Study of financial market dynamic and stability using "complex system's" approach and methodologies

Ngoc Kim Khanh Nguyen

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Pratique des Hautes Études

Étude de la dynamique et de la stabilité des marchés financiers à l'aide de méthodologies "systèmes complexes"

Soutenue par
Ngoc Kim Khanh NGUYEN
Le 30 septembre 2022

École doctorale n° 472
École doctorale de l'École Pratique des Hautes Études

Spécialité
Informatique, mathématique et applications

Composition du jury :

Mhand HIFI Professeur, Université d'Amiens	*Président*
Ahmed BOUNEKKAR Professeur, Université de Lyon	*Rapporteur*
Hi-Duc PHAM Professeur, ECE PARIS Ecole d'Ingenieurs	*Examinateur*
Soufian BENAMOR Maître de conférences, Université de Versailles Saint-Quentin	*Examinateur*
Marc BUI Professeur, École Pratique des Hautes Études	*Directeur de thèse*

École Pratique des Hautes Études | PSL

Acknowledgments

First and foremost, I would like to thank my advisor, Prof. Marc Bui. His enthusiastic supports and instructions are very helpful for me to adjust my works and capture my study results. Additionally, I sincerely thank him for giving me the opportunity to discuss online with other members in our laboratory in EPHE in the complicated context of the Covid-19 pandemic.

Secondly, I'm deeply grateful to Prof. Soufian Ben-Amor and Prof. Hi Duc Pham for their precious time and valuable comments in following my research process annually.

In addition, I would like to thank all of my friends in the EA 4004 Human and Artificial Cognition (CHArt) Laboratory, especially Ms. Thi Kim Thoa Ho, for their kindness, friendliness and the happy time I got when working with them in this laboratory.

Finally, special thanks go to Quang, my husband and also my colleague, who not only always encourages me through out the doctoral course but also collaborate with me on research. Also, I would like to thank my mother, Phuong, for her useful advises for my life as well as her care of my children when I was in France.

Hochiminh, 09/2021,
Ngoc Kim Khanh Nguyen

Résumé

Dans cette thèse, nous étudions les comportements collectifs des marchés boursiers, les primaires où se concentrent la plupart des ressources financières. Donné un bourse, pour comprendre ses caractéristiques de comportement collectif et mécanisme, nous analysons de manière approfondie le marché dans de nombreux aspects, y compris sa structure de réseau, sa résistance aux défaillances de composants, son facteur de marché déterminant principalement les rendements des avoirs sous-jacents, l'évolution de la défaillance en cascade et la dynamique de son indice représentatif. Étant donné que les marchés financiers peuvent être considérés comme des systèmes complexes, nous utilisons différentes techniques issues de la science complexe pour étudier les marchés boursiers dans de tels aspects, notamment la science des réseaux, la théorie des matrices aléatoires, la théorie de la prétopologie et l'analyse topologique des données.

Mots clés : marchés boursiers, réseaux complexes, théorie des matrices aléatoires, théorie de la prétopologie, analyse topologique des données

Abstract

In this thesis, we study the collective behaviors of stock markets, the primary ones where most of financial resources concentrate. Given a stock market, to understand its collective behavior's characteristics and mechanism, we comprehensively analyze the market in many aspects, including its network structure, its resilience under component fails, its market factor primarily driving the returns of the underlying holdings, the cascading failure's evolution, and its representative index's dynamics. Because financial markets can be considered as complex systems, we use different techniques employed from complex science to investigate stock markets in such aspects, including network analysis, random matrix theory, pretopology theory, and topological data analysis.

Keywords : stock markets, complex networks, random matrix theory, pretopology theory, topological data analysis

Contents

List of Figures

Acronyms

FSE *Frankfurt Stock Exchange*

HSX *Hochiminh Stock Exchange*

IB *initial betweenness centrality of nodes*

ID *initial degrees of nodes*

MCP *most correlated portfolio*

MST *minimum spanning tree*

NYSE *New York Stock Exchange*

PC *principal component*

PCA *principal component analysis*

RB *ecalculated betweenness centrality of nodes*

RD *recalculated degrees of nodes*

RMT *random matrix theory*

TDA *topological data analysis*

TMFG *triangulated maximally filtered graph*

WSE *Warsaw Stock Exchange*

List of Symbols

Notation	Meaning
$\langle . \rangle$	sample mean
$\|A\|$	number of elements of a set A
A'	transposition of a matrix A
$\|x - y\|$	a metric between x and y
N	number of stocks
T	length of time series
$S_i(t), i = \overline{1, N}$	price of stock i at time t
$r_i(t), i = \overline{1, N}$	log-returns of stock i at time t
$\mathbf{r} = (r_i)_{i=\overline{1,N}}$	random vector of log-returns of stocks
$\sigma_i, i = \overline{1, N}$	standard deviation of r_i
$\tilde{r}_i, i = \overline{1, N}$	normalized return of stock i
$\mathbf{r}^* = (r_i^*)_{i=\overline{1,N}}$	random vector of standardized stock returns
$\mathbf{C} = (c_{ij})_{i=\overline{1,N}, j=\overline{1,N}}$	cross-correlation matrix of stocks
$\mathbf{D} = (d_{ij})_{i=\overline{1,N}, j=\overline{1,N}}$	distance matrix of stocks
$P(k)$	degree distribution
q	fraction of removed nodes
q_c	critical threshold of q
η	allometric exponent
$\lambda_i, i = \overline{1, N}$	eigenvalues of the sample cross-correlation matrix
$\mathbf{u}_i, i = \overline{1, N}$	unit eigenvector associated with λ_i
$u_i^{(j)}, j = \overline{1, N}$	j-th component of eigenvector $\mathbf{u}_i$
z_i	i-th PC of $\mathbf{r}$
$a(.)$	pseudoclosure
$i(.)$	interior
$\mathbf{F}(A)$	closure of a set A
$\mathbf{O}(A)$	opening of a set A
E	set of all listed stocks
$\mathcal{P}(E)$	set of all subsets of a set E
H	set of failed stocks
i_0	initial stock of a cascading failure process
t_0	time of failure of i_0
τ	time-delay
d	embedding dimension
$\mathbf{y}_t^{\tau,d}$	reconstructed vectors with the time-delay τ and the embedding dimension d
$\mathbf{y}_{t^*}^{\tau,d}$	the nearest neighbor of $\mathbf{y}_t^{\tau,d}$
$\check{\text{C}}\text{ech}_\alpha(\mathbf{V})$	Čech complex spanned by a point cloud $\mathbf{V}$
$\text{Rips}_\alpha(\mathbf{V})$	Vietoris-Rips complex spanned by a point cloud $\mathbf{V}$
∂_k	k-boundary map
H_k	k-dimensional homology group of a simplicial complex $\mathbf{G}$

Introduction

The financial market plays an important role in any economy, and therefore all market players, including the banks, investors, regulators..., are very concern about its evolution and stability. A financial crisis is often followed by a long economic downturn which is always painful. For example, the recent financial crisis of 2007-2008 had a serious impact on the worldwide economy. Many questions have been asked:

- Why did it happen? How can we predict it?

- How was it spreading throughout the world financial systems?

- How can we improve the financial market stability?

- ...

From the modern economic theory, a financial market's collective behavior is neither deterministic nor converges to a fixed equilibrium point as suggested in classical theories of economics. Nevertheless, although the collective behavior is complicated, it is rational instead of random due to the strong relationships between the market's components and their abilities to learn and adapt to the change of the environment. Especially, it can be extremely different from the components' behaviors. Extreme behaviors of a financial market illustrate the non-linearity characteristic. The non-linearity is a complicated phenomenon in natural science such as physics and is perhaps even more important in the social sciences (which include finance). It is the consequence of the collective interactions between different sub-components of a system. Such systems are often called complex systems, and their study becomes very demanding recently. In order to answer the above questions, we need to consider financial markets as complex systems and study their non-linearity aspects.

In this thesis, we study the collective behaviors of stock markets, the primary ones where most financial resources concentrate on in economies. Given a stock market, knowledge about its mechanism and characteristics is essential to prevent dramatic recessions. The structure of this thesis includes five chapters performing the following contents:

Chapter 1 provides some basic concepts of complex systems and some common approaches used to study such systems. Besides, details of the view that considers financial markets as complex systems are also presented in this chapter.

In Chapter 2, we introduce the correlation-based network, which is often used to model the mutual interactions between a complex system's components. In our financial context, this network helps model the co-movement of stock prices in a stock market. With two special

subnetworks, the minimum spanning tree and the correlation-based threshold network, we can use graph theory's tools and the allometric scaling relation to study the market's geometrical structure and its resilience under random failures and intentional attacks. The result is important to get information about the market's stability as well as its robustness.

If we afraid of getting noises when calculating empirical cross-correlation matrices, *random matrix theory*, introduced in Chapter 3, helps find the "true" interaction between a stock market's components. It is often used to study the spectrum of an empirical cross-correlation matrix. According to this theory, the deviation of this matrix from the Wishart matrix gives information about the nature of the correlations. We especially focus on the largest eigenvalue, which is often significantly deviated from the theoretical spectral distribution, and its associated eigenvector having the unit module. To examine their roles in our financial problem, we use the method of *principal component analysis*. Because the first principal component of stock returns explains most of the stock returns' variances, it helps identify the market factor. We, therefore, provide not only further analyzes of the first principal component but also an estimate of its loadings in this chapter.

On the other hand, the collective behavior of a complex system is sometimes caused by a cascading failure, a process in which the failure of several components triggers the failure of their most correlated components and continues to spread to the entire system. This process is caused by the strong relationships between the components, not by any attacks on the system. To capture the cascading failure's evolution, we use pretopology theory introduced in Chapter 4. In this chapter, we propose a pretopological framework to model the diffusion of distress stocks in a stock market.

Finally, in Chapter 5, we investigate how to detect abnormal dynamics of a stock market's collective behavior. This question takes great interest from market managers, businesses to individual investors, especially after the world financial crisis 2007 – 2008. Although studying the dynamics of the network structure or the first principal component of the assets' fluctuations can solve this problem, we use another approach basing on the market's representative index because this data is transparent, continuously updated, and free. Even though a market index is not always replicate well the corresponding market's collective behavior because of the index's calculating method, the market's liquidity, etc. However, since the indexes are always constructed to be able to capture the variation of the corresponding stock markets as much as possible, it is deficient if we neglect the indices in studying the collective behaviors of such markets. Therefore, we solve the anomalies detection problem by recognizing any significant changes in the dynamics of market indices. To figure out important features of a market index's dynamics, we use *topological data analysis* combined with the time-delay embedding to get topological information of the dynamics' state space. The result is expected to give warnings about crises without analyzing a lot of micro and macro statistics.

In addition, in each chapter, we present empirical studies to examine our results in real markets. For more specific, in chapters 2 and 3, we carry out empirical studies in the U.S. stock market and the Vietnamese stock market to compare our results in two different cases - a developed market and an emerging market. Especially, we use the S&P 500 Index's components

and the VN Index's components to represent the two markets. The cause of our selection is that the S&P 500 Index, which includes the common stocks of 500 large-cap companies, is widely regarded as the best single gauge for U.S. equities since its components possess about 80% of the available market capitalization. Similarly, the VN Index, which includes all companies listed on the *Hochiminh Stock Exchange* (HSX), contains most large-cap companies in the Vietnamese stock market. This index's components also possess more than 80% of the market capitalization. However, in chapters 4 and 5, we only do empirical studies for the U.S. stock market. The reason is that the market's information about recessions and historical stock prices before mergers, acquisitions, bankruptcies, and removal decisions from the committee are more complete and transparent. We need the database to test the capability of our research methods presented in these chapters. All our empirical works, including data processing, modeling, analyzing statistical results, and plotting, are implemented using the R language.

Our result helps understand the characteristics of stock markets, and generally, financial markets, such as the geometrical structure, the phase transition, the robustness, the market factor's approximation, the cascading failure's evolution, and the markets' dynamics. These insights are important to get an overview of a stock market's dynamics and stability, and subsequently, help construct useful tools for managing the systemic risk.

Chapter 1

Introduction to Complex Systems

Objective

This chapter provides basic knowledge about complex systems and some common approaches of studying such systems in complex science. In addition, details of the scientific point of view that financial markets are complex systems are also discussed.

Contents

1 What Is a Complex System?

Complex science is a relatively new approach that helps study behaviors and properties of a large variety of systems from physical to social systems, called complex systems. Even though many aspects of this research topic has been studied for decades (dynamic system theory, chaos theory, self-organization, cybernetics, agent-based modeling, computational modeling, data-mining...), the definition of complex systems is still unconcise. Frequently, a complex system is supposed as a system whose large populations of units can self-organize into aggregations that generate patterns, store information, and engage in collective decision-making [Parrish, 1999]. More specifically, it mainly has the following features [Foote, 2007; Ladyman, 2013; McCarthy, 2000; Newman, 2011]:

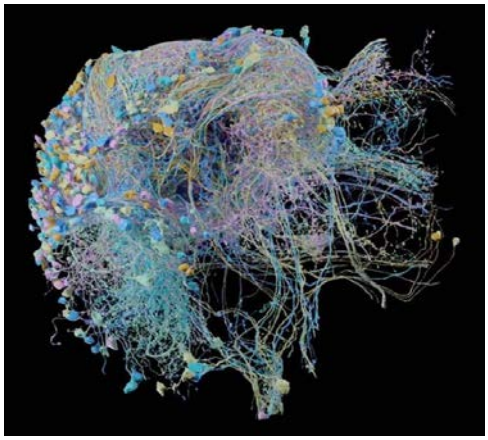
- Non-linearity: The system may respond in different ways to the same input depending on their current state or context, and small changes might have large effects in a nonlinear manner, while large ones could have little or no effect. This feature makes the system difficult to predict. For example, in chemistry, a single step in the multistep synthesis of a simple organic substance might include a lot of molecules of several types (each comprising many anharmonically oscillating bonds). The synthesis might proceed by different strategies for making and breaking bonds, and for generating the intermediate compounds that ultimately result in the final compound; each strategy might have a large number of possible variants differing in synthetic detail [Whitesides, 1999]. Sometimes, this feature makes complex systems confused with chaotic systems whose future state is also very sensitive to the initial state. In fact, complex systems are chaotic systems, but the opposite is generally not true.

- Adaption: The system's components constantly interact and change their behaviors in reaction to those of others and external conditions. For example, the course of each member of a flock of birds depends on the proximity and bearing of the birds around, but after one member adjusts its course, its neighbors also change their flight plans in response in part to its trajectory. Similarly, in technology, alterations in the maximum power of the engine of an automobile alter an optimal tire, suspension, and even highway design [Kauffman, 1995].

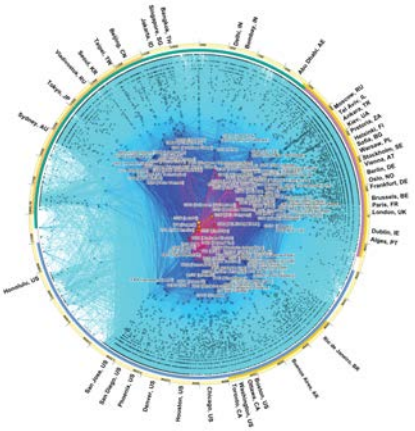
- Emergence: The interactions between the system's components and their responses to the environment can generate behaviors on the macro scale, which might be different from the local scale behaviors. In other words, the system's overall behavior may have an extreme level of magnitude, be qualitatively different from that of its parts, and usually not predictable. For example, the weather is an emergent property of air, moisture, and land interactions; global political dynamics are emergent from innumerable social, economic and political interactions; animal aggregations function as an integrated whole, displaying a complex set of behaviors not possible at the level of the individual organism such as the ability to build a nest or thermoregulate the hive of bees and termites [Parrish, 1999].

- Self-organization: The system that is formed and operates through many mutually adapting components is called self-organizing because no entity designs it or directly controls it. Due to the ability to self-organize and feedback mechanisms, the system will adapt autonomously to the environment's changes including changes imposed by policymakers. The best example of such a system is an ecosystem or the whole system of life on Earth.

Other reviews of researches in complex systems could be found in [Beinhocker, 2006; Kirman, 2011; Mitchell, 2011; Newman, 2003]. A few complex systems are known, such as weather systems, ecosystems, the brain, the immune system, flocking or schooling behavior in birds or fishes, condensed matter systems, the economy, financial markets, granular materials, road traffic, insect colonies, the Internet, social networks, transportation, and engineering infrastructure systems [Newman, 2011].



(a) A fly's brain including 25000 neurons and 20 million synapses rolling between them [1]



(b) The Internet IPv4 contained 47,610 autonomous systems and 148,455 links (according to the Feb 2017 Internet Topology Data Kit) [2]

Figure 1.1: Real complex systems.

2 Complex Science

The popularity and importance of complex systems in real life, along with the development of technologies and data sciences, provide opportunities for substantial recent advances in the study of these systems. It is an interdisciplinary domain that requires contributions from many diverse disciplines, including statistical physics, information theory, nonlinear dynamics, anthropology, computer science, meteorology, sociology, economics, psychology, and biology (Figure 1.2).

The studies of complex systems can be divided into two approaches. The first one includes studies of the systems' structure such as connectivity, rank-order correlation, clustering, structural change over time, huge systems whose sizes are unknown, visual representation... The second one includes studies of the systems' dynamic process such as forming and decomposing

[1] http://www.webmarketshop.com/25000-neurons-this-image-is-the-best-map-ever-made-of-a-brain/
[2] http://www.caida.org/research/topology/as_core_network/2017/

process, percolation process, containment control problem, phase transition, emergencies modeling, prediction... These two approaches can be combined and complement each other because the better a complex system's structure is understood, the more exactly its dynamic is described.

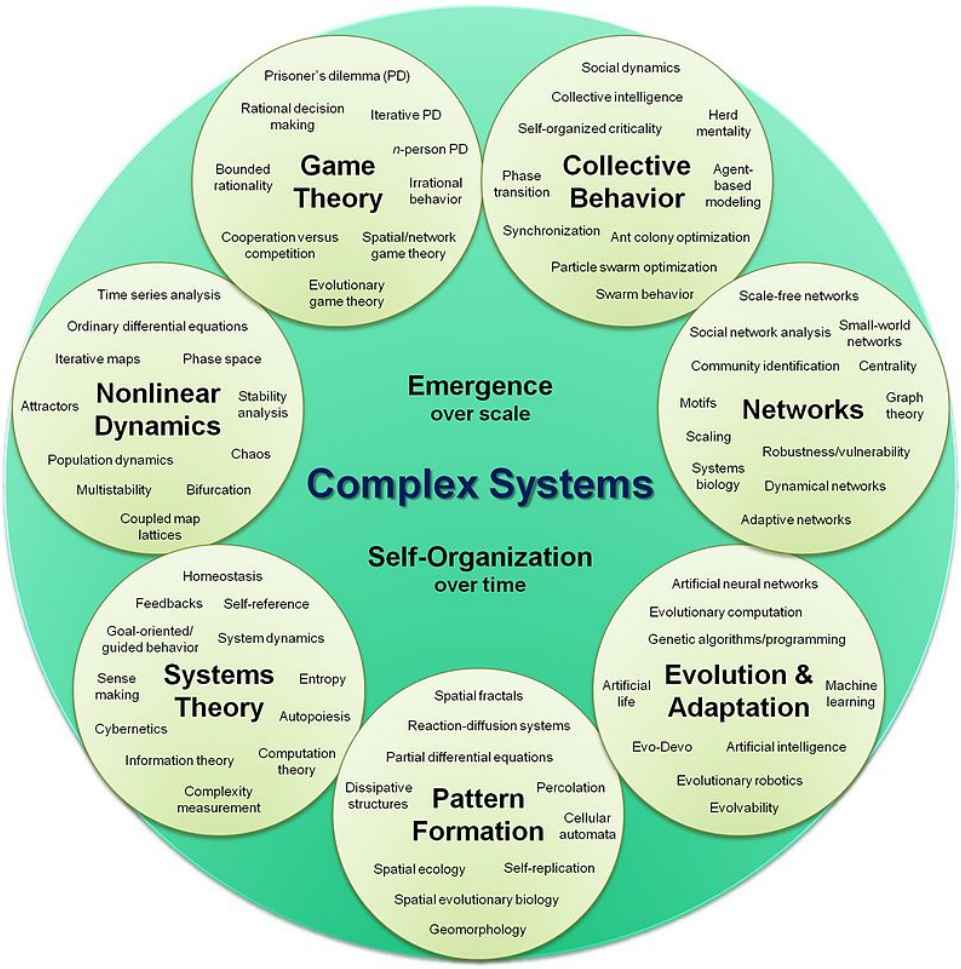


Figure 1.2: Complex science as an interdisciplinary domain.[3]

3 Fundamental Tools of Complex System Analysis

Because the collective behavior of a complex system cannot be determined by understanding the individual behaviors of the system's components, tools that help to capture not only the interaction between its parts but also its behavior as a whole are required. According to the review of Newman [Newman, 2011], to create and study simplified mathematical models abstracting the most important qualitative elements in a real complex system, there are some useful tools including dynamical systems theory, information theory, cellular automata, networks, computational complexity theory, and numerical methods. Meanwhile, for creating and studying more comprehensive and realistic models that represent the interacting parts or components of a complex system to observe and measure its emergent behaviors, agent-based simulation and Monte

[3] https://en.wikipedia.org/wiki/Complex_system

Carlo simulation are available tools.

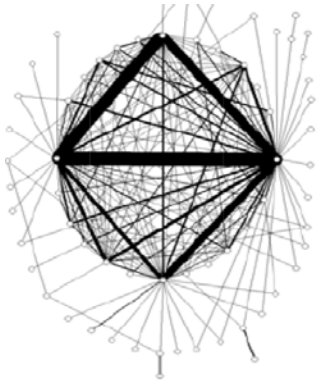
Among these fundamental tools, network and agent-based modeling are typical approaches that can be used to represent a complex system from its local scales to its global scale. More details about these approaches are given in the following paragraphs.

3.1 Agent-based Modeling

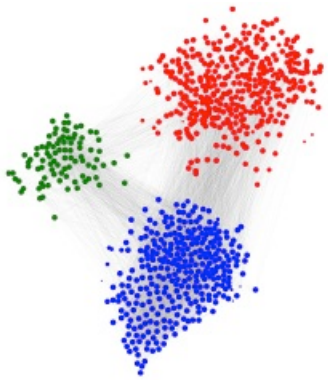
In order to describe the particular behavior of a complex system, for example, an emerging event as the "Black Monday" stock market fall in 1987, we need a new model that goes beyond the limits of the classical models with rational agents. In recent years, agent-based modeling with heterogeneous agents is developed to model the complex features of real systems. This is a "bottom-up" approach that separately and individually simulates the agents in a complex system and their interactions, allows the emergent behaviors of the system to appear naturally rather than puts them in by hand [Berry, 2002]. The simulation results will produce a virtual system that can be structured, or continue to be used to simulate the system's dynamic process. The agent-based modeling helps the researchers to conduct experiments, in terms of computer simulations, to test hypotheses, and to validate ideas and conjectures. This approach has become an important tool for understanding how real complex systems work, such as ecosystems [Grimm, 2005], social systems [Gilbert, 2008], the economy [Farmer, 2009]. Nevertheless, its disadvantage is the lack of supporting theories and models since it mainly depends on artificial intelligence and computer simulation. Hence, it is saved for our later researches.

3.2 Network Analysis

A complex system consists of many interacting components. So, a simplified representation of such a system can be a graph whose nodes represent the system's components, and each edge represents the interaction between two components (Figure 1.3). The graph is called a complex network. According to the research target, such networks can be directed or not. In the simplest form, one can assume that nodes are homogeneous, i.e., the system consists of components having the same natures, and an edge is defined between two nodes if they have any kind of interaction. Such a simplified representation has both advantages and disadvantages. One of its advantages is the support of many tools such as graph theory's tools to make researches on the structure and characteristics of the network. Another advantage is the universality because many diverse systems in physics, biology, engineering, economics, social science. . . , can be modeled as complex networks. Some of the most well-known complex networks are the Internet [Albert, 1999], the World Wide Web [Faloutsos, 1999], and metabolic networks [Jeong, 2000]. However, the huge number of a complex system's components, their multi-relations, as well as the heterogeneity of the components cause troubles in constructing their network representations.



(a) Interbank payment flow [Soramäki, 2007]



(b) Protein interaction network of the bacterium Borrelia hermsii HS1 [Martin, 2016]

Figure 1.3: Real complex systems represented by: (a) a weighted network, and (b) an unweighted network.

Another useful theory for complex network analysis is random matrix theory. It helps not only predict spectral properties of a complex network successfully but also understand the statistical properties of the empirical cross-correlation matrix computed by the multivariate time series of the system's components [Jalan, 2007]. For example, the empirical correlation matrix of price fluctuations in a stock market [Laloux, 2000; Plerou, 2002], EEG data of brain [Šeba, 2003], the variation of basic atmospheric parameters that characterize the state of the atmosphere [Santhanam, 2001]...

On the other hand, to overcome the weakness of the network representation, complex systems are also studied by using pretopology theory in recent years. This theory can be considered as an extension of graph theory because it allows processing multiple relations among a complex system's components or its parts [Belmandt, 2011]. Consequently, under the new approach, a complex system can be considered as a hypergraph, and each of the system's components can be modeled with its own nature. The most important tool of this theory is a map, called pseudoclosure, that helps to expand a certain set of the system's components by their relations (Figure 1.4). This map and minimal closed subset, another pretopological concept, can help model a proximity concept between a complex network's subsets based on different types of relations at the same time as well as examine the evolution of the network in each individual step. Therefore, pretopology theory can be used to study the dynamic structure of a complex network in many aspects; for example, formalizing the neighborhood concept to generalize the systems' percolation processes [Ben-Amor, 2006], modeling the systems' dynamic processes such as the information diffusion when combining with random set theory [Bui, 2019], understanding the structure and dynamics of web communities [Levorato, 2010]...

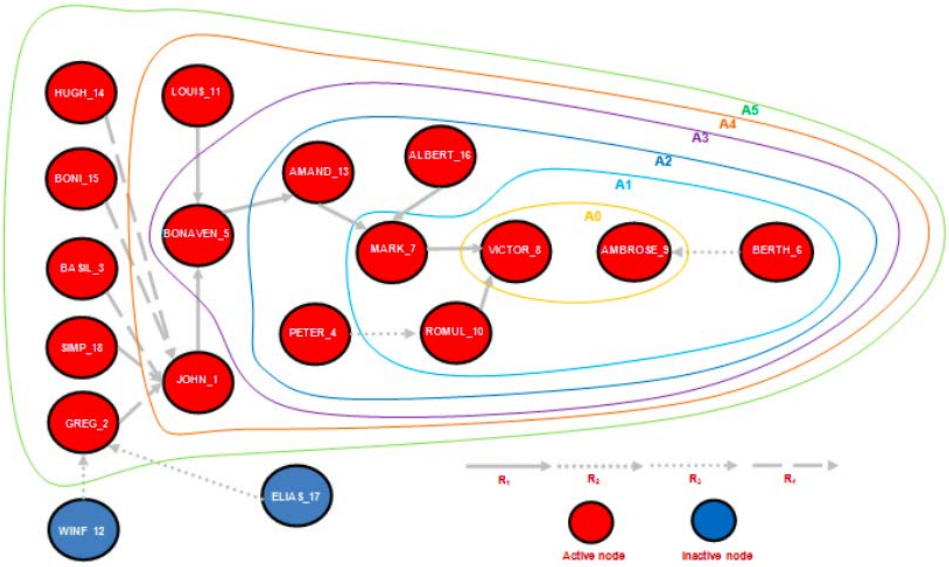


Figure 1.4: Pretopology Cascade Models on a network of 4 relations R_1, R_2, R_3, R_4: the result of the diffusion process originated from the set A0 of 2 nodes [Bui, 2019].

As a result, because of the complexity of a complex system, we should study it by different approaches to get a comprehensive of its structure and dynamics. Although there are many tools to carry out this task, in this work, we propose graph theory, random matrix theory, and pretopology theory as efficient approaches because these theories provide an extensive set of mathematical, computational, and statistical tools that can be used for analyzing, modeling, and understanding complex systems.

4 Financial Markets as Complex Systems

Before the mid-twentieth century, economists view the economy as an equilibrium system whose equilibrium point only changes under external forces. By contrast, modern economic theory considers the economy as a system that is dynamic and complex, and that moves from equilibrium point to equilibrium point over time, propelled along by shocks from technology, politics, changes in consumer tastes, and other external factors. In other words, the economy is a good example of complex systems. The literature of this perspective is mentioned in [Beinhocker, 2006; Kirman, 2011; Newman, 2011].

Similarly, from the scientific point of view, the financial market, which plays an important role in any economy, is also a complex system. In fact, a financial market is a collection of many constituents such as bonds, stocks, derivatives, currencies, banks, commodities that interact with each other and have the ability to learn and change behaviors from their experiences, for example, changes in security prices can affect each other and can be affected by negative information on saving and loans institutions. In addition, even though the market's collective behaviors are the result of its constituents' interactions as well as its response to the environment and external impacts, they can be far in order-of-magnitude and degree-of-complexity of the characteristics of its constituents. For example, the falling housing-related assets contributed to the collapse

of many of the United States' largest financial institutions and led to one of the greatest crisis in history in 2008 [Williams, 2010]; the market can fall into a long economic downturn after a policy of the government such as what happened in the stagflation in 1973 in the United States [Merrill, 2007]. Extreme behaviors of a financial system illustrate its non-linearity which makes its characteristic is too difficult to predict. However, similar to other complex systems, they are still expected to have real attractors rather than theoretically anticipated attractors [Lewin, 1994]. Therefore, considering financial markets as complex systems provides a new set of theories and techniques for understanding or explaining the mechanism and effects of economic phenomena such as phase transitions, the "fat-tail" price return distribution, volatility clustering phenomena, cascading failures, financial crises, dynamism rather than equilibrium... Especially after the worldwide financial crisis of 2008– 2009, rare-but-extreme volatile situations of financial markets has got more notices. More understanding of these special situations can help to improve the markets' stability, predict the worst-case scenarios, or evaluate potential policies.

Although a financial market contains different parts, including stock markets, bond markets, commodity markets, derivatives markets, futures markets, insurance markets, foreign exchange markets, and mortgage markets, but stock markets are our selection because of the two following reasons. Firstly, stock markets are the primary ones where most of the financial resources concentrate. Secondly, their historical data are always take down frequently and transparently. This is very important for our empirical studies, especially studies of developing markets like Vietnamese market.

Chapter 2

Financial Markets under Network Representations

Objective

In this chapter, we learn about the correlation-based network and its application in modeling the co-movement of stock prices in a stock market. Especially, we study its two special subnetworks, the minimum spanning tree and the correlation-based threshold network to get the most important information of the relationship between components of a stock market. Under this network representation, we can apply graph theory's tools and the allometric scaling relation to study the market's geometrical structure and its characteristic including its scale-free property, its resilience under failures and attacks, and its phase transitions in stress periods. Our results are examined on the Vietnamese and U.S. stock markets.

Contents

1 Correlation-based Networks in Financial Markets

As mentioned in Section 3 of Chapter 1, complex networks, the graph representations of complex systems are one of direct tools to model the relationships or interactions between such systems' components. This approach is applied in a large range of real-life systems, from biology to medicine, sociology, economics, and engineering [Mitchell, 2006; Jeong, 2000; Liljeros, 2001; Onnela, 2003b; Cohen, 2001]. Similarly, a financial system can be modeled by a network that is a collection of nodes (or vertices) and edges, where nodes represent the system's components and edges represent the relationships of nodes. For example, a network of corporations can be represented by a graph whose nodes are corporations, and each edge connects a pair of nodes if they have some common characteristic such as managers, investors, corporations. Also, a network of banks can be a weighted directed graph whose nodes are banks, and each weighted edge links a bank to another, where the weight is the value of the loan that the former write for the latter.

One of the popular network representations of financial systems is the correlation-based network [Bonanno, 2004]. This network helps infer the structure of cross-correlations among a set of time series. It means that, for a given complex system, we take out the time series of its components' behaviors. Then, two components are connected by an edge whose weight is a function of the correlation coefficient between the two corresponding time series. For example, a correlation-based network of futures contracts can be a complete network such that edges' weights depend on the correlation coefficients of pairs of contracts' price fluctuations [Lautier, 2013]. Also, a correlation-based network of indexes of worldwide stock exchanges can be obtained by the cross-correlations of the indexes' fluctuations [Bonanno, 2000]... For stock markets, such network is constructed such that nodes represent stocks while the correlation between two nodes is the correlation coefficient between the logarithm differences of the two corresponding stocks' prices as proposed by many works such as [Lux, 1999; Onnela, 2002; Plerou, 1999; Zheng, 2012]. More specifically, for N stocks $i = \overline{1,N}$, let $S_i(t)$ be the price of stock i at time t ($i = \overline{1,N}$), then:

Definition 2.1. *The $N \times N$ matrix $\mathbf{C} = (c_{ij})$ is called the cross-correlation matrix of the given stocks if*

$$c_{ij} = \frac{\langle r_i(t).r_j(t)\rangle - \langle r_i(t)\rangle.\langle r_j(t)\rangle}{\sigma_i \sigma_j}, \quad i,j = \overline{1,N} \tag{2.1}$$

where $r_i(t) = \ln(S_i(t)) - \ln(S_i(t-1))$ is the log-return of stock i at time t; $\langle . \rangle$ denotes the temporal average of the inside variable; σ_i and σ_j are the standard deviation of r_i and r_j, respectively. We call r_i the return of stock i and c_{ij} the correlation coefficient between stock i and stock j.

However, because the correlation coefficient of two stocks does not satisfy the three axioms of a metric's definition. Indeed, although it is nonnegative and symmetric, but it can miss the triangle condition. Therefore, a metric distance basing on the correlation coefficient is necessary to get a topological arrangement of the stock system. In this study, we use the distance discussed in [Gower, 1966]:

Definition 2.2. *The distance between stock i and stock j is defined by the following non-linear transformation of the correlation coefficient c_{ij} between these stocks:*

$$d_{ij} = \sqrt{2\left(1 - c_{ij}\right)} \tag{2.2}$$

The $N \times N$ matrix $\mathbf{D} = (d_{ij})$ is called the distance matrix.

Proposition 2.1.

(i) $0 \leq d_{ij} \leq 2$ *for all* i, j,

(ii) The set of N stocks associated with the distance measure given in Definition 2.2 is a metric space.

Proof.

(i) For all i, j, since $-1 \leq c_{ij} \leq 1$, it's clearly that $0 \leq d_{ij} \leq 2$.

(ii) Let consider the three axioms of being a metric:

- From (i) we already get that d_{ij} is nonnegative for all i, j.
  On the other hand, $d_{ij} = 0$ if and only if $c_{ij} = 1$. In addition, the latter happens if and only if $i = j$ in all empirical cross-correlation matrices of stock markets.
- The symmetry of d_{ij} is the result of the symmetry of c_{ij}.
- Let $\tilde{r}_i$ is the normalization of r_i, $\forall i = \overline{1, N}$, i.e.

$$\tilde{r}_i(t) = \frac{r_i(t) - \langle r_i(t) \rangle}{\sigma_i}, \ \forall t = \overline{1, T} \tag{2.3}$$

  Then, $\langle \tilde{r}_i(t) \rangle = 0$, $\left\langle \left(\tilde{r}_i(t)\right)^2 \right\rangle = 1$ and $c_{ij} = \langle \tilde{r}_i(t) . \tilde{r}_j(t) \rangle$ for all $i, j = \overline{1, N}$.
  Therefore, we obtain

$$\begin{aligned} d_{ij} &= \sqrt{2 - 2c_{ij}} = \sqrt{\left\langle \left(\tilde{r}_i(t)\right)^2 \right\rangle + \left\langle \left(\tilde{r}_j(t)\right)^2 \right\rangle - 2\left\langle \tilde{r}_i(t) . \tilde{r}_j(t) \right\rangle} \\ &= \sqrt{\left\langle \left(\tilde{r}_i(t) - \tilde{r}_j(t)\right)^2 \right\rangle} = \frac{1}{\sqrt{T}} \sqrt{\sum_{t=1}^{T} \left(\tilde{r}_i(t) - \tilde{r}_j(t)\right)^2} = \frac{1}{\sqrt{T}} \left\| \tilde{r}_i - \tilde{r}_j \right\| \end{aligned} \tag{2.4}$$

  where $\left\| \tilde{r}_i - \tilde{r}_j \right\|$ is the Euclidean distance of two vectors $\tilde{r}_i = \left(\tilde{r}_i(t)\right)_t$ and $\tilde{r}_j = \left(\tilde{r}_j(t)\right)_t$.
  As a result, for all i, j, k $(i, j, k = \overline{1, N})$,

$$d_{ik} + d_{kj} = \frac{1}{\sqrt{T}} \left(\left\| \tilde{r}_i - \tilde{r}_k \right\| + \left\| \tilde{r}_k - \tilde{r}_j \right\| \right) \geq \frac{1}{\sqrt{T}} \left\| \tilde{r}_i - \tilde{r}_j \right\| = d_{ij} \tag{2.5}$$

■

According to Definition 2.2, the more correlated two stocks are, the smaller their distance is. Consequently, the distance measure helps infer the topological arrangement of the stock market

through the level of synchronous evolution of stock returns. Therefore, the stock system can be represented by the following network:

Definition 2.3. *The correlation-based network of given stocks is the graph whose nodes represent stocks, and the adjacency matrix is the distance matrix constructed from the cross-correlation matrix of the stocks.*

In summary, the correlation-based network is computed by using Algorithm 1:

Algorithm 1 Compute the correlation-based network of stocks

Require: Time series of stock prices $(S_i(t))_{t=\overline{1,T}}$, $i = \overline{1,N}$

procedure STOCK_CORRELATION_BASED_NETWORK($\left\{ (S_i(t))_{t=\overline{1,T}} \middle| i = \overline{1,N} \right\}$)
  ▷ compute the stock returns
  for $i \in \overline{1,N}$ **do**
    $r_i(t) \leftarrow \ln(S_i(t)) - \ln(S_i(t-1))$ for all $t = \overline{2,T}$
  end for
  ▷ compute the empirical cross-correlation matrix
  for $(i,j) \in \overline{1,N} \times \overline{1,N}$ **do**
    $c_{ij} \leftarrow \frac{\langle r_i(t).r_j(t)\rangle - \langle r_i(t)\rangle.\langle r_j(t)\rangle}{\sigma_i \sigma_j}$
  end for
  ▷ compute the distance matrix
  for $(i,j) \in \overline{1,N} \times \overline{1,N}$ **do**
    $d_{ij} \leftarrow \sqrt{2(1-c_{ij})}$
  end for
  ▷ build the network's adjacency matrix
  $AdjacencyMatrix \leftarrow (d_{ij})_{i=\overline{1,N}, j=\overline{1,N}}$ ▷ Output
end procedure

The idea of transforming the cross-correlation matrix $\mathbf{C}$ into the distance matrix $\mathbf{D}$ was first introduced in [Mantegna, 1999] but in a different formula from Definition 2.2. However, since the distance in Definition 2.2 is an Euclidean distance, as demonstrated in Proposition 2.1, this measure is more convenient to reflect the topological and geometrical structure of the stock network. Also, because the correlation-based network of stocks is fully connected, it contains all possible co-movements of pairs of asset values and their strengths in the stock system. Therefore, such network is the subject of many studies about financial markets such as [Bonanno, 2004; Lautier, 2013; Mantegna, 2007; Onnela, 2003b]... In our researches, we also implement empirical studies about the correlation-based network of stocks (see [Nguyen, 2018; Nguyen, 2019b; Nguyen, 2019c]).

In the remainder of this thesis, we agree on the following points. Firstly, because we don't know the exact correlations between stocks, the notation "cross-correlation matrix" refers to the empirical cross-correlation matrix obtained from the historical data of assets. Secondly, we only pay attention to the daily fluctuation of stock prices, so, in all of the empirical examples below, except the ones referenced from other studies, the database is the daily closing prices of stocks. Finally, all networks or graphs discussed in the following statements of this proposal are undirected unless there's additional information.

2 Important Subgraphs of a Correlation-based Network

It is not easy to observe the topological structure of a correlation-based network or study its

dynamics. The main problem comes from the network's huge size. Indeed, because the network is complete, its number of nodes is N, the size of the underlying system, which is very large in most complex systems. Furthermore, its number of edges is $N(N-1)/2$. Therefore, in order to have a subgraph that contains enough important information of the relationship between the original network's nodes, we construct the following subgraphs: the *minimum spanning tree* (MST) of the network and the subgraph of highly connected nodes.

2.1 Minimum Spanning Tree

The MST of the correlation-based network is favored in many studies about financial markets. It is a concept of graph theory [West, 2001] and is defined as follows:

Definition 2.4. *A minimum spanning tree of a weighted network is a subgraph that is*

(i) connected, i.e., the subgraph contains all nodes of the original network, and there is a path to reach out from any node to another,

(ii) formed a tree, i.e., the subgraph doesn't have any node which loops back to itself, and

(iii) satisfied (i) and (ii) with the minimum total edge weight.

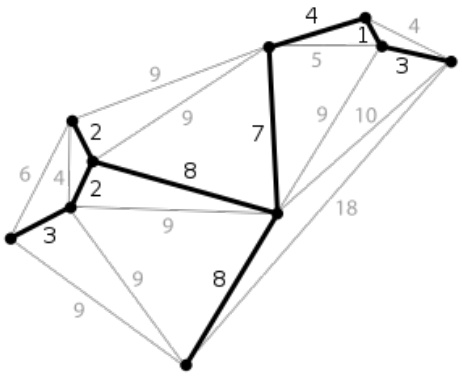


Figure 2.1: A network and its MST (bold).[1]

For constructing a MST of a given weighted network G containing N nodes, we can use a simple procedure called Kruskal's algorithm [West, 2001]. This algorithm is described as follows. At first, let M be a fully disconnected network associated with G, i.e., M includes all nodes of G but edges. Next, we order the set of edges of G based on their weights increasingly. By this order, we sequentially add each edge of G into M such that the additional edge doesn't create any cycle when combining with the edges of M. The adding process will stop when M has $N-1$ edges because, according to graph theory, this is the number of edges of a spanning tree if the corresponding network contains N nodes.

Moreover, if all edges' weights are mutually different, the network has only one MST. This condition is satisfied by many real correlation-based networks including the stock networks.

[1] https://en.wikipedia.org/wiki/Minimum_spanning_tree

Algorithm 2 Kruskal's algorithm for finding the MST of a weighted network G

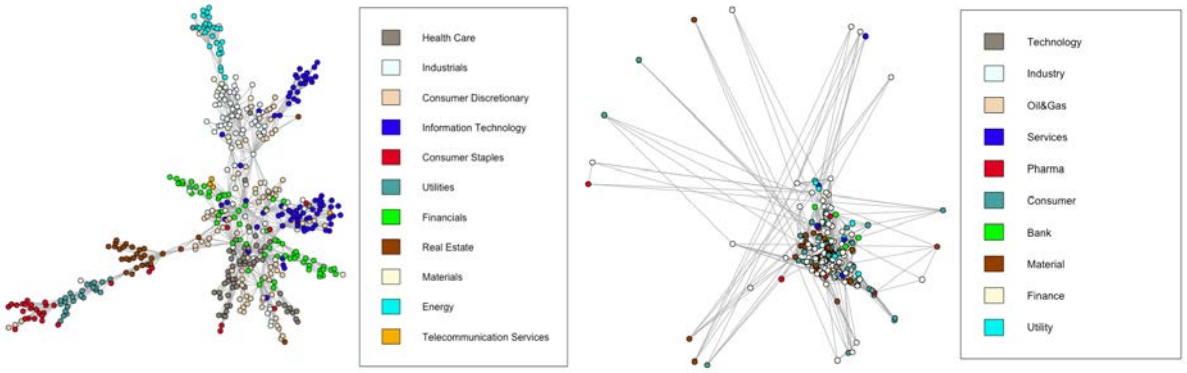
Require: Adjacency matrix $\mathbf{D} = (d_{ij})_{i=\overline{1,N},j=\overline{1,N}}$
procedure MST($\mathbf{D}$)
  ▷ set $M = (Nodes, Edges)$ as a fully disconnected network
  $Nodes \leftarrow \overline{1,N}$
  $Edges \leftarrow \varnothing$
  ▷ add edges by the increasing order of edge weights
  for $(i,j) \in \overline{1,N} \times \overline{1,N}$ ordered by increasing d_{ij} **do**
    If $(edge(i,j) \notin Edges)$ **and** $(\{k | \{edge(i,k), edge(k,j)\} \subset Edges\} = \varnothing)$ **then**
      $Edges \leftarrow Edges \cup \{edge(i,j)$ with weight $d_{ij}\}$
  end for
  $M \leftarrow (Nodes, Edges)$ ▷ Output
end procedure

Theoretically, the correlation c of two certain stocks can be any real number in the interval $[-1, 1]$, so the probability of a specific value of c is zero. In reality, because we must consider a time window that is long enough when calculating the empirical cross-correlation matrix to prevent short-time noises, the chance to get two pairs of stocks having the same correlation is null. Consequently, the MST of a correlation-based network of stocks is practically unique.

When using the MST of the original correlation-based network to represent a financial market, there are many advantages. The first one is the MST's simplicity because it has only N nodes and $N-1$ edges. In addition, since the MST of a correlation-based network corresponds to the shortest path covering all nodes of the original network without loops, it is expected to help extract the most important information contained in the network. Indeed, one of significant information is the indexed hierarchical tree associated with the MST which exhibits a meaningful economic taxonomy [Bonanno, 2004; Mantegna, 1999; Onnela, 2003c]. Furthermore, the MST is the most probable path of a price shock's propagation in the corresponding network [Lautier, 2013]. The reason is clearly that the MST prefers edges whose weights d_{ij} are smaller than others, except when meeting loops back, so it prioritizes edges corresponding to strong stock correlations c_{ij}. It means that the MST models the fastest path that a price shock can spread over the whole network. Consequently, the structure of the MST can provide meaningful information about the properties of a market such as its clustering, stability, different market's states... Therefore, studying the MST's structure and its dynamics in real financial networks becomes an attractive research in recent decades. More details are given in Section 4.

However, the MST of a correlation-based network of stocks has a considerable weakness: some edges associated with small weights, i.e., high stock correlations, may not belong to the tree. This weakness is the result of the acyclic condition. Because of this disadvantage, although the MST can provide an overall taxonomy of the market, the connections it creates may be misinterpreted to be more meaningful than they are [Onnela, 2004]. So, other subgraphs are proposed to filter information in the original complex network such that the corresponding system's clusters are well-defined such as the average linkage minimum spanning tree [Tumminello, 2007], the planar maximally filtered graph [Tumminello, 2005], the directed bubble hierarchical tree [Song, 2011; Song, 2012], and the *triangulated maximally filtered graph* (TMFG) [Massara, 2017]. Empirically, these subgraphs are proved to be able to model communities of economic sectors and sub-sectors in a stock network slightly better than the MST does in developed markets such as the *New York Stock Exchange* (NYSE) [Tumminello, 2005; Tumminello, 2007]. Nevertheless, in

emerging markets, we recognize that these subgraphs might not well clarify the stock network's partition. The reason is that the correlations between the listed stocks and some outstanding stocks, such as stocks of financial companies (brokerage companies, banks...), can be higher than the intra-correlations of economic sectors [Nguyen, 2019b; Nguyen, 2019c]. This remark is clearly illustrated in Figure 2.2. This figure shows the TMFG of the correlation-based network of 486 common stocks of large companies comprised in the S&P 500 Index and the TMFG of the correlation-based network of all stocks listed on the HSX (262 stocks). The filter graphs are constructed from the daily closing prices of stocks listed on the market from 04/01/2015 to 04/01/2020. It is easy to see that the clusters of the Vietnamese stock network are not well-defined clearly as in the U.S. stock network whose clusters associate to economic sectors.



(a) Stocks comprised in the S&P 500 Index (b) Stocks listed on the HSX

Figure 2.2: The TMFG of the correlation-based network of stocks listed from 04/01/2015 to 04/01/2020.

In summary, we propose that the classical MST is a subgraph that is simple but efficient enough to infer important information of stock markets, especially, emerging markets where the market clusters are not well-defined. Moreover, understanding the MST's structure is really important for managing the systemic risk because the MST is the most probable path that makes the transmission of a price shock spread throughout the market. Further investigation about using the MST for risk management in financial systems can be found in the review [Marti, 2021].

2.2 Correlation-based Threshold Network

Although the MST helps get an overview about the synchronic price fluctuation's capability of constituents in a stock network, it can sophisticate the network's robustness under errors of the elements. So, the network seems more fragile than it is under the MST representation. We can explain this problem as follows. The MST might include edges associated with very small correlation coefficients while neglecting other edges linking highly correlated stocks. This happens because the MST must contain all nodes of the original network to be a spanning tree. Thus, another interesting subgraph of the correlation-based network is the correlation-based threshold network. It means that we only keep stocks and edges associated with enough high

stock correlations. Works of [Garas, 2008; Onnela, 2003b; Onnela, 2004] are good examples of this approach. In those studies, the authors first order the edge weights d_{ij} of their correlation-based networks increasingly. Then, according to this arrangement and starting with an empty subgraph, they respectively add stocks and edges to the subgraph until its number of edges is $N-1$. They call it the asset graph. So, the asset graph has the same number of edges as the MST. This similarity helps compare the MST with the asset graph. In general, the asset graph is just a correlation-based threshold network because we can replace the condition about the number of edges by a lower bound for the correlation coefficients c_{ij}. This subgraph is computed by Algorithm 3.

Algorithm 3 Compute the correlation-based threshold network

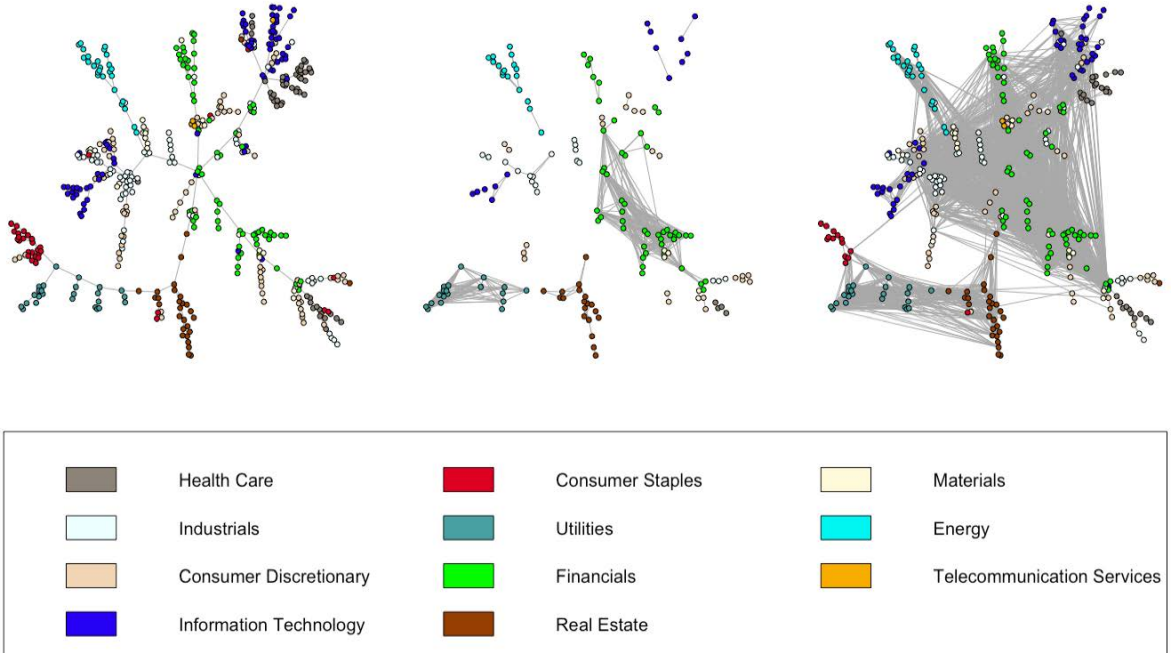
Require: Cross-correlation matrix $\mathbf{C} = (c_{ij})_{\overline{1,N}\times\overline{1,N}}$, threshold c_0
procedure THRESHOLD_NETWORK($\mathbf{C}, c_0$)
  ▷ set $M = (Nodes, Edges)$ as an empty graph
  $Nodes \leftarrow \varnothing$
  $Edges \leftarrow \varnothing$
  ▷ Add edges and nodes associated with high correlation coefficients
  for $(i,j) \in \overline{1,N} \times \overline{1,N}$ **do**
    If $c_{ij} > c_0$ **then**
      $Nodes \leftarrow Nodes \cup \{i,j\}$
      $d_{ij} \leftarrow \sqrt{2\,(1-c_{ij})}$
      $Edges \leftarrow Edges \cup$ {edge (i,j) with weight d_{ij} }
    end if
  end for
  $M \leftarrow (Nodes, Edges)$ ▷ Output
end procedure

Although the asset graph seems reflect the stock network's partition associated with outstanding economic sectors better than the MST [Onnela, 2003b; Onnela, 2004], the number of nodes in the asset graph is extremely smaller than the one in the original network. For example, let's consider the correlation-based network constructed from the same data as the network in Figure 2.2a, we show its MST and its asset graph in Figure 2.3a and Figure 2.3b, respectively. As mentioned above, the asset graph in Figure 2.3b can be considered as the graph of stocks corresponding to correlation coefficients that are higher or equal to 0.78. This value is really high in stock markets. Therefore, although the number of nodes of the original network is 486, the asset graph includes only 167 nodes, approximately 34.36% of the former. Because the asset graph lacks a considerable amount of stocks, it only represents the market's communities, whose elements tightly correlate to each other, but it misses other information about the entire market. In fact, as we can see in Figure 2.3b, the asset graph well models the intra-correlations of some sectors such as financials, utilities, real estate and energy but mostly neglects other sectors.

As a result, when constructing the correlation-based threshold network of stocks, in order to not miss market information due to neglecting too many nodes, a suitable threshold for the stock correlations is very important. The selected threshold must help reduce the size of the original correlation-based network by removing unimportant connections but still build a representative graph for the market. Especially, when analyzing a stock system's characteristics, the correlation-based threshold network helps avoid noises caused by unstable connections. Figure 2.3c shows the graph corresponding to the threshold 0.63 for the correlation coefficients of stocks with the same database as Figure 2.3a and 2.3b. This value approximates the 97%-quantile of

the empirical stock correlations. The graph has 349 nodes, approximating 71.81% of the number of nodes of the original correlation-based network. The graph's number of edges is 3300. It means that the original network can be figured well by keeping a small number of its edges (only 3% in this example) such that the selected edges corresponding to the most important connections. In addition, we propose that the threshold for the stock correlations should be selected variously in different markets. For example, in [Nguyen, 2018], a study about the Vietnamese stock market, the 97%-quantile of the empirical stock correlations is only 0.25 since the stock correlations in an emerging market are usually smaller than the ones in a developed market. Thus, in that work, we chose the threshold of 0.25.

In general, for the MST of a correlation-based network of stocks, we have a simple graph spreading over the network by the shortest path to study the network's overall structure and the shock prices' propagation problem. Meanwhile, for the correlation-based threshold network of stocks with a suitable threshold, we have a graph that is more efficient to study the robustness of the network.



(a) The MST (b) The asset graph (c) The correlation-based threshold network where the threshold is the 97%-quantile of the empirical stock correlations.

Figure 2.3: Some subgraphs of the correlation-based network of stocks listed on the NYSE from 04/01/2015 to 04/01/2020.

3 Structural Measures of Financial Networks

In order to analyze the structure and characteristics of the correlation-based network of stocks and its subgraphs, we use different measures provided by graph theory such as the degree

distribution, the average shortest-path length, the betweenness centrality, the giant component's size, and the allometric scaling. We also study the change of the structure over time by considering the dynamics of these characteristics as well as the dynamics of the single-step survival ratio and the same sector ratio. Besides, because the correlation-based network of stocks that we study is undirected, the below concepts are introduced in case of undirected graphs only as mentioned in previous section.

3.1 Degree Distribution

A simple but powerful tool to measure a network's structure is the node degrees defined as follows:

Definition 2.5. *In a network, the degree of a node is the number of edges connected to it.*

The degree of a node helps measure the node's level of connectivity (Figure 2.4). Consequently, in the representative network of a financial system, a node with a high degree plays an important role in the system's connectivity.

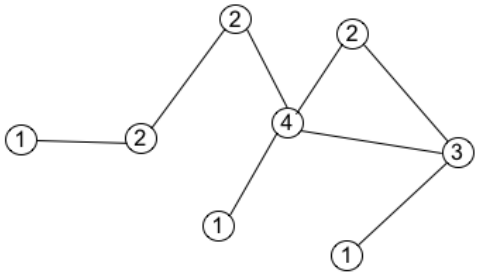


Figure 2.4: A graph having nodes labeled by degree.

Definition 2.6. *In a network, let $P(k)$ be the fraction of the number of nodes with degree k. A histogram of $P(k)$ is called the degree distribution of the network.*

Equivalently, we can define $P(k)$ as the probability that a node in the network has degree of k.

In a random graph, where each edge presents with the same probability, the degree distribution is binomial or Poisson if the graph's size is too large. However, in real-world complex networks, their degree distributions are almost far from a Poisson distribution because of their long right tail [Newman, 2003]. Therefore, a study about a financial network's degree distribution is really necessary to understand the network's structure constructed from its elements' relationship as well as its complexity.

3.2 Average Shortest-path Length

In a correlation-based network, to evaluate the effect level of the network's nodes to each other, or more generally, to measure the ability that information spreads between two nodes, we need to compute the shortest path length between every pair of nodes. In graph theory, we have the following definitions:

Definition 2.7. *A path connecting a pair of nodes in a network is a sequence of edges which joins the two nodes. The length of the path is the total weight of the edges belonging to the path if the network is weighted, and equals the number of these edges if the network is unweighted.*

Definition 2.8. *The average shortest-path length of a network is the average length of shortest paths for all possible node pairs in a network, i.e.,*

$$L = \frac{\sum_{(i,j)} l(i,j)}{N(N-1)} \tag{2.6}$$

where N is the number of nodes and $l(i,j)$ is the shortest path length from node i to node j.

The shortest path connecting two nodes in a correlation-based network of stocks can be considered as the most probable path that the stocks affects each other. Meanwhile, the average shortest-path length gives an expected distance between two randomly chosen nodes. Thus, it is an intuitive characterization of how sensitive the current market is under a shock. That the reason why this measure is an important factor to examine the network's stability.

3.3 Betweenness Centrality

In graph theory, betweenness centrality is a measure of a network's centrality based on the fraction of shortest paths that go through each node. Its definition is given below:

Definition 2.9. *The betweenness centrality of node i is given by:*

$$b\,(i) = \sum_{j \neq i \neq k} \frac{s^i_{jk}}{s_{jk}} \tag{2.7}$$

where s_{jk} is the number of the shortest paths connecting node j and node k, and s^i_{jk} is the number of those paths that pass through node i (not where i is an endpoint).

Clearly, the node with the highest betweenness centrality is the node that connects "regions" of dense nodes. In other words, the betweenness centrality of a node helps decide the node's importance in the percolation problem rather than concentrate on the node's neighborhood only as its degree. Indeed, let's see Figure 2.5. We can see that a higher node degree does not imply a higher node betweenness. Besides, let's assume that we attack the network in this figure by removing a node. If the node's betweenness is highest, the network then breaks into pieces of which the largest size is 4. However, if the node's degree is highest, the largest size of the pieces is 5. As a result, by removing the node with the highest betweenness, we break the network's connectivity better.

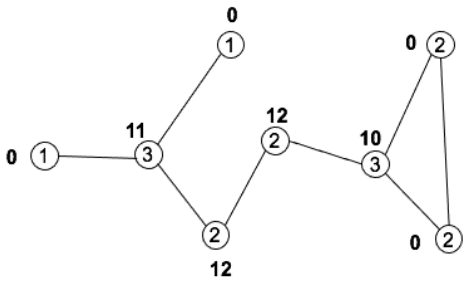


Figure 2.5: Differences between node degree (given inside the circles) and the number of shortest paths going through each node (given outside the circles).

Consequently, nodes with high betweenness play important roles in transaction-based networks such as the banking network. In case of the correlation-based networks of financial systems, this measure helps find the components that play significant roles in making a downtrend broaden to other asset values in a recession or make the systems broken when the components cannot maintain their functions. This information is really useful in managing the systemic risk.

3.4 Giant Component

In reality, with a significant number of damaged nodes, many complex networks are unable to keep their normal operations. When a node is damaged, the node and its links are deleted from the original network. Therefore, percolation theory, a theory studies the behavior of a network when nodes or links are removed, is important to help characterize a network's robustness and fragility. For example, in a population system whose each potential host for a disease is represented by a node, a node is occupied if the corresponding host is susceptible to the disease; then, the percolation theory helps get more knowledge about the disease's contagion to avoid a pandemic. Similarly, the vaccination to get the immunity of a community, the information's propagation in a social system or a communication system, the financial shock's spreading in a financial system, etc., are other real percolation problems.

A key prediction of percolation theory is that the decomposition of a network under node removal is not a gradual process with the fraction q of removed nodes. With a wide range of q, the network may still keep its normal operation but its integrity changes when q is larger than a critical threshold q_c. To identify the threshold, we can depend on the existence of the network's giant component after the node removal. In percolation theory, this terminology is defined as follows:

Definition 2.10. *A giant component or giant cluster is a connected cluster of a network that contains a significant proportion of the entire nodes in the network even when the network's size increases.*

Remind that, in graph theory, a cluster or component of a network is a subgraph that there is a path between every pair of nodes, but no node in the cluster can have an edge to another cluster. Therefore, according to Definition 2.10, a node in a giant component can be reachable to so many nodes of the current network even when the network's size changes. Typically, the

giant component of a network is understood loosely as the biggest cluster (see Figure 2.6). In all cases, the component must be the cluster that is much bigger than others.

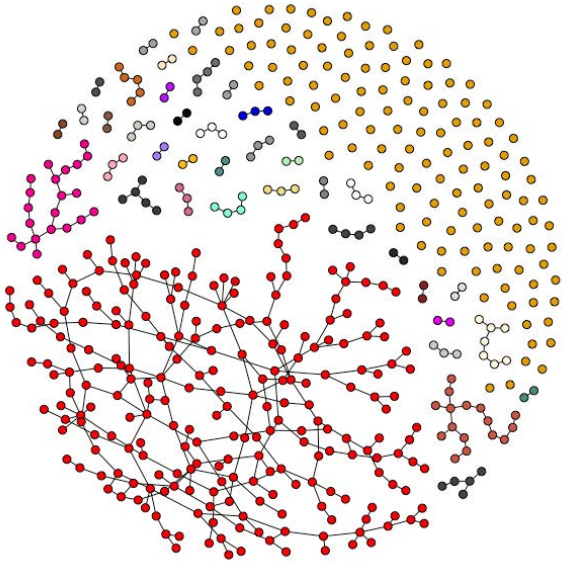


Figure 2.6: Giant component (red) of a random network.

Because after deleting a fraction q of nodes from a network, if the network fragments into many significantly small clusters, its global connectivity will break up. So, the critical threshold q_c can be considered as the value such that the giant component is destroyed when q goes over. The following theorem provides the Molloy-Reed criterion [Cohen, 2000; Molley, 1995], a popular criterion for the existence of a giant component in a random uncorrelated network, i.e., the network whose degrees of all nodes are independent, random integers drawn from a specified distribution $P(k)$.

Theorem 2.1. *In a random uncorrelated network with degree distribution $P(k)$, a giant component exists if*

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} \geq 2 \tag{2.8}$$

where $\langle k \rangle$ and $\langle k^2 \rangle$ are the first and the second moment of $P(k)$.

Proof. Let's consider a random uncorrelated network with an arbitrary degree distribution $P(k)$. If the network has a giant component and if loops of connected nodes are neglected, the percolation transition takes place when a node i, connected to a node j in the component, also links to at least another node. Otherwise, the component is fragmented. Therefore, the degree of node i cannot be less than 2. Since the component must take up a large proportion of the network, for a percolation transition to take place, the average degree of a node i in the network must be at least 2 given that it is connected to another node j, i.e.,

$$\langle k_i | \, i \leftrightarrow j \rangle = \sum_{k_i} k_i P \left(k_i | \, i \leftrightarrow j \right) \geq 2 \tag{2.9}$$

where k_i is the degree of node i, and $P(k_i | i \leftrightarrow j)$ is the conditional probability that node i has degree k_i, given that it is connected to node j.

Use Bayes' rule, we obtain

$$P(k_i | i \leftrightarrow j) = \frac{P(k_i, i \leftrightarrow j)}{P(i \leftrightarrow j)} = \frac{P(i \leftrightarrow j | k_i) P(k_i)}{P(i \leftrightarrow j)} \tag{2.10}$$

where $P(k_i, i \leftrightarrow j)$ is the joint probability that node i has degree k_i and that it is connected to node j. Without degree correlations and loops, because of the fact that we can choose between $N-1$ nodes to link to, each with probability $1/(N-1)$ where N is the number of nodes of the network, and that we can try this k_i times, we get:

$$P(i \leftrightarrow j | k_i) = \frac{k_i}{N-1} \tag{2.11}$$

and

$$P(i \leftrightarrow j) = \frac{\langle k_i \rangle}{N-1} \tag{2.12}$$

Substitute (2.11) and (2.12) into (2.10), we obtain

$$P(k_i | i \leftrightarrow j) = \frac{k_i . P(k_i)}{\langle k_i \rangle} \tag{2.13}$$

Consequently, the inequation (2.9) is rewritten as:

$$\sum_{k_i} \frac{k_i^2 . P(k_i)}{\langle k_i \rangle} \geq 2 \tag{2.14}$$

The left-hand side of (2.14) is κ, so the theorem is valid. ■

Using the Molloy-Reed criterion, Cohen et al. [Cohen, 2000] shown the relation between q_c and κ as follows:

Theorem 2.2. *In a random uncorrelated network, we obtain:*

$$1 - q_c = \frac{1}{\kappa_0 - 1} \tag{2.15}$$

where $\kappa_0 = \frac{\langle k_0^2 \rangle}{\langle k_0 \rangle}$ *is computed from the initial distribution before the random breakdown.*

Proof. Let $P(k)$ be the initial degree distribution of a random uncorrelated network. After randomly removing a fraction q of the nodes, each node appears in the updated network with probability $1-q$, independently with other nodes. Therefore, for a node with initial degree k_0, the distribution of its number of connectivity must follow the binomial distribution with parameters k_0 and $1-q$, i.e., the probability that the new degree of the node is k ($k \leq k_0$) equals $\binom{k_0}{k}(1-q)^k q^{k_0-k}$. Consequently, the new degree distribution is

$$P_{new}(k) = \sum_{k_0=k}^{\infty} P(k_0) \binom{k_0}{k} (1-q)^k q^{k_0-k} \tag{2.16}$$

Without loss of generality, we can assume that the smallest possible connectivity is 1. From (2.16), we can identify the expected value of the new degree distribution:

$$\begin{aligned}\langle k\rangle_{new} &= \sum_{k=1}^{\infty} kP_{new}(k) = \sum_{k_0=1}^{\infty}\sum_{k=1}^{k_0} kP\left(k_0\right)\binom{k_0}{k}(1-q)^k\, q^{k_0-k} \\ &= \sum_{k_0=1}^{\infty} P\left(k_0\right)\sum_{k=1}^{k_0} k\binom{k_0}{k}(1-q)^k\, q^{k_0-k}\end{aligned} \tag{2.17}$$

The inside sum is the expected value of a random variable drawn from the binomial distribution with parameters k_0 and $1-q$, so the sum equals $k_0(1-q)$. Replace this result into (2.17), we obtain:

$$\langle k\rangle_{new} = \sum_{k_0=1}^{\infty} P\left(k_0\right) k_0(1-q) = (1-q)\left\langle k_0\right\rangle \tag{2.18}$$

Similarly, we can compute the second moment of the new degree distribution as follows:

$$\left\langle k^2\right\rangle_{new} = \sum_{k=1}^{\infty} k^2 P_{new}(k) = \sum_{k_0=1}^{\infty} P\left(k_0\right)\sum_{k=1}^{k_0} k^2\binom{k_0}{k}(1-q)^k\, q^{k_0-k} \tag{2.19}$$

The inside sum is the second moment of a random variable drawn by the binomial distribution with parameters k_0 and $1-q$, so the sum equals $k_0\left(1-q\right)q + k_0^2\left(1-q\right)^2$. Replacing this result into (2.19), we obtain:

$$\left\langle k^2\right\rangle_{new} = \sum_{k_0=1}^{\infty} P\left(k_0\right)\left[k_0\left(1-q\right)q + k_0^2\left(1-q\right)^2\right] = (1-q)\,q\left\langle k_0\right\rangle + (1-q)^2\left\langle k_0^2\right\rangle \tag{2.20}$$

At the critical threshold q_c, because the network closes its giant component when q goes over q_c, according to the Molloy-Reed criterion (2.8) and equations (2.18), (2.20), we get

$$\kappa = \frac{\left\langle k^2\right\rangle_{new}}{\langle k\rangle_{new}} = q_c + (1-q_c)\frac{\left\langle k_0^2\right\rangle}{\left\langle k_0\right\rangle} = q_c + (1-q_c)\kappa_0 = 2 \tag{2.21}$$

Hence, we obtain (2.15). ■

Theorem 2.2 helps measure the robustness of a network under random failures of nodes. However, when we have more information about the network's structure, we can break the network more efficiently. Our study on the resilience of a stock network under both random failures and intentional attacks is provided in Section 4.

3.5 Allometric Scaling Relation

Besides classical measures of graph theory, we also focus on another specific factor that is helpful to measure the hierarchical degree of a network's structure: the allometric scaling relation. This relation, which takes the form of a power law $C = A^{\eta}$, is often used to model a large number of relationships between size and rate in a biological or physical process, where η is a constant to specify the relationship. In fact, in biology, C can be body mass and A can

be the biological property of interest, for example, rates of resource used in individual plants scale as about the 3/4 power of body mass, which is the same as metabolic rates of animals. Consequently, the above relationship can link to the geometrical and topological properties of a distribution network sustaining the supply for metabolic activity [McMahon, 1983; Schmidt-Nielsen, 1984; West, 1997]. Similarly, this relation can use to consider the general structure of branching networks (without loops) serving a particular volume in inanimate systems, for example, in the drainage network of river basins, A stands for the total water flow coming from the sub-basin area around each node and C stands for the total water flow that goes through this node through the drainage direction [Banavar, 1999; Rodriguez-Iturbe, 2001]. Also, the allometric scaling laws are the subject in other complex networks in different fields such as the food webs [Garlaschelli, 2003], the world trade webs [Duan, 2007], the world investment networks [Song, 2009], and so forth.

In economics, using allometric scaling relation to characterize the financial network complexity is a novel idea and, according to our knowledge, only a few works have been done, for instance, [Duan, 2007; Lautier, 2012; Lautier, 2013; Qian, 2010; Song, 2009]. Qian et al. [Qian, 2010] analyzed the MST of the visibility graph constructed from the time series of 30 worldwide stock market indices where each data point is a node and an edge is drawn to connect two nodes according to the rule that the two corresponding data points can see each other in the diagram of the time series. Meanwhile, Lautier et al. [Lautier, 2013] analyzed the MST constructed from future contracts in 14 derivatives markets, which is a subset of a larger graph of 250 future contracts with different maturities [Lautier, 2012].

The original model of the allometric scaling on a spanning tree was developed by Banavar et al. [Banavar, 1999]. For studying the structural property of the tree through the allometric scaling relation, we must firstly assign a direction for each edge if the tree is not directed. The rule is that the edges connecting a node and the hub with the highest degree must reach out from the hub. Other edges must reach out from the node that connect to the hub with a less number of edges (see Figure 2.7). We temporarily call the result of this direction assignment as the directed spanning tree. Then, the allometric scaling relation is picked out by the power-law relation between two variables A and C computed for each node of the network. These variables are found in an iterative manner as follows [Qian, 2010]:

Definition 2.11. *For each node i in the directed spanning tree, let*

$$A_i = \sum_j A_j + 1, \qquad C_i = \sum_j C_j + A_i, \tag{2.22}$$

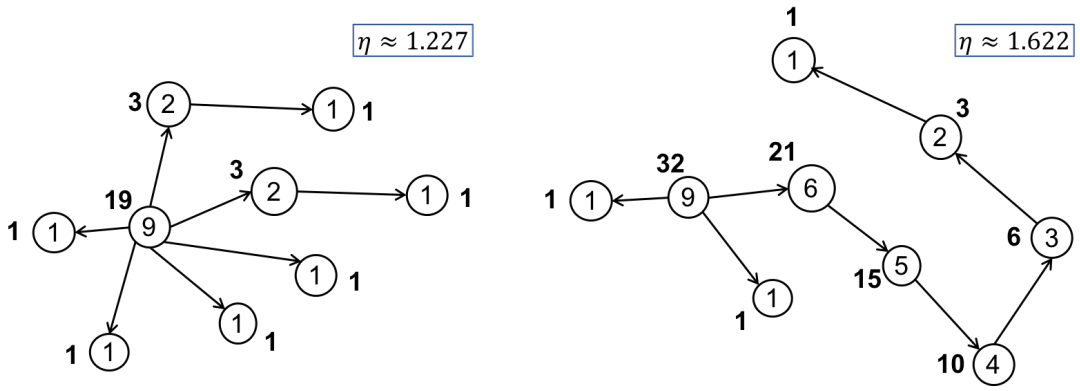
where j stands for all nodes linked from node i. Then, the allometric exponent η is the fitting power of the following expression:

$$C \sim A^{\eta} \tag{2.23}$$

where the leaf nodes with $A = C = 1$ have to be ejected from fitting the exponent.

The allometric exponent η represents the complexity of the MST and lies between 1 and 2 for two extreme network structures: star network and chain network [Garlaschelli, 2003; Qian,

2010]. This is demonstrated as follows. In a star network, all nodes (except the hub) connect to only one single center-node which is the hub. So, all nodes (except the hub) are leaf nodes and are ejected when fitting the allometric scaling exponent. Meanwhile, for the hub, we have $A = N$ and $C = 2N - 1$ where N is the number of nodes. Hence, according to (2.23), we get $\eta = 1$ when $N \to \infty$. By contrast, in a chain network where all nodes are linked one after another, there's only one leaf node which is the final in the tree in the direction starting from the root. Then, according to (2.22), the value C of a node whose $A = k$ is $k(k+1)/2$. So, when $N \to \infty$, we get the fitting exponent η must be 2. Consequently, the allometric exponent must satisfy $1 < \eta < 2$ where $\eta = 1^+$ for a star-like trees and $\eta = 2^-$ for a chain-like trees. For example, in Figure 2.7, the tree on the left looks like a star structure more than the rest. This topological information is reflected well through the allometric exponent: $\eta \approx 1.227$ for the tree on the left and $\eta \approx 1.622$ for the rest. Clearly, the allometric exponent of the tree on the left is closer to 1.



(a) Star-like structure

(b) Chain-like structure

Figure 2.7: Allometric scaling computation where A is inside the nodes and C is nearby the nodes.

Besides, for the direction originated from the largest hub, according to Definition 2.11, we can see that variable A of a node i stands for the total number of nodes that can be reached from i (including itself). Therefore, variable C of node i can measure the total impact of the node toward the network through its k-nearest neighbors, where the closeness level k goes to infinity.

Moreover, it is empirically demonstrated that the allometric scaling relation really appears in the MST network of a stock system. In Figure 2.8, we give some examples of well-fitting equation (2.23) to the nodes of three MSTs associated with the HSX in different periods: 03/31/2009 – 10/19/2010, 05/16/2012 – 12/02/2013 and 01/14/2014 – 08/18/2019. The figure shows the log-log plot of the relation between two variables A and C of each node. The number of nodes of the MSTs are 140, 232, and 249, respectively, for each tree.

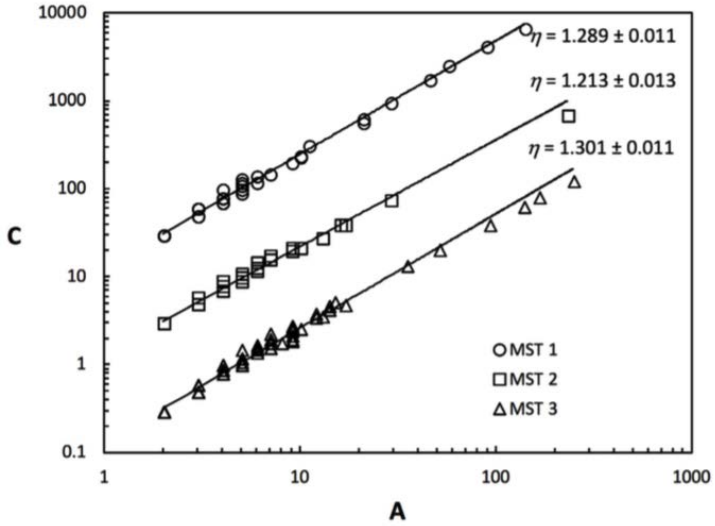


Figure 2.8: The allometric scaling behaviors of MSTs representing the Vietnamese stock system in three periods: 03/31/2009 – 10/19/2010, 05/16/2012 – 12/02/2013 and 01/14/2014 – 08/18/2019.

As a result, the allometric scaling relation of the MST associated with a financial system can help quantify the global "shape" of the system and determine the influence level of each constituent on others in the system. Hence, this relation plays an important role in studying the system's stability.

3.6 Survival Ratio

When studying a complex network, an outstanding question is how the network's structure changes over time. To measure its structure's stability, we simply measure the number of common edges found in the graphs representing the network in two consecutive periods. After diving the result by the number of edges in the graph of the later period, we get a measure call single-step survival ratio, or survival ratio for short, introduced in [Garas, 2008; Onnela, 2003b; Onnela, 2003c]. However, in our study, we replace the denominator by the average number of edges of the two consecutive graphs. Our reason is that the size of a stock network often increases over time because more companies join in the market in general. Consequently, without our adjustment, the survival ratio becomes smaller because of the denominator's increase. The ratio's decrease does not usually relate to the changes in the network's connectivity because most new comers rarely have ability to change the old connectivity. For more specific, we define this measure as follows:

Definition 2.12. *Let G_t and G_{t-1} be two consecutive graphs representing a complex network and E_t and E_{t-1} be the set of edges of G_t and G_{t-1}, respectively. Then, the survival ratio between G_t and G_{t-1} is defined by the following expression:*

$$S\left(G_t, G_{t-1}\right) = \frac{2\left\|E_t \cap E_{t-1}\right\|}{\|E_t\| + \|E_{t-1}\|} \tag{2.24}$$

where $\|.\|$ denotes the size of the inside set, i.e. the number of the set's elements.

Proposition 2.2. *The survival ratio between two consecutive graphs G_t and G_{t-1} ranges between 0 and 1. It equals 1 if and only if E_t and E_{t-1} are the same, and equals 0 if and only if the two graphs have no common edge.*

Proof. Clearly, $S(G_t, G_{t-1}) \geq 0$, according to formula (2.24). In addition, for any t, since $\|E_t \cap E_{t-1}\| \leq \|E_t\|$ and $\|E_t \cap E_{t-1}\| \leq \|E_{t-1}\|$, we have $2\|E_t \cap E_{t-1}\| \leq \|E_t\| + \|E_{t-1}\|$, which implies $S(G_t, G_{t-1}) \leq 1$.

Besides, we have $S(G_t, G_{t-1}) = 1$ if and only if $\|E_t \cap E_{t-1}\| = \|E_t\| = \|E_{t-1}\|$, i.e., E_t and E_{t-1} are the same.

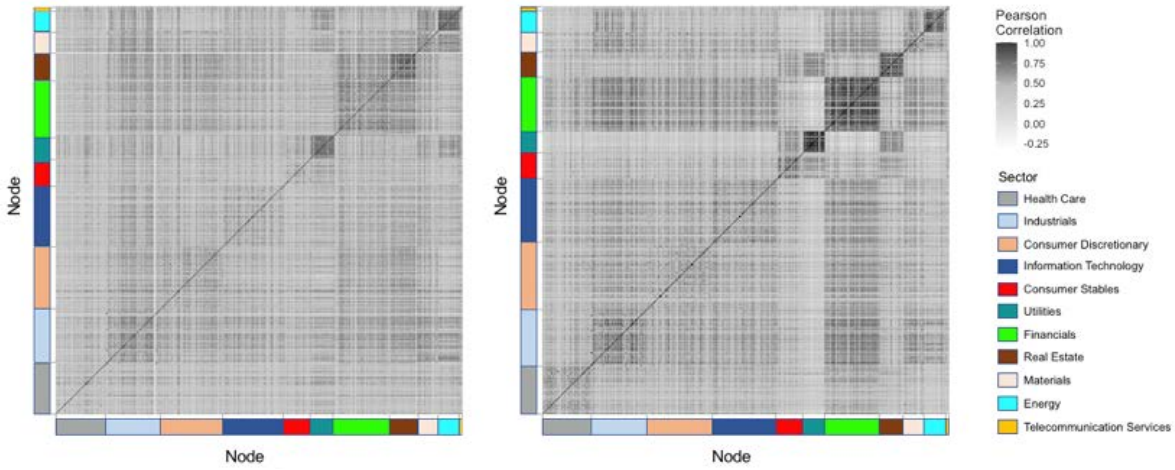
On the other hand, $S(G_t, G_{t-1}) = 0$ if and only if $\|E_t \cap E_{t-1}\| = 0$, i.e. $E_t \cap E_{t-1} = \varnothing$. ■

3.7 Same Sector Ratio

For a stock network, to study the network's structure and its dynamics, we also pay attention to another measure – the same sector ratio.

Definition 2.13. *The same sector ratio of the stock network is the fraction of the number of edges that connect two stocks belonging to the same business sector.*

Especially when using the MST to represent the correlation-based network of stocks, this ratio also plays an important role in studying the stability of the tree's structure because of the following observation: stocks in the same business sector are generally more correlated than stocks in different sectors, except some special sector such as the financial sector; however, in crises, stock prices become more sensitive with many factors that can come from many other sectors or external factors of the market. Consequently, in crises, the intra-correlation of a sector is not almost higher than the inter-correlation. Therefore, this observation shows that significant changes of the same sector ratio of the MST can give useful information about crucial changes of the MST's structure. Figure 2.9 visualizes the empirical cross-correlation matrix of stocks comprised in the S&P 500 Index in two periods corresponding to the two states of the market: the stress period and the normal period. The former is the time when the Great Recession 2007 – 2008 happened.



(a) Data in the stress period (01/01/2007 - 01/01/2009) (b) Data in the normal period (01/01/2015 - 01/01/2017)

Figure 2.9: Heatmap of the empirical cross-correlation matrix of stocks listed on the NYSE by business sectors.

In brief, with many tools introduced in this section, we can have a various analysis about the structure of the correlation-based networks of stock markets. The empirical results when we study real stock networks are given in the next section.

4 Characteristics of Stock Networks

Let's remind that to represent a stock system, instead of using its correlation-based network, we can use the MST of the network or the correlation-based threshold network. Which one is better? The answer depends on our research purpose. In this section, we focus on the network's stability and robustness. While the MST network is more suitable to study the network's stability, the correlation-based threshold networks is often used to study the second problem. Both of these networks of a stock system have an essential characteristic, the scale-free property. This is the common property of many complex systems.

4.1 Scale-free Property

Most real complex networks have a common property called the scale-free property, although they can be constructed from objects of different natures [Newman, 2003]. In addition, networks that are scale-free have a number of intriguing properties. Therefore, such property is of our particular interest.

Definition 2.14. *A scale-free network is a network whose degree distribution follows a power law, i.e.,*

$$P(k) \sim k^{-\gamma} \tag{2.25}$$

The positive constant γ is called the degree exponent of the distribution.

The term "scale-free" comes from the fact, in many real scale-free networks, when the network's size goes to infinity, the second and higher moments of the degree distribution also go to infinity (see [Barabási, 2016] for more details).

Now, let's take a logarithm of (2.25), we obtain

$$\log P(k) \sim -\gamma \log k \tag{2.26}$$

So, for a scale-free network, $\log P(k)$ is expected to depend linearly on $\log k$ where the slope of the line is $-\gamma$.

Here, we just highlight some properties that will be relevant for our study about stock networks:

Remark. A scale-free network with $\gamma > 1$ has the following characteristics:

(i) It could have central nodes with extremely high degrees (often called "hubs").

(ii) The largest hub's degree grows with the network's size.

(iii) Comparing with random networks having the same expected value, it lacks of internal scale.

To demonstrate the first and second characteristics, because real networks are finite, let $k_{\min}$ and $k_{\max}$ be the lower and upper cutoffs for the node degree of a network, respectively. If the network is scale-free, we can approximate the distribution (2.25) to a continuum which exacts for $1 \ll k_{\min} \ll k_{\max}$, and preserves the essential features of the discrete distribution even for small $k_{\min}$. Then, for the degree exponent $\gamma > 1$, due to the normalization that

$$\int_{k_{\min}}^{\infty} P(k)\, dk = \int_{k_{\min}}^{\infty} ck^{-\gamma} dk = 1 \tag{2.27}$$

we obtain

$$c = (\gamma - 1) k_{\min}^{\gamma - 1} \tag{2.28}$$

To calculate $k_{\max}$, we assume that the probability to have a node degree that is greater than $k_{\max}$ is $1/N$ where N is the network size, i.e.,

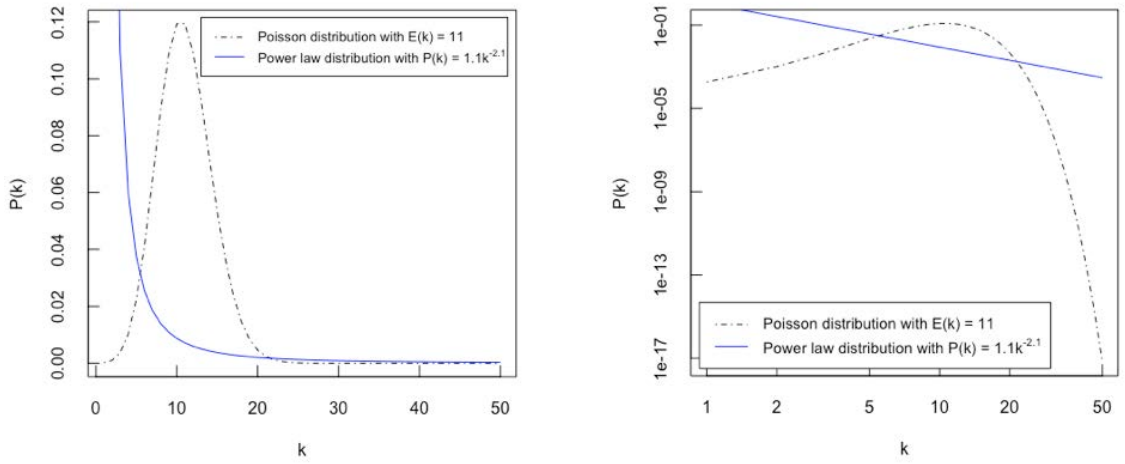
$$\int_{k_{\max}}^{\infty} P(k)\, dk = \int_{k_{\max}}^{\infty} ck^{-\gamma} dk = \frac{1}{N} \tag{2.29}$$

If $\gamma > 1$, from (2.28) and (2.29), we get

$$k_{\max} = N^{\frac{1}{\gamma - 1}} k_{\min} \tag{2.30}$$

Therefore, the largest hub's degree can be extremely large with the growth of the network size. It implies that the first characteristic is also valid. This makes the right tails of such networks' the degree distributions fatter than the ones of random networks whose degree distributions follow Poisson distribution if their sizes are large. The difference is illustrated in Figure 2.10 where we shows the probability density functions of two distributions having the same expected value: the

Poisson distribution with parameter $\lambda = 11$ and the power law distribution $P(k) = 1.1k^{-2.1}$. The figure particularly points out the exceedingly large probability of nodes with small degrees. It also shows that although having the same expected value, the power law distribution admits a higher probability of central nodes than the one of the Poisson distribution.



(a) Normal plot

(b) log-log plot

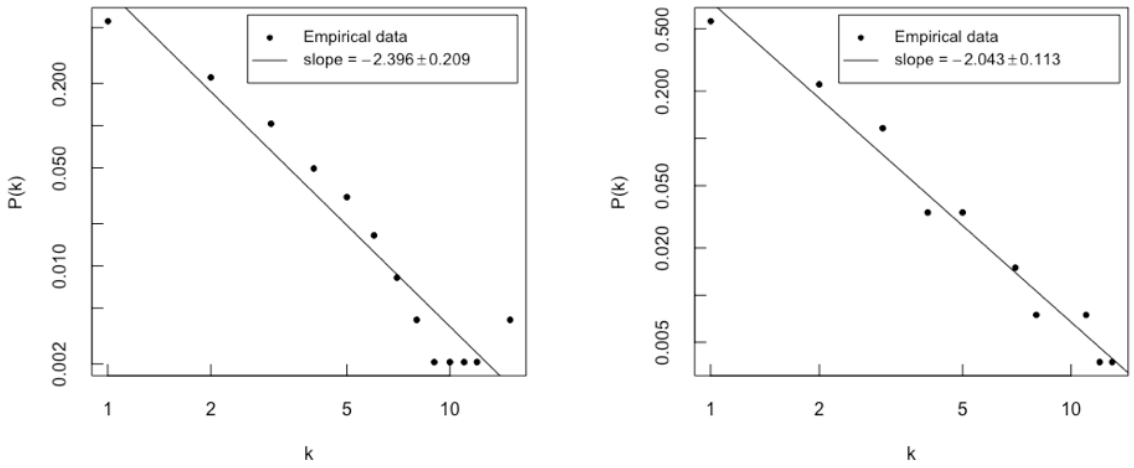
Figure 2.10: Poisson distribution vs. power law distribution.

Also, it is easy to indicate the third characteristic if the first is right. Indeed, according to Definition 2.14, the probability of having a node with small degree is very large. Then, if the first's right, the network's node degrees are widely diverse with a few of large degrees. Consequently, when we randomly choose a node in the scale-free network, the node's degree could be very far from the average degree and tends to bias to the average's left side. This fact is quite different from a random network whose node degrees vary in a narrow range (Figure 2.10). Therefore, the scale-free network lacks of internal scale more than random networks having the same expected value.

Real complex networks generally satisfy these three characteristics of a scale-free network. So, their empirical degree distributions almost follows a power law distribution with the degree exponent mostly ranges from 1.8 to 3.2 [Barabási, 2016]. The network of film actors [Amaral, 2000; Watts, 1998], the network of sexual contacts [Liljeros, 2003; Liljeros, 2001], the network of word co-occurrence [Cancho, 2001; Dorogovtsev, 2001], Internet [Chen, 2002; Faloutsos, 1999], peer-to-peer network [Adamic, 2001; Ripeanu, 2002] and metabolic network [Jeong, 2000] are some examples of scale-free networks.

Similarly, when consider the MST of the correlation-based network as the representative network for a financial system, there are some hubs with very high connections in the tree. The hubs represent the components having high correlations with others. In stock market, they could be common stocks of large corporations in business sectors or stocks of principal financial organizations. Empirical researches demonstrate that the stock networks constructed by the

MST method are almost scale-free. In fact, without crises, the MST of the correlation-based networks of stocks listed on many exchanges such as the NYSE [Onnela, 2003b; Vandewalle, 2001], the Athen stock exchange [Garas, 2007], the *Warsaw Stock Exchange* (WSE) [Sienkiewicz, 2013] and the *Frankfurt Stock Exchange* (FSE) [Wiliński, 2013] are some examples of scale-free networks of stocks. In [Nguyen, 2019c], we also found a similar result on the Vietnamese stock market, an emerging market, when analyzing the MST of the correlation-based network of stocks listed on the HSX. Figure 2.11 provides two examples of scale-free networks of stocks modeled by the MST method in a developed market and an emerging market. The figure displays the log-log plot of the networks' degree distribution, where our database is the closing prices of stocks comprised in the S&P 500 Index and stocks listed on the HSX from 01/01/2017 to 01/01/2019. The empirical distributions on both of these markets are fitted well with the power law distribution with the R square of 0.92 for the U.S. stocks and 0.98 for the Vietnamese stocks, approximately. The degree exponent approximates 2.40 for the U.S. market and 2.04 for the Vietnamese market.

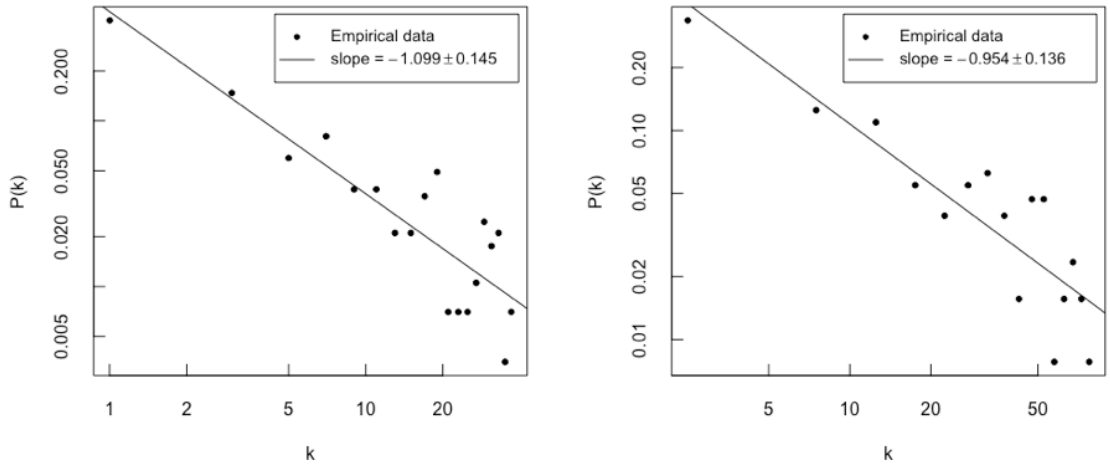


(a) Networks of stocks comprised in the S&P 500 Index

(b) Networks of stocks listed on the HSX

Figure 2.11: Degree distribution of stocks networks modeled by the MST method in the period 01/01/2017 – 01/01/2019.

In addition, we also study the correlation-based threshold network of stocks. As discussed in Section 2, the network constructed by a suitable threshold of stock correlations can represent the corresponding market. Moreover, with a large enough threshold, the network likely has the scale-free property, even though that is not as clear as the MST network. In fact, when constructing stock networks by keeping high correlations which are larger than 0.63 for the U.S. stocks and 0.25 for the Vietnamese stocks, these networks' degree distributions as histograms of $P(k)$ likely fit a power law distribution as shown in Figure 2.12. The database used to construct the figure is the same as the database used in Figure 2.11. We presented a similar result in [Nguyen, 2018].



(a) Degree distribution of the network of stocks comprised in the S&P 500 Index with the thresh- old of 0.63

(b) Degree distribution of the network of stocks listed on the HSX with the threshold of 0.2

Figure 2.12: Degree distributions of the correlation-based threshold network of stocks on the U.S. market and on the Vietnamese market in the period 01/01/2017 – 01/01/2019.

In general, both representative networks of a stock system that we study, the MST of the correlation-based network and the correlation-based threshold network, are almost scale-free, especially, the former. Furthermore, the emerging market's degree exponent is always smaller than the developed market's one, so the former's structure is denser than the latter's one. On the other hand, the property of scale-free networks that we're most interested in is the presence of hubs as well as their roles in the networks' stability and robustness. Especially, in financial distressing periods, due to the unpredicted behaviors of a stock system's components as well as the appearance of its collective behaviors, which can be significantly different from the components' behaviors, these networks might lose the scale-free property. All of these issues will be discussed in the next sections.

4.2 Network Resilience

For many real complex systems, errors or failures of a few components can make the systems hard to operate normally; for instance, the failure of a part in a car's engine, a wiring error in a computer chip. By contrast, many other systems still operate well if some of their components fail from an accident. For example, the Internet still functions if some routers are down somewhere, the economy still runs if some corporations file for bankruptcy due to some management mistakes. In this section, using a graph representation of a financial system, we will quantify the system's ability to keep its operation despite the faults of some components. The ability of a network to provide and maintain an acceptable level of service in the face of faults and challenges to its normal operation is known as *network resilience*. In the literature, the level of network resilience is a good deal subject in many fields, such as the network of actors collaboration and

science citations [Gallos, 2006], the Internet, the Word-Wide Web [Albert, 2000; Cohen, 2001], the metabolic networks [Jeong, 2001], food webs [Dunne, 2002]... Although some works study the resilience of a network by the increase of the mean distance of node pairs (see more details in the review of Newman [Newman, 2003]), we quantify the level of network resilience as the fraction of node removal such that the network still keeps its global connectivity as discussed in [Callaway, 2000; Cohen, 2000; Molley, 1995]. Obviously, if a node is removed from the network, the related links are also deleted, and the network's average short-test path length must increase consequently. So, the two approaches are similar.

In economics, the vulnerability of a financial network under improper operations of some of its parts becomes an important subject due to the requirement of systemic risk management to prevent financial crises. In this section, we measure the level of a stock network's resilience under two types of node removals: failures and attacks. A network under failures of nodes means that an arbitrary part of its nodes is damaged. By contrast, a network is under attack if some of its nodes are damaged intentionally. In this case, the most probable damaged nodes are the most important ones in the network. Frequently, a considerable attack strategy is removing the most highly connected nodes since they often damage the integrity of the network most. Besides, we also focus on another attack strategy – damaging nodes with the largest betweenness centrality. This strategy is expected to rapidly fragment the network into many pieces because the nodes help connect regions of dense nodes.

As discussed in Section 3, the level of a network's resilience is computed by the critical threshold q_c, the largest fraction of removed nodes such that the network's giant component is undestroyed. Besides, according to Theorem 2.2, q_c of a network under random breakdown can be calculated directly from the ratio of second-to first-moment of the network's degree distribution, κ. We're especially interested in the value of κ in scale-free networks, the type of network corresponding to our financial networks. Basing on κ, Cohen et al. [Cohen, 2000] theoretically found the relation between the resilience and degree exponent γ of a scale-free network under a random removal of its nodes for $1 < \gamma < 3$ and $\gamma \neq 2$. We briefly present their result with additional information for the case $\gamma = 2$:

Theorem 2.3. *In a large scale-free network with degree exponent γ, under a random removal of its nodes,*

- *for $\gamma > 3$, the critical threshold q_c approximates $1 - \left(\frac{2-\gamma}{3-\gamma} k_{\min} - 1\right)^{-1}$, where $k_{\min}$ are the smallest possible connectivity.*
- *for $1 < \gamma < 3$, the critical threshold q_c approximates 1.*

Proof. Let $P(k) = ck^{-\gamma}$ $\left(k = \overline{k_{\min}, k_{\max}}, \gamma > 1\right)$ be the degree distribution of the network and $k_{\max}$ be the largest possible connectivity. Using a continuum approximation which is valid in the limit $1 \ll k_{\min} \ll k_{\max}$ and preserves the essential features of the discrete distribution even for small $k_{\min}$, for $\gamma \neq 2$ and $\gamma \neq 3$, we get:

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{\int_{k_{\min}}^{k_{\max}} k^2 P(k) dk}{\int_{k_{\min}}^{k_{\max}} k P(k) dk} = \frac{\int_{k_{\min}}^{k_{\max}} c k^{2-\gamma} dk}{\int_{k_{\min}}^{k_{\max}} c k^{1-\gamma} dk} = \frac{2-\gamma}{3-\gamma} \cdot \frac{k_{\max}^{3-\gamma} - k_{\min}^{3-\gamma}}{k_{\max}^{2-\gamma} - k_{\min}^{2-\gamma}} \tag{2.31}$$

Let N be the network's size. According to equation (2.30), we can express $k_{\max}$ in the above equation in term of N and $k_{\min}$, then:

$$\kappa = \frac{2-\gamma}{3-\gamma}.\frac{N^{\frac{3-\gamma}{\gamma-1}}k_{\min}^{3-\gamma} - k_{\min}^{3-\gamma}}{N^{\frac{2-\gamma}{\gamma-1}}k_{\min}^{2-\gamma} - k_{\min}^{2-\gamma}} = \frac{2-\gamma}{3-\gamma}.\frac{N^{\frac{3-\gamma}{\gamma-1}} - 1}{N^{\frac{2-\gamma}{\gamma-1}} - 1}.k_{\min} \tag{2.32}$$

For $\gamma > 3$, as N approaches infinity, since both $N^{\frac{3-\gamma}{\gamma-1}}$ and $N^{\frac{2-\gamma}{\gamma-1}}$ approach 0, equation (2.32) implies that κ approaches $(2-\gamma)/(3-\gamma)k_{\min}$. Then, because $q_c = 1 - (\kappa - 1)^{-1}$, according to Theorem 2.2, we get that q_c approaches $1 - \left(\frac{2-\gamma}{3-\gamma}k_{\min} - 1\right)^{-1}$ as N approaches infinity.

By contrast, for $2 < \gamma < 3$, as N approaches infinity, since $N^{\frac{2-\gamma}{\gamma-1}}$ approaches 0 and $N^{\frac{3-\gamma}{\gamma-1}}$ approaches infinity, κ approaches infinity. So, according to Theorem 2.2, q_c must approaches 1 as N approaches infinity.

Similarly, for $1 < \gamma < 2$, because $\lim_{N\to\infty} \frac{N^{\frac{3-\gamma}{\gamma-1}}-1}{N^{\frac{2-\gamma}{\gamma-1}}-1} = \infty$, κ also diverges. This implies that q_c approaches 1 as N approaches infinity.

For $\gamma = 2$, we calculate the value of κ as follows:

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{\int_{k_{\min}}^{k_{\max}} cdk}{\int_{k_{\min}}^{k_{\max}} ck^{-1}dk} = \frac{k_{\max} - k_{\min}}{\ln \frac{k_{\max}}{k_{\min}}} \tag{2.33}$$

Using equation (2.30) to express $k_{\max}$ in term of N and $k_{\min}$, we rewrite equation (2.33):

$$\kappa = \frac{N^{\frac{1}{\gamma-1}}k_{\min} - k_{\min}}{\ln \frac{N^{\frac{1}{\gamma-1}}k_{\min}}{k_{\min}}} = \frac{N-1}{\ln N}k_{\min} \tag{2.34}$$

Since $\lim_{N\to\infty} \frac{N-1}{\ln N} = \infty$, κ diverges. Consequently, q_c approaches 1 as N approaches infinity. ■

Theorem 2.3 plays an important role in determining the resilience level of many real complex networks against random removal of nodes. Indeed, many of them, such as the film actors network, the email messages network, the word co-occurrence, the Internet, the peer-to-peer network, the metabolic network, the protein interactions are scale-free networks with degree exponents mostly ranges from 1^+ to 3^- (see the summary provided in [Cohen, 2010]. Therefore, this theorem confirms that these networks' giant components still exist with large fractions of randomly removed nodes. It means that these real networks have an extremely high level of resilience under failures. For example, over 99% of the Internet's nodes must be damaged to destroy the network's giant component [Cohen, 2000].

On the other hand, if attackers know a network's structure, they usually attack nodes playing important roles in the network first. In many cases, the robustness of a node may depend on its connectivity. That's the reason why attacking nodes with the highest degrees is a strategy worth considering. Under this intentional attack, many studies, such as [Albert, 2000; Callaway, 2000; Crucitti, 2004b; Gallos, 2006], found that a large scale-free network can be broken by a comparatively small fraction of removed nodes. Cohen et al. [Cohen, 2001] demonstrated

theoretically the vulnerability of a scale-free network against this attack strategy:

Theorem 2.4. *In a large scale-free network with degree exponent γ, under an intentional attack to the most highly connected nodes, the probability that an edge links to a deleted node approximates 1 for $1 < \gamma \leq 2$, and approximates $q^{\frac{2-\gamma}{1-\gamma}}$ for $\gamma > 2$, where q is the fraction of attacked nodes.*

Proof. Let $P(k) = ck^{-\gamma}\left(k = \overline{k_{\min}, k_{\max}}, \gamma > 1\right)$ where $1 \ll k_{\min} \ll k_{\max}$. After the attackers damage the most highly connected nodes, the nodes and their links are removed from the network. Thus, the network has a new cutoff degree $\tilde{k}_{\max} < k_{\max}$. Since the removal fraction q is the probability that a node has a degree larger than $\tilde{k}_{\max}$, using the hypothesis (2.29) and the normalized constant c given in equation (2.28), we obtain:

$$q = \int_{\tilde{k}_{\max}}^{k_{\max}} P(k)dk = \int_{\tilde{k}_{\max}}^{\infty} P(k)dk - \int_{k_{\max}}^{\infty} P(k)\,dk = (\gamma - 1)\,k_{\min}^{\gamma-1} \int_{\tilde{k}_{\max}}^{\infty} k^{-\gamma}dk - \frac{1}{N} \tag{2.35}$$

Consequently, as N approaches infinity, from equation (2.35), we can estimate the value of $\tilde{k}_{\max}$ as follows:

$$\tilde{k}_{\max} \approx k_{\min} q^{\frac{1}{1-\gamma}} \tag{2.36}$$

Let $\tilde{q}$ be the probability that an edge links to a deleted node. Then, $\tilde{q}$ equals the fraction of edges belonging to deleted nodes, i.e., for $\gamma \neq 2$,

$$\tilde{q} = \frac{\int_{\tilde{k}_{\max}}^{k_{\max}} kP(k)dk}{\int_{k_{\min}}^{k_{\max}} kP(k)dk} = \frac{k_{\max}^{2-\gamma} - \tilde{k}_{\max}^{2-\gamma}}{k_{\max}^{2-\gamma} - k_{\min}^{2-\gamma}} \tag{2.37}$$

For large networks, we replace $k_{\max}$ and $\tilde{k}_{\max}$ in the above equation by the expression given in equations (2.30) and (2.36), respectively, to get:

$$\tilde{q} \approx \frac{N^{\frac{2-\gamma}{\gamma-1}} k_{\min}^{2-\gamma} - k_{\min}^{2-\gamma} q^{\frac{2-\gamma}{1-\gamma}}}{N^{\frac{2-\gamma}{\gamma-1}} k_{\min}^{2-\gamma} - k_{\min}^{2-\gamma}} = \frac{N^{\frac{2-\gamma}{\gamma-1}} - q^{\frac{2-\gamma}{1-\gamma}}}{N^{\frac{2-\gamma}{\gamma-1}} - 1} \tag{2.38}$$

Let N approach infinity in the right-hand side of equation (2.38). Then, for $\gamma > 2$, $N^{\frac{2-\gamma}{\gamma-1}}$ approaches 0, so we can approximate $\tilde{q}$ as follows:

$$\tilde{q} \approx q^{\frac{2-\gamma}{1-\gamma}} \tag{2.39}$$

By contrast, for $1 < \gamma < 2$, it's easy to see that the limit of the right-hand side of equation (2.38) is 1, as N approaches infinity. So, $\tilde{q} \approx 1$ for large N.

In the case of $\gamma = 2$, we have to rewrite the equation (2.37) as follows:

$$\tilde{q} = \frac{\int_{\tilde{k}_{\max}}^{k_{\max}} kP(k)dk}{\int_{k_{\min}}^{k_{\max}} kP(k)dk} = \frac{\int_{\tilde{k}_{\max}}^{k_{\max}} k^{-1}dk}{\int_{k_{\min}}^{k_{\max}} k^{-1}dk} = \frac{\ln\left(\frac{k_{\max}}{\tilde{k}_{\max}}\right)}{\ln\left(\frac{k_{\max}}{k_{\min}}\right)} \tag{2.40}$$

For large networks, we replace $k_{\max}$ and $\tilde{k}_{\max}$ in the above equation by the expression given

in equations (2.30) and (2.36), respectively. Then,

$$\tilde{q} \approx \frac{\ln(Nq)}{\ln(N)} \tag{2.41}$$

The limit of the right-hand side of equation (2.41) is 1, as N approaches infinity, So, $\tilde{q} \approx 1$ for large N. ■

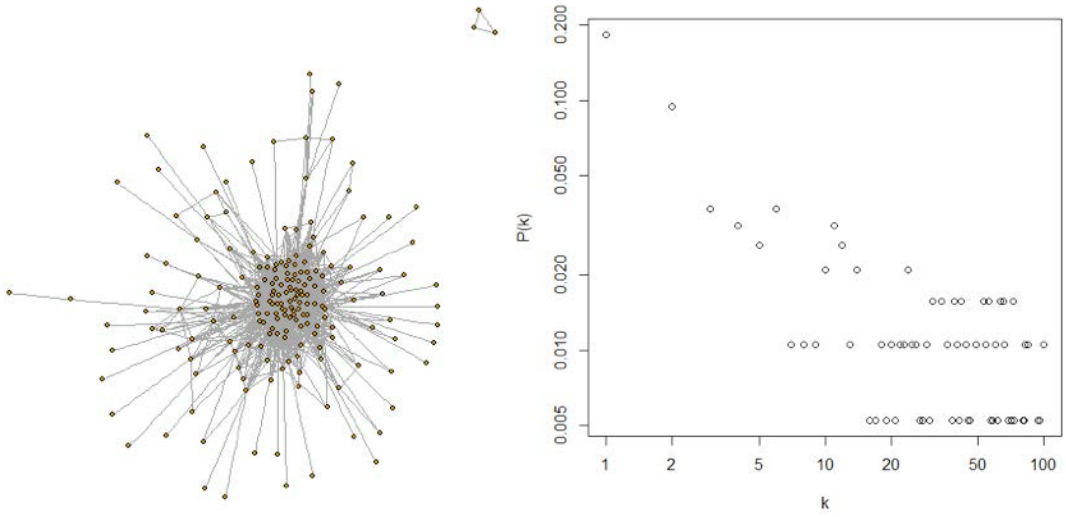
For a large scale-free network with degree exponent $1 < \gamma \leq 2$, Theorem 2.4 shows that, with an arbitrary fraction q of removed nodes, we can destroy most of the edges when the removed nodes have the highest degrees. Thus, attackers only need a little knowledge of these hubs to break the network entirely.

The results of Theorem 2.3 and Theorem 2.4 can be explained as follows. There are a few highly connected nodes in a scale-free network because of the inhomogeneity of its degree distribution. These hubs control the network's connectedness. Therefore, random node removal does little damage due to the fact that the chances of selecting randomly one of the few hub is negligible. By contrast, under the attack to the most connected nodes, the hubs' removal dramatically changes the network's topology and decreases the ability of the remaining nodes to communicate with each other. As a result, the network is extremely robust under random failure but very fragile under the attack.

In our financial context, we already know that two representative graphs of a stock system, the MST network and the correlation-based threshold network, can be considered scale-free networks. To study the vulnerability of the market under improper operations of some of its parts, we use the correlation-based threshold network. The reason is that, with a suitable threshold, the network helps get an overview of the research market by avoiding neglecting too many connections as the MST network, and helps reduce noises from small stock correlations, which are usually unstable over time. However, the degree exponent of a correlation-based threshold network of stocks is often low, especially the networks associated with emerging markets (let's see examples given in Figure 2.12). Hence, as an emerging market, the Vietnamese market's robustness under failures of its arbitrary components has to be studied carefully. In [Nguyen, 2018], we perform the random breakdown of the correlation-based threshold network of stocks listed on the HSX in the period from 01/01/2015 to 05/19/2017. The selected threshold is 0.25, which equals the 97% -quantile of the empirical stock correlations [2]. We plot the network in Figure 2.13. Its size is $N = 191$, and its degree exponent of 1.3 is also close to 1.[3]

[2] We also try with different thresholds such as 0.3, 0.35, 0.4... and get similar results of the market's robustness.

[3] To get an estimation for a power-law distribution from empirical data, we need a histogram of node degrees. Hence, the estimated degree exponent depends on the bins' size. However, the estimation doesn't affect our results about the network's resilience because we break the network based on its real connections.



(a) The network with the threshold of 0.63 (b) Degree distribution

Figure 2.13: The correlation-based threshold network of stocks listed on the HSX in the period 01/01/2017 – 01/01/2019 and its degree distribution.

Firstly, we in turn removing an arbitrary node until the giant component is destroyed, using the Molloy-Reed criterion (2.8). We repeat this process many times to get the average critical threshold according to the Monte-Carlo method (see Algorithm 4). As a result, we found that the network still exceedingly robust under random breakdown with the critical threshold of 95% approximately. In other words, one needs to randomly destroy over 95% of this network's nodes to make the spanning cluster collapse. So, Theorem 2.3 is valid for our financial network even though the number of its nodes is much less than other real scale-free networks such as the Internet, the protein interactions,...

Algorithm 4 Compute the critical threshold q_c of network resilience under random failure

Require: Network $G = (Nodes, Edges)$
procedure RESILIENCE_RANDOM_FAILURE(G)
  $N \leftarrow$ size of $Nodes$
  ▷ Take the Monte Carlo simulation with a large number M of iterations
  $sum\_qc \leftarrow 0$
  for $m \in \overline{1, M}$ **do**
    $G1 \leftarrow G$
    $P(k) \leftarrow$ degree distribution of $G1$.
    $number\_removed\_nodes \leftarrow 0$
    while $\frac{\langle k^2 \rangle}{\langle k \rangle} \geq 2$ **do**
      $i \leftarrow$ a node chosen randomly from nodes of $G1$.
      $G1 \leftarrow G1$ after removing i and edges linked to i.
      $P(k) \leftarrow$ degree distribution of $G1$
      $number\_removed\_nodes \leftarrow number\_removed\_nodes + 1$
    end while
    $sum\_qc \leftarrow sum\_qc + number\_removed\_nodes/N$
  end for
  $q_c = sum\_qc/M$ ▷ Output
end procedure

Next, we observe the resilience of the Vietnamese stock network under different attack strategies, including the strategy based on the *initial degrees of nodes* (ID), the strategy based on the *initial betweenness centrality of nodes* (IB), the strategy based on the *recalculated degrees*

of nodes (RD), and the strategy based on the *ecalculated betweenness centrality of nodes* (RB). For more specific, we perform these strategies as follows:

- The ID strategy: We remove, in turn, a node in the network in the descending order of the degree distribution of the original correlation-based network until the Molloy-Reed criterion (2.8) is invalid in the current network. This process imitates the attack to highly connected nodes we've just discussed.
- The IB strategy: Similar to the ID strategy but the nodes are eliminated sequentially in the descending order of their betweenness centrality. This strategy is expected to break a large scale-free network into many small pieces more rapidly than the first strategy because the betweenness centrality of a node can model the node's ability to connect regions of dense nodes in the network.
- The RD strategy: A node is deleted from the current network if its degree is largest in the current network. It means that we have to recalculate the degree distribution every time we want to remove one node.
- The RB strategy: A node is deleted from the current network if its betweenness centrality is largest in the current network. It means that we have to recalculate the betweenness centrality of nodes every time we want to remove one node.

In particular, similar to Algorithm 4, Algorithm 5 given below is used to find the critical threshold q_c based on the Molloy-Reed criterion (2.8) when we attack the highly connected nodes of a network. This algorithm is applied for both cases: using the arrangement by node degree of the initial network and using the updated arrangement after deleting a node. For the strategies based on the betweenness centrality, we replace the arrangements by nodes' degrees with the corresponding arrangements by nodes' betweenness.

Algorithm 5 Compute the critical threshold q_c of network resilience under attacks to the highly connected nodes.

Require: Network $G = (Nodes, Edges)$
procedure RESILIENCE_ATTACK_BY_DEGREE(G)
  $N \leftarrow$ size of $Nodes$
  $degree\_order \leftarrow$ the decreasing arrangement of nodes in G by their degrees
  $G1 \leftarrow G$
  $P(k) \leftarrow$ degree distribution of $G1$
  $number\_removed\_nodes \leftarrow 0$
  while $\frac{\langle k^2 \rangle}{\langle k \rangle} \geq 2$ **do**
    If G is attacked by the ID strategy **then**
      $i \leftarrow degree\_order$'s first node that appears in $G1$
    Else
    If G is attacked by the RD strategy **then**
      $i \leftarrow$ the node with the most connections in $G1$
    $G1 \leftarrow G1$ after removing i and edges linked to i.
    $P(k) \leftarrow$ degree distribution of the network $G1$
    $number\_removed\_nodes \leftarrow number\_removed\_nodes + 1$
  end while
  $q_c = number\_removed\_nodes/N$ ▷ Output
end procedure

The ID and IB strategies may less effective than the others since they use outdated information, which may be very different from the current network's structure. Our result on the Vietnamese stock market shown in Figure 2.14 confirms this conjecture, and it is similar to the one of Nie et al. [Nie, 2015]. In this figure, we plot the ratio of the giant component's size P_∞ to the initial network's size N after removing a fraction q of nodes randomly or removing nodes intentionally by the attack strategies. Obviously, the decrease of each recalculated strategy is much sharper than ones of the strategy basing on similar structural information got from the initial network. Consequently, while the ID and IB strategies make the network completely broken with the critical threshold of 49.74% and 45.55%, respectively, the RD and RD strategies only require a smaller critical threshold of respectively 36.65% and 43.98%. It means that we have to remove more nodes to destroy the network's global connectivity if we only use the network's initial information.

The considerable difference between the ID and RD strategies can be explained due to the fact that just a few highly connected nodes in a scale-free network control its entire connectivity. Thus, the removal of the most connected nodes makes extreme changes to the network's topology. As a consequence, the initial degree distribution no longer effectively reflects the new structure. This issue also explains why the ID strategy is less efficient than the IB strategy.

On the other hand, we can see that the initial ranking by nodes' betweenness only changes a little, so the difference between the IB and RB strategies is relatively negligible. The reason is that when removing nodes with the highest betweenness, we remove the paths that go through these nodes, but other paths still exist, then the overall betweenness' ranking is less affected.

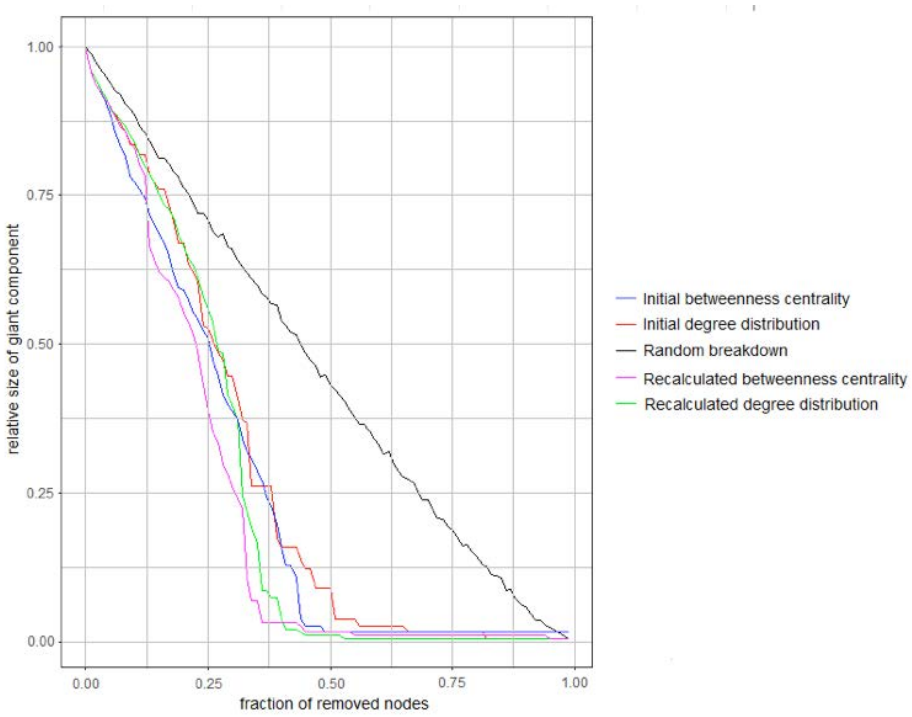


Figure 2.14: The relative size of the giant component as a function of the fraction of removed nodes under the random failure of nodes and different attack strategies to the correlation-based threshold network of stocks listed on the HSX in the period 01/01/2017 – 01/01/2019.

Besides, when comparing the strategies using recalculated structural information, we found an interesting phenomenon. Clearly, Figure 2.14 shows that the RB strategy almost reduces the giant component's size more rapidly than the RD strategy does. However, near the end of the giant component's disappearance, the RB strategy becomes slower in destroying the component. Finally, the RB strategy stops to break the giant component of the network after removing 43.98% of nodes. This critical threshold is much larger than the one of 36.65% under the RD strategy. To understand this phenomenon, let's remind that the RB strategy prefers cutting nodes playing important roles in connecting concentrated clusters. Therefore, at the beginning steps of the attack, the RB strategy breaks the network into many sub-networks. By this method, the giant component's size reduces more significantly. However, near the critical point, the RD strategy gets stuck in a highly connected cluster, which has many nodes but low betweenness. Meanwhile, the RB strategy removes nodes in other clusters, and the maximum size remains unchanged. We illustrate this explanation in Figure 2.15. The figure plots the network structure after removing 25%, 34%, 36%, and 41% of nodes, respectively, using the two strategies.

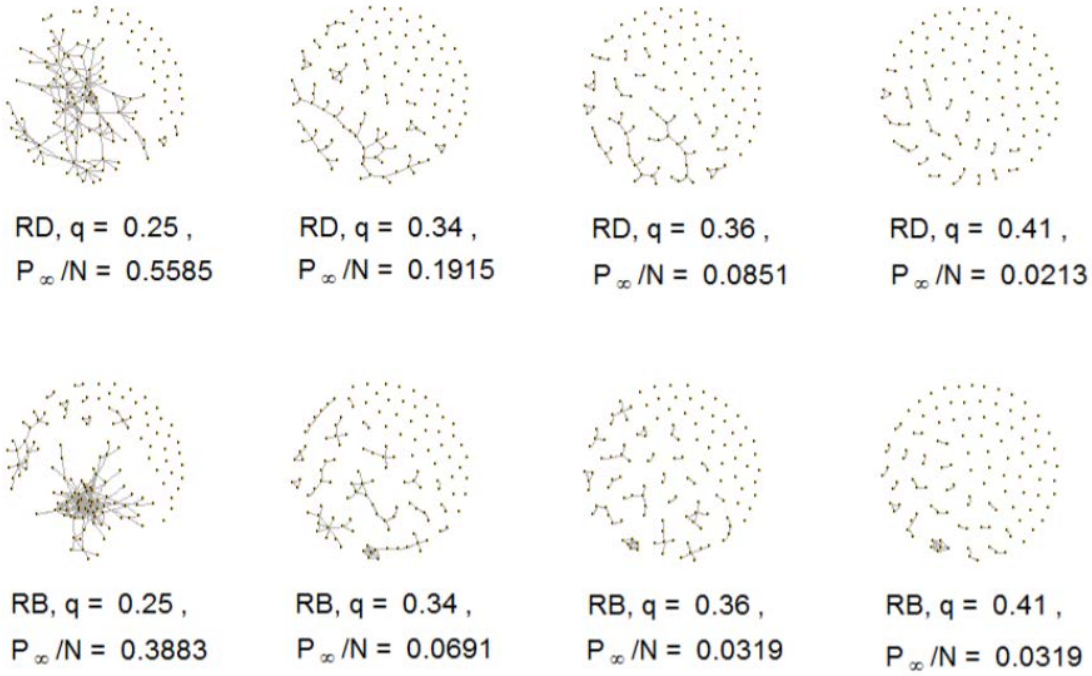


Figure 2.15: The correlation-based threshold network of stocks listed on the HSX in the period 01/01/2017 – 01/01/2019 after removing a fraction q of nodes by the RD and RB strategies. P_∞ is the giant component's size.

However, the critical thresholds of the Vietnamese stock network under the research attack strategies are drastically larger than the ones of theoretical scale-free networks, given by Theorem 2.4, and the ones of other real complex networks, which are often less than 10% [Albert, 2000; Cohen, 2001]. For example, the critical threshold of the Internet, the WWW and the temozolomide resistant network are 0.03, 0.067, and 0.02, respectively [Albert, 2000; Azevedo, 2015]. The cause of the less vulnerability of the Vietnamese stock network under attacks is the low degree

exponent of the network, which corresponding to a denser structure for the network's topology. Moreover, the perturbation at the right tail of the network's empirical degree distribution also makes Theorem 2.4 not perfectly available for this case.

Briefly, we used the correlation-based threshold network to study the robustness of a stock market when its constituents get errors. Despite the low degree exponents of such networks, we demonstrated that the networks behave similarly to other real scale-free networks due to the presence of a few hubs: the networks are significantly robust under random failures of their constituents, but much fragile under intentional attacks, especially under the attack to the most connected nodes determined by recalculating the degrees of remaining nodes. However, if we want to damage a stock network to only a level of its size rather than completely destroy it, the RB strategy can be a better option. These results help construct a steady stock market and protect it efficiently. For example, we should alert significant decreases in prices of stocks corresponding to even a small fraction of nodes with high degrees or high betweenness centrality because they may cause a considerable fall of the entire market.

4.3 Phase Transitions

In order to study the spreading of a price shock of one stock to the entire market, the MST of the correlation-based network is more suitable than the correlation-based threshold network since the MST provides the most probable path for the spreading. However, as a complex system, the behavior of a stock market can be very complicated. So, the topological arrangement of its MST network often changes due to the changes in the constituents' behaviors and their relationships, the environment's variation, or the impacts of other external factors. When the MST's structure alters drastically, this especially affects the ability of a shock to spread overall the corresponding market. Therefore, understanding phase transitions of the MST's structure becomes an attractive approach in evaluating the market's stability and controlling the systemic risk. In this section, our research subject is the change of the MST network's structure over time.

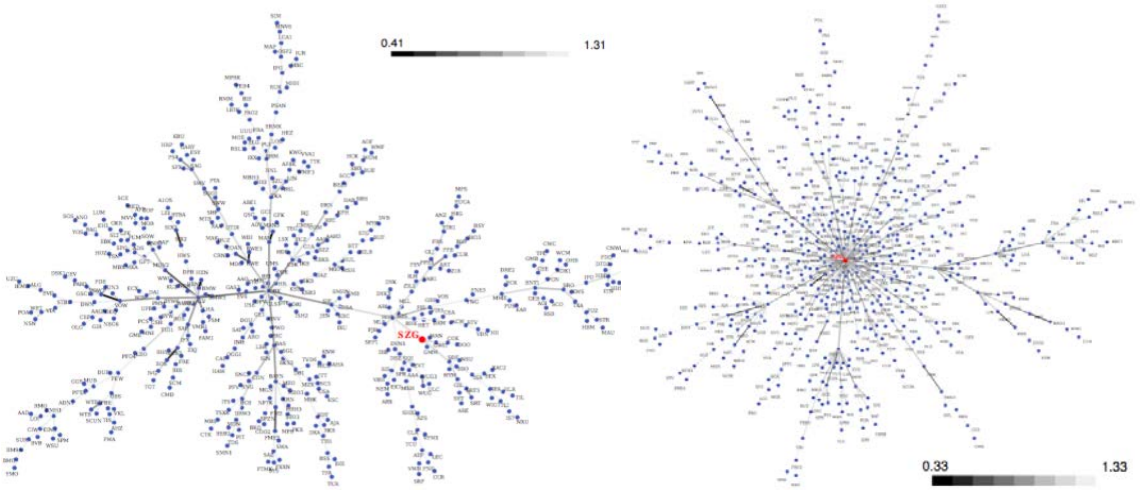
In [Nguyen, 2019c], we found that the change is really homogenous to different states of the market. Similar results are found on the *Frankfurt Stock Exchange* (FSE) [Wiliński, 2013] and the *Warsaw Stock Exchange* (WSE) [Sienkiewicz, 2013]. The following remark gives more details about this issue.

Remark The dynamic of the MST of a correlation-based network of stocks goes through three phases:

- phase of hierarchical MST – a (relatively) stable stock market state
- phase of the superstar-like MST – a transient market state
- phase of hierarchical MST decorated by few local star-like trees – a (relatively) stable stock market state.

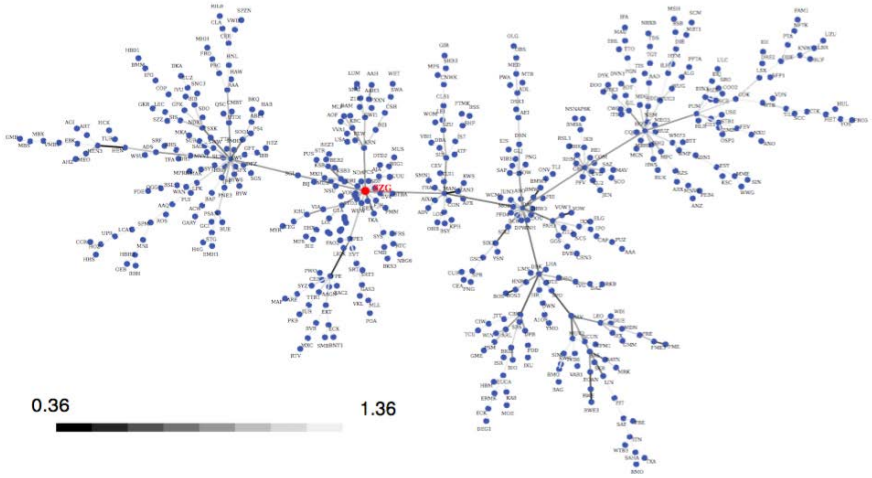
Figure 2.16 shows the MST associated with the FSE in three different periods. Each tree in the figure represents the general structure of the MST in every phase above. The result is

provided in [Wiliński, 2013]. Besides, instead of reflecting the trees by their exact geometric distance, a link between two nodes is in dark grey if the nodes' distance is small. According to the authors, this presentation makes the tree more readable. Meanwhile, Figure 2.17 shows a similar result observed in an emerging market – the HSX, provided in our study [Nguyen, 2019c]. The degree distributions of the MSTs shown in this figure are provided in Figure 2.18 and Figure 2.19c, respectively.



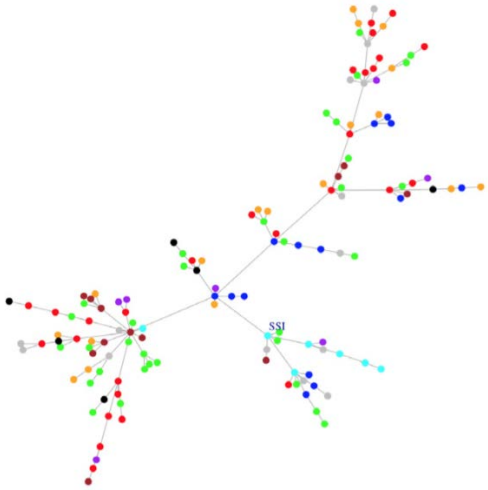
(a) The hierarchical MST for the period from 01/03/2005 to 03/09/2006

(b) The superstar-like MST for the period from 04/20/2006 to 10/31/2007

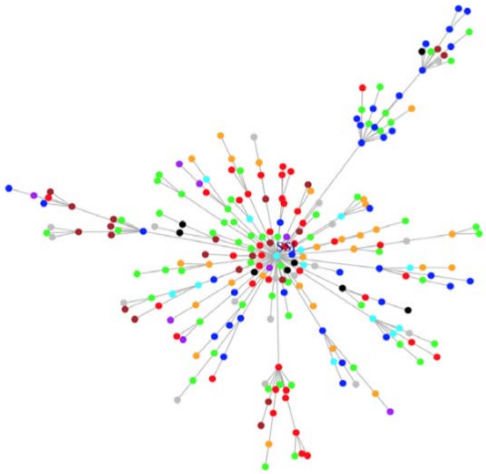


(c) The hierarchical MST decorated by few local star-like trees for the period from 06/01/2007 to 08/12/2008

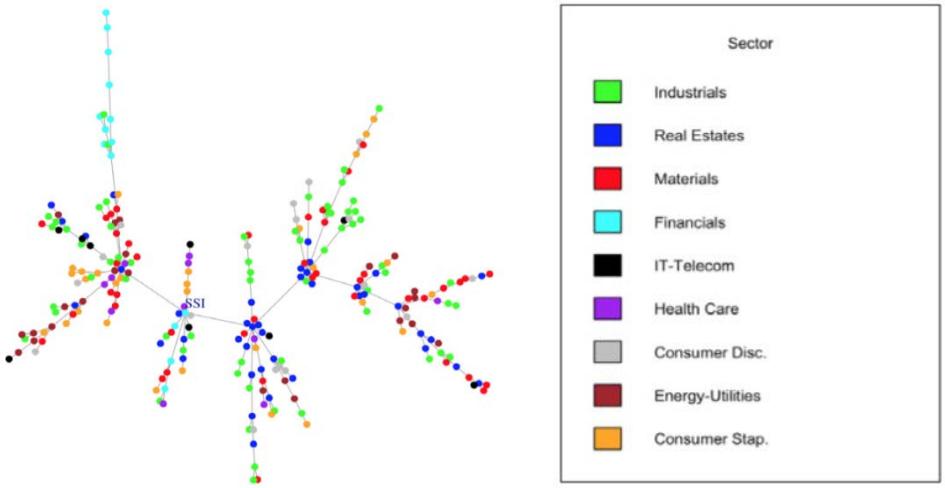
Figure 2.16: Structural change of the MST network of stocks listed on the FSE [Wiliński, 2013].



(a) The hierarchical MST for the period from 03/31/2009 to 10/19/2010

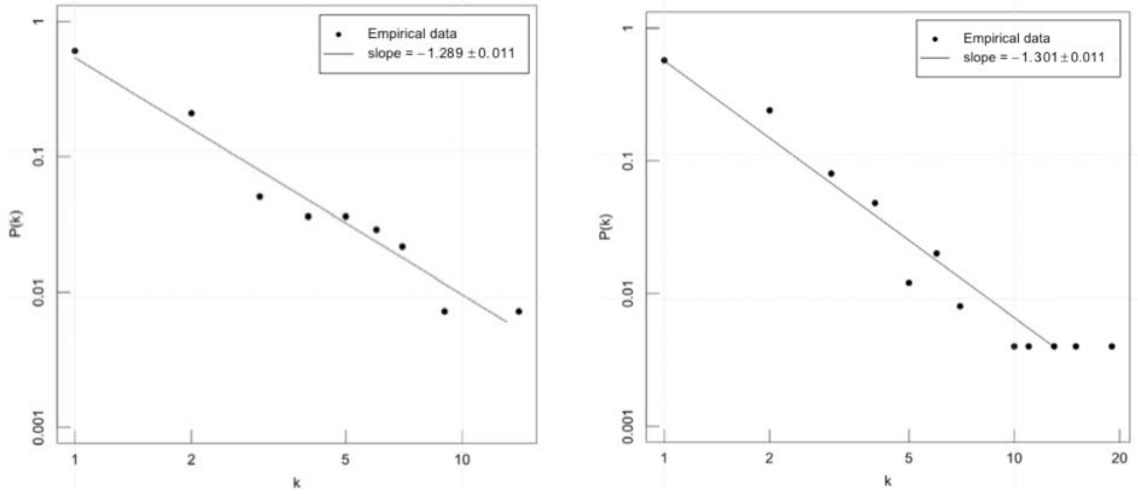


(b) The superstar-like MST for the period from 05/16/2012 to 12/02/2013



(c) The hierarchical MST decorated by few local star-like trees for the period from 01/14/2014 to 08/18/2015

Figure 2.17: Structural change of the MST network of stocks listed on the HSX.



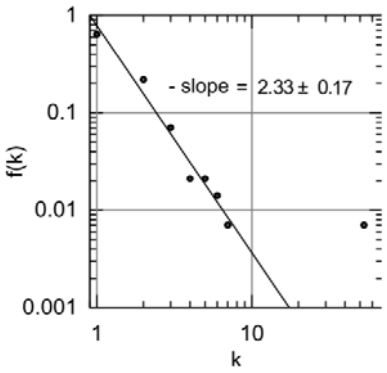
(a) The degree distribution of the hierarchical MST for the period from 03/31/2009 to 10/19/2010

(b) The degree distribution of the hierarchical MST decorated by few local star-like trees for the period from 01/14/2014 to 08/18/2015

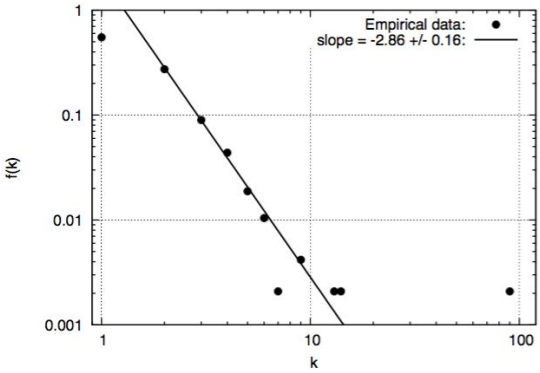
Figure 2.18: Degree distribution of the hierarchical MST network of stocks listed on the HSX.

The MST in the first and third phases has the common structure of stock networks, the scale-free structure. Especially when the MST is in the third phase, its local hubs are reflected on the right tail of its degree distribution by the presence of more points near the line fitting the power law (see Figure 2.18 where we plot the degree distributions of the hierarchical MSTs shown in Figure 2.17a, 2.17c). Generally, in both of the two phases, because there is no node with an extremely large degree comparing with others in the MST, the network has a hierarchical structure. In this case, the market is relatively stable under a shock of price fluctuations starting from a small group of stocks because the shock's propagation needs a long process to reach the entire market. Besides, since the networks' global connectivity doesn't mostly depend on only one stock as what happens in the star-like structure, it's not easy to damage the network's global connectivity.

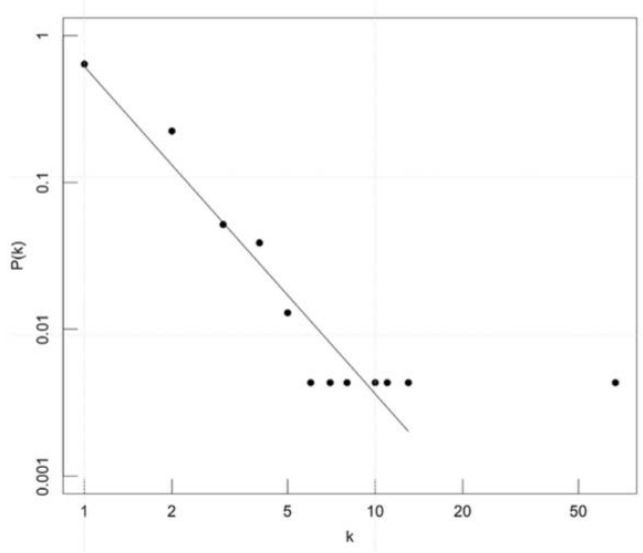
Now, let's focus on the second phase, where the MST network has a star-like structure. In this case, the network has a super hub, the largest hub whose degree is extremely higher than the ones of other nodes. Remind that, in an MST, the number of connections equals $N - 1$. So, the presence of the super hub implies the lack of connections between pairs of other nodes. Because the usual structure of the MST network is absent, the network can close its scale-free property. Indeed, in the scatter plot of the empirical degree distribution of a star-like MST after neglecting the point representing the super hub, the remaining points fit well a power law. Since this point is very far from the fitted line of the power law in the log-log plot, its existence dissolves the scale-free property. This empirical result is illustrated in Figure 2.19, where we plot the degree distributions of the star-like MST shown in Figure 2.17b of stocks listed on the HSX and the star-like MSTs found on other developed markets, including the FSE and the WSE.



(a) Degree distribution of the MST network on the WSE for the period from 06/01/2007 to 08/12/2008 [Sienkiewicz, 2013]



(b) Degree distribution of the MST network on the FSE for the period from 04/20/2006 to 10/31/2007 [Wiliński, 2013]



(c) Degree distribution of the MST network on the HSX for the period from 05/16/2012 to 12/02/2013

Figure 2.19: Degree distribution of the star-like MST and the fitted line of a power law after neglecting the super hub.

Consequently, a star-like structure of the MST network is a crucial sign informing us that we are dealing with an exceptional event. An important hypothesis is that the event is a coming stock market crash. This argument is compatible with our observation on the HSX and other empirical studies on developed markets. For more specific, in [Nguyen, 2019c], we found that the star-like MST appears in the period when the Vietnamese economy is under serious stressing. During the stressing, the interest rate was particularly high when it went from 10% to as high as 30%/year in a short time of 8 months. Similarly, in [Wiliński, 2013] and [Sienkiewicz, 2013], the authors show that the star-like MST network of stocks listed on the FSE and the WSE occurs in the early period of the worldwide financial crash 2007 – 2008, so the authors speculate that the star-like MST plays the role of a crash precursor for the corresponding market. Because of the special role, the three following questions attract our attention. The first is why the star-like structure likely relates to an unstable state of the corresponding market. Secondly, we wonder how we can quantify the change in an MST's structure from a chain-like one to a star-like one.

The last is about the super hub's role.

For the first question, although the star-like MST is exceptional, this strange is not enough to confirm the connection between this special structure and financial instability. Instead, let's see how a stock market fragile when its MST network looks like a star. In this case, the most probable paths connecting pairs of nodes in the correlation-based network are almost shorter. Then, a price shock of one stock can transmit to the entire market more easily after a few steps through the super hub. Therefore, with this structure, the network is more sensitive to shocks than usual. For example, Figure 2.20 shows the synchronization between the small average shortest-path length of the MST constructed by stocks listed on the HSX and the severe decline period of the Vietnamese economy. In this figure, we construct the MST network in different time windows of the same length in the period from 01/09/2008 to 12/31/2017. The length is 390, the number of trading days. The sliding time is 60 trading days. We also highlight the points corresponding to the three MSTs illustrated in Figure 2.17. Besides, to normalize the length, we divide it by the number of nodes of the corresponding MST. Without this normalization, the average shortest-path length of the MST constructed in a period can be larger than the one of the MST constructed in the preceding periods when more companies are listed. At that time, this increment of the average shortest-path length is just a mechanical rise due to the increase of the network's size instead of reflecting meaningful information about the MST's structural variation.

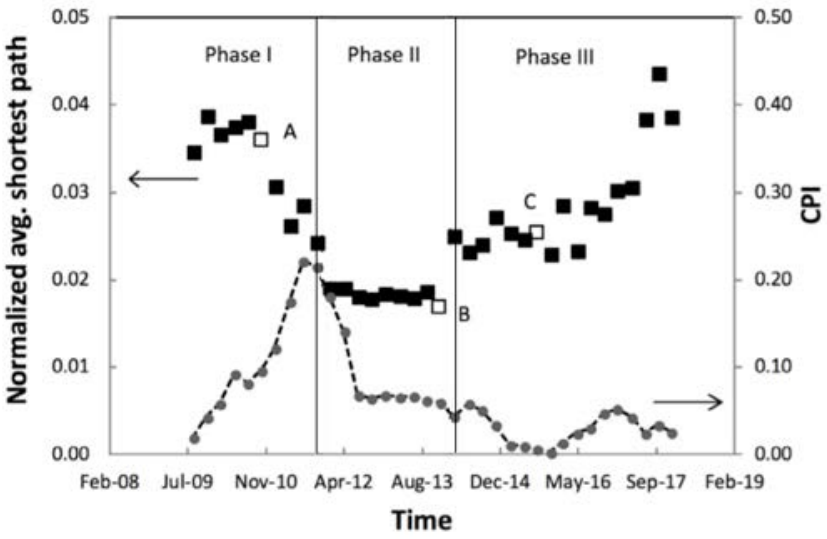


Figure 2.20: Synchronization between the small normalized average shortest-path length of the MST network and the severe decline period of the Vietnamese economy in Phase II.

Due to the connection between a star-like MST and financial recessions, quantifying the variation of the MST network from a hierarchical structure to a star-like structure is useful for many problems such as setting automatic trading strategies, managing the systematic risk... To measure the change of the MST network's structure over time, the survival ratio introduced in Definition 2.12 can be a useful tool. Because the ratio provides the proportion of common

connections between two MSTs constructed in two consecutive periods, the network goes through a phase transition of its structure at time t if the ratio at this time point is significantly smaller than the ones nearby. For example, using the same database as Figure 2.20, Figure 2.21 shows the dynamics of the survival ratio of the MST network of stocks listed on the HSX. We can see that the phase transitions of the market in Figure 2.20 are also well-defined in Figure 2.21.

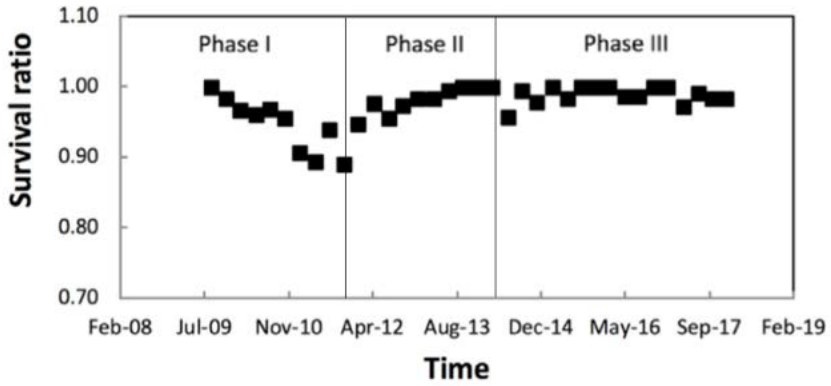


Figure 2.21: Synchronization between the depression of the survival ratio of the MST network constructed on the HSX and the phase transitions of the Vietnamese economy.

However, the MST's survival ratio is not a clearly measure to detect the star-like structure because we have to observe the ratio for a long period to prevent confusing the time of phase transitions. Therefore, we also use another measure, the allometric exponent, to easily get the geometrical information of the MST network in a certain time window. As introduced in Section 3, the allometric scaling relation appears in the MST network of a stocks system with the allometric exponent η ranging from 1 to 2. The exponent is closer to 1 if the MST is more similar to a star. For example, η equals 1.289 ± 0.011, 1.213 ± 0.013, and 1.301 ± 0.011, respectively, for each MST drawn in Figure 2.17. Obviously, the allometric exponent of the second tree, which has a star-like structure, is closer to 1 than the exponents of the others in the figure. Using this allometric scaling relation, we can confirm that Phase II in Figure 2.20 is corresponding to a different state of the MST's structure, the star-like structure because of the low exponent of the allometric scaling relation in the phase. Figure 2.22 shows the simultaneous variation between the dynamics of the allometric exponent and the normalized average shortest-path length, which was demonstrated to compatible with the depressed period of the corresponding market in Figure 2.20.

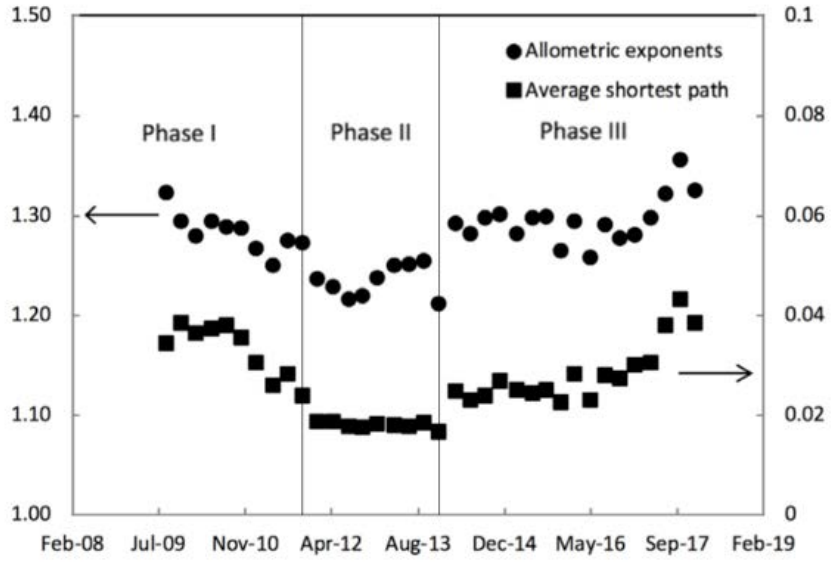


Figure 2.22: Synchronization between the normalized average shortest-path length's decline and the allometric exponent's decline of the MST network constructed on the HSX.

In addition, remember that the allometric scaling relation can measure the hierarchical degrees of nodes in the MST's structure through variable C. For a stock market, the value C of a node represents how important the corresponding stock influences others in the market if a crisis occurs. Therefore, the value helps explain the role of the super hub in a star-like MST. For example, let's draw the two MSTs in Figure 2.17b and 2.17c again such that the node size is an increasing function of C. We found that there is only one stock having an extremely high impact C on the star-like network in Phase II (Figure 2.23a), while the total impact toward the whole market is distributed to many stocks in Phase III (Figure 2.23b). Then, if a shock occurs at the super hub, a propagation of this shock will occur almost instantly over the entire market. By contrast, for the hierarchical tree, the cascading failure process performs more slowly because the shock has to transfer to other important stocks before spreading throughout the network.

(a) Superstar-like MST for the period from 05/16/2012 to 12/02/2013

(b) Hierarchical MST decorated by few local star-like trees for the period from 01/14/2014 to 08/18/2015

Figure 2.23: MSTs constructed on the MST with $\log(C)$ as the node size.

Furthermore, the super hub's emergence in phase II also makes changes in the usual relationships between stocks belonging to the same business sector. Indeed, in a star-like MST, the number of a sector's intra-connections must decrease since stocks prefer connecting the super hub to connecting other stocks in the same sector. For example, for the same database with Figure 2.20, Figure 2.24 demonstrates the low same sector ratio of the MST in Phase II. However, although this remark gives more information about the role of the super hub, the low same sector ratio is not enough to confirm the instability of an arbitrary stock market. The reason comes from the lack of sectors' intra-connections in emerging markets, for instance, the Vietnamese market (see Figure 2.17), which is very different from the plentiful of such intra-connections in developed markets, for instance, the U.S. market (see Figure 2.3).

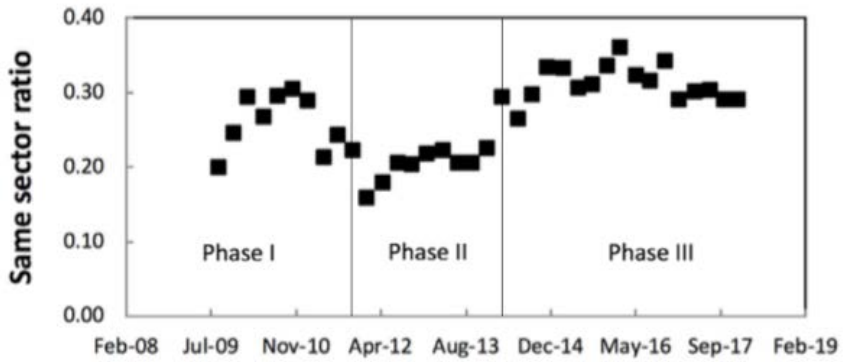


Figure 2.24: Synchronization between the decline of the same sector ratio of the MST network constructed on the HSX and the Vietnamese economy's unstable period.

Finally, it is meaningful to understand why a marginal company in the MST network becomes a central node in financial stressing periods. A common result in many studies is that the central node is relevant to companies providing financial services. Indeed, in the star-like MST constructed on the HSX, we found that the central node is corresponding to the Saigon Securities Incorporation – a stock brokerage company. Also, in such MST network constructed on the WSE, the central node is corresponding to the Capital Partners – an investment company [Sienkiewicz, 2013]. Similarly, in [Onnela, 2002; Onnela, 2003a], the central node of the U.S. network represents the General Electric Co. – a technology and financial services company. The companies' evolution from a marginal one to a central one in Phase II supposedly originates from the attractive financial products and available investment advice that the companies offered only in this period. In distressing periods of an economy, the products and advice can be considered as directions for corporations and investors to lean on to overcome the period. However, in [Wiliński, 2013], the central node of the star-like MST constructed on the FSE is corresponding to the Salzgitter AG–Stahl und Technologie, a company that manufactures steel and associated products. Thus, a deep analysis of the central company's characteristics should be taken in the future.

In conclusion, the MST network's structure of a stock system in both cases of developed markets and emerging markets goes over 3 phases. Its normal structure is hierarchical, but the structure becomes likely a star when the market is unstable. After the stress period, the MST's structure comes back to a hierarchical tree decorated with a few local hubs. The changes in the MST's structure from a hierarchical tree to a star-like tree and vice versa imply the changes in its scale-free property, which is a crucial characteristic in the corresponding market's robustness under failures. The changes also affect the market's ability of information's spread. Especially, we can determine the star-like structure of an MST network by the allometric exponent close to 1. Furthermore, the allometric scaling relation can help certify the roles of a stock in attending the propagation of a price shock to the entire network. The meaning of variables A and C of this relation should be studied more carefully. Besides, the relation of the central company to companies providing financial services is a far going hypothesis that we should study more. We can see the importance of financial companies in contributing to the collective behavior of a stock market in the next chapter.

Chapter 3

Spectral Property of Stocks' Cross-correlation Matrix

Objective

In this chapter, we study the spectral property of the correlation-based network, i.e., the spectrum of the cross-correlation matrix of stock returns. Because we only have the sample matrix, we use random matrix theory to get the "true" stock correlations. In particular, we pay attention to the sample matrix's largest eigenvalue, which is always extremely larger than the largest one predicted by the theory in our financial problems. Also, we use Principal Component Analysis to investigate the role of the unit eigenvector associated with the eigenvalue in reflecting the collective behavior of a stock market and the affect of an individual stock to others. Our results are empirically demonstrated in the Vietnamese and U.S. stock markets.

Contents

In the previous chapter, we used the cross-correlation matrix of stocks to construct representative networks for a stock market. However, we don't know the exact correlations of stocks. Instead, the correlations are empirically computed by finite time series of historical values of stock prices. Then, the observed period's length and the underlying market's noises considerably affect the estimation of the true correlations. Although the knowledge about the market can help reduce noises, it doesn't dissolve the problem completely. Meanwhile, *random matrix theory* (RMT) and P*principal component analysis* (PCA) support an available solution for this problem. They helps understand the spectral property of a sample cross-correlation matrix. From this matrix, we can estimate the "true" adjacency matrix of the correlation-based network of a stock system. The sample matrix's spectral property is very useful to capture important information about the network's structural characteristics [Cvetković, 1998]. In this chapter, we use the spectral properties analyzed by RMT and PCA to understand not only a stock network's structural properties but also the common interaction between entities of the underlying market.

1 Random Matrix Theory Applied to Stock Systems

RMT is a physics theory that helps get the accurate cross-correlation matrix of numerous entities. It's applied in many works, namely [Laloux, 1999; Laloux, 2000; Lux, 1999; Mehta, 2004; Plerou, 1999; Plerou, 2002]... It was developed to deal with the statistics of energy levels of complex quantum systems when physicists had difficulties in interpreting the spectra of the nuclei because the exact nature of the interactions was unknown. In particular, in 1951, Wigner used a real symmetric matrix with independent random elements to make predictions representing an average over all possible interactions [Wigner, 1951]. The deviations of the sample cross-correlation matrix compared to its RMT prediction are proposed to provide true information of the interactions because the deviations identify non-random properties of the research system [Guhr, 1998; Mehta, 2004]. Because of this benefit, the RMT's prediction becomes popular to analyze the spectral distributions of sample cross-correlation matrices in many complex systems such as the EEG data of brain [Šeba, 2003], the variation of basic atmospheric parameters that characterize the state of the atmosphere [Santhanam, 2001]. Similarly, in this study, we use RMT to understand the nature of the correlations of stock price fluctuations in a stock market.

From this point of view, let's consider a system of N stocks. According to Definition 2.1, we can compute the cross-correlation matrix $\mathbf{C} = (c_{ij})$ of stocks from the log-price changes, also call the stock returns, r_i, $i = \overline{1, N}$. Let $\tilde{r}_i$ be the normalized return of stock i, i.e.

$$\tilde{r}_i = \frac{r_i - \langle r_i \rangle}{\sigma_i} \tag{3.1}$$

Obviously, $\tilde{r}_i$ has zero mean and unit variance. Now, we can rewrite the formula of the empirical correlation between stock i and stock j in Definition 2.1 as follows:

$$c_{ij} = \langle \tilde{r}_i . \tilde{r}_j \rangle \tag{3.2}$$

Therefore, with T observations $\tilde{r}_i(t)$, $t = \overline{1,T}$, the correlation between stocks i and j is empirically estimated by the time average of the scalar product of two vectors $(\tilde{r}_i(t))_{t=\overline{1,T}}$ and $(\tilde{r}_j(t))_{t=\overline{1,T}}$. Let $\mathbf{X}$ be the matrix whose rows are N vectors $(\tilde{r}_i(t))_{t=\overline{1,T}}$, $i = \overline{1,N}$, then the cross-correlation matrix $\mathbf{C}$ can be expressed as

$$\mathbf{C} = \frac{1}{T}\mathbf{XX'} \tag{3.3}$$

From equation (3.3), in RMT, to consider a null hypothesis that the stock returns are strictly uncorrelated, we assume that the cross-correlation matrix $\mathbf{C}$ is equivalent to a purely random matrix $\mathbf{W}$ obtained from standard normally distributed i.i.d. time series. Such matrix is in the ensemble of Wishart matrix introduced by Wishart in 1928 [Wishart, 1928]. In particular, we're only interested in the Wishart matrix of real entries because financial data only contains real numbers.

Definition 3.1. *A real Wishart matrix is a random symmetric matrix* $\mathbf{W}$ *of the form:*

$$\mathbf{W} = \frac{1}{T}\mathbf{MM'} \tag{3.4}$$

where ' denotes matrix transposition, and $\mathbf{M}$ is a random matrix of size $N \times T$ such that:

- $(M_{ij})_{1\leq j\leq T}$ are independent samples of a real-value random variable m_i.
- $(m_1, \ldots, m_N)$ is a Gaussian vector with given covariance matrix $\mathbf{K}$.

T is called the degree of freedom.

Matrices $\mathbf{W}$ and $\mathbf{K}$ are both of size $N \times N$. Besides, in our financial context, $(m_1, \ldots, m_N)$ is a standard normal vector with covariance $\mathbf{K} = \mathrm{diag}(1, \ldots, 1)$ to interpret the null model of $\mathbf{C}$. This null hypothesis helps identify the effects of the randomness of $\mathbf{C}$. In other words, the difference of the spectral distribution of $\mathbf{C}$ from the one of $\mathbf{W}$ indicates the presence of meaningful information about true correlation of the assets.

Especially in a large system, it's demonstrated that the distribution of the spectrum of a Wishart matrix $\mathbf{W}$ with $\mathbf{K} = \mathrm{diag}(\sigma, \ldots, \sigma)$ follows a certain distribution, the Marčenko - Pastur distribution [Marčenko, 1967]. More detail is given in the below theorem, while its result is illustrated in Figure 3.1.

Theorem 3.1. *If* $N \to \infty, T \to \infty$ *in such a way that* $\frac{T}{N}$ *approaches a fixed number* $\alpha \geq 1$*, the empirical spectral distribution of the Wishart matrix* $\mathbf{W}$ *converges weakly, in probability, to the Marčenko - Pastur distribution with density* ρ *supported on* $[\lambda_-; \lambda_+]$ *and given by*

$$\rho(\lambda) = \frac{\alpha}{2\pi\lambda\sigma^2}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}, \qquad \forall\lambda \in [\lambda_-; \lambda_+] \tag{3.5}$$

where $\lambda_\pm = \sigma^2\left(1 \pm \alpha^{-\frac{1}{2}}\right)^2$.

Sketch of proof. As well as this theorem's application, its proof is also an attractive subject of many works such as [Dyson, 1971; Marčenko, 1967; Sengupta, 1999; Stein, 1969; Yaskov, 2016]

using various techniques, for instance, the method of moments, free probability method, the singular value decomposition, modification of the standard Cauchy–Stieltjes resolvent method. The most natural but long method to prove this theorem is the moment method, i.e., we compare the moments of the two distributions, the empirical distribution of the spectrum of $\mathbf{W}$ and the Marčenko - Pastur distribution. The sketch of this proof is that, firstly, we show that the k-th moment of the Marčenko - Pastur density ρ is

$$\int \lambda^k \rho(\lambda)\, d\lambda = \sigma^{2(k+1)} \sum_{r=0}^{k} \frac{\alpha^{-r} \binom{k-1}{r} \binom{k}{r}}{r+1} \tag{3.6}$$

On the other hand, let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ be the eigenvalues of a sample of $\mathbf{W}$, then the empirical cdf of the spectrum of $\mathbf{W}$ is

$$l(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{[\lambda_i, \infty)}(\lambda) \tag{3.7}$$

So, the k-th moment of the empirical density distribution of the spectrum of $\mathbf{W}$ is

$$\int \lambda^k \frac{dl}{d\lambda} d\lambda = \left\langle N^{-1} \mathrm{tr} \mathbf{W}^k \right\rangle \tag{3.8}$$

The proof of Theorem 3.1 is accomplished by showing that:

- for each positive integer k, the expectation of $N^{-1}\mathrm{tr}\mathbf{W}^k$ converges to the right-hand side of equation (3.6), and
- the variance of $\mathrm{tr}\mathbf{W}^k$ converges to 0. ■

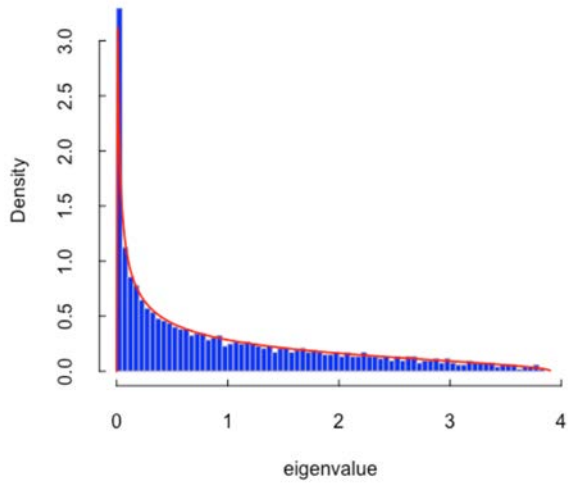


Figure 3.1: Compatibility between the Marčenko - Pastur distribution (red line) and the spectral distribution of the Wishart matrix obtained from $N = 1000$ i.i.d. standard normal random vectors with $\alpha = 1.05$.

In reality, many works found that most of eigenvalues of the cross-correlation matrices of asset price changes in world-wide financial markets agree surprisingly well with the range provided in Theorem 3.1. Some examples of these cross-correlation matrices can be listed as the matrix of daily returns of stocks comprised in the S&P 500 Index [Laloux, 1999], the matrix of daily returns of stocks corresponding to the largest companies traded on the NYSE [Plerou, 1999], the matrix of daily returns of the most actively traded German stocks [Rosenow, 2008], the matrix of daily returns of 20 world stock indices [Nobi, 2013], the matrix of daily returns of Korean stocks [Nobi, 2013], and the matrix of daily returns of Vietnamese stocks in our study [Nguyen, 2019b]. In Figure 3.2, we provide an example of this agreement on the cross-correlation matrix of stocks quoted on the HSX from 01/01/2017 to 01/01/2020. There are $N = 271$ stocks in this database observed in $T = 749$ trading days, i.e., $\alpha \approx 3.1$.

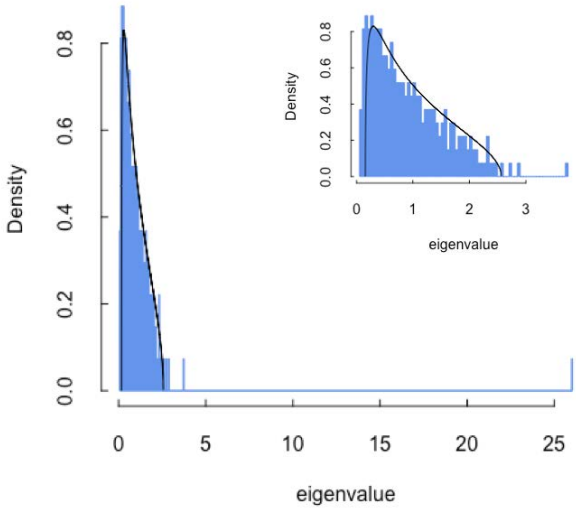


Figure 3.2: Explanation of the spectral distribution predicted by RMT (black line) for a large part of the spectral distribution of the cross-correlation matrix of stocks quoted on the HSX from 01/01/2017 to 01/01/2020 (insert: these two distributions when zooming in the eigenvalues without the largest one).

Furthermore, since the largest eigenvalues of $\mathbf{C}$ are inconsistent with the null hypothesis, we may consider another null hypothesis assuming that the matrix is purely random except for some eigenvalues significantly exceeding the theoretical upper limit λ_+. Then, we adjust the parameter σ to be less than 1 to subtract the contribution of these largest eigenvalues to the normal value $\sigma = 1$, as proposed by [Laloux, 1999]. Figure 3.3 provides an example given in [Laloux, 1999]. This figure shows the remarkable compatibility of the spectral distribution predicted by the new null model and the spectral distributions of the cross-correlation matrix extracted by the daily closing price of $N = 406$ stocks of the S&P 500 in $T = 1309$ trading days during the year 1991-1996.

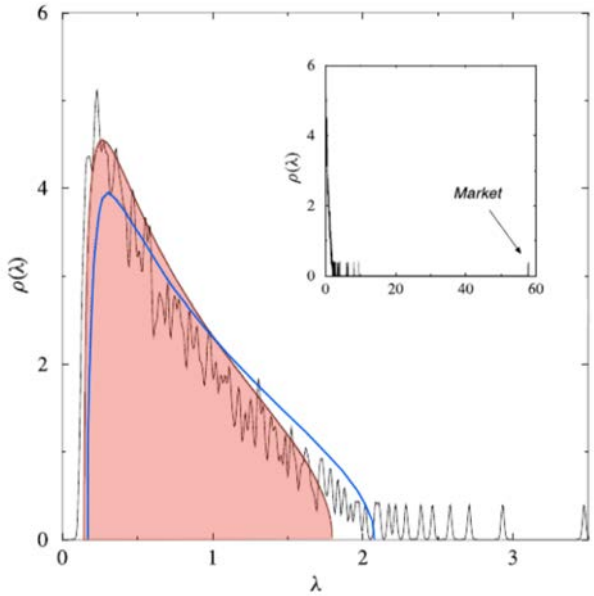


Figure 3.3: Compatibility of the spectral distribution of the cross-correlation matrix of stocks comprised in the S&P 500 during the year 1991-1996 (black line) and the spectral distribution predicted by RMT. The blue curve is the predicted distribution with $\sigma^2 = 0.85$ obtained from assuming that the matrix is purely random except its largest eigenvalue. The pink curve is the RMT's prediction that optimally fits the empirical distribution with $\sigma^2 = 0.74$ obtained from assuming that the matrix is purely random, except 6% the largest of its spectrum [Laloux, 1999].

Let's notice that the theoretical upper bound λ_+ given by Theorem 3.1 is obtained in the limit $N \to \infty$, while an empirical cross-correlation matrix always has finite numbers of N and T. Therefore, when N is large, the empirical largest eigenvalue that relates to the randomness of the cross-correlation matrix may only be approximately the theoretical bound λ_+. Therefore, only eigenvalues that are substantially larger than λ_+ are expected to store information about the exact inter-correlations of the system's components that cannot be explained by the randomness. These significantly large eigenvalues, as well as their associated eigenvectors, attract much attention to understand the collective behavior of a system since they reflect the true correlations of the system's components.

In fact, many economically meaningful results are found from this point of view. One of them is the role of the largest eigenvalue of the cross-correlation matrix of stocks. This eigenvalue is always extremely larger than others regardless of whether the underlying market is a developed one [Laloux, 1999; Plerou, 1999; Rosenow, 2008], a developing one [Nobi, 2013], or an emerging one [Nguyen, 2019b] (see Figure 3.2 – 3.3). Especially, the largest eigenvalue's predominance becomes more outstanding during market crashes in different stock markets [Drożdz, 2000; Nobi, 2013]. We can explain this fact by a general observation that stock price fluctuations tend to change simultaneously in a crisis. Because the stocks become more correlated, the effect of randomness reflected by most of the remaining eigenvalues becomes less. It implies a narrower distribution of the spectrum in the entire spectral distribution of the cross-correlation matrix. By contrast, the distribution of the largest eigenvalues must be wider. A similar observation is verified by Zheng et al. [Zheng, 2012] when they study the sum of the largest eigenvalues

of the cross-correlation matrix contributed by the indexes of ten major sectors in the U.S. and another similar matrix in Europe. They found that a steep increase in this sum can be used as an indicator of systemic risk, in which the largest eigenvalue in the case of U.S. indexes accounts for nearly 60% of the total variation. Besides, the assets' true interactions estimated by the RMT framework are very helpful in risk management and portfolio optimization [Laloux, 2000]. The important roles of the eigenvalues and their associated eigenvectors are explained deeply through PCA. More details are given in the next section.

2 Principal Components of Stock Returns

PCA is a well-known technique of multivariate analysis. The main idea of PCA is to reduce the dimensionality of a dataset consisting of many correlated variables while retaining the present variance of the dataset as much as possible. If we simply fix some initial variables, we can reduce the dimensionality, but a considerable amount of the fixed variables' variations can be lost. Instead, we compose a new variable that is a linear combination of N initial variables with maximum variance. The first new variable is called the first *principal component* (PC). Similarly, we continue finding the second new variable, called the second PC, such that it is still a linear combination of the initial variables, has a maximum variance, and is uncorrelated with the first PC. Generally, the i-th PC is a linear combination of the initial variables such that it has the maximum variance and is uncorrelated with the first $(i-1)$ PCs. As a result, with PCA, we transform a set of correlated random variables into a set of uncorrelated random variables PCs. Furthermore, most of the initial dataset's variation can be expressed by the first m PCs, where $m \ll N$. Consequently, using the first m PCs, we can reduce the initial dimensionality significantly while keeping most of the variations in the initial dataset. From this main idea, it turns out that the PCs relate to the eigenvalues and the eigenvectors of the covariance matrix of the initial variables. This issue is explained in the following paragraphs as proposed in [Jolliffe, 1986]. Moreover, we present the result for our situation, the cross-correlation matrix of stocks.

In particular, for a stock system containing N stocks, let $\mathbf{r} = (r_i)_{i=\overline{1,N}}$, where r_i is the return of stock i, and $\mathbf{r}^* = (r_i^*)_{i=\overline{1,N}}$, where $r_i^* = \frac{r_i}{\sigma_i}$ is the standardized return of stock i. The cross-correlation matrix $\mathbf{C}$ of variables in random vector $\mathbf{r}$, also the cross-correlation matrix of variables in random vector $\mathbf{r}^*$, can be replaced by the sample cross-correlation matrix. According to the Spectral Theorem in linear algebra, because the matrix is real, symmetric and positive semidefinite, it has N nonnegative eigenvalues (not necessary to be distinct), and there is an orthonormal basis of the vector space $\mathbb{R}^N$ consisting of eigenvectors of $\mathbf{C}$. With financial data, we suppose that the rank of matrix $\mathbf{C}$ equals the number of its columns, so $\mathbf{C}$ has N distinct eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_N$. Then, the idea of PCs is concreted as follows:

Definition 3.2. *Let* $\mathbf{u}_i$ $(i = \overline{1,N})$ *be the eigenvector associated with the* i*-th eigenvalue* λ_i *such that* $\|\mathbf{u}_i\| = \mathbf{u}_i\text{'}\mathbf{u}_i = 1$, $\mathbf{A}$ *is the* $N \times N$ *matrix whose columns are* $\mathbf{u}_i s$*, and* $\mathbf{z} = \mathbf{A}\text{'}\mathbf{r}^*$*. Let* z_i *be the* i*th component of* $\mathbf{z}$*. Then,* z_i *is called the* i*-th PC of* $\mathbf{r}$*.*

Let's notice that, in the following statements, when we mention a certain eigenvector associated with a given eigenvalue, this vector is the unit one, i.e., the vector whose elements' sum of

squares equals 1. Furthermore, if $\mathbf{u}$ is a unit eigenvector associated with an eigenvalue, vector $-\mathbf{u}$ is also a unit eigenvector associated with the eigenvalue. For simplicity, we only focus on the unit eigenvector that the number of nonnegative components are not smaller than the number of negative components. The following theorem indicates that PCs defined in Definition 3.2 are consistent with the idea of dimensionality reduction above. Besides, we notice that this theorem is valid for any set of uncorrelated variables.

Theorem 3.2. *PCs of the random vector* $\mathbf{r}$ *satisfy the following properties:*

(i) The PCs are uncorrelated.

(ii) The variance of each PC equals the corresponding eigenvalue. i.e.,

$$\mathrm{Var}\left(z_i\right) = \lambda_i, \qquad \forall i = \overline{1,N} \tag{3.9}$$

(iii) For all linear combination $\mathbf{v'r}^*$ *of variables in* $\mathbf{r}^*$*, where* $\mathbf{v}$ *is a unit vector, the first PC's variance is the largest, i.e.,*

$$\max_{\|\mathbf{v}\|=1} \mathrm{Var}\left(\mathbf{v'r}^*\right) = \mathrm{Var}\left(z_1\right) \tag{3.10}$$

(iv) For all non-zero linear combination $\mathbf{v'r}^*$ *of variables in* $\mathbf{r}^*$*, the first PC is the one that maximizes the sum of squares of the Pearson correlation coefficients with each of stock returns, i.e.,*

$$\max_{y=\mathbf{v'r}^*, y\neq 0} \sum_{i=1}^{N} \left(\rho_{i,y}\right)^2 = \sum_{i=1}^{N} \left(\rho_{i,z_1}\right)^2 \tag{3.11}$$

where $\rho_{i,y}$ *is the correlation coefficient between* y *and* r_i*, and* ρ_{i,z_1} *is the correlation coefficient between* z_1 *and* r_i*,* $i = \overline{1,N}$*.*

Proof. According to Definition 3.2, the i-th PC of $\mathbf{r}$ is

$$z_i = \mathbf{u}_i\mathbf{'r}^*, \quad \forall i = \overline{1,N} \tag{3.12}$$

Besides, since $\mathbf{u}_i$ is the eigenvector associated with eigenvalue λ_i, we have

$$\mathbf{C}\mathbf{u}_i = \lambda_i\mathbf{u}_i, \quad \forall i = \overline{1,N} \tag{3.13}$$

From (3.12) and (3.13), we obtain the following statements:

(i) Remind that the set of eigenvectors $\left(\mathbf{u}_i\right)_{i=\overline{1,N}}$ is an orthonormal basis of the vector space $\mathbb{R}^N$. So, for any numbers $i \neq j$ $\left(i, j = \overline{1,N}\right)$, $\mathbf{u}_i\mathbf{'u}_j = 0$, then

$$\begin{aligned} \mathrm{Cov}\left(z_i, z_j\right) &= E\left(\left(\mathbf{u}_i\mathbf{'r}^*\right)\left(\mathbf{u}_j\mathbf{'r}^*\right)\right) - E\left(\mathbf{u}_i\mathbf{'r}^*\right).E\left(\mathbf{u}_j\mathbf{'r}^*\right) \\ &= E\left(\left(\mathbf{r}^*\right)'\mathbf{u}_i\mathbf{u}_j\mathbf{'r}^*\right) - \left(E\left(\mathbf{r}^*\right)\right)'\mathbf{u}_i.\mathbf{u}_j\mathbf{'}E\left(\mathbf{r}^*\right) = 0 \end{aligned} \tag{3.14}$$

(ii) For any $i = \overline{1,N}$, because the standardized return r_i^* has unit variance

$$\text{Var}\,(z_i) = \text{Var}\,(\mathbf{u}_i\text{'}\mathbf{r}^*) = \mathbf{u}_i\text{'}\mathbf{C}\mathbf{u}_i = \mathbf{u}_i\text{'}\,(\lambda_i\mathbf{u}_i) = \lambda_i\,\|\mathbf{u}_i\| \tag{3.15}$$

Since $\mathbf{u}_i$ is a unit vector, we easily obtain (3.9) from (3.15).

(iii) Because $(\mathbf{u}_i)_{i=\overline{1,N}}$ is an orthonormal basis of the vector space $\mathbb{R}^N$, any vector $\mathbf{v} \in \mathbb{R}^N \backslash \{\mathbf{0}\}$ can be written in a unique way as a linear combination of $\mathbf{u}_i$s, i.e., $\mathbf{v} = \sum_{i=1}^N \alpha_i\mathbf{u}_i$, where α_is $(i = \overline{1,N})$ are real constants that are not all simultaneously equal to zero.

So, for any vector $\mathbf{v} \in \mathbb{R}^N$,

$$\begin{aligned}\mathbf{Var}\,(\mathbf{v}\text{'}\mathbf{r}^*) = \mathbf{v}\text{'}\mathbf{C}\mathbf{v} &= \left(\sum_{i=1}^N \alpha_i\mathbf{u}_i\right)\text{'}\,\mathbf{C}\left(\sum_{i=1}^N \alpha_i\mathbf{u}_i\right) = \sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j\mathbf{u}_i\text{'}\mathbf{C}\mathbf{u}_j \\ &= \sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j\mathbf{u}_i\text{'}\lambda_j\mathbf{u}_j = \sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j\lambda_j\,(\mathbf{u}_i\text{'}\mathbf{u}_j) = \sum_{i=1}^N \alpha_i\alpha_i\lambda_i\,\|\mathbf{u}_i\| \\ &= \sum_{i=1}^N \alpha_i^2\lambda_i \le \lambda_1\sum_{i=1}^N \alpha_i^2\end{aligned} \tag{3.16}$$

Moreover, if $\mathbf{v}$ is a unit vector, then

$$\begin{aligned}1 = \|\mathbf{v}\| = \left(\sum_{i=1}^N \alpha_i\mathbf{u}_i\right)\text{'}\left(\sum_{i=1}^N \alpha_i\mathbf{u}_i\right) &= \sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j\mathbf{u}_i\text{'}\mathbf{u}_j \\ &= \sum_{i=1}^N \alpha_i^2\,\|\mathbf{u}_i\| = \sum_{i=1}^N \alpha_i^2\end{aligned} \tag{3.17}$$

From (3.14) and (3.15), we get that

$$\text{Var}\,(\mathbf{v}\text{'}\mathbf{r}^*) \le \lambda_1 \tag{3.18}$$

Obviously, equation (3.10) is obtained directly from (3.9) and (3.18). The equality sign occurs when $\mathbf{v} = \mathbf{u}_1$.

(iv) For any $i = \overline{1,N}$, the correlation coefficient between z_1 and r_i is

$$\begin{aligned}\rho_{i,z_1} &= \frac{E\,(r_iz_1) - E\,(r_i)\,E\,(z_1)}{\sigma_i\sqrt{\text{Var}\,(z_1)}} = \frac{E\,(r_i\mathbf{u}_1\text{'}\mathbf{r}^*) - E\,(r_i)\,E\,(\mathbf{u}_1\text{'}\mathbf{r}^*)}{\sigma_i\sqrt{\lambda_1}} \\ &= \frac{\mathbf{u}_1\text{'}\,[E\,(r_i\mathbf{r}^*) - E\,(r_i)\,E\,(\mathbf{r}^*)]}{\sigma_i\sqrt{\lambda_1}} = \frac{\mathbf{u}_1\text{'}\mathbf{C}_i\sigma_i}{\sigma_i\sqrt{\lambda_1}} = \frac{\mathbf{C}_i\text{'}\mathbf{u}_1}{\sqrt{\lambda_1}} = \frac{\lambda_1u_1^{(i)}}{\sqrt{\lambda_1}} \\ &= \sqrt{\lambda_1}u_1^{(i)}\end{aligned} \tag{3.19}$$

where $\mathbf{C}_i$ is the i-th column of matrix $\mathbf{C}$ and $u_1^{(i)}$ is the i-th component of vector $\mathbf{u}_1$.

Consequently, because $\mathbf{u}_1$ is a unit vector, we obtain

$$\sum_{i=1}^{N} (\rho_{i,z_1})^2 = \lambda_1 \sum_{i=1}^{N} \left(u_1^{(i)}\right)^2 = \lambda_1 \|\mathbf{u}_1\| = \lambda_1 \tag{3.20}$$

Similarly, for all non-zero linear combination $y = \mathbf{v'r}^*$ of variables in $\mathbf{r}^*$, remind that $\mathbf{v} = \sum_{i=1}^{N} \alpha_i \mathbf{u}_i$, where α_is ($i = \overline{1,N}$) are real constants that are not all simultaneously equal to zero, so

$$\begin{aligned} \rho_{i,y} &= \frac{\mathbf{v'}\left[E\left(r_i\mathbf{r}^*\right) - E\left(r_i\right)E\left(\mathbf{r}^*\right)\right]}{\sigma_i\sigma_y} = \frac{\mathbf{v'C}_i\sigma_i}{\sigma_i\sigma_y} = \frac{\mathbf{C}_i\mathbf{'v}}{\sigma_y} \\ &= \frac{1}{\sigma_y}\left(\mathbf{C}_i\mathbf{'}\sum_{j=1}^{N}\alpha_j\mathbf{u}_j\right) = \frac{1}{\sigma_y}\left(\sum_{j=1}^{N}\alpha_j\mathbf{C}_i\mathbf{'u}_j\right) = \frac{1}{\sigma_y}\left(\sum_{j=1}^{N}\alpha_j\lambda_j u_j^{(i)}\right) \end{aligned} \tag{3.21}$$

where σ_y is the standard deviation of random variable y, and $u_j^{(i)}$ is the i-th component of eigenvector $\mathbf{u}_j$. Because unit vectors $\mathbf{u}_i$s, $i = \overline{1,N}$, are mutually orthogonal, equation (3.21) implies that

$$\begin{aligned} \sum_{i=1}^{N}(\rho_{i,y})^2 &= \frac{1}{\sigma_y^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\alpha_j\alpha_k\lambda_j\lambda_k u_j^{(i)}u_k^{(i)} \\ &= \frac{1}{\sigma_y^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\alpha_j\alpha_k\lambda_j\lambda_k\left(\mathbf{u}_j\mathbf{'u}_k\right) \\ &= \frac{1}{\sigma_y^2}\sum_{j=1}^{N}\alpha_j^2\lambda_j^2\|\mathbf{u}_j\| = \frac{1}{\sigma_y^2}\sum_{j=1}^{N}\alpha_j^2\lambda_j^2 \le \frac{\lambda_1}{\sigma_y^2}\sum_{j=1}^{N}\alpha_j^2\lambda_j \end{aligned} \tag{3.22}$$

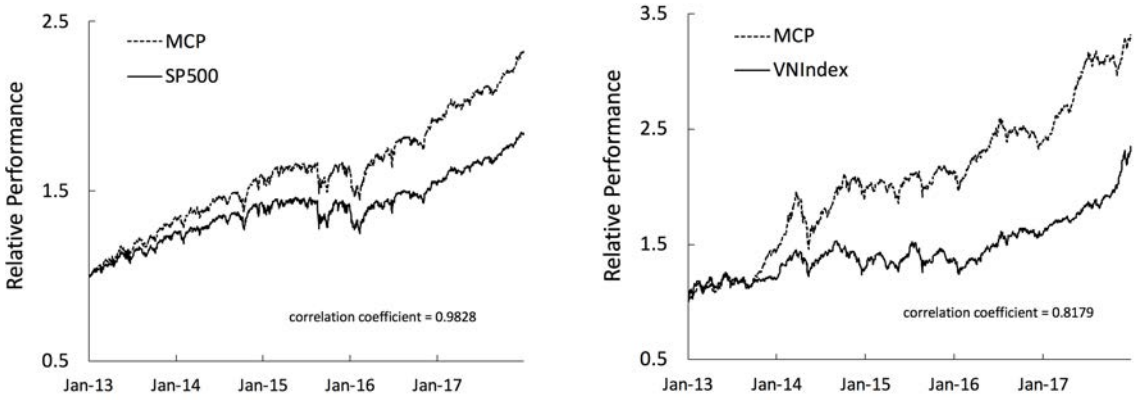
Let's replace $\sigma_y^2 = \mathrm{Var}\left(\mathbf{v'r}^*\right) = \sum_{i=1}^{N}\alpha_i^2\lambda_i$, as demonstrated in (3.16), we get

$$\sum_{i=1}^{N}(\rho_{i,y})^2 \le \lambda_1 \tag{3.23}$$

Obviously, equation (3.11) is obtained directly from (3.20) and (3.23). The equality sign occurs when $\mathbf{v} = \mathbf{u}_1$. ■

According to Theorem 3.2, the first PC explains most of the variance in the data. For investment and risk management, PCs of $\mathbf{r}$, especially the first PC, have remarkable meanings. Indeed, if every nonzero vector $\mathbf{v}$ represents an investment portfolio where the i-th component $v^{(i)}$ of $\mathbf{v}$ is the capital invested to stock i, $\mathbf{v'r}$ is the portfolio's return. Hence, the j-th PC z_j is the return of a portfolio whose the loading of stock i is the fraction $\mathbf{u}_j^{(i)}/\sigma_i$. We call this portfolio is the j-th eigen-portfolio. Theorem 3.2 provides that the eigen-portfolios are uncorrelated to each other, and the variances of their returns equal the corresponding eigenvalues. So, the first eigen-portfolio is the one whose return has the largest variance. It means that this portfolio is the riskiest eigen-portfolio. In general, according to (3.10), the first eigen-portfolio is the riskiest of all portfolios satisfying the standardized condition $\sum\left(v^{(i)}\sigma_i\right)^2 = 1$, where $v^{(i)}$ is the loading

of stock i ($i = \overline{1, N}$). Consequently, the first PC is equivalent to the market factor in the capital asset pricing model (CAPM), a well-known framework for pricing the return of an asset [Plerou, 2002]. Indeed, the market factor of the model monitors the state of the overall stock market as a whole, so it is the primary driver of the stock market returns. Meanwhile, the first PC represents the linear combination of all stock returns that explains most of the returns' volatility because of the dominance of the first eigenvalue in the spectrum. In addition, since the first eigen-portfolio is the portfolio with the largest sum of squares of correlation coefficients with all stock returns, according to the final property given in Theorem 3.2, it correlates the most with the entire market. Thus, in [Nguyen, 2019b], we call it the *most correlated portfolio* (MCP). As a result, this portfolio's return tightly correlates with the return of a market portfolio represented by a capitalization-weighted index. Figure 3.4 plots examples of this linear relationship in two cases of markets having different development levels, the U.S market and the Vietnamese market. The database used to obtain the cross-correlation matrices are the closing prices of the S&P 500 Index's component stocks from the U.S. market and the VN Index's component stocks from the Vietnamese market, for a period of 5 years from 01/01/2013 to 12/31/2017. In these empirical examples, we found a high correlation coefficient between the first eigen-portfolio and the market index, which is about 0.983 for the U.S. market and about 0.818 for the Vietnamese market. Therefore, from the portfolio management's point of view, if one is risk-averse, she can choose the eigen-portfolio associated with the smallest eigenvalue. Meanwhile, if one is bench-marked by the overall market performance, the first eigen-portfolio is a considerable alternative. Obviously, although both the market portfolio and the first eigen-portfolio diversify the idiosyncratic risks of individual stocks, the former bases on the corporations' capitals while the latter is just a pure exposure to systematic risk - the risk affected by many external factors such as monetary, fiscal policy, growth expectations, political risk, regulatory risk and so forth.



(a) The S&P 500 Index vs the most correlated portfolio constructed on the index's components

(b) The VN Index vs the most correlated portfolio constructed on the index's components

Figure 3.4: The relative performance of the simulated most correlated portfolio (dash line) vs. the corresponding market index from 2013 to the end of 2017.

The strong relationship between the largest eigenvalue and the correlations within the entire market is also verified by Nguyen [Nguyen, 2013]. The author found that the eigenvalue ap-

proximates the product of the number of stocks and the average stock correlations. As a result, the largest eigenvalue and its associated eigenvector can offer an alternative way to study the mechanism and evolution of a financial crisis, then construct indicators for the systemic risk. On the other hand, the eigenvectors associated with the remaining eigenvalues, which significantly exceed the upper bound predicted by RMT, also have interesting meanings. It is empirically demonstrated that we can get predictions about a market's clustering from these eigenvalues because they contain information about business sectors or groups of stocks that exhibit common trends [Jiang, 2012; Plerou, 2002; Utsugi, 2004]. Furthermore, from these results, we can compose portfolios that replicate the market index and sector indices such that their returns are uncorrelated.

3 Loadings of the First Principal Component of Stock Returns

Because a few largest eigenvalues, which much deviate from the RMT's upper bound, can reflect the nature of the stock correlations and the PCs associated with these eigenvalues have remarkably economic meanings as discussed in the previous section, a question is how individual stocks contribute to these PCs. Since a PC is a linear combination of the standardized stock returns, we focus on the combination coefficients or the components of the corresponding eigenvector.

Definition 3.3. *For any number $i = \overline{1,N}$, the components $u_i^{(j)} (j = \overline{1,N})$ of eigenvector $\mathbf{u}_i$ are called the loadings of the i-th PC.*

Basing on the loadings of the first PCs associated with the largest eigenvalues deviating from the RMT's upper bound, except λ_1, we can indicate that these PCs mainly contributes by distinct groups such as the group of stocks with large market capitalization, the group of stocks of firms in the same sector or a mixture of some sectors [Jiang, 2012; Plerou, 2002; Utsugi, 2004].

In this section, the first PC's loadings are of our interest. Because the first PC is the market factor which is the primary driver of the stock returns, its loadings provide meaningful information about the common interaction among stocks in the underlying market and main elements that affect the market's collective behavior. Most studies found that the loadings have the same sign [Gopikrishnan, 2001; Nguyen, 2013; Pan, 2007; Plerou, 2002]. Because the sign of a linear regression's coefficient indicates that the corresponding independent variable is positively correlated or negatively correlated with the dependent variable, loadings of the first PC are generally all of the same sign, whereas this is not the case for any of other PCs, confirms that there is a dominant systematic factor totally impacting all or most of stocks in the market. Figure 3.5 illustrates the same sign of the loadings in the U.S market and Vietnamese market. The database used in this figure is the daily stock prices from 01/01/2013 to 12/31/2017, i.e., the same database with Figure 3.4. Figure 3.5 also demonstrates that the loadings are uncorrelated with the market capitalization of the corresponding firms. Therefore, using market capitalization to weigh a stock is not the best way to have a portfolio capturing the common behavior of a stock market. Let's notice that in the figure, we plot the first eigenvector $\mathbf{u}_1$'s components divided by

$\sqrt{\lambda_1}$. This division helps express how much the volatility of an individual stock contributes to one unit of the market factor's volatility. Thus, we have a better visual comparison between the two markets.

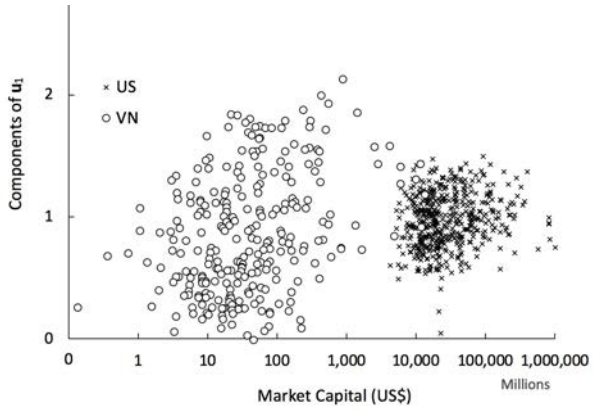


Figure 3.5: Components of $\mathbf{u}_1$ (scaled by $\sqrt{\lambda_1}$) against the market capitalization of the corresponding stocks in the S&P 500 Index and the VN Index in the period from 2013 to the end of 2017.

Obviously, a deep analysis of the loadings is necessary to study the strong common interaction between components of our complex systems - stock markets.

4 Stocks' Influence Reflected through the First Principal Component of Stock Returns

According to property (iv) of Theorem 3.2, the first PC is corresponding to the portfolio that most correlates with the overall market. On the other hand, the correlation between the first PC and the return of any stock is generally positive. Consequently, we can guess that a stock has a positively large loading in the first PC if the stock correlates positively and highly with most stocks in the market. This is demonstrated by our following result:

Theorem 3.3. *If the largest eigenvalue is extremely larger than other eigenvalues, the loading of the first PC on a component nearly relates linearly to the average of correlation coefficients between the stock corresponding to the component and other stocks.*

Proof. By the spectral decomposition of the cross-correlation matrix $\mathbf{C}$, we can express the matrix as $\mathbf{C} = \sum_{i=1}^{N} \lambda_i \mathbf{u}_i \mathbf{u}_i$'. However, because the eigenvectors are unit vectors and the largest eigenvalue λ_1 is dominant remaining eigenvalues, we can approximate the matrix by the first term of the sum, i.e.,

$$\mathbf{C} \approx \lambda_1 \mathbf{u}_1 \mathbf{u}_1\text{'} \tag{3.24}$$

Using (3.24), for any stock i ($i = \overline{1, N}$) the average $\bar{c}_i$ of correlation coefficients between the

stock and others are estimated as follow:

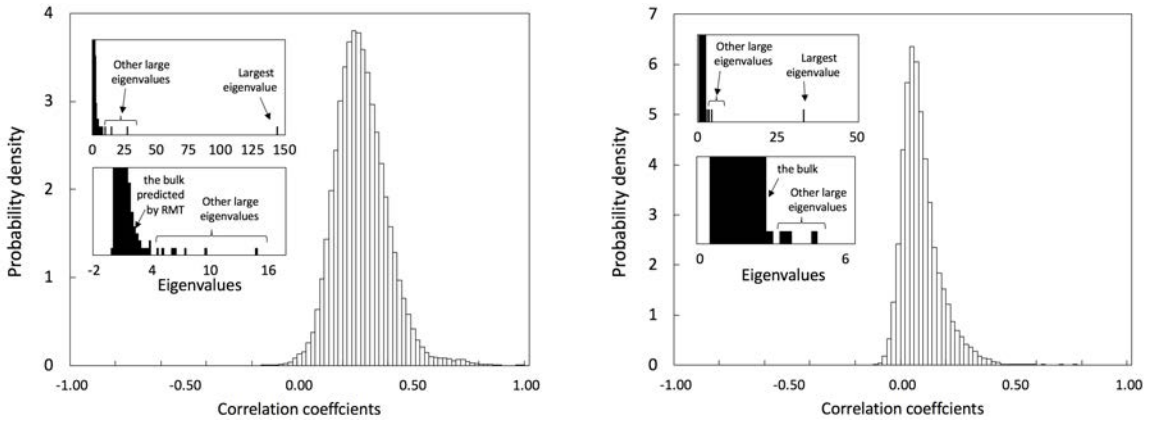
$$\begin{aligned}\overline{c}_i &= \frac{1}{N-1}\sum_{j\neq i} c_{ij} = \frac{1}{N-1}\left(\sum_j c_{ij} - 1\right) \\ &\approx \frac{1}{N-1}\left(\sum_j \lambda_1 u_1^{(i)} u_1^{(j)} - 1\right) \\ &= \frac{1}{N-1}\left(\lambda_1 u_1^{(i)} \sum_j u_1^{(j)} - 1\right) = \frac{N}{N-1}\lambda_1 \overline{u}_1 u_1^{(i)} - \frac{1}{N-1}\end{aligned} \tag{3.25}$$

where $\overline{u}_1$ is the average of the components of $\mathbf{u}_1$. Especially, when $N \to \infty$, we obtain

$$\overline{c}_i \approx \lambda_1 \overline{u}_1 u_1^{(i)} \tag{3.26}$$

■

Theorem 3.3 is valid in our financial context due to the fact that a stock system is always large, and the first eigenvalue of the cross-correlation matrix of stocks is often at least one order of magnitude larger than other eigenvalues (see Figure 3.2 – 3.3). For example, with the database used in Figure 3.4 – 3.5, the number of stocks equals 482 and 274, while the largest eigenvalue λ_1 is 144 and 33.1 for the S&P 500 Index and the VN Index, respectively. Figure 3.6 shows that the value of λ_1 is nearly one order of magnitude larger than others in both the U.S. case and the Vietnamese case. As a consequence, we can see that the components of $\mathbf{u}_1$ are linearly dependent on the average correlations of individual stocks, as mentioned in Theorem 3.3 (Figure 3.7). Furthermore, Figure 3.6 also illustrates the empirically general observation that most stock correlations are positive while their distribution does not always follow a normal distribution.



(a) $\mathbf{C}$ is computed from the S&P 500 Index's components (b) $\mathbf{C}$ is computed from the VN Index's components

Figure 3.6: Probability density of the cross-correlation matrix $\mathbf{C}$ obtained in the period from 2013 to the end of 2017 and its spectrum (upper insert: all eigenvalues; lower insert: all eigenvalues excluding the largest).

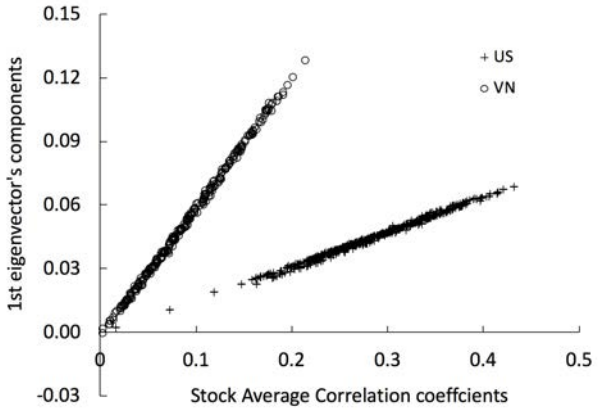
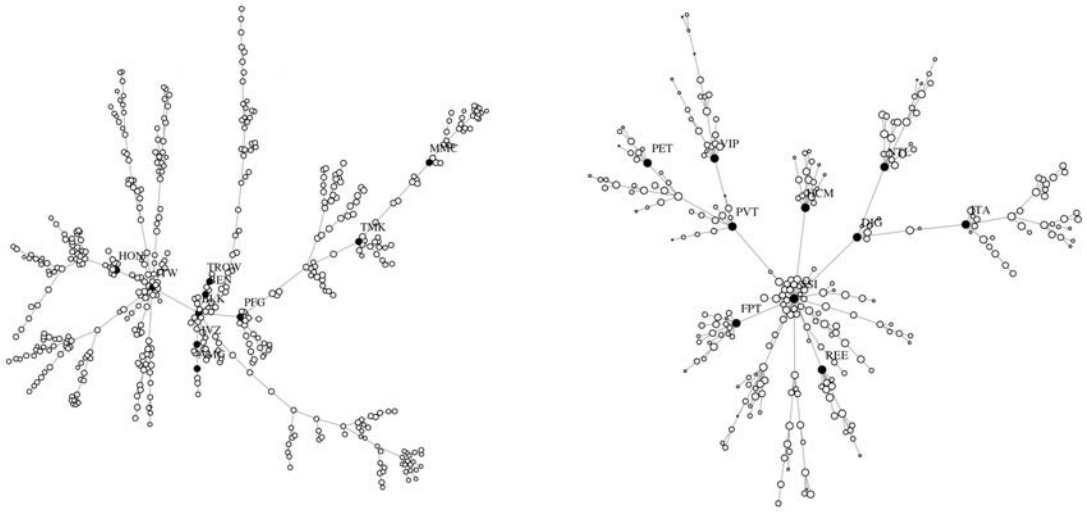


Figure 3.7: Relationship between the first eigenvector's components and the average correlation coefficient of the corresponding stocks for the S&P 500 Index and the VN Index in the period from 2013 to the end of 2017.

As a result, if a stock is highly correlated with others on average, the loading of the first PC on the stock is large. Since the PC is the market factor that most correlates with the overall market and explains most of the variance of stock returns, the stock must have a strong influence on the market stability. Indeed, for the database used in the previous figure, Figure 3.8 shows that stocks with the largest loadings of the first PC are at the center hubs of the MST of the corresponding cross-correlation network. It means that the stocks play an important role when a systemic breakdown happens. Meanwhile, stocks with small loadings of the first PC are leaf nodes mostly. The various magnitudes of the loadings of the first PC on different stocks are illustrated by the diversification of the node sizes in Figure 3.8b. This observation is not obvious in Figure 3.8a because the figure only focuses on selected stocks in the S&P 500 Index, which is mainly composed of about 500 blue-chip stocks out of nearly 3000 stocks listed on the NYSE. Meanwhile, the stocks composed in VN Index are all of the stocks listed on the HSX. Hence, the influence of each stock on the entire market is clearly decentralized.



(a) **C** is computed from the S&P 500 Index's stock components

(b) **C** is computed from the VN Index's stock components

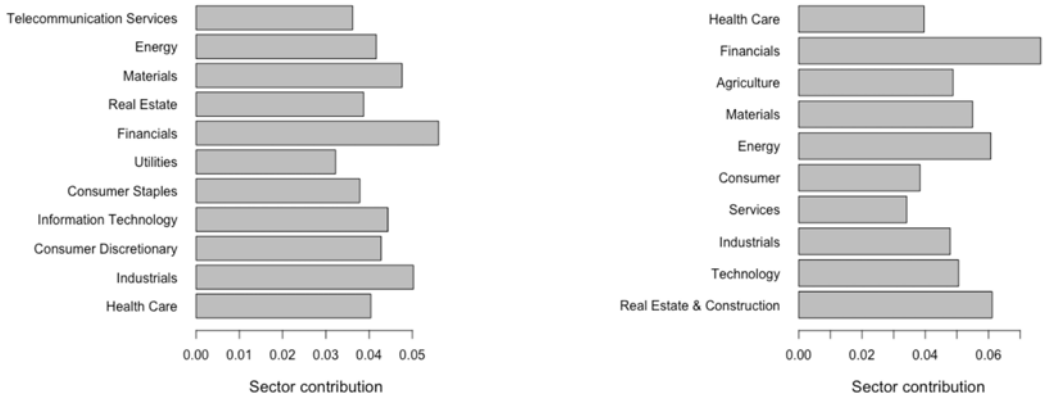
Figure 3.8: The MST obtained in the period from 2013 to the end of 2017 with node size as the logarithm of the first PC's loading on the corresponding stock (tickers of the stocks corresponding to the top 10 loadings are shown and their corresponding nodes are filled).

On the other hand, we realize that the largest loadings of the first PC correspond to stocks of financial services (brokerage or investment advisory firms) in both these cases. Especially, let's measure the contribution a_s of a business sector s for the first PC by the average loading of the first PC on stocks belonging to the sector[1], i.e.,

$$a_s = \frac{1}{n_s} \sum_{j=1}^{N} u_1^{(j)} \delta(j, s) \tag{3.27}$$

where $\delta(j, s) = 1$ if stock j belongs to sector s and $\delta(j, s) = 0$ otherwise; n_s is the number of stocks belonging to sector s. Then, we found that the financials sector's contribution for the first PC is the largest (see Figure 3.9). As a consequence, Figure 3.8 and Figure 3.9 show the essentials of financial companies in contributing to the common behavior of a stock market. This helps explain why a company of financial services can rise to a super-hub when the MST network approaches its unstable state – a star-like structure – as discussed in Chapter 2.

[1]We use the first PC's loadings instead of the square of the loadings as suggested in [Conlon, 2007; Coronnello, 2005] to evaluate sector contributions. This choice comes from the positive values of most of the loadings. Moreover, we want to evaluate the co-movement between a sector and the entire market rather than just compute how much the sector can influence the market positively or negatively on average.



(a) The first PC is constructed from the S&P 500 Index's stock components

(b) The first PC is constructed from the VN Index's stock components

Figure 3.9: Sector contributions for the first PC of stock returns obtained in the period from 2013 to the end of 2017.

In brief, because the first PC of stock returns are most correlated with the overall market and equivalent to the market factor, its loadings can reflect the influence of a stock or a group of stocks on the overall market. We demonstrate, both theoretically and empirically, the positive correlation between the loadings and the average correlations of individual stocks. We also found that financial companies tend to have prominent loadings in the market factor. These results are very useful in investment and managing systemic risk. Especially, the relationship between the loading of the first PC on a stock and the role of the stock in the MST network shows that the PC contains meaningful information about the MST's structure. It means that PCA can be a useful method to analyze the MST network. Moreover, because PCA encodes the whole data of a cross-correlation matrix, it is expected to provide more information than the MST analysis.

Chapter 4

Cascading Failure in Financial Systems and Its Pretopological Model

Objective

The collective behavior of a complex system sometimes emerges from a cascading failure caused by the strong relationships between the system's components. In this chapter, we use pretopology theory to construct a framework that can capture the cascading failure's evolution. The framework takes into account the relationship between each pair of stocks, the relationship between each stock and a group, as well as the crowd effect. It helps predict the impact magnitude of a stock's price fluctuation on others when the stock is in a negative price trend and also helps predict stocks not affected by the stressed stock. Then, we apply our framework to test the effect of the common stock of Merrill Lynch & Co. on others on the NYSE. We also compare the effectiveness of our pretopological framework with ones of the MST network and the correlation-based threshold network.

Contents

1 Cascading Failure in Complex Systems

In Chapter 2, we studied the complex networks or the graph representations of a complex system basing on the inter-correlation between its components and discussed the resilience of a complex network under intentional attacks and the random failure of nodes. Being different from these failures whose percolation processes only depend on the role of each component in the system's structure or occur randomly, sometimes a complex system can be damaged by the fault of just one or a few components because of the strong relationship between its components. In particular, when one or a few components fail, some components most relating to the failed components are first infected. Then, these components continue to trigger the failure of others and so on. This diffusion process through the relationship of a system's components is called the cascading failure. It occurs in different complex systems such as computer systems, transportation systems, power systems, social systems, biological systems [Schäfer, 2018; Sun, 2012]. In these systems, when a components fails, some others must compensate for the fail components. This leads these components to overload, then infect others (see Figure 4.1). One of the well-known cascading failures is the India blackouts that happened in July 2012. In this event, a circuit breaker tripped, and power failures cascaded through the grid such that over 400 million people were affected [1]. Another example is the Internet congestion collapse that happened in October 1986. At that time, traffic was rerouted to bypass malfunctioning routers, eventually leading to an avalanche of overloads on other routers that were not equipped to handle the extra traffic. Consequently, the system's performance becomes extremely poor when the connection speed between the two places of only 200 meters apart dropped by a factor of 100 [Crucitti, 2004a]. The present Covid-19 pandemic is also an illustration of the cascading failure because the contagion passes from a person to whom having close contacts and rapidly spreads to the community. Similarly, the cascading failure also appears in financial systems when a financial institution's failure may cause its counterparts to fail and evenly spread throughout the market. For example, a decline of contracts in the equity market leads to a U.S trillion-dollar stock market crash in May 2010 [2].

[1] See details on the site: https://en.wikipedia.org/wiki/2012_India_blackouts
[2] See details on the site: https://en.wikipedia.org/wiki/2010_flash_crash

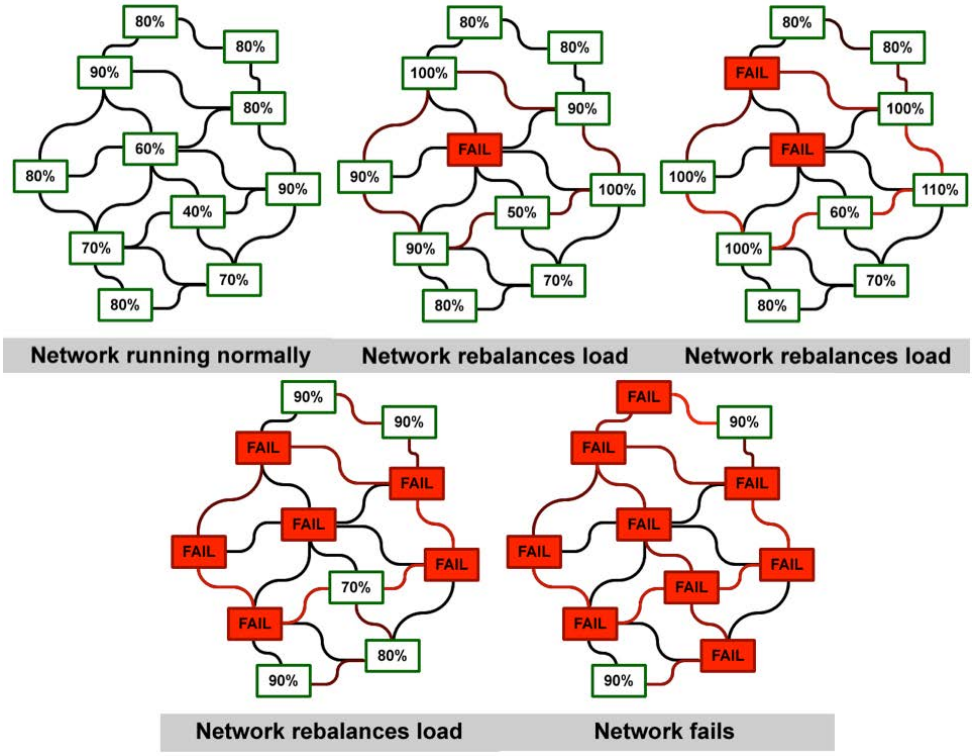


Figure 4.1: Cascading failure in a network. [3]

Because of its serious impact, the cascading failure becomes an important research subject. One of the first approaches to get more analysis about its mechanism and impacts is modeling how it happens inside a system. In physical systems such as computer systems and transportation systems, by considering the systems as complex networks, this failure can be modeled by the contagion of a node's failure to its neighborhood. This model uses a flow concept that goes through network connections and damages a node if the flow's intensity of the node is higher than the node's capacity due to the failures of its neighbors [Crucitti, 2004a; Schäfer, 2018]. However, this idea is not suitable to apply for a stock system under its correlation-based network because this network is complete but not all stock correlations are high enough to make stocks affect each other. The MST of the network is also inappropriate since it just takes into account the most probable path for the failure spreading from a node to the entire market instead of reflecting all potential paths of the infection. For this point of view, the correlation-based threshold network is more convenient because, with a suitable threshold, it keeps more connections that correspond to meaningful stock correlations in the original correlation-based network. However, using this threshold network, we still have trouble in modeling the cascading failure from a group of stocks to another stock because the network only reflects the relationship between each pair of stocks. Meanwhile, a stock's price is often affected by the dramatic changes of a group of other stocks' prices rather than by the price fluctuation of only one stock. This can be caused by the changes of some common factors with influence the entire market or a certain part of the market, such as stocks belonging to the same sector. Another reason is the crowd effect that makes more

[3] https://en.wikipedia.org/wiki/Cascading_failure

and more investors, governed by greed and fear, engage in buying or selling stocks frantically and consequently creates economic bubbles or stock market crash. Therefore, in this chapter, to model the cascading failure, in addition to continuing to focus on high correlations between stocks as the correlation-based threshold network, we construct a new model that takes into account both the relationship between each pair of stocks and the relationship between each stock and a group. The new model is based on pretopology theory with more details given in the next section.

2 Pretopology Theory

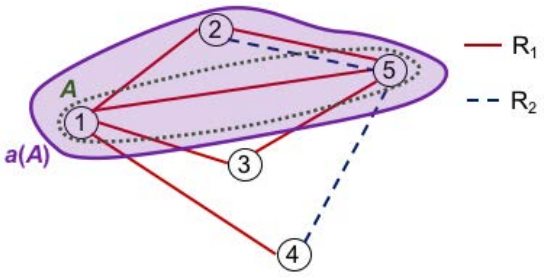
Pretopology theory was developed with objective to tracking the evolution of a diffusion process and how it contributes to the final result [Belmandt, 2011]. In network analysis, a diffusion model is usually defined via node's neighbors and follows a diffusion rule. Similarly, in pretopology theory, we start with a map represented the rule – pseudoclosure. Let E be a nonempty set and $\mathcal{P}(E)$ be the set of all its subsets.

Definition 4.1. *We call pseudoclosure defined on E any map a from $\mathcal{P}(E)$ into $\mathcal{P}(E)$ such that:*

(i) $a(\varnothing) = \varnothing$, *and*

(ii) $\forall A \subset E, A \subset a(A)$.

Obviously, pseudoclosure is a more general concept defined simply as any expansion rule for nonempty subsets of a system's elements instead of basing only on the node connections as diffusion rules studied in network analysis. With pseudoclosure, we can consider multi-relation between the system's elements or between an element and a group of elements to model an appropriate proximity concept for each particular diffusion problem. Figure 4.2 illustrates an example of a pseudoclosure a defined by two binary relations R_1 and R_2 on five elements such that $a(A)$ is the union of A and the set of elements that have relation R_1 with all elements of A and have relation R_2 with at least one of elements of A. For instance, if $A = \{1, 5\}$, $a(A) = \{1, 2, 5\}$.



$$a(A) = A \cup \left\{ i = \overline{1,5} | (\forall j \in A, i\mathrm{R}_1 j) \wedge (\exists j \in A, i\mathrm{R}_2 j) \right\}$$

Figure 4.2: A pseudoclosure defined by two relations R_1 and R_2, to model the proximity concept between an element and a group of elements.

By contrast, in order to model the dilution of a given set, we need a rule that helps to remove the most irrelevant elements with the remaining ones. In pretopology theory, any map that transfers a set to a smaller one or keeps the set as itself is defined as the below concept:

Definition 4.2. *We call interior defined on E any map i from $\mathcal{P}(E)$ into $\mathcal{P}(E)$ such that:*

(i) $i(E) = E$, *and*

(ii) $\forall A \subset E, i(A) \subset A$.

For simplicity, an interior defined on E is usually defined as the c-duality of a pseudoclosure defined on E, i.e.:

Definition 4.3. *We call interior defined on E any map i from $\mathcal{P}(E)$ into $\mathcal{P}(E)$ such that:*

$$i(A) = E \backslash a(E \backslash A), \qquad \forall A \subset E \tag{4.1}$$

where a is a pseudoclosure defined on E.

In this study, we consider i as a c-duality of a. Then, a pretopological space is defined simply as the following:

Definition 4.4. *A pretopological space is a pair $(E, a(.))$ where a is a pseudoclosure defined on the nonempty set E.*

The first advantage of a pretopological space is that the successive computations of a pseudoclosure to a given set A helps model the evolution of a dilation process starting from A. Meanwhile, the result of applying the interior successively to A can model a dilution process starting from the set in individual steps (see Figure 4.3). In addition, we also take into account the limits of these processes if they exist to get the processes' final results.

Definition 4.5. *Given a pretopological space $(E, a(.))$, for any subset A of E,*

(i) A is said to be a closed subset of E if and only if $A = a(A)$.

(ii) A is said to be an open subset of E if and only if $A = i(A)$.

Definition 4.6. *Given a pretopological space $(E, a(.))$, for any subset A of E,*

(i) we call closure of A, denoted by $\mathbf{F}(A)$, the smallest closed subset of E that contains A if the subset exists.

(ii) we call opening of A, denoted by $\mathbf{O}(A)$, the biggest open subset of E that is included in A if the subset exists.

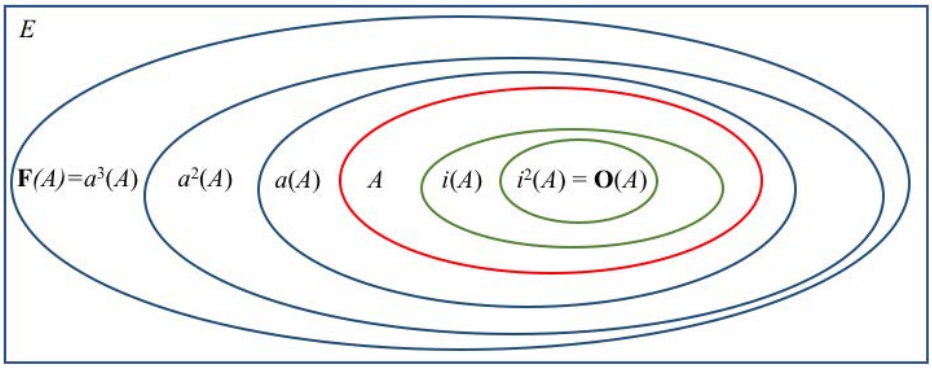


Figure 4.3: Successive computations of pseudoclosure and interior.

Especially, any topological space is a pretopological space whose pseudoclosure is the closure operator. In addition, a pretopological space can be considered an extension of a usual graph or even a hypergraph, a generalization of a graph in which an edge can join any number of nodes. Indeed, according to [Dalud-Vincent, 2011], it is always possible to associate a pretopological space to a hypergraph; furthermore, we can construct a pretopological space from a given hypergraph such that their connectivity properties are equivalent:

Proposition 4.1. *For any hypergraph (E, H) where nonempty set E is the set of nodes and $H \subset \mathcal{P}(E)\backslash\{\varnothing\}$ is the set of hyperedges, let R is a binary relationship defined on H by, for any hyperedges $e_i,\ e_j \in H$,*

$$e_i R e_j \Leftrightarrow e_i \bigcap e_j \neq \varnothing \tag{4.2}$$

Let a be a map defined on H by

$$a(A) = A \cup \{e \in H \,|R_e \cap A \neq \varnothing\}, \qquad \forall A \subset H \tag{4.3}$$

where $R_e = \{e^\star \in H \,|eRe^\star\ \}$. Then, we have:

(i) $(H, a(.))$ is a pretopological space.

(ii) Assume that there is no isolated node in (E, H), then (E, H) is a connected hypergraph if and only if $(H, a(.))$ is a strongly connected pretopological space, i.e., for any nonempty set $A \subset H, \mathbf{F}(A) = H$.

Proof.

(i) Obviously, from (4.3) we can see that $A \subset a(A)$ for any subset A of H. Besides, since $R_e \cap \varnothing = \varnothing$ for any $e \in H$, we have $a(\varnothing) = \varnothing$. Thus, a is a pseudoclosure defined on H, i.e., (H, a) is a pretopological space.

(ii) The hypergraph (E, H) is connected if and only if any two of its distinct nodes connect to each other. Consequently, if A is a subset of H, for any $e \in A$ and for any $e^\star \in H\backslash A$, there is at least one path from a node in e to a node in $e^\star$, i.e., there exists a sequence

of nodes $(u_i)_{i=\overline{0,k}} \subset E$ and a sequence of hyperedges $(e_i)_{i=\overline{1,k}} \subset H$ such that $u_0 \in e$, $u_k \in e^\star$, and $\{u_0, u_1\} \subset e_1, \{u_1, u_2\} \subset e_2, \ldots, \{u_{k-1}, u_k\} \subset e_k$. Hence, we have eRe_1, $e_1Re_2, \ldots, e_{k-1}Re_k$ and $e_kRe^\star$. Then, $e^\star$ must belong to the set obtained from at most k successive computations of a to the set A in pretopological space $(H, a(.))$. This implies that $e^\star \in \mathbf{F}(A)$. So, $H \backslash A \subset \mathbf{F}(A)$. Besides, since $A \subset \mathbf{F}(A)$, we get $H \subset \mathbf{F}(A)$. Hence, $\mathbf{F}(A) = H$.

By contrast, let assume that $(H, a(.))$ is a strongly connected pretopological space. For any pair of distinct nodes $u, v \in E$, let e and $e^\star$ are hyperedges in H such that $u \in e$ and $v \in e^\star$. Because $\mathbf{F}(\{e\}) = H$, so $e^\star \in \mathbf{F}(\{e\})$. Then, there exists a number k such that $e^\star \in a^k(\{e\})$. It means that we can find a sequence of hyperedges $(e_i)_{i=\overline{1,k}} \subset H$ such that eRe_1, e_1Re_2, $\ldots, e_{k-1}Re_k$ and $e_kRe^\star$. The sequence is the path connecting two nodes u and v. Therefore, (E, H) is a connected hypergraph.

■

By contrast, a pretopological space is not always associated with a graph or a hypergraph. For example, for the pretopological space given in Figure 4.2, we have to construct two graphs to model two relations R_1, R_2 of nodes, respectively.

As a result, graphs, hypergraphs, and topological spaces are particular cases of pretopological spaces. Furthermore, pretopology theory still allows considering basic concepts of graph theory such as degree, path, closeness centrality, betweenness centrality, and basics concepts of topology theory such as closure, opening, neighborhood, filter [Belmandt, 2011; Levorato, 2014]. Therefore, when graphs and hypergraphs are inappropriate to study a system including elements' multi-relation and relationships between a group of elements and one element, or when the evolution of dilation and/or diminution process is the research target, pretopology theory can be a valuable solution. It is also used in different fields to investigate complex systems, for instance, pollution modeling [Ben-Amor, 2010; Lamure, 2009], macroeconomics analysis [Auray, 1979], image analysis [Bonnevay, 2009], Smart Grid model [Guérard, 2015; Petermann, 2012]...

3 Pretopological Framework of Cascading Failure in Stock Markets

For a complex system, if its elements are related to each other by a valued relation, one of the common ways to build a pseudoclosure a is that for any subset A of the elements, $a(A)$ is composed of A and all other elements whose relation to A is greater than a threshold. In particular problems, the relation of an element x to a group can be defined by different measures, for instance, the sum, the mean, the maximum, or minimum of relations between x and elements of the group. In this way, we transfer the valued relation between pairs of elements to a binary relation between one element and a group of other elements. This supports the main target: constructing a proximity concept so that the spread of a special behavior from a part of the system to the rest can be reflected as much as possible. For example, in [Ben-Amor, 2010; Lamure, 2009], to construct a stochastic pretopological model of the spreading pollution in a

geographic area, the authors use a lower limit for the concentration of an emitter's pollutants on a point to verify whether the emitter contaminates the point. Then, the pseudoclosure is built by expanding a given subset of the area with all points polluted by at least one emitter of the subset. Also, in [Auray, 1979], to establish a pretopological space to reflect the input-output relation between activity sectors of an economy, a lower threshold of this relation is considered to define a binary reflexive relation for each pair of sectors. Then, the pseudoclosure is built to expand a group of sectors by other sectors having the binary relation with at least one sector of the group.

Similarly, in a stock system, to model the cascading failure of a stock's price shock, we choose a lower limit of stock correlation to consider if stocks correlate highly enough to each other so that a stock's price fluctuation can affect others. Base on the threshold, we construct a pseudoclosure to determine the closeness of a stock to be affected by the price changes of a group of other stocks. In contrast to the works in [Auray, 1979; Ben-Amor, 2010; Lamure, 2009], we don't use two assumptions: a constant threshold of the relation between a stock and a group, the spreading condition that the stock only needs to correlate highly enough to at least one element of the group. The reason is that we want a model describing the group effect, which usually takes place in financial markets due to the impact of some common factors such as the market factor, sector factor... The three following assumptions are used in our model:

(i) If a stock has a price shock, stocks highly correlate with it are directly influenced.

(ii) The impact of a group on an outside stock is higher if the group is larger.

(iii) A change in the size of a group makes the group have more impact on an outside stocks if the group's size is larger.

The second assumption means that when more stocks have significant changes, the probability that the stocks affect the price fluctuations of others is also higher. Especially if the number of negative stocks is large enough, this can create a dramatic crash in the entire market due to the psychological fear of investors. Then, a stock is also caught up in the crash despite its low correlations to other stocks. Therefore, the last assumption is essential: the size increase of a large group makes more impact than the size increase of a small group. Consequently, the last two assumptions imply that the impact threshold of a group of stocks is a decreasing and concave function of the group size. In addition, because we adjust the impact threshold according to the group size, to identify if a stock is "close" enough to a group, we compare the impact threshold of the group with the average correlation coefficient between the stock and the group's elements. As a result, if E be the set of all listed stocks in a stock market, and N be the number of these stocks, we use the following pseudoclosure for our cascading failure problem:

Proposition 4.2. *Let f be a decreasing and concave function from $[1, N]$ into $[0, 1)$. Let a be a map from $\mathcal{P}(E)$ into $\mathcal{P}(E)$ such that $a(\varnothing) = \varnothing$ and*

$$a(A) = A \cup \left\{ k \in E \backslash A \middle| \frac{1}{\|A\|} \sum_{j \in A} c_{jk} \geq f(\|A\|) \right\}, \ \forall A \in \mathcal{P}(E) \backslash \{\varnothing\} \tag{4.4}$$

where $\|A\|$ is the size of A. Then, $(E, a(.))$ is a pretopological space with the interior i such that $i(E) = E$ and

$$i(A) = \left\{ k \in A \middle| \frac{1}{N - \|A\|} \sum_{j \in E \backslash A} c_{jk} < f(N - \|A\|) \right\}, \ \forall A \in \mathcal{P}(E) \backslash \{E\} \tag{4.5}$$

Proof. It's clearly that a is a pseudoclosure by Definition 4.1, so $(E, a(.))$ is a pretopological space.

According to Definition 4.3, we have $i(A) = E \backslash a(E \backslash A)$ for any $A \subset E$. So, if $A = E$, we get $i(E) = E \backslash a(E \backslash E) = E \backslash a(\varnothing) = E \backslash \varnothing = E$.

Conversely, if $A \neq E$, since $E \backslash A \neq \varnothing$, we obtain

$$\begin{aligned}
i(A) &= E \backslash \left((E \backslash A) \cup \left\{ k \in A \middle| \frac{1}{\|E \backslash A\|} \sum_{j \in E \backslash A} c_{jk} \geq f(\|E \backslash A\|) \right\} \right) \\
&= A \cap \left(E \backslash \left\{ k \in A \middle| \frac{1}{\|E \backslash A\|} \sum_{j \in E \backslash A} c_{jk} \geq f(\|E \backslash A\|) \right\} \right) \\
&= \left\{ k \in A \middle| \frac{1}{\|E \backslash A\|} \sum_{j \in E \backslash A} c_{jk} < f(\|E \backslash A\|) \right\} \\
&= \left\{ k \in A \middle| \frac{1}{N - \|A\|} \sum_{j \in E \backslash A} c_{jk} < f(N - \|A\|) \right\}.
\end{aligned}$$

■

Thus, in our cascading problem, if we consider A as a set of failed elements, the pseudoclosure defined by formula (4.4) satisfies the three assumptions of our model for the impact of A on other stocks. On the other hand, when $A \neq E$, the corresponding interior $i(A)$ is an extenuation of A such that remaining stocks have small average correlations with the stocks outside A. The remaining condition is determined by an upper limit which is equal to $f(N - \|A\|)$. So, differently from the pseudoclosure a, the interior's threshold is an increasing and convex function of the group size. Consequently, $i(A)$ reflects the A's subset of stocks mostly affected by stocks of A more than others; moreover, the stock impact synchronizes with the size of A. Especially when A is E, it means that the entire market is totally broken. As a result, with this model, by finding the closure of a group of stocks, we can predict the impact magnitude of these stocks' price fluctuations on others when these stocks are in a negative price trend. Inversely, the opening of the compensation of the group can help predict the stocks not affected by the group's negative price trend. With the pretopological space considered in Proposition 4.2, we can find the closure and opening of a set of stocks by Algorithm 6 and Algorithm 7, respectively.

Algorithm 6 Compute the closure of a set in the pretopological model of stocks.

Require: stock market E, subset A of E, decreasing and concave function f
procedure CLOSURE(E, A, f)
  $pseudoclosure \leftarrow A$
  $A \leftarrow \varnothing$
  while $pseudoclosure \neq A$ **do**
    $A \leftarrow pseudoclosure$
    $pseudoclosure \leftarrow A \cup \left\{ k \in E \backslash A \middle| \frac{1}{\|A\|} \sum_{j \in A} c_{jk} \geq f(\|A\|) \right\}$
  end while
  $closure \leftarrow pseudoclosure$ ▷ Output
end procedure

Algorithm 7 Compute the opening of a set in the pretopological model of stocks.

Require: stock market E, subset A of E, decreasing and concave function f
procedure OPENING(E, A, f)
  $interior \leftarrow A$
  $A \leftarrow E$
  while $interior \neq A$ **do**
    $A \leftarrow interior$
    $interior \leftarrow \left\{ k \in A \middle| \frac{1}{N - \|A\|} \sum_{j \in E \backslash A} c_{jk} < f(N - \|A\|) \right\}$
  end while
  $opening \leftarrow interior$ ▷ Output
end procedure

4 Empirical Results on the NYSE

In this section, we empirically study how the pseudoclosure introduced in Proposition 4.2 can model the spreading of a price shock in a real stock market, the NYSE, and vice versa, how the corresponding interior can help to predict stocks not affected by the shock.

4.1 Database

We examine the cascading failure starting from the common stock of Merrill Lynch & Co., which traded under the ticker MER on the NYSE, to others composed in the S&P 500 Index. The cooperation is selected because of its important position in the U.S market for decades. It was a bank with a remarkable brokerage network such that it could move stocks, securities, and bonds base on its interests and those of its clients before the mortgage crisis of 2007. However, it lost the position after serious losses from the second of 2007 and was finally acquired on 12/31/2008 [4]. We denote MER as stock i_0.

The day when i_0 got a considerable decrease in its price is defined as when the price fell more than 70% within one year before its consolidated day. The day is 09/12/2008 and denoted by t_0 (see Figure 4.4). Let's consider the set E of all components of the S&P 500 Index to represent the U.S. stock market. From the viewpoint that E is a complex system, we say that the stock fails at time t_0. We use the daily closing prices of all research stocks in 2 years before t_0 to calculate their correlations. To give a backtest for the efficiency of our pretopological framework in modeling the cascading failure starting from i_0, we use the daily closing prices of the index's components in 6 months from t_0, i.e., from 09/12/2008 to 03/12/2009. Similarly, in this period, a component of the index is considered to be failed if its price drops more than 70%. The set of

[4] See details on the site: https://en.wikipedia.org/wiki/Merrill_Lynch_%26_Co

failed stocks in this period is denoted by H. The size of E and the size of H are 489 and 102, respectively.

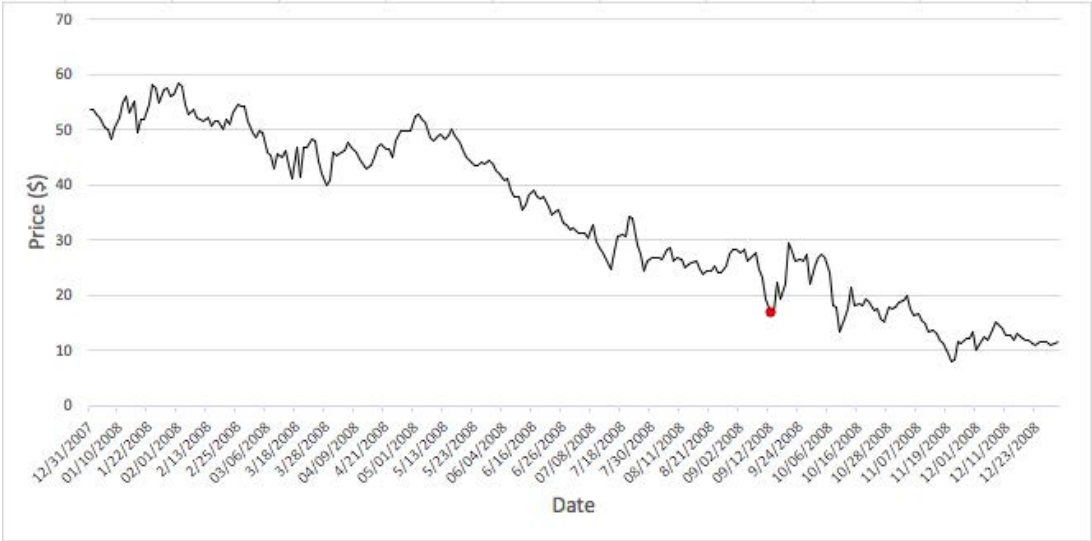


Figure 4.4: Adjusted daily closing price of MER from 12/31/2007 to 12/31/2008 (the point corresponding with the price at time t_0 is colored in red).

4.2 Research Method

In order to construct the pretopological space $(E, a(.))$ as mentioned in Proposition 4.2, we need a decreasing and concave function f from $[1, N]$ into $[0, 1)$ to verify the impact threshold of a group of elements on another element. In [Nguyen, 2019a] and [Nguyen, 2021b], we try different forms for this threshold function. In [Nguyen, 2019a], the following function is used:

$$f(x) = \theta \left(\frac{N}{N-1}\right)^{\gamma} \left(\frac{1}{x-N-1}+1\right)^{\gamma}, \quad \forall x \in [1, N] \tag{4.6}$$

where $\gamma > 0$ and $0 < \theta < 1$. Meanwhile, in [Nguyen, 2021b], we use the below function:

$$f(x) = 1 - \theta e^{\gamma x}, \quad \forall x \in [1, N] \tag{4.7}$$

where $0 < \theta < 1$ and $0 < \gamma < -N^{-1} \ln \theta$.

To predict stocks influenced by the failure of i_0, we use the closure of i_0. We quantify the prediction's efficiency by two measures: the precision and recall of the prediction. In particular, the prediction's precision is the fraction of failed stocks in $\mathbf{F}(\{i_0\})$ except i_0, while its recall is the fraction of failed stocks predicted by $\mathbf{F}(\{i_0\})$. Furthermore, the successive computations of pseudoclosure to get $\mathbf{F}(\{i_0\})$ are expected to help study the evolution of the failure's propagation. On the other hand, according to the meaning of the interior mentioned in Section 3, the opening of $E \backslash \{i_0\}$ is expected to predict stocks not affected by the failure of i_0.

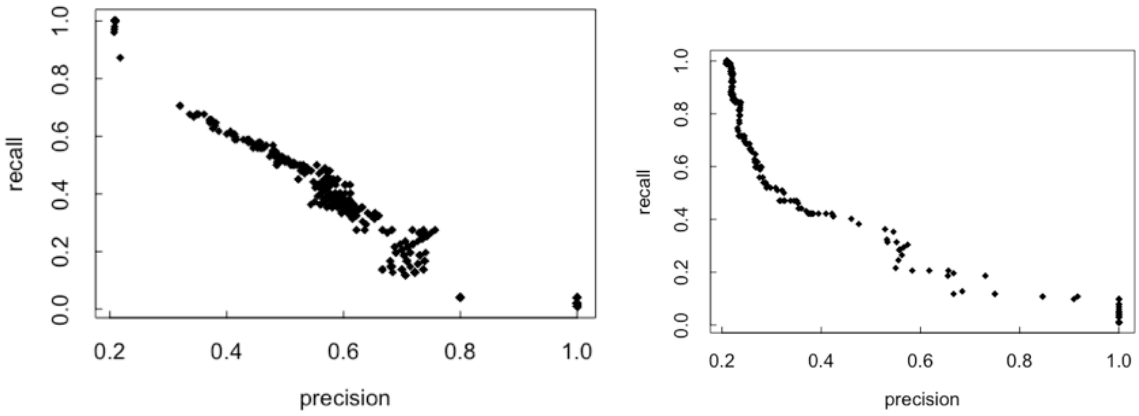
4.3 Transmission Process of a Price Shock

We found that it doesn't matter what form of threshold function, our pretopological framework with appropriate parameters can model the failure's transmission better than the common stock networks mentioned in Chapter 2.

Indeed, with $\gamma = 0$, we have a simple pretopological model for the transmission of the failure of i_0 with a constant threshold of the group impact. Another stock is included to $a(\{i_0\})$ if the correlation between it and i_0 is greater than or equal to the threshold. In general, regardless of the size of a given group of failed stocks affected by i_0, another stock j is affected by these stocks' behaviors if the average correlation between it and these stocks is not smaller than the threshold. Then, there exists a stock i of the group such that the correlation between i and j is not smaller than the constant. Thus, any stock j of $\mathbf{F}(\{i_0\})$ is a constituent of the correlation-based threshold network with the same threshold; moreover, there exist a path connecting i_0 and j in the network. As a result, the set of stocks affected by i_0 in our pretopological model is a subset of a cluster containing i_0 in this network. However, this network only focuses on the relationship between any two stocks but neglects the role of the group impact. Under the impact, the larger number of failed stocks implies the higher ability that these stocks' negative fluctuations influence another stock's fluctuation. The impact is represented by the positive γ in our model. This parameter plays an important role in deciding the flexure of the graph of f. With a given θ, the larger γ is, the larger the magnitude of the instantaneous rate of change of the function is because of its concavity. Therefore, we use γ to adjust the change of the group impact corresponding to the change of the group size.

Let's consider the dilation process from $\{i_0\}$ to $\mathbf{F}(\{i_0\})$ in the pretopological space defined in Proposition 4.2. Note that we want to use $\mathbf{F}(\{i_0\})$ to predict stocks affected by the failure of i_0. In [Nguyen, 2019a] and [Nguyen, 2021b], we found that although different values of θ and γ lead to different levels of the prediction's efficiency as illustrated in Figure 4.5a, our pretopological framework is better than the MST of the correlation-based network in modeling the cascading failure caused by the price shock of i_0. Indeed, when using the MST to model the cascading failure, the propagation of the shock of i_0 to another stock j must spread through the path connecting the two stocks. However, while the MST neglects too many stock connections in the correlation-based network, our pretopological framework is more efficient since it depends on all of the connections. For more details, Figure 4.5b shows the precision and recall of predicting stocks affected by i_0 with different values of impact distance in the MST. We can see that at the same level of recall, the precision of the prediction based on the MST's connection in Figure 4.5a is mostly less than the precision of the prediction based on $\mathbf{F}(\{i_0\})$ in Figure 4.5b. In addition, successive computations of our pseudoclosure to get $\mathbf{F}(\{i_0\})$ can help forecast the evolution of the failure contagion starting from i_0. Especially, Figure 4.6 shows that the contagion can reach distanced stocks in the MST [5]. We assume that this happens because the MST may contain edges corresponding to small stock correlations to comprise all stocks of the system under the acyclic condition.

[5] In Figure 4.5 and Figure 4.6, we use the function f defined by equation (4.7) as in [Nguyen, 2021b]. A similar result is found in [Nguyen, 2019a] when f is defined by equation (4.6).



(a) (b)

Figure 4.5: Relationship between the precision and the recall of predicting stocks influenced by the price shock of i_0 when (a) using $\mathbf{F}(\{i_0\})$, and (b) using the MST network.

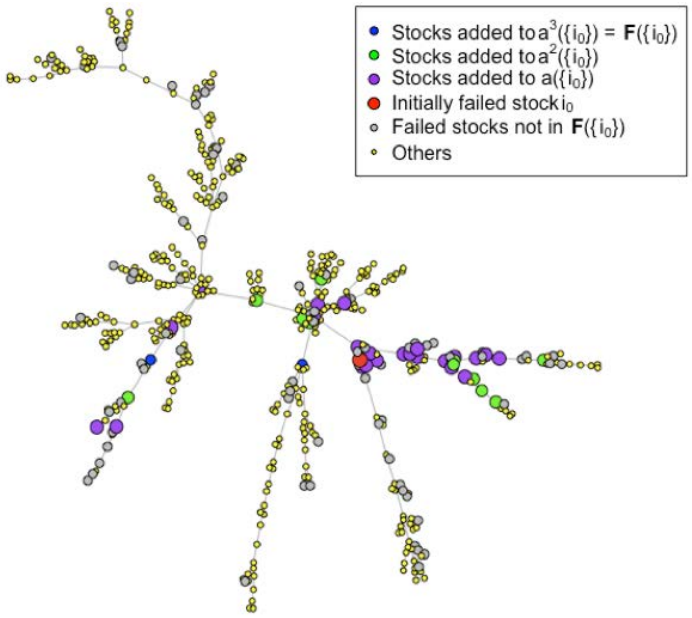


Figure 4.6: Distanced stocks in the MST network reached in the dilation process from $\{i_0\}$ to $\mathbf{F}(\{i_0\})$ with $\theta = 0.34$ and $\gamma = 5 \times 10^{-4}$.

A question is which values of parameters θ and γ are appropriate. For both threshold functions given in equations (4.6) and (4.7), θ plays a major role in determining the magnitude of the group impact, especially the one of the group containing only one stock i_0. Let's remind that according to Proposition 4.2, we have

$$a(\{i_0\}) = \{i_0\} \cup \{k \neq i_0 | c_{ki_0} \geq f(1)\} \tag{4.8}$$

Therefore, if the dilation process from $\{i_0\}$ to $\mathbf{F}(\{i_0\})$ starts with an extremely large impact threshold $f(1)$, there are very few stocks in $a(\{i_0\})$. Due to the concavity of f, the threshold

decreases slowly when the group size is small. So, the process stops quickly with very few stocks in $\mathbf{F}(\{i_0\})$. As a result, when we use this closure to predict stocks affected by i_0, the precision is large because of the significantly tight relationships between the price fluctuations of stocks in $\mathbf{F}(\{i_0\})$ and i_0. However, since it is hard to extend $\{i_0\}$ with a too large lower threshold, the recall is small. This remark is illustrated on the right of Figure 4.5a, where the precision evenly reaches 80 – 100%, but the recall of lower than 3% is trivial. On the contrary, a too small threshold at the first step of the process is also inappropriate because it is invaluable if many stocks lowly correlating to i_0 are considered to be affected by i_0.

In our opinion, since we are not sure that the fail of i_0 causes the fails of all stocks of H, the precision of the prediction is more important than the recall. Therefore, we're interested in choosing suitable values for θ and γ such that the impact threshold of expanding $\{i_0\}$ to $a(\{i_0\})$ in the first step of the dilation process is neither too large nor too small. In particular, if we use equation (4.6), θ, which equals the threshold at the first step, should range from 0.65 to 0.72; if we use equation (4.7), the first threshold equals $1-\theta e^{\gamma}$, and θ should range from 0.28 to 0.35. Then, the precision is acceptable enough, and the recall is not too small. For example, the precision is 75% and the recall is 26.73% if we use equation (4.6) with $\theta = 0.66$ and $42.5 \leq \gamma \leq 62.5$; the precision is 75.7% and the recall is 27.5% if we use equation (4.7) with $\theta = 0.34$ and $\gamma = 5 \times 10^{-4}$.

Inversely, in [Nguyen, 2021b], we use the opening of $E\backslash\{i_0\}$ in the pretopological space given in Proposition 4.2, where f is verified by equation (4.7), to predict stocks not affected by the price shock of i_0. In our database, there are 387 stocks whose prices did not decrease more than 70% during 6 months after the crash day t_0 of i_0. We consider these stocks as usable nodes of the stock system in the research period. These nodes are our objectives in this prediction. Figure 4.7 shows the precision and recall of the prediction with different values of θ and γ. There are two extreme cases. Firstly, when the impact threshold at the first step of the extenuation process from $\{i_0\}$ to $\mathbf{O}(E\backslash\{i_0\})$ is too large, then $\mathbf{O}(E\backslash\{i_0\}) = E\backslash\{i_0\}$. Indeed, according to Proposition 4.2, we have

$$i(E\backslash\{i_0\}) = \{k \neq i_0 | c_{ki_0} < f(1)\} \tag{4.9}$$

Therefore, if the impact threshold $f(1)$ is larger than the maximum of the correlations between i_0 and other stocks, we obtain $i(E\backslash\{i_0\}) = E\backslash\{i_0\}$. So, the process stops at the first extenuating step. Then, all of the usable nodes belong to the opening $\mathbf{O}(E\backslash\{i_0\})$. Consequently, the recall is always equal to 100%, but it has no meaning because it's equivalent to make no prediction. Contrarily, if the impact threshold $f(1)$ at the first step is too small, there are few stocks, or even no stocks, satisfy the inequality in (4.9). Because f is decreasing, the impact threshold decreases step by step. Consequently, the opening $\mathbf{O}(E\backslash\{i_0\})$ contains very few stocks or even becomes empty. Thus, the recall is extremely small or even equal to zero. As a result, we should consider suitable values for θ and γ such that the impact threshold is neither too large nor too small. For example, we propose $0.2 < \theta < 0.36$, then the precision of the prediction is mostly from 79.3% to 88.9%, while the recall is from 60.5% to 100%.

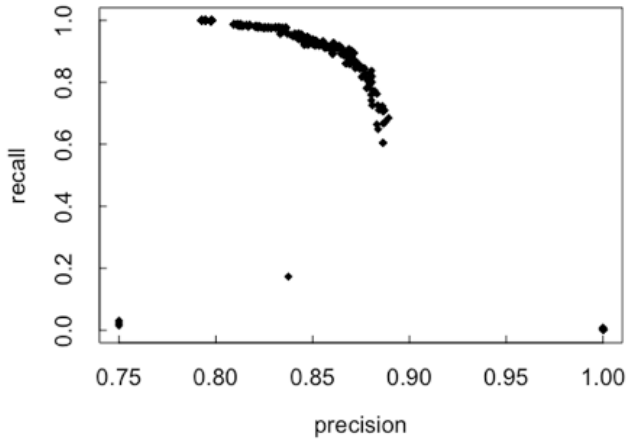


Figure 4.7: Relationship between the precision and the recall of predicting stocks not influenced by the price shock of i_0 by $\mathbf{O}(E \backslash \{i_0\})$.

As a result, instead of using graph representations of a stock system to study the cascading failure starting from a part of the system, we propose that pretopology theory is a better option that provides necessary tools to consider multi-relation as well as the crowd effect. By assuming that the number of failed stocks increases the impact of these stocks on another one, the evolution of the cascading failure is modeled by our pseudoclosure better than by connections in the MST network and the correlation-based threshold network. Meanwhile, the interior can help solve the inverse problem. Although there is a little difficulty in choosing a particular function determining the group impact's threshold, our pretopology framework can be effective when the function's value at 1 is neither too small nor too large. To improve our pretopological framework, we can try other threshold functions. Also, we can try additional combinations of stocks' relations that affect the co-movement of their prices, such as business sectors, investors, market capitalization, because pretopology theory allows processing multiple relations between a component with other components. Besides, in order to use other benefit tools of pretopology theory, we can improve our pretopological space of a stock system such that it is one of the special types introduced in the next section.

5 Types of Pretopological Spaces

There are four types of pretopological spaces: $\mathcal{V}$-type spaces, $\mathcal{V}_D$-type spaces, $\mathcal{V}_S$-type spaces, and topological spaces.

Definition 4.7. *Let a pretopological space $(E, a(.))$, we say that it is:*

(i) a $\mathcal{V}$-type space if $(A \subset B \Rightarrow a(A) \subset a(B))$ for any subsets A, B of E.

(ii) a $\mathcal{V}_D$-type space if $a(A \cup B) = a(A) \cup a(B)$ for any subsets A, B of E.

(iii) a $\mathcal{V}_S$-type space if $a(A) = \bigcup_{x \in A} a(\{x\})$ for any subset A of E.

The relationships of these types of pretopological spaces are demonstrated below:

Proposition 4.3. *Any $\mathcal{V}_D$-type space is a $\mathcal{V}$-type space.*

Proof. Let $(E, a(.))$ be a $\mathcal{V}_D$-type space. Then, for any subsets A, B of E such that $A \subset B$, since $A \cup B = B$, we have:

$$a(B) = a(A \cup B) = a(A) \cup a(B) \tag{4.10}$$

This implies that $a(A) \subset a(B)$, so $(E, a(.))$ is of $\mathcal{V}$-type. ■

Proposition 4.4. *Any $\mathcal{V}_S$-type space is a $\mathcal{V}_D$-type space.*

Proof. Let $(E, a(.))$ be a $\mathcal{V}_S$-type space. Then, for any subsets A, B of E, we have:

$$a(A \cup B) = \bigcup_{x \in A \cup B} a(\{x\}) = \left(\bigcup_{x \in A} a(\{x\}) \right) \cup \left(\bigcup_{x \in B} a(\{x\}) \right) = a(A) \cup a(B) \tag{4.11}$$

Thus, $(E, a(.))$ is of $\mathcal{V}_D$-type. ■

The last level in pretopological spaces is the level of topological spaces.

Proposition 4.5. *A pretopological space $(E, a(.))$ is a topological space if and only if it is of $\mathcal{V}_D$-type and $a(a(A)) = a(A)$ for any subset A of E.*

Proof. a is a pseudoclosure, hence by Definition 4.1, we have $a(\varnothing) = \varnothing$ and $A \subset a(A)$. Then, $(E, a(.))$ is a topological space $\Leftrightarrow$ a is a Kuratowski closure operator [Kuratowski, 1922] $\Leftrightarrow$ $a(a(A)) = a(A)$ and $a(A \cup B) = a(A) \cup a(B)$ for any subsets A, B of E. ■

The proposition shows that if the pseudoclosure a cannot extend $a(A)$ for any subset A of E, a becomes a closure operator, and $(E, a(.))$ becomes a topological space with more special characteristics than pretopological spaces. One of the most important concepts of topological spaces is the neighborhood which helps define many essential tools to study a complex network such as connectedness, limits, continuity. However, the last level of pretopological spaces can't help describe the expanding process from A to its closure $\mathbf{F}(A)$ in individual steps as others. As topology theory, the concept of neighborhoods in pretopology theory is defined as follows:

Definition 4.8. *Given a pretopological space $(E, a(.))$, the family defined by*

$$\mathcal{U}(x) = \{B \subset E \,|x \in i(B)\}, \qquad \forall x \in E \tag{4.12}$$

is called the family of neighborhoods of x.

However, despite Definition 4.8, we still have trouble in constructing the concept of connectivity in pretopological spaces. Indeed, in a general pretopological space, if U is a neighborhood of an element x of E, and U is included in $V \subset E$, because $U \subset V$ does not always imply $i(U) \subset i(V)$, we're not sure that V is also a neighborhood of x. Therefore, $\mathcal{V}$-type spaces play a particular role in pretopological spaces since the preservation of inclusion relation through the interior is equivalent to the same preservation through the pseudoclosure. Moreover, according

to Proposition 4.3 and Proposition 4.4, $\mathcal{V}_S$-type spaces and $\mathcal{V}_D$-type spaces are $\mathcal{V}$-type spaces. In a $\mathcal{V}$-type space, we have many necessary tools to construct a proximity concept, such as basics of neighborhoods of an element, connectedness, and minimal closed subsets. More details are given in [Belmandt, 2011]. Therefore, we propose to improve the pretopological space in Section 3 such that it becomes a $\mathcal{V}$-type space. For example, instead of considering the average of the correlations between a stock and a given group of others as mentioned in Section 3, we can use the maximum of them:

Proposition 4.6. *Let f be a decreasing and concave function from $[1, N]$ into $[0, 1)$. Let a be a map from $\mathcal{P}(E)$ into $\mathcal{P}(E)$ such that $a(\varnothing) = \varnothing$ and*

$$a(A) = A \cup \left\{ k \in E \backslash A \middle| \max_{j \in A} c_{jk} \geq f(\|A\|) \right\}, \qquad \forall A \in \mathcal{P}(E) \backslash \{\varnothing\} \tag{4.13}$$

where $\|A\|$ is the size of A. Then, $(E, a(.))$ is a pretopological space of $\mathcal{V}$-type.

Proof. It's clearly that the map a is a pseudoclosure by Definition 4.1, so $(E, a(.))$ is a pretopological space.

Besides, for any subsets A, B of E, if $A \subset B$, we get $\max_{j \in A} c_{jk} \leq \max_{j \in B} c_{jk}$. On the other hand, because $\|A\| \leq \|B\|$ and f is decreasing, we obtain $f(\|B\|) \leq f(\|A\|)$. Therefore, for any stock $k \in a(A)$, if $k \in E \backslash A$ and $\max_{j \in A} c_{jk} \geq f(\|A\|)$ then $\max_{j \in B} c_{jk} \geq f(\|B\|)$, so $k \in a(B)$. By contrast, if $k \in A$, we get directly $k \in a(B)$ since $k \in B$. Consequently, $a(A) \subset a(B)$.

Hence, $(E, a(.))$ is a pretopological space of $\mathcal{V}$-type. ■

Nevertheless, with the pseudoclosure defined in Proposition 4.6, the impact of a group of stocks on another stock k depends only on the group's size and the stock most correlated to k in A.

As a result, thanks to pretopological spaces' advantage of describing multi-relation between a complex system's components, modeling relationships between a component and a group, and studying the evolution of diffusion processes and condensation processes taking place in the system, we propose to use pretopology theory more in our future works about stock systems in addition to network analysis. Meanwhile, deeper analysis of pretopological frameworks should be studied. Especially, we plan to focus on the frameworks where the corresponding spaces is at least of $\mathcal{V}$-type spaces to use more beneficial tools of this theory.

Chapter 5

Topological Anomalies of Market Indices' Dynamics

Objective

Because the collective behavior of a stock market is usually well captured through the fluctuation of its representative index, our target in this final chapter is to detect abnormal behaviors of a stock market through anomalies in the dynamics of its representative index's return. To figure out important features of the dynamics, we use the approach of *topological data analysis* combined with the method of time-delay embedding to get topological information about the dynamics' state space. Our method is demonstrated to be efficient in the case of the S&P 500 Index.

Contents

1 Market Indexes as Representations of Stock Markets' Collective Behaviors

As a complex system with many components and complicated relationships, the movement of a stock market as a whole is not easy to predict. It is also difficult to distinguish the movement's magnitude because the collective behavior of a stock market is not a simple synthetic of its components' behaviors. For example, a market crash may lead to a recession like the crash of 1929 that only occurred in over four trading days but drowned out the market into a 10-year depression [1]. In general, people often conjecture the market's stability and then guess its upturn or downturn. The conjecture usually bases on many macro factors that can directly affect all of the market's components and drive their movements in the same direction. These factors can be the political situation, the government's important policies and procedures, the infrastructure, the import and export values, the monetary... Although the macro factors can provide a valuable prediction of the market development, this method requires deep knowledge about the economy and take our time to analyze many statistics in different types. So, it is not suitable for individual investors. In addition, because the macro statistics are reported periodically, this data may not be suitable for evaluating the market's current situation if there is suddenly any significant change, such as the occurrence of an epidemic or a disaster. Therefore, in addition to using macro statistics, we can observe the current developments of the market directly through technical analysis based on price fluctuations of the underlying holdings.

Especially, to get an overview of a stock market intuitively, people often depend on market indexes. We know that a market index is a hypothetical portfolio of investment holdings. It can be composed of all listed stocks or a basket of representative stocks satisfying many conditions about market capitalization, liquidity, public float, earnings... Many stock market indexes are capitalization-weighted indexes [2] included the two indexes studied in this thesis, the S&P 500 Index and the VN Index. Let's remind that in Chapter 3, the fluctuations of the two indexes are respectively found to correlate highly with the fluctuation of all stock returns' first PC. Meanwhile, the first PC explains most of the stock returns' variances and has the highest sum of square of correlation coefficients with all of them, so its movement can represent the movement of the market's collective behavior. Consequently, we propose that a market index's fluctuation can be used to gauge the market's collective behavior if it highly correlates with the first PC's fluctuation. The indexes' occurrence made stock markets different from most complex systems. Indeed, frequently in such systems, their collective behaviors are not easy to capture by a measurement that is transparent, continuously updated, and provided free like market indexes of stock systems. Therefore, in this chapter, our target is to detect abnormal behaviors of a stock market through anomalies in the dynamics of its representative index's fluctuation, i.e., the index's return calculated by the log-difference of the index's daily values.

[1] See details on the site: https://www.fdic.gov/about/history/timeline/1920s.html

[2] Other methods for weighting a stock market index are the price-weighted, fundamental-weighted, and equal-weighted index construction methods.

2 Time-delay Embedding of a Time Series

For discovering anomalies in the dynamics of an index return, we need to apprehend its different states. In addition, we can't avoid getting noises in the timing data of the index return. To solve the first problem, Ruelle [Ruelle, 1979] and Packard et al. [Packard, 1980] introduced a simple method named the time-delay embedding. When analyzing the point cloud got from this method by its persistence diagram, an efficient tool given in the *topological data analysis* (TDA), we can deal with the second problem.

2.1 Delay Reconstruction

According to the embedding theorem of Takens [Takens, 1981], a chaotic series can be perfectly modeled by a smooth function when it is correctly embedded. From this point of view, the main goal of the time-delay embedding method is converting a time series into a point cloud of a higher-dimensional space such that it can capture different states of the time series' dynamics. At first, let's make acquaintance with the following concept:

Definition 5.1. *A reconstructed vector obtained from a time series* (x_t) *is defined by, for all* t*,*

$$\mathbf{y}_t^{\tau,d} = \left(x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(d-1)\tau}\right) \tag{5.1}$$

We call τ *the time-delay and* d *the embedding dimension.*

In [Sauer, 1991], Sauer et al. found that, with probability one, there are suitable values for parameters of the time-delay and the embedding dimension to fulfill the goal above. Hence, $\mathbf{Y} = (\mathbf{y}_t^{\tau,d})$ is called the phase/state space. In time series analysis, this approach is used to estimate a dynamical system's attractor as a set of numerical values toward which the system tends to evolve for various starting conditions [Kantz, 2003]. Therefore, in such a system, an appropriate time-delay embedding of some system variable's scalar observations over time helps transform the one-dimensional data into a point cloud of a d-dimensional space to capture the system's deterministic properties, especially topological properties. For example, let's consider a time series (x_t) such that $x_1 = 1$, and $x_t = \alpha_t x_{t-1} + 2\sin t + \beta_t$ for any integer $t > 1$, where α_ts are i.i.d. random normal variables with unit mean and standard deviation of 10^{-3} and β_ts are i.i.d. random normal variables with zero mean and standard deviation of 10^{-2}. Figure 5.1 shows a sample of (x_t) with 200 sample points and how we construct the state space of (x_t) with $\tau = 2$ and $d = 3$. In this example, we map each reconstructed vector with a point of $\mathbb{R}^3$ by the vector's components. Then, we get the time series' topological feature on a certain scale, a circle. Topological features will be discussed more in the next section.

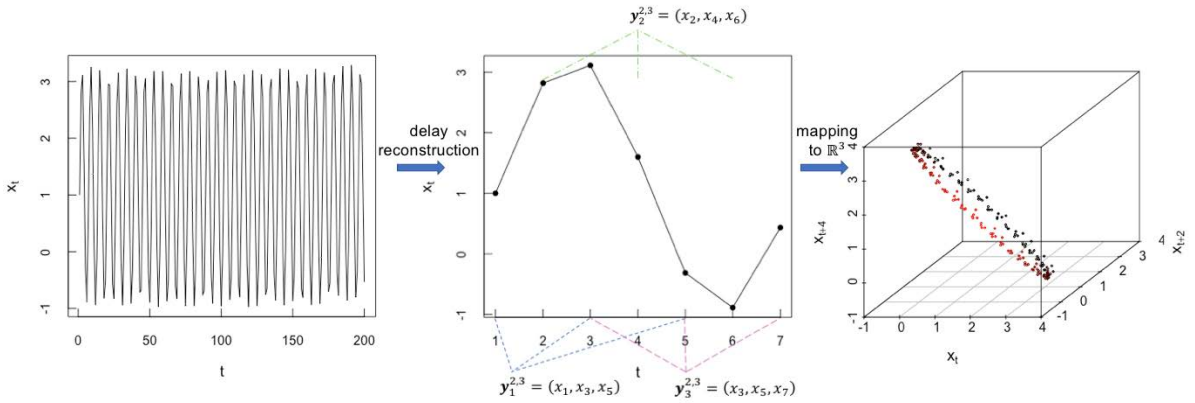


Figure 5.1: Time-delay embedding of a time series helps get the series' topological features.

2.2 Selecting Time-delay

In practice, time-delay τ and embedding dimension d should be selected based on the data itself and the objective of the analysis. Frequently, an appropriate selection of τ should make the two series (x_t) and $(x_{t+\tau})$ are independent of each other as much as possible. Hence, the measures below are common to choose time-delay parameters: autocorrelation and average mutual information [Abarbanel, 1993].

- *Autocorrelation*: Time-delay τ is selected as the first zero of the linear autocorrelation function below:

$$A(\tau) = \frac{\langle (x_{t+\tau} - \overline{x})(x_t - \overline{x}) \rangle_t}{\left\langle (x_t - \overline{x})^2 \right\rangle_t} \tag{5.2}$$

  where $\langle . \rangle_t$ is the average over time and $\overline{x} = \langle x_t \rangle_t$. A disadvantage of this method is that the zero autocorrelation only confirms the linear independence between (x_t) and $(x_{t+\tau})$, while (x_t) may contain nonlinear dependence.

- *Average mutual information*: According to information theory [Gallager, 1968], we can quantify how much information one can learn about $x_{t+\tau}$ from measuring x_t by the mutual information of the two measurements as follows:

$$I(x_t, x_{t+\tau}) = \log_2 \frac{\hat{p}(x_t, x_{t+\tau})}{\hat{p}_1(x_t)\,\hat{p}_2(x_{t+\tau})} \tag{5.3}$$

  where $\hat{p}$ is the estimated joint probability distribution of x_t and $x_{t+\tau}$; $\hat{p}_1$ and $\hat{p}_2$ are the estimated marginal function of the joint probability distribution. Then, the average mutual information of x_t and $x_{t+\tau}$ for a given time-delay τ, denoted $\mathrm{AMI}(\tau)$, is the average of all possible measurements of $I(x_t, x_{t+\tau})$, i.e.,

$$\mathrm{AMI}(\tau) = \sum_t \hat{p}(x_t, x_{t+\tau})\, I(x_t, x_{t+\tau}) \tag{5.4}$$

  Hence, to get information of $x_{t+\tau}$ from observing x_t as little as possible, a good hint of a

choice for τ is the time lag where the first minimum of AMI occurs. This method is more popular than the previous one because it works well with both nonlinear and linear time series.

2.3 Selecting Embedding Dimension

Different methods, usually dynamical tests and geometrical tests, are proposed to get the optimal embedding dimension. In a dynamical system, reconstructed vectors are used to identify the system's deterministic properties, which do not depend on initial conditions or perturbations. Hence, dynamical tests look for an embedding dimension that provides a unique future for every data point. From this point of view, the following assumption is proposed:

$$\frac{d\mathbf{y}_t^{\tau,d}}{dt} = F(\mathbf{y}_t^{\tau,d}) \tag{5.5}$$

or equivalently, the concrete form of (5.5) is:

$$\mathbf{y}_{t+1}^{\tau,d} = f(\mathbf{y}_t^{\tau,d}) \tag{5.6}$$

So, dynamical tests are carried out by increasing the embedding dimension until the typical behavior of the time series appears. Lyapunov exponents' estimation [Eckmann, 1986] is an example of such tests. Another approach of dynamical tests is singular-value analysis [Broomhead, 1986], where reconstructed vector $\mathbf{y}_t^{\tau,d}$ is assumed to be composed by the typical behavior $\mathbf{z}_t^{\tau,d}$ plus some contamination $\mathbf{c}$, i.e.,

$$\mathbf{y}_t^{\tau,d} = \mathbf{z}_t^{\tau,d} + \mathbf{c}_t \tag{5.7}$$

Then, by analyzing the eigenvalue spectrum of the $d \times d$ sample covariance matrix of the reconstructed vectors' components, we can recognize which eigenvalues represent the noise $\mathbf{c}$. So, the number of remaining eigenvalues is a good choice for the optimal embedding dimension.

However, in our opinion, the geometrical tests which depend on the distance between points of the state space, i.e., the reconstructed vectors, are more suitable for real-world time series. The reason is that the main target of the reconstruction is to provide an Euclidean space $\mathbb{R}^d$ which is large enough so that the attractor obtained from the embedding can be unfolded without ambiguity. In other words, if two points of the state space close to each other, this is caused by the state space's property instead of the small value of d. Therefore, the geometrical tests are direct approaches to the reconstruction's goal. The test of the saturation of some system invariant with the embedding dimension's change and the method of false nearest neighbor are some examples of such tests.

- *Saturation of system invariants*: The method looks for the embedding dimension that provides independence with some function depending on distances between the state space's

points. A familiar example of such function is provided in [Grassberger, 1983]:

$$C_{p,r} = \left\langle n\left(r, \mathbf{y}_t^{\tau,d}\right)^{p-1} \right\rangle_t \tag{5.8}$$

where $n\left(r, \mathbf{y}_t^{\tau,d}\right)$ is the number of the state space's points in the ball with center of $\mathbf{y}_t^{\tau,d}$ and radius of r, $\langle . \rangle_t$ is the average over time. So, $C_{p,r}$ is the average over the attractor of moments of the number density $n\left(r, \mathbf{y}_t^{\tau,d}\right)$ and depends on the embedding dimension because of the distances' calculations between the state space's points. Hence, for an observed time-series, we compute $C_{p,r}$ as a function of embedding dimension d and select the necessary embedding dimension when the variation of $C_{p,r}$ with d is small enough.

- *False nearest neighbor*: Instead of basing on some functions associated with distances between the state space's points, Kennel et al. [Kennel, 1992] proposed this method as a straightforward approach to the reconstruction's goal because the distances are directly considered. This method argues that if d is the optimal embedding dimension, the attractor is unfolded in the state space with dimension d and higher. So, this method looks for d as the smallest number such that for any point of the d-dimensional reconstructed space, its nearest point is still close enough in the $(d+1)$–dimensional reconstructed space. In particular, for each reconstructed vector $\mathbf{y}_t^{\tau,d}$, let the reconstructed vector $\mathbf{y}_{t^*}^{\tau,d}$ be the nearest neighbor of $\mathbf{y}_t^{\tau,d}$ with nearest in the sense of some distance [3]. Then, if the two vectors move apart when the dimension increases, $\mathbf{y}_{t^*}^{\tau,d}$ is called the false nearest neighbor of $\mathbf{y}_t^{\tau,d}$, and d is not an appropriate embedding dimension. Hence, we need a given threshold R_τ to identify a false nearest neighbor, and d should be the smallest number such that no false nearest neighbor exists, i.e.,

$$a\left(t,d\right) = \frac{\left\|\mathbf{y}_t^{\tau,d+1} - \mathbf{y}_{t^*}^{\tau,d+1}\right\|}{\left\|\mathbf{y}_t^{\tau,d} - \mathbf{y}_{t^*}^{\tau,d}\right\|} < R_\tau, \qquad \forall t \tag{5.9}$$

where $\|.\|$ is some measurements of distance. Without loss of generality, we use the Euclidean distance in this work. The function $a\left(t,d\right)$ measures the variation of the distance between $\mathbf{y}_t^{\tau,d}$ and its nearest neighbor from d to $d+1$. Although this method is the most common to select the embedding dimension, it has some drawbacks. Firstly, it is too sensitive to the value of R_τ; secondly, the threshold may have different values for different time series. To avoid these problems, in [Cao, 1997], the author suggested a modification of this method by using the average of $a\left(t,d\right)$ over time and investigating the change of the following measure:

$$E_1\left(d\right) = \frac{\left\langle a\left(t,d+1\right)\right\rangle_t}{\left\langle a\left(t,d\right)\right\rangle_t} \tag{5.10}$$

$E_1\left(d\right)$ measures the variation of the average $\left\langle a\left(t,d\right)\right\rangle_t$ from d to $d+1$. Especially for different time series, it is found to stop changing when d is large enough. Therefore, in

[3] If $\mathbf{y}_{t^*}^{\tau,d} \equiv \mathbf{y}_t^{\tau,d}$, we take $\mathbf{y}_{t^*}^{\tau,d}$ as the second nearest neighbor of $\mathbf{y}_t^{\tau,d}$.

practice, when its change becomes trivial when the embedding dimension is larger than some number d_0, we take $d_0 + 1$ as the optimal embedding dimension. Figure 5.2 shows how we use this modification of the method of false nearest neighbor to select the optimal embedding dimension of the time series plotted in Figure 5.1, given $\tau = 2$.

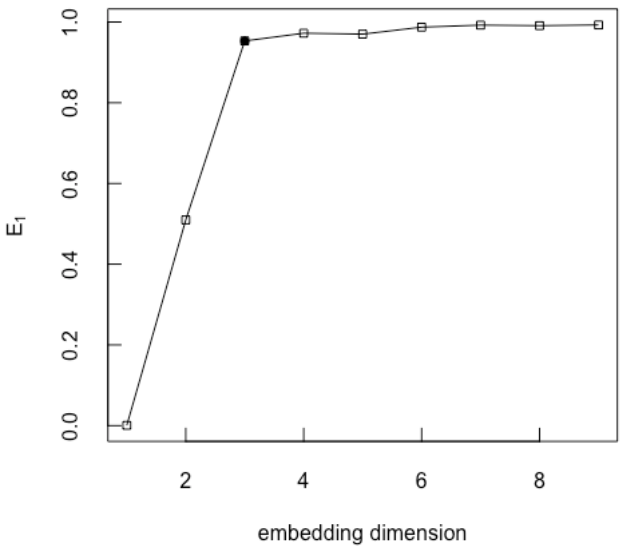


Figure 5.2: Selecting the optimal embedding dimension (filled point) by looking for the stability of the mean of distance's variation between a reconstructed vector and its nearest neighbor when the dimension increases.

3 Persistent Homology

In our financial context discussed in Section 1, the time series of an index return is a sequence of scalar measurements representing the collective behavior of the corresponding stock system. So, the point cloud $\mathbf{Y} = (\mathbf{y}_t^{\tau,d})$ is expected to define a state space such that each point in this space specifies a state of the system. However, because time-series data tend to be considerably noisy, there are numerous points in the state space, so the points' coordinates don't have more meaning than the points' arrangement. In this sense, we propose using persistence homology, the main technique of TDA [Chazal, 2021; Edelsbrunner, 2002]. TDA is an approach that provides topological and geometrical tools to infer information about the structure of a point cloud in a metric space at different spatial resolutions. The below paragraphs provide the principal notions of TDA. The final result helps understand the "shape" of the time series in different spatial resolutions to get a valuable conclusion of the underlying system's behavior without worrying about noises.

3.1 Simplicial Complexes

The state space resulted from the time-delay embedding method is a point cloud of $\mathbb{R}^d$.

This data is discrete. Hence, to study the points' arrangement, an intuitive approach is merging the points into a set of connected components such that the set's structure gives meaningful information about the points' arrangement. In TDA, the set must be a simplicial complex to easily study its topological features.

Let $\mathbf{V} = \{v_0, v_1, \ldots, v_k\} \subset \mathbb{R}^d$ be a set of affinely independent points.

Definition 5.2. *A k-dimensional simplex σ spanned by $\mathbf{V}$ is the convex hull of $\mathbf{V}$, i.e.,*

$$\sigma = \left\{ \sum_{i=0}^{k} \alpha_i v_i \,\middle|\, \sum_{i=0}^{k} \alpha_i = 1 \ \wedge \ 0 \leq \alpha_i \leq 1 \right\} \tag{5.11}$$

$v_0, v_1, \ldots, v_k$ are called vertices of σ. The convex hull of any subset of $\mathbf{V}$ is also a simplex called a face of σ.

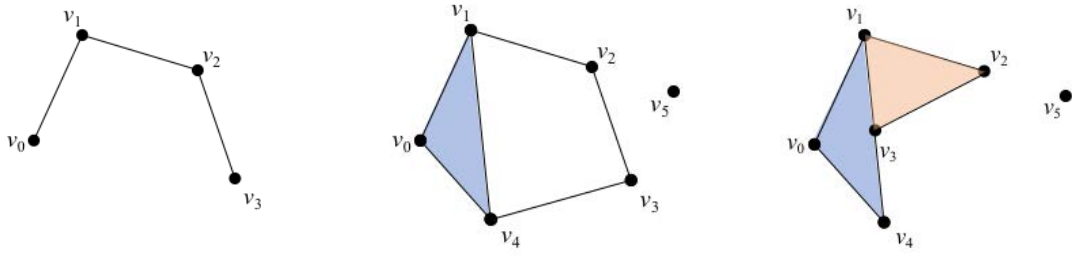
Intuitively, in $\mathbb{R}^3$, a 0-dimensional simplex is a point, a 1-dimensional simplex is a line segment, a 2-dimensional simplex is a triangle, a 3-dimensional simplex is a tetrahedron, a 4-dimensional simplex is a cell...

Definition 5.3. *A simplicial complex $\mathbf{G}$ is a finite collection of simplices, such that:*

(i) Any face of a simplex of $\mathbf{G}$ is a simplex of $\mathbf{G}$.

(ii) The intersection of any two simplices of $\mathbf{G}$ is either empty or a common face of both.

For example, Figure 5.3a shows a collection of simplices including 0-dimensional simplices: points v_0, v_1, v_2, and v_3, 1-dimensional simplices: line segments v_0v_1, v_1v_2, and v_2v_3. Because the faces of a line segment are its starting and ending point, the collection obviously satisfies two properties (i) and (ii) proposed in Definition 5.3, so it is a simplicial complex. Hence, simplicial complexes can be seen as higher-dimensional generalizations of graphs. Similarly, Figure 5.3b shows the collection of simplices including points v_0, v_1, v_2, v_3, v_4, and v_5, line segments: v_0v_1, v_1v_2, v_2v_3, v_3v_4, v_1v_4, and v_4v_0, and the only 2-dimensional simplex: triangle $v_0v_1v_4$. Because the faces of a triangle are its vertices and edges, the collection is a simplicial complex. However, the collection of simplices in Figure 5.3c is not a simplicial complex since the intersection of two triangles $v_1v_2v_3$ and $v_0v_1v_4$ is not a face of the latter.



(a) A simplicial complex (b) A simplicial complex (c) Not a simplicial complex

Figure 5.3: Example and counterexample of simplicial complexes.

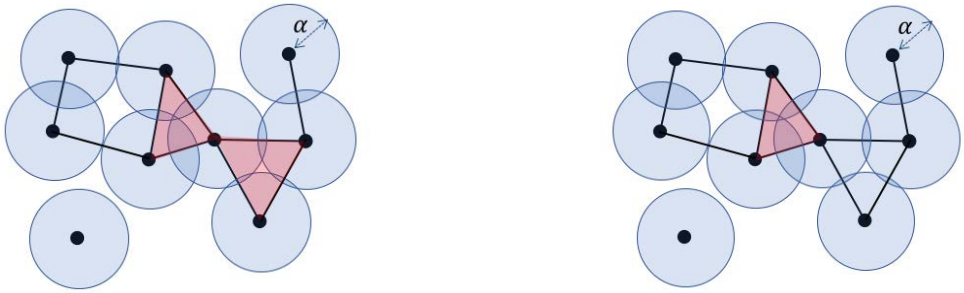
There are some ways to convert the set of affinely independent points $\mathbf{V}$ of a metric space into a simplicial complex to summarize, visualize and explore the point cloud's arrangement. We

introduce two familiar simplicial complexes constructed from a given point cloud: the Vietoris-Rips complex and the Čech complex.

Definition 5.4. *Given a number α, the Čech complex $\check{\text{C}}\text{ech}_\alpha(\mathbf{V})$ is the set of simplices spanned by subsets of $\mathbf{V}$ such that: for any simplex $\sigma \in \check{\text{C}}\text{ech}_\alpha(\mathbf{V})$, the closed balls $B(v_i, \alpha)$ for all vertex v_i of σ have a non-empty intersection.*

Definition 5.5. *Given a number α, the Vietoris-Rips complex (also called Vietoris complex or Rips complex) $\text{Rips}_\alpha(\mathbf{V})$ is the set of simplices spanned by subsets of $\mathbf{V}$ such that: for any simplex $\sigma \in \text{Rips}_\alpha(\mathbf{V})$, $\|v_i - v_j\| \leq \alpha$ for any vertices $v_i,\ v_j$ of σ.*

Vietoris-Rips complexes and Čech complexes are simplicial complexes. Figure 5.4 shows an example of constructing these complexes of 9 points of $\mathbb{R}^2$. Besides, since we must find the intersection of balls when constructing $\check{\text{C}}\text{ech}_\alpha(\mathbf{V})$, the number of calculations becomes numerous if the dimension or the number of points increases. Therefore, Vietoris-Rips complexes are easier to construct because we only have to compute the distances between points. Furthermore, for any number α, $\check{\text{C}}\text{ech}_\alpha(\mathbf{V}) \subset \text{Rips}_{2\alpha}(\mathbf{V})$ (see Figure 5.4).



(a) $\text{Rips}_{2\alpha}(\mathbf{V})$ (b) $\check{\text{C}}\text{ech}_\alpha(\mathbf{V})$

Figure 5.4: $\check{\text{C}}\text{ech}_\alpha(\mathbf{V})$ as a subset of $\text{Rips}_{2\alpha}(\mathbf{V})$.

3.2 Homology Groups

To discover topological information of a simplicial complex, homology is a powerful approach that helps distinguish the complex's structures through detecting its holes. To understand the notion holes in algebraic topology, we must know the notion boundary [Munkres, 1993]. Let's denote $[v_0, v_1, \ldots, v_k]$ as a simplex spanned by points $v_0, v_1, \ldots, v_k$ together with an orientation of the vertices. We call it an oriented simplex.

Definition 5.6. *The k-boundary map $\partial_k : C_k \to C_{k-1}$ $(k > 0)$ is defined by:*

(i) for any oriented simplex $\sigma = [v_0, v_1, \ldots, v_k]$,

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, v_{i-1}, v_{i+1}, ..v_k] \tag{5.12}$$

and

(ii) for any k-dimensional simplices $\sigma_1, \ldots, \sigma_p$*, and coefficients* $\alpha_1, \ldots \alpha_p \in \mathbb{Z}$*,*

$$\partial_k \left(\sum_{i=1}^{p} \alpha_i \sigma_i \right) = \sum_{i=1}^{p} \alpha_i \partial_k (\sigma_i) \tag{5.13}$$

where C_k *is the set of k-chains with coefficients in* $\mathbb{Z}$

For example,

- $\partial_0 (v_0) = 0$,
- $\partial_1 ([v_0, v_1]) = v_1 - v_0$,
- $\partial_2 ([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1]$.
- For any closed polygonal curve $c = [v_0, v_1] + [v_1, v_2] + \ldots + [v_{k-1}, v_k] + [v_k, v_0]$: $\partial_1 (c) = \partial_1 ([v_0, v_1]) + \ldots + \partial_1 ([v_k, v_0]) = v_1 - v_0 + v_2 - v_1 + \ldots + v_0 - v_k = 0$.

Definition 5.7. *Elements of* $\ker (\partial_k)$ *are called k-cycles.*

For example, as demonstrated above, points are 0-cycles, closed polygonal curves are 1-cycles.

Definition 5.8. *A k-dimensional hole is a k-cycle that is not a boundary of a* $(k+1)$*-dimensional simplicial complex.*

For example, let c_1 and c_2 be the simplicial complex in Figure 5.3a and 5.3b, respectively. Then:

- For c_1, since $\ker (\partial_1) = \varnothing$, c_1 doesn't contain any 1-cycle.
- For c_2, let $c_2^\star = [v_1, v_2] + [v_2, v_3] + [v_3, v_4] + [v_4, v_1]$. Because $c_2^\star$ is a closed polygonal curve, it is a 1-cycle. On the other hand, for the only 2-dimensional simplex $[v_0, v_1, v_4]$ of c_2, we get

  $$\partial_2 ([v_0, v_1, v_4]) = [v_1, v_4] - [v_0, v_4] + [v_0, v_1] \neq c_2^\star$$

  So, $c_2^\star \notin \mathrm{Im} (\partial_2)$. We conclude that $c_2^\star$ is a 1-dimensional hole. By contrast, the closed polygonal curve $c_2^{\star\star} = [v_0, v_1] + [v_1, v_4] + [v_4, v_0]$ is not a 1-dimensional hole although it is a 1-cycle. The reason is $c_2^{\star\star} = \partial_2 ([v_0, v_1, v_4])$.

The k-dimensional holes are the topological features that we pay attention to when distinguishing the shape of a simplicial complex. Definition 5.8 suggests detecting k-dimensional holes by homology groups defined by:

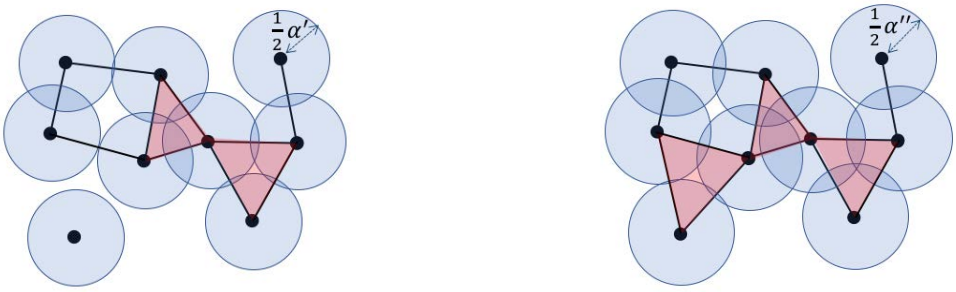
Definition 5.9. *Given a simplicial complex* $\mathbf{G}$*, the k-dimensional homology group of* $\mathbf{G}$ *is*

$$H_k (\mathbf{G}) = \ker (\partial_k) / Im (\partial_{k+1}) \tag{5.14}$$

Therefore, the 0-dimensional homology group H_0 represents the connected components of the complex, the 1-dimensional homology group H_1 represents the 1-dimensional holes or loops, the 2-dimensional homology group H_2 represents the 2-dimensional holes or cavities,... For example, in Figure 5.4a, the complex has two 0-dimensional features and one 1-dimensional feature.

3.3 Persistence Diagram

With homology, we have a way to detect topological features in the arrangement of a point cloud. However, even though we use only one method, ex., the Vietoris-Rips complex, to convert the point cloud to a simplicial complex, the result is too sensitive to the selected spatial resolution. For instance, different values of α lead to different Vietoris-Rips complexes whose homology characteristics can distinct from each other, as illustrated in Figure 5.5. In this figure, when α increases from α' to α", the number of 0-dimensional features decreases from two to one.



(a) $\mathrm{Rips}_{\alpha'}(\mathbf{V})$ (b) $\mathrm{Rips}_{\alpha''}(\mathbf{V})$

Figure 5.5: Topological changes of $\mathrm{Rips}_{\alpha}(\mathbf{V})$ when α changes.

To get an overview of homology groups' appearance in a simplicial complex when the spatial resolution changes, we can use the main tool of TDA, the persistence diagram.

Definition 5.10. *A filtration is a sequence of simplicial complexes* $(\mathbf{G}_\alpha)_{\alpha\in I\subset\mathbb{R}}$ *ordered by inclusion, i.e.,* $\mathbf{G}_{\alpha'} \subset \mathbf{G}_{\alpha''}$ *if* $\alpha' \leq \alpha''$ *for any numbers* α', α'' *of* I.

Definition 5.11. *A persistence diagram of a filtration* $(\mathbf{G}_\alpha)_{\alpha\in I\subset\mathbb{R}}$ *is the diagonal* $\{(x,y)\in\mathbb{R}^2 \,|\, x=y\}$ *together with a set of points* $\{(b,d)\in\mathbb{R}^2 \,|\, b<d\}$ *such that each point* (b,d) *corresponds to a topological feature as follows:* b *is the smallest value of* $\alpha\in I$ *such that the feature appears in* $\mathbf{G}_\alpha$*, and* d *is the smallest value of* $\alpha\in I$ *such that* $\alpha > b$ *and the feature disappears in* $\mathbf{G}_\alpha$.

We call b *the birth scale, and* d *the death scale of the feature. The difference* $d-b$ *is called the persistence of the feature.*

Given a point cloud, to merge the points into connected components, we should construct simplicial complexes for different spatial resolutions such that the complexes compose a filtration. Because the filtration's persistence diagram encodes topological information's change of the point cloud's arrangement when the spatial resolution changes, we know how "long" (for scale) a topological feature persists before it is filled in. So, we can get the arrangement's principal features, which are less affected by noises, to acquire the point cloud's structural characteristics such as the classification, the attractor...

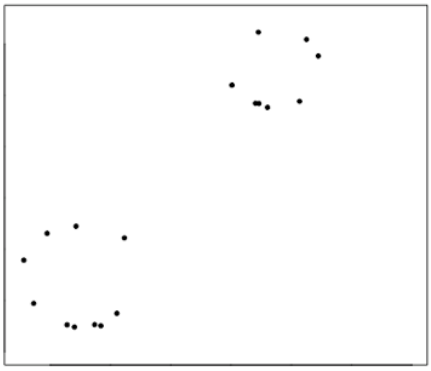
There are many ways to construct a filtration that covers a certain point cloud. The Vietoris-Rips complexes and the Čech complexes are familiar options.

Theorem 5.1. *Given a set of affinely independent points* $\mathbf{V}$ *of a metric space, the families of* $(\text{Rips}_\alpha(\mathbf{V}))_{\alpha\geq 0}$ *and* $\left(\check{\text{C}}\text{ech}_\alpha(\mathbf{V})\right)_{\alpha\geq 0}$ *are filtrations.*

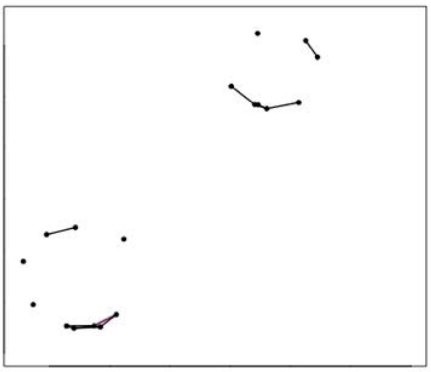
Proof. Let numbers $\alpha', \alpha'' \in \mathbb{R}$ such that $0 \leq \alpha' \leq \alpha''$.

- According to Definition 5.5, for any simplex $\sigma \in \text{Rips}_{\alpha'}(\mathbf{V})$, we get $\|v_i - v_j\| \leq \alpha' \leq \alpha''$ for any vertices $v_i,\ v_j$ of σ. So, $\sigma \in \text{Rips}_{\alpha''}(\mathbf{V})$. Consequently, $\text{Rips}_{\alpha'}(\mathbf{V}) \subset \text{Rips}_{\alpha''}(\mathbf{V})$.
- On the other hand, according to Definition 5.4, for any simplex $\sigma \in \check{\text{C}}\text{ech}_{\alpha'}(\mathbf{V})$, if σ is spanned by $\{v_i\}_{i\in K} \subset \mathbf{V}$, then $\varnothing \neq \bigcap_{i\in K} B(v_i, \alpha') \subset \bigcap_{i\in K} B(v_i, \alpha'')$. So, $\sigma \in \check{\text{C}}\text{ech}_{\alpha''}(\mathbf{V})$. Consequently, $\check{\text{C}}\text{ech}_{\alpha'}(\mathbf{V}) \subset \check{\text{C}}\text{ech}_{\alpha''}(\mathbf{V})$. ■

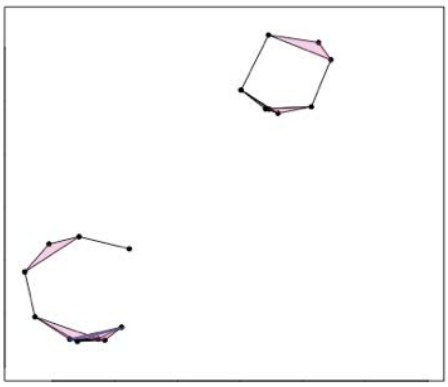
Figure 5.6 illustrates how to construct the persistence diagram of $(\text{Rips}_\alpha(\mathbf{V}))_{\alpha\geq 0}$, where $\mathbf{V}$ is the set of 18 points in Figure 5.6a. At first, when $\alpha = 0$, there are 18 0-dimensional features in $\text{Rips}_\alpha(\mathbf{V})$, i.e., 18 connected components corresponding to these points. So, the birth scales of these features are 0. When α increases a little, points closed together can be included in a sub-simplex of $\text{Rips}_\alpha(\mathbf{V})$, as illustrated in Figure 5.6b. In this case, some first connected components are merged together. So, such values of α become the death scales of some first 0-dimensional features. This change of $\text{Rips}_\alpha(\mathbf{V})$ is tracked by the circle points whose x-coordinates equal 0 in the persistence diagram (Figure 5.6d). When α increases to a certain value such that $\text{Rips}_\alpha(\mathbf{V})$ first contains a 1-dimensional feature, as shown in Figure 5.6c, this value is the feature's birth scale. When α continues to increase, the feature's death scale is the first value of α that the loop is filled in $\text{Rips}_\alpha(\mathbf{V})$. Its birth and death scales are coordinates of the first triangle point (from the left to the right) in the persistence diagram (Figure 5.6d). Obviously, from points corresponding to 1-dimensional features in the persistence diagram, we get that, with suitable scales, we can respectively have 2 loops. However, although there is also one point corresponding to a 2-dimensional feature in the persistence diagram, the feature doesn't give meaningful information about the point cloud's arrangement. The reason is that the point is too closed to the diagonal $\{(x, y) \in \mathbb{R}^2 | \, x = y\}$, so the feature's persistence is very small with respect to the scale's change. In general, features represented by points near the diagonal can be considered as noises. On the other hand, in Figure 5.6d, we can divide points corresponding to 0-dimensional features into 2 groups: one group of the point on the top and one group of the remaining points. This reflects the fact that the original point cloud $\mathbf{V}$ can be divided into 2 groups: one group of eight points on the top right corner and one group of the remaining points. Indeed, since the groups are only merged with a large enough α, the death scale of one of the first connected components is so larger than the others. In general, the group classification of 0-dimensional features on the persistence diagram associated with a point cloud provides a good hint for the point cloud's classification problem.



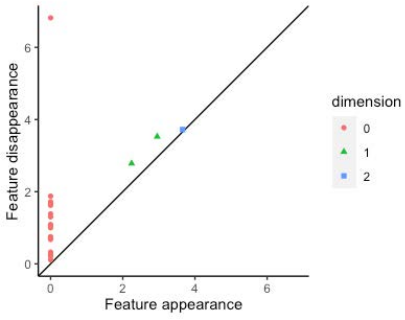
(a) A point cloud $\mathbf{V}$



(b) $\mathrm{Rips}_\alpha(\mathbf{V})$ where some connected components are merged



(c) $\mathrm{Rips}_\alpha(\mathbf{V})$ where the first loop appears



(d) The persistence diagram

Figure 5.6: Constructing the persistence diagram of $(\mathrm{Rips}_\alpha(\mathbf{V}))_{\alpha\geq 0}$.

In time series analysis, by constructing the persistence diagram associated with the state space $\mathbf{Y} = (\mathbf{y}_t^{\tau,d})$ of a time series, we can extract the state space's topological information, which is robustness to noises [Cohen-Steiner, 2007]. As a result, we can draw a meaningful conclusion for the underlying system's movement. For example, the groups of dense 0-dimensional features on the persistence diagram help classify the system's behaviors, while 1-dimensional features having high persistence relate to the periodic trend of the system's dynamics.

3.4 Bottleneck Distance and Wasserstein Distance

A question is how to compare the "shapes" of two different time series. This leads to comparing their corresponding persistence diagrams of filtrations of the same type. An approach to measuring the similarity between two persistence diagrams is constructing a distance function that gets a smaller value if the diagrams are more "similar" to each other. A common way to build the function is firstly matching every point not belonging to the diagonal in one persistence diagram to only one point in the other persistence diagram; then, using a metric to aggregate differences between matched points. After considering all possible pairs of matched points between the two diagrams, the distance between the diagrams corresponds to the best matching,

which is intuitively the matching providing the infimum of the matched points' aggregate difference. From this point of view, there are two familiar distance functions of persistence diagrams: the Bottleneck distance and the Wasserstein distance [Chazal, 2021].

Definition 5.12. *The Bottleneck distance between two persistence diagrams D_1 and D_2 is defined by:*

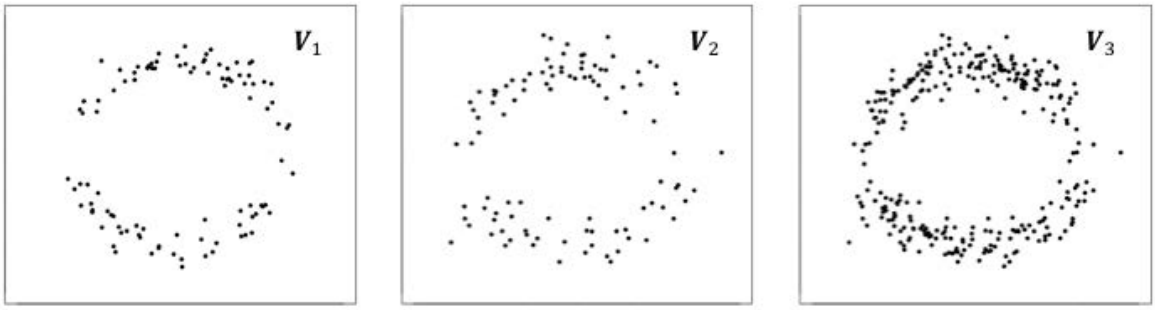
$$W_\infty(D_1, D_2) = \inf_{matching\ m} \sup_{(u,v)\in m} \|u - v\|_\infty \tag{5.15}$$

Definition 5.13. *The Wasserstein distance between two persistence diagrams D_1 and D_2 is defined by:*

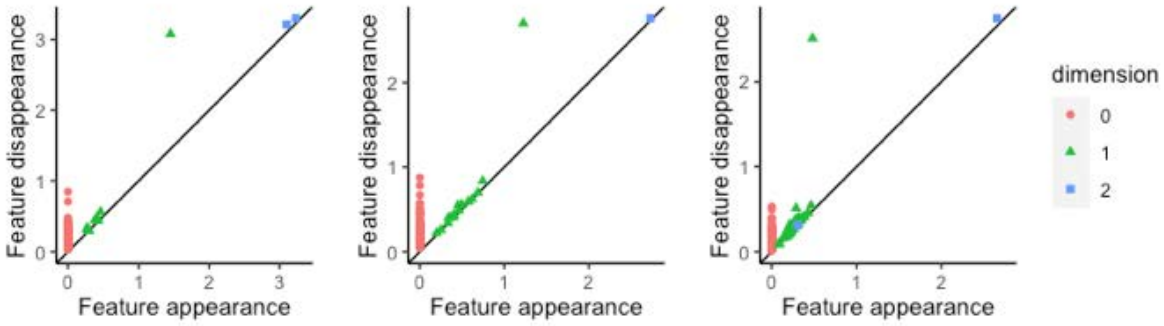
$$W_p(D_1, D_2) = \inf_{matching\ m} \left(\sum_{(u,v)\in m} \|u - v\|_\infty^p \right)^{\frac{1}{p}} \tag{5.16}$$

where $\|s\|_\infty = \max_{i=\overline{1,d}} |s_i|$ for any $s = (s_i) \in \mathbb{R}^d$.

However, we concern that the metrics are not appropriate to measure the difference between two persistence diagrams if their numbers of points, except points on the diagonal, are too different. In this case, evenly for the best matching of the diagrams, too many points in the denser diagram are matched to the diagonal because there are no other options.



(a) Point clouds



(b) The persistence digrams of $(\mathrm{Rips}_\alpha(\mathbf{V}_1))_\alpha$, $(\mathrm{Rips}_\alpha(\mathbf{V}_2))_\alpha$ and $(\mathrm{Rips}_\alpha(\mathbf{V}_3))_\alpha$, respectively

Figure 5.7: A counterexample of the Bottleneck distance and the Wasserstein distance.

For instance, let's observe point clouds $\mathbf{V}_1, \mathbf{V}_2$ and $\mathbf{V}_3$ in Figure 5.7. $\mathbf{V}_1$ is the set of 100 points determined by adding noises to the coordinates of 100 points drawn randomly on

a circumference of radius 2. The noises are i.i.d normal random variables of zero mean and standard deviation of 0.2. $\mathbf{V}_2$ is the set of 100 points drawn similarly to $\mathbf{V}_1$ but with the noises' standard deviation of 0.3. $\mathbf{V}_3$ includes 300 points such that 2/3 of them oscillate around the circumference with noises similar to $\mathbf{V}_1$ while others are points of $\mathbf{V}_2$. The persistence diagrams of the Vietoris-Rips complex filtration of these point clouds are denoted D_1, D_2 and D_3, respectively (see Figure 5.7b). Although both $\mathbf{V}_2$ and $\mathbf{V}_3$ have the same attractor as $\mathbf{V}_1$, from the statistical view, it's easier to get the attractor of $\mathbf{V}_1$ from $\mathbf{V}_3$ than from $\mathbf{V}_2$. So, D_1 must more "similar" to D_3 than D_2. Nevertheless, the Bottleneck distance and the Wasserstein distance between D_3 and D_1 approximate 0.971 and 0.977, respectively. They are too larger than such distances between D_2 and D_1, which is just about 0.377 and 0.389, respectively. This irrational result comes from remarkable differences between the numbers of points of D_1, D_2, and D_3, which equal 110, 115, and 349, respectively.

4 Detecting Anomalies of a Market Index's Dynamics from its Topological Characteristics

TDA combined with the time-delay embedding method has recently used in many studies about time series' characteristics, such as the periodicity of biological time series [Perea, 2015], the global behavior of biological aggregations [Topaz, 2015], the classification problem of volatile time series [Umeda, 2017], analyzing a bridge's deterioration based on its vibration data [Umeda, 2019]... Similarly, in this section, we use this method to detect anomalies in a stock system's behavior.

4.1 Research Methods

As a complex system, a stock market has a collective behavior that is complicated but rationally instead of randomly. Therefore, we can suppose the existence of its attractor, although the attractor is dynamical rather than fixed. This viewpoint is suggested in modern economic theory [Beinhocker, 2006; Kirman, 2011; Lewin, 1994]. Even though the attractor is changeable to adapt to the internal and external factors' movement, its change is usually not too dramatic if there are no significant impacts on the market. Because the fluctuation of a stock market's representative index can store meaningful information about the entire market's behavior, as discussed in Section 1, in [Nguyen, 2021a], we study the time series of a market index's return and consider the return's dramatically strange dynamics as the market's anomalies.

To recognize whether the index return fluctuates too differently from its historical variation, we compare the topological structure of the index return with its previous topology. This leads to comparing persistence diagrams that encode the topological information of its state space in the present period and previous periods. This information helps get principal characteristics of the index return's dynamics in the periods. Similar to machine learning, we use the terms "test data" and "training data" for the index return's time series that we want to detect anomalies and its time series in previous periods, respectively. Obviously, the period used to get the training data should be close to the periods used to get the test data.

Firstly, we use the time-delay embedding method to construct the training data's state space. The state space is a set of points in $\mathbb{R}^d$, where d is the embedding dimension. Similarly, we find the test data's state space using the same parameters of the time-delay and the embedding dimension.

Next, we divide the training data's state space into s consecutive segments having the same size as the test data's state space. Our target is detecting significant differences in the topological structures between these segments and the test data's state space. In other words, we compare the persistence diagrams associated with s point clouds received from the training data and the persistence diagram associated with the point cloud received from the test data, where all of these point clouds have the same number of points, denoted by m. So, we have s historical samples to test the current dynamics of the index return. The same size of these point clouds enables the proper observation of periodic property or timing pattern of the index return's dynamics in a period of a certain length.

For simplicity, we can compute the maximum or the average distance of the persistence diagram constructed from a historical sample and the persistence diagram constructed from the test data's state space, using some metrics such as the Bottleneck distance or the Wasserstein distance. However, we're afraid that this method is less statistical because s is not large enough. Indeed, to confirm whether a stock market behaves dramatically to fall into a recession, people often observe its circumstance in about 6 months to neglect its transient states. Hence, we use this time length for the test data in our empirical study presented below. Meanwhile, because of stock markets' adaption, like other complex systems, we shouldn't use data taken in periods that distance too much from the test period. Consequently, in our empirical study, we observe the index return within three years before the test period. Then, s is even smaller than 10 to divide the training data's state space into segments such that each segment has the same size as the test data's state space.

Accordingly, our method is merging persistence diagrams constructed from segments of the training data's state space. The result is called the total diagram. It provides all the index return's topological features that appeared in the nearest periods. Like the case demonstrated in Figure 5.7, we shouldn't directly compare the persistence diagram constructed from the test data's state space with the total diagram because the number of points of the latter can be 10 times greater than the former's. Instead, we first divide the total diagram's points, except the diagonal, into k small clusters based on their locations and the homology groups of their corresponding features. Then, for any of these clusters, each one is used to define a region of space $\mathbb{R}^2$. The rest of the space is the last region. With s persistence diagrams constructed from the training data, we easily approximate the empirical probability P_i that a point selected randomly from a persistence diagram associated with the index return in a 6-months period belongs to a certain region of $\mathbb{R}^2$ as follows:

$$P_i = \left\langle \frac{n_{ij}}{n_j} \right\rangle_{j=\overline{1,s}}, \ i = \overline{1, k+1} \tag{5.17}$$

where n_{ij} is the number of points belonging to region i in persistence diagram j, n_j is the number

of points in persistence diagram j, except the diagonal, and $\langle . \rangle_j$ is the average over all of s persistence diagrams constructed from the training data. According to our region classification, we get $P_{k+1} = 0$.

Obviously, P_is provide an empirical point distribution of a persistence diagram associated with the index return in a 6-months period, using the training data. Hence, to determine if the persistence diagram constructing from the test data implies any considerably strange topological information, we calculate the diagram's point distribution Q_i based on the same region classification, then quantify the difference between the two point distributions by the following measure:

$$\delta = \sqrt{\sum_{i=1}^{k+1} (P_i - Q_i)^2} \tag{5.18}$$

As a result, δ helps measure the deviation of the index return's topological structure from its earlier structures. A larger value of δ implies more variation of the index return's dynamics from the test period to the previous ones. Our method is summarized in Algorithm 8.

Algorithm 8 Compute the topological structure's deviation δ of the index return's dynamics from a certain period to the previous ones.

Require: index return $(x_t)_{t=\overline{1,T}}$ as training data, index return $\left(x_t^{'}\right)_{t=\overline{1,T'}}$ as test data,

procedure TOPOLOGICAL_STRUCTURE_VARIATION((x_t), $(x_t^{'})$)
  $\tau \leftarrow$ the optimal time-delay of (x_t)
  $d \leftarrow$ the optimal embedding dimension of (x_t)
  ▷ compute the time-delay embedding of (x_t) and $\left(x_t^{'}\right)$
  $\mathbf{y}_t \leftarrow \left(x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(d-1)\tau}\right),\ t = \overline{1, T-(d-1)\tau}$
  $\mathbf{y}_t^{'} \leftarrow \left(x_t^{'}, x_{t+\tau}^{'}, x_{t+\tau}^{'}, \ldots, x_{t+(d-1)\tau}^{'}\right),\ t = \overline{1, T'-(d-1)\tau}$
  $m \leftarrow T' - (d-1)\tau$ ▷ the number of vectors $\mathbf{y}_t^{'}$s
  ▷ divide $\mathbf{y}_t$ into s consecutive segments of length m
  $segment_j \leftarrow \left(\mathbf{y}_{1+(j-1)m}, \mathbf{y}_{2+(j-1)m}, \ldots, \mathbf{y}_{jm}\right),\ j = \overline{1,s}$
  ▷ compute persistence diagrams
  $dgm \leftarrow$ the persistence diagram constructed from $\left(\mathbf{y}_t^{'}\right)$
  $dgm_j \leftarrow$ the persistence diagram constructed from $segment_j,\ j = \overline{1,s}$
  $total\_dgm \leftarrow$ merging all persistence diagrams $dgm_j,\ j = \overline{1,s}$
  ▷ compute the point distribution of the persistence diagrams
  $cluster_i \leftarrow$ points assigned to the i-th group after partitioning points out of the diagonal of $total\_dgm$ into k clusters based on the points' locations and their corresponding homology groups
  for $i \in \overline{1, k+1}$ **do**
    $region_i \leftarrow$ the region of $\mathbb{R}^2$ identified by $cluster_i$, where the last region is the rest of the space
    $P_i \leftarrow \left\langle \frac{n_{ij}}{n_j} \right\rangle_{j=\overline{1,s}}$, where n_{ij} and n_j are the number of points in dgm_j belonging to $region_i$ and the number of points in dgm_j, except the diagonal, respectively
    $Q_i \leftarrow \frac{n_i^{'}}{n^{'}}$, where $n_i^{'}$ is the number of points in dgm belonging to $region_i$, and n' is the number of points in dgm, except the diagonal, respectively
  end for
  ▷ compute how the point distributions of the diagrams constructed from the test data and the training data are different
  $\delta \leftarrow \sqrt{\sum_{i=1}^{k+1} (P_i - Q_i)^2}$ ▷ Output
end procedure

4.2 Empirical Results with the S&P 500 Index

In [Nguyen, 2021a], we check our method's efficiency in the case of the S&P 500 Index. We use the daily closing values of the index to compute its daily returns from 12/18/1972 to

08/04/2020. In this period, we consider each time window of 132 trading days with 22 rolling trading days. As a result, we get 541 time windows. The index return's dynamics in each of these time windows is compared with its dynamics of 750 trading days ago. Approximately, we compare the index return's dynamics of a 6-months period with its dynamics of 3 previous years with 1-month sliding.

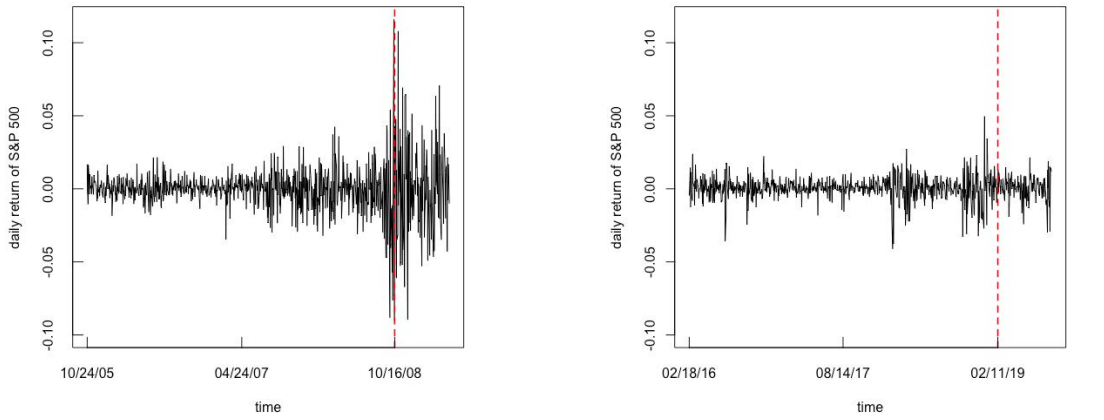
For each of these time windows, to construct the suitable time-delay embedding of the corresponding training data, we find the ideal time-delay by using the average mutual information to not neglect any nonlinear dependence. Meanwhile, the ideal embedding dimension is selected by finding the stop changing of the function defined in formula (5.10) (see subsection 2.3).

For less computation, we use the family of the Vietoris-Rips complexes to construct the persistence diagrams mentioned in Algorithm 8. As a result, we observe the lack of points representing 2-dimensional features on the diagrams. The features have small persistence and high birth scales, so they can be considered as noises. Hence, we only focus on the 0-dimensional and 1-dimensional features.

In order to partition points of the total diagram, except the diagonal, into small clusters, we need to use a clustering algorithm. Because the partition also bases on the homological dimensions of the features corresponding to the points, we can embed the points into $\mathbb{R}^3$, where the homological information is considered the third coordinate. Due to the simple arrangement of these points in $\mathbb{R}^3$, we use the k-mean algorithm [Hartigan, 1979] to fast solve the problem and get an acceptable result, as illustrated in Figure 5.9a and Figure 5.10a.

Next, we have to partition $\mathbb{R}^3$ into regions such that each contains one of the clusters above, except the last region. Equivalently, the problem is how we can assign a point in $\mathbb{R}^3$ to a given cluster. Intuitively, for any point of the persistence diagram constructed from the test data, except the diagonal, after embedding it into $\mathbb{R}^3$, we assign the embedded point to its nearest cluster. However, if the feature's persistence corresponding to the point is too different from the ones corresponding to the cluster's points, the point should be assigned to the last region. Hence, in our empirical study, we only assign a point to its nearest cluster if its corresponding persistence is not greater than the sum of the average persistence of the cluster's corresponding features and 3 times of the persistence's standard deviation.

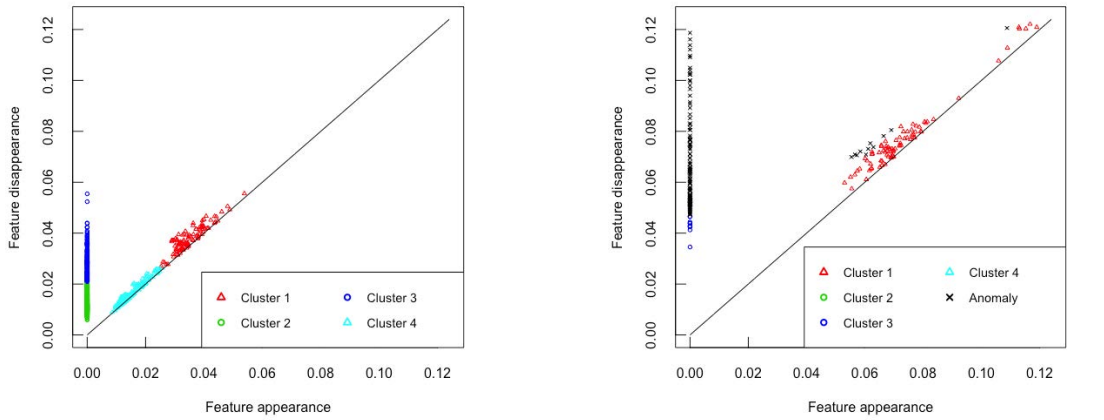
In Figure 5.8, we present two of 541 samples of our test data. The two examples demonstrate that our method can detect the significant difference by measuring the deviation of the persistence diagram constructed by the test data and the total diagram. In deed, when the test data's dynamics is too different from the historical dynamics in three years ago (Figure 5.8a), then there are so much differences between the two diagram (Figure 5.9). So, we get a large value of δ which equals 83.7%. Inversely, when the test data mostly has no dramatically different fluctuation (Figure 5.8b), the persistence diagram constructed by the test data is nearly a subset of the total diagram (Figure 5.10). As a result, we have δ is 11.9%, a very small value. Therefore, we think that δ is efficient in measuring the difference of the index return's behavior between a certain period and consecutively previous periods.



(a) Daily return of the S&P 500 Index from 10/24/2005 to 04/27/2009

(b) Daily return of the S&P 500 Index from 02/18/2016 to 08/19/2019

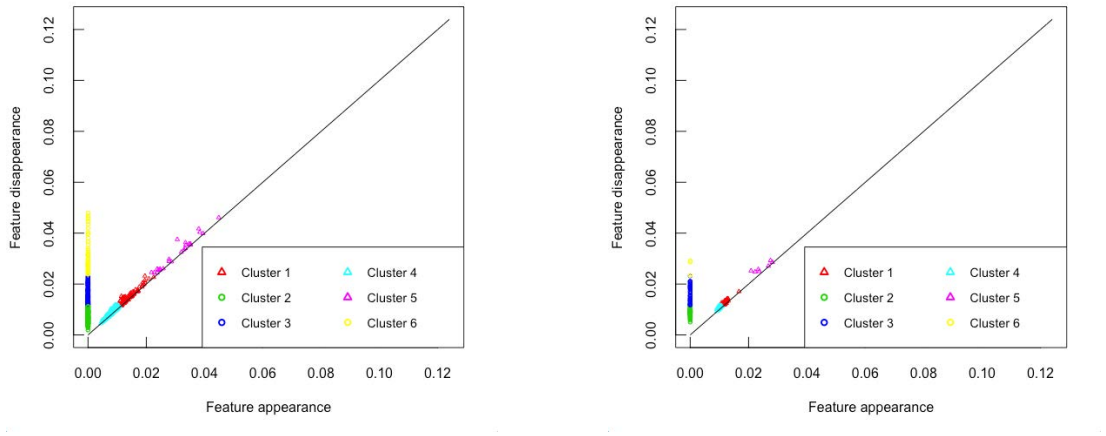
Figure 5.8: Two sample databases where the test data is on the right of the dashed line and the training data is on the left.



(a) Total diagram

(b) Persistence diagram constructed by the test data

Figure 5.9: Detecting topological anomalies of the test data in the database illustrated in Figure 5.8a. Circles represent 0-dimensional features, and triangles represent 1-dimensional features. The black sign $\times$ denotes abnormal features that cannot be assigned to any clusters of the total diagram.

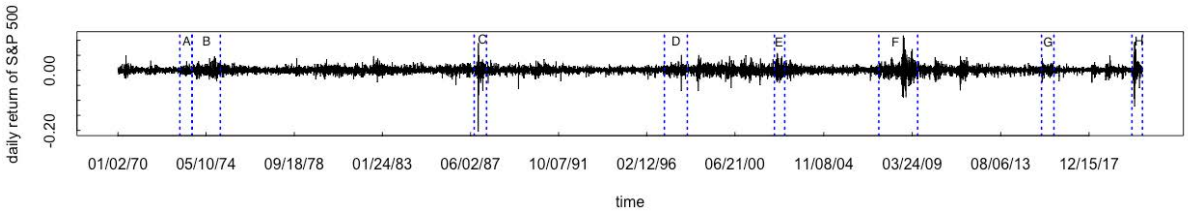


(a) Total diagram

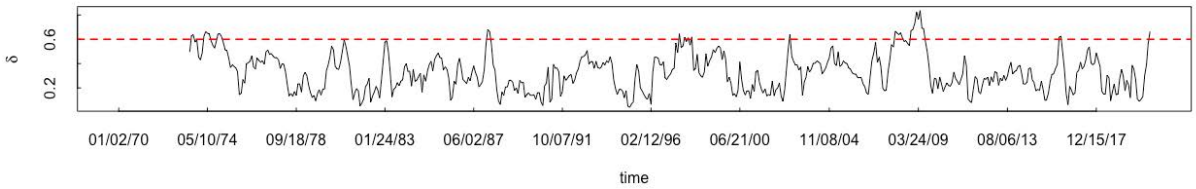
(b) Persistence diagram constructed by the test data

Figure 5.10: There is no abnormal feature in the persistence diagram constructed by the test data illustrated in Figure 5.8b. Circles represent 0-dimensional features, and triangles represent 1-dimensional features.

Especially for all 541 samples of our test data, we found that whenever δ is greater than 60%, there are serious market crashes or recessions in the corresponding test period. Figure 5.11b shows the value of δ on the corresponding test period's last day. The periods corresponding to the values greater than 60% are named from A to H in Figure 5.11a.



(a) Daily return of the S&P 500 Index from 01/02/1972 to 08/04/2020



(b) Value of δ plotted at the last days of test periods

Figure 5.11: Dynamics of δ and the S&P 500 Index's return.

In particular, in [Nguyen, 2021a], we show that the strange dynamics of the S&P 500 Index's daily return discovered in periods A and B are corresponding to the 1973 – 1974 stock market crash spreading from January 1973 to December 1974. Besides, periods B also relates to the

1970s stagflation, where the OPEC oil embargo signed on 10/19/1973 is widely blamed for causing the stagflation. Similarly, period C relates to the "Black Monday", a rapid and severe stock market crash of U.S. stock prices in late October 1987. In addition, this period is sensitive with the 1989 savings and loan crisis where more than 1000 of the country's savings and loans had failed. In fact, the crisis is an outcome of uncontrollable bad loans and losses for a long time, especially after the Federal Savings and Loan Insurance Corporation, an institution that administered deposit insurance for savings and loan institutions in the United States, had become insolvent by 1987. How about period D? It just contains a fast crash in October 1997. The crash is considered as the beginning of the end of the 1990s economic boom in the U.S. Meanwhile, period E is corresponding to the stock market downturn of 2002, also known as the internet bubble bursting with a dramatic decline in July and September 2002. In fact, the crash is just the worst result of the dot-com crash 2000 – 2002. Especially, the longest period, period F, is clearly related to the 2008 financial crisis, the worst crisis in the U.S. from the Great Depression of 1929. The crisis officially lasted from December 2007 to Jun 2009, and the bankruptcy of the investment bank Lehman Brothers in September 2008 is often thought to play a major role in the unfolding of the crisis. Period G relates to a stock market selloff occurring from August 2015 to Jun 2016. Finally, the last period is corresponding to the COVID-19 recession, which started in February 2020. In most of these recessions, except the dot-com crash 2000 – 2002, we find that although δ doesn't get such high values before the recessions occur, it still helps measure the severity of the problem. This explains why the measure can help recognize most of the crises above at the beginning when its value become greater than 60%.

As a result, we propose our method as an efficient tool to detect anomalies in the dynamics of a market index. Its result provides a simple way to recognize the beginning of a financial crisis through analyzing the corresponding stock market's representative index instead of getting a full analysis of many micro and macro statistics. Therefore, we suggest that the topological deviation δ of an index return's dynamics can be an effective measure of the systemic risk. It is especially appropriate for individual investors and auto-trading systems.

Conclusion

In this thesis, we used various techniques from complex science, including network analysis, RMT, pretopology theory, and TDA, to investigate the characteristics and mechanism of a stock market' collective behavior in different aspects. Concretely, we studied actual markets, including the U.S stock market and the Vietnamese stock market, and we used the R language to implement our empirical works.

In particular, we found that the MST of a stock market's correlation-based network is common to summarize the network's structure because the MST provides the most probable path in which a stock price shock spreads to the entire market. By contrast, the correlation-based threshold network is more appropriate than the MST in studying the market's resilience because of the MST's disadvantage in neglecting many stock correlations, which can be very large. However, since the cross-correlation matrix is computed from historical stock prices, we only get the sample matrix. According to RMT, the largest eigenvalue of the sample matrix and its associated unit eigenvector can give information about the "true" correlations of stocks because the eigenvalue is extremely greater than the upper limit of the Marčenko - Pastur distribution. So, we can use the first PC whose loadings are the eigenvector's components to study the market's collective behavior. This behavior is also reflected in the market index's dynamics.

With these tools, we provided a comprehensive analysis of the dynamics and stability of a stock market in this thesis. Firstly, after studying the dynamics of the market's MST network, we confirmed that the market's unstable state can reflect on the star-like structure of the network or the goner of the network's scale-free property. Also, this state can be quantified by the remarkable decline of different measures such as the shortest path length, the survival ratio, the same sector ratio, and the allometric coefficient. In addition, as a scale-free network, we also established by using real data that the correlation-based threshold network remains robust under random failure but very fragile under intentional attacks to its most connected nodes or its most loading nodes. This result demonstrated a stock market's robustness when some companies go bankrupt because of their wrong management. However, when the companies' common stocks play important roles in the network's structure, for instance, they are the most connected nodes or the most loading nodes, the bankruptcies will damage the network's connectivity. This negatively impacts the markets' stability.

Next, we studied the largest eigenvalue of the empirical cross-correlation matrix of stocks and its associated unit eigenvector. While other works found that the eigenvalue becomes larger

in financial crises, we suggested composing the most correlated portfolio from the first PC of stock returns. Since the eigenvalue is always dominant in the matrix's spectrum, the first PC explains most variances of stock returns. So, it can be considered as the market factor and highly correlates with the corresponding market index. In addition, we established a simple formula to approximate its loadings based on the loadings' asymptotically linear relationship with the stocks' average correlation coefficients.

On the other hand, we empirically showed the principal role of financial companies in a stock market's stability because the companies usually stay at hubs in the MST network, especially the star-like network. Also, the financial sector is dominant in the first PC's loadings.

In addition, since a market's collective behavior can be caused by a cascading failure, we proposed a method to study the failure's evolution. We considered a stock as a failed component if its price declines dramatically. With the assumption that the number of failed stocks increases the impact of these stocks on another stock and triggers its failure if the impact is large enough, we designed a pretopological space in which the pseudoclosure models the contagion of a stock group's failure. By contrast, the opening of the group's compensation can be used to predict stocks not affected by the failed stocks. We found that our pretopological framework is more efficient than the MST network and the correlation-based threshold network in modeling the cascading failure's evolution. The efficiency comes from taking into account all the stocks' correlations, obviously illustrating the contagion in individual steps and noting not only the relationship of stocks but also the relationship between a stock and a group.

Finally, we suggest a method to detect anomalies in a stock market's collective behavior. Since the market factor represented by the first PC of the stock returns often correlates highly with the market's index return, the index return's dynamic is suitable data to study the market's collective behavior. We establish a measure to recognize how topological features of the index's time series got in a certain period are different from the ones of the index's time series got in previous periods. This measure is tested in the case of the S&P 500 Index. We found that the deviation measure really helps detect significant crashes in the U.S. market when it is greater than 60%. Because it often takes such a large value from the beginning of financial crises, this value can be a warning of crises instead of spending much time analyzing many economic statistics.

As a result, this thesis helps get deep knowledge of stock markets' evolution, geometrical structures and signs of stability which are extremely valuable in controlling the systemic risk. We can improve the above researches with more appropriate models, such as pretopological spaces of $\mathcal{V}$-type for the cascading failure of stock markets, or improve the deviation measure of an index market's topological features with other tools of TDA. In addition, we also plan to study more about the first PC's role in calculating the β coefficient of a stock. In general, the scientific point of view that financial markets are complex systems opens up new theories and technologies for researching such markets' characteristics and dynamics. Therefore, this approach will continue taking more interest in our future works with other methods of complex science such as agent-based modeling.

List of Publications

Peer-reviewed articles

Quang Nguyen, N. K. Khanh Nguyen, L. H. Ngoc Nguyen. "*Dynamic topology and allometric scaling behavior on the Vietnamese stock market*". Physica A: Statistical Mechanics and its Applications.
DOI: 10.1016/j.physa.2018.09.061

Quang Nguyen, N. K. Khanh Nguyen. "*Composition of the first principal component of a stock index – A comparison between SP500 and VNIndex*". Physica A: Statistical Mechanics and its Applications.
DOI: 10.1016/j.physa.2019.04.216

N. K. Khanh Nguyen, Marc Bui. "*Modeling cascading failures in stock markets by a pretopological framework*". Vietnam Journal of Computer Science.
DOI: 10.1142/S2196888821500019

N. K. Khanh Nguyen, Marc Bui. "*Detecting anomalies in the dynamics of a market index with Topological Data Analysis*". International Journal of Systematic Innovation.
DOI: 10.6977/IJoSI.202112_6(6).0005

International conferences

N. K. Khanh Nguyen, Quang Nguyen. "*Resilience of stock cross-correlation network to random breakdown and intentional attack*". Proceedings of the 1st International Econometric Conference of Vietnam, Ho Chi Minh, Vietnam, 2018.
DOI: 10.1007/978-3-319-73150-6_44

N. K. Khanh Nguyen, Quang Nguyen, Marc Bui. "*Mining stock market time series and modeling stock price crash using a pretopological framework*". Proceedings of the 11th International Conference on Computational Collective Intelligence, Hendaye, France, 2019.
DOI: 10.1007/978-3-030-28377-3_53

Appendix. Thesis Abstract in French

Introduction

Le marché financier joue un rôle important dans toutes les économies, et donc tous les acteurs du marché sont très préoccupés par son évolution et sa stabilité. Avec des théorie économique moderne, les systèmes financiers peuvent être considérés comme des systèmes complexes.

Dans cette thèse, en utilisant des techniques de science complexe, nous étudions les comportements collectifs des marchés boursiers, les components principaux sur lesquels se concentrent la plupart des ressources financières dans les économies. La structure de cette thèse comprend cinq chapitres réalisant les contenus suivants:

Le chapitre 1 fournit la litérature des systèmes complexes et des approches communes utilisées pour étudier de tels systèmes. En outre, l'idée que les marchés financiers sont considérés comme des systèmes complexes est également présentés en détails dans ce chapitre.

Dans le chapitre 2, nous introduisons le réseau basé sur la corrélation, qui est souvent utilisé pour modéliser les interactions mutuelles entre les composants d'un système complexe. Dans notre contexte financier, ce réseau permet de modéliser le co-mouvement des prix boursiers dans un marché. Avec deux sous-réseaux spéciaux, l'arbre couvrant minimal et le réseaux à seuils basé sur la corrélation, nous pouvons utiliser les outils de la théorie des graphes et la relation d'échelle allométrique pour étudier la structure géométrique du marché et sa résilience en cas de défaillances aléatoires et d'attaques intentionnelles. Le résultat est important pour obtenir des informations sur la stabilité du marché ainsi que sa robustesse.

Comme nous travaillons toujours avec des matrices empiriques de corrélation croisée des actions, la théorie des matrices aléatoires, présentée au chapitre 3, aide à trouver l'interaction "essentiele" entre les composants d'un marché boursier. Cette théorie est utile pour étudier le spectre d'une matrice empirique de corrélation croisée. Nous nous concentrons particulièrement sur la plus grande valeur propre et son vecteur propre associé ayant le module unitaire. Pour examiner leurs rôles dans notre problème financier, nous utilisons la méthode de l'analyse en composantes principales. Étant donné que la première composante principale des rendements boursiers joue le rôle du facteur de marché, nous fournissons non seulement des analyses profonds de la première composante principale, mais également une estimation de ses chargements dans ce chapitre.

D'autre part, le comportement collectif d'un système complexe est parfois causé par une défaillance en cascade. Pour capturer l'évolution de l'échec en cascade, nous utilisons la théorie de la prétopologie présentée au chapitre 4. Dans ce chapitre, nous proposons un modèle pré-

topologique pour modéliser la diffusion des actions de détresse dans un marché boursier.

Enfin, au chapitre 5, nous étudions comment détecter les dynamiques anormales du comportement collectif d'un marché boursier. Bien que l'étude de la dynamique de la structure du réseau ou de la première composante principale des fluctuations des actifs puisse résoudre ce problème, nous utilisons une autre approche basée sur l'indice représentatif du marché car ces données sont transparentes, mises à jour en continu et gratuites. Pour comprendre les caractéristiques importantes de la dynamique d'un indice de marché, nous utilisons l'analyse topologique des données combinée à l'intégration de retard pour obtenir des informations topologiques sur l'espace d'état de la dynamique. Le résultat devrait donner des avertissements sur les crises sans analyser beaucoup de statistiques micro et macro.

De plus, nous menons des études empiriques sur le marché boursier américain et le marché boursier vietnamien pour comparer nos résultats dans deux cas différents - un marché développé et un marché émergent. Plus concrètement, nous utilisons les composantes de l'indice S&P 500, y compris les actions ordinaires de 500 sociétés à grande capitalisation sur la bourse américaine, et les composantes de l'indice VN, y compris toutes les actions cotées de la bourse de Hochiminh (HSX), pour représenter les deux marchés. Ces deux indices détiennent plus de 80% de la capitalisation des marchés correspondants. Tous nos travaux empiriques, y compris le traitement des données, la modélisation, l'analyse des résultats statistiques et le traçage, sont mis en œuvre à l'aide du langage R.

Notre résultat permet de comprendre le mécanisme et les caractéristiques des marchés boursiers, et plus généralement des marchés financiers, tels que la structure géométrique, la transition de phase, la robustesse, l'approximation du facteur de marché, l'évolution des défaillances en cascade et la dynamique des marchés. Ces informations sont importantes pour obtenir une vue d'ensemble de la dynamique et de la stabilité d'un marché boursier et, par la suite, aider à construire des outils utiles pour gérer le risque systémique afin d'éviter des récessions dramatiques.

Chapitre 1

Introduction aux systèmes complexes

1 Qu'est-ce qu'un système complexe?

Fréquemment, un système complexe est supposé d'être un grand système ayant les caractéristiques suivantes [Foote, 2007; Ladyman, 2013; McCarthy, 2000; Newman, 2011]: non-linéarité, adaptation, émergence et auto-organisation. Il existe un certain nombre de systèmes complexes, varié des systèmes physiques aux systèmes sociaux [Newman, 2011].

2 Sciences complexes

La science complexe est un domaine interdisciplinaire qui nécessite des contributions de nombreuses disciplines diverses, notamment la physique statistique, la théorie de l'information, la dynamique non linéaire, l'anthropologie, l'informatique, la météorologie, la sociologie, l'économie, la psychologie et la biologie.

Les études de systèmes complexes peuvent être divisées en deux approches. La première comprend des études sur la structure des systèmes. La seconde comprend des études sur le processus dynamique des systèmes. Ces deux approches peuvent se combiner et se compléter car mieux la structure d'un système complexe est comprise, plus sa dynamique est décrite avec précision.

3 Outils fondamentaux de l'analyse des systèmes complexes

Pour modéliser des systèmes complexes et étudier leur dynamique, les réseaux et la modélisation à base d'agents sont des approches typiques.

3.1 Modélisation à base d'agents

La modélisation à base d'agents est une approche "bottom-up" qui simule séparément et individuellement les agents d'un système complexe et leurs interactions, permet aux comportements émergents du système d'apparaître naturellement plutôt que de les mettre en place à la main [Berry, 2002]. L'inconvénient de cette approche est le manque de théories et de modèles à l'appui, car elle dépend principalement de l'intelligence artificielle et de la simulation informatique. Par conséquent, il est conservé pour nos recherches ultérieures.

3.2 Science des réseaux

Une représentation simplifiée d'un système complexe peut être un graphe dont les nœuds représentent les composants du système, et chaque arête représente l'interaction entre deux composants. Le graphe est appelé un réseau complexe. Cependant, le grand nombre de composants d'un système complexe, leurs multi-relations, ainsi que l'hétérogénéité des composants posent des problèmes dans la construction de leurs représentations en réseau.

En raison de la complexité d'un système complexe, nous devons l'étudier par différentes approches pour obtenir une compréhension complète de sa structure et de sa dynamique. Dans ce travail, nous proposons la théorie des graphes, la théorie des matrices aléatoires et la théorie de la prétopologie comme approches efficaces car ces théories fournissent un ensemble complet d'outils mathématiques, informatiques et statistiques qui peuvent être utilisés pour analyser, modéliser et comprendre des systèmes complexes.

4 Les marchés financiers en tant que systèmes complexes

Du point de vue scientifique, les marchés financiers peuvent être considérés comme des systèmes complexes. En fait, un marché financier est un ensemble de nombreux composants tels que des obligations, des actions, des produits dérivés, des devises, des banques, des matières premières qui interagissent les uns avec les autres et ont la capacité d'apprendre et de changer les comportements à partir de leurs expériences.

Considérer les marchés financiers comme des systèmes complexes fournit un nouvel ensemble de théories et de techniques pour comprendre ou expliquer le mécanisme et les effets des phénomènes économiques tels que les transitions de phase, la distribution des rendements de prix "fat-tail", les phénomènes de regroupement de volatilité, les défaillances en cascade, les crises, dynamisme plutôt qu'équilibre... Une meilleure compréhension de ces phénomènes peut permettre d'améliorer la stabilité des marchés, de prévoir les pires scénarios ou d'évaluer les politiques potentielles.

Bien qu'un marché financier contienne différentes parties, les marchés boursiers sont notre sélection pour les deux raisons suivantes. Premièrement, les marchés boursiers sont les principaux où se concentrent la plupart des ressources financières. Deuxièmement, leurs données historiques sont mise-a-jour et de manière transparente. Ceci est très important pour nos études empiriques, en particulier les études sur les marchés en développement.

Chapitre 2

Marchés financiers sous représentations en réseau

1 Réseaux basés sur la corrélation dans les marchés financiers

L'une des représentations de réseau populaires des systèmes financiers est le réseau basé sur la corrélation [Bonanno, 2004]. Concrètement, pour N actions $i = \overline{1,N}$, soit $S_i(t)$ le prix de l'action i à l'instant t ($i = \overline{1,N}$).

Définition 2.1. *La matrice $N \times N$ $\mathbf{C} = (c_{ij})$ est appelée la matrice de corrélation croisée des actions données si*

$$c_{ij} = \frac{\langle r_i(t).r_j(t)\rangle - \langle r_i(t)\rangle.\langle r_j(t)\rangle}{\sigma_i \sigma_j}, \quad i,j = \overline{1,N} \tag{2.1}$$

où $r_i(t) = \ln(S_i(t)) - \ln(S_i(t-1))$ est le log-changement du action i à l'instant t; $\langle . \rangle$ désigne la moyenne temporelle de la variable interne; σ_i et σ_j sont l'écart type de r_i et r_j, respectivement. Nous appelons r_i le rendement de l'action i et c_{ij} le coefficient de corrélation entre l'action i et l'action j.

Sur la base du coefficient de corrélation, une distance métrique est construite pour obtenir un arrangement topologique du système d'action. Dans cette étude, nous utilisons la distance discutée dans [Gower, 1966]:

Définition 2.2. *La distance entre l'action i et l'action j est définie par la transformation non linéaire suivante du coefficient de corrélation c_{ij} entre ces actions:*

$$d_{ij} = \sqrt{2(1 - c_{ij})} \tag{2.2}$$

La matrice $N \times N$ $\mathbf{D} = (d_{ij})$ est appelée la matrice des distances.

Définition 2.3. *Le réseau basé sur la corrélation de valeurs données est le graphe dont les nœuds représentent les valeurs, et la matrice d'adjacence est la matrice de distance construite à partir de la matrice de corrélation croisée des valeurs.*

Parce que le réseau d'actions basé sur la corrélation est entièrement connecté, il contient tous les co-mouvements possibles de paires de valeurs d'actifs et leurs points forts dans le système d'actions.

Dans la thèse, nous sommes d'accord sur les points suivants. Premièrement, parce que nous ne connaissons pas les corrélations exactes entre les actions, la notation "matrice de corrélation croisée" fait référence à la matrice de corrélation croisée empirique obtenue à partir des données historiques des actifs. Deuxièmement, nous ne prêtons attention qu'à la fluctuation quotidienne des cours des actions, donc, dans tous les exemples empiriques ci-dessous, à l'exception de ceux référencés dans d'autres études, la base de données correspond aux cours de clôture quotidiens des actions. Enfin, tous les réseaux ou graphiques abordés dans les énoncés suivants de cette proposition ne sont pas directioné, à moins qu'il n'y ait des informations supplémentaires.

2 Sous-graphes importants d'un réseau basé sur la corrélation

Pour avoir un sous-graphe contenant suffisamment d'informations importantes sur la relation entre les nœuds du réseau d'origine, nous construisons les sous-graphes suivants: l'arbre couvrant minimal (MST) du réseau et le sous-graphe des nœuds hautement connectés.

2.1 Arbre couvrant minimal

Définition 2.4. *Un arbre couvrant minimal d'un réseau pondéré est un sous-graphe qui est*

(i) connected, c'est-à-dire que le sous-graphe contient tous les nœuds du réseau d'origine et qu'il existe un chemin pour aller de n'importe quel nœud à un autre,

(ii) formé un arbre, c'est-à-dire que le sous-graphe n'a pas de nœud qui revient sur lui-même, et

(iii) satisfait (i) et (ii) avec le poids de bord total minimum.

Le MST est le chemin le plus probable qui fait que la transmission d'un choc de prix se propage à travers le marché [Lautier, 2013; Marti, 2021]. Cependant, en raison de la condition acylique, le MST d'un réseau d'actions basé sur la corrélation présente une faiblesse considérable: certaines arêtes associées à de faibles poids et certaines arêtes associées à des corrélations élevées peuvent être négligés. Par conséquent, les connexions de l'arbre peuvent ne pas bien définir les clusters du système correspondant qui suivent souvent les secteurs d'activité.

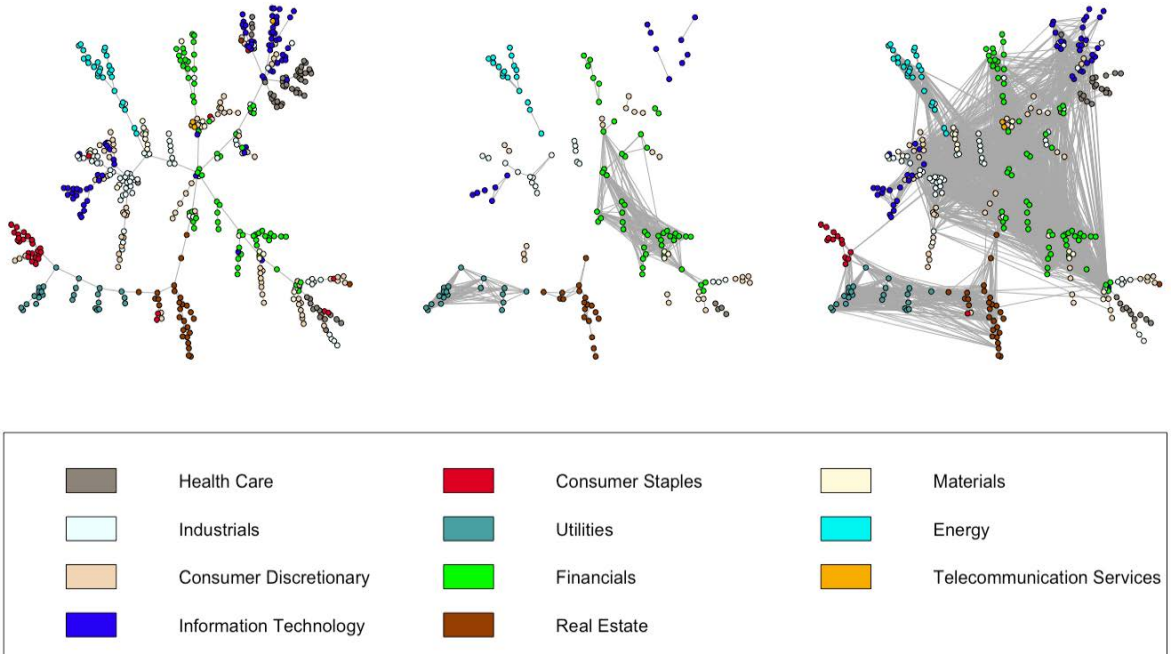
Nous proposons que le MST soit un simple sous-graphe du réseau basé sur la corrélation qui soit suffisamment efficace pour déduire des informations importantes sur les marchés boursiers, en particulier les marchés émergents où les corrélations entre les actions cotées et certaines actions importants/particuliers peuvent être supérieures aux intra-corrélations des secteurs économiques [Nguyen, 2019b; Nguyen, 2019c]. Comprendre la structure du MST est vraiment important pour gérer le risque systémique.

2.2 Réseaux à seuils basé sur la corrélation

Sous la représentation MST, un réseau basé sur la corrélation semble plus fragile qu'il ne

l'est. Ainsi, nous considérons le réseaux à seuils basé sur la corrélation où nous ne conservons que les actions et les arêtes associés à des corrélations d'actions suffisamment élevées. Le réseaux à seuils basé sur la corrélation qui a le même nombre d'arêtes que le MST est appelé le graphe d'actifs [Garas, 2008; Onnela, 2003b; Onnela, 2004]. Bien que le graphe des actifs semble mieux refléter la partition du réseau boursier associée aux secteurs économiques exceptionnels que le MST [Onnela, 2003b; Onnela, 2004], le graphe des actifs manque d'une quantité considérable d'actions du réseau d'origine. Par conséquent, il néglige d'autres informations sur l'ensemble du marché (la figure 2.1b).

Un seuil approprié pour les corrélations d'actions permet de réduire la taille du réseau basé sur la corrélation d'origine tout en créant un graphique représentatif du marché. En particulier, lors de l'analyse des caractéristiques d'un système d'action, le réseaux à seuils basé sur la corrélation permet d'éviter les bruits causés par des connexions instables. Dans notre étude, le seuil est le quantile à 97% des corrélations empiriques des actions. Par exemple, la figure 2.1c montre le réseau basé sur la corrélation du marché américain où le seuil est de 0.63. Le réseau contient 71.81% de nœuds du réseau basé sur la corrélation d'origine alors qu'il ne contient que 3% des connexions les plus importantes du réseau d'origine. Le quantile à 97% des corrélations empiriques entre les actions du marché boursier vietnamien n'est inférieur que de 0.25, car les corrélations boursières sur les marchés émergents sont généralement plus faibles que celles des marchés développés.



(a) Le MST (b) Le graphique des actifs (c) Le réseaux à seuils basé sur la corrélation où le seuil est le quantile à 97% des corrélations empiriques d'action.

Figure 2.1: Quelques sous-graphes du réseau basé sur la corrélation des actions cotées sur le NYSE du 01/04/2015 au 01/04/2020.

En général, le MST est utile pour étudier la structure globale du réseau et le problème de propagation des prix des chocs. Pendant ce temps, le réseaux à seuil basé sur la corrélation avec un seuil approprié est plus efficace pour étudier la robustesse du réseau.

3 Mesures structurelles des réseaux financiers

Nous utilisons différentes mesures pour analyser la structure et les caractéristiques d'un réseau boursier:

3.1 Répartition des degrés

Définition 2.5. *Dans un réseau, le degré d'un nœud est le nombre d'arêtes qui lui sont connectées.*

Le degré d'un nœud permet de mesurer le niveau de connectivité du nœud.

Définition 2.6. *Dans un réseau, soit $P(k)$ la fraction du nombre de nœuds de degré k. Un histogramme de $P(k)$ est appelé la distribution des degrés du réseau.*

De manière équivalente, nous pouvons définir $P(k)$ comme la probabilité qu'un nœud du réseau ait un degré de k.

3.2 Longueur moyenne du plus court chemin

Définition 2.7. *Un chemin reliant une paire de nœuds dans un réseau est une séquence d'arêtes qui relie les deux nœuds. La longueur du chemin est le poids total des tronçons appartenant au chemin si le réseau est pondéré, et est égale au nombre de ces tronçons si le réseau est non pondéré.*

Définition 2.8. *La longueur moyenne du chemin le plus court d'un réseau est la longueur moyenne des chemins les plus courts pour toutes les paires de nœuds possibles dans un réseau, c'est-à-dire,*

$$L = \frac{\sum_{(i,j)} l(i,j)}{N(N-1)} \tag{2.3}$$

où N est le nombre de nœuds et $l(i,j)$ est le chemin le plus court du nœud i au nœud j.

La longueur moyenne du plus court chemin est une caractérisation intuitive de la sensibilité du marché actuel à un choc.

3.3 Centralité intermédiaire

Définition 2.9. *La centralité d'intermédiarité du nœud i est donnée par:*

$$b\,(i) = \sum_{j \neq i \neq k} \frac{s^i_{jk}}{s_{jk}} \tag{2.4}$$

où s_{jk} est le nombre de chemins les plus courts reliant le nœud j, et le nœud k et s^i_{jk} est le nombre de ces chemins qui passent par le nœud i (pas où i est un point de terminaison).

Le nœud avec la centralité d'intermédiarité la plus élevée est le nœud qui relie les «régions» de nœuds denses.

3.4 Composant géant

Définition 2.10. *Un composant géant ou cluster géant est le cluster connecté d'un réseau qui contient une proportion importante de l'ensemble des nœuds du réseau, même lorsque la taille du réseau augmente.*

En générale, le composant géant d'un réseau est vaguement associé au plus grand cluster. Le théorème suivant fournit le critère de Molloy-Reed [Cohen, 2000; Molley, 1995] pour l'existence d'une composante géante dans un réseau aléatoire non corrélé:

Théorème 2.1. *Dans un réseau aléatoire non corrélé avec une distribution de degrés $P(k)$, une composante géante existe si*

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} \geq 2 \tag{2.5}$$

où $\langle k \rangle$ et $\langle k^2 \rangle$ sont le premier et le deuxième moment de $P(k)$.

Une prédiction clé de la théorie de la percolation est que la décomposition d'un réseau par suppression de nœuds n'est pas un processus graduel avec la fraction q de nœuds supprimés. En réalité, avec un nombre important de nœuds endommagés, de nombreux réseaux complexes sont incapables de maintenir leur fonctionnement normal. Le seuil q_c peut être considéré comme la valeur telle que la composante géante est détruite au franchissement de q. En utilisant le critère de Molloy-Reed, Cohen et al. [Cohen, 2000] a montré la relation entre q_c et κ comme suit:

Théorème 2.2. *Dans un réseau aléatoire non corrélé, on obtient:*

$$1 - q_c = \frac{1}{\kappa_0 - 1} \tag{2.6}$$

où $\kappa_0 = \frac{\langle k_0^2 \rangle}{\langle k_0 \rangle}$ est calculé à partir de la distribution initiale avant la répartition aléatoire.

3.5 Relation d'échelle allométrique

Pour étudier la propriété structurelle de l'arbre à travers la relation d'échelle allométrique, il faut d'abord attribuer une direction à chaque arête si l'arbre n'est pas orienté. La règle est que les bords reliant un nœud et le hub avec le degré le plus élevé doivent s'étendre à partir du hub. D'autres arêtes doivent atteindre le nœud qui se connecte au concentrateur avec un nombre inférieur d'arêtes. Nous appelons temporairement le résultat de cette affectation de direction l'arbre couvrant dirigé. Ensuite, la relation d'échelle allométrique est choisie par la relation de loi de puissance entre deux variables A et C calculées pour chaque nœud du réseau. Ces variables sont trouvées de manière itérative comme suit [Qian, 2010] :

Définition 2.11. *Pour chaque nœud i dans l'arbre couvrant dirigé, soit*

$$A_i = \sum_j A_j + 1, \qquad C_i = \sum_j C_j + A_i, \tag{2.7}$$

où j représente tous les nœuds liés à partir du nœud i. Alors, l'exposant allométrique η est la puissance d'ajustement de l'expression suivante:

$$C \sim A^{\eta} \tag{2.8}$$

où les nœuds feuilles avec $A = C = 1$ doivent être éjectés de l'ajustement de l'exposant.

L'exposant allométrique η est compris entre 1 et 2 pour deux structures de réseau extrêmes: le réseau en étoile et le réseau en chaîne. Ainsi, un arbre en forme d'étoile a $\eta = 1^{+}$ tandis que pour un arbre en forme de chaîne a $\eta = 2^{-}$ [Garlaschelli, 2003; Qian, 2010]. En particulier, la valeur de C à un nœud peut mesurer l'impact total du nœud vers le réseau à travers ses k-voisins les plus proches, où le niveau de proximité k va à l'infini.

Nous démontrons empiriquement que la relation d'échelle allométrique apparaît réellement dans le réseau MST d'un système d'action. Ainsi, la relation d'échelle allométrique du MST associée à un système financier peut aider à quantifier la "forme" globale du système et à déterminer le niveau d'influence de chaque constituant sur les autres dans le système.

3.6 Taux de survie

Définition 2.12. *Soient G_t et G_{t-1} deux graphes consécutifs représentant un réseau complexe et E_t et E_{t-1} l'ensemble des arêtes de G_t et G_{t-1}, respectivement. Ensuite, le taux de survie entre G_t et G_{t-1} est défini par l'expression suivante:*

$$S\left(G_t, G_{t-1}\right) = \frac{2\left\|E_t \cap E_{t-1}\right\|}{\|E_t\| + \|E_{t-1}\|} \tag{2.9}$$

où $\|.\|$ désigne la taille de l'ensemble intérieur, c'est-à-dire le nombre d'éléments de l'ensemble.

Le taux de survie permet de mesurer l'évolution de la structure du réseau dans le temps. Cette mesure a été introduite dans [Garas, 2008; Onnela, 2003b; Onnela, 2003c]. Cependant, dans notre étude, nous remplaçons leur dénominateur par le nombre moyen d'arêtes des deux graphes consécutifs pour éviter les bruits causés par l'augmentation de la taille du réseau.

3.7 Ratio du même sector

Définition 2.13. *Le ratio du même sector du réseau boursier est la fraction du nombre d'arêtes qui relient deux cours appartenant au même secteur d'activité.*

Fréquemment, les actions du même secteur d'activité sont généralement plus corrélées que les actions de différents secteurs, à l'exception de certains secteurs particuliers tels que les services financiers. Cependant, dans certaines crises, il arrive que l'intra-corrélation d'un secteur n'est presque pas plus élevée que l'inter-corrélation. Par conséquent, des changements significatifs du même ratio sectoriel du MST peuvent donner des informations utiles sur les changements cruciaux de la structure du MST ainsi que sur la phase du marché.

4 Caractéristiques des réseaux boursiers

La plupart des réseaux complexes réels ont une propriété commune appelée propriété sans échelle, bien qu'ils puissent être construits à partir d'objets de natures différentes [Newman, 2003].

4.1 Propriété sans échelle

Définition 2.14. *Un réseau sans échelle est un réseau dont la distribution des degrés suit une loi de puissance, c'est-à-dire,*

$$P(k) \sim k^{-\gamma} \tag{2.10}$$

La constante positive γ est appelée le degré exposant de la distribution.

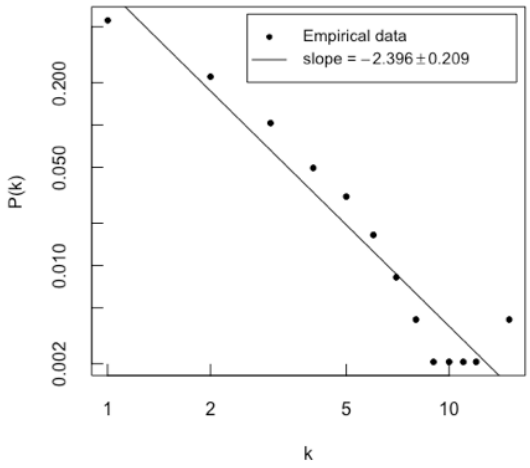
Remarque. Un réseau sans échelle avec $\gamma > 1$ a les caractéristiques suivantes :

(i) Il pourrait avoir des nœuds centraux avec des degrés extrêmement élevés (souvent appelés "hubs").

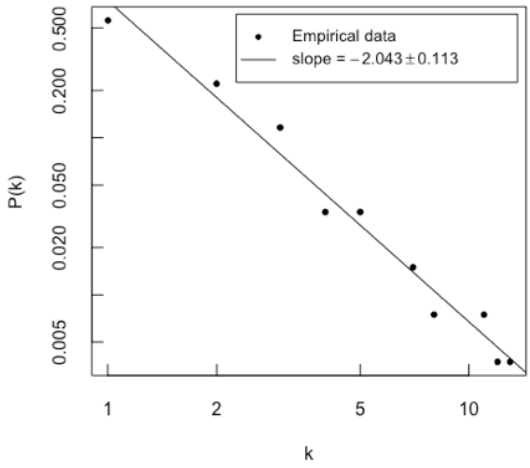
(ii) Le degré du plus grand hub croît avec la taille du réseau.

(iii) Comparé à des réseaux aléatoires ayant la même valeur attendue, il manque d'échelle interne

Semblable aux marchés développés [Onnela, 2003b; Sienkiewicz, 2013; Wiliński, 2013], dans [Nguyen, 2018; Nguyen, 2019c], nous avons constaté que le MST et le réseaux à seuils basé sur la corrélation des actions vietnamiennes sont presque sans échelle dans sa situation normale, en particulier le MST. De plus, l'exposant de degré du marché émergent est toujours plus petit que celui du marché développé, donc la structure du premier est plus dense que celle du second (les figures 2.2 et 2.3).

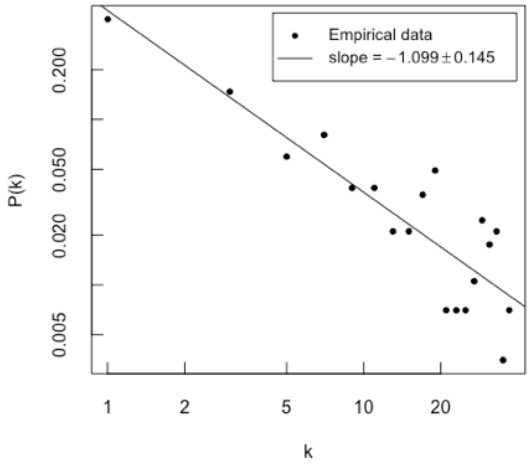


(a) Réseaux d'actions compris dans l'indice S&P 500

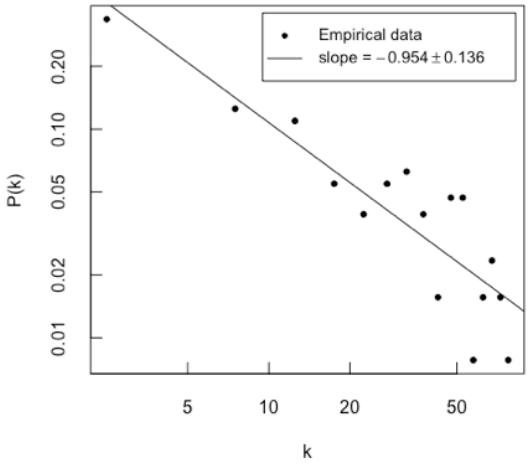


(b) NRéseaux d'actions cotées sur le HSX

Figure 2.2: Distribution en degrés des réseaux d'actions modélisés par la méthode MST sur la période 01/01/2017 – 01/01/2019.



(a) Degré de distribution du réseau d'actions compris dans l'indice S&P 500 avec le seuil de 0.63



(b) Degré de répartition du réseau d'actions cotées sur le HSX avec le seuil de 0.25

Figure 2.3: Distributions en degrés du réseau de seuils de corrélation des actions sur le marché américain et sur le marché vietnamien sur la période 01/01/2017 – 01/01/2019.

4.2 Résilience du réseau

Dans cette section, à l'aide d'une représentation graphique d'un système financier, nous quantifierons la capacité du système à maintenir son fonctionnement malgré les défauts de certains composants. Cette capacité est connue sous le nom de *résilience du réseau*. Nous

quantifions le niveau de résilience du réseau comme la fraction de suppression de nœuds telle que le réseau conserve sa connectivité globale, comme indiqué dans [Callaway, 2000; Cohen, 2000; Molley, 1995]. Nous mesurons le niveau de résilience d'un réseau d'action sous deux types de suppressions de nœuds: les pannes aléatoires et les attaques. Un réseau en panne aléatoires signifie qu'une partie arbitraire de ses nœuds est endommagée. En revanche, un réseau est attaqué si certains de ses nœuds les plus importants sont intentionnellement endommagés.

Comme discuté dans la section 3, le niveau de résilience est calculé par le seuil critique q_c. En utilisant le théorème 2.2, nous présentons le résultat de Cohen et al. [Cohen, 2000] avec des informations supplémentaires pour le cas $\gamma = 2$ comme suit:

Théorème 2.3. *Dans un réseau sans grande échelle avec un exposant de degré γ, sous une suppression aléatoire de ses nœuds,*

- *pour $\gamma > 3$, le seuil critique q_c se rapproche de $1 - \left(\frac{2-\gamma}{3-\gamma} k_{\min} - 1\right)^{-1}$, où $k_{\min}$ sont la plus petite connectivité possible.*
- *pour $1 < \gamma < 3$, le seuil critique q_c se rapproche de 1.*

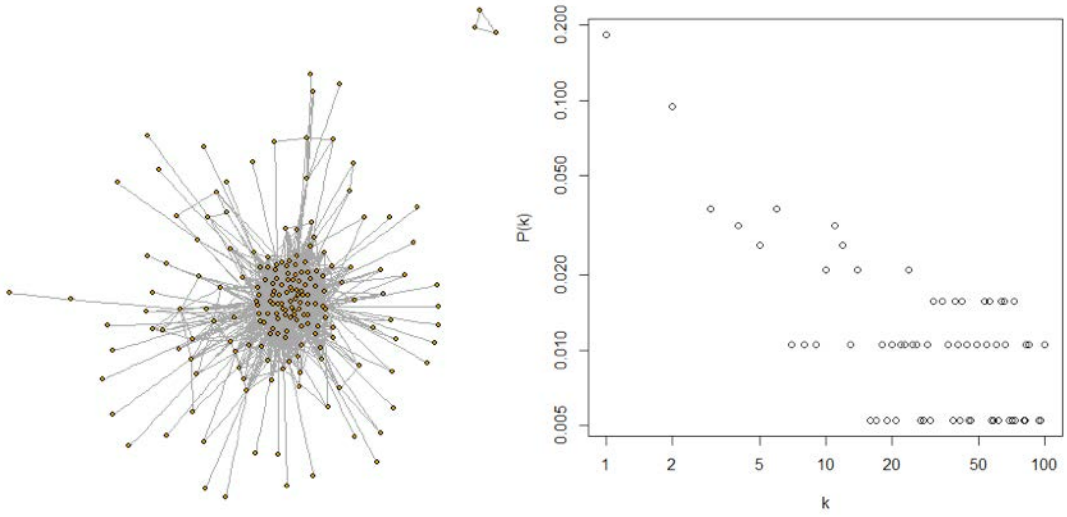
De nombreux réseaux complexes réels sont des réseaux sans échelle avec des exposants de degré allant principalement de 1^+ à 3^- [Cohen, 2010]. Par conséquent, le théorème 2.3 confirme que les réseaux ont un niveau de résilience extrêmement élevé en cas de pannes. En revanche, Cohen et al. [Cohen, 2001] a démontré théoriquement la vulnérabilité d'un réseau sans échelle contre l'attaque intentionnelle des nœuds les plus connectés:

Théorème 2.4. *Dans un réseau sans grande échelle avec un exposant de degré γ, sous une attaque intentionnelle vers les nœuds les plus connectés, la probabilité qu'un bord se connecte à un nœud supprimé est d'environ 1 pour $1 < \gamma \leq 2$, et se rapproche de $q^{\frac{2-\gamma}{1-\gamma}}$ pour $\gamma > 2$, où q est la fraction de nœuds attaqués.*

Pour un réseau sans grande échelle avec un exposant de degré $1 < \gamma \leq 2$, le théorème 2.4 montre que les attaquants n'ont besoin que d'une petite connaissance des hubs du réseau pour le casser entièrement.

Dans notre contexte financier, pour étudier la résilience d'un système boursier, nous utilisons le réseau de seuils basé sur la corrélation. Dans [Nguyen, 2018], nous effectuons une simulation de deffaillant aléatoire d'un tel réseau d'actions cotées sur le HSX (la figure 2.4). Le seuil sélectionné est de 0.25, ce qui équivaut au quantile à 97% des corrélations empiriques d'action [1].

[1] Le résultat similaire est trouvé lorsque nous vérifions avec différents seuils tels que 0.3, 0.35, 0.4...



(a) Le réseau avec le seuil de 0.63 (b) Répartition des dégré

Figure 2.4: Le réseaux à seuils basé sur la corrélation des actions cotées sur le HSX dans la période 01/01/2017 – 01/01/2019 et sa distribution de degré.

Nous utilisons la méthode de Monte-Carlo et le critère de Molley-Reed pour calculer le seuil critique q_c. Nous avons constaté que le réseau était extrêmement robuste en cas de panne aléatoire avec le seuil critique de 95% malgré le faible exposant de 1.3. Ensuite, nous attaquons les nœuds les plus importants du réseau, en utilisant respectivement les mesures suivantes pour définir le niveau affectant d'un nœud: les degrés initiaux des nœuds (ID), la centralité d'intermédiarité initiale des nœuds (IB), les degrés recalculés des nœuds après chaque suppression de nœud (RD) et la centralité d'intermédiarité recalculée des nœuds après chaque suppression de nœud (RB). De ce fait, nous avons constaté que le réseau est fragile sous les attaques intentionnelles, notamment sous la stratégie ID (la figure 2.5). Cependant, nous proposons que si nous voulons endommager un réseau d'action uniquement à un niveau de sa taille plutôt que de le détruire complètement, la stratégie RB peut être une meilleure option. Ces résultats aident à construire un marché boursier stable et à le protéger efficacement.

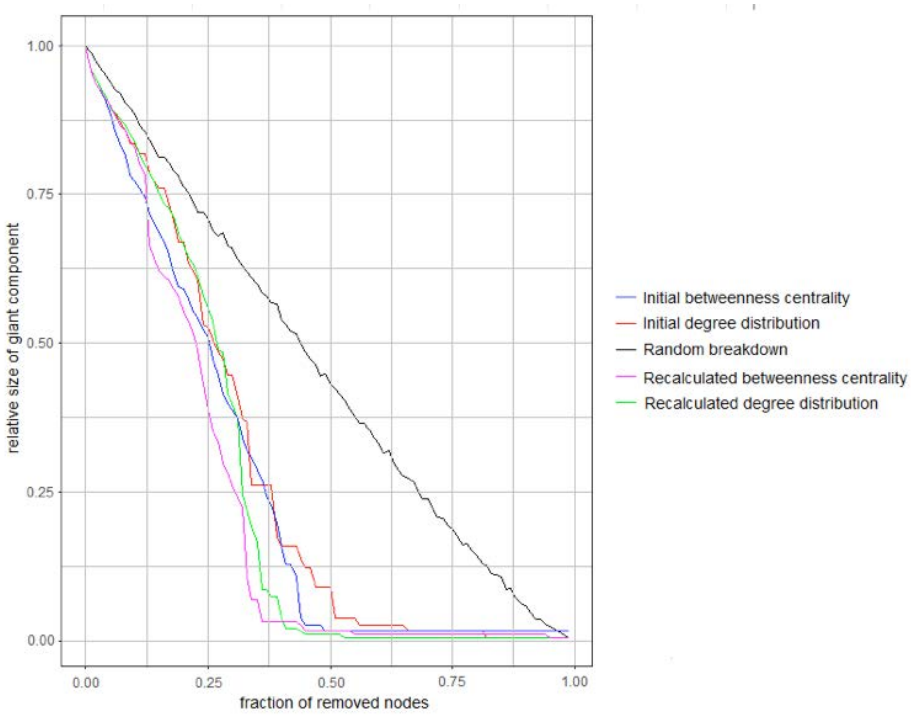


Figure 2.5: La taille relative de la composante géante en fonction de la fraction de nœuds supprimés sous l'échec aléatoire des nœuds et des différentes stratégies d'attaque au réseau de seuil basé sur la corrélation des actions cotées sur le HSX dans la période 01/01/2017 – 01 /01/2019.

4.3 Transitions de phase

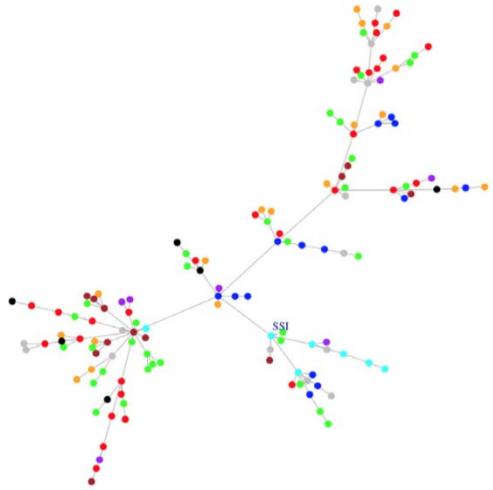
Pour étudier la propagation d'un choc de prix d'une action à l'ensemble du marché, le réseau MST est plus adapté que le réseaux à seuil basé sur la corrélation. Dans cette section, notre sujet de recherche est l'évolution de la structure du réseau MST dans le temps. Comprendre les transitions de phase de la structure du MST permet d'évaluer la stabilité du marché et de contrôler le risque systémique.

Dans certaines marchés développés, on a observé que [Wiliński, 2013; Sienkiewicz, 2013]:

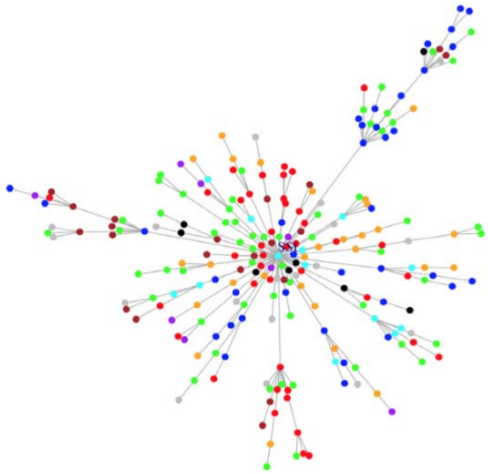
Remarque La dynamique du MST d'un réseau d'actions basé sur la corrélation passe par trois phases:

- phase de MST hiérarchique - un état boursier (relativement) stable
- phase de la superstar MST - un état de marché transitoire
- phase de MST hiérarchique décorée par quelques arbres en forme d'étoiles locales – un état boursier (relativement) stable.

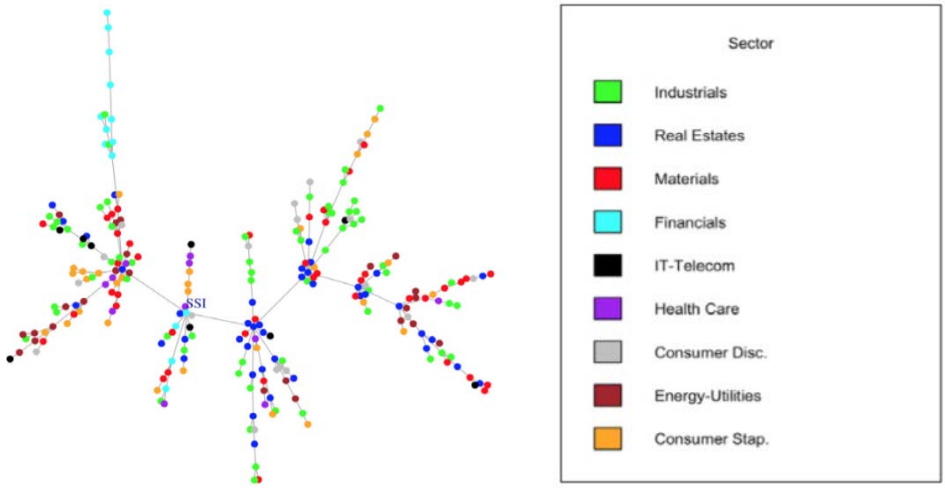
Dans [Nguyen, 2019c], nous avons trouvé le même résultat sur le marché vietnamien (la figure 2.6).



(a) Le MST hiérarchique pour la période du 31/03/2009 au 19/10/2010

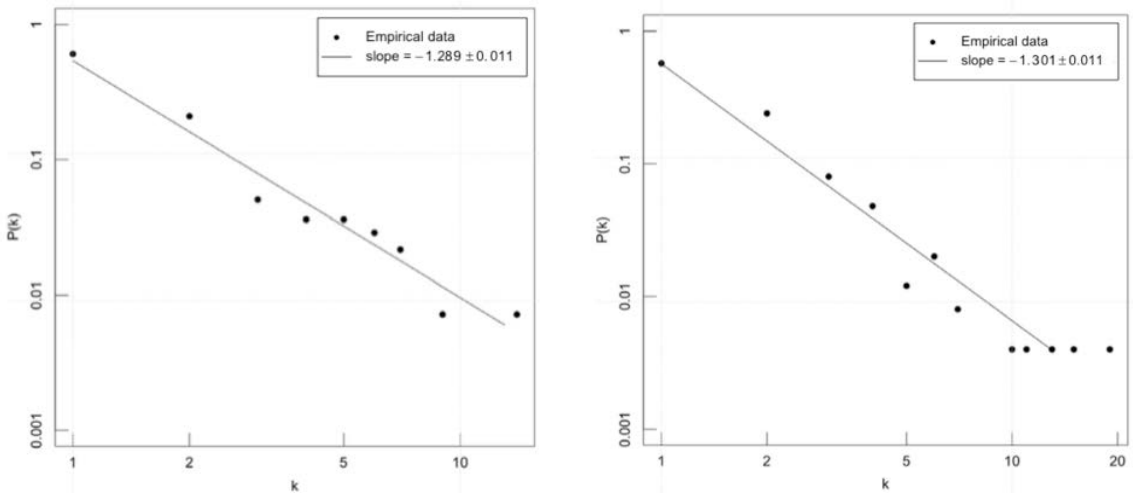


(b) Le MST aux allures de superstar pour la période du 16/05/2012 au 02/12/2013



(c) Le MST hiérarchique décoré de quelques arbres étoilés locaux pour la période du 14/01/2014 au 18/08/2015.

Figure 2.6: Changement structurel du réseau d'actions MST cotées sur le HSX.



(a) La distribution des degrés du MST hiérarchique pour la période du 31/03/2009 au 19/10/2010

(b) La distribution en degrés du MST hiérarchique décorée par quelques arbres étoilés locaux pour la période du 14/01/2014 au 18/08/2015

Figure 2.7: Degré de répartition du réseau hiérarchique MST des actions cotées à la HSX.

Le MST dans les première et troisième phases a la structure commune des réseaux d'action, la structure sans échelle (la figure 2.7). Par contre, dans la deuxième phase, le MST a un super hub, ce qui implique l'absence de la plupart des connexions entre les paires d'autres nœuds et le réseau perd sa propriété scale-free (la figure 2.8).

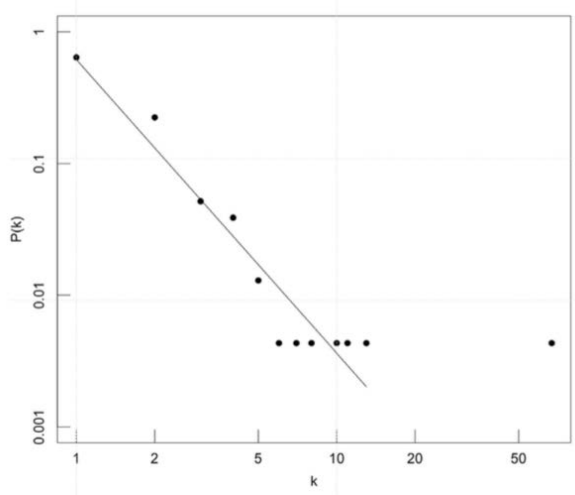


Figure 2.8: Distribution des degrés du MST en étoile sur le HSX pour la période du 16/05/2012 au 02/12/2013 et la droite ajustée d'une loi de puissance après avoir négligé le super hub.

Par conséquent, une structure en étoile du réseau MST est un signe crucial signifiant un événement exceptionnel. Il a été montré dans [Wiliński, 2013], [Sienkiewicz, 2013] et dans notre étude [Nguyen, 2019c] que le MST en forme d'étoile apparaît lorsque l'économie est soumise à un stress important. La raison est donnée lors de l'analyse de la longueur moyenne du plus court chemin du MST.

Pour quantifier le changement dans la structure d'un MST d'une structure en chaîne à une

structure en étoile, nous pouvons utiliser le rapport de survie, le même rapport sectoriel et, plus directement, l'exposant allométrique introduit dans la section 4.2 (la figure 2.9).

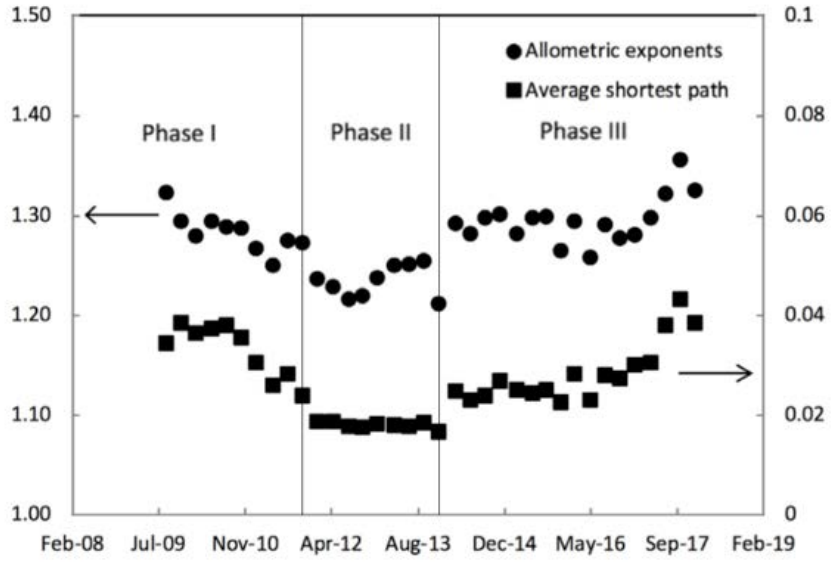


Figure 2.9: Synchronisation entre le déclin de la longueur moyenne normalisée du plus court chemin et le déclin de l'exposant allométrique du réseau MST construit sur le HSX.

De plus, lors de l'analyse de la relation d'échelle allométrique, la valeur C d'un nœud représente l'importance de l'influence du action correspondant sur les autres sur le marché en cas de crise. Nous avons constaté qu'il n'y a qu'un seul action ayant un impact extrêmement élevé C sur le réseau en forme d'étoile dans la phase II (la figure 2.10a), tandis que l'impact total sur l'ensemble du marché est distribué à de nombreux actions dans Phase III (la figure 2.10b). Par conséquent, le réseau en étoile est extrêmement sensible à un choc de son hub.

(a) Superstar-like MST pour la période du 16/05/2012 au 02/12/2013

(b) MST hiérarchique décoré de quelques arbres étoilés locaux pour la période du 14/01/2014 au 18/08/2015

Figure 2.10: MST construits sur le HSX avec $\log(C)$ comme taille de nœud.

En outre, un résultat courant dans de nombreuses études est que le nœud central d'un MST en forme d'étoile est pertinent pour les entreprises fournissant des services financiers [Sienkiewicz, 2013; Onnela, 2002; Onnela, 2003a]. Nous obtenons le même résultat dans notre étude empirique sur le marché vietnamien. Notre nœud central correspond à la Saigon Securities Incorporation, une société de courtage en valeurs mobilières.

Chapitre 3

Propriété spectrale de la matrice de corrélation croisée des actions

Dans ce chapitre, nous utilisons les propriétés spectrales analysées par la théorie des matrices aléatoires (RMT) et l'analyse en composantes principales (PCA) pour comprendre non seulement les propriétés structurelles d'un réseau boursier, mais également l'interaction commune entre les entités du marché.

1 Théorie des matrices aléatoires appliquée aux systèmes d'action

Le RMT est une théorie physique qui permet d'obtenir la matrice de corrélation croisée précise de nombreuses entités. Dans cette étude, nous utilisons le RMT pour comprendre la nature des corrélations des fluctuations des cours boursiers sur un marché boursier.

Considérons un système d'actions N. Selon la définition 2.1, nous pouvons calculer la matrice de corrélation croisée $\mathbf{C} = (c_{ij})$ des actions à partir des rendements boursiers, r_i, $i = \overline{1, N}$. Dans le RMT, pour considérer une hypothèse nulle selon laquelle les rendements boursiers sont strictement non corrélés, nous supposons que la matrice de corrélation croisée $\mathbf{C}$ est équivalente à une matrice purement aléatoire $\mathbf{W}$ obtenue à partir de la norme normalement IID distribué des séries chronologiques. Une telle matrice est dans l'ensemble de la matrice de Wishart [Wishart, 1928].

Définition 3.1. *Une vraie matrice de Wishart est une matrice symétrique aléatoire* $\mathbf{W}$ *de la forme:*

$$\mathbf{W} = \frac{1}{T}\mathbf{MM'} \tag{3.1}$$

où ' désigne la transposition matricielle, et $\mathbf{M}$ est un matrice de taille $N \times T$ telle que:

- $(M_{ij})_{1 \leq j \leq T}$ sont des échantillons indépendants d'une variable aléatoire à valeur réelle m_i.
- $(m_1, \ldots, m_N)$ est un vecteur gaussien avec une matrice de covariance donnée $\mathbf{K}$.

T est appelé degré de liberté.

De plus, dans notre contexte financier, $(m_1, \ldots, m_N)$ est un vecteur normal standard de covariance $\mathbf{K} = \text{diag}(1, \ldots, 1)$ pour interpréter le modèle nul de $\mathbf{C}$. La différence de la distribution spectrale de $\mathbf{C}$ de celle de $\mathbf{W}$ indique la présence d'informations significatives sur la véritable corrélation des actifs. Surtout dans un grand système, pour $\mathbf{K} = \text{diag}(\sigma, \ldots, \sigma)$, nous avons [Marčenko, 1967]:

Théorème 3.1. *Si $N \to \infty, T \to \infty$ de telle sorte que $\frac{T}{N}$ se rapproche d'un nombre fixe $\alpha \geq 1$, le spectre empirique distribution de la matrice de Wishart $\mathbf{W}$ converge faiblement, en probabilité, vers la distribution Marčenko - Pastur avec la densité ρ supportée sur $[\lambda_-; \lambda_+]$ et donné par*

$$\rho(\lambda) = \frac{\alpha}{2\pi\lambda\sigma^2}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}, \qquad \forall\lambda \in [\lambda_-; \lambda_+] \tag{3.2}$$

où $\lambda_\pm = \sigma^2\left(1 \pm \alpha^{-\frac{1}{2}}\right)^2$.

En réalité, de nombreux travaux ont montré que la plupart des valeurs propres des matrices de corrélation croisée des variations des prix des actifs sur les marchés financiers mondiaux concordent étonnamment bien avec l'intervalle fournie dans le théorème 3.1, mais la plus grande valeur propre est nettement supérieure à λ_+ que le marché soit développé [Laloux, 1999; Plerou, 1999; Rosenow, 2008], en développement [Nobi, 2013] ou émergent [Nguyen, 2019b] (Figure 3.1). Selon le RMT, la plus grande valeur propre permet de refléter les véritables corrélations des composants du système. En particulier, la prédominance de la plus grande valeur propre devient plus importante lors des "crashes" boursiers [Drożdz, 2000; Nobi, 2013; Zheng, 2012].

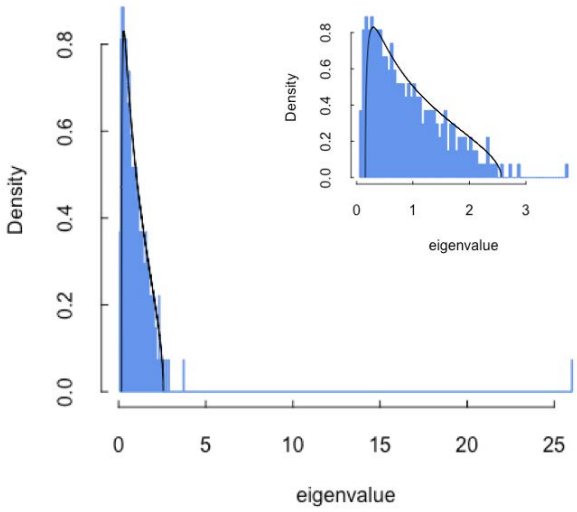


Figure 3.1: Explication de la distribution spectrale prédite par le RMT (trait noir) pour une grande partie de la distribution spectrale de la matrice de corrélation croisée des valeurs cotées sur le HSX du 01/01/2017 au 01/01/2020 (insérer: ces deux distributions lors du zoom sur les valeurs propres sans la plus grande).

2 Composantes principales des rendements boursiers

L'idée principale de l'PCA est de réduire la dimensionnalité d'un ensemble de données composé de nombreuses variables corrélées en transformant l'ensemble de données en un ensemble de variables aléatoires non corrélées, appelées composantes principales (PC), tout en conservant la plupart des variations dans l'ensemble de données initial [Jolliffe, 1986]. Dans notre contexte financier, pour un système d'actions contenant des actions de N, soit $\mathbf{r} = (r_i)_{i=\overline{1,N}}$ et $\mathbf{r}^* = (r_i^*)_{i=\overline{1,N}}$, où $r_i^* = \frac{r_i}{\sigma_i}$ est le rendement standardisé de l'action i. Soit $\lambda_1 > \lambda_2 > \ldots > \lambda_N$ N valeurs propres distinctes de $\mathbf{C}$.

Définition 3.2. *Laisser* $\mathbf{u}_i$ $(i = \overline{1,N})$ *soit le vecteur propre associé à la* i*-ième valeur propre* λ_i *tel que* $\|\mathbf{u}_i\| = \mathbf{u}_i\,\mathbf{'u}_i = 1$, $\mathbf{A}$ *est la matrice* $N \times N$ *dont les colonnes sont* $\mathbf{u}_i s$*, et* $\mathbf{z} = \mathbf{A}\,\mathbf{'r}^*$*. Soit* z_i *le iième composant de* $\mathbf{z}$*. Ensuite,* z_i *est appelé le* i*-ième PC de* $\mathbf{r}$*.*

Dans les énoncés suivants, lorsque nous mentionnons un certain vecteur propre associé à une valeur propre donnée, ce vecteur est l'unité.

Théorème 3.2. *Les PC du vecteur aléatoire* $\mathbf{r}$ *satisfont les propriétés suivantes:*

(i) Les PC ne sont pas corrélés.

(ii) La variance de chaque PC est égale à la valeur propre correspondante, c'est à dire.,

$$\mathrm{Var}\,(z_i) = \lambda_i, \qquad \forall i = \overline{1,N} \tag{3.3}$$

(iii) Pour toute combinaison linéaire $\mathbf{v'r}^*$ *de variables dans* $\mathbf{r}^*$*, où* $\mathbf{v}$ *est un vecteur unitaire, la variance du premier PC est la plus grande, c'est-à-dire*

$$\max_{\|\mathbf{v}\|=1} \mathrm{Var}\,(\mathbf{v'r}^*) = \mathrm{Var}\,(z_1) \tag{3.4}$$

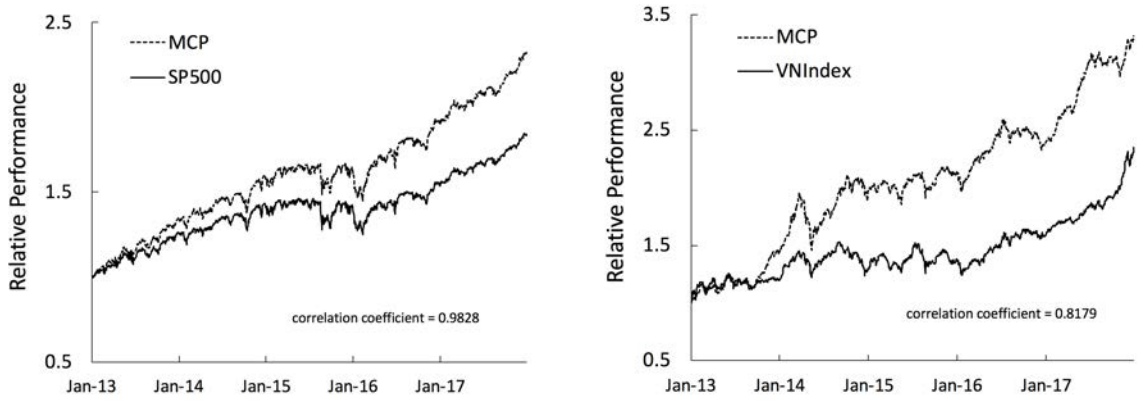
(iv) Pour toute combinaison linéaire non nulle $\mathbf{v'r}^*$ *de variables dans* $\mathbf{r}^*$*, le premier PC est celui qui maximise la somme des carrés des coefficients de corrélation de Pearson avec chacun des rendements boursiers, c'est-à-dire*

$$\max_{y=\mathbf{v'r}^*, y\neq 0} \sum_{i=1}^{N} (\rho_{i,y})^2 = \sum_{i=1}^{N} (\rho_{i,z_1})^2 \tag{3.5}$$

où $\rho_{i,y}$ *et* ρ_{i,z_1} *sont le coefficient de corrélation entre* y *et* r_i *et le coefficient de corrélation entre* z_1 *et* r_i*,* $i = \overline{1,N}$*, respectivement.*

Si chaque vecteur non nul $\mathbf{v}$ représente un portefeuille d'investissement où la i-ième composante $v^{(i)}$ de $\mathbf{v}$ est le capital investi pour action i, $\mathbf{v'r}$ est le rendement du portefeuille. Ainsi, le j-ième PC z_j est le rendement d'un portefeuille dont le chargement de l'action i est la fraction $\mathbf{u}_j^{(i)}/\sigma_{je}$. Ce portefeuille est appelé le j-th eigen-portfolio. Le théorème 3.2 propose que le premier PC soit équivalent au facteur de marché dans le modèle CAPM bien connu [Plerou, 2002].

La figure 3.2 montre la synchronisation du premier portefeuille propre et de l'indice de marché correspondant.



(a) Indice S&P 500 vs le portefeuille le plus corrélé construit sur les composants de l'indice

(b) Indice VN vs le portefeuille le plus corrélé construit sur les composantes de l'indice

Figure 3.2: La performance relative du portefeuille simulé le plus corrélé (ligne pointillée) par rapport à l'indice de marché correspondant de 2013 à la fin de 2017.

3 Charges de la première composante principale des rendements boursiers

Définition 3.3. *Pour tout nombre* $i = \overline{1, N}$*, les composants* $u_i^{(j)} (j = \overline{1, N})$ *du vecteur propre* $\mathbf{u}_i$ *sont appelés les chargements du* i*- e PC.*

Dans cette section, les premiers chargements du PC sont de notre intérêt car les chargements aident à comprendre comment les actions individuelles contribuent au facteur de marché. De nombreuses études ont montré que les chargements ont généralement le même signe [Gopikrishnan, 2001; Nguyen, 2013; Pan, 2007; Plerou, 2002] alors que ce n'est le cas pour aucun des autres PC. Cela confirme qu'il existe un facteur systématique dominant impactant totalement la totalité ou la plupart des actions du marché. Par ailleurs, les chargements sont décorrélés de la capitalisation boursière de l'entreprise correspondante (la figure 3.3). Par conséquent, l'utilisation de la capitalisation boursière pour peser une action n'est pas la meilleure façon d'avoir un portefeuille capturant le comportement commun d'un marché boursier.

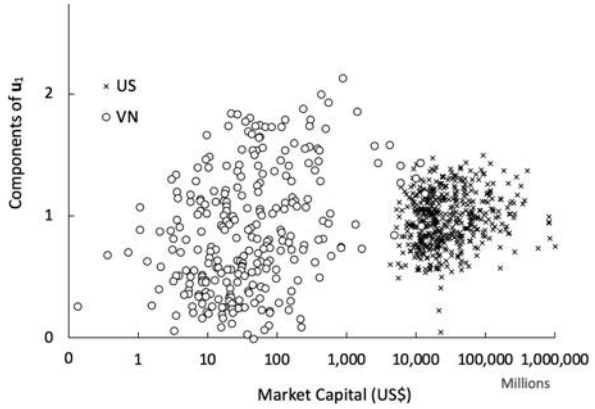


Figure 3.3: Les composantes de $\mathbf{u}_1$ (relevé par $\sqrt{\lambda_1}$) par rapport à la capitalisation boursière des actions correspondantes dans l'indice S&P 500 et l'indice VN en la période de 2013 à fin 2017.

4 Influence des actions reflétée par la première composante principale des rendements des actions

Théorème 3.3. *Si la plus grande valeur propre est extrêmement supérieure aux autres valeurs propres, le chargement du premier PC sur un composant est presque linéairement lié à la moyenne des coefficients de corrélation entre l'action correspondant au composant et les autres actions.*

Concrètement, pour tout action i ($i = \overline{1, N}$), soit $\overline{c}_i$ la moyenne des coefficients de corrélation entre l'action et les autres, ensuite nous avons:

$$\overline{c}_i \approx \lambda_1 \overline{u}_1 u_1^{(i)} \tag{3.6}$$

Nous montrons empiriquement que le théorème 3.3 est valide dans notre contexte financier (la figure 3.4).

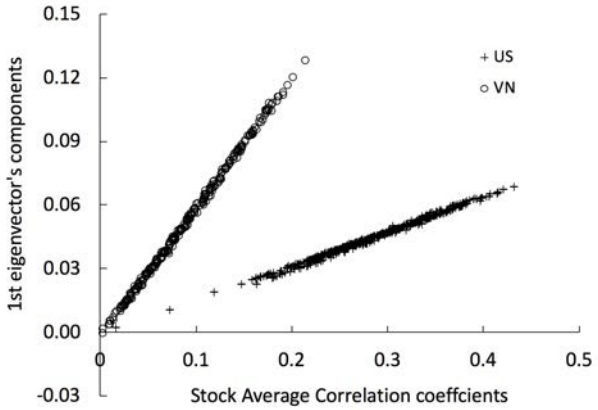
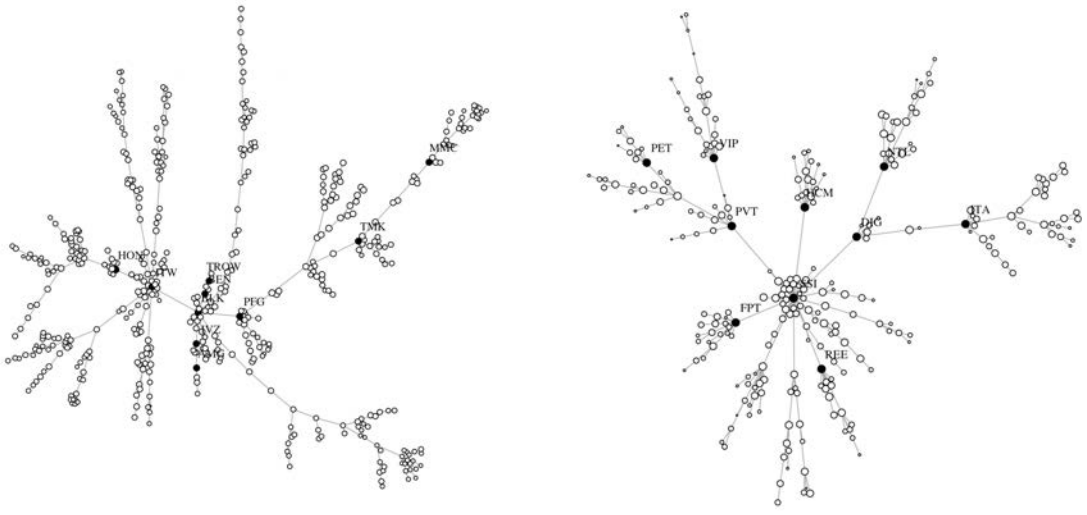


Figure 3.4: Relation entre les premières composantes du vecteur propre et le coefficient de corrélation moyen des actions correspondantes pour l'indice S&P 500 et l'indice VN sur la période de 2013 à fin 2017.

Surtout, la relation entre le chargement du premier PC sur une action et le rôle du action dans le réseau MST montre que le PC contient des informations significatives sur la structure MST (la figure 3.5). Cela signifie que l'PCA peut être une méthode utile pour analyser le réseau MST.



(a) **C** est calculé à partir des composants boursiers de l'indice S&P 500

(b) **C** est calculé à partir des composants boursiers de l'indice VN

Figure 3.5: Le MST obtenu dans la période de 2013 à fin 2017 avec la taille du nœud comme logarithme du chargement du premier PC sur l'action correspondant (les tickers des actions correspondant aux 10 premiers chargements sont affichés et leurs nœuds correspondants sont remplis).

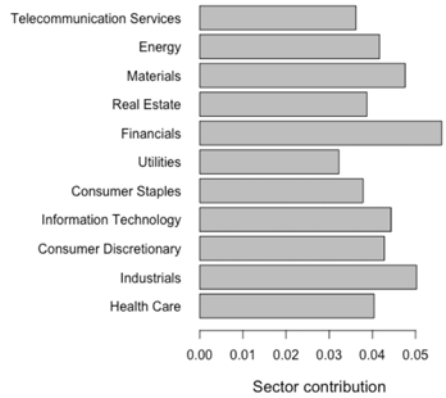
De plus, nous avons construit la mesure a_s pour quantifier la contribution d'un secteur

d'activité s dans le premier vecteur propre:

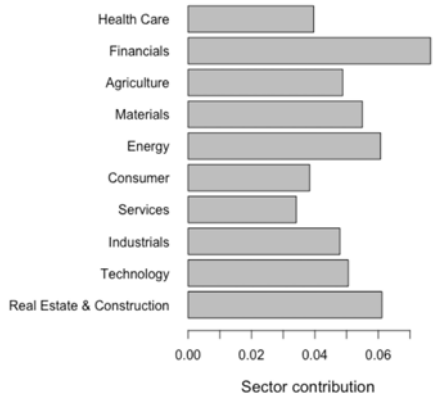
$$a_s = \frac{1}{n_s} \sum_{j=1}^{N} u_1^{(j)} \delta(j, s) \tag{3.7}$$

où $\delta(j, s) = 1$ si l'action j appartient au secteur s et $\delta(j, s) = 0$ sinon; n_s est le nombre d'actions appartenant au secteur s.

Nous avons constaté que les sociétés financières ont tendance à avoir des charges importantes dans le facteur de marché (la figure 3.6). Cela aide à expliquer pourquoi une entreprise de services financiers peut devenir un super-hub lorsque le réseau MST approche de son état instable – une structure en étoile – comme discuté dans le chapitre 2



(a) Le premier PC est construit à partir des composants boursiers de l'indice S&P 500



(b) Le premier PC est construit à partir des composants boursiers de l'indice VN

Figure 3.6: Contributions sectorielles pour le premier PC des rendements boursiers obtenus dans la période de 2013 à fin 2017.

Chapitre 4

Défaillances en cascade dans les systèmes financiers et son modèle prétopologique

1 Défaillances en cascade dans les systèmes complexes

Parfois, un système complexe peut être endommagé par la défaillance d'un seul ou de quelques composants, car les composants les plus liés aux composants défaillants sont d'abord infectés et continuent de déclencher la défaillance d'autres, etc. Ce processus de diffusion à travers la relation entre les composants d'un système est appelé défaillance en cascade. De même, la défaillance en cascade apparaît également dans les systèmes financiers lorsque la défaillance d'une institution financière peut entraîner la défaillance de ses homologues et se propager uniformément sur le marché. Pour les systèmes d'action, nous proposons d'utiliser la théorie de la prétopologie pour modéliser la défaillance en cascade en raison des caractéristiques du système et de la faiblesse de la modélisation du réseau.

2 Théorie de la prétopologie

La théorie de la prétopologie a été développée dans le but de suivre l'évolution d'un processus de diffusion et comment il contribue au résultat final [Belmandt, 2011]. Soit E un ensemble non vide et $\mathcal{P}(E)$ soit l'ensemble de tous ses sous-ensembles.

Définition 4.1. *On appelle pseudofermeture définie sur E toute carte a de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ tel que:*

(i) $a(\varnothing) = \varnothing$, *et*

(ii) $\forall A \subset E, A \subset a(A)$.

Définition 4.2. *Nous appelons l'intérieur défini sur E toute application i de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ telle que:*

(i) $i(E) = E$, *et*

(ii) $\forall A \subset E, i(A) \subset A$.

Dans cette étude, nous considérons i comme une c-dualité de a.

Définition 4.3. *Un espace prétopologique est un couple $(E, a(.))$ où a est une pseudo-fermeture définie sur l'ensemble non vide E.*

Les calculs successifs d'une pseudofermeture à un ensemble A donné permettent de modéliser l'évolution d'un processus de dilatation à partir de A. Pendant ce temps, le résultat de l'application successive de l'intérieur à A peut modéliser un processus de dilution à partir de A.

Définition 4.4. *Étant donné un espace prétopologique $(E, a(.))$, pour tout sous-ensemble A de E,*

(i) A est dit être un sous-ensemble fermé de E si et seulement si $A = a(A)$.

(ii) A est un sous-ensemble ouvert de E si et seulement si $A = i(A)$.

Définition 4.5. *Étant donné un espace prétopologique $(E, a(.))$, pour tout sous-ensemble A de E,*

(i) nous appelons fermeture de A, notée $\mathbf{F}(A)$, le plus petit sous-ensemble fermé de E qui contient A si le sous-ensemble existe.

(ii) nous appelons ouverture de A, notée $\mathbf{O}(A)$, le plus grand sous-ensemble ouvert de E qui est inclus dans A si le sous-ensemble existe.

Un espace prétopologique est une extension d'un hypergraphe [Dalud-Vincent, 2011]. Pour étudier un système incluant la multi-relation d'éléments et les relations entre un groupe d'éléments et un élément, ou lorsque l'évolution du processus de dilatation et/ou de diminution est l'objectif de la recherche, la théorie de la prétopologie peut être une solution intéressante.

3 Cadre prétopologique des défaillances en cascade des marchés boursiers

Similaire à [Auray, 1979; Ben-Amor, 2010; Lamure, 2009] pour résoudre le problème de propagation de la pollution, nous construisons un espace prétopologique des actions en considérant $a(A)$ comme une composition de A et tous les autres éléments dont la relation avec A est supérieur à un seuil, pour tout ensemble d'actions A. Cependant, nous n'utilisons pas de seuil constant. Les trois hypothèses suivantes sont utilisées dans notre modèle:

(i) Si une action subit un choc de prix, les actions qui ont des corrélations élevées avec elle sont directement influencées.

(ii) L'impact d'un groupe sur les autres actions est plus important si le groupe est plus grand.

(iii) Un changement dans la taille d'un groupe fait que le groupe a plus d'impact sur une action extérieur si la taille du groupe est plus grande.

Par conséquent, si E est l'ensemble de toutes les actions cotées sur un marché boursier, et N est le nombre de ces actions, nous utilisons la pseudo-fermeture suivante pour notre problème d'échec en cascade:

Proposition 4.1. *Soit f une fonction décroissante et concave de $[1, N]$ dans $[0, 1)$. Soit a une application de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ telle que $a(\varnothing) = \varnothing$ et*

$$a(A) = A \cup \left\{ k \in E \backslash A \middle| \frac{1}{\|A\|} \sum_{j \in A} c_{jk} \geq f(\|A\|) \right\}, \ \forall A \in \mathcal{P}(E) \backslash \{\varnothing\} \tag{4.1}$$

où $\|A\|$ est la taille de A. Alors, $(E, a(.))$ est un espace prétopologique d'intérieur i tel que $i(E) = E$ et

$$i(A) = \left\{ k \in A \middle| \frac{1}{N - \|A\|} \sum_{j \in E \backslash A} c_{jk} < f(N - \|A\|) \right\}, \ \forall A \in \mathcal{P}(E) \backslash \{E\} \tag{4.2}$$

Avec ce modèle, en trouvant la fermeture d'un groupe d'actions, nous pouvons prédire l'ampleur de l'impact des fluctuations de prix de ces actions sur les autres lorsque ces actions sont dans une tendance de prix négative. A l'inverse, l'ouverture de la rémunération du groupe peut permettre de prédire les valeurs non impactées par l'évolution négative des cours du groupe.

4 Résultats empiriques sur le NYSE

Dans cette section, nous étudions empiriquement comment la pseudo-fermeture introduite dans la proposition 4.1 peut modéliser la propagation d'un choc de prix dans un marché boursier réel, le NYSE, et vice versa, comment l'intérieur correspondant peut aider à prédire les actions non touché par le choc.

4.1 Base de données

Nous examinons l'échec en cascade à partir de MER, l'action ordinaire de Merrill Lynch & Co. sur le NYSE, à d'autres composés dans l'indice S&P 500. Cet action est sélectionnée en raison de sa position importante sur le marché américain pendant des décennies, mais elle a perdu la position à partir du deuxième de 2007 et a finalement été acquise le 31/12/2008. Nous désignons MER comme action i_0. Le jour où i_0 a connu une baisse considérable de son prix est défini comme le moment où le prix a chuté de plus de 70% dans l'année précédant son jour de consolidation, noté t_0. On dit que l'action échoue au temps t_0.

Considérons l'ensemble E de toutes les composantes de l'indice S&P 500 pour représenter le marché boursier américain. Nous utilisons les cours de clôture quotidiens des actions de E dans 2 ans avant t_0 pour calculer les corrélations des actions. On note H comme l'ensemble des composants défaillants du système complexe E pendant une période de 6 mois après t_0.

4.2 Méthode de recherche

Dans [Nguyen, 2019a], pour construire l'espace prétopologique $(E, a(.))$ selon la proposition

4.1, les fonctions ci-dessous sont utilisées pour définir le seuil affecté:

$$f(x) = \theta \left(\frac{N}{N-1}\right)^{\gamma} \left(\frac{1}{x-N-1}+1\right)^{\gamma}, \quad \forall x \in [1, N] \tag{4.3}$$

où $\gamma > 0$ et $0 < \theta < 1$.

Pendant ce temps, dans [Nguyen, 2021b], nous utilisons la fonction ci-dessous:

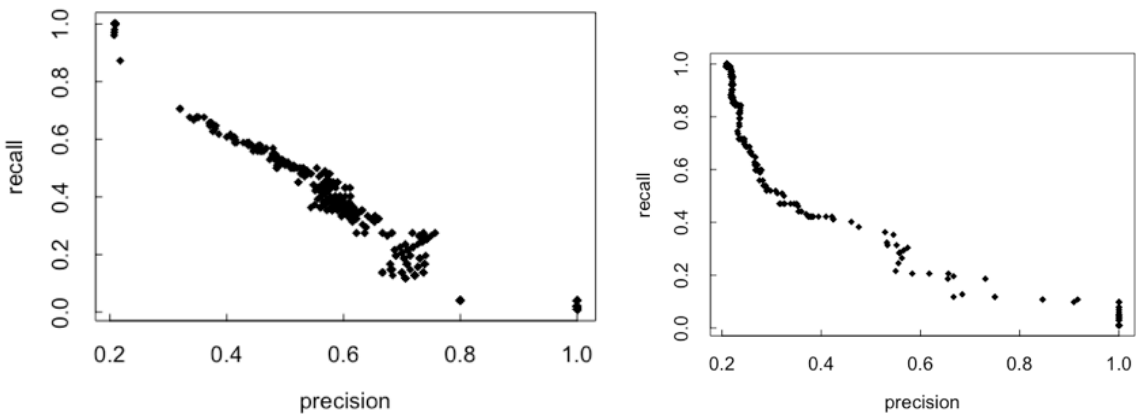
$$f(x) = 1 - \theta e^{\gamma x}, \quad \forall x \in [1, N] \tag{4.4}$$

où $0 < \theta < 1$ et $0 < \gamma < -N^{-1} \ln \theta$.

Nous utilisons $\mathbf{F}(\{i_0\})$ pour prédire les actions influencées par l'échec de i_0 tandis que $\mathbf{O}(E \backslash \{i_0\})$ est utilisé pour prédire les actions non affectés par la défaillance de i_0. Nous quantifions l'efficacité de la prédiction par deux mesures: la précision et le rappel de la prédiction.

4.3 Processus de transmission d'une chute du prix d'action

Nous avons constaté que notre cadre prétopologique avec des paramètres appropriés peut mieux modéliser la transmission de la chute à partir de i_0 que le réseau MST. En effet, la précision de la prédiction basée sur la connexion du MST dans la figure 4.1a est généralement inférieure à la précision de la prédiction basée sur $\mathbf{F}(\{i_0\})$ dans la figure 4.1b au même niveau de rappel. De plus, la contagion des échecs modélisée par les calculs successifs de pseudofermeture pour obtenir $\mathbf{F}(\{i_0\})$ à partir de i_0 peut atteindre des actions éloignés dans le MST (la figure 4.2).



(a) (b)

Figure 4.1: Relation entre la précision et le rappel de la prédiction des actions influencées par le choc de prix de i_0 lorsque (a) en utilisant $\mathbf{F}(\{i_0\})$, et (b) en utilisant le réseau MST.

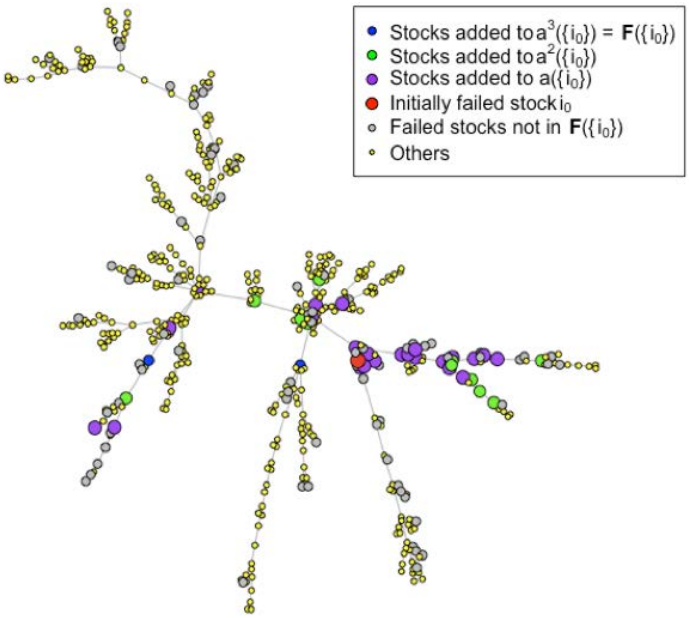


Figure 4.2: Actions distants dans le réseau MST atteints dans le processus de dilatation de $\{i_0\}$ à $\mathbf{F}(\{i_0\})$ avec $\theta = 0.34$ et $\gamma = 5 \times 10^{-4}$.

Une question est de savoir quelles valeurs de θ et γ sont appropriées. Nous proposons de choisir des valeurs appropriées pour eux tels que le seuil d'impact de l'expansion de $\{i_0\}$ à $a(\{i_0\})$ dans la première étape du processus de dilatation n'est ni trop grand ni trop petit si la précision de la prédiction est plus importante que le rappel.

Inversement, dans [Nguyen, 2021b], on utilise $\mathbf{O}(E\backslash\{i_0\})$ dans l'espace prétopologique donné dans la proposition 4.1, où f est vérifié par (4.4), pour prédire les actions non affectées par le choc de prix de i_0 (la figure 4.3). Nous avons trouvé que θ et γ devraient être choisis de telle sorte que le seuil d'impact ne soit ni trop grand ni trop petit.

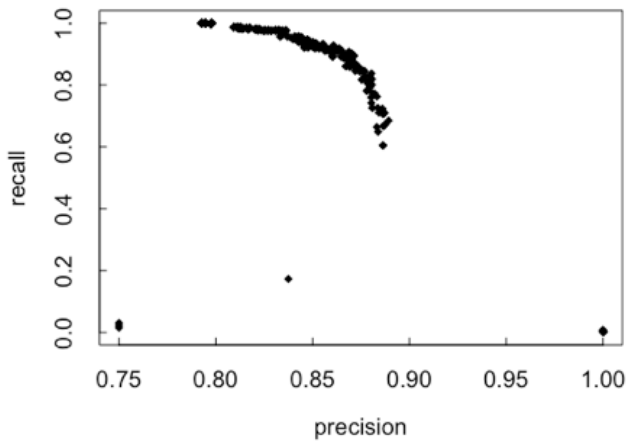


Figure 4.3: Relation entre la précision et le rappel des actions prédites non influencées par le choc de prix de i_0 par $\mathbf{O}(E\backslash\{i_0\})$.

5 Types d'espaces Prétopologiques

Dans cette section, nous introduisons quelques types particuliers d'espaces prétopologiques et leurs relations.

Définition 4.6. *Soit un espace prétopologique $(E, a(.))$, on dit que c'est:*

(i) un espace de type $\mathcal{V}$ si $(A \subset B \Rightarrow a(A) \subset a(B))$ pour tout sous-ensemble A, B de E.

(ii) un espace de type $\mathcal{V}_D$ si $a(A \cup B) = a(A) \cup a(B)$ pour tout sous-ensemble A, B de E.

(iii) un espace de type $\mathcal{V}_S$ si $a(A) = \bigcup_{x \in A} a(\{x\})$ pour tout sous-ensemble A de E.

Proposition 4.2. *Tout espace de type $\mathcal{V}_D$ est un espace de type $\mathcal{V}$.*

Proposition 4.3. *Tout espace de type $\mathcal{V}_S$ est un espace de type $\mathcal{V}_D$.*

Proposition 4.4. *Un espace prétopologique $(E, a(.))$ est un espace topologique si et seulement s'il est de type $\mathcal{V}_D$ et $a(a(A)) = a(A)$ pour tout sous-ensemble A de E.*

Dans un espace de type $\mathcal{V}$, nous disposons de nombreux outils nécessaires pour construire un concept de proximité, tels que les bases des voisinages d'un élément, la connexité et les sous-ensembles fermés minimaux [Belmandt, 2011]. Par conséquent, nous proposons d'améliorer l'espace prétopologique de la section 4.3 pour qu'il devienne un espace de type $\mathcal{V}$, par exemple:

Proposition 4.5. *Soit f une fonction décroissante et concave de $[1, N]$ dans $[0, 1)$. Soit a une application de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ telle que $a(\varnothing) =$ rien et*

$$a(A) = A \cup \left\{ k \in E \backslash A \middle| \max_{j \in A} c_{jk} \geq f(\|A\|) \right\}, \qquad \forall A \in \mathcal{P}(E) \backslash \{\varnothing\} \tag{4.5}$$

où $\|A\|$ est la taille de A. Alors, $(E, a(.))$ est un espace prétopologique de type $\mathcal{V}$.

Chapitre 5

Anomalies topologiques de la dynamique des indices de marché

1 Les indices boursiers comme représentations des comportements collectifs des marchés boursiers

En tant que système complexe avec de nombreux composants et des relations compliquées, le mouvement d'un marché boursier dans son ensemble n'est pas facile à prévoir. Pour avoir une vue d'ensemble intuitive d'un marché boursier, les gens dépendent souvent des indices boursiers. Selon l'étude du chapitre 3, nous proposons que la fluctuation d'un indice de marché puisse être utilisée pour évaluer le comportement collectif du marché si elle est fortement corrélée avec la fluctuation du premier PC. L'occurrence des indices a rendu les marchés boursiers différents de la plupart des systèmes complexes car nous pouvons facilement saisir les comportements collectifs des marchés par une mesure transparente, mise à jour en permanence et fournie gratuitement. Par conséquent, dans ce chapitre, notre objectif est de détecter les comportement anormaux d'un marché boursier à travers des anomalies dans la dynamique du rendement de son indice représentatif.

2 Incorporation à retardement d'une série temporelle

Pour découvrir des anomalies dans la dynamique d'un retour d'indice, il faut appréhender ses différents états. Nous utilisons la méthode d'intégration à retardement [Packard, 1980; Ruelle, 1979] pour résoudre ce problème.

2.1 Retarder la reconstruction

Définition 5.1. *Un vecteur reconstruit obtenu à partir d'une série temporelle* (x_t) *est défini par, pour tout* t*,*

$$\mathbf{y}_t^{\tau,d} = \left(x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(d-1)\tau}\right) \tag{5.1}$$

Nous appelons τ *le délai et* d *la dimension de plongement.*

L'objectif principal de la méthode d'intégration à retardement est de convertir une série

chronologique en un nuage de points d'un espace de dimension supérieure afin qu'il puisse capturer différents états de la dynamique de la série chronologique. Alors, $\mathbf{Y} = (\mathbf{y}_t^{\tau,d})$ est appelé l'espace des phases/états et nous pouvons obtenir l'attracteur de la série temporelle à partir de la topologie de l'espace fonctionnalité.

2.2 Sélection du délai

Nous introduisons les méthodes courantes pour choisir le paramètre de délai [Abarbanel, 1993]:

- *Autocorrélation*: τ est sélectionné comme premier zéro de la fonction d'autocorrélation linéaire:

$$A(\tau) = \frac{\langle (x_{t+\tau} - \overline{x})(x_t - \overline{x}) \rangle_t}{\left\langle (x_t - \overline{x})^2 \right\rangle_t} \tag{5.2}$$

  où $\langle . \rangle_t$ est la moyenne dans le temps et $\overline{x} = \langle x_t \rangle_t$.

- *Information mutuelle moyenne*: τ est sélectionné comme le décalage dans le temps que le premier minimum de l'information mutuelle moyenne (AMI) de x_t et $x_{t+\ tau}$, où

$$\mathrm{AMI}(\tau) = \sum_t \hat{p}(x_t, x_{t+\tau}) \log_2 \frac{\hat{p}(x_t, x_{t+\tau})}{\hat{p}_1(x_t)\, \hat{p}_2(x_{t+\tau})} \tag{5.3}$$

  $\hat{p}$ est la distribution de probabilité conjointe estimée de x_t et $x_{t+\tau}$; $\hat{p}_1$ et $\hat{p}_2$ sont la fonction marginale estimée de la distribution de probabilité conjointe.

2.3 Sélection de la dimension d'intégration

Nous introduisons deux approches pour choisir une dimension de plongement appropriée: les tests dynamiques et les tests géométriques. Les tests dynamiques sont effectués en augmentant la dimension de plongement jusqu'à ce que le comportement typique de la série temporelle apparaisse [Broomhead, 1986; Eckmann, 1986] tandis que les tests géométriques dépendent de la distance entre les points de l'espace d'état. Nous pensons que les tests géométriques sont plus adaptés aux séries temporelles du monde réel car de tels tests se rapprochent directement de l'objectif de la reconstruction. Quelques exemples courants de tests géométriques:

- *Saturation des invariants du système*: la méthode recherche la dimension de plongement qui fournit l'indépendance avec une certaine fonction en fonction des distances entre les points de l'espace d'états [Grassberger, 1983].

- *Faux voisin le plus proche*: cette méthode recherche d comme le plus petit nombre tel que pour tout point de l'espace reconstruit de dimension d, son point le plus proche soit encore assez proche dans le $(d+1)$– espace reconstruit dimensionnel [Kennel, 1992]. De plus, dans notre étude, nous utilisons la version modifiée proposée dans [Cao, 1997]. En particulier, pour chaque vecteur reconstruit $\mathbf{y}_t^{\tau,d}$, soit le vecteur reconstruit $\mathbf{y}_{t^*}^{tau,d}$ être le plus proche voisin de $\mathbf{y}_t^{\tau,d}$ avec le plus proche dans le sens d'une certaine distance [1].

[1] If $\mathbf{y}_{t^*}^{\tau,d} \equiv \mathbf{y}_t^{\tau,d}$, on prend $\mathbf{y}_{t^*}^{\tau,d}$ comme deuxième voisin le plus proche de $\mathbf{y}_t^{\tau,d}$.

Soit

$$a\left(t,d\right)=\frac{\left\|\mathbf{y}_{t}^{\tau,d+1}-\mathbf{y}_{t^{*}}^{\tau,d+1}\right\|}{\left\|\mathbf{y}_{t}^{\tau,d}-\mathbf{y}_{t^{*}}^{\tau,d}\right\|}<R_{\tau},\qquad\forall t \tag{5.4}$$

et

$$E_{1}\left(d\right)=\frac{\left\langle a\left(t,d+1\right)\right\rangle_{t}}{\left\langle a\left(t,d\right)\right\rangle_{t}} \tag{5.5}$$

où $\|.\|$ est quelques mesures de distance. Sans perte de généralité, nous utilisons la distance euclidienne. Lorsque le changement de E_1 devient trivial lorsque la dimension d'intégration est supérieure à un certain nombre d_0, nous prenons d_0+1 comme dimension d'intégration optimale.

3 Homologie persistante

Pour étudier les informations topologiques de l'espace d'état $\mathbf{Y}=(\mathbf{y}_t^{\tau,d})$, nous proposons d'utiliser l'homologie persistante, la principale technique d'analyse topologique des données(TDA) [Chazal, 2021].

3.1 Complexes simpliciels

Soit $\mathbf{V}=\{v_0,v_1,\ldots,v_k\}\subset\mathbb{R}^d$ un ensemble de points affinement indépendants.

Définition 5.2. *Un simplexe k-dimensionnel σ englobé par $\mathbf{V}$ est l'enveloppe convexe de $\mathbf{V}$, c'est-à-dire,*

$$\sigma=\left\{\sum_{i=0}^{k}\alpha_i v_i\middle|\sum_{i=0}^{k}\alpha_i=1\ \wedge\ 0\leq\alpha_i\leq 1\right\} \tag{5.6}$$

$v_0,v_1,\ldots,v_k$ sont appelés sommets de σ. L'enveloppe convexe de tout sous-ensemble de $\mathbf{V}$ est aussi un simplexe appelé face de σ.

Définition 5.3. *Un complexe simplicial $\mathbf{G}$ est une collection finie de simplexes, telle que:*

(i) Toute face d'un simplexe de $\mathbf{G}$ est un simplex de $\mathbf{G}$.

(ii) L'intersection de deux simples de $\mathbf{G}$ est soit vide, soit une face commune des deux.

Nous introduisons deux complexes simpliciaux familiers construits à partir d'un nuage de points donné: le complexe de Vietoris-Rips et le complexe Čech.

Définition 5.4. *Étant donné un nombre α, le complexe Čech* $\check{\mathrm{C}}\mathrm{ech}_\alpha(\mathbf{V})$ *est l'ensemble des simples par des sous-ensembles de $\mathbf{V}$ tels que: pour tout simplex $\sigma\in\check{\mathrm{C}}\mathrm{ech}_\alpha(\mathbf{V})$, les boules fermées $B(v_i,\alpha)$ pour tout sommet v_i de σ ont un intersection.*

Définition 5.5. *Étant donné un nombre α, le complexe de Vietoris-Rips (appelé aussi complexe de Vietoris ou complexe de Rips)* $\mathrm{Rips}_\alpha(\mathbf{V})$ *est l'ensemble des simplexes englobés par des sous-ensembles de $\mathbf{V}$ tels que: pour tout simplexe $\sigma\in\mathrm{Rips}_\alpha(\mathbf{V})$, $\|v_i-v_j\|\leq\alpha$ pour tout sommet v_i, v_j de σ.*

3.2 Groupes d'homologie

Pour découvrir les informations topologiques d'un complexe simplicial, l'homologie est une approche puissante qui permet de distinguer les structures du complexe en détectant ses trous.
Définition 5.6. *La carte des limites k* $\partial_k : C_k \to C_{k-1}$ $(k > 0)$ *est défini par:*

(i) pour tout simplex orienté $\sigma = [v_0, v_1, \ldots, v_k]$,

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, v_{i-1}, v_{i+1}, ..v_k] \tag{5.7}$$

et

(ii) pour tout simplex de dimension k $\sigma_1, \ldots, \sigma_p$, *et coefficients* $\alpha_1, \ldots \alpha_p \in \mathbb{Z}$,

$$\partial_k \left(\sum_{i=1}^{p} \alpha_i \sigma_i \right) = \sum_{i=1}^{p} \alpha_i \partial_k(\sigma_i) \tag{5.8}$$

où C_k *est l'ensemble des k-chaînes avec des coefficients dans* $\mathbb{Z}$

Définition 5.7. *Les éléments de* $\ker(\partial_k)$ *sont appelés k-cycles.*

Définition 5.8. *Un trou de dimension k est un cycle de k qui n'est pas une frontière d'un complexe simplicial de dimension* $(k+1)$.

Les trous de dimension k peuvent être détectés par des groupes d'homologie définis par:

Définition 5.9. *Étant donné un complexe simplicial* $\mathbf{G}$, *le groupe d'homologie k-dimensionnel de* $\mathbf{G}$ *est*

$$H_k(\mathbf{G}) = \ker(\partial_k) / Im(\partial_{k+1}) \tag{5.9}$$

Par conséquent, le groupe d'homologie à 0 dimension H_0 représente les composantes connexes du complexe, le groupe d'homologie à 1 dimension H_1 représente les trous ou boucles à 1 dimension, le groupe d'homologie à 2 dimensions H_2 représente les trous ou cavités en 2 dimensions,...

3.3 Diagramme de persistance

Définition 5.10. *Une filtration est une séquence de complexes simpliciaux* $(\mathbf{G}_\alpha)_{\alpha \in I \subset \mathbb{R}}$ *ordonnés par inclusion, c'est-à-dire* $\mathbf{G}_{\alpha'} \subset \mathbf{G}_{\alpha''}$ *si* $\alpha' \leq \alpha''$ *pour n'importe quel nombre* α', α'' *de I.*

Définition 5.11. *Un diagramme de persistance d'un filtrage* $(\mathbf{G}_\alpha)_{\alpha \in I \subset \mathbb{R}}$ *est la diagonale* $\{(x, y) \in \mathbb{R}^2 | \, x = y\}$ *avec un ensemble de points* $\{(b, d) \in \mathbb{R}^2 | \, b < d\}$ *tel que chaque point* (b, d) *correspond à une caractéristique topologique comme suit: b est la plus petite valeur de* $\alpha \in I$ *telle que la caractéristique apparaisse dans* $\mathbf{G}_\alpha$, *et d est la plus petite valeur de* $\alpha \in I$ *telle que* $\alpha > b$ *et la caractéristique disparaisse dans* $\mathbf{G}_\alpha$.

Nous appelons b l'échelle de naissance, et d l'échelle de mort du trait. La différence $d - b$ *est appelée la persistance de la caractéristique.*

En analyse de séries temporelles, en construisant le diagramme de persistance associé à l'espace d'état $\mathbf{Y} = (\mathbf{y}_t^{\tau,d})$ d'une série temporelle, on peut extraire les informations topologiques de l'espace d'état intrinsèque sous le changement de la résolution spatiale. Ainsi, l'information est la robustesse aux bruits [Cohen-Steiner, 2007]. Les filtrages $(\text{Rips}_\alpha(\mathbf{V}))_{\alpha\geq 0}$ et $\left(\check{\text{C}}\text{ech}_\alpha(\mathbf{V})\right)_{\alpha\geq 0}$ sont souvent utilisés pour construire le diagramme de persistance.

3.4 Distance du goulot d'étranglement et distance de Wasserstein

Il existe deux mesures usuelles pour quantifier la similarité entre deux diagrammes de persistance: la distance du goulot d'étranglement et la distance de Wasserstein [Chazal, 2021].

Définition 5.12. *La distance du goulot d'étranglement entre deux diagrammes de persistance D_1 et D_2 est définie par:*

$$W_\infty(D_1, D_2) = \inf_{matching\ m} \sup_{(u,v)\in m} \|u - v\|_\infty \tag{5.10}$$

Définition 5.13. *La distance de Wasserstein entre deux diagrammes de persistance D_1 et D_2 est définie par:*

$$W_p(D_1, D_2) = \inf_{matching\ m} \left(\sum_{(u,v)\in m} \|u - v\|_\infty^p \right)^{\frac{1}{p}} \tag{5.11}$$

où $\|s\|_\infty = \max_{i=\overline{1,d}} |s_i|$ pour tout $s = (s_i) \in \mathbb{R}^d$.

Cependant, nous montrons que les métriques ne sont pas appropriées pour mesurer la différence entre deux diagrammes de persistance si leurs nombres de points en dehors de la diagonale sont trop différents.

4 Détection d'anomalies de la dynamique d'un indice de marché à partir de ses caractéristiques topologiques

Dans [Nguyen, 2021a], nous utilisons le TDA combiné à la méthode d'intégration à retardement pour détecter des anomalies dans le comportement d'un marché boursier.

4.1 Méthodes de recherche

Nous étudions la série chronologique du rendement d'un indice de marché et considérons la dynamique dramatiquement étrange du rendement comme des anomalies du marché. Cela conduit à comparer des diagrammes de persistance construits à partir du rendement de l'indice dans la période courante et les périodes précédentes. Nous utilisons les termes "données de test" et "données d'entraînement" pour la série chronologique du retour de l'indice que nous voulons détecter des anomalies et sa série chronologique dans les périodes précédentes, respectivement. Notre méthode est résumée dans l'algorithme 9.

δ permet de mesurer l'écart de la structure topologique du retour d'index par rapport à ses structures antérieures. Une valeur plus élevée de δ implique une plus grande variation de la dynamique de retour de l'indice de la période de test aux précédentes.

Algorithm 9 Calculer l'écart de la structure topologique δ de la dynamique de retour de l'indice d'une certaine période aux précédentes.

Require: retour d'index $(x_t)_{t=\overline{1,T}}$ comme données d'apprentissage, retour d'index $\left(x_t^{'}\right)_{t=\overline{1,T'}}$ comme données de test,

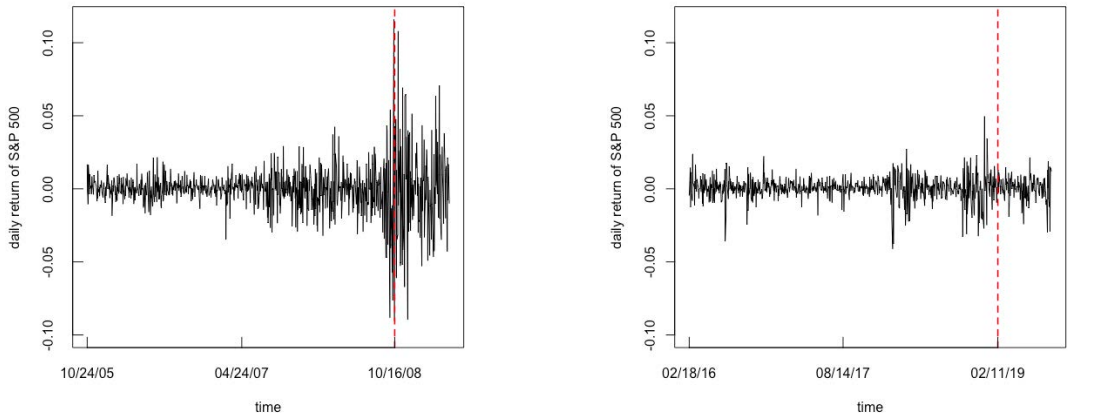
procedure VARIATION_DE_LA_STRUCTURE_TOPOLOGIQUE((x_t), $(x_t^{'})$)
  $\tau \leftarrow$ le délai optimal de (x_t)
  $d \leftarrow$ la dimension d'encastrement optimale de (x_t)
  ▷ calculer le temps d'intégration de (x_t) et $\left(x_t^{'}\right)$
  $\mathbf{y}_t \leftarrow \left(x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(d-1)\tau}\right),\ t = \overline{1, T-(d-1)\tau}$
  $\mathbf{y}_t^{'} \leftarrow \left(x_t^{'}, x_{t+\tau}^{'}, x_{t+\tau}^{'}, \ldots, x_{t+(d-1)\tau}^{'}\right),\ t = \overline{1, T'-(d-1)\tau}$
  $m \leftarrow T'-(d-1)\tau$ ▷ le nombre de vecteurs $\mathbf{y}_t^{'}$s
  ▷ diviser $\mathbf{y}_t$ en s segments consécutifs de longueur m
  $segment_j \leftarrow \left(\mathbf{y}_{1+(j-1)m}, \mathbf{y}_{2+(j-1)m}, \ldots, \mathbf{y}_{jm}\right),\ j = \overline{1,s}$
  ▷ calculer des diagrammes de persistance
  $dgm \leftarrow$ le diagramme de persistance construit à partir de $\left(\mathbf{y}_t^{'}\right)$
  $dgm_j \leftarrow$ le diagramme de persistance construit à partir de $segment_j,\ j = \overline{1,s}$
  $total\_dgm \leftarrow$ fusionner tous les diagrammes de persistance $dgm_j,\ j = \overline{1,s}$
  ▷ calculer la distribution ponctuelle des diagrammes de persistance
  $cluster_i \leftarrow$ points attribués au i-ème groupe après partitionnement des points hors de la diagonale de $total\_dgm$ en k clusters sur la base des emplacements des points et de leurs groupes d'homologie correspondants
  for $i \in \overline{1, k+1}$ **do**
    $region_i \leftarrow$ la région de $\mathbb{R}^2$ identifiée par $cluster_i$, où la dernière région est le reste de l'espace
    $P_i \leftarrow \left\langle \frac{n_{ij}}{n_j} \right\rangle_{j=\overline{1,s}}$, où n_{ij} et n_j sont respectivement le nombre de points de dgm_j appartenant à $region_i$ et le nombre de points de dgm_j, sauf la diagonale
    $Q_i \leftarrow \frac{n_i^{'}}{n^{'}}$, où $n_i^{'}$ est le nombre de points dans dgm appartenant à $region_i$, et n' est le nombre de points dans dgm, à l'exception de la diagonale, respectivement
  end for
  ▷ calculer comment les distributions ponctuelles des diagrammes construits à partir des données de test et des données de formation sont différentes
  $\delta \leftarrow \sqrt{\sum_{i=1}^{k+1} (P_i - Q_i)^2}$ ▷ Output
end procedure

4.2 Résultats empiriques avec l'indice S&P 500

Nous vérifions l'efficacité de notre méthode dans le cas de l'indice S&P 500 du 18/12/1972 au 04/08/2020. Nous utilisons l'AMI pour trouver le délai idéal et la méthode donnée dans [Cao, 1997] pour trouver la dimension de plongement idéale. Pour moins de calcul, nous utilisons la famille des complexes de Vietoris-Rips pour construire les diagrammes de persistance. Étant donné que les caractéristiques bidimensionnelles de nos diagrammes de persistance peuvent être considérées comme des bruits, nous nous concentrons uniquement sur les caractéristiques 0 et 1 dimension. Pour partitionner les points du diagramme total, à l'exception de la diagonale, en petits clusters, nous utilisons l'algorithme k-mean [Hartigan, 1979] pour résoudre rapidement le problème et obtenir un résultat acceptable.

Pour tout point du diagramme de persistance construit à partir des données de test, à l'exception de la diagonale, nous attribuons le point à son cluster le plus proche ayant le même groupe d'homologie si la persistance de la caractéristique correspondant au point n'est pas trop différente de celles correspondant aux points du cluster .

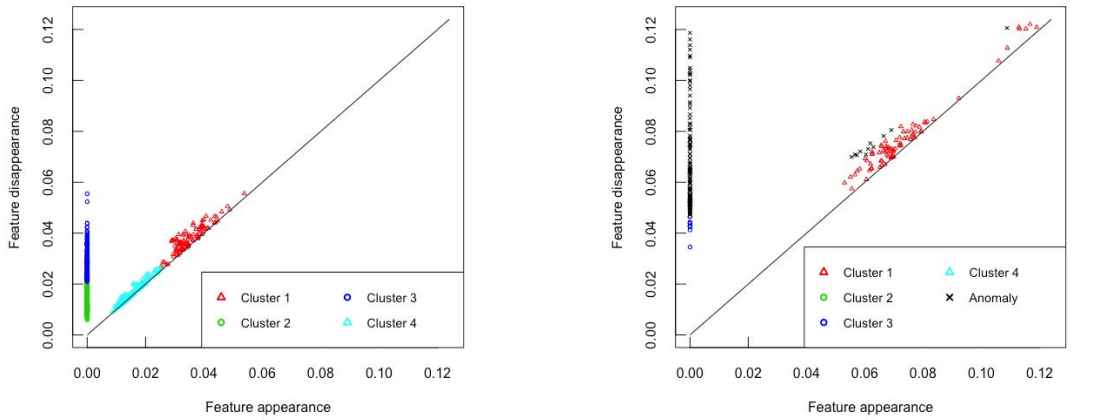
Nous montrons que notre méthode peut détecter la différence significative entre le comportement du retour d'index et son comportement historique (les figures 5.1, 5.2 et 5.3). En fait, δ vaut 83.7% dans le cas de la figure 5.1a mais seulement 11.9% dans le cas de la figure 5.1b.



(a) Rendement quotidien de l'indice S&P 500 du 24/10/2005 au 27/04/2009.

(b) Rendement quotidien de l'indice S&P 500 du 18/02/2016 au 19/08/2019.

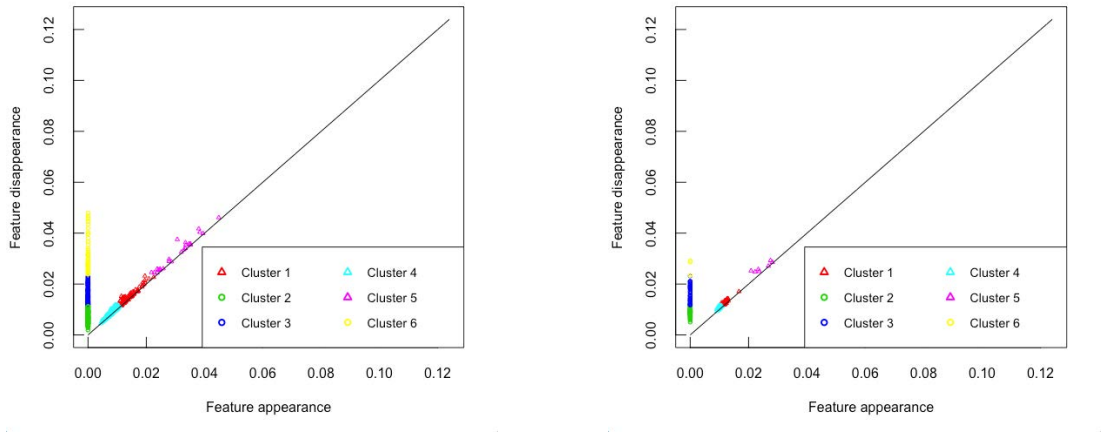
Figure 5.1: Deux exemples de bases de données où les données de test sont à droite de la ligne pointillée et les données d'apprentissage sont à gauche.



(a) Diagramme total

(b) Diagramme de persistance construit à partir des données de test

Figure 5.2: Détection d'anomalies topologiques des données de test dans la base de données illustrée à la figure 5.1a. Les cercles représentent les entités à 0 dimension et les triangles représentent les entités à 1 dimension. Le signe noir × indique des caractéristiques anormales qui ne peuvent être attribuées à aucun groupe du diagramme total.

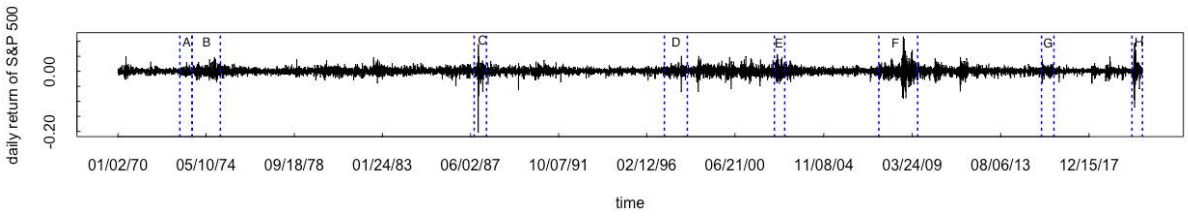


(a) Diagramme total

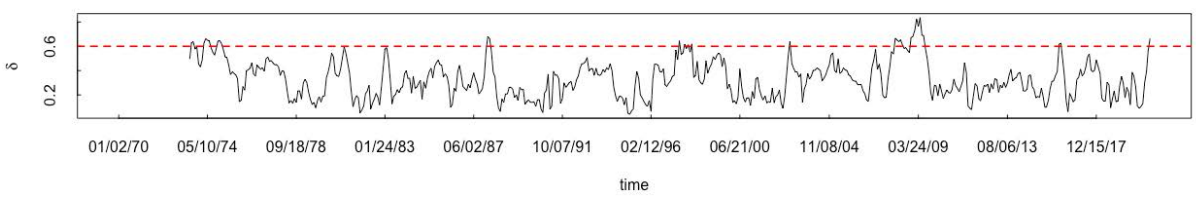
(b) Diagramme de persistance construit à partir des données de test

Figure 5.3: Il n'y a aucune caractéristique anormale dans le diagramme de persistance construit à partir des données de test dans la base de données illustrée dans la figure 5.1b. Les cercles représentent les entités à 0 dimension et les triangles représentent les entités à 1 dimension.

Surtout pour l'ensemble de nos 541 données de test, nous avons constaté que chaque fois que δ est supérieur à 60%, il y a de graves krachs ou récessions du marché, y compris le krach boursier de 1973-1974, la stagflation des années 1970, le "lundi noir" de 1987, la crise de l'épargne et du crédit de 1989, le début du boom économique de la fin des années 1990 aux États-Unis, le ralentissement du marché boursier de 2002 - le pire résultat du krach de la dot-com 2000 - 2002, la crise financière de 2008 - la pire crise aux États-Unis depuis la Grande Dépression de 1929, la récession COVID-19. Les périodes correspondant à ces grandes valeurs de δ sont nommées de A à H dans la figure 5.4b,



(a) Rendement quotidien de l'indice S&P 500 du 01/02/1972 au 08/04/2020.



(b) Value of δ plotted at the last days of test periods

Figure 5.4: Dynamique de δ et rendement de l'indice S&P 500.

En conséquence la méthode que nous proposons est un outil efficace pour détecter des anomalies dans la dynamique d'un indice de marché. Son résultat fournit un moyen simple de reconnaître le début d'une crise financière en analysant l'indice représentatif du marché boursier correspondant au lieu d'obtenir une analyse complète de nombreuses statistiques micro et macro. Par conséquent, nous suggérons que l'écart topologique δ de la dynamique de rendement d'un indice peut être une mesure efficace du risque systémique. Il est particulièrement approprié pour les investisseurs individuels et les systèmes de négociation automatique.

Conclusion

Dans cette thèse, nous avons utilisé diverses techniques issues de la science des ssytèmes complexes, notamment la science des réseaux, la théorie des matrices aléatoires, la théorie de la prétopologie et la TDA, pour étudier les caractéristiques et le mécanisme du comportement collectif d'un marché boursier sous différents aspects. Concrètement, nous avons étudié des marchés réels, dont la bourse américaine et la bourse vietnamienne, et nous avons utilisé le langage R pour mettre en œuvre nos travaux empiriques.

En particulier, nous avons constaté que le MST d'un réseau basé sur la corrélation d'un marché boursier est courant pour résumer la structure du réseau, car le MST fournit le chemin le plus probable dans lequel un choc boursier se propage à l'ensemble du marché. En revanche, le réseaux à seuil basé sur la corrélation est plus approprié que le MST pour étudier la résilience du marché en raison de l'inconvénient du MST à négliger de nombreuses corrélations d'actions, qui peuvent être très importantes. Cependant, étant donné que la matrice de corrélation croisée est calculée à partir des cours boursiers historiques, nous n'obtenons que la matrice d'échantillon. Selon le RMT, la plus grande valeur propre de la matrice d'échantillonnage et son vecteur propre unitaire associé peuvent donner des informations sur les "vraies" corrélations des actions car la valeur propre est extrêmement supérieure à la limite supérieure de la distribution de Marčenko - Pastur. Ainsi, nous pouvons utiliser le premier PC dont les chargements sont les composantes du vecteur propre pour étudier le comportement collectif du marché, comportement qui se reflète également dans la dynamique de l'indice de marché.

Avec ces outils, nous avons fourni une analyse complète de la dynamique et de la stabilité d'un marché boursier dans cette thèse. Tout d'abord, après avoir étudié la dynamique du réseau MST du marché, nous avons confirmé que l'état instable du marché peut se refléter sur la structure en étoile du réseau ou sur la disparition de la propriété sans échelle du réseau. De plus, cet état peut être quantifié par le déclin remarquable de différentes mesures telles que la longueur du chemin le plus court, le taux de survie, le même rapport de secteur et le coefficient allométrique. De plus, en tant que réseau sans échelle, nous avons également établi en utilisant des données réelles que le réseaux à seuil basé sur la corrélation reste robuste en cas de défaillance aléatoire mais très fragile en cas d'attaques intentionnelles contre ses nœuds les plus connectés ou ses nœuds les plus chargés. Ce résultat a démontré la robustesse d'un marché boursier lorsque certaines entreprises font faillite à cause de leur mauvaise gestion. Cependant, lorsque les actions ordinaires des entreprises jouent un rôle important dans la structure du réseau, par exemple,

elles sont les nœuds les plus connectés ou les nœuds les plus chargés, les faillites nuiront à la connectivité du réseau. Cela a un impact négatif sur la stabilité des marchés.

Ensuite, nous avons étudié la plus grande valeur propre de la matrice de corrélation croisée empirique des actions et son vecteur propre unitaire associé. Alors que d'autres travaux ont montré que la valeur propre devient plus grande dans les crises financières, nous avons suggéré de composer le portefeuille le plus corrélé à partir du premier PC des rendements boursiers. Étant donné que la valeur propre est toujours dominante dans le spectre de la matrice, le premier PC explique la plupart des variances des rendements boursiers. Ainsi, il peut être considéré comme le facteur de marché et est fortement corrélé à l'indice de marché correspondant. De plus, nous avons établi une formule simple pour approximer ses charges en fonction de la relation asymptotiquement linéaire des charges avec les coefficients de corrélation moyens des actions.

D'autre part, nous avons montré empiriquement le rôle principal des sociétés financières dans la stabilité d'un marché boursier car les sociétés restent généralement dans les hubs du réseau MST, en particulier le réseau en étoile. Aussi, le secteur financier est dominant dans les chargements des premiers PC.

De plus, comme le comportement collectif d'un marché peut être causé par une défaillance en cascade, nous avons proposé une méthode pour étudier l'évolution de la défaillance. Nous avons considéré une action comme un composant défaillant si son prix baisse de façon spectaculaire. En partant de l'hypothèse que le nombre d'actions défaillants augmente l'impact de ces actions sur un autre action et déclenche sa défaillance si l'impact est suffisamment important, nous avons conçu un espace prétopologique dans lequel la pseudo-fermeture modélise la contagion de la défaillance d'un groupe d'actions. En revanche, l'ouverture de la rémunération du groupe permet de prédire les actions non impactés par les actions défaillantes. Nous avons constaté que notre cadre prétopologique est plus efficace que le réseau MST et le réseaux à seuil basé sur la corrélation pour modéliser l'évolution de la défaillance en cascade. L'efficacité vient de la prise en compte de toutes les corrélations des actions, illustrant évidemment la contagion par étapes individuelles et notant non seulement la relation des actions mais aussi la relation entre une action et un groupe.

Enfin, nous proposons une méthode pour détecter les anomalies dans le comportement collectif d'un marché boursier. Étant donné que le facteur de marché représenté par le premier PC des rendements boursiers est souvent fortement corrélé au rendement de l'indice du marché, la dynamique du rendement de l'indice est une donnée appropriée pour étudier le comportement collectif du marché. Nous établissons une mesure pour reconnaître comment les caractéristiques topologiques de la série chronologique de l'indice obtenues au cours d'une certaine période sont différentes de celles de la série chronologique de l'indice obtenues au cours des périodes précédentes. Cette mesure est testée dans le cas de l'indice S&P 500. Nous avons constaté que la mesure de l'écart aide vraiment à détecter les crashs importants sur le marché américain lorsqu'il est supérieur à 60%. Parce qu'elle prend souvent une telle valeur dès le début des crises financières, cette valeur peut être un avertissement de crises au lieu de passer beaucoup de temps à analyser de nombreuses statistiques économiques.

En conséquence, cette thèse permet d'acquérir une connaissance approfondie de l'évolution

des marchés boursiers, des structures géométriques et des signes de stabilité qui sont extrêmement précieux pour contrôler le risque systémique. Nous pouvons améliorer les recherches ci-dessus avec des modèles plus appropriés, tels que les espaces prétopologiques de type $\mathcal{V}$ pour l'échec en cascade des marchés boursiers, ou améliorer la mesure de l'écart des caractéristiques topologiques d'un marché indiciel avec d'autres outils de TDA. En outre, nous prévoyons également d'étudier davantage le rôle du premier PC dans le calcul du coefficient β d'une action. En général, le point de vue scientifique selon lequel les marchés financiers sont des systèmes complexes ouvre de nouvelles théories et technologies pour étudier les caractéristiques et la dynamique de ces marchés. Par conséquent, cette approche continuera à prendre plus d'intérêt dans nos futurs travaux avec d'autres méthodes de science complexe telles que la modélisation à base d'agents.

Bibliography

[Abarbanel, 1993] Henry D. I. Abarbanel, Reggie Brown, John J. Sidorowich, and Lev Sh. Tsimring. "The analysis of observed chaotic data in physical systems". *Reviews of Modern Physics* 65.4 (1993), pp. 1331–1392 (cit. on pp. 92, 150).

[Adamic, 2001] Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. "Search in power law networks". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 64.4 (2001), p. 046135 (cit. on p. 34).

[Albert, 1999] Réka Albert, Hawoong Jeong, and Albert-László Barabási. "Diameter of the World-Wide Web". *Nature* 401 (1999), pp. 130–131 (cit. on p. 9).

[Albert, 2000] Réka Albert, Hawoong Jeong, and Albert-László Barabási. "Error and attack tolerance of complex networks". *Nature* 406 (2000), pp. 378–382 (cit. on pp. 37, 38, 44).

[Amaral, 2000] Luís A. N. Amaral, Atonio Scala, Marc Barthélémy, and H. Eugene Stanley. "Classes of small-word networks". *Proceedings of the National Academy of Sciences of the USA*. Vol. 97. 21. 2000, pp. 11149–11152 (cit. on p. 34).

[Auray, 1979] Jean-Paul Auray, Gérard Duru, and Michel Mougeot. "A pretopological analysis of the input-output model". *Economics Letters* 2.4 (1979), pp. 343–347 (cit. on pp. 78, 79, 144).

[Azevedo, 2015] Hátylas Azevedo and Carlos A. Moreira-Filho. "Topological robustness analysis of protein interaction networks reveals key targets for overcoming chemotherapy resistance in glioma". *Scientific Reports* 5 (2015), p. 16830 (cit. on p. 44).

[Banavar, 1999] Jayanth R. Banavar, Amos Maritan, and Andrea Rinaldo. "Size and form in efficient transportation networks". *Nature* 399 (1999), pp. 130–132 (cit. on p. 28).

[Barabási, 2016] Albert-László Barabási. *Network Science*. 1st edition. Cambridge University Press, 2016 (cit. on pp. 33, 34).

[Beinhocker, 2006] Eric D. Beinhocker. *The Origin of Wealth: Evolution, Complexity and the Radical Remaking of Economics*. Harvard Business Review Press, 2006 (cit. on pp. 7, 11, 103).

[Belmandt, 2011] ZT Belmandt. *Basics of Pretopology*. HERMANN, 2011 (cit. on pp. 10, 75, 78, 88, 143, 148).

[Ben-Amor, 2010] Soufian Ben-Amor, Marc Bui, and Michel Lamure. "Modeling urban aerial pollution using stochastic pretopology". *Africa Mathematics Annals* 1.1 (2010), pp. 7–19 (cit. on pp. 78, 79, 144).

[Ben-Amor, 2006] Soufian Ben-Amor, Ivan Lavallée, and Marc Bui. "Percolation, pretopology and complex systems modeling". *Complex Systems Modeling and Cognition Eurocontrol and EPHE Joint Research Lab* (2006) (cit. on p. 10).

[Berry, 2002] Brian J. L. Berry, L. Douglas Kiel, and Euel Elliott. "Adaptive agents, intelligence, and emergent human organization: Capturing complexity through agent-based modeling". *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 99. 3. 2002, pp. 7187–7188 (cit. on pp. 9, 119).

[Bonanno, 2004] Giovanni Bonanno, Guido Caldarelli, Fabrizio Lillo, Salvatore Miccichè, Nicolas Vandewalle, and Rosario N. Mantegna. "Networks of equities in financial markets". *The European Physical Journal B* 38 (2004), pp. 363–371 (cit. on pp. 14, 16, 18, 121).

[Bonanno, 2000] Giovanni Bonanno, Nicolas Vandewalle, and Rosario N. Mantegna. "Taxonomy of stock market indices". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 62.6 (2000), R7615(R) (cit. on p. 14).

[Bonnevay, 2009] Stephane Bonnevay. "Pretopological operators for gray-level image analysis". *Studia Informatica Universalis, Hermann* 7.1 (2009), pp. 27–44 (cit. on p. 78).

[Broomhead, 1986] D. S. Broomhead and Gregory P. King. “Extracting qualitative dynamics from experimental data”. *Physica D: Nonlinear Phenomena* 20.2-3 (1986), pp. 217–236 (cit. on pp. 93, 150).

[Bui, 2019] Q. Vu Bui, Soufian Ben-Amor, and Marc Bui. “Stochastic pretopology as a tool for complex networks analysis”. *Journal of Information and Telecommunication* 3.2 (2019), pp. 135–155 (cit. on pp. 10, 11).

[Callaway, 2000] Duncan S. Callaway, M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. “Network robustness and fragility: Percolation on random graphs”. *Physical Review Letters* 85.25 (2000), pp. 5468–5471 (cit. on pp. 37, 38, 129).

[Cancho, 2001] Ramon Ferrer I. Cancho and Ricard V. Solé. “The small word of human language”. *Proceedings of the Royal Society of London B: Biological Sciences*. Vol. 268. 1482. 2001, pp. 2261–2265 (cit. on p. 34).

[Cao, 1997] Liangyue Cao. “Practical method for determining the minimum embedding dimension of a scalar time series”. *Physica D: Nonlinear Phenomena* 110.1-2 (1997), pp. 43–50 (cit. on pp. 94, 150, 154).

[Chazal, 2021] Frédéric Chazal and Bertrand Michel. “An introduction to Topological Data Analysis: Fundamental and practical aspects for data scientists”. *Frontiers in Artificial Intelligence* 4 (2021), p. 667963 (cit. on pp. 95, 102, 151, 153).

[Chen, 2002] Qian Chen, Hyunseok Chang, Ramesh Govindan, Sugih Jamin, Scott J. Shenker, and Walter Willinger. “The origin of power laws in internet topologies revisited”. *Proceedings of the Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002)*. Ed. by Parviz Kermani, David Lee, and Ariel Orda. Vol. 2. 2002, pp. 608–617 (cit. on p. 34).

[Cohen, 2000] Reuven Cohen, Keren Erez, Daniel ben-Avraham, and Shlomo Havlin. “Resilience of the internet to random breakdowns”. *Physical Review Letters* 85.21 (2000), pp. 4626–4628 (cit. on pp. 25, 26, 37, 38, 125, 129).

[Cohen, 2001] Reuven Cohen, Keren Erez, Daniel ben-Avraham, and Shlomo Havlin. “Breakdown of the internet under intentional attack”. *Physical Review Letters* 86.16 (2001), pp. 3682–3685 (cit. on pp. 14, 37, 38, 44, 129).

[Cohen, 2010] Reuven Cohen and Shlomo Havlin. *Complex Networks: Structure, Robustness and Function*. 1st. Cambridge University Press, 2010 (cit. on pp. 38, 129).

[Cohen-Steiner, 2007] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. “Stability of persistence diagrams”. *Discrete & Computational Geometry* 37.1 (2007), pp. 103–120 (cit. on pp. 101, 153).

[Conlon, 2007] Thomas Conlon, Heather J. Ruskin, and Martin Crane. “Random matrix theory and fund of funds portfolio optimization”. *Physica A: Statistical Mechanics and its Applications* 382.2 (2007), pp. 565–576 (cit. on p. 70).

[Coronnello, 2005] Claudia Coronnello, Michele Tumminello, Fabrizio Lillo, Salvatore Micciché, and Rosario N. Mantegna. “Sector identification in a set of stock return time series traded at the London Stock Exchange”. *Acta Physica Polonica B* 36.9 (2005), pp. 2653–2679 (cit. on p. 70).

[Crucitti, 2004a] Paolo Crucitti, Vito Latora, and Massimo Marchiori. “Model for cascading failures in complex networks”. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 69.4 Pt 2 (2004), 045104(R) (cit. on pp. 73, 74).

[Crucitti, 2004b] Paolo Crucitti, Vito Latora, Massimo Marchiori, and Andrea Rapisarda. “Error and attack tolerance of complex networks”. *Physica A: Statistical Mechanics and its Applications* 340.1-3 (2004), pp. 388–394 (cit. on p. 38).

[Cvetković, 1998] Dragoš M. Cvetković, Michael Doob, and Horst Sachs. *Spectra of Graphs: Theory and Applications*. 3rd. Wiley-VCH, 1998 (cit. on p. 56).

[Dalud-Vincent, 2011] Monique Dalud-Vincent, Marcel Brissaud, and Michel Lamure. “Pretopology, Matroïdes and hypergraphs”. *International Journal of Pure and Applied Mathematics* 67.4 (2011), pp. 363–375 (cit. on pp. 77, 144).

[Dorogovtsev, 2001] Sergey N. Dorogovtsev and José Fernando F. Mendes. “Language as an evolving word web”. *Proceedings of the Royal Society of London B: Biological Sciences*. Vol. 268. 1485. 2001, pp. 2603–2606 (cit. on p. 34).

[Drożdz, 2000] Stanislaw Drożdz, Frank Grümmer, Andrzej Z. Górski, F. Ruf, and Josef Speth. “Dynamics of competition between collectivity and noise in the stock market”. *Physica A: Statistical Mechanics and its Applications* 287.3-4 (2000), pp. 440–449 (cit. on pp. 60, 137).

[Duan, 2007] W. Q. Duan. “Universal scaling behaviour in weighted trade networks”. *European Physical Journal B* 59.2 (2007), pp. 271–276 (cit. on p. 28).

[Dunne, 2002] Jennifer A. Dunne, Richard J. Williams, and Neo D. Martinez. "Network structure and biodiversity loss in food webs: Robustness increases with connectance". *Ecology Letters* 5.4 (2002), pp. 558–567 (cit. on p. 37).

[Dyson, 1971] F. J. Dyson. "Distribution of eigenvalues for a class of real symmetric matrices". *Revista Mexicana de Física* 20.4 (1971), pp. 231–237 (cit. on p. 57).

[Eckmann, 1986] Jean-Pierre Eckmann, Sylvie O. Kamphorst, D. Ruelle, and Sergio Ciliberto. "Liapunov exponents from time series". *Physical Review A - Atomic, Molecular, and Optical Physics* 34.6 (1986), pp. 4971–4979 (cit. on pp. 93, 150).

[Edelsbrunner, 2002] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. "Topological persistence and simplification". *Discrete & Computational Geometry* 28 (2002), pp. 511–533 (cit. on p. 95).

[Faloutsos, 1999] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. "On power-law relationships of the Internet topology". *ACM SIGCOMM Computer Communication Review*. Ed. by Craig Partridge and Roch Guerin. Vol. 29. 4. 1999, pp. 251–262 (cit. on pp. 9, 34).

[Farmer, 2009] J. Doyne Farmer and Duncan Foley. "The economy needs agent-based modelling". *Nature* 460 (2009), pp. 685–686 (cit. on p. 9).

[Foote, 2007] Richard Foote. "Mathematics and complex systems". *Science* 318.5849 (2007), pp. 410–412 (cit. on pp. 6, 119).

[Gallager, 1968] Robert G. Gallager. *Information Theory and Reliable Communication*. New York, Wiley, 1968 (cit. on p. 92).

[Gallos, 2006] Lazaros K. Gallos, Reuven Cohen, Fredrik Liljeros, Panos Argyrakis, Armin Bunde, and Shlomo Havlin. "Attack strategies on complex networks". *Workshop on "Networks: structure and dynamics" in Conference ICCS, Lecture Notes in Computer Science*. Ed. by Vassil N. Alexandrov, Geert Dick van Albada, Peter M. A. Sloot, and Jack Dongarra. Vol. 3993. 2006, pp. 1048–1055 (cit. on pp. 37, 38).

[Garas, 2007] Antonios Garas and Panos Argyrakis. "Correlation study of the Athens Stock Exchange". *Physica A: Statistical Mechanics and its Applications* 380 (2007), pp. 399–410 (cit. on p. 35).

[Garas, 2008] Antonios Garas, Panos Argyrakis, and Shlomo Havlin. "The structural role of weak and strong links in a financial market network". *European Physical Journal B* 63.2 (2008), pp. 265–271 (cit. on pp. 20, 30, 123, 126).

[Garlaschelli, 2003] Diego Garlaschelli, Guido Caldarelli, and Luciano Pietronero. "Universal scaling relations in food webs". *Nature* 423 (2003), pp. 165–168 (cit. on pp. 28, 126).

[Gilbert, 2008] Nigel Gilbert. *Agent-based Models*. SAGE Publications, Inc., 2008 (cit. on p. 9).

[Gopikrishnan, 2001] Parameswaran Gopikrishnan, Bernd Rosenow, Vasiliki Plerou, and H. Eugene Stanley. "Quantifying and interpreting collective behavior in financial markets". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 64.3 Pt 2 (2001), p. 035106 (cit. on pp. 66, 139).

[Gower, 1966] John C. Gower. "Some distance properties of latent root and vector methods used in multivariate analysis". *Biometrika* 53.3-4 (1966), pp. 325–338 (cit. on pp. 14, 121).

[Grassberger, 1983] Peter Grassberger and Itamar Procaccia. "Characterization of strange attractors". *Physical Review Letters* 50.5 (1983), pp. 346–349 (cit. on pp. 94, 150).

[Grimm, 2005] Volker Grimm and Steven F. Railsback. *Individual-based Modeling and Ecology*. 1st. Princeton University Press, Princeton, New Jersey, 2005 (cit. on p. 9).

[Guérard, 2015] Guillaume Guérard, Soufian Ben-Amor, and Alain Bui. "A context-free Smart Grid model using pretopologic structure". *Proceedings of the 4th International Conference on Smart Cities and Green ICT Systems - Volume 1: SMARTGREENS*. Ed. by Markus Helfert, Karl-Heinz Krempels, Brian Donnellan, and Cornel Klein. Vol. 1. SciTePress, Lisbon, Portugal, 2015, pp. 335–341 (cit. on p. 78).

[Guhr, 1998] Thomas Guhr, Axel Müller-Groeling, and Hans A. Weidenmüller. "Random-matrix theories in quantum physics: Common concepts". *Physics Reports* 299.4-6 (1998), pp. 189–425 (cit. on p. 56).

[Hartigan, 1979] John A. Hartigan and M. Anthony Wong. "Algorithm AS 136: A k-means clustering algorithm". *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 28.1 (1979), pp. 100–108 (cit. on pp. 106, 154).

[Jalan, 2007] Sarika Jalan and Jayendra N. Bandyopdhyay. "Random matrix analysis of complex networks". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 76.4 Pt 2 (2007), p. 046107 (cit. on p. 10).

[Jeong, 2001] Hawoong Jeong, S. P. Mason, Albert-Laszlo Barabási, and Zoltan N. Oltvai. "Lethality and centrality in protein networks". *Nature* 411.6833 (2001), pp. 41–42 (cit. on p. 37).

[Jeong, 2000] Hawoong Jeong, B. Tombor, Rita R. Albert, Zoltan N. Oltvai, and Albert-Laszlo Barabási. "The large-scale organization of metabolic networks". *Nature* 407 (2000), pp. 651–654 (cit. on pp. 9, 14, 34).

[Jiang, 2012] Xiongfei Jiang and Biaowen Zheng. "Anti-correlation and subsector structure in nancial systems". *Europhysics Letters* 97.4 (2012), p. 48006 (cit. on p. 66).

[Jolliffe, 1986] Ian T. Jolliffe. *Principal Component Analysis*. 1st. Springer New York, NY, 1986 (cit. on pp. 61, 138).

[Kantz, 2003] Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. 2nd. Cambridge University Press, Cambridge, Massachusetts, 2003 (cit. on p. 91).

[Kauffman, 1995] Stuart Kauffman and William MacReady. "Technological evolution and adaptive organizations". *Complexity* 1.2 (1995), pp. 26–43 (cit. on p. 6).

[Kennel, 1992] Matthew B. Kennel, Reggie Brown, and Henry D. I. Abarbanel. "Determining embedding dimension for phase-space reconstruction using a geometrical construction". *Physical Review A - Atomic, Molecular, and Optical Physics* 45.6 (1992), pp. 3403–3411 (cit. on pp. 94, 150).

[Kirman, 2011] Alan Kirman. *Complex Economics: Individual and Collective Rationality*. 1st. Routledge, 2011 (cit. on pp. 7, 11, 103).

[Kuratowski, 1922] Kazimierz Kuratowski. "Sur l'opération ā de l'analysis situs". *Fundamenta Mathematicae* 3 (1922), pp. 182–199 (cit. on p. 87).

[Ladyman, 2013] James Ladyman, James Lambert, and Karoline Wiesner. "What is a complex system?" *European Journal for Philosophy of Science* 3.1 (2013), pp. 33–67 (cit. on pp. 6, 119).

[Laloux, 1999] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. "Noise dressing of financial correlation matrices". *Physical Review Letters* 83.7 (1999), p. 1467 (cit. on pp. 56, 59, 60, 137).

[Laloux, 2000] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. "Random matrix theory and financial correlations". *International Journal of Theoretical and Applied Finance* 3.3 (2000), pp. 391–397 (cit. on pp. 10, 56, 61).

[Lamure, 2009] Michel Lamure, Stephane Bonnevay, Marc Bui, and Soufian Ben-Amor. "A stochastic and pretopological modeling aerial pollution of an urban area". *Studia Informatica Universalis* 7.3 (2009), pp. 410–426 (cit. on pp. 78, 79, 144).

[Lautier, 2012] Delphine Lautier and Franck Raynaud. "Systemic risk in energy derivative markets: A graph-theory analysis". *The Energy Journal* 33.3 (2012), pp. 215–239 (cit. on p. 28).

[Lautier, 2013] Delphine Lautier and Franck Raynaud. "Systemic risk and complex systems: A graph-theory analysis". *Econophysics of Systemic Risk and Network Dynamics. New Economic Windows*. Ed. by Frédéric Abergel, Bikas K. Chakrabarti, Anirban Chakraborti, and Asim Ghosh. Springer, Milano, 2013, pp. 19–37 (cit. on pp. 14, 16, 18, 28, 122).

[Levorato, 2014] Vincent Levorato. "Group measures and modeling for social networks". *Journal of Complex Systems* 2014.3 (2014), p. 354385 (cit. on p. 78).

[Levorato, 2010] Vincent Levorato and Marc Bui. "Modeling the complex dynamics of distributed communities of the web with pretopology". *10th International Conference on Innovative Internet Community Systems (I2CS)- Jubilee Edition 2010*. Ed. by Gerald Eichler, Peter Kropf, Ulrike Lechner, Phayung Meesad, and Herwig Unger. Bonn: Gesellschaft für Informatik e.V., 2010, pp. 306–320 (cit. on p. 10).

[Lewin, 1994] Roger Lewin. *Complexity Life at the Edge of Chaos*. Simon & Schuster, 1994 (cit. on pp. 12, 103).

[Liljeros, 2003] Fredrik Liljeros, Christofer R. Edling, and Luís A. Nunes Amaral. "Sexual networks: Implications for the transmission of sexually transmitted infections". *Microbes and Infection* 5.2 (2003), pp. 189–196 (cit. on p. 34).

[Liljeros, 2001] Fredrik Liljeros, Christofer R. Edling, Luís A. Nunes Amaral, H. Eugene Stanley, and Yvonne Åberg. "The web of human sexual contacts". *Nature* 411 (2001), pp. 907–908 (cit. on pp. 14, 34).

[Lux, 1999] Thomas Lux and Michele Marchesi. "Scaling and criticality in a stochastic multi-agent model of a financial market". *Nature* 397 (1999), pp. 498–500 (cit. on pp. 14, 56).

[Mantegna, 1999] Rosario N. Mantegna. "Hierarchical structure in financial markets". *European Physical Journal B - Condensed Matter and Complex Systems* 11 (1999), pp. 193–197 (cit. on pp. 16, 18).

[Mantegna, 2007] Rosario N. Mantegna and H. Eugene Stanley. *An Introduction to Econophysics: Correlations and Complexity in Finance*. 1st. Cambridge University Press, Cambridge, 2007 (cit. on p. 16).

[Marčenko, 1967] Vladimir A. Marčenko and Leonid A. Pastur. "Distributions of eigenvalues of some sets of random matrices". *Mathematics of the USSR-Sbornik* 1.4 (1967), pp. 457–483 (cit. on pp. 57, 137).

[Marti, 2021] Gautier Marti, Frank Nielsen, Mikołaj Bińkowski, and Philippe Donnat. "A review of two decades of correlations, hierarchies, networks and clustering in financial markets". *Progress in Information Geometry. Signals and Communication Technology*. Ed. by Frank (eds) Nielsen. Springer, Cham, 2021, pp. 245–274 (cit. on pp. 19, 122).

[Martin, 2016] Travis Martin, Brian Ball, and Mark E. J. Newman. "Structural inference for uncertain networks". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 93.1 (2016), p. 012306 (cit. on p. 10).

[Massara, 2017] Guido P. Massara, T. Di Matteo, and Tomaso Aste. "Network filtering for big data: Triangulated maximally filtered graph". *Journal of Complex Networks* 5.2 (2017), pp. 161–178 (cit. on p. 18).

[McCarthy, 2000] Ian P. McCarthy, Thierry Rakotobe-Joel, and Gerry Frizelle. "Complex systems theory: Implications and promises for manufacturing organisations". *International Journal of Manufacturing Technology and Management* 2.1-7 (2000), pp. 559–579 (cit. on pp. 6, 119).

[McMahon, 1983] Thomas A. McMahon and John T. Bonner. *On Size and Life*. 1st. Scientific American Books - W. H. Freeman & Co., 1983 (cit. on p. 28).

[Mehta, 2004] Madan L. Mehta. *Random Matrices*. 3rd. Vol. 142. Academic Press, Boston, 2004 (cit. on p. 56).

[Merrill, 2007] Karen R. Merrill. *The Oil Crisis of 1973–1974: A Brief History with Documents*. 1st. Bedford/St. Martin's, 2007 (cit. on p. 12).

[Mitchell, 2006] Melanie Mitchell. "Complex systems: Network thinking". *Artificial Intelligence* 170.18 (2006), pp. 1194–1212 (cit. on p. 14).

[Mitchell, 2011] Melanie Mitchell. *Complexity: A Guided Tour*. 1st. Oxford University Press, 2011 (cit. on p. 7).

[Molley, 1995] Michael Molley and Bruce Reed. "A critical point for random graphs with a given degree sequence". *Random Structures and Algorithms* 6.2-3 (1995), pp. 161–180 (cit. on pp. 25, 37, 125, 129).

[Munkres, 1993] James R. Munkres. *Elements of Algebraic Topology*. 1st. CRC Press, 1993 (cit. on p. 97).

[Newman, 2003] Mark E. J. Newman. "The structure and function of complex networks". *Society for Industrial and Applied Mathematics (SIAM Review)* 45.2 (2003), pp. 167–256 (cit. on pp. 7, 22, 32, 37, 127).

[Newman, 2011] Mark E. J. Newman. "Complex systems: A survey". *American Journal of Physics* 79.8 (2011), pp. 800–810 (cit. on pp. 6–8, 11, 119).

[Nguyen, 2021a] N. K. Khanh Nguyen and Marc Bui. "Detecting anomalies in the dynamics of a market index with Topological Data Analysis". *International Journal of Systematic Innovation* 6.6 (2021), pp. 37–50 (cit. on pp. 103, 105, 108, 153).

[Nguyen, 2021b] N. K. Khanh Nguyen and Marc Bui. "Modeling cascading failures in stock markets by a pretopological framework". *Vietnam Journal of Computer Science* 8.1 (2021), pp. 23–38 (cit. on pp. 82, 83, 85, 146, 147).

[Nguyen, 2018] N. K. Khanh Nguyen and Quang Nguyen. "Resilience of stock cross-correlation network to random breakdown and intentional attack". *Proceedings of International Econometric Conference of Vietnam (ECONVN 2018). Studies in Computational Intelligence*. Ed. by Ly H. Anh, Le S. Dong, Vladik Kreinovich, and Nguyen N. Thach. Vol. 760. Springer, Cham, 2018, pp. 553–561 (cit. on pp. 16, 21, 35, 40, 127, 129).

[Nguyen, 2019a] N. K. Khanh Nguyen, Quang Nguyen, and Marc Bui. "Mining stock market time series and modeling stock price crash using a pretopological framework". *Proceedings of International Conference on Computational Collective Intelligence (ICCCI 2019), Lecture Notes in Computer Science*. Ed. by Ngoc Thanh Nguyen, Richard Chbeir, Ernesto Exposito, Philippe Aniorté, and Bogdan Trawiński. Vol. 11683. Springer, Cham, 2019, pp. 638–649 (cit. on pp. 82, 83, 145).

[Nguyen, 2013] Quang Nguyen. "One-factor model for the cross-correlation matrix in the Vietnamese stock market". *Physica A: Statistical Mechanics and its Applications* 392.13 (2013), pp. 2915–2923 (cit. on pp. 65, 66, 139).

[Nguyen, 2019b] Quang Nguyen and N. K. Khanh Nguyen. "Composition of the first principal component of a stock index – A comparison between SP500 and VNIndex". *Physica A: Statistical Mechanics and its Applications* 536 (2019), p. 120980 (cit. on pp. 16, 19, 59, 60, 65, 122, 137).

[Nguyen, 2019c] Quang Nguyen, N. K. Khanh Nguyen, and Le H.N. Nguyen. "Dynamic topology and allometric scaling behavior on the Vietnamese stock market". *Physica A: Statistical Mechanics and its Applications* 514 (2019), pp. 235–243 (cit. on pp. 16, 19, 35, 45, 46, 49, 122, 127, 131, 133).

[Nie, 2015] Tingyuan Nie, Zheng Guo, Kun Zhao, and Zhe-Ming Lu. "New attack strategies for complex networks". *Physica A: Statistical Mechanics and its Applications* 424 (2015), pp. 248–253 (cit. on p. 43).

[Nobi, 2013] Ashadun Nobi, Seong E. Maeng, Gyeong G. Ha, and Jae W. Lee. "Random matrix theory and cross-correlations in global financial indices and local stock market indices". *Journal of the Korean Physical Society* 62.4 (2013), pp. 569–574 (cit. on pp. 59, 60, 137).

[Onnela, 2002] Jukka-Pekka Onnela, Anirban Chakraborti, Kimmo Kaski, and Janos Kertész. "Dynamic assets trees and portfolio analysis". *European Physical Journal B* 30.3 (2002), pp. 285–288 (cit. on pp. 14, 54, 135).

[Onnela, 2003a] Jukka-Pekka Onnela, Anirban Chakraborti, Kimmo Kaski, and Janos Kertész. "Dynamic assets trees and Black Monday". *Physica A: Statistical Mechanics and its Applications* 324 (2003), pp. 247–252 (cit. on pp. 54, 135).

[Onnela, 2003b] Jukka-Pekka Onnela, Anirban Chakraborti, Kimmo Kaski, Janos Kertész, and Antti Kanto. "Asset trees and asset graphs in financial markets". *Physica Scripta* 2003.T106 (2003), pp. 48–54 (cit. on pp. 14, 16, 20, 30, 35, 123, 126, 127).

[Onnela, 2003c] Jukka-Pekka Onnela, Anirban Chakraborti, Kimmo Kaski, Janos Kertész, and Antti Kanto. "Dynamics of market correlations: Taxonomy and portfolio analysis". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 68.5 (2003), p. 056110 (cit. on pp. 18, 30, 126).

[Onnela, 2004] Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertész. "Clustering and information in correlation based financial networks". *European Physical Journal B* 38.2 (2004), pp. 353–362 (cit. on pp. 18, 20, 123).

[Packard, 1980] Norman H. Packard, James P. Crutchfield, J. Doyne Farmer, and Rob S. Shaw. "Geometry from a time series". *Physical Review Letters* 45.9 (1980), pp. 712–716 (cit. on pp. 91, 149).

[Pan, 2007] Raj Kumar Pan and Sitabhra Sinha. "Collective behavior of stock price movements in an emerging market". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 76.4 (2007), p. 046116 (cit. on pp. 66, 139).

[Parrish, 1999] Julia K. Parrish and Leah Edelstein-Keshet. "Complexity, pattern, and evolutionary trade-offs in animal aggregation". *Science* 284.5411 (1999), pp. 99–101 (cit. on p. 6).

[Perea, 2015] Jose A. Perea, Anastasia Deckard, Steve B. Haase, and John Harer. "Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data". *BMC Bioinformatics* 16 (2015), p. 257 (cit. on p. 103).

[Petermann, 2012] Coralie Petermann, Soufian Ben-Amor, and Alain Bui. "A pretopological multi-agents based model for an efficient and reliable Smart Grid simulation". *14th International Conference on Artificial Intelligence (ICAI)*. CSREA Press, USA, 2012, pp. 354–360 (cit. on p. 78).

[Plerou, 1999] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A. Nunes Amaral, and H. Eugene Stanley. "Universal and nonuniversal properties of cross correlations in financial time series". *Physical Review Letters* 83.7 (1999), pp. 1471–1474 (cit. on pp. 14, 56, 59, 60, 137).

[Plerou, 2002] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A. Nunes Amaral, and H. Eugene Stanley. "Random matrix approach to cross correlations in financial data". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 65.6 (2002), p. 066126 (cit. on pp. 10, 56, 65, 66, 138, 139).

[Qian, 2010] Meng-Cen Qian, Zhi-Qiang Jiang, and Wei-Xing Zhou. "Universal and nonuniversal allometric scaling behaviors in the visibility graphs of world stock market indices". *Journal of Physics A: Mathematical and Theoretical* 43.33 (2010), p. 335002 (cit. on pp. 28, 125, 126).

[Ripeanu, 2002] Matei Ripeanu, Ian Foster, and Adriana Iamnitchi. "Mapping the Gnutella network: Properties of large-scale peer-to-peer systems and implications for system design". *IEEE Internet Computing* 6.1 (2002), pp. 50–57 (cit. on p. 34).

[Rodriguez-Iturbe, 2001] Ignacio Rodriguez-Iturbe and Andrea Rinaldo. *Fractal River Basins: Chance and Self-Organization*. Cambridge University Press, New York, 2001 (cit. on p. 28).

[Rosenow, 2008] Bernd Rosenow. "Determining the optimal dimensionality of multivariate volatility models with tools from random matrix theory". *Journal of Economic Dynamics & Control* 32.1 (2008), pp. 279–302 (cit. on pp. 59, 60, 137).

[Ruelle, 1979] David Ruelle. "Ergodic theory of differentiable dynamical systems". *Publications Mathématiques de L'Institut des Hautes Scientifiques* 50 (1979), pp. 27–58 (cit. on pp. 91, 149).

[Santhanam, 2001] M. S. Santhanam and Prabir K. Patra. "Statistics of atmospheric correlations". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 64.1 Pt 2 (2001), p. 016102 (cit. on pp. 10, 56).

[Sauer, 1991] Tim Sauer, James A. Yorke, and Martin Casdagli. "Embedology". *Journal of Statistical Physics* 65.3-4 (1991), pp. 579–616 (cit. on p. 91).

[Schäfer, 2018] Benjamin Schäfer, Dirk Witthaut, Marc Timme, and Vito Latora. "Dynamically induced cascading failures in power grids". *Nature Communications* 9 (2018), p. 1975 (cit. on pp. 73, 74).

[Schmidt-Nielsen, 1984] Knut Schmidt-Nielsen. *Scaling: Why is Animal Size so Important?* 1st. Cambridge University Press, 1984 (cit. on p. 28).

[Šeba, 2003] Petr Šeba. "Random matrix analysis of human EEG data". *Physical Review Letters* 91.19 (2003), p. 198104 (cit. on pp. 10, 56).

[Sengupta, 1999] Anirvan M. Sengupta and Partha P. Mitra. "Distributions of singular values for some random matrices". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 60.3 (1999), pp. 3389–3392 (cit. on p. 57).

[Sienkiewicz, 2013] A. Sienkiewicz, Tomasz Gubiec, Ryszard Kutner, and Zbigniew Struzik. "Dynamic structural and topological phase transitions on the Warsaw Stock Exchange: A phenomenological approach". *Acta Physica Polonica Series A* 123.3 (2013), pp. 615–620 (cit. on pp. 35, 45, 49, 54, 127, 131, 133, 135).

[Song, 2009] Dong-Ming Song, Zhi-Qiang Jiang, and Wei-Xing Zhou. "Statistical properties of world investment networks". *Physica A: Statistical Mechanics and its Applications* 388.12 (2009), pp. 2450–2460 (cit. on p. 28).

[Song, 2011] Won-Min Song, Tiziana Di Matteo, and Tomaso Aste. "Nested hierarchies in planar graphs". *Discrete Applied Mathematics* 159.17 (2011), pp. 2135–2146 (cit. on p. 18).

[Song, 2012] Won-Min Song, Tiziana Di Matteo, and Tomaso Aste. "Hierarchical information clustering by means of topologically embedded graphs". *PLoS ONE* 7.3 (2012), e31929 (cit. on p. 18).

[Soramäki, 2007] Kimmo Soramäki, Morten L. Bech, Jeffrey Arnold, Robert J. Glass, and Walter E. Beyeler. "The topology of interbank payment flows". *Physica A: Statistical Mechanics and its Applications* 379.1 (2007), pp. 317–333 (cit. on p. 10).

[Stein, 1969] Charles M. Stein. *Multivariate analysis I.* Department of Statistics, Stanford University, Stanford, 1969 (cit. on p. 57).

[Sun, 2012] Longxiao Sun, Shudong Wang, Kaikai Li, and Dazhi Meng. "Analysis of cascading failure in gene networks". *Frontiers in Genetics* 3 (2012), p. 292 (cit. on p. 73).

[Takens, 1981] Floris Takens. "Detecting strange attractors in turbulence". *Dynamical Systems and Turbulence, Warwick 1980, Lecture Notes in Mathematics.* Ed. by D. Rand and L.-S. Young. Vol. 898. Springer, Berlin, Heidelberg, 1981, pp. 366–381 (cit. on p. 91).

[Topaz, 2015] Chad M. Topaz, Lori Ziegelmeier, and Tom Halverson. "Topological data analysis of biological aggregation models". *PloS ONE* 10.5 (2015), e0126383 (cit. on p. 103).

[Tumminello, 2005] Michele Tumminello, Tomaso Aste, Tiziana Di Matteo, and Rosario N. Mantegna. "A tool for filtering information in complex systems". *Proceedings of the National Academy of Sciences of the United States of America.* Vol. 102. 30. 2005, pp. 10421–10426 (cit. on p. 18).

[Tumminello, 2007] Michele Tumminello, Claudia Coronnello, Fabrizio Lillo, Salvatore Micciché, and Rosario N. Mantegna. "Spanning trees and bootstrap reliability estimation in correlation-based networks". *International Journal of Bifurcation and Chaos* 17.7 (2007), pp. 2319–2329 (cit. on p. 18).

[Umeda, 2017] Yuhei Umeda. "Time series classification via topological data analysis". *Transactions of the Japanese Society for Artificial Intelligence* 32.3 (2017), D-G72_1–12 (cit. on p. 103).

[Umeda, 2019] Yuhei Umeda, Junji Kaneko, and Hideyuki Kikuchi. "Topological data analysis and its application to time-series data analysis". *Fujitsu Scientific and Technical Journal* 55.2 (2019), pp. 65–71 (cit. on p. 103).

[Utsugi, 2004] Akihiko Utsugi, Kazusumi Ino, and Masaki Oshikawa. "Random matrix theory analysis of cross correlations in financial markets". *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics* 70.2 (2004), p. 026110 (cit. on p. 66).

[Vandewalle, 2001] Nicolas Vandewalle, F. Brisbois, and X. Tordoir. "Non-random topology of stock markets". *Quantitative Finance* 1.3 (2001), pp. 372–374 (cit. on p. 35).

[Watts, 1998] Duncan J. Watts and Steven H. Strogatz. "Collective dynamics of 'small word' networks". *Nature* 393 (1998), pp. 440–442 (cit. on p. 34).

[West, 2001] Douglas B. West. *Introduction to Graph Theory.* 2nd. Prentice-Hall, Englewood Cliffs, NJ, 2001 (cit. on p. 17).

[West, 1997] Geoffrey B. West, James H. Brown, and Brian J. Enquist. "A general model for the origin of allometric scaling laws in biology". *Science* 276.5309 (1997), pp. 122–126 (cit. on p. 28).

[Whitesides, 1999] George M. Whitesides and Rustem F. Ismagilov. "Complexity in chemistry". *Science* 284.5411 (1999), pp. 89–92 (cit. on p. 6).

[Wigner, 1951] Eugene P. Wigner. "On a class of analytic functions from the quantum theory of collisions". *Annals of Mathematics* 53.1 (1951), pp. 36–67 (cit. on p. 56).

[Wiliński, 2013] Mateusz Wiliński, A. Sienkiewicz, Tomasz Gubiec, Ryszard Kutner, and Zbigniew R. Struzik. "Structural and topological phase transitions on the German Stock Exchange". *Physica A: Statistical Mechanics and its Applications* 392.23 (2013), pp. 5963–5973 (cit. on pp. 35, 45, 46, 49, 54, 127, 131, 133).

[Williams, 2010] Mark T. Williams. *Uncontrolled Risk: Lessons of Lehman Brothers and How Systemic Risk Can Still Bring Down the World Financial System*. 1st. McGraw-Hill Education, 2010 (cit. on p. 12).

[Wishart, 1928] John Wishart. "The generalised product moment distribution in samples from a normal multivariate population". *Biometrika* 20A.1-2 (1928), pp. 32–52 (cit. on pp. 57, 136).

[Yaskov, 2016] Pavel Yaskov. "A short proof of the Marchenko–Pastur theorem". *Comptes Rendus Mathematique* 354.3 (2016), pp. 319–322 (cit. on p. 57).

[Zheng, 2012] Zeyu Zheng, Boris Podobnik, Ling Feng, and Baowen Li. "Changes in cross-correlations as an indicator for systemic risk". *Scientific Reports* 2 (2012), p. 888 (cit. on pp. 14, 60, 137).

RÉSUMÉ

Dans cette thèse, nous étudions les comportements collectifs des marchés boursiers, les primaires où se concentrent la plupart des ressources financières. Donné un bourse, pour comprendre ses caractéristiques de comportement collectif et mécanisme, nous analysons de manière approfondie le marché dans de nombreux aspects, y compris sa structure de réseau, sa résistance aux défaillances de composants, son facteur de marché déterminant principalement les rendements des avoirs sous-jacents, l'évolution de la défaillance en cascade et la dynamique de son indice représentatif. Étant donné que les marchés financiers peuvent être considérés comme des systèmes complexes, nous utilisons différentes techniques issues de la science complexe pour étudier les marchés boursiers dans de tels aspects, notamment la science des réseaux, la théorie des matrices aléatoires, la théorie de la prétopologie et l'analyse topologique des données.

MOTS CLÉS

marchés boursiers, réseaux complexes, théorie des matrices aléatoires, théorie de la prétopologie, analyse topologique des données

ABSTRACT

In this thesis, we study the collective behaviors of stock markets, the primary ones where most of financial resources concentrate. Given a stock market, to understand its collective behavior's characteristics and mechanism, we comprehensively analyze the market in many aspects, including its network structure, its resilience under component fails, its market factor primarily driving the returns of the underlying holdings, the cascading failure's evolution, and its representative index's dynamics. Because financial markets can be considered as complex systems, we use different techniques employed from complex science to investigate stock markets in such aspects, including network analysis, random matrix theory, pretopology theory, and topological data analysis.

KEYWORDS

stock markets, complex networks, random matrix theory, pretopology theory, topological data analysis