Using Non-parametric Stochastic Block Models to Analyze Dynamic Network in the Stock Market Through Covid

Xiangyu Tian

Master of Science
School of Informatics
University of Edinburgh
2023

Abstract

This project explores the impact of the COVID-19 pandemic on the stock market network and its community structure. Firstly, we create a temporal bipartite network of stocks and investors, and attempt to use NPSBM to detect the communities of time sliced subnetworks. Due to difficulties encountered in attempting community detection, the project adopted a simulated annealing method, but has yet to achieve results. We decided to use ABC to calculate weights to create a dynamic stock network. Representative selections were made for three sub networks in 2018, 2020, and 2022, and community detection and dynamic analysis were conducted. The results show that the emergence of Covid-19 leads to frequent disappearance and restructuring of communities, indicating that a lot of investors choose to update their investment portfolios at the beginning of the epidemic. During the pandemic, investors chose a co-holding strategy to reduce risk. After the end of the pandemic, its impact on the stock market continued, reflecting persistent changes in investor psychology and behavior. This project also discusses the limitations of network analysis and NPSBM, while providing suggestions for future development directions.

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.

(*Xiangyu Tian*)

Acknowledgements

First and foremost, I would like to deeply thank my supervisor Valerio Restocchi. Throughout my academic journey, you have provided invaluable advice, unwavering support, and patient guidance. Your professionalism and teachings have immensely benefitted me, and your insights will be a significant guide in my future career endeavors.

I also want to express my gratitude to my dear parents. Your unconditional love and support have brought me to where I am today. Your encouragement and faith have always been the driving force as I pursued my academic and personal goals. I am grateful for the strong backbone you provided whenever I faced challenges and setbacks.

Table of Contents

1 Introduction **1**

1.1 Motivation . . . 1

1.2 Problem Statement . . . 2

1.3 Hypothesis and Objectives . . . 2

1.3.1 Hypothesis . . . 2

1.3.2 Objectives . . . 3

1.4 Dissertation Outline . . . 3

2 Related work **4**

2.1 Temporary bipartite network . . . 4

2.1.1 Bipartite network . . . 5

2.1.2 Temporary Network . . . 6

2.2 ABC . . . 6

2.2.1 Background . . . 6

2.2.2 Formula and Value . . . 8

2.3 Non-parametric Stochastic Block Model (NPSBM) . . . 9

2.3.1 SBM . . . 9

2.3.2 NPSBM . . . 11

2.3.3 NPSBMs' Advantages . . . 12

2.4 Network Analysis Methods . . . 12

3 Data **15**

3.1 Data Source . . . 15

3.2 Operation for Extracting Data . . . 15

4 Methodology **17**

4.1 Investor and Stock Network . . . 17

4.1.1 Network Construction . . . 17

4.1.2 Community Detection 19
4.1.3 Difficult and Suggestion 19
4.2 Stock Network 21
4.2.1 Data pre-processing 21
4.2.2 Network Construction (by ABC) 22
4.2.3 Model selection 23
4.2.4 Community Detection 25

5 Result and Evaluation 27
5.1 Community Analysis 27
5.1.1 Result Analysis 27
5.1.2 Visualization 29
5.1.3 Dynamic changes in the community 30
5.2 Network Analysis 31

6 Conclusion and Future Work 37
6.1 Conclusion 37
6.2 Limitation 38
6.3 Future Work 39

Bibliography 40

Chapter 1

Introduction

1.1 Motivation

In the past few decades, financial markets have undergone tremendous changes, with unprecedented growth in the complexity and speed of stock trading. Previous models often found it difficult to capture the true structure of the market. Therefore, new non-parametric techniques must be used to improve stock market analysis. Non-parametric stochastic block model is a method used to infer structures in complex networks. It can use non-parametric Bayesian methods based on random block models to adaptively determine the number of communities in the network[25]. This means that you don't need to pre-set the number of communities, it can adjust parameters to infer the most suitable number and structure of communities. At present, NPSBM is widely used in various fields, such as bioinformatics[33], social network analysis[1], financial network analysis[37], etc. Due to its excellent performance, it is widely used to help people analyze complex network models.

At the same time, with the development of computer computing power and the deepening of people's understanding of the stock market, dynamic networks have gradually entered people's vision. The stock market can be seen as a complex network, where nodes represent different financial entities, while edges represent their interactive relationships[29]. In the early days, people mainly used static networks to describe various relationships in the stock market. These static networks capture the market structure at specific time points, such as certain correlations in the stock market on a certain day or period. However, this method has limitations as it cannot capture features generated by time changes. A dynamic network is a network model that reflects the changes in system state over time. It adds a time tag to the static network, making it a time slice network, and countless time slice networks combine to form a dynamic network. Dynamic networks often have better insights, as they can better understand the interaction mechanisms and potential influencing factors in the market by observing the changes in the network on a time scale. Building such a network might can enhance investment portfolios and raise the possibility that returns will benefit different stakeholders, such as individual and institutional investors[23].

The COVID has brought unprecedented difficulties and adjustments, and the stock market is no exception. Undoubtedly, there has been a significant structural change in the stock market, which provides us with a unique opportunity to analyze the stock market network before, during, and after the pandemic. Through these analyses, we can try to understand how the market responds to such a major shock event and what specific changes it has made after being shocked.

1.2 Problem Statement

The crisis caused by the Covid-19 pandemic has had a profound impact on the stock market, and in this context, deepen analysis and understanding of the stock market has become particularly important. However, traditional network analysis methods may not be able to fully capture the rapid and complex changes in the stock market caused by the Covid-19 pandemic. Therefore, using NPSBM to analyze dynamic networks established using data from this period has become very important. They provide us with a new perspective to study the dynamic changes of the stock market, thereby avoiding further economic losses in the future.

1.3 Hypothesis and Objectives

1.3.1 Hypothesis

We have three research hypotheses for this project: Assumption 1: The data we obtain is objective and accurate, and there is indeed a certain relationship in the data. This relationship and some special attributes of the data can also be reflected in the dynamic network. Assumption 2: During the epidemic, there is a stable community structure in the stock market network, and there are some similarities between nodes in the same community. Assumption 3: The NPSBM is very effective in detecting potential community structures in dynamic networks.

1.3.2 Objectives

Based on the provided data, establish a dynamic network representing the stock market. Using a non-parametric random block model to identify the core communities and subnetworks in dynamic networks during the epidemic, as well as how they evolve and change over time. observe the impact of the COVID-19 pandemic on the dynamic network and its community structure. The project aims to provide valuable insights into complex relationships in the stock market, which may lead to the development of more informed investment strategies. However, it is also possible that the algorithms of non-parametric statistic block models may not work on certain network structures in certain situations, resulting in the inability to accurately obtain the correct number of communities. This has happened in our project, and I will provide a detailed introduction to this issue in 4.1.3 of the paper.

1.4 Dissertation Outline

This dissertation is divided into 6 chapters. Chapter 2 will introduce the releted work of the Temporary Bipartite Network, Association Beyond Chance, and non-parametric stochastic block models, as well as some well-known network analysis methods. The methods of data acquisition and preprocessing will be introduced in Chapter 3. In Chapter 4, we will mainly demonstrate the methods used in this project, including the two constructed networks, model selection, community detection, etc. The results of the project, as well as the analysis and evaluation of the results, will be presented in Chapter 5. Finally, we will present the conclusion, limitations, and future work in Chapter 7.

Chapter 2

Related work

2.1 Temporary bipartite network

A temporal bipartite network is a network where the interactions between two different sets of nodes evolve over time. This type of network provides a richer representation of the system, where not only are there two types of interactive entities, but the interactions themselves also change over time. In practical terms, imagine a network composed of authors and scientific papers. The bipartite property captures the relationship between the author and their paper (assuming that the author and author are not directly related, and the paper is also not related to other papers). If this network still captures the writing time of each paper and displays the progress of each author's publication over time, then it becomes a temporal bipartite network. (As shown in Figure 2.1). This network integrates two important concepts in network theory: the temporal evolution characteristics of dynamic networks and the dual node set structure of bipartite networks. This fusion will address some of the pain points of these two type networks and provide a powerful framework for analyzing complex system. Below, we will introduce bipartite networks and temporal networks respectively.

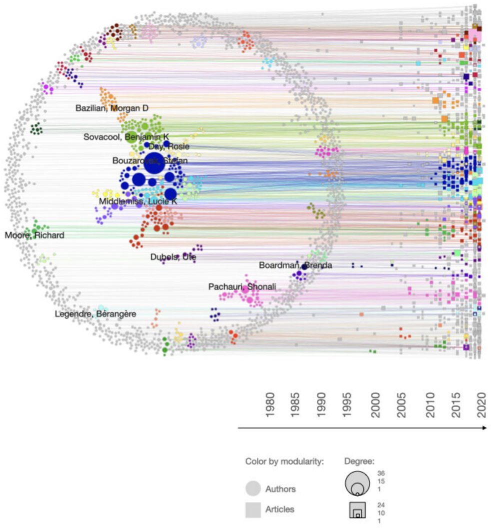


Figure 2.1: [16]

2.1.1 Bipartite network

Bipartite Networks is a core concept in network science that describes the relationship between two sets of vertices. A bipartite network will divide nodes into two groups according to their attributes or people's requirements, and then add edges between the two groups of nodes, ensuring that there are no edges between the nodes within this group. In other words, all links only occur between these two different groups. The characteristic of bipartite networks that can only connect different groups of vertices has led to their prominent position in various scientific research. As early as the last century, mathematicians had already begun to study this special type of graph, which was systematically elaborated by König in the early 20th century[27]. But it wasn't until the emergence of Hungarian algorithms that this powerful concept truly had a profound impact on graph theory and combinatorial optimization. The Hungarian algorithm was proposed by Harold Kuhn in 1955 based on the work of König and Egerváry[28]. This algorithm is mainly used to solve allocation problems, utilizing the characteristics of bipartite graphs to find the best match. Its success has attracted widespread attention in the field of graph theory and has been used in many practical applications. Over time, financial scholars and analysts are increasingly adopting the application of bipartite networks in the stock market to study the interaction and correlation between stocks. For example, Guillaume and Latapy explored interbank exposures by leveraging a bipartite network approach, which uncovers the intricate relationships between banks based on the strength and structure of their mutual exposures[17].

However, a bi partite network still belongs to a static network, which lacks consideration for the time dimension, which means it cannot capture the dynamic interaction of nodes in the system, such as the activity of nodes, the establishment and disappearance of relationships, etc.

2.1.2 Temporary Network

Temporal networks are also dynamic networks, referring to networks that connect and evolve over time, providing dynamic changes for traditional static network models. The difference between these two types of networks is that the links between nodes in static networks are fixed, while these links in temporal networks disappear or reconnect over time, and the strength or properties of edges also change over time. Early foundational work, such as the work of Holme and Saramäki, established the core concepts and methods for analyzing temporal networks[22]. Later, Perra et al. explored the dynamic processes of temporal networks and elucidated how these temporal characteristics can significantly change people's understanding of networks[36]. Temporal networks have many practical application scenarios, which can describe the spread of infectious diseases over time and record financial transaction records that change over time. The characteristic of dynamic networks being able to record time labels makes these networks more complex but also fascinating, as they may be able to uncover unique patterns and behaviors that are still masked in static network analysis[8]. However, a simple temporal network may have the disadvantage of being too complex in structure, as it may contain complex connection patterns where one node may be connected to multiple other nodes. In contrast, temporal bipartite networks have a clearer structure, where nodes in one set are only connected to nodes in another set, making them easier to analyze and interpret.

2.2 ABC

2.2.1 Background

Within the scope of statistical evaluations, understanding the potency and relevance of relationships between variables is pivotal for sound conclusions. Various measures have emerged over time to articulate this degree of correlation, with relative risk being a prominent example[49]. This article presents a novel metric, Association Beyond Chance (ABC), formulated within a Bayesian context[31]. ABC's primary goal is to offer a deeper understanding of associations, emphasizing co-occurrences linked to shared risk factors.

ABC delineates the proportion comparing the joint occurrence tied to shared risks and the incidental appearance associated with separate factors. At its core, this measure tries to identify associations that are unlikely to have occurred purely by chance. Such a measure becomes crucial in studies where it's vital to distinguish between genuine associations and those that might be mere coincidences.

Traditional statistical methods rely on fixed probabilities and often use the frequentist framework. However, Bayesian statistics introduces the concept of ”prior beliefs”, allowing researchers to incorporate prior knowledge about a parameter. It then updates this belief based on new evidence (data) to provide a ”posterior” belief. ABC leverages the Bayesian approach, focusing on the co-occurrence of events beyond what is expected by chance. This Bayesian framework is more flexible, accommodating prior information and new evidence to produce more nuanced results.

One of the unique challenges in statistical modeling is dealing with extreme cases, especially when data is sparse. For situations with limited observations, results can be heavily influenced by outliers. ABC addresses this by incorporating weakly informative priors. These priors play a regulating role, ensuring that the model doesn't produce extreme values based purely on a limited set of data points. The prior essentially acts as a gentle guide[53], ensuring that the ABC remains grounded. In ABC, as more empirical evidence becomes available, the influence of the prior diminishes. This is a testament to the dynamic nature of Bayesian statistics, where data continually informs the assumptions, refining them with additional information.

Incorporating prior beliefs can be a double-edged sword. While they provide context and direction, there's always a risk of introducing bias, especially if the priors are too strong or misinformed. The ABC approach emphasizes the use of weakly informative priors. This choice is deliberate, ensuring that the results are not unduly influenced by the priors. It reflects the exploratory nature of the framework, prioritizing new findings and associations over preconceived notions.

At the same time, an innovative aspect of the ABC framework is the incorporation of hyperpriors. Though beneficial, Priors still bring along the challenge of choosing appropriate values. Hyperpriors, priors on priors, offer an added layer of flexibility. By including hyperpriors, ABC reduces the arbitrariness in the choice of priors, making the model more robust and adaptable.

The Association Beyond Chance (ABC) framework presents a promising approach to association analysis, especially in scenarios where discerning genuine associations from mere chance is challenging. By seamlessly blending the Bayesian approach with strategic choices of priors and hyperpriors, ABC offers a more refined, flexible, and context-aware tool for researchers. It stands as an exemplar of how modern statistical methods can evolve to address contemporary research challenges.

2.2.2 Formula and Value

iven two conditions, we propose that the occurrence of each condition for an individual can be represented by a binary variable, Ci, which can take values 0 or 1. We also posit that there exist specific independent risk factors, represented by Fi, influencing the presence of each condition individually. In addition, there is a shared risk factor, denoted by Fij, that simultaneously impact a combination of two conditions. Based on conditional independence and Bayesian prior calculations, the following ABC formula[39] is obtained:

$$ABC = (Pij - Pi * Pj) * (1 - Pi) * (1 - Pj)/(Pi - Pij)/(Pj - Pij)/(1 - Pi - Pj + Pij)$$

Pi represents the likelihood of event i taking place. Pj denotes the chance of event j happening. Pij signifies the probability of both events i and j occurring together.

As previously noted, RR and ABC bear a close resemblance, leading to analogous interpretations of their values. We first introduce what is RR and its value meanings.

Relative risk stands out as an essential statistical tool when analyzing data from a variety of studies, as pointed out by [49]. This encompasses ecological, cohort, medical, and intervention-based research. The primary objective is to gauge the relationship between specific exposures, such as medical interventions or potential risk factors, and the outcomes they lead to. Essentially, (as shown in figure 2.2) the computation involves comparing the incidence rate among the exposed, denoted as I_e, with the incidence rate among the unexposed, symbolized as I_u. This measure becomes particularly insightful when assessing the possible detrimental effects linked to medical treatments, especially when juxtaposed against scenarios without such treatments, or in contexts of environmental risk evaluations.

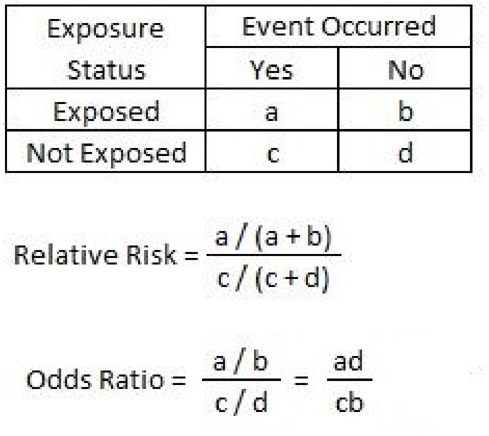
One can consider the following guidelines when interpreting the direct relationship between exposure and outcome via relative risk:

- A relative risk of 1 indicates the exposure leaves the outcome unchanged.

- A value below 1 denotes that the exposure diminishes the outcome's likelihood, acting as a safeguard.

- Conversely, a value exceeding 1 suggests that the exposure amplifies the risk of the outcome, labeling it a potential hazard.

ABC values span from approximately -1 to a conceivable positive infinity. Within this spectrum, -1 signifies a completely opposing relationship while positive infinity indicates a fully aligned association. A value of 0 suggests that the two elements are unrelated, showing no association. If the denominator becomes 0, it reflects an impeccable correlation, resulting in an boundlessly positive value.



Exposure Status	Event Occurred	
	Yes	No
Exposed	a	b
Not Exposed	c	d

$$\text{Relative Risk} = \frac{a/(a+b)}{c/(c+d)}$$

$$\text{Odds Ratio} = \frac{a/b}{c/d} = \frac{ad}{cb}$$

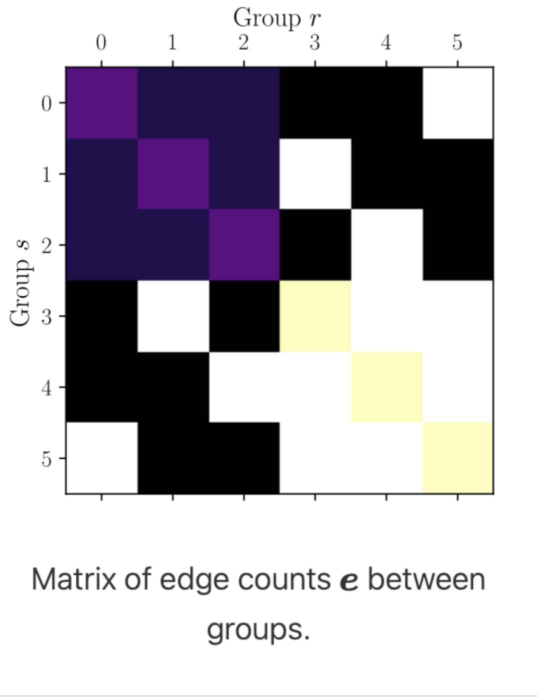
Figure 2.2: [49]

2.3 Non-parametric Stochastic Block Model (NPSBM)

2.3.1 SBM

Stochastic block model (SBM) stands out as a significant category of statistical models used in the study of social networks. It has also risen to prominence as a pivotal instrument in deciphering the complex architecture of networks. It presents a probabilistic framework for graphs, wherein nodes are organized into specific groups or ”blocks”. By adopting an SBM, network nodes are grouped into distinct clusters, with nodes within each cluster exhibiting consistent connectivity behaviors. This concept is referred to as stochastic equivalence[21].The fundamental principle behind the SBM is simple yet powerful: nodes belonging to the same cluster tend to connect more frequently than those in separate clusters. This methodology has proven invaluable in shedding light on community patterns across diverse real-world networks, from biological pathways to human social dynamics. When an SBM is applied and statistical analysis conducted, it not only reveals the previously undisclosed group affiliations but also often accomplishes community identification or assortativity. This indicates that nodes within a particular cluster are stochastically alike and closely linked, while inter-group connections remain predominantly sparse.

Central to the SBM is a probabilistic graph model that embeds predefined clusters. Two primary elements define an SBM: the count of clusters or blocks and a probability matrix that signifies the chances of connections between these clusters. This matrix is the heart of the SBM (as shown in figure 2.3), illustrating the likelihood of a connection between two nodes based on their affiliated blocks. For example, considering two clusters A and B, the matrix might indicate probabilities such as P(A, A) (the chance of a connection between two nodes within block A) and P(A, B) (the chance of a connection between a node from block A and another from block B).



Matrix of edge counts $\boldsymbol{e}$ between groups.

Figure 2.3: [12]

2.3.2 NPSBM

To effectively delve into the concealed structures of networks, it's imperative to utilize generative models, followed by a systematic deduction of their parameters from the data at hand. When these hidden structures manifest as modules or ”communities,” the stochastic block model (SBM) emerges as the go-to strategy. In this framework, nodes find themselves categorized into groups, and the presence of edges is conditioned upon these group associations. In our current exploration, we leverage a Bayesian non-parametric methodology[35] to decode the layered architecture of empirical networks, shedding light on both the count and hierarchical arrangement of these modules. We also call this methodology- non-parametric stochastic block model (NPSBM).

Delving deeper, nonparametric methods, especially within the Bayesian paradigm, offer a robust toolkit to navigate the complex terrains of data, with network analyses being a prime application. Our proposal of a Bayesian adaptation of the SBM aims to amplify our grasp over the nuanced, multi-tiered organization intrinsic to networks. The Dirichlet process[48], a cornerstone in Bayesian non-parametrics, predominantly features in nonparametric data representations, shining particularly in the realms of Dirichlet process mixture models, also termed as infinite mixture models. In essence, this process serves as a distribution over distributions, implying that each extraction from the Dirichlet process encapsulates a unique distribution. The nomenclature ”Dirichlet” arises from its intrinsic property: its finite-dimensional marginal distributions are Dirichlet-distributed. This bears a resemblance to the Gaussian process, another heavyweight in the nonparametric space, characterized by Gaussian-distributed finite-dimensional marginals. Notably, while distributions sampled from the Dirichlet process are inherently discrete, they aren't confined by a static parameter set, justifying their nonparametric tag. Although the Dirichlet process was originally envisioned as a prior with expansive support coupled with manageable posterior computations, it's not without its challenges, especially its predilection for discrete distributions. The transition to more flexible priors seemed unattainable until the revolutionary wave of MCMC techniques. This evolution marked the dawn of a plethora of innovations spanning inference mechanisms, model expansions, theoretical explorations, and tangible applications.

In summary, implementing a non-parametric SBM requires sophisticated probabilistic modeling methods, frequently using concepts like the Chinese Restaurant Process[7] or the Dirichlet Process to account for the indeterminate number of blocks.

Computing these models can be resource-intensive, and specialized techniques, such as Gibbs sampling[15], are typically employed for data interpretation.

2.3.3 NPSBMs' Advantages

The Nonparametric Stochastic Block Model (NPSBM) stands as an advanced technique, exhibiting several notable advantages over traditional community detection approaches and the standard Stochastic Block Model (SBM). Firstly, the nonparametric nature of NPSBM offers a significant edge when handling intricate networks. Traditional community detection strategies typically necessitate the pre-setting of certain parameters, like the number of communities, or they rely on some form of prior knowledge regarding the network's attributes. However, in the real world, these parameters or prior knowledge might be unknown or challenging to ascertain. NPSBM, employing a nonparametric approach, overcomes this limitation, allowing the data itself to guide the discovery of community structures[34]. Secondly, in comparison to traditional SBM, NPSBM can identify and manage communities of varying sizes and shapes. In numerous practical applications, community sizes and densities within networks might significantly differ. The flexibility of NPSBM ensures capturing this diversity more accurately instead of compellingly dividing the network into uniformly sized blocks. Thirdly, many traditional community detection methods are tailored for specific network types, such as undirected networks. In contrast, NPSBM can effortlessly be applied to undirected, directed, weighted, and multilayered networks, expanding its applicability across a variety of network data. Lastly, in comparison to certain conventional methods, NPSBM demonstrates enhanced robustness when confronting noise and uncertainties in the network. This aspect becomes especially salient as real-world network data is frequently incomplete or noisy. NPSBM can furnish stable and precise community divisions even under less than ideal data conditions.

2.4 Network Analysis Methods

In an age characterized by the proliferation of data, the ability to represent and understand vast datasets in meaningful structures is pivotal. The foundational steps towards understanding networks began with research into regular and stochastic networks. However, the intricacies of real-world networks remained elusive to these traditional models. Pioneering research by Watts and Strogatz[52] lifted the veil on the concept of small-world networks, revealing a distinctive blend of short average path lengths between nodes and high clustering. Meanwhile, Barabasi and Albert's groundbreaking study[4] unveiled the notion of scale-free networks, wherein node connectivities astoundingly obey a power-law distribution. These theoretical models have laid an immensely important foundation for network analysis and have aided researchers in better depicting and analyzing various real-world networks. Network analysis is a powerful tool to study complex relationship in the network by breaking them down into simpler components to understand. A network consists of nodes and edges, where nodes represent entities and edges represent the relationships or interactions between them. The number of nodes and edges in a network directly indicates its size and the extent of relationships.

One of the fundamental metrics in network analysis is the degree of a node[41], which refers to the number of connections or edges a node has. The average degree of a network offers a macro-level perspective, suggesting how interconnected, on average, the nodes are. By considering the maximum and minimum degree, one can identify the most connected (hubs) and least connected nodes, respectively. These hubs often play crucial roles in network functionality and robustness.

Network density, another crucial metric, measures the proportion of actual connections to the total number of possible connections[32]. A dense network implies that the nodes are closely connected, facilitating rapid information or substance flow. In contrast, a sparser network suggests fewer connections and possibly slower dissemination rates.

Average path length, a measure of efficiency in a network, calculates the average number of steps it takes to travel between any two nodes[14]. In a 'small-world' network, characterized by a short average path length, information or influence can spread quickly across the network. It provides insights into how efficiently entities in the network can communicate or interact with each other.

The clustering coefficient quantifies the likelihood that two neighbors (nodes connected to the same node) of a node will also be neighbors of each other[45]. It reveals the extent to which nodes in a network tend to cluster together. A high clustering coefficient indicates a pronounced tendency for nodes to form tightly-knit groups, characteristic of many real-world networks where entities tend to form communities or clusters.

Lastly, degree distribution[18] is a fundamental concept in network theory that describes how the number of links (or edges) connecting to nodes (or vertices) is distributed in a network. In other words, it provides a statistical measure of how many nodes have a certain degree.

As a result, network analysis, through its suite of metrics, provides a holistic approach to understanding complex systems.

Chapter 3

Data

3.1 Data Source

In this project, the main dataset comes from https://www.dataroma.com/m/home.php. The website records portfolio (investor) data of well-known investors. It also includes the top 20 holding histories of each investment portfolio. The comprehensive dataset includes data from 67 quarters, starting from the fourth quarter of 2006 and ending in the second quarter of 2023. We hope to divide the data into 67 Excel files based on time, and the file name of each Excel file is the corresponding time for this set of data. In Excel, the first column contains the names of all investors at that time, while the remaining columns represent the top 20 stocks held by the investor during that time. However, one challenge encountered is that this set of data is not directly presented on the webpage, so we need to click on the hyperlink to view a certain investment portfolio and its top 20 stock holding historical data, and then need to traverse all the hyperlinks. Next, we will proceed with this operation.

3.2 Operation for Extracting Data

We initially set a user agent to emulate a browser for web content requests. The next step is to use the ‘requests‘ library to accesses the manager list page and then employ ‘BeautifulSoup‘[38] for HTML parsing. Based on the parsed content, the URL links for each portfolio are fetched which can lead to each investor’s specific portfolio page. On these pages, the links which is related to portfolio history continue to be retrieved and revisits them to extract detailed information about various holdings. For every holding in a portfolio, the stock name, its specific details, and its percentage in the portfolio are extracted. All this information is organized into a list and stored in ‘pf_list‘, with each row representing a specific held stock. Transitioning to the second phase, we mainly perform the task of segmenting the data, previously scraped from the Dataroma website, by quarter and then exporting the data for each quarter into separate Excel files. The first job is iterating through the data list, identifying and gathering all distinct quarters. Subsequently, it is important to create a new data sublist for each identified quarter, containing only the relevant data for that specific quarter. Lastly, the data for each quarter is converted into a DataFrame format and saved as an Excel file, named after the corresponding quarter, stored in the ”/content/” directory.

Chapter 4

Methodology

4.1 Investor and Stock Network

In this section, we will establish a temporal bipartite network of stocks and investors, and conduct community detection on all time slice networks. We will focus on the time slice network and its community structure before, during and after the epidemic to achieve our Objectives.

4.1.1 Network Construction

We first introduced 67 Excel data and constructed an undirected graph ’g’, in this graph, nodes represent either investors or stocks. Each node is assigned specific properties: the ”name” property ensures easy identification, and the ”type” property serves as a crucial differentiator, classifying nodes either as investors (label ’0’) or stocks (label ’1’). Edges in this network describe relationships, each edge signifies an investor’s holding of a specific stock, and to capture when this holding occurred, it’s given a ”time_label” property. A data slice is a block of data from different time frames and is then used to populate the graph. For each data slice, each investor and their holdings are checked, the code checks the graph to determine if the investor already exists as a node, the process adds new investors and identifies existing investors. The same procedure applies to stock, and for each stock held by the investor, the chart is checked for existence. New stocks are added as nodes, while existing ones are identified. Each investor - stock pair creates an edge in the graph that represents the investment relationship with the ”time_label” attribute. This process ensures that nodes or edges are not duplicated to ensure clarity and accuracy. The result is a dynamic bipartite network[19]: one side is the investor and the other is the stock (As shown in Figure 4.1). The introduction of ”time_label” provides a dynamic perspective on the investor’s investment patterns over time, the popularity of stocks, and the consistency of their portfolios. Since we need to perform community detection on 67 time slice networks, we will continue to process our newly obtained time bipartite network. We hope to filter out a subgraph from a large graph based on specific time labels, which only contains edges that match a given time label and vertices associated with these edges. We first defined a function ‘filter_ By_ time_ label’, this function first creates a Boolean edge for the ‘edge_ filter’ attribute is used to determine whether to retain the edge in the filtered graph. At the same time, a Boolean vertex was also created for the ‘vertex_ filter’ attribute determines whether to preserve vertices related to the specified timestamp. After defining these attributes, the function starts traversing all edges in the original graph g and checks each edge ‘time_ label’ attribute. As long as the time label of an edge matches the specified timestamp, the edge will be retained and the two vertices connected to it will also be marked as preserved. If it does not match, the edge will not be retained. After this series of operations, using the ‘GraphView’ class, a new filtered graph view filtered is generated from the original graph g based on these reserved tags. Only those Vertices and edges with a filter attribute of True will appear in the new graph view. The function will ultimately return the filtered graph view.

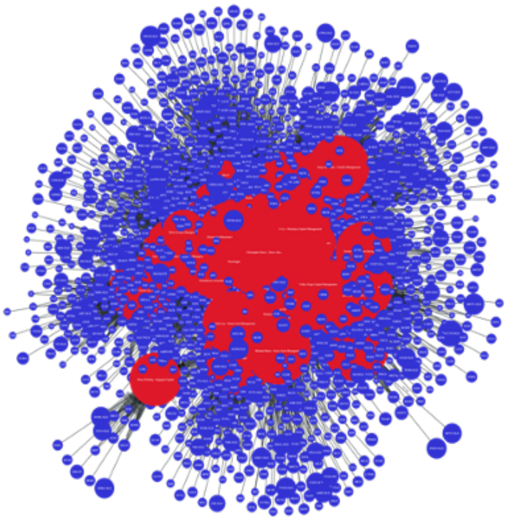


Figure 4.1: Stock and investor temporal bipartite network diagram, red dot for investor, blue dot for stock, time information on every edge.

4.1.2 Community Detection

Community detection has always been an important topic in complex network research, especially in dynamic graphs or scenes with time series data. The method we have chosen is a time based subgraph and Non-parametric Stochastic Block Model (NPSBM). This method aims to detect and track the community structure over time.

Firstly, we obtain subgraphs for a specific time. A subgraph is extracted from the original network and only contains nodes and edges that exist at a given time point. This method allows me to focus on the network structure at a specific time, while ignoring other irrelevant information. This is particularly valuable for understanding the temporal evolution of networks, as it can reveal how communities form or disappear within a certain period of time. After checking the subgraph, we must determine whether it contains valid information. If the subgraph does not have vertices or edges, it is actually an empty graph and there is no need for community detection. In this case, I will return a message like ’No vertices or edges for this time label’. This not only provides me with a clear indication of data integrity, but also allows me to avoid wasting computing resources unnecessarily. However, if the subgraph does contain vertices and edges, I will execute a Non-parametric Stochastic Block Model. NPSBM is a powerful community detection tool that can automatically identify and allocate communities without the need to know the number of communities in advance. This is a huge advantage, especially when facing large and complex networks, where the number and structure of communities may be unknown or changing. To save the results at each time point, we used a structure called ‘npsbm_ states’. This is a collection that stores the output of NPSBM for each timestamp. This allows me to not only save the current results, but also easily trace and view past results, which is very valuable for analyzing the time evolution of the network. Finally, once we have completed the community detection of all subgraphs, we will traverse the ‘npsbm_ states’. We can use time labels as reference points to ensure that the results are viewed in chronological order. This step provides us with a complete time series view, showcasing how the community has developed and changed over time.

4.1.3 Difficult and Suggestion

After using the NPSBM model for community detection, we encountered unexpected issues. For each time slice network, the model is divided into two main communities: one containing only investors, and the other containing only stocks (shown as figure

4.2). This result is clearly not what we expected. Considering that it may be a local optimization problem in the model, I further attempted to use simulated annealing technology to optimize community partitioning, but unfortunately, such an attempt did not change the initial results. One possible reason is the structural characteristics of the bipartite network I am dealing with. In a bipartite network, there are two different types of nodes, and their connections form the basic framework of the network. When using models like NPSBM for community detection, the common result is to divide these two types of nodes into a community, as this division is usually the most obvious way structurally. Another possible reason is data volume issues, especially when the data volume of a single time slot network may be too small. The amount of data is crucial for revealing the substructure and community patterns of a network. Only when the data is sufficiently rich can we expect to see more communities and more complex structural patterns. For this issue, this article provides the following solution suggestions: 1. Network pre-processing: Before starting community testing, some pre-processing steps can be taken. One possible method is to use network projection, which can convert a bipartite network into a single mode network. In a single mode network, similar nodes are linked based on their connection with other nodes, which may help reveal the community structure hidden in the bipartite network. 2. Consider other community detection methods: Although NPSBM is an effective tool in many cases, it may be necessary to consider other community detection methods when dealing with specific types of networks, such as bipartite networks. For example, bi spectral clustering can be applied to bipartite networks, which can better capture the community structure in the network. Tensor decomposition[26][44] is another possible method that can handle complex networks containing multiple types of nodes or links, and has the potential to reveal more community patterns.

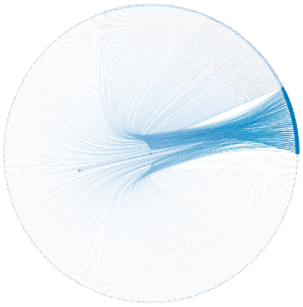


Figure 4.2: Community situation after NPSBM, dark blue dots belong to investor community and light blue dots belong to stock community.

4.2 Stock Network

Due to the difficulties encountered in establishing the Investor and Stock Network that we cannot solve, and addressing these difficulties is not the main focus of this paper, we have decided to use ABC to construct a stock network and conduct community detection on the network to achieve our goal. The stock network has a more simplified network structure, which may make community detection and other analysis more intuitive and effective. In addition, a single stock network may be more likely to reveal the impact of the epidemic on the economy. For example, certain stocks may exhibit strong correlations during the pandemic, which are not significant in other periods.

4.2.1 Data pre-processing

The data format we possess has the first entry of each row as the investor, while the subsequent entries correspond to the stocks associated with that investor. This does not meet our requirements for constructing a stock network. Additionally, to discern trends more easily, we aim to consolidate quarterly data into annual data, necessitating data transformation and pre-processing. Firstly, we must import data from all four quarters, integrating them into a complete annual dataset. Next, to build the desired stock network, it becomes crucial to extract each investor and their associated stocks from this yearly dataset. This involves navigating through each row, pairing the initial entry (the investor) with each subsequent entry (the stocks). In doing so, we can distinctly associate every investor with their respective stocks. To ensure the uniqueness of the data, we add each of these pairings to a set, leveraging the set's intrinsic property of automatically eliminating any duplicate items. Once we've established these investor-stock pairings, the subsequent step entails gathering all investors for each stock. This implies a shift from a many-to-many relationship to a one-to-many relationship, given that a single stock might be associated with multiple investors, and conversely, an investor might invest in multiple stocks. We employ a dictionary structure to achieve this: within this dictionary, the stock name serves as the key, and its corresponding value comprises a list of all investors linked to that stock. Finally, after the aforementioned data consolidation and transformation processes, we have a structured dataset, illustrating the precise relationship between each investor and their invested stocks. To guarantee data persistence and to facilitate future operations, we export this restructured data into a new Excel file. This offers a clear, structured data source for further analysis, visualization, or other related tasks. To establish a stock network that demands datasets in both ”stock-to-investor” and ”investor-to-stock” forms, we must perform similar operations on the recently obtained Excel file to yield the annual data of investors corresponding to stocks.

4.2.2 Network Construction (by ABC)

To devise a stock network through the ABC methodology, one must first grasp the significance of stock correlations, particularly those shaped by overlapping ownership structures, within stock networks. The ABC methodology presents itself as a pioneering tool to capture these correlations by spotlighting the probabilities of co-ownership across stocks. The initial step entails gathering data on the stock portfolios of individual investors, ensuring a clear picture of which investors own which stocks. From this data, one can then calculate the probabilities Pi,Pj, and Pij for every duo of stocks present in the market. Within this stock network backdrop, Pi illuminates the probability that an investor owns stock i, while Pj conveys the odds of this investor holding stock j. Together, Pij encapsulates the combined probability of an investor owning both stocks i and j. Subsequent to this, for each stock pair, determine the ABC value through the given equation. This numerical value becomes an instrumental weight, shedding light on the co-owners-based relationship between two stocks. In essence, the ABC value crystallizes the correlation in ownership patterns of two stocks. A nearing -1 value implies a divergent owners trend between two stocks, hinting that stock i holders might seldom invest in stock j and vice versa. In contrast, a value tilting towards positive infinity signifies a near-perfect alignment in the owners of the two stocks, suggesting a scenario where the owners of one stock almost guarantees the possession of the other. On the other hand, a zero value delineates an absence of any noticeable correlation in the owners dynamics of the two stocks. the ABC methodology offers a nuanced lens through which one can decode and interpret stock ownership intricacies. Embedding co-ownership probabilities into this network structure paves the way for richer insights, intertwining investor choices with overarching stock market ebbs and flows. When building the network, if for any stock pair, the value of Pij exceeds 0, an edge is drawn between them, bearing a weight equivalent to its associated ABC value. This strategy of edge creation provides a granular understanding of relationships between stocks. The entire process is iteratively executed for each year present in ‘time_labels’. As the procedure unfolds for a specific year, the constructed network for that year is appended to the ‘temporal_networks’ dictionary. By iterating through each year and preserving the generated networks in ‘temporal_networks’, we achieve a dynamic stock network. At the same time, with ‘temporal_networks[time_label]’, it’s possible to instantly call upon a subgraph for any given time slice, providing a clear snapshot of the stock relationships for that particular period. The meticulous construction of this dynamic stock network ensures that stock relationships are not only captured accurately but are also easily accessible for any desired timeframe, which offers a holistic yet detailed view of the stock market’s changing dynamics over time, making it a potent tool for analyses and predictions.

4.2.3 Model selection

4.2.3.1 Nested Models and Non-nested Models

Within the domain of non-parametric stochastic block models, two predominant categories exist: the nested models and the non-nested models.

Nested Models[24] operate under the assumption that community structures within a network inherently exhibit a hierarchical nature. This hierarchical approach[50] is reminiscent of the way an extensive tree might branch into smaller sub-trees. The fundamental strength of nested models lies in their ability to discern and represent intricate network structures with great detail. For instance, in a vast network, like a biological system, you can observe macro-structures like organ systems, which can further be divided into individual organs, tissues, cells, and so on. Similarly, in social networks, large communities could represent countries, which further consist of states, cities, neighborhoods, and eventually individual households. This ability to delineate fine-grained hierarchical structures enables the nested models to capture both broad and nuanced interrelations, making them invaluable for deeply interconnected systems.

Non-nested Models[51], juxtaposing their nested counterparts, work under the premise that all nodes exist coherently on a single level. Non-nested Models view the network as a flat structure. So, instead of delving deep into hierarchical nuances, non-nested models segregate the network into distinct non-overlapping communities or blocks. This model is akin to looking at a forest and classifying based on types of trees without concerning the individual branches and leaves. While this might seem oversimplified for intricate networks, its strength lies in its simplicity, making it exceptionally efficient for less complex networks where hierarchical breakdown might be redundant or unnecessary.

Distinguishing the two, the nested models stand out with their layered hierarchical organization, allowing them to depict networks with varying levels of granularity. In contrast, non-nested models, with their single-level portrayal, are more straightforward and computationally less demanding. This simplicity, however, might come at the cost of depth in representation, especially in multifaceted networks.

For a stock network, nested models appear to be the optimal choice for several reasons. Firstly, because our stock network is multi-level and interrelated, any stock in the network may have edges with other stocks. Furthermore, stock market dynamics often reflect cascading effects: global economic trends influence sectors, which in turn affect individual companies and eventually the stock prices of these entities. This cascading, multi-tiered influence underscores the intricate hierarchical nature of stock markets. Thus, utilizing nested models for stock networks ensures a more detailed, nuanced, and, most importantly, accurate representation of these inherent hierarchies and interrelations, facilitating a superior analysis.

4.2.3.2 Degree Correction

Degree correction is an important concept in complex network analysis, especially during community detection[54][9]. In network science, the ”degree” of a node refers to the number of edges connected to it[42]. In certain networks, especially social or financial ones, there’s a vast disparity in degree distribution, meaning that some nodes (referred to as hub nodes or high-degree nodes) connect to many other nodes, while most others connect to only a few. During community detection, these high-degree nodes can disproportionately influence the results. To mitigate this influence and achieve a more meaningful community structure, the concept of ”degree correction” is introduced. The goal of degree correction is to ensure that each node’s influence in community partitioning is proportional to its degree, thus avoiding biases that arise due to the presence of high-degree nodes. By considering each node’s degree and adjusting its weight in the community partitioning process accordingly, degree correction provides a more authentic and meaningful way to understand and reveal the community structure of a network. For our project, aiming to analyze the impact of the pandemic on the stock market, three time-sliced networks from 2018, 2020, and 2022 were chosen as samples before, during, and after the pandemic respectively. This is because these years played pivotal roles in the timeline of the COVID-19 outbreak. 2018 provides a benchmark as a typical year before the pandemic, reflecting the normal circumstances of the stock market. 2020 represents the peak of the pandemic when most countries were severely affected. By 2022, many nations began recovering and adjusting to the new norm, offering insights into the stock market's recovery and shifts post-pandemic. We employed the NPSBM nested model for community detection on each time-sliced network with and without degree correction. After conducting degree-corrected and non-degree-corrected community detections on the three time-sliced networks, we compared the entropies of the two community detections for the same time-sliced network. The results indicated that when degree correction was applied, the entropy of the community detection was smaller. entropy can be understood as a metric that measures the complexity of community partitioning. A smaller entropy typically indicates a more structured and clearer community partition. This implies that by incorporating degree correction, a more structured, clearer, and meaningful community partition result was achieved. This further validates that applying degree correction is essential for your stock market network data.

4.2.4 Community Detection

In this project, we employed NPSBM for community detection within stock networks. This approach is very similar to the section 4.1.2 of the paper. The difference is that we need to transform the newly created stock network from a Network X graph to a GraphTool graph[2] so that we can conduct community detection. A unique variation we introduced was a novel iterative detection procedure to enhance the reliability and stability of our findings. Specifically, rather than merely conducting a single community detection for each temporal slice network, we implemented it tenfold. This iterative approach was instituted to capture variations potentially induced by initial conditions or other stochastic elements. Each detection iteration yields a community partition result and its corresponding entropy value, with the latter serving as a metric to evaluate the quality and complexity of the community division. Ultimately, we elected the detection with the smallest entropy value as our final community detection result, as a lower entropy typically signifies a more structured and stable community partition.

However, our exploration didn't end there. Upon the completion of NPSBM community detection, we delved into the temporal sequence data of the detection results. Initially, we calculated the number of communities each year, aiming to discern possible structural shifts within the stock market. Our focus subsequently shifted to the dynamic alterations between communities across consecutive years.

For instance, examining the span from 2018 to 2019, we ventured beyond just quantifying communities. Our analysis deeply probed the stability, dissolution, and emergence of communities during this period. To be specific:

Stable Communities: These are communities present in both 2018 and 2019, maintaining their structural integrity over the two years. Disappearing Communities: These communities, present in 2018, vanished by 2019. This might suggest a weakening or loss of interconnections between the stocks within these communities in 2019.

Emerging Communities: In contrast to the disappearing ones, these communities manifested in 2019 but were absent in 2018. This might indicate certain stocks beginning to exhibit new analogous trends or mutual connections.

In summation, through the NPSBM community detection and the subsequent in-depth analysis, we aim to uncover the structural nuances of the stock market and its dynamic evolution over time.

Chapter 5

Result and Evaluation

5.1 Community Analysis

5.1.1 Result Analysis

Drawing insights from the stock network structure through the lens of NPSBM community detection across the years 2018, 2020, and 2022 allows for a keen understanding of market dynamics in the face of the global turbulence caused by COVID-19. Prior to the pandemic's onset in 2018, the stock network portrayed a graph of diversity and a well-defined hierarchical depth, initiating with 716 stocks that bifurcated into 133 vibrant communities. Transitioning into 2020 when the COVID-19 was raging, the network expanded, accommodating 813 nodes. This expansion can be attributed to a possible surge in companies gaining prominence or opting for public listing, with sectors like health tech and telecommuting software gaining traction due to their relevance in a pandemic-stricken world[11]. The consistency of the hierarchical depth across the years 2018, 2020, and 2022, with all presenting a level of 5, offers insights into the structural properties of the network. Despite the expected fluctuations and changes that might occur in the market due to various factors, including the influence of the COVID-19 pandemic, the overall structural hierarchy of the market remains unchanged. This might indicate that stock market has a certain degree of adaptability and elasticity, which can withstand short-term shocks and changes. Even in the face of significant challenges such as the pandemic, the market may maintain its community level through internal adjustments and optimizations.

Moreover, 2022 year's result signifies a world recuperating from the pandemic's major onslaughts, the network manifested a marginal shrinkage to 794 stocks. This may be because investors feel that the market is starting to recover, uncertainty is decreasing, and people no longer need to buy more stocks to share risks, resulting in a decrease in the number of stocks. While the hierarchy remains the same in 2022 compared to other years, the actual constituents or characteristics of the individual levels might undergo changes. The consistent numbers in the foundational communities, oscillating between 125 and 150, suggest an enduring significance of certain sectors or investment philosophies. By analyzing the structure of the stock network in 2018, 2020 and 2022, we can gain insight into the impact of the global COVID-19 pandemic on the market. Despite challenges such as the pandemic in the market, the overall structure and hierarchy of the network remained stable. This highlights the stock market's resilience to external challenges and its ability to maintain community levels through internal adjustments and optimizations.

Before Covid

```
Results for year 2018:
l: 0, N: 716, B: 133
l: 1, N: 133, B: 52
l: 2, N: 52, B: 16
l: 3, N: 16, B: 5
l: 4, N: 5, B: 1
l: 5, N: 1, B: 1
```

During Covid

```
Results for year 2020:
l: 0, N: 813, B: 150
l: 1, N: 150, B: 51
l: 2, N: 51, B: 13
l: 3, N: 13, B: 4
l: 4, N: 4, B: 1
l: 5, N: 1, B: 1
```

After Covid

```
Results for year 2022:
l: 0, N: 794, B: 125
l: 1, N: 125, B: 48
l: 2, N: 48, B: 13
l: 3, N: 13, B: 3
l: 4, N: 3, B: 1
```

l: 5, N: 1, B: 1

5.1.2 Visualization

Visualizing the NPSBM results presents an intuitive grasp of the stock market's intricate landscape. Each node and its hierarchical level can represent conglomerations of stocks that possess certain common features or characteristics to draw the interest of a group of investors. The breadth and depth of these hierarchies provide a direct insight into the diversity and specificity of investment strategies.

Each community or node in the hierarchy might represent dominant sectors or industries[30]. By observing which nodes (sectors or industries) persist across different time frames, one can infer which sectors are more stable or robust against various economic challenges. New nodes or communities appearing in the hierarchy might suggest the emergence of innovative sectors or industries, which could be a result of technological advancements, changing consumer preferences, or new market opportunities. In addition, by examining the connections between these nodes, one can derive information about the interplay between different industries. For instance, strong ties between the technology and finance nodes might indicate the rise of fintech. It is worth mentioning that the strength and number of interconnections between communities can provide insights into collective investor behaviors. Tight-knit communities might reflect a unified investor base with a shared belief in particular stock combinations. When we examine the weakening or strengthening of these connections across the pandemic phases, these changes reflect the evolution of investor sentiment, risk preferences, and confidence levels.

A comparison of the three phases uncovers the evolution of market dynamics. The pre-pandemic period in 2018 shows six community levels, indicating diverse investment strategies and a deeper market penetration, suggesting a robust economic scenario. The year 2020 during the height of the pandemic keep the same community hierarchy as 2018. This suggests that despite the tumultuous changes brought about by the pandemic, the underlying structural depth of the market remained intact. Such persistence indicates that, while individual investment strategies might have shifted during the crisis, the overall complexity and depth of the market's structure persisted. Come 2022, the network, although reminiscent of 2018 in its primary layer, indicates a richer, more varied investment landscape, suggesting market recovery and diversification post-pandemic.[40]

Studying patterns and shifts in the community hierarchy over time can empower economists and analysts to predict future trends. If certain community structures or behaviors consistently precede economic booms or recessions, these become invaluable predictive tools. The pandemic, in this regard, can serve as a significant case study for future crisis management. For institutional and retail investors, the visualized hierarchy can serve as a roadmap. By identifying dominant communities or nodes, they can align their portfolios with prevailing market trends. Additionally, the visualization can serve as a tool for identifying outlier communities that might represent untapped investment opportunities or high-risk areas.

(a) 2018

(b) 2020

(c) 2022

5.1.3 Dynamic changes in the community

The evolution of stock market communities between 2018, 2020, and 2022, as characterized by the NPSBM community detection, mirrors the dynamic nature of the financial world in response to broader socio-economic circumstances. In 2018, with 133 communities, the market presented a multifaceted landscape reflecting a myriad of investment patterns, yet, by 2019, a significant 113 of these communities vanished, replaced almost in kind by 114 new ones. This means that with the advent of the COVID-19 epidemic, most investors may be changing their investment models. Transitioning into 2020, a pivotal year marked by the zenith of the COVID-19 crisis, there was a noticeable swell in the number of communities to 150, possibly indicating heightened market fragmentation or diversified strategies, yet this surge was accompanied by a staggering disappearance of 115 communities from the previous year, supplanted by an even more pronounced influx of 131 new communities. This fluidity, in part, can be attributed to the global tumult and uncertainties brought about by the pandemic, reshuffling investor priorities and sectoral focuses. However, as the world started grappling with the aftermath of COVID-19, there was a subsequent contraction to 134 communities in 2021, and eventually, a further decline to 125 communities in 2022. This sequential reduction, juxtaposed with the loss of 115 communities and the emergence of 106 new ones from 2021 to 2022, perhaps hints at a gradual return to stability, with the market slowly recalibrating itself. Yet, with only 19 stable communities in 2022 compared to 2018's 20, it underscores the profound reshaping of the stock market landscape over these years, with enduring communities potentially representing resilient sectors or steadfast investment approaches amidst rapid change. In 2018, there are 133 communities.

In 2019, there are 134 communities.

In 2020, there are 150 communities.

In 2021, there are 134 communities.

In 2022, there are 125 communities.

From 2018 to 2019:

Stable communities: 20

Communities that disappeared: 113

New communities: 114

From 2019 to 2020:

Stable communities: 19

Communities that disappeared: 115

New communities: 131

From 2020 to 2021:

Stable communities: 24

Communities that disappeared: 126

New communities: 110

From 2021 to 2022:

Stable communities: 19

Communities that disappeared: 115

New communities: 106

5.2 Network Analysis

Let's focus on the stock network's evolution across three distinct periods related to COVID-19: before COVID-19 (2018), during COVID-19 (2020), and after COVID-19 (2022).

1. Node and Edge Dynamics:

The dynamics of nodes and edges from 2018 to 2022 depict a market that was in flux, rapidly responding to the externalities of the pandemic. The 13.5% uptick in nodes between 2018 and 2020 mirrors a marketplace that was both reactive and opportunistic. As sectors like healthcare and technology became frontline defenders against the pandemic, they likely saw increased investor interest, translating to new stocks entering the network[5]. Similarly, the boom in e-commerce, propelled by global lockdowns and a paradigm shift in consumer behavior, further expanded the network[6][47]. The substantial 25.% surge in edges in the same timeframe paints a picture of investors actively reshaping their portfolios. This doesn't just reflect an increased interest in specific sectors, but also a broader strategy to diversify portfolios. However, the descent in the number of nodes and edges by 2022 raises intriguing questions. The data might be indicative of a market that's becoming more discerning. After the initial reactive phase, there seems to be a strategic retreat. This could be a sign of investors consolidating their positions, focusing on stocks that promise long-term stability over short-term gains. It's also plausible that the market, having ridden the volatile waves of the pandemic, is now veering towards a more balanced and matured stance. This transition from a reactive to a proactive approach in investment strategy is emblematic of a market that has learned, adapted, and is now more prepared for unforeseen challenges in the future.

2. Degree Analysis:

The upward trajectory in the average degree during 2020 underscores a market that was evolving rapidly. The increase in average degree could be influenced by two reasons. Firstly, there are more direct connections between nodes, suggesting that most stocks have a growing number of common investors. The term ”co-holding patterns” typically refers to the scenario where different investors or institutions have the same or similar stocks in their portfolios. With escalating uncertainty, the surge in such co-holding patterns can be seen as a safety net cast by investors in turbulent waters. This is because most investors are likely to make similar decisions during market shifts, and having a portfolio similar to other major investors can mitigate the risks of price fluctuations. The second reason is the potential emergence of a hub in the network[3]. As the pandemic progresses, more and more investors invest in this hub stock, leading to an increase in the network's average degree. This hub stock is most likely in the medical sector. This too represents a form of co-holding pattern.

Although the average degree decreased slightly in 2022, it still maintains a high level, which is particularly indicative. It suggests that the habits formed or strategies adopted during the stormy phase of 2020 might have entrenched themselves into the investment psyche. This could be indicative of a more systemic and structural shift in the market's modus operandi. Investors might now be prioritizing a broad-based approach over a concentrated one, perhaps influenced by the lessons of the pandemic, where over-reliance on a specific sector or stock could have led to significant losses.

In addition, the wide range of degrees in 2020 is board, which might imply a more inclusive market, where even lesser-known or traditionally less popular stocks found favor in portfolios. Such broad-based diversification is emblematic of a market bracing for the unknown, ensuring that all bets aren't placed in one basket, no matter how lucrative or stable that basket may have seemed in pre-pandemic times.

3. Density and Path Length:

Network density provides a robust insight into how interconnected a system is. The observed decrease in density in both 2020 and 2022 suggests a more sprawling but less densely interconnected stock network. This reflects a market phenomenon where there is the entry of new stocks, which may be related to emerging industries or they are adjusting to adapt to the new environment. But this does not mean that there has been a corresponding increase in proportional connectivity. These stocks might have been welcomed into portfolios, but the relationships between them and the existing stocks weren't as deeply established as among the pre-existing entities.

Several factors might account for this phenomenon. For instance, these newly introduced stocks may come from emerging industries and represent new technologies, and they may not be suitable for the established co-holding model[20]. Furthermore, these stocks, due to their novelty or the inherent risks associated with untested technology during turbulent times, might have been approached with caution, resulting in fewer immediate connections within the network.

The constancy in the average path length shows the resilience in the stock network. In network analysis, the average path length serves as an indicator of the network's efficiency, elucidating how quickly information or influence can spread. Despite the volatility introduced by the pandemic, the unaltered average path length suggests that the overall efficiency of the stock market remained intact. This indicates that while investors diversified and the network expanded, the fundamental interconnectedness of the system held firm. It's as if the foundational threads binding the market together were elastic, stretching but not breaking. This is a testament to the agility and resilience of the global financial system. In the face of an unprecedented crisis, while the landscape of investments shifted and morphed, the core structure demonstrated adaptability. The steadfast average path length showcases that, despite the turbulence of the pandemic, the market's inherent ability to communicate, to transfer value, and to respond remained unscathed.

4. Clustering Coefficient:

The clustering coefficient is an insightful metric, shedding light on how closely stocks tend to cluster or group together[46]. A rising clustering coefficient during these periods suggests that investors are not just buying stocks at random, but are rather meticulously curating their portfolios to reflect specific themes, sectors, or trends. The continuous ascent of this coefficient during the periods of our interest is indicative of a market that's becoming progressively more discerning and strategic. It's plausible to surmise that, confronted by the uncertainties brought about by COVID-19, investors sought solace in solidarity. They converged around those stocks or sectors that showcased either proven resilience or offered a promise of growth amidst the upheaval. Such convergence is emblematic of a risk-averse strategy, where investors cluster around known safe havens or emerging growth stories to mitigate potential downturns. For instance, as conventional business models were challenged, investors might have gravitated towards innovative sectors, such as cloud computing or telehealth, recognizing their transformative potential in a post-COVID world.

The heightened clustering also speaks to a shared investment sentiment. In turbulent times, market sentiment often gets shaped by a blend of economic forecasts, emerging narratives, and collective investor psychology. The increasing clustering could be a manifestation of this collective sentiment, where shared beliefs and strategies lead to similar investment behaviors, creating tighter and more defined clusters. In essence, the evolving clustering coefficient captures a narrative of adaptation and collective strategy. In times where uncertainty was the only certainty, the stock network's ability to maintain its functional structure showcases the strength and resilience embedded in the very fabric of the global financial ecosystem.

5.Degree distribution

The degree distribution of a network offers profound insights into its structure and the behavior of its members, and in the realm of a stock network, the degree of a node showcases the quantity of other stocks it's typically co-held with by investors[10]. A pronounced peak in this distribution hints at prevalent co-holding patterns within the market. Interestingly, the distributions for 2018, 2020, and 2022 all prominently peak at a degree of approximately 25, revealing a consistent co-holding pattern among investors. Despite the myriad of challenges and disruptions ushered in by COVID-19, this consistency underscores the remarkably unwavering nature of investor behavior in terms of diversification.

Furthermore, an intriguing evolution emerges when we delve into the frequencies at this degree: while the degree's peak remains a constant, the frequency has ascended from 240 in 2018 to a notable 340 by 2022. This evolution suggests an expanding pool of stocks aligning with this typical co-holding behavior, pointing towards a potential trend of normalization or even homogenization in investor behavior over the years. Such a trend might be a reaction, possibly shaped by factors ranging from nuanced market analyses and widely endorsed investment practices, to broader external catalysts, a prime example being the global responses and adaptations to the COVID-19 pandemic. This steadfast peak in the degree not only mirrors the consistency but also the resilience entrenched within investment strategies. The multifaceted repercussions of COVID-19 on financial terrains were varied and profound, with diverse sectors undergoing different magnitudes of impact. Nevertheless, this persistent pattern reflects that the quintessential strategy of diversification held its ground as a pivotal element in investment decisions. Adding another layer to this observation, it's pertinent to note the tools and resources investors lean on. A significant portion of investors weave their decisions based on tools, advisory platforms, and analyses. This persistently observed consistency in co-holding patterns might be an indicator that these prevalent analytical tools or platforms exert a standardizing influence over investor behavior, culminating in a degree distribution that has remained steadfast across the years.

Statistics for 2018:

num_nodes: 716

num_edges: 15328

avg_degree: 42.815642458100555

max_degree: 292

min_degree: 4

density: 0.05988201742391686

avg_path_length: 2.2511622455756535

clustering_coefficient: 0.8052209263648691

Statistics for 2020:

num_nodes: 813

num_edges: 19274

avg_degree: 47.41451414514145

max_degree: 393

min_degree: 3

density: 0.058392258799435284

avg_path_length: 2.2092959845854616

clustering_coefficient: 0.8104428831614624

Statistics for 2022:

num_nodes: 794

num_edges: 17379

avg_degree: 43.77581863979849

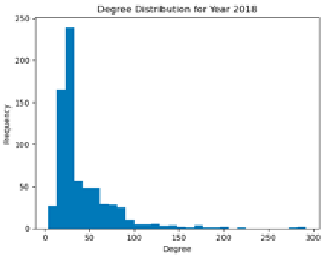
max_degree: 360

min_degree: 3

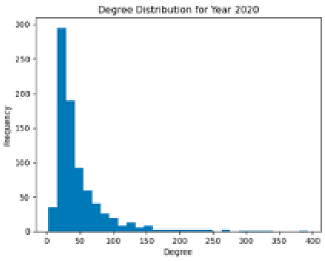
density: 0.05520279778032596

avg_path_length: 2.2545319403724657

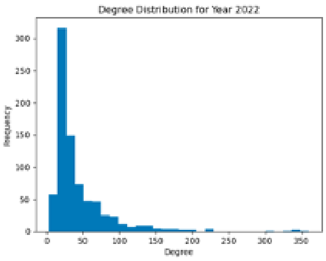
clustering_coefficient: 0.8138486023103274



(a) 2018



(b) 2020



(c) 2022

In general, the data offers a snapshot into how the stock market network's topology evolved in response to the challenges and uncertainties brought about by COVID-19. During the height of the pandemic in 2020, there was evident growth in network complexity and interconnectedness, potentially as investors sought refuge in diversification. As the world began adjusting to a post-COVID-19 reality in 2022, the network saw some degree of consolidation, though it did not revert entirely to its pre-pandemic state. This underlines the lasting impact of global events on market structures and investor behaviors, with COVID-19 serving as a significant case study in market adaptability and resilience.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this project, our goal is to observe the impact of the COVID-19 pandemic on dynamic networks and their community structures. We first created a temporal binary network about stocks and investors using the ‘graph.tool’ tool, and wanted to use NPSBM to perform community detection on the time sliced subnetwork of the network. However, we were not able to successfully partition the community for any time sliced network, and we attempted to use simulated annealing to complete effective community detection, but the results were still unsuccessful. This may be a problem caused by the structural characteristics of the binary network. We have provided our solution suggestions and decided to build a stock network to achieve the goals of this paper. We used ABC to calculate weights and created a dynamic stock network. Subsequently, we selected representatives from three sub networks before, during, and after the pandemic, 2018/2020/2022. We selected the optimal model for these three time slice networks and used npsbm for community detection. Then, we evaluated the results of community detection and conducted dynamic analysis on these three time slice networks. By analyzing and evaluating these three time slot networks and their community structures, we conclude that the emergence of the epidemic has led to a significant and frequent disappearance and restructuring of communities, which means that a large number of investors have updated their investment portfolios. Faced with increased uncertainty, a large number of investors have chosen co holding strategies to reduce their own risks. However, risks often coexist with opportunities, and the epidemic not only brings turbulence to the market, but also brings opportunities to some investors. We believe that the abnormal state of the stock network during the epidemic period is likely due to the addition of medical and information technology related stocks. We also found that the impact of the epidemic did not end with the end of the epidemic, as the network in 2022 still showed relevant trends. The psychological changes brought by the epidemic to investors continue to affect their investment strategy choices and investment styles.

6.2 Limitation

Firstly, the NPSBM's approach, grounded in its ability to model communities within networks, inherently views the stock market as a collection of interacting entities. This perspective is incredibly advantageous for capturing overarching community interactions, but the granularity of individual stock behavior might sometimes become secondary. The stock market, with its myriad of events, announcements, policy changes, and global cues, responds not just at the community level but also at the level of individual stocks. When NPSBM focuses on identifying clusters of similar stock behaviors, it might inadvertently gloss over the unique stories and journeys of individual stocks that diverge from the group.

Another consideration is the model's reliance on past data. Financial markets are notoriously forward-looking. Investors often make decisions based on expectations of future earnings, dividends, and macroeconomic events, among other factors. Hence, while NPSBM might identify historical patterns of stock communities effectively, it might not always predict how these communities will reconfigure in the face of new information. For instance, disruptive technologies can reshape stock communities in ways that past data might not necessarily suggest.

Lastly, the inherently quantitative nature of the NPSBM might not always account for qualitative aspects that play a significant role in stock market dynamics. Factors such as market sentiment, investor psychology, and narrative-driven movements are hard to quantify but have a profound impact on the market. An over-reliance on NPSBM might result in underappreciating these qualitative factors.

While the Non-parametric Stochastic Block Model offers a novel and powerful lens to understand the community structures within the stock market, it should be viewed as one tool among many. To truly grasp the intricacies of the ever-evolving financial markets, a multifaceted approach that blends quantitative modeling with qualitative insights and incorporates both historical patterns and forward-looking perspectives is imperative.

6.3 Future Work

In light of the insights derived and the limitations acknowledged from the use of the Non-parametric Stochastic Block Model (NPSBM) in analyzing the stock market's dynamics during the pandemic, several avenues for future work emerge.

In constructing a temporal stock network, we can combine machine learning technology to train models using historical data to predict future dynamics of the stock market, including price, trading volume, and potential changes in community structure[43]. Combining machine learning technology[55] can provide a new way to interpret and predict market dynamics. Furthermore, while a broad overview of the market is invaluable, there's a growing imperative for sector-specific analyses. Each sector, from technology and healthcare to finance and energy, faced unique challenges and opportunities during the pandemic. A more granulated examination of these individual sectors could unearth nuanced strategies, innovations, and adaptations that a broader lens might miss. Such sector-focused insights could be instrumental in tailoring more effective investment strategies and predicting future sectorial growth trajectories.

Another promising avenue involves integrating NPSBM with other analytical models and tools. Each model, with its inherent strengths and biases, provides a unique perspective. A fusion of these perspectives, through a multi-model approach, could yield a richer and more comprehensive understanding of market dynamics, mitigating the blind spots that any single model might possess. Lastly, and perhaps most intriguingly, is the integration of behavioral analysis into future studies. The stock market, while often perceived as a realm of numbers and logic, is profoundly influenced by human behaviors and psychologies[13]. By intertwining the rigorous structural insights of models like NPSBM with principles from behavioral finance, we could delve into the underlying psychological factors steering market changes. Such an amalgamation could offer unprecedented insights, merging the quantitative with the qualitative, and painting a more complete portrait of the market's performance in the face of global disruptions like COVID-19.

Building stock networks by ABC and the application of nonparametric stochastic block models provide a structured understanding of the stock market's response to the challenge of COVID-19. While the insights derived are invaluable, it's crucial to acknowledge their limitations. Future endeavors can build upon this foundation, aiming for a more holistic grasp of the market's dynamics during COVID-19.

Bibliography

[1] Edo M Airoldi, David Blei, Stephen Fienberg, and Eric Xing. Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21, 2008.

[2] Nadeem Akhtar and Mohd Vasim Ahamad. Graph tools for social network analysis. In *Research Anthology on Digital Transformation, Organizational Change, and the Impact of Remote Work*, pages 485–500. IGI Global, 2021.

[3] Sibel Alumur and Bahar Y Kara. Network hub location problems: The state of the art. *European journal of operational research*, 190(1):1–21, 2008.

[4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[5] Sailee Bhambere, B Abhishek, and H Sumit. Rapid digitization of healthcare—a review of covid-19 impact on our health systems. *Int. J. All Res. Educ. Sci. Methods*, 9:1457–1459, 2021.

[6] Anam Bhatti, Hamza Akram, Hafiz Muhammad Basit, Ahmed Usman Khan, Syeda Mahwish Raza, Muhammad Bilal Naqvi, et al. E-commerce trends during covid-19 pandemic. *International Journal of Future Generation Communication and Networking*, 13(2):1449–1452, 2020.

[7] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):1–30, 2010.

[8] Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. Time-varying graphs and dynamic networks. In *Ad-hoc, Mobile, and Wireless Networks: 10th International Conference, ADHOC-NOW 2011, Paderborn, Germany, July 18-20, 2011. Proceedings 10*, pages 346–359. Springer, 2011.

[9] Yudong Chen, Xiaodong Li, and Jiaming Xu. Convexified modularity maximization for degree-corrected stochastic block models. 2018.

[10] K Tse Chi, Jing Liu, and Francis CM Lau. A network perspective of the stock market. *Journal of Empirical Finance*, 17(4):659–667, 2010.

[11] Rafel Crespí-Cladera, Alfredo Martín-Oliver, and Bartolomé Pascual-Fuster. Financial distress in the hospitality industry during the covid-19 disaster. *Tourism Management*, 85:104301, 2021.

[12] Tiago de Paula Peixoto. graph-tool documentation(2.58).

[13] Abderrazak Dhaoui, Saad Bourouis, and Melek Acar Boyacioglu. The impact of investor psychology on stock markets: Evidence from france. *Journal of Academic Research in Economics*, 5(1), 2013.

[14] Agata Fronczak, Piotr Fronczak, and Janusz A Hołyst. Average path length in random networks. *Physical Review E*, 70(5):056110, 2004.

[15] Alan E Gelfand. Gibbs sampling. *Journal of the American statistical Association*, 95(452):1300–1304, 2000.

[16] Zeus Guevara. Bipartite temporal network, 2023.

[17] Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):795–813, 2006.

[18] Michael Hay, Chao Li, Gerome Miklau, and David Jensen. Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE International Conference on Data Mining*, pages 169–178. IEEE, 2009.

[19] Tobias Hecking, Laura Steinert, Tilman Göhnert, and H Ulrich Hoppe. Incremental clustering of dynamic bipartite networks. In *2014 European Network Intelligence Conference*, pages 9–16. IEEE, 2014.

[20] Andy Hodder. New technology, work and employment in the era of covid-19: reflecting on legacies of research. *New technology, work and employment*, 35(3):262–275, 2020.

[21] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[22] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[23] Min Hu, Dayong Zhang, Qiang Ji, and Lijian Wei. Macro factors and the realized volatility of commodities: a dynamic network analysis. *Resources Policy*, 68:101813, 2020.

[24] Michael I Jordan. Hierarchical models, nested models and completely random measures. *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*, pages 207–217, 2010.

[25] Charles Kemp and Joshua B Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

[26] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[27] Dénes Konig. Graphok és alkalmazásuk a determinánsok és a halmazok elméletére. *Mathematikai és Természettudományi Ertesito*, 34:104–119, 1916.

[28] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21, 2005.

[29] Yujie Lai and Yibo Hu. A study of systemic risk of global stock markets under covid-19 based on complex financial networks. *Physica A: Statistical Mechanics and its Applications*, 566:125613, 2021.

[30] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New journal of physics*, 11(3):033015, 2009.

[31] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[32] Peter V Marsden. The reliability of network density and composition measures. *Social Networks*, 15(4):399–421, 1993.

[33] Tiago P Peixoto. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807, 2015.

[34] Tiago P Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1):012317, 2017.

[35] Tiago P Peixoto. Nonparametric weighted stochastic block models. *Physical Review E*, 97(1):012306, 2018.

[36] Nicola Perra, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. Activity driven modeling of time varying networks. *Scientific reports*, 2(1):469, 2012.

[37] Francesco Pozzi, Tiziana Di Matteo, and Tomaso Aste. Spread of risk across financial markets: better to invest in the peripheries. *Scientific reports*, 3(1):1665, 2013.

[38] Leonard Richardson. Beautiful soup documentation, 2007.

[39] 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, Aug 2023.

[40] Daniele Schilirò. Towards digital globalization and the covid-19 challenge. 2020.

[41] Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983.

[42] Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983.

[43] Shunrong Shen, Haomiao Jiang, and Tongda Zhang. Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, pages 1–5, 2012.

[44] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on signal processing*, 65(13):3551–3582, 2017.

[45] Sara Nadiv Soffer and Alexei Vazquez. Network clustering coefficient without degree-correlation biases. *Physical Review E*, 71(5):057101, 2005.

[46] Sara Nadiv Soffer and Alexei Vazquez. Network clustering coefficient without degree-correlation biases. *Physical Review E*, 71(5):057101, 2005.

[47] Ms K Susmitha. Impact of covid 19 on e-commerce. *Journal of Interdisciplinary Cycle Research*, 12(9):1161–1165, 2021.

[48] Yee Whye Teh et al. Dirichlet process. *Encyclopedia of machine learning*, 1063:280–287, 2010.

[49] Steven Tenny and Mary R Hoffman. Relative risk. 2017.

[50] Robertus Van Nes. Design of multimodal transport networks: A hierarchical approach. 2004.

[51] Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: journal of the Econometric Society*, pages 307–333, 1989.

[52] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[53] Robert L Winkler. The assessment of prior distributions in bayesian analysis. *Journal of the American Statistical association*, 62(319):776–800, 1967.

[54] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. 2012.

[55] Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021.