Network analysis of famous investors' portfolio - finding associations between investors

Juan Wang

Master of Science
School of Informatics
University of Edinburgh
2023

Abstract

Predicting stock market movements has always been a hot topic in the financial field. Traditional stock price prediction methods and modern machine-learning techniques have made much progress. However, the high complexity, nonlinearity, and many external factors of the stock market make stock price prediction a challenging problem. This project bypasses the complex analysis of technical indicators and various influencing factors but analyses the associations between famous investors through network analysis. This project aims to determine whether any early signs of financial shocks can be spotted based on network signals and whether we can build an effective portfolio based on the portfolios of famous investors. This project applies network science to build a famous investors' network and conducts standard network analysis to study the correlation between the portfolios of famous investors.

This project extracts famous investors' portfolio data for 64 quarters from the website www.dataroma.com to build an investor network in which nodes (vertexes) are investors, links (edges or connections) between two nodes mean they invest in the same stock, and ABC (associations beyond chance) value is set as the weight (strength) of the link. ABC represents the likelihood of two investors holding the same stocks beyond what is expected by chance. This project analyses the evolution of nodes, links, average weight, and weighted clustering coefficient of the networks. NPSBM (non-parametric stochastic block models), a family of Bayesian algorithms is used to do community detection analysis.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics Committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Juan Wang*)

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Valerio Restocchi for the guidance and support throughout the project. I would also like to thank Ognyan Simeonov for his help on the SBM model and thank Dr. Guillermo Romero Moreno for his help on the ABC model. They provided thoughtful feedback on my question.

I would like to thank Linzhi Xu and Xiangyu Tian, my classmates in the same project. Their humorous words eased my anxiety.

Finally, I would like to thank my classmates Zhenyu Zhang, Jiao Liu and Wanghao Yu. They helped me a lot with the project. Every discussion with them can bring me a lot of harvest and help me better understand the content of relevant literature, which gives me more confidence to complete the project.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Accurate stock market prediction can help investors make proper decisions and make significant profits. Stock market prediction is, therefore, always an attractive field for investors, industrialists, and academics. Both traditional financial analysis and more recent data-driven techniques have made a lot of progress, such as Technical Analysis, Fundamental Analysis, Quantitative Analysis, Machine Learning and Artificial Intelligence, Sentiment Analysis, Time Series Analysis, Macroeconomic Indicators, News-based Analysis and Event-driven Strategies. They mainly focus on analysing historical trading data or external factors affecting stock prices. They are roughly divided into two categories: prediction-based techniques and clustering-based techniques[4]. However, due to the high complexity, nonlinearity, and many factors affecting stock prices, including traders' expectations, financial conditions, administrative events, and market trends, especially some irrational factors, stock market prediction is a very complex and challenging task.

This project bypasses the complex analysis of technical indicators and various influencing factors and studies this task from network science. It aims to determine whether network signals can identify early signs of financial shocks and whether famous investors' portfolios can be adapted to build a successful portfolio. Network science was invented before the term data science was introduced, starting with mathematics. Nowadays, it is an important interdisciplinary academic field, and it is applied to various fields, including social media, banking, finance, international trade, fraud detection, counterterrorism, biology, and so on. Network Analysis focuses on studying patterns of connection in a wide range of physical and social phenomena[5]. The

focal point is the interaction and connection between the elements in the network. Network analysis can help us find the associations between investors, namely the factors that influence them to make the same decision.

This is a new perspective. Famous investors have already considered many influencing factors and applied a series of stock market prediction methods when making investment decisions, which, to some extent, has guaranteed a relatively high success probability. Applying network analysis to find the internal connection of their decisions and get valuable investment portfolios is equivalent to analysing only a small amount of data but considering all the factors that influence the stock market. It provides valuable insights into the investment strategies of famous investors and their potential impact on the stock market.

1.2 Objectives and Hypotheses

This project aims to determine whether network signals can identify early signs of financial shocks and whether famous investors' portfolios can be adapted to build a successful portfolio by analysing the associations between famous investors in the network. The associations beyond chance (ABC) model is applied to calculate the likelihood that two investors will invest in the same stock, and non-parametric SBM is used to detect non-trivial communities of investors.

This project thinks that famous investors in the market are rational and make value investments based on professional analysis. Their opinion on a certain stock will impact the stock price. However, the stock market itself is a complex network, and there are also many irrational investors in it. They trade based on market news or price fluctuations. The influence of these irrational factors cannot be eliminated. Therefore, the stock market is notoriously difficult to predict, and past performance does not always indicate future results. It is essential to approach this project with a healthy dose of scepticism and caution.

1.3 Dissertation Structure

This paper includes five chapters:

1. **Introduction**: This section introduced the motivation, objectives, and hypothesis of the project.

2. **Background and Related Work**: This section presented the research and technical background of the project. The purpose of this section is to provide readers with a better understanding of the project and its work.

3. **Design and Implementation**: This section explained the methodology and steps for implementing the project.

4. **Results and Evaluation**: This section analysed and verified the running results and discussed the meaning behind the results

5. **Conclusion and Future Works**: This section summarized the final project analysis results and what further work needs to be done in the future

Chapter 2

Background and Related Work

2.1 Network Analysis

2.1.1 Overview

Network analysis is also known as network theory or network science, a field of study involving complex network analysis. It is a method of studying the structure of a network, the relationship between nodes, and the evolution of the network. It encompasses a range of techniques and methods used to study networks' structure, dynamics, and behaviour, which can be social, biological, technological, or physical. It focuses on the interactions between elements in the network[1]. A network is typically represented as a graph consisting of nodes (vertices) and links (edges or connections) which connect nodes. By analysing the graph's topology, such as the degree distribution, clustering coefficient, and centrality measures, network analysts can gain insights into the underlying structure of the network and its properties, such as robustness, resilience, and vulnerability[2]. Network analysis has applications in various fields, including sociology, computer science, biology, economics, and transportation. It is used to model social networks, analyse the spread of infectious diseases, study the structure of the internet, and understand the behaviour of financial markets, among other things[1].

2.1.2 Key Concepts and Techniques

Key concepts and techniques for network analysis include:

1. **Nodes and links**: The basic building blocks of a network. For example, in this project, nodes are investors, and links between investors mean they invest in the

same stock[8].

2. **Degree**: indicates the number of connections of a node. In a directed network, degrees can be divided into incoming and outgoing degrees. Degree analysis provides us with an initial understanding of the activity or importance of each node in the network[8].

3. **Clustering Coefficient**: The probability that an edge exists among neighbours of a node. The clustering Coefficient measures the local clustering characteristics or tightness of nodes in a network. Specifically, it measures how connected or tight a node's neighbours are to each other[8].

4. **Centrality**: This is a set of metrics that measure the importance of a node in a network. Common centrality indexes include degree centrality, closeness centrality and betweenness centrality. Degree centrality measures a node's direct connectivity. Closeness centrality measures a node's average "distance" from other nodes in the network. Betweenness centrality measures a node's role as a "bridge" or "intermediary" in the network[8].

5. **Community Structure**: In a network, certain nodes may be more closely linked together to form a community. Community detection methods attempt to identify these tightly connected groups of nodes. In some networks, such as biological networks or social networks, structure may be closely related to actual function or behaviour. For example, in a protein interaction network, a community may represent a functionally related protein group; In a social network, a community may represent a group of people with common interests or backgrounds[8].

6. **Scale-free Property**: Some nodes in the network (called centres or hubs) have many connections, while most nodes have only a few connections. The degree distribution follows a power law. Identifying hubs can help a researcher or decision-maker identify a network's most important or influential elements to develop strategies or conduct further research[8].

7. **Network dynamics**: The study of the evolution of networks over time or the processes that occur on them, such as the diffusion of information or the spread of disease[8].

2.1.3 Community Detection

Community detection is a process of identifying densely connected groups or communities of nodes in a network. A community is a group of nodes that are more closely connected to one another than to other nodes. By detecting communities in a network, we can gain valuable insight into the function and organisation of the system being studied because the network's modular structure can be understood[7, 14]. Community detection, or graph clustering, partitions graph vertices into more densely connected clusters. From a more general point of view, community structures may also refer to groups of vertices that connect similarly to the rest of the graphs without necessarily having a higher inner density, such as disassortative communities with higher external connectivity [9]. Community detection may also be performed on graphs where edges have labels or intensities. If these labels represent similarities among data points, the problem may be called data clustering or communities[3].

Based on purpose and application, community detection can be broadly divided into two categories:

Descriptive Methods: The core goal of descriptive community detection methods is to identify and describe patterns of community structure in the network rather than explain how these patterns are formed. These methods mainly rely on finding and identifying patterns in the network without considering the generation mechanism behind these patterns[12].

Inferential Methods: The primary purpose is to understand the network structure by inferring the underlying generative mechanisms. These methods are often based on statistical models, such as random graph models or Bayesian methods. An inferential approach would try to understand which structures will most likely generate the observed network data. The results are often probabilistic inferences about possible community structures or generative mechanisms. This means that the results not only give a community partition but also give an estimate of the uncertainty of this partition[12].

In short, descriptive approaches focus primarily on finding significant community structures in the network, while inferential methods seek to understand the underlying mechanisms that generate these structures. In this project, our focus is the associations between famous investors, which is the mechanism that produces communities. Therefore, an inferential method – non-parametric stochastic block model will be used for community detection, a Bayesian algorithm family[12].

2.1.4 Non-parametric SBM

The **stochastic block model (SBM)** is the simplest generative process based on the notion of groups of nodes. The basic (or traditional) version takes the microcanonical formulation as parameters, based on a partition of N nodes into B groups, given by the vector b with entries $bi \in \{1,...B\}$ specifying the group membership of node i and $B \times B$ matrix of edge count e, where e_{rs} is the number of edges between groups r and s. Based on these constraints, the edges are then randomly placed. As a result, nodes in the same group have the same probability of being connected with other network nodes and tend to have very similar degrees. This is a poor model for most empirical networks, which possess highly heterogeneous degree distributions. To solve this problem, an enhanced model **degree-corrected SBM** emerged, and it added the degree sequence $k = \{k_i\}$ of the graph as an additional parameter.

The above model requires the number of groups B to be known in advance. Instead, we wish to formulate a nonparametric framework where the number of groups, as well as any other model parameter, is determined from the data itself. That is **Non-parametric SBM** which inferred the parameters of the model from the data by **non-parametric statistical inference**: we define a model that generates a network A with a probability $P(A|\theta,b)$, where θ are additional model parameters(such as e,k in SBM) that control how the node partition affects the structure of the network. Therefore, if we observe a network A, the likelihood that it was generated by a given partition b is obtained via the Bayesian posterior probability:

$$P(b|A) = \frac{\sum_\theta P(A|\theta,b)P(\theta,b)}{P(A)} \quad (1)$$

where $P(\theta,b)$ is the prior probability of the model parameters, and

$$P(A) = \sum_{\theta,b} P(A|\theta,b)P(\theta,b) \quad (2)$$

is called the evidence and corresponds to the total probability of the data summed over all model parameters. The particular types of model that will be considered here have “hard constraints”, such that there is only one choice for the remaining parameters that is compatible with the generated network, which means Eq. (1) simplifies to

$$P(b|A) = \frac{P(A|\theta,b)P(\theta,b)}{P(A)} \quad (3)$$

with θ above being the only choice compatible with A and b. The inference procedures considered here will consist in either finding a network partition that maximizes Eq. (3), or sampling different partitions according to its posterior probability [9, 6, 10, 11].

2.2 Associations beyond Chance

Based on[13], this project introduces the concept of ABC (associations beyond chance) to indicate the possibility of two investors investing in the same stock and will set ABC as the weight of the edges in the network of investors. ABC is a metric used to determine the strength of association in a 2x2 contingency table. It provides a measure to understand if the observed associations (or co-occurrences) in the table significantly differ from what one would expect by chance alone.

Here is the calculation formula:

$ABC = [n11 - (n10 + n11)(n01 + n11)](1 - n10 - n11)(1 - n01 - n11)/n01n10n00$

where n11, n10, n01, n00 are the cells of the contingency table.

- n11 refers to the number of occurrences where both events A and B occur.
- n10 refers to the number of occurrences where event A occurs but B doesn't.
- n01 refers to the number of occurrences where event B occurs but A doesn't.
- n00 refers to the number of occurrences where neither event A nor B occurs.

or analogously with

$ABC = (P12 - P1 * P2) * (1 - P1) * (1 - P2)/(P1 - P12)/(P2 - P12)/(1 - P1 - P2 + P12)$

where P12, P1, P2 are the joint and marginal probabilities.

- P12 is the joint probability of A and B occurring together.
- P1 is the probability of A occurring (irrespective of B).
- P2 is the probability of B occurring (irrespective of A).

The ABC ranges from -1 to potential infinity, where -1 implies mutual exclusion (perfect negative correlation), infinity indicates perfect correlation, and zero indicates independence (no correlation).

This project will use the second formula to calculate ABC. And,

- N = Total number of stocks in a dataset for building a network
- P1 = Number of stocks invested by Investor 1 / N
- P2 = Number of stocks invested by Investor 2 / N
- P12 = Number of stocks invested by Investor 1 and Investor 2 / N

This project will focus on combinations/communities that have strong associations

Chapter 3

Design and Implementation

3.1 Environment

1. **Data Resource**: This project extracts famous investors' portfolio data from https://www.dataroma.com/m/home.php, which is a website that tracks top hedge fund managers in the United States and changes in their portfolios. Dataroma focuses on investments in U.S.-listed stocks by these top investors. We can view these top fund managers' latest and past stock holdings and position movements on Dataroma.

2. **Python Environment**: This project uses Python 3.10.12 for programming and uses Colab (Google Colaboratory) as the IDE platform, which is a cloud service provided by Google that allows users to write and execute Python code in a browser, especially code for machine learning and data analysis. The data comes from a public website, so there is no data security issue for working on the cloud platform.

3. **Library for Network Analysis**: This project uses Graph-tool for network analysis and other libraries for calculation, exporting and analysis. Graph-tool is a powerful Python module for handling and counting complex networks. It is written in C++ and provides its functionality through the Python interface, which means it is performance-efficient while maintaining Python's ease of use.

4. **Install Graph-tool**: Installing Graph-tool on Colab is simple and quick, but because it's a cloud platform, it can't keep the environment configuration. I need to reinstall graph-toll every time I reconnect to the cloud platform. Here is the install command:

```
!echo "deb http://downloads.skewed.de/apt jammy main" >> /etc/apt/sources.list
!apt-key adv --keyserver keyserver.ubuntu.com --recv-key 612DEFB798507F25
!apt-get update
!apt-get install python3-graph-tool python3-matplotlib python3-cairo
```

```
!apt purge python3-cairo
!apt install libcairo2-dev pkg-config python3-dev
!pip install --force-reinstall pycairo
!pip install zstandard
```

3.2 Design

1. **Build Network**: This project will build a single-layer, undirected, weighted investors network based on the website's famous investors' portfolio data. Investors are network nodes; edges will be added between two investors if they invested in the same stock, and the ABC value of two investors is the weight of an edge. The calculation formula of ABC refers to section 2.2.

2. **Evolution Analysis**: This project will analyse the evolution of nodes, links, average weight and average weighted cluster coefficient, as well as the evolution of communities. These evolution analyses try to find relationships between changes in investor networks and market trends.

3. **Centrality Analysis**: This project attempts to find critical investors in the network through centrality analysis, mainly focusing on betweenness centrality and closeness centrality. Investors with high betweenness centrality are likely to act as a "bridge" between multiple investment groups or strategies, which means that these investors may have diversified investment strategies and connections to various investment groups. Their investment choices may reflect market trends and information across multiple investment groups. Investors with high closeness centrality may have some connection or similarity to most other investors, which means that these investors may have a good understanding of the overall market trend. The strategies and choices of these investors may represent the prevailing views and strategies of the market. Degree centrality will not be analysed because investors with a high degree can be institutions with high market influence or followers.

4. **Community Analysis**: This project will find investor communities with similar investment strategies through community detection, analyse the industries' distribution in the community, and judge whether an effective investment portfolio can

be obtained based on the common selection of investors in the community. On the other hand, this project will analyse the changes in investor communities before and after the financial shock to evaluate whether it can predict the occurrence of an economic shock through the changes in the investor communities.

This project will use Non-parametric SBM to detect communities in the network. In this project, ABC value indicates the likelihood of two investors holding the same stocks beyond what is expected by chance. When applying non-parametric SBM, weights (ABC value) will be treated as covariates.

3.3 Implementation

3.3.1 Data Extraction

This project extracted two kinds of data from the website through a crawler program.

1. **Investors' portfolio data** The website has 77 investors' portfolio data for the 67 quarters from the fourth quarter of 2006 to the second quarter of 2023. 77 is the total number of investors, but only some quarters's data includes all investors. They were crawled to a Portfolios.xlsx file and then split into 67 Excel files for 67 quarters (file name is the quarter of the data, such as 2007 Q1.xlsx) under the folder MScX.

   Dataset of Networks: The final task is to generate quarterly data for creating the network. Based on these 67 Excel files, 67 CSV files were generated (in the folder "MScC") with the same name as the Excel files to create the network. The format is shown in Figure 3.1.

Investor1	Num1	Investor2	Num2	ABC	number of stocks	stocks			
Bill Nygren	7	Charles Bobrir	11	2.938119	4	IMS	HRB	JPM	DELL-OLD
Bill Nygren	7	David Katz	22	0.554187	2	GPS	TWX		
Bill Nygren	7	Dodge & Cox	26	0.554187	2	MCD	TWX		

Figure 3.1: CSV file for building Network

Each row in Figure 3.1 means an edge in the network.

Columns ***Investor1*** and ***Investor2*** are the names of the two investors that are connected.

Columns ***Num1*** and ***Num2*** represent the two investors' index number, which is used as the node's name to solve the problem that the investor's name is too long to display in the network graph. Except for being used as the node's name, the index

number is also used to control the size of a node. The higher the portfolio's value of an investor, the larger the investor's index number, and then the larger the size of its node in the network. It can help us to analyse the distribution of super investors in each community. To do this, we extracted a list of investors and numbered them according to the value of their portfolio in ascending order (refer to Figure 3.2).

Column ***number of same stocks*** is the number of the same stocks that the two investors invest in.

Carl Icahn - Icahn Capital Management	70	Carl Icahn	$21.8 B	21,800,000,000
Terry Smith - Fundsmith	71	Terry Smith	$21.9 B	21,900,000,000
Chris Hohn - TCI Fund Management	72	Chris Hohn	$29.9 B	29,900,000,000
Polen Capital Management	73	Polen	$34.2 B	34,200,000,000
Bill & Melinda Gates Foundation Trust	74	Bill & Melinda Gates	$35.7 B	35,700,000,000
First Eagle Investment Management	75	First Eagle	$36.2 B	36,200,000,000
Dodge & Cox	76	Dodge & Cox	$87 B	87,000,000,000
Warren Buffett - Berkshire Hathaway	77	Warren Buffett	$299 B	299,000,000,000

Figure 3.2: Index of Investors

Summary table: A summary table was generated while processing quarterly data, which would help us understand the overall situation of the quarterly data and help us decide how to do data screening (refer to Figure 3.3).

Quarter	Number of Investors	Number of Stocks	Max ABC	Min ABC	Number of Single Investors	Name of single Investors	Comments
2006 Q4	20	266	9.27381	-0.37201	1	['FPA']	
2007 Q1	20	290	6.713335	-0.30517	0		
2007 Q2	23	289	9.392876	-0.30795	1	['FPA']	
2007 Q3	23	292	9.516348	-0.29961	1	['FPA']	

Figure 3.3: Summary Table

2. **Stocks Industry Information**

The table of stock industry information was exported to analyse the industry distribution of the stocks in a community. There are 1500 stocks that were exported to the Stocks.xlsx file (please refer to Figure 3.4). Column ***Sector*** represents the industry to which the stock belongs.

Abbr	**Name**	**Sector**
MSFT	Microsoft Corp. (MSFT)	Information Technology
BRK.A	Berkshire Hathaway CL A (BRK.A)	Financials
GOOG	Alphabet Inc. CL C (GOOG)	Technology

Figure 3.4: Industry Information of Stocks

3.3.2 Data Screening and Processing

To reduce the complexity of the networks and remove noise, I conducted the following data screening:

1. **Data Range**: This project only focused on the full 16 years (64 quarters) of data from 2007 to 2022 (did not consider the data of 2006 Q4 and 2023 Q1 & Q2).

2. **Remove Single Investors**: It can be seen that there are 13 quarters of data (refer to Figure 3.5) that include a single investor whose degree is 0 in the network based on the summary table. A node with zero degree has no relationship with other nodes, which is useless for analysing the associations between investors. So, they were removed from the network.

Quarter	Number of Investors	Number of Stocks	Max ABC	Min ABC	Number of Single Investors	Name of single Investors
2006 Q4	20	266	9.27381	-0.37201	1	['FPA']
2007 Q2	23	289	9.392876	-0.30795	1	['FPA']
2007 Q3	23	292	9.516348	-0.29961	1	['FPA']
2008 Q1	26	329	7.900375	-0.19682	1	['FPA']
2008 Q3	27	342	5.749579	-0.16075	1	['FPA']
2014 Q1	43	470	24.50535	0	1	['Bill Ackman']
2014 Q3	43	476	76.34869	0	1	['FPA']
2014 Q4	44	488	156.3722	0	1	['FPA']
2015 Q1	45	483	200.0417	0	1	['FPA']
2015 Q2	45	471	135.2394	0	1	['FPA']
2019 Q2	73	603	734.1202	0	1	['Glenn Welling']
2022 Q1	77	638	999	0	1	['Glenn Welling']
2022 Q2	77	647	999	0	1	['Michael Burry']
2007 Q1	20	290	6.713335	-0.30517	0	

Figure 3.5: Summary - sorted by Number of Single Investors

3. **Set negative Weight to 0**: The weight of an edge is ABC value in this project. As the background section explains, ABC indicates the possibility of two investors investing in the same stock. A negative ABC value means a negative correlation, which is not what this project is concerned about. This project focuses only on positive correlation, so all negative ABC values were set to 0.

3.3.3 Evolution Analysis

This project generated 64-quarter evolution graphs of 6 standard indicators of network analysis, including **nodes**, **links**, **number of stocks**, **average degree**, **average weight**

and **average weighted cluster coefficient**, which would help us analyse the impact of financial shocks on the investor network from a macro perspective.

3.3.4 Centrality Analysis

This project calculated weighted centralities of betweenness and closeness, got their top ten lists of centralities and then compared the two lists to get the common investors in the two lists through the VeNN (Vector-Enhanced Nearest Neighbor) graph. These common investors are the key investors in the network. They have diversified investment strategies and a good understanding of overall market trends.

3.3.5 Community Analysis

1. **Function Selection**

   This project uses Graph-tool to conduct community detection. There are two functions **minimize_blockmodel_dl()** and **minimize_nested_blockmodel_dl()** in Graph-tool that can be used to detect communities. The nested function has advantages in detecting large networks and can show hierarchical structure, but it needs more running time. The networks built in this project are all single-layer, undirected, and weighted networks, and the number of nodes in the networks is small (less than 100). Therefore, the first non-nested function is enough. However, compared with the first function, the image detected by the nested function is more intuitive (refer to Figure 3.6; the left is detected by the nested function and the right is detected by the non-nested function). Due to the small amount of data, I do not need to consider performance and running time, so this project adopts the nested function **minimize_nested_blockmodel_dl()**.

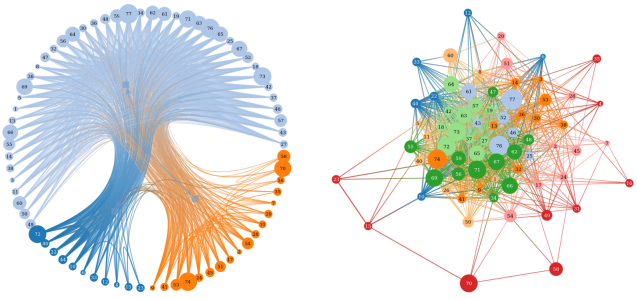

Figure 3.6: Graph of Communities for 2019 Q2

2. **Model Selection**

As the background section explains, most empirical networks are better fitted by degree-corrected models, but there are still some exceptions [graph-tool]. In this project, we compared the data length of the degree-corrected model and the normal model for Q4 from 2007 to 2022 and found that the degree-corrected model fitted the networks better most of the time. We ran the comparison program about five times and found that the degree-corrected model was better than normal each time, especially in the networks with a relatively large number of nodes (networks after 2016). So, a degree-corrected model will be used in this project. Figure 3.7 shows the result of one comparison.

```
Improvement of 2007 Q4 : 4.445736816253543
Improvement of 2008 Q4 : -12.220102449425553
Improvement of 2009 Q4 : 14.256330091120049
Improvement of 2010 Q4 : 18.11494401178504
Improvement of 2011 Q4 : 46.43294859801574
Improvement of 2012 Q4 : 9.912514482196457
Improvement of 2013 Q4 : -0.6092870977256553
Improvement of 2014 Q4 : 7.925484101862821
Improvement of 2015 Q4 : 7.881588047420337
Improvement of 2016 Q4 : -4.8402693410955635
Improvement of 2017 Q4 : -10.619764228276836
Improvement of 2018 Q4 : -15.375480085728668
Improvement of 2019 Q4 : -124.2001605478863
Improvement of 2020 Q4 : -14.744951621645669
Improvement of 2021 Q4 : -137.43694288651204
Improvement of 2022 Q4 : -86.66709814679325
Degree-corrected is more better in 9 times of total 16 times.
```

Figure 3.7: Comparison between degree-corrected and normal model

3. **Refinements**

As described in Graph-tool documentation, we can perform extra MCMC sweeps to improve the detection result. However, we tested this operation and found that this operation took a long time. Considering that the functions **minimize_blockmodel_dl()** and **minimize_nested_blockmodel_dl()** have employed an agglomerative multilevel Markov chain Monte Carlo (MCMC) algorithm [2014] and the network of investors is a small network with less than 100 nodes, we did not conduct this optimisation algorithm in this project.

Chapter 4

Results and Discussion

4.1 Evolution Analysis

4.1.1 Nodes, Links, and Average Degree

Note: The number of stocks data is got from the summary table (Figure 3.3).

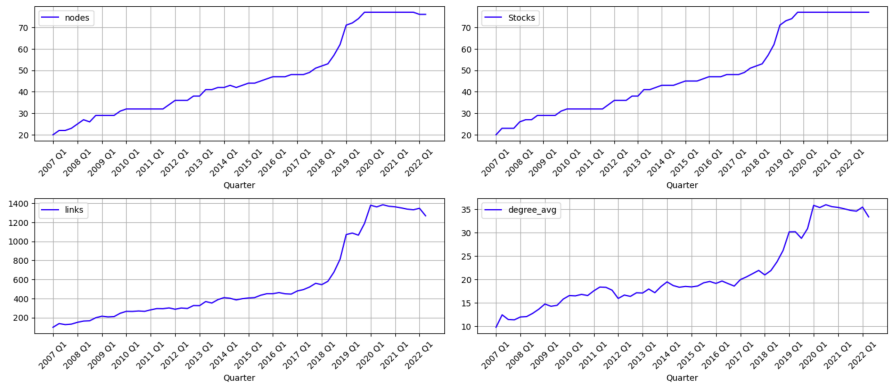

Figure 4.1: Nodes, Links, Density and Average Degree

From Figure 4.1, it can be seen that before the outbreak of COVID-19 in 2020, the economy was in a rising stage, the number of investors in the network continued to increase, and the co-investment behaviour among them was also increasing. After the outbreak of COVID-19, the economy stagnated, no new investors joined, and the investment behaviour (links) did not change significantly. The number of nodes (or investors) and the number of stocks keep a similar growth curve, indicating that each investor's average number of stocks is stable.

4.1.2 Average Weight

Figure 4.2: Average Weight

From Figure 4.2, the average weight increased significantly in 2020 when COVID-19 appeared, which means that investors are more likely to co-invest in a particular stock. The pandemic has led to dramatic changes in global economic and market conditions, and specific industries and companies (e.g., medical, technology, telecommuting-related companies) may be getting a lot of attention. Investors may generally view these sectors or companies as having better growth prospects, leading to large inflows, which explains the rise in weighted averages - that is, more investors are choosing the same or similar investments.

4.1.3 Weighted Cluster Coefficient

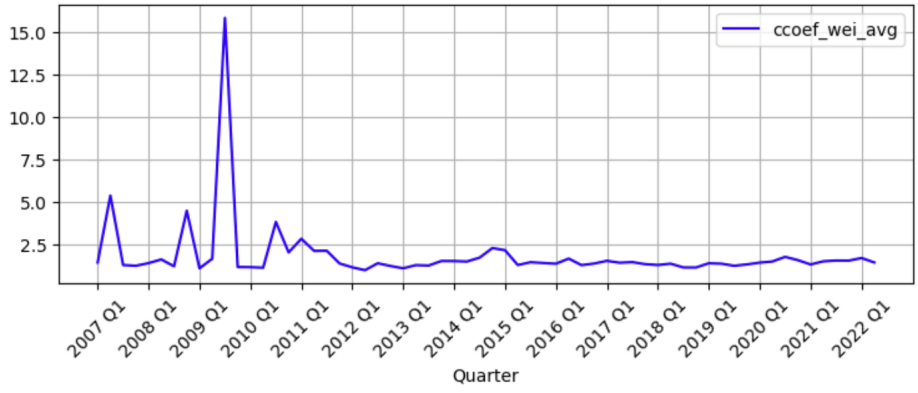

Figure 4.3: Weighted Cluster Coefficient

From Figure 4.3, the weighted clustering coefficient increased significantly in 2009, which means that there were more similar investment strategies or investment be-

haviours among these investors. This is just after the 2008 financial crisis, which shows that investors' investment choices converge when financial shocks occur.

Analysis of average weight and weighted clustering coefficient is very interesting. The financial crisis in 2008 led to an increase in the weighted clustering coefficient, while the pandemic led to an increase in the average weight. Both the financial crisis and the COVID-19 pandemic are big shocks to the global economy, but their impact on investor behaviour is likely to be unique. After the financial crisis, investors may be more inclined to pursue safety, resulting in homogenisation of behaviour. While the pandemic may have led to a general interest in certain sectors or companies with growth prospects, there may be greater differences in the choice of investment strategies.

4.1.4 Communities

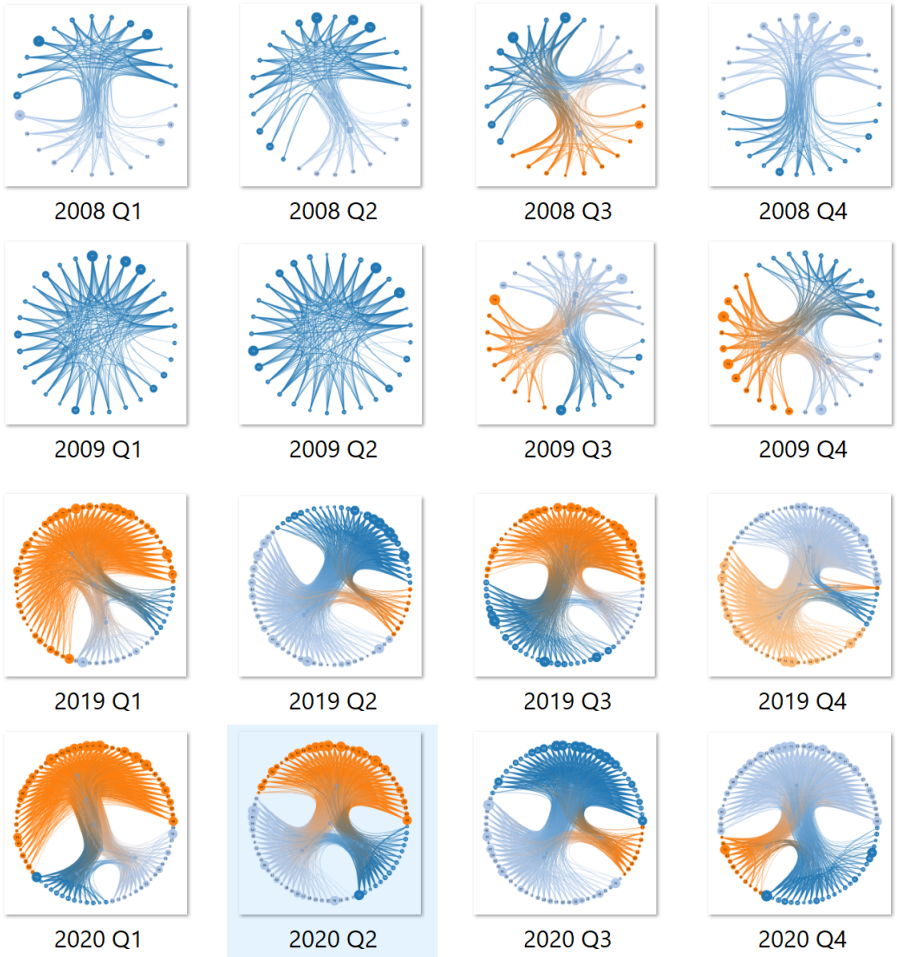

Figure 4.4: Evolution of Communities

I ran a total of 64 quarterly community structure graphs. Here are the comparison charts before and after the 2008 financial crisis and the COVID-19 epidemic in early 2020. It can be seen from Figure 4.4 that the community disappeared in the first and second quarters of 2009 after the 2008 financial crisis, which represents the convergence of investment strategies of all investors. This result is consistent with the previous analysis of the Weighted Cluster Coefficient in section 4.1.3. However, in contrast, the network structure did not change significantly before and after 2019, which also shows that COVID-19 did not lead to the convergence of investment strategies of all investors.

4.2 Centrality Analysis

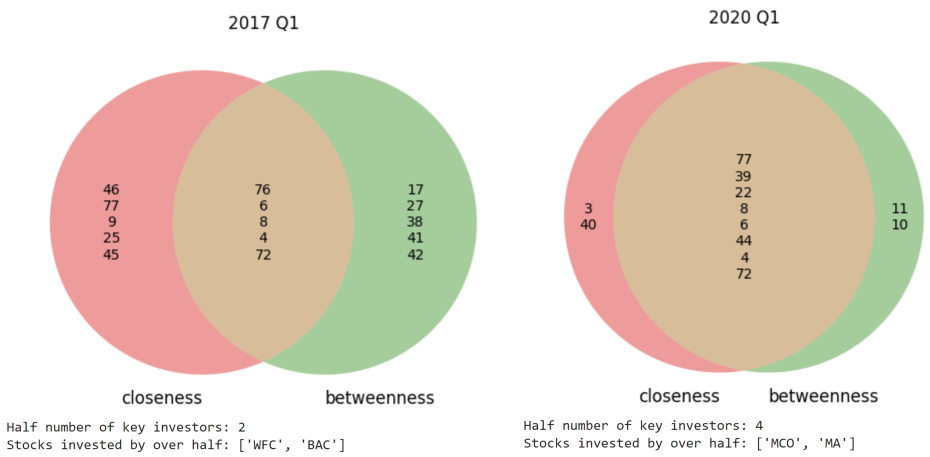

Figure 4.5: Intersection of Top ten closeness and betweenness

Figure 4.5 shows the intersection of the top ten nodes of closeness centrality and the top ten betweenness centrality nodes through the VeNN (Vector-Enhanced Nearest Neighbor) graph. The numbers on the graph represent the unique number of investors. The nodes in the intersection are the network's critical investors, who have diversified investment strategies and a good understanding of overall market trends. It also shows the list of stocks invested by over half of the nodes in the intersection.

I selected 2017 Q1 (before COVID-19) and 2020 Q1 (after COVID-19) data to evaluate the return of these stocks in 1 year. I checked the trend (from the website DATAROMA) of the four stocks (WFC, BAC, MCO, MA), which over half critical investors invested in the network. From Figure 4.6, we can see that these stocks all went up in one year. This result shows that this method is feasible to some extent

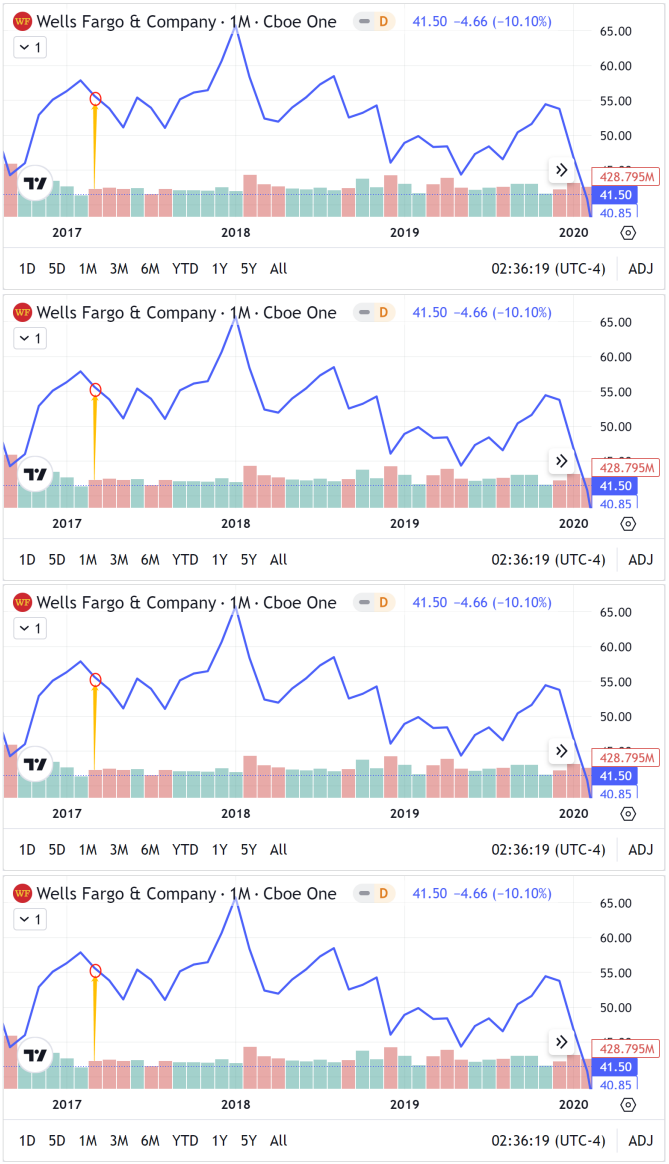

Figure 4.6: Trend of Stocks

despite the limited amount of data analysed.

4.3 Community Analysis

I selected two time periods before and after the epidemic (2019 Q3 and 2021 Q2) for analysis (refer to Figure 4.7). From this Figure, it can be seen:

1. The top 10 investors by portfolio value (index number from 68 to 77) are scattered across different communities, which may mean that the top hedge funds in the market employ various investment strategies. Different communities may represent

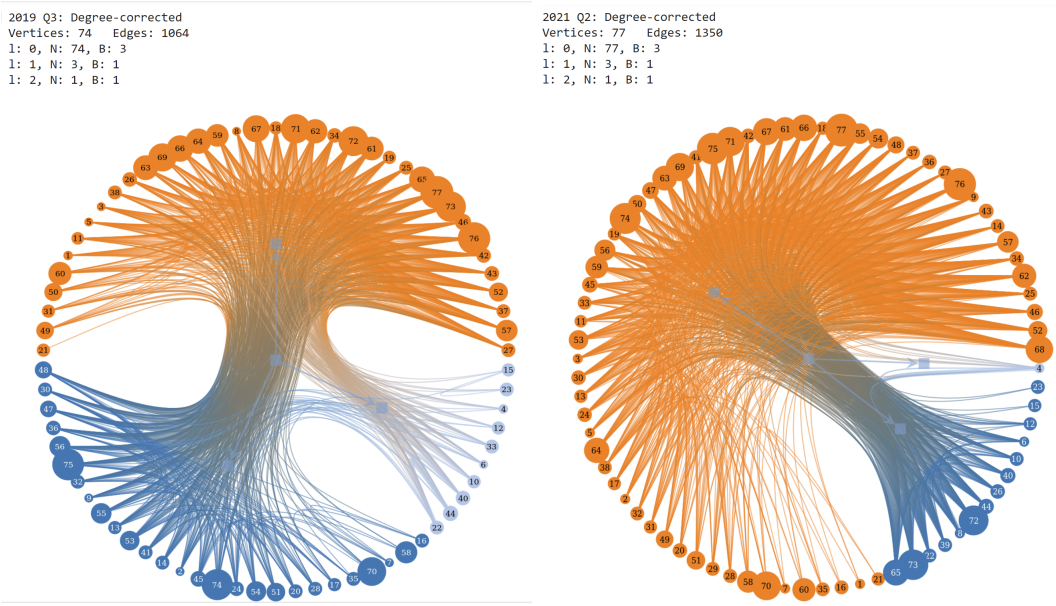

Figure 4.7: Communities of 2019 Q3 and 2021 Q2

different investment styles, strategies, or market positioning, suggesting that there is no ”single, best” investment strategy, but rather multiple strategies co-exist and can all be successful.

2. When multiple large investors are dispersed among different communities, systemic risk in the market may be reduced. If the majority of large investors are concentrated in the same community (with similar investment strategies), then any market change that affects that strategy can cause a larger market shock.

3. Since the top investors are scattered among different communities, we can pick representative stocks or investment strategies from each community to build a diversified portfolio. This way, the portfolio can draw on each community's characteristics and strengths to improve overall and risk-adjusted returns.

I compiled information about the sectors in which each community invested. How many investors have invested in each industry is counted and shown in order from largest to smallest in a bar chart, and the comparison of the top 3 industry preferences in different communities also was shown through the VeNN (Vector-Enhanced Nearest Neighbor) graph (refer to Figure 4.8 and Figure 4.9).

From the VeNN graph in the two Figures, it can be seen:

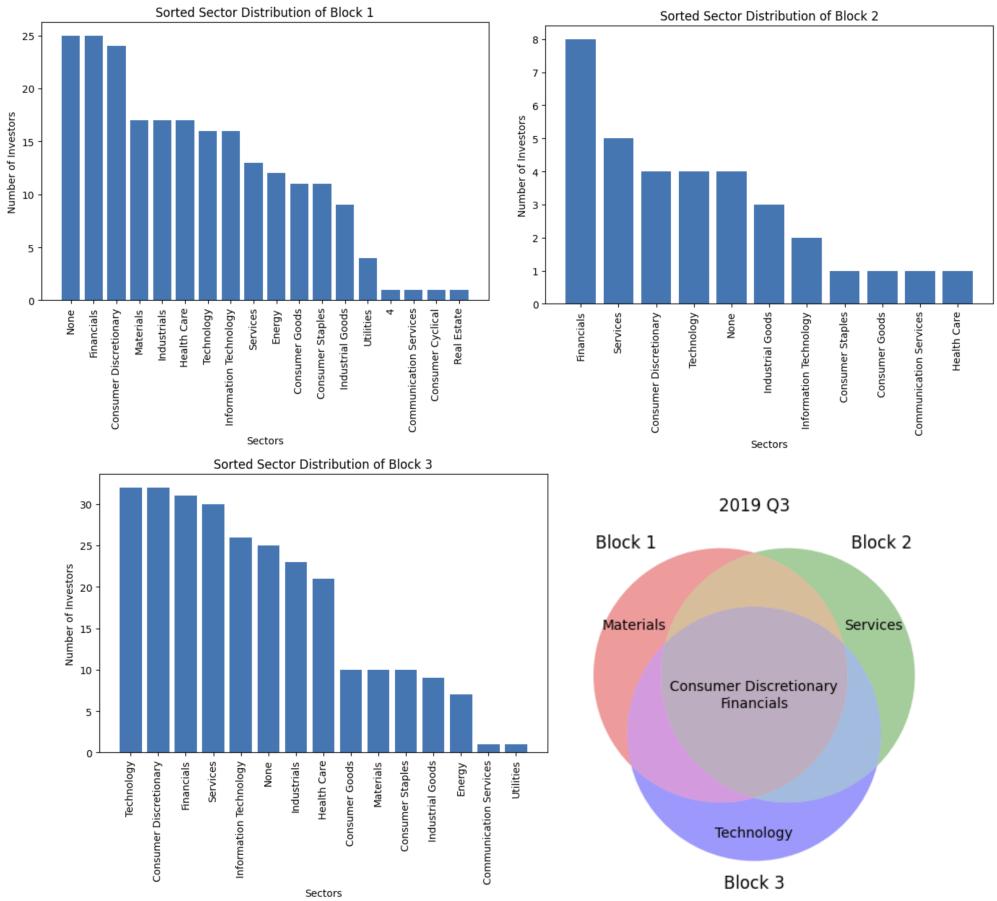

Figure 4.8: Sector distribution in blocks - 2019 Q3

1. Each community has unique preferences, but there are also some common industries. An industry unique to each community may represent a particular investment style or strategy for that community. Some common industries exist in multiple communities, which may mean that the market holds a certain consensus on the prospects of these industries. These sectors may be experiencing some well-known macroeconomic trends, technological innovations, or policy support that have attracted the attention of many investors. From the industry preferences of different communities, we can mine which industries may be undervalued or overvalued. For example, if a particular industry is prevalent in one community but not common in others, it may be that the community has unique information or a deeper understanding of the industry. We can consider industries that follow the market consensus while looking for opportunities in certain industries that are highly regarded by specific communities to build a portfolio.

2. Comparing the VeNN graph of 2019 Q3 and 2021 Q2, we can see that the com-

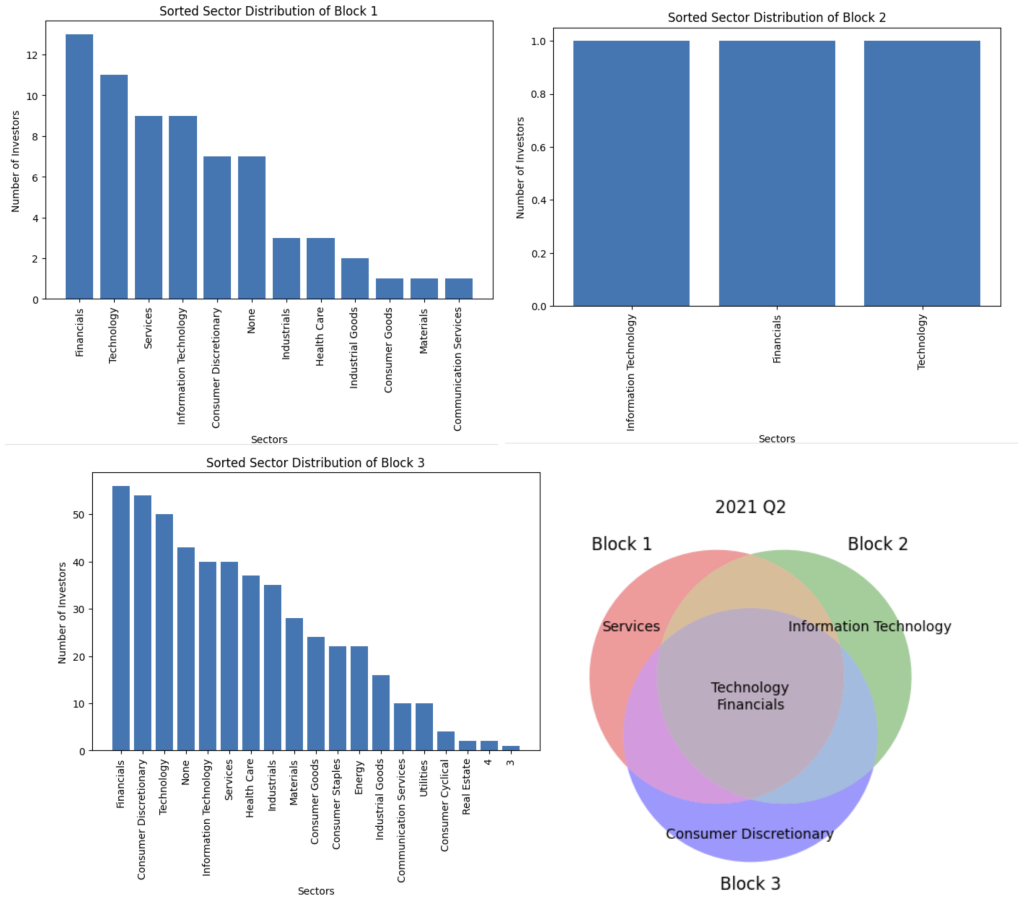

Figure 4.9: Sector distribution in blocks - 2021 Q2

monality and personality parts change over time, which indicates that the market, investment strategy and external environment are dynamic. The market is an evolving entity, influenced by various factors, including macroeconomic, policy, technological progress and social change. As these factors change, investors' views and strategies will adjust accordingly.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

This project aims to determine if network signals can detect early signs of financial shocks and if famous investors' portfolios can be adapted to build successful portfolios. For early signs of financial shocks, this project conducted an evolution analysis of nodes, links, average weight, and weighted clustering coefficients of the networks and the evolution of communities. According to evolution analysis, we found that no new investors joined, and investment behaviour (links) did not change significantly after the outbreak of COVID-19, which indicates the economy stagnated; the financial crisis in 2008 led to an increase in the weighted clustering coefficient, while the pandemic led to an increase in the average weight. These are all signals of a financial shock's occurrence on the evolution graph. However, they are almost all lagging, and we have yet to find a signal that can predict the occurrence of a crisis in advance, which may be because the evolution analysis in this project is too macro, and we need some microanalysis.

For building a compelling portfolio, this project conducted centrality and community analysis. We identified the key investors in the network through the analysis of betweenness centrality and closeness centrality. By comparing their portfolios, we found stocks in which more than half of the critical investors were invested. We checked the performance of these stocks over the next year. They did increase over the next year, which is a promising finding. Although our data sample is small, it is an entirely random selection. Therefore, this is a surprising finding, which means that it is possible to build a portfolio based on the holdings of those famous investors. This project also analysed the industry investment preferences in each community and compared the

similarities and differences in industry preferences among different communities. We found that each community has unique investment preferences, which suggests that the results of community testing are meaningful. We can consider industries that adhere to market consensus while looking for opportunities in specific industries that are well-regarded by particular communities to build a portfolio. We can also mine which industries may be undervalued or overvalued from the industry preferences of different communities. The current analysis results show that it is possible to build an adequate portfolio based on the choice of famous investors.

5.2 Future Works

Firstly, the community analysis was primarily focused on the industry in this project. We can analyse stocks in the specific industry to gain valuable information in the future. Secondly, this project introduced ABC (associations beyond chance) value as the weight of edges in the network, which indicates the likelihood of two investors holding the same stocks beyond what is expected by chance. However, it was only used as a covariate when conducting centrality analysis and community detection. We may get a portfolio by analysing the weight distribution in the network. Analysing those edges with high weight can help identify stocks in the current market that are considered to have high value or potential. Through the analysis of weights, it is possible to identify which investors frequently invest in particular stocks together, which may suggest that they have similar investment strategies or sources of information. This can be a concentrated risk point if many highly weighted investors are concentrated in a particular stock or sector. Identifying these risk points is essential for risk management and asset allocation.

Bibliography

[1] Ulrtk Brandes, Garry Robins, Ann Mrcr Anif, and Stanley Wasserman. What is network science? *Network Science*, 1:1–15, 2013.

[2] Hocine Cherifi, Gergely Palla, Boleslaw K Szymanski, and Xiaoyan Lu. On community structure in complex networks: challenges and opportunities. *Applied Network Science*, 4(1):1–35, 2019.

[3] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[4] Dattatray P Gandhmal and K Kumar. Systematic analysis and review of stock market prediction techniques. *Computer Science Review*, 34:100190, 2019.

[5] Derek L. Hansen, Ben Shneiderman, and Marc A. Smith. Introduction to social media and social networks. *Analyzing Social Media Networks with NodeXL*, pages 3–9, 1 2011.

[6] Brian Karrer and M. E.J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 83, 1 2011.

[7] Clement Lee and Darren J Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50, 2019.

[8] Filippo Menczer, Santo Fortunato, and Clayton A. Davis. *Small Worlds*, page 36–65. Cambridge University Press, 2020.

[9] Tiago P. Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4, 2014.

[10] Tiago P. Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95, 1 2017.

[11] Tiago P Peixoto. Bayesian stochastic blockmodeling. *Advances in network clustering and blockmodeling*, pages 289–332, 2019.

[12] Tiago P Peixoto. Descriptive vs. inferential community detection in networks: pitfalls, myths, and half-truths. *arXiv preprint arXiv:2112.00183*, 2021.

[13] Guillermo Romero Moreno, Valerio Restocchi, Jacques D Fleuriot, Atul Anand, Stewart Mercer, and Bruce Guthrie. Associations between morbidities in small but important subgroups: A novel bayesian approach for robust multimorbidity analysis with small sample sizes. *Available at SSRN 4515875*, 2023.

[14] Felipe Vaca-Ramírez and Tiago P. Peixoto. Systematic assessment of the quality of fit of the stochastic block model for empirical networks. *Physical Review E*, 105, 5 2022.