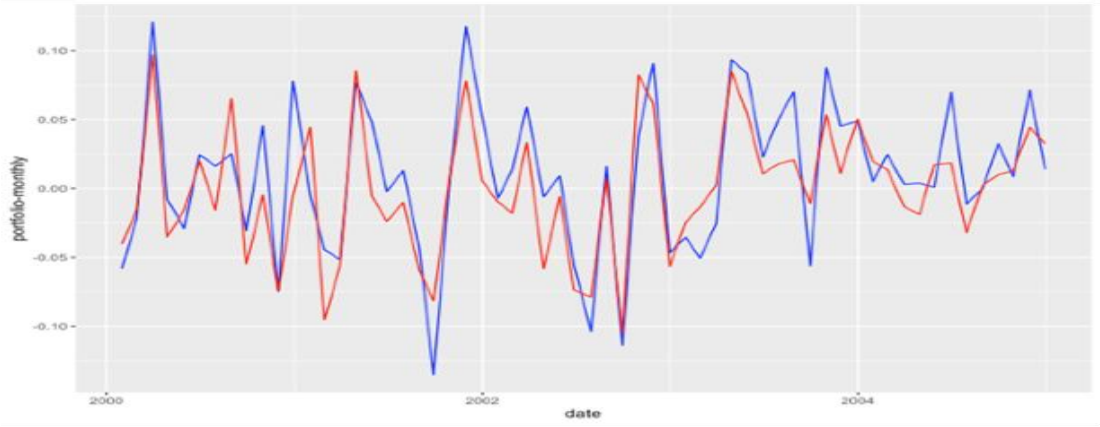


Figure 23-Community 4-Low Centrality



Portfolio Selection- 2000-2004 (in-sample): Least Diverse Community

Figure 24-Community 1-High Centrality

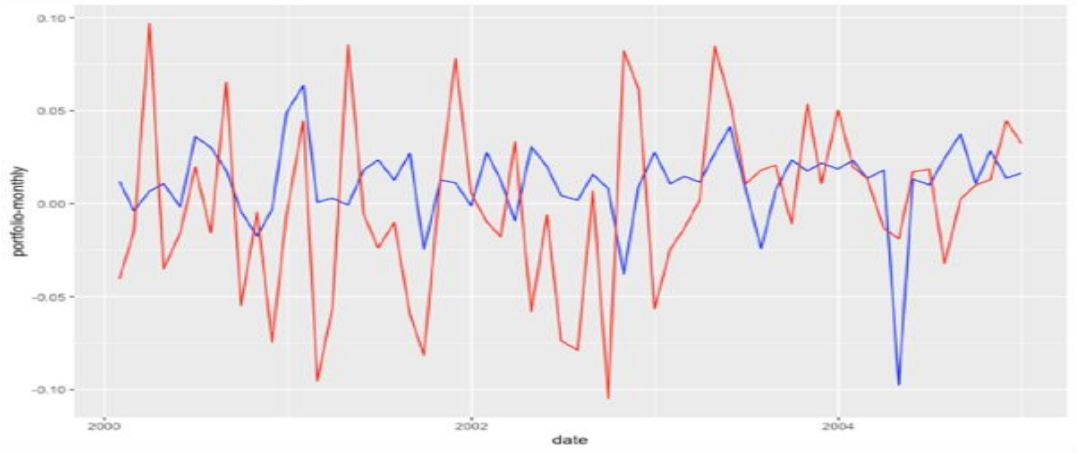
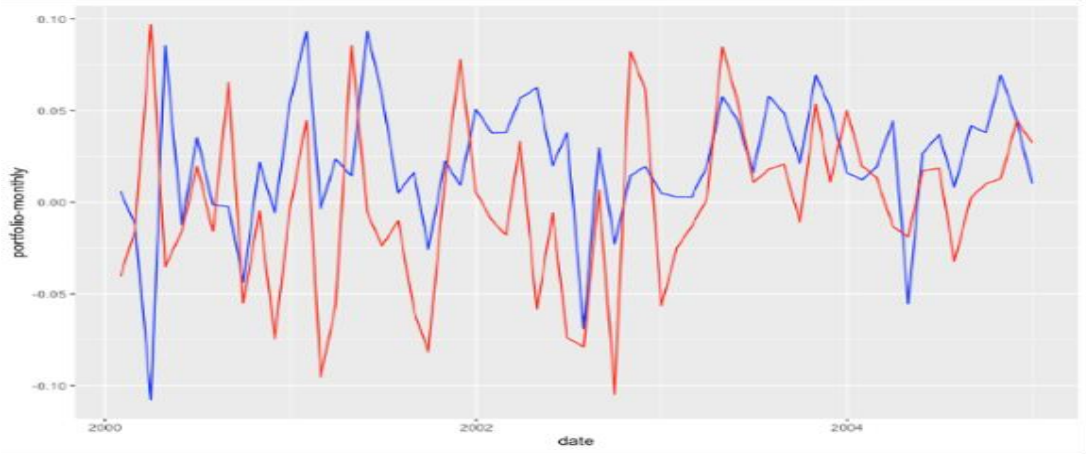


Figure 25-Community 1-Low Centrality



Portfolio Selection- 2005-2009 (in-sample): Most Diverse Community

Figure 26-Community 4-High Centrality

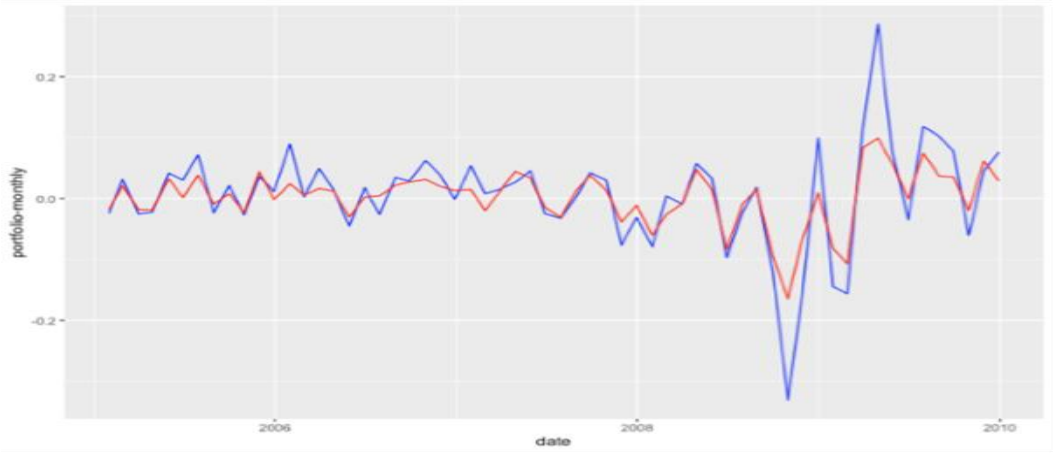
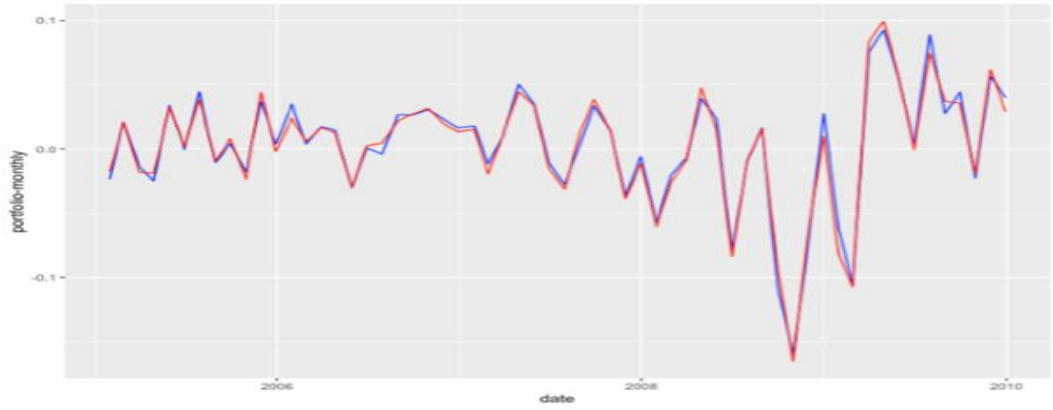


Figure 27-Community 4-Low Centrality



Portfolio Selection- 2005-2009 (in-sample): Least Diverse Community

Figure 28-Community 2-High Centrality

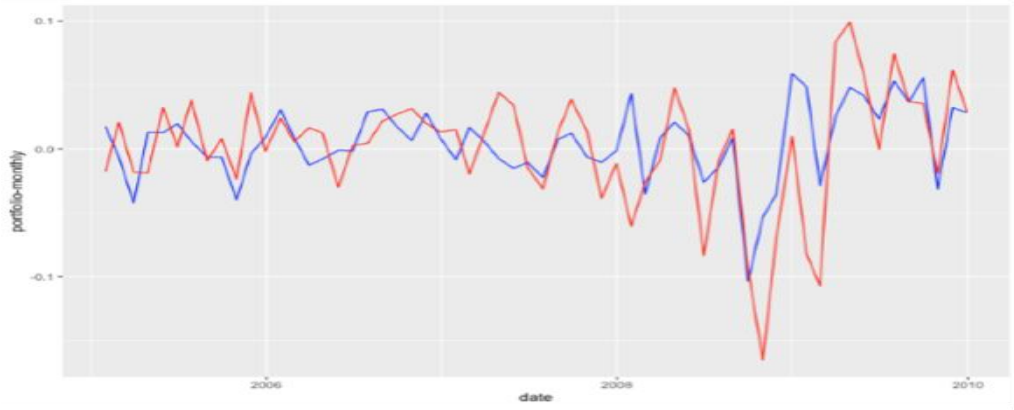
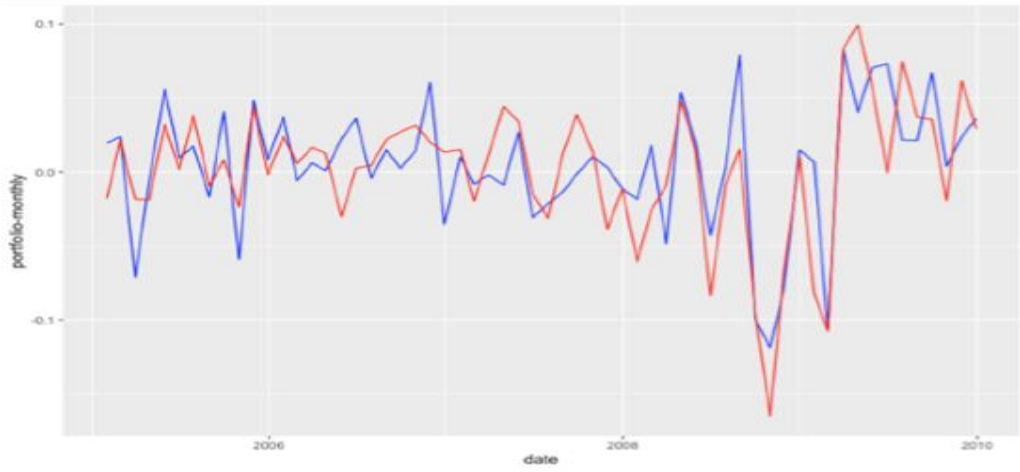


Figure 29-Community 2-Low Centrality



Portfolio Selection- 2000-2009(Out-Of-Sample): Most Diverse Community

Figure 30-Community 2-High Centrality

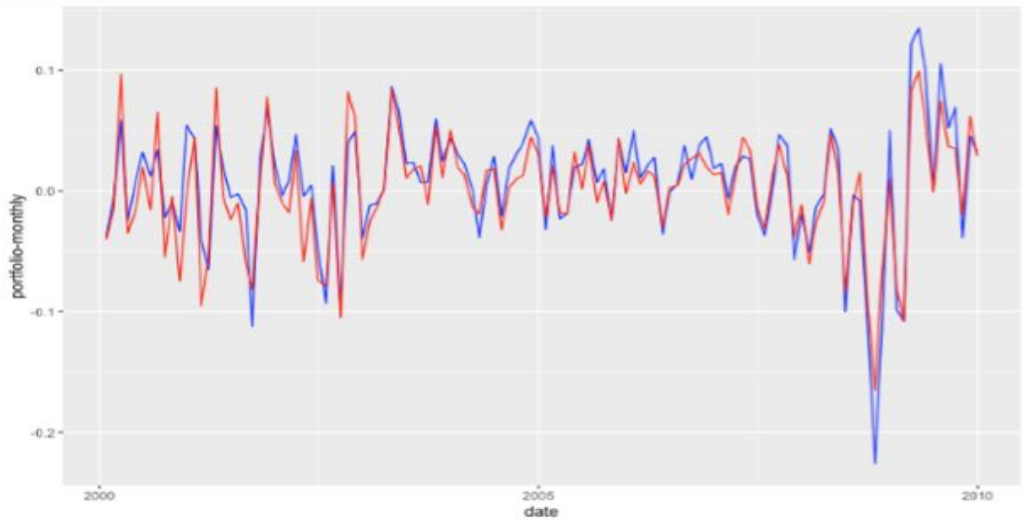
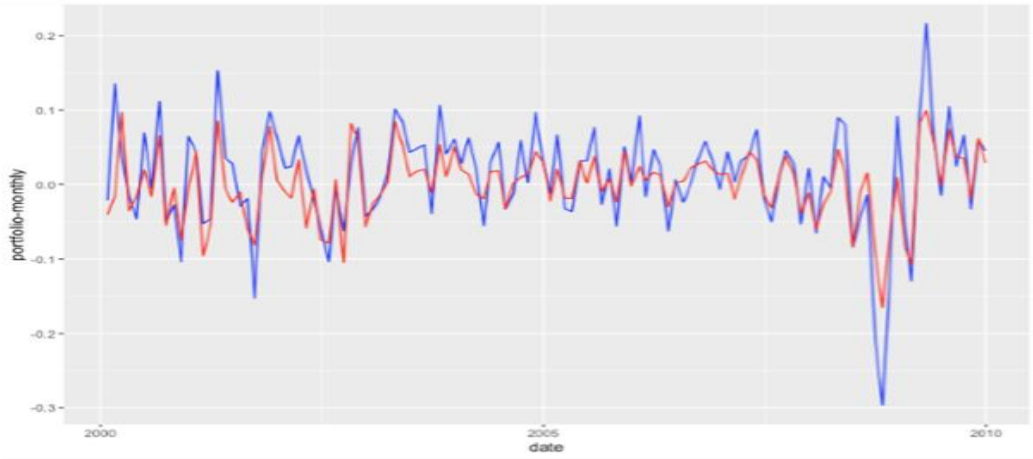


Figure 31-Community 2-Low Centrality



Portfolio Selection-2000-2009(Out-Of-Sample): Least Diverse Community

Figure 32-Community 3-High Centrality

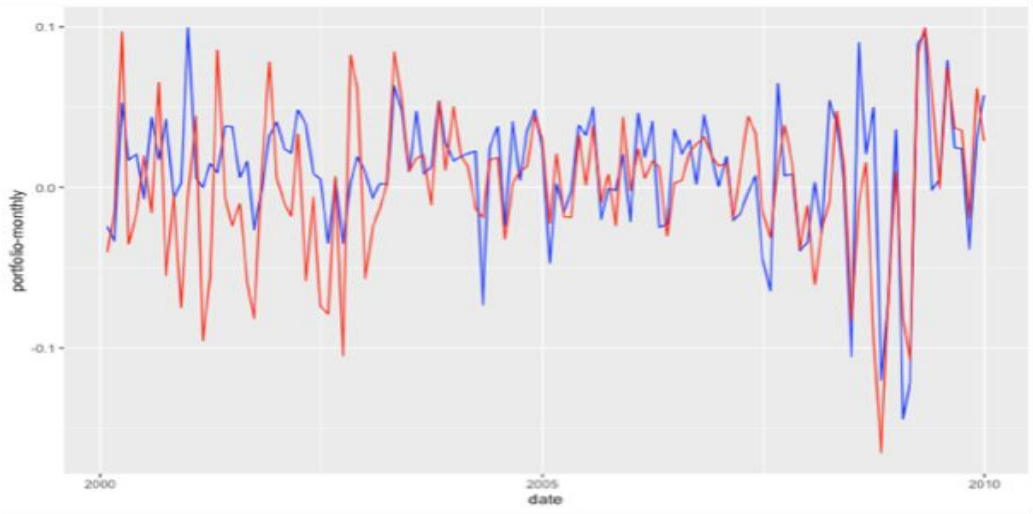
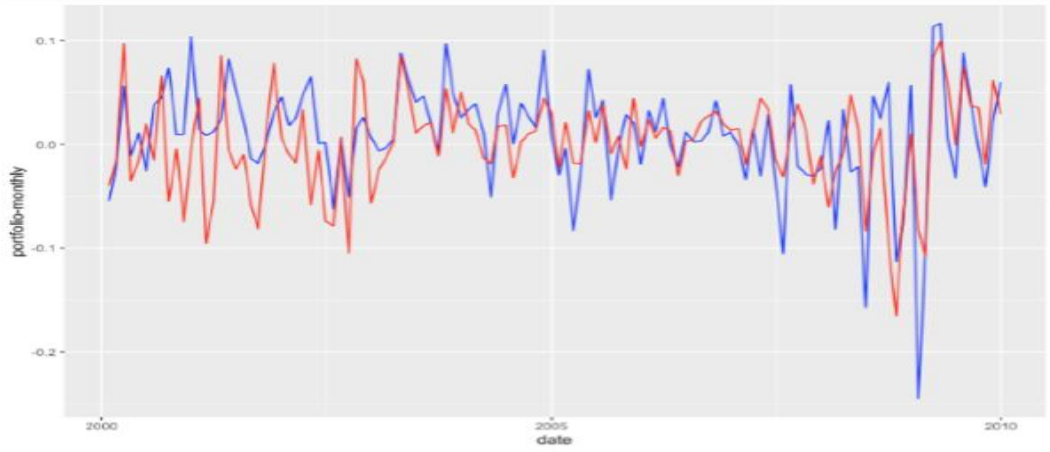


Figure 33-Community 3-Low Centrality



Portfolio Selection- 2010-2012 (Out-Of-Sample): Most Diverse Community

Figure 34-Community 15-High Centrality

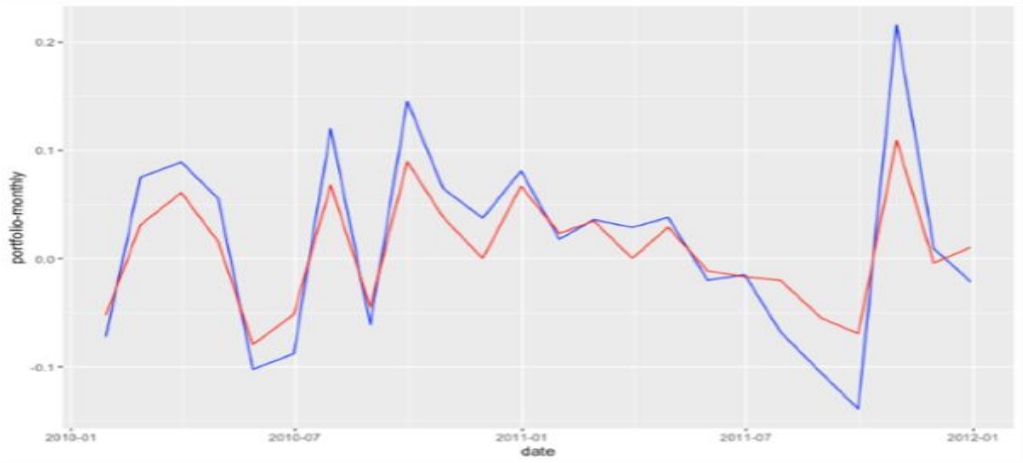
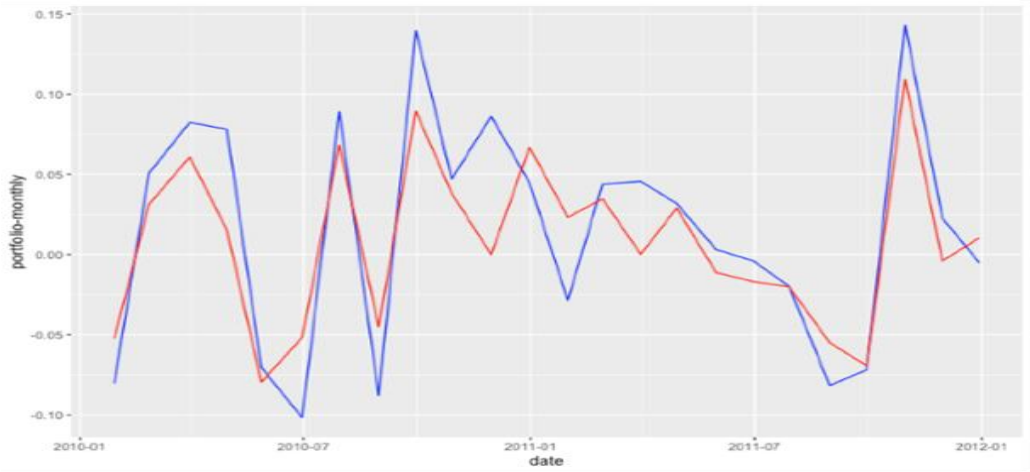


Figure 35-Community 15-Low Centrality



Portfolio Selection-2010-2012 (Out-Of-Sample): Least Diverse Community

Figure 36-Community 17-High Centrality

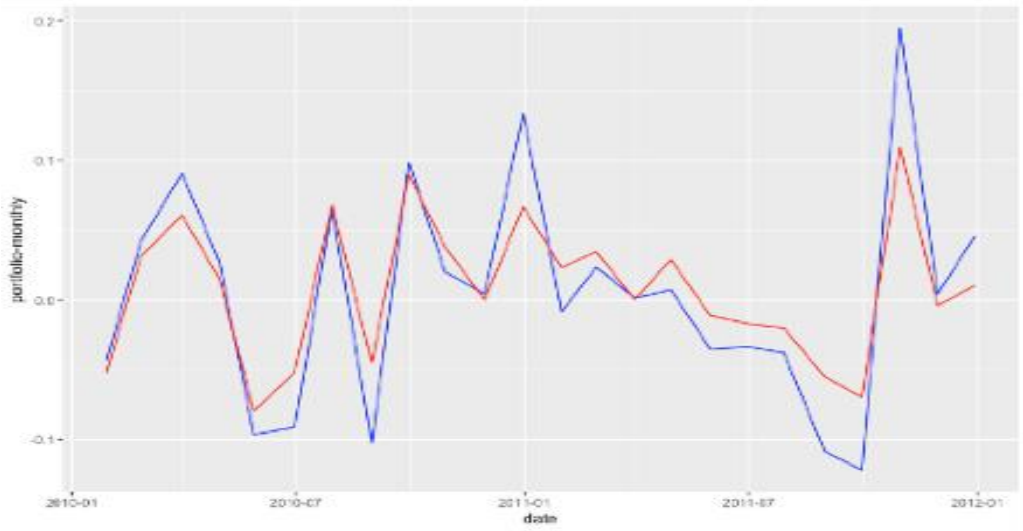
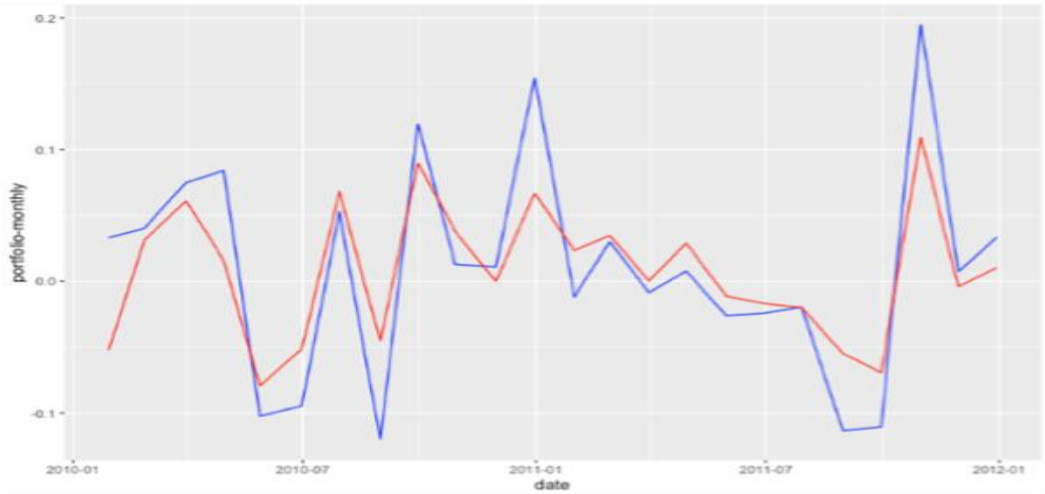


Figure 37-Community 17-Low Centrality



8.6 Conclusion & Discussion

The importance of a portfolio and how to select stocks is always a challenging task for investors. Researchers always attempt to apply different methodologies to create

the best portfolio, from the traditional portfolio theory to the modern portfolio theory. In this research, based on the concept of Modern Portfolio Theory (MPT), a combination of approaches was applied to the stocks' datasets. In this study, the concept of population analysis employed network and enrichment analysis to retrieve hidden information from the complex stock market. The stock market is a complex network, and to reveal information in it; community extraction is needed. The objective of population analysis is to compare the behavior of a group of stocks to other groups. Therefore, in this research, after creating the similarity/correlation network based on stocks' excess returns, communities were extracted by using the Louvain algorithm. Enrichment analysis was used to find the significant parameters in each community and its stocks to determine what was special about the community.

What was discovered was that nodes in the network that had lower degrees of centrality led to developing a portfolio of lower risk, with acknowledgment to the Modern Portfolio Theory. The empirical result in this case study revealed that the meaning of centrality measurement in network analysis in the stock market has a different meaning compared to social network analysis. In most networks, high central entities are the most important entities in the network; however, in this case study, we learned that high centrality was not something that researchers should look for when developing and building a portfolio with low risk. This research aimed to create a portfolio based on MPT, which prefers stocks with lower risks. Therefore, in our analysis, we tried to find the portfolio that could outperform the famous index such as S&P 500. We tested different datasets under in-sample and out-of-sample analysis, and

our analysis proved that selecting low central stocks in the most diverse communities could outperform the S&P 500.

CHAPTER 9 EVALUATING PORTFOLIO PERFORMANCE BY HIGHLIGHTING NETWORK PROPERTY AND THE SHARPE RATIO IN THE STOCK MARKET

9.1 Introduction

With the advent of the new portfolio theory in the late 1960s and the shift of industry owners investing in diversified assets to mitigate the consequences of risk, the competitive environment in the dynamic business world gradually narrowed. In addition, the dramatic growth of the level of communication and rapid exchange of information, along with the various complexities of the coming decades, intensified business competition. The term portfolio is a combination of stocks with other assets that an investor has purchased. In simpler terms, the “portfolio” means forming a combination of different shares and not investing in one share, which is an intelligent measure to reduce the risk of investing in the stock market. To succeed in the corporate stock market, choosing the right approach and maintaining coherence and order, like any other economic market, is important. Without a strategy, investing will only be unplanned buying and selling, affecting the investor’s capital and profit or loss. Therefore, portfolio investing is a critical and vital decision for individuals and legal entities, and portfolio diversification is a technique for managing risk and capital. In this case study, we used population analysis by employing a correlation network model to extract communities and select a potential portfolio that could outperform the market.

The economy's behavior is not isolated from the behavior of individuals; therefore, the economy as a network works as a state of transformation. Analysis of the network involves the recognition of which vertices are connected to others in a graph. Each stock is assumed as the graph's vertices, and edges represent the relationship between vertices. Applying population analysis helps us compare individual data points with other data points in different communities regarding different performance levels. Population analysis allows us to compare two or more communities of companies with respect to one or more enrichment parameters. The result of this analysis enables us to discover the parameters that significantly affect a community [34]. In this case study, different correlation networks were created based on the different datasets in different time periods. The potential portfolios were selected based on network properties, centrality measurements, and the Sharpe ratio. In the next step, the performance of the potential portfolios was compared to the S&P500 to check if those potential portfolios can predict the market. This case study attempted to evaluate the presented model's ability to identify the stocks with the most diversification in terms of economic sectors and company sizes.

9.2 Methodology

This research is an interdisciplinary work that took advantage of big data analysis associated with knowledge in the financial domain. The methodology overview is as shown below:

- Computational analysis: Creating a correlation network model and assessing its property.

- Financial analysis: Examining the Sharpe ratio focusing on the financial theoretical framework.
- Comparison: Comparing the portfolio performance with the benchmark.

To summarize, first, the centrality measurements (Betweenness, Closeness, Eigen centrality) were calculated for stocks in communities extracted from the correlation network. Second, based on assessment of centrality scores, a collection of stocks as the potential portfolio were selected, and finally, the portfolios' Sharpe ratio was calculated to check the portfolio performance against the benchmark.

9.3 Data Collection and Procedures

For analysis, the initial dataset was divided into four datasets for every five years, starting from 2000 and ending with 2019. Different correlation networks were created from different time periods depending on different sets of data. For example, one correlation network was created based on the excess returns of the companies for the years 2000-2004 inclusively, and another was a network for companies from the year 2005-2009. To avoid bias selection, all companies existing in each five-year dataset were included in the analysis. Therefore, there was a range of 4000 to 6000 companies from different economic sectors and sizes in each dataset. Each dataset contained companies' Ticker, excess return, and companies' economic sectors.

9.4 Network and Community Detections

To test the result of our proposed model and check to see how companies and their returns volatility behaved during the time, we divided the data into different sets of 5 years. After constructing correlation networks and applying the Louvain algorithm, the

communities were extracted from the network [76] . The table below shows the number of nodes in each network followed by selected communities' nodes. For example, in 2000-2004, there were 4280 nodes in the network and out of 4280 nodes, 3389 nodes were grouped in different communities. Communities 1,2 and 3 were selected for further analysis since out of 3389 nodes, 3269 nodes were distributed in those three communities.

Table 20-Number of nodes in each network (N.N), number of nodes in selected communities (N.C), Selected Communities (S.C) and total number of nodes in selected communities (T.N.S.C)

Year	N.N	N.C	S.C	T.N.S.C
2000-2004	4280	3389	1,2,3	3269
2005-2009	4083	3590	1,2,3	3539
2010-2014	4489	1495	1,2	884
2015-2019	5211	3481	1,3	2228

This research relied on population analysis based on enrichment analysis on communities extracted from the network. For the robustness analysis, different correlation coefficients, Betweenness, Closeness, and Eigen centralities were all examined within each community. Enrichment analysis as the in-depth analysis was

applied in each community to get more information about stocks characteristics, such as size and economic sectors.

9.5 Results

In this study, five correlation networks were created for datasets: 2000-2004, 2005-2009, 2010-2014, and 2015-2019. In the process of creating the correlation networks, different correlation coefficients were tested to find the network containing stocks that had the highest similarities in their excess returns. Since the datapoints were distributed normally, the Pearson correlation coefficient was used in constructing the networks [74]. For avoiding sample bias issues, in the process of filtering and cleaning the data, stocks that existed in each five-year range were selected regardless of presenting in another datasets. Extracting knowledge from complicated networks is not an easy task, therefore the Louvain algorithm was applied in each network and communities with the highest number of nodes selected for the further analysis. After measuring the degree of the centralities, another filtering step was performed for each community. Stocks in each community were divided into subcommunities with high and low degrees of centrality. According to the stocks' characteristics in each of the subcommunities, subcommunities with high and low degrees of centrality were considered as potential portfolios. Further analysis, based on enrichment analysis, showed that subcommunities with low degrees of centrality contained stocks that had higher diversity in the sense of companies' size and economic sectors, meaning that subcommunities with a low degree of centralities had stocks that belonged to most of economic sectors (range of 12 economic sectors) and a fair range of large and small size of companies.

The potential portfolios were compared against the benchmark. The results showed that portfolios containing stocks with low degree of centrality could outperform the benchmark compared to higher-central stocks. Figure 38 shows that portfolios containing stocks with low degrees of centrality could outperform the benchmark compared to portfolios containing stocks with high degrees of centrality (Figure 39).

In Figure 38 and 39, Blue Line is a potential portfolio and red line is benchmark.

Figure 38-Low-central stocks-2000-2004

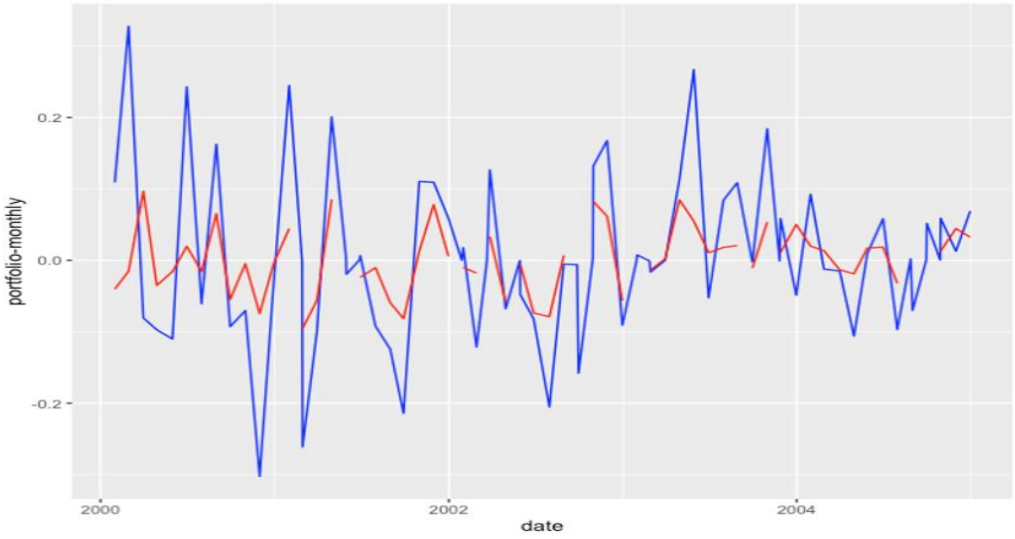
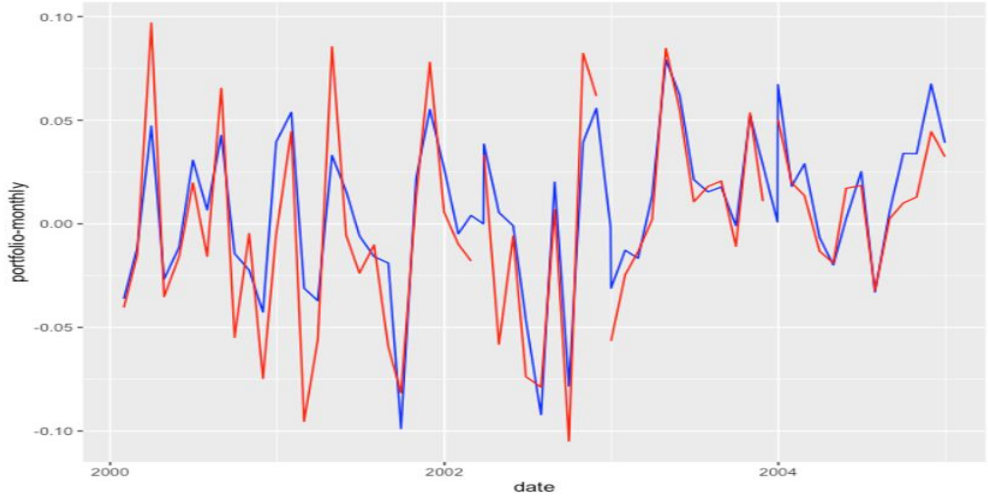


Figure 39-High-central stocks-2000-2004



This result was consistent for all potential portfolios selected from correlation networks for the time periods 2005-2009, 2010-2014, and 2015-2019 and their communities.

In the next step, portfolios' Sharpe ratios were measured for each high- and low-central subcommunity. The result showed that the Sharpe ratio for low-central stocks was slightly lower than high-central stocks; meaning that they were less profitable than higher-central stocks. However, they could outperform the market with less deviation from market movements.

9.6 Discussion and Conclusion

Applying population analysis employing correlation networks, community detection algorithms, and enrichment analysis proved that this model could predict the benchmark trend. Different centralities were measured for stocks in each community, and a specific algorithm was constructed based on the equal weightage to create the final

centrality score for each stock. We examined different portfolios categorized based on low and high final centrality scores. Enrichment analysis showed that low central stock portfolios had higher diversity in size and economic sectors. Our model identified portfolios with low final centrality scores and greater diversity as candidate portfolios that could predict the market trend better than portfolios with high final centrality. This research concludes that the general statement about the meaning of centrality measurements is not always correct, meaning that nodes with high centrality scores are not always the important entities. This research found that the importance of nodes did not rely on high centrality measurement, but also on the network model structure. In this regard, to reduce the risk of portfolio profitability and be able to predict the market, we must choose stocks for the portfolio that have a low degree of centralities. Since the Sharpe ratios for portfolios containing stocks with high degrees of centrality are slightly higher than portfolios containing stocks with low degrees of centrality, future research should focus on the optimization strategy as it changes the weight of stocks in low-end portfolios to increase Sharpe's ratio. Therefore, the portfolio will predict the market well and have a good amount of profitability.

CHAPTER 10 VALIDATION AND ROBUSTNESS ANALYSIS

After designing different case studies and stablishing a well-defined model, the next step is exploring the remaining building blocks of this study. The proposed model can construct a weighted portfolio containing group of stocks that can outperform the benchmark. In this section, the final stages of the method—designing and testing final dataset, method validations, and robustness analysis—are presented.

10.1 Test the Model 2019-2021

The same procedures highlighted in case study five were repeated for years 2019-2021 inclusively. There were 4867 companies that were part of the network; after applying community detection, communities 2, 4 and 16 were selected for further analysis. Between these three communities, community 16 was the most diverse community in terms of economic sector and size. In the next step, different centralities were measured, and another filtering step was performed for community 16. Stocks in this community were divided into subcommunities with high and low degrees of centrality. In the subcommunities, companies with high and low degrees of centrality were considered as potential portfolios. The potential portfolios were compared against the benchmark. Proving the method's functionality, the results showed that portfolios containing stocks with low degrees of centrality could outperform the benchmark, compared to higher-central stocks. Also, enrichment analysis showed that subcommunities with low degrees of centrality contained stocks that had higher diversity in terms of companies' size and economic sectors, meaning that subcommunities with

low degrees of centralities had stocks that belonged to most economic sectors (range of 12 economic sectors), and a fair range of large and small sized of companies.

Figure 40 shows that potential portfolios containing stocks with low degrees of centrality could outperform the benchmark, compared to portfolios containing stocks with high degrees of centrality (Figure 41). (Blue line= Our model, Red line=S&P 500)

Figure 40-Low-central stocks-2019-2021

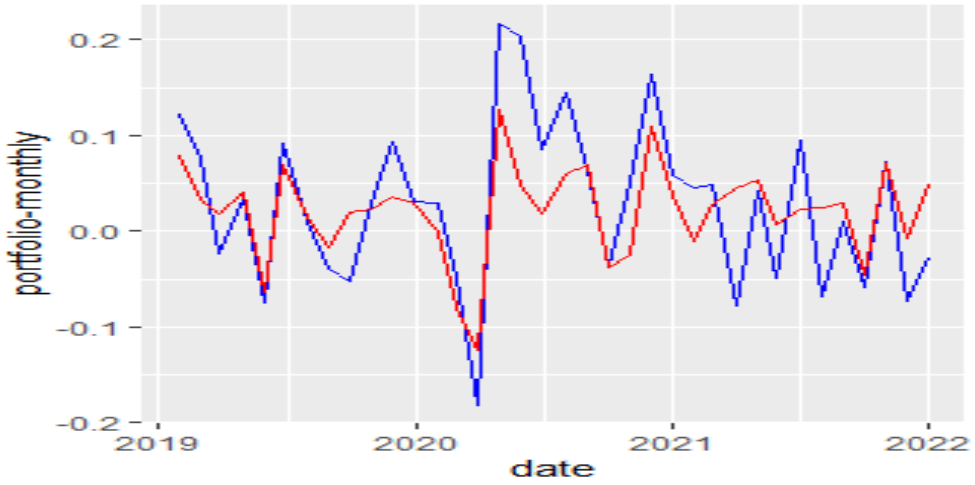
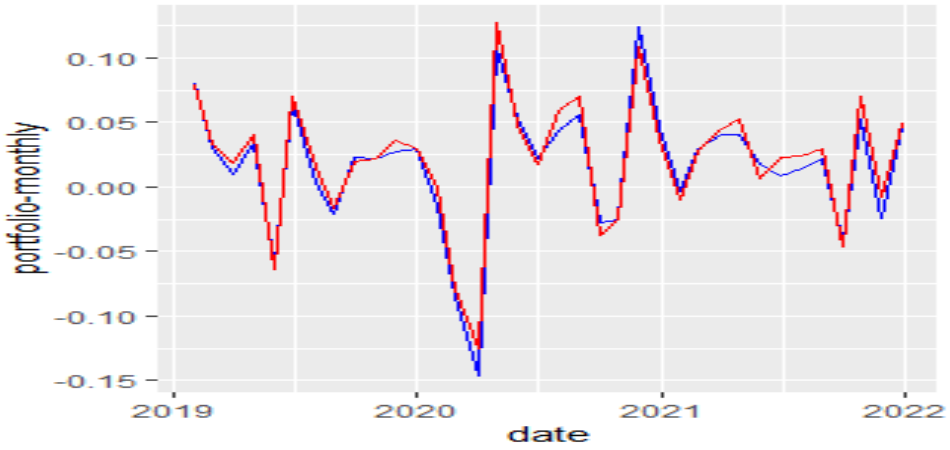


Figure 41-High-central stocks-2019-2021



10.2 Sharpe Ratio Evidence

After finding the potential portfolio and testing them against the benchmark, the Sharpe ratio was calculated for each subcommunity. The results of the Sharpe ratio calculations are presented in the table below.

Table 21-Sharpe Ratio

Year	2000-2004	2005-2009	2010-2014	2015-2019	2019-2021
Low Central	0.4394562	0.03462224	0.2478249	0.126506	-0.35
High Central	0.4465492	0.03950506	0.3159342	0.1013317	-0.53

Table 21 shows the Sharpe ratio calculations for potential portfolios in low and high central subcommunities during the years from 2000 to 2021. The result shows that even though portfolios containing low central stocks can outperform the benchmark, their

Sharpe ratio is lower than portfolios containing high central stocks. It is worth noting that the Sharpe ratio for low central stocks in less diverse subcommunities was lower than diverse subcommunities. For example, from 2005-2009, the Share ratio for low central stocks in diverse subcommunities was small, but still positive; and the Sharpe ratio for less diverse subcommunities in the portfolio containing low central stocks was -0.04325032 (not presented in Table 21). Therefore, the result of the analysis showed that the proposed model validates its notion that investors need to select the stocks for their portfolio from low central subcommunities that carry the highest amount of diversity. Now the questions that arise may concern how we can improve the Sharpe ratio for low central stocks. The answer for this question is presented in the chapter 3, the section on weighted portfolios. By using the weighted portfolio formula, the Sharpe ratio for portfolios containing low central stocks increased more than portfolios containing high central stocks.

10.3 Robustness Analysis

To check the robustness and validate the result, networks were constructed based on different datasets for different periods**.** The datasets were designed to be continuous and overlapping. The proposed approach examined robustness based on different datasets in-sample and out-of-sample to avoid bias issues. For example, in case study three, different datasets were formed to check the method's capability of handling different datasets while there were overlaps. For example, the method tested on data from 2000-2004, 2005-2009, 2000-2010, 2000-2012, etc. The purpose of robustness analysis was to show

that the obtained significant parameters were not dependent on the correlation coefficient, specific clustering algorithms and fixed period datasets. The validation for this study was tested based on the results of previous studies and experts' knowledge. Their cross-checks involved checking the in-sample and out-of-sample analysis to utilize the structural stock market network to develop a portfolio selection process. Therefore, the validation and robustness analysis proposed a correlation network model that tested different datasets and correlation coefficients, tested different clustering algorithms, applied graph properties, used population analysis, and performed the analysis result against well-known benchmarks, such as the S&P 500.

CHAPTER 11 DISCUSSION

11.1 Conclusion

This study is a comprehensive study based on different perspectives, either financial, big data, and so forth. In the present study, the novel approach titled population analysis was introduced as a comprehensive method for analyzing the stock market. This new approach could transform industries from an ad hoc approach to a more futuristic, data driven approach. This approach moves away from individual analysis and towards community/group analysis. Population analysis employs correlation networks, community detections, and enrichment analysis to find enriched parameters in various communities. The networks and graphs are analogous to the structure of the data in a way in which extraction of information is easier.

The methodology for this research was developed based on the results of the analysis from different case studies. We tested the model in different circumstance, such as the 2008 economic crash and the Covid-19 pandemic. After testing the model, we then tried to implement different approaches of portfolio construction as basic components of the portfolio management process, and to define a realistic method to determine the optimal strategy. People who are looking to invest and form portfolios in investment companies usually have a certain risk tolerance threshold; and portfolio managers are required to form their investment portfolios in a way in which the risk level does not exceed a certain level over time.

In this research, stock centralization features were used as criteria to determine stocks outperforming benchmark. To determine the outperformed portfolios, the subcommunities extracted from communities were categorized as low central and high central stocks. These high and low central stocks were then considered as candidate portfolios. The next step was examining the performance of these candidate portfolios and comparing them with the S&P 500 benchmark. The subcommunities containing low central stocks proved that they have more diversity in the terms of economic sectors and size and could outperform the above-mentioned benchmark, while the high central stocks proved otherwise.

As further evidence, the results of the analysis in [78]–[81] showed that higher centralized stocks have a stronger relationship in terms of the correlation of returns with other existing stocks. Therefore, it can be expected that changes in stock returns in the market network will have a more significant impact on the returns of highly centralized stocks. The result of this argument is the prediction of systematic risk and higher total risk for highly concentrated stocks. Therefore, if the investor forms a portfolio within the theoretical framework (Markowitz theory), he will move his capital away from stocks with higher risk. Consequently, it can be concluded that there is an inverse and significant relationship between centrality and shared weight in the framework of the Markowitz model. This statement is not just a prediction, but proof that the inverse relationship between stock shared weight in the Markowitz model and its centrality can be proved mathematically.

In the next step, calculating the Sharpe ratio showed that the Sharpe ratio for potential portfolios containing low central stocks was slightly lower than portfolios containing high central stocks. Therefore, at the final stage of our analysis, our proposed model moved from equal weighted portfolio to weighted portfolio. Meaning that, the weight for the stocks in potential portfolios changed in a way so that subcommunities with lower central stocks had higher Sharpe ratios, compared to subcommunities containing high central stocks.

Using this approach, investors can choose stocks for portfolio in more knowledgeable and calculated way, as compared to randomly checking benchmarks. It was found that the application of this strategy is more trustable than previous research suggests. Comparing the findings of this study with the findings of previous studies, the results obtained regarding the relationship between low and high centrality, the comparison with the benchmark, and the weight of stocks within the Markowitz framework are consistent. This important finding shows that the meaning of centralization is different than previously defined in other studies, such as social networks. According to our research findings, we advise capital market participants to construct their portfolios using network analysis, enrichment analysis, and centrality criteria.

In sum, the novel population analysis helps to:

- Better visualize the big data associated with financial datasets.
- Identify the hidden patterns that cannot be uncovered with the existing methods

- Identify the companies' behavior among the groups of companies that perform the same for a long timeframe.
- Study the behavior of isolated companies that act differently than others.
- Identify significantly enriched input parameters that affect the outcome excess return.
- Provide new insights of the bigdata associated with CRSP dataset.

11.2 Limitation

The data was mainly collected from CRSP and was partially collected from Bloomberg. This led to a few limitations as follows: there were a limited number of variables in the CRSP dataset; and for case studies related to Covid-19, we had to collect data from Bloomberg datasets that contained specific variables. Collecting the data for post-pandemic was difficult in terms of availability and time consumption. However, we made sure to evaluate the proposed method with overlapping datasets from different time periods in past crises. Another limitation with our analysis was that our finding about centrality is new and to the best of our knowledge, there are not many studies about measuring centrality in financial markets. If there were, then the results of our analysis could have been more interpretable.

11.3 Future Studies

According to the results of this study and to better understand the issue of portfolio balancing, it is recommended to address the following issues in future studies:

1- Researchers are encouraged to do more research on centrality.
2- Investigate the nature of centrality network strategy by eliminating the assumption that the high central stocks are the ones that affect the market the most.
3- Try to find new functions for centrality as well as design more useful and efficient strategies.
4- Examine the conditions for adding or removing assets from the portfolio over time based on centrality measurements and adjust the weightage of stocks following with technical analysis.
5- Compare potential portfolios benchmarks other than S&P500.

CHAPTER 12 REFERENCES

[1] E. Bikas, D. Jurevičienė, P. Dubinskas, and L. Novickytė, "Behavioural finance: The emergence and development trends," *Procedia-Soc. Behav. Sci.*, vol. 82, pp. 870–876, 2013.

[2] W. A. Brock, "An integration of stochastic growth theory and the theory of finance, Part I: The growth model," in *General equilibrium, growth, and trade*, Elsevier, 1979, pp. 165–192.

[3] E. E. Carter, "The behavioral theory of the firm and top-level corporate decisions," *Adm. Sci. Q.*, pp. 413–429, 1971.

[4] B. W. Roberts, "The macroeconomic impacts of the 9/11 attack: evidence from real-time forecasting," *Peace Econ. Peace Sci. Public Policy*, vol. 15, no. 2, pp. 341–367, 2009.

[5] T. A. Garrett, "Pandemic economics: The 1918 influenza and its modern-day implications," *Fed. Reserve Bank St Louis Rev.*, vol. 90, no. March/April 2008, 2008.

[6] J. L. Bettman, S. J. Sault, and E. L. Schultz, "Fundamental and technical analysis: substitutes or complements?," *Account. Finance*, vol. 49, no. 1, pp. 21–36, 2009.

[7] A. Gupta and B. Dhingra, "Stock market prediction using hidden markov models," in *2012 Students Conference on Engineering and Systems*, 2012, pp. 1–4.

[8] A. B. Mehrabi, C. A. Ligozio, A. T. Ciolko, and S. T. Wyatt, "Evaluation, rehabilitation planning, and stay-cable replacement design for the hale boggs bridge in Luling, Louisiana," *J. Bridge Eng.*, vol. 15, no. 4, pp. 364–372, 2010.

[9] M. Barthelemy, “Betweenness centrality in large complex networks,” *Eur. Phys. J. B*, vol. 38, no. 2, pp. 163–168, 2004.

[10] K. Okamoto, W. Chen, and X.-Y. Li, “Ranking of closeness centrality for large-scale social networks,” 2008, pp. 186–195.

[11] P. Bonacich, “Some unique properties of eigenvector centrality,” *Soc. Netw.*, vol. 29, no. 4, pp. 555–564, 2007.

[12] Z. Hatami, M. Hall, and N. Thorne, “Identifying Early Opinion Leaders on COVID-19 on Twitter,” in *HCI International 2021 - Late Breaking Papers: Design and User Experience*, Cham, 2021, pp. 280–297.

[13] P. Chetti and H. Ali, “Analyzing the Structural Health of Civil Infrastructures Using Correlation networks and Population Analysis,” in *Proceedings of the Eighth International Conference on Data Analytics*, 2019, pp. 12–19.

[14] K. Cooper, W. Hassan, and H. Ali, “Identification of temporal network changes in short-course gene expression from C. elegans reveals structural volatility,” *Int. J. Comput. Biol. Drug Des.*, vol. 12, no. 2, pp. 171–188, 2019.

[15] K. M. Dempsey and H. H. Ali, “Identifying aging-related genes in mouse hippocampus using gateway nodes,” *BMC Syst. Biol.*, vol. 8, no. 1, p. 62, 2014, doi: 10.1186/1752-0509-8-62.

[16] A. Fuchsberger and H. Ali, “A Correlation Network Model for Structural Health Monitoring and Analyzing Safety Issues in Civil Infrastructures,” 2017.

[17] S. Kim, I. Thapa, and H. H. Ali, “A Graph-Theoretic Approach for Identifying Bacterial Inter-correlations and Functional Pathways in Microbiome Data,” in *2018*

IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, Dec. 2018, pp. 405–411. doi: 10.1109/BIBM.2018.8621179.

[18] E. Rastegari, S. Azizian, and H. Ali, "Machine learning and similarity network approaches to support automatic classification of parkinson's diseases using accelerometer-based gait analysis," 2019.

[19] P. Chetti and H. Ali, "Estimating the Inspection Frequencies of Civil Infrastructures using Correlation Networks and Population Analysis," *Int. J. Adv. Intell. Syst.*, vol. 13, no. 1 & 2, pp. 125–136, 2020.

[20] A. Batushansky, D. Toubiana, and A. Fait, "Correlation-Based Network Generation, Visualization, and Analysis as a Powerful Tool in Biological Studies: A Case Study in Cancer Cell Metabolism," *BioMed Res. Int.*, vol. 2016, pp. 1–9, 2016, doi: 10.1155/2016/8313272.

[21] P. Shannon *et al.*, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, 2003.

[22] A. Gupta and B. Dhingra, "Stock market prediction using hidden markov models," in *2012 Students Conference on Engineering and Systems*, 2012, pp. 1–4.

[23] G. Bonanno, G. Caldarelli, F. Lillo, S. Micciche, N. Vandewalle, and R. N. Mantegna, "Networks of equities in financial markets," *Eur. Phys. J. B*, vol. 38, no. 2, pp. 363–371, 2004.

[24] M. Tumminello, T. Di Matteo, T. Aste, and R. N. Mantegna, "Correlation based networks of equity returns sampled at different time horizons," *Eur. Phys. J. B*, vol. 55, no. 2, pp. 209–217, 2007.

[25] G.-J. Wang, C. Xie, and H. E. Stanley, “Correlation structure and evolution of world stock markets: Evidence from Pearson and partial correlation-based networks,” *Comput. Econ.*, vol. 51, no. 3, pp. 607–635, 2018.

[26] K. T. Chi, J. Liu, and F. C. Lau, “A network perspective of the stock market,” *J. Empir. Finance*, vol. 17, no. 4, pp. 659–667, 2010.

[27] S. Vitali, J. B. Glattfelder, and S. Battiston, “Supporting Information: The Network of Global Corporate Control.” Switzerland, 2011.

[28] D. Y. Kenett, T. Preis, G. Gur-Gershgoren, and E. Ben-Jacob, “Dependency network and node influence: Application to the study of financial markets,” *Int. J. Bifurc. Chaos*, vol. 22, no. 07, p. 1250181, 2012.

[29] J. E. Stiglitz, “The Current Economic Crisis and Lessons for Economic Theory,” *East. Econ. J.*, vol. 35, no. 3, pp. 281–296, Jun. 2009, doi: 10.1057/eej.2009.24.

[30] D. Y. Kenett, T. Preis, G. Gur-Gershgoren, and E. Ben-Jacob, “Dependency network and node influence: Application to the study of financial markets,” *Int. J. Bifurc. Chaos*, vol. 22, no. 07, p. 1250181, 2012.

[31] J. L. Bettman, S. J. Sault, and E. L. Schultz, “Fundamental and technical analysis: substitutes or complements?,” *Account. Finance*, vol. 49, no. 1, pp. 21–36, Mar. 2009, doi: 10.1111/j.1467-629X.2008.00277.x.

[32] K. KHAN, H. ZHAO, H. ZHANG, H. YANG, M. H. SHAH, and A. JAHANGER, “The impact of COVID-19 pandemic on stock markets: An empirical analysis of world major stock indices,” *J. Asian Finance Econ. Bus.*, vol. 7, no. 7, pp. 463–474, 2020.

[33] V. Kukreti, H. K. Pharasi, P. Gupta, and S. Kumar, “A Perspective on Correlation-Based Financial Networks and Entropy Measures,” *Front. Phys.*, vol. 8, p. 323, Aug. 2020, doi: 10.3389/fphy.2020.00323.

[34] Z. Hatami, P. Chetti, H. Ali, and D. Volkman, “A Novel Population Analysis Approach for Analyzing Financial Markets under Crises–2008 Economic crash and Covid-19 pandemic,” 2022.

[35] W. Sun, C. Tian, and G. Yang, “Network analysis of the stock market.” Stanford Univ., Stanford, CA, USA, Tech. Rep, 2015.

[36] J. Crotty, “Structural causes of the global financial crisis: a critical assessment of the ‘new financial architecture,’” *Camb. J. Econ.*, vol. 33, no. 4, pp. 563–580, Jul. 2009, doi: 10.1093/cje/bep023.

[37] J. Mazurek and E. Mielcová, “The Evaluation of Economic Recession Magnitude: Introduction and Application,” *Prague Econ. Pap.*, vol. 22, no. 2, pp. 182–205, Jan. 2013, doi: 10.18267/j.pep.447.

[38] M. B. Aalbers, “The financialization of home and the mortgage market crisis,” *Compet. Change*, vol. 12, no. 2, pp. 148–166, 2008.

[39] Y. Karmakar, “Early warning signal system for economic crisis: a threshold and indicators approach,” *Pac. Bus. Rev. Int.*, vol. 6, no. 8, pp. 60–70, 2014.

[40] Y. Karmakar, “Early warning signal system for economic crisis: a threshold and indicators approach,” *Pac. Bus. Rev. Int.*, vol. 6, no. 8, pp. 60–70, 2014.

[41] G. D. Green, “The economic impact of the stock market boom and crash of 1929,” in *Federal Reserve Bank of Boston, Consumer Spending and Monetary Policy: The Linkages, Monetary Conference*, 1971, pp. 189–220.

[42] J. Cannell, “The financial crisis and its impact on the electric utility industry,” *Edison Electr. Inst.*, 2009.

[43] G. M. Hodgson, “The great crash of 2008 and the reform of economics,” in *The Handbook of Globalisation, Third Edition*, Edward Elgar Publishing, 2019.

[44] F. S. Mishkin and E. N. White, “US stock market crashes and their aftermath: implications for monetary policy,” National bureau of economic research, 2002.

[45] R. C. Altman, “The great crash, 2008,” *Foreign Aff.*, vol. 88, p. 1, 2009.

[46] E. N. White, “The stock market boom and crash of 1929 revisited,” *J. Econ. Perspect.*, vol. 4, no. 2, pp. 67–83, 1990.

[47] T. Hanspal, A. Weber, and J. Wohlfart, “Exposure to the COVID-19 stock market crash and its effect on household expectations,” *Rev. Econ. Stat.*, pp. 1–45, 2020.

[48] S. vanDongen, “A cluster algorithm for graphs,” *Inf. Syst. INS*, no. R 0010, 2000.

[49] G. Peralta and A. Zareei, “A network approach to portfolio selection,” *J. Empir. Finance*, p. 24, 2016.

[50] K. Dimitrios and O. Vasileios, “A Network Analysis of the Greek Stock Market,” *Procedia Econ. Finance*, vol. 33, pp. 340–349, 2015, doi: 10.1016/S2212-5671(15)01718-9.

[51] E. F. Fama, “Efficient capital markets: A review of theory and empirical work,” *J. Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[52] S. Kapoor and J. M. Prosad, “Behavioural finance: A review,” *Procedia Comput. Sci.*, vol. 122, pp. 50–54, 2017.

[53] M. Inuiguchi and T. Tanino, “Portfolio selection under independent possibilistic information,” *Fuzzy Sets Syst.*, vol. 115, no. 1, pp. 83–92, Oct. 2000, doi: 10.1016/S0165-0114(99)00026-3.

[54] G. Curtis, “Modern portfolio theory and behavioral finance,” *J. Wealth Manag.*, vol. 7, no. 2, pp. 16–22, 2004.

[55] F. J. Fabozzi, F. Gupta, and H. M. Markowitz, “The Legacy of Modern Portfolio Theory,” *Institutional Invest.*, p. 16, 2002.

[56] V. Ricciardi and H. K. Simon, “What is behavioral finance?,” *Bus. Educ. Technol. J.*, vol. 2, no. 2, pp. 1–9, 2000.

[57] H. Markowitz, “de Finetti scoops Markowitz,” *J. Invest. Manag.*, vol. 4, no. 3, 2006.

[58] F. A. Sortino and L. N. Price, “Performance measurement in a downside risk framework,” *J. Invest.*, vol. 3, no. 3, pp. 59–64, 1994.

[59] F. A. Sortino, R. Van Der Meer, and A. Plantinga, “The dutch triangle,” *J. Portf. Manag.*, vol. 26, no. 1, pp. 50–57, 1999.

[60] M. E. Mangram, “A simplified perspective of the Markowitz portfolio theory,” *Glob. J. Bus. Res.*, vol. 7, no. 1, pp. 59–70, 2013.

[61] A. Beja, “On systematic and unsystematic components of financial risk,” *J. Finance*, vol. 27, no. 1, pp. 37–45, 1972.

[62] D. Chisholm, S. O’Reilly, and M. Betro, “Equity Sectors: Essential Building Blocks for Portfolio Construction,” *Fidel. Leadersh. Ser. Invest. Insights*, 2013.

[63] G. Su, A. Kuchinsky, J. H. Morris, D. J. States, and F. Meng, “GLay: community structure analysis of biological networks,” *Bioinformatics*, vol. 26, no. 24, pp. 3135–3137, 2010.

[64] J. Miśkiewicz, “Analysis of time series correlation. The choice of distance metrics and network structure,” *Acta Phys Pol A*, vol. 121, p. B-89, 2012.

[65] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’networks,” *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[66] S. S. Group, *Risk management lessons from the global banking crisis of 2008*. Senior Superviors Group, 2009.

[67] M. B. Aalbers, *The Financialization of Housing: A political economy approach*, 1st ed. Abingdon, Oxon ; New York, NY : Routledge, 2016.: Routledge, 2016. doi: 10.4324/9781315668666.

[68] S. vanDongen, “A cluster algorithm for graphs,” *Inf. Syst. INS*, no. R 0010, 2000.

[69] M. L. Hale, I. Thapa, and D. Ghersi, “FunSet: an open-source software and web server for performing and displaying Gene Ontology enrichment analysis,” *BMC Bioinformatics*, vol. 20, no. 1, p. 359, 2019.

[70] S. T. Straetmans, W. F. Verschoor, and C. C. Wolff, “Extreme US stock market fluctuations in the wake of 9/11,” *J. Appl. Econom.*, vol. 23, no. 1, pp. 17–42, 2008.

[71] S. R. Baker, N. Bloom, S. J. Davis, K. Kost, M. Sammon, and T. Viratyosin, “The unprecedented stock market reaction to COVID-19,” *Rev. Asset Pricing Stud.*, vol. 10, no. 4, pp. 742–758, 2020.

[72] S. Brohee and J. Van Helden, “Evaluation of clustering algorithms for protein-protein interaction networks,” *BMC Bioinformatics*, vol. 7, no. 1, p. 488, 2006.

[73] V. Boginski, S. Butenko, and P. M. Pardalos, “Statistical analysis of financial networks,” *Comput. Stat. Data Anal.*, vol. 48, no. 2, pp. 431–443, Feb. 2005, doi: 10.1016/j.csda.2004.02.004.

[74] Z. Hatami, M. Hall, and N. Thorne, “Identifying Early Opinion Leaders on COVID-19 on Twitter,” in *HCI International 2021 - Late Breaking Papers: Design and User Experience*, Cham, 2021, pp. 280–297.

[75] M. E. J. Newman, “A measure of betweenness centrality based on random walks,” *Soc. Netw.*, vol. 27, no. 1, pp. 39–54, Jan. 2005, doi: 10.1016/j.socnet.2004.11.009.

[76] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, “Generalized louvain method for community detection in large networks,” 2011, pp. 88–93.

[77] V. Zakamouline and S. Koekebakker, “Portfolio performance evaluation with generalized Sharpe ratios: Beyond the mean and variance,” *J. Bank. Finance*, vol. 33, no. 7, pp. 1242–1254, 2009.

[78] S. Fallahpour and A. Ghahramani, “An Analysis of Centrality’s Features as a New Measure for Network Analysis, Risk Measurement & Portfolio Selection,” *Financ. Res. J.*, vol. 23, no. 2, pp. 158–171, 2021.

[79] C. Demone, O. Di Matteo, and B. Collignon, “Classical decomposition of Markowitz portfolio selection,” Bank of Canada Staff Working Paper, 2020.

[80] F. Pozzi, T. Di Matteo, and T. Aste, “Spread of risk across financial markets: better to invest in the peripheries,” *Sci. Rep.*, vol. 3, no. 1, pp. 1–7, 2013.

[81] C. Katsouris, “Optimal Portfolio Choice and Stock Centrality for Tail Risk Events,” *ArXiv Prepr. ArXiv211212031*, 2021.