EXAMINATION OF INSTITUTIONAL INVESTOR NETWORK PATTERNS IN CONTEXT OF MAJOR CRASHES IN US STOCK MARKETS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ERSIN DEMİREL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

JANUARY 2024

EXAMINATION OF INSTITUTIONAL INVESTOR NETWORK PATTERNS IN CONTEXT OF MAJOR CRASHES IN US STOCK MARKETS

submitted by **ERSIN DEMİREL** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems Department, Middle East Technical University** by,

Date: 18.01.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ersin DEMİREL

Signature :

ABSTRACT

EXAMINATION OF INSTITUTIONAL INVESTOR NETWORK PATTERNS IN CONTEXT OF MAJOR CRASHES IN US STOCK MARKETS

DEMİREL, Ersin

M.S., Department of Information Systems

Supervisor: Prof. Dr. Banu GÜNEL KILIÇ

January 2024, 82 pages

In the dynamic landscape of the stock market, accredited investors in the U.S. are required to report quarterly holdings in their portfolios to the Securities and Exchange Commission (SEC) using Form 13F. Although these filings disclose holdings of each investor, it is challenging to extract valuable insights because of the complex structure of the market and the natural tendency of investors to keep their strategies discreet.

This thesis presents an innovative approach to process and enrich historical 13F filings into a dynamic bipartite graph of stocks and investors with rich edge and node attributes in order to shed light into investor behaviors in context of major crashes in the U.S. stock markets. The analysis revealed that the clustering coefficient, significant sale and buy counts, late submitted filings and specific motif counts of the generated graph exhibit substantial changes during market crashes. These findings highlight the potential to use 13F filings and network methods for deeper understanding of market dynamics during downturns, and to be used as useful features for market monitoring systems.

Keywords: Financial Analysis, Stock Markets, Stock Market Crashes, Network Analysis, Network Motif Discovery

ÖZ

ABD HİSSE SENEDİ PİYASALARINDAKİ BÜYÜK ÇÖKÜŞLER BAĞLAMINDA KURUMSAL YATIRIMCI AĞI ÖRÜNTÜLERİNİN İNCELENMESİ

DEMİREL, Ersin

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Prof. Dr. Banu GÜNEL KILIÇ

Ocak 2024, 82 sayfa

Hisse senedi piyasasının dinamik ortamında akredite yatırımcılar, portföylerindeki varlıkları her üç ayda bir Form 13F raporu ile Menkul Kıymetler ve Borsa Komisyonu (SEC) kurumuna bildirmektedir. Her ne kadar 13F raporları yatırımcıların borsadaki varlıklarını açıklasa da, piyasanın karmaşık yapısı ve yatırımcıların stratejilerini gizli tutma yönündeki doğal eğilimleri nedeniyle bu raporlardan değerli içgörüler elde etmek zordur.

Bu tez, ABD hisse senedi piyasalarındaki büyük çöküşler bağlamında yatırımcı davranışlarına ışık tutmak amacıyla, tarihsel 13F belgelerini, zengin uç ve düğüm özelliklerine sahip hisse senetleri ve yatırımcılardan oluşan dinamik iki taraflı bir ağ halinde işlemek ve zenginleştirmek için yenilikçi bir yaklaşım sunmaktadır. Analiz, oluşturulan ağın kümelenme katsayısı, önemli satış ve satın alma sayıları, geç teslim edilen 13F raporları ve spesifik motif sayılarının piyasa çöküşleri sırasında belirgin değişimler gösterdiğini ortaya çıkarmıştır. Bu bulgular, kriz sırasında piyasa dinamiklerinin daha derinlemesine anlaşılması ve piyasa izleme sistemlerinde kullanışlı öznitelikler olarak 13F raporlarının ve ağ analizi yöntemlerinin kullanılma potansiyelini vurgulamaktadır.

Anahtar Kelimeler: Finansal Analiz, Hisse Senedi Piyasaları, Hisse Senedi Piyasa Çöküşleri, Ağ Analizi, Ağ Motif Tespiti

I dedicate this work to my family, who has always supported and encouraged me in my decisions. Freely exploring, learning, and specializing in my career and the subjects I was truly interested in was not something I could do alone.

I hope I made all of you proud.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my social network analysis teacher and advisor, Banu Günel Kılıç. Her guidance has not only made my work possible but also provided me with academic research skills, knowledge, and perspectives that I couldn't find anywhere else.

I am also grateful for the METU Informatics Institute, a supportive and highly research-driven environment where I consistently find answers to all of my questions. As someone from the design background and deeply impressed by the power of network science, I genuinely believe that METU Informatics Institute provides a multidisciplinary ecosystem. I believe that this environment catalyzes exponential growth in innovative ideas that will to shape the future by simply enabling collaboration of curious minds from various domains to create something bigger from themselves.

Thanks for everything.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ v

DEDICATION vi

ACKNOWLEDGMENTS vii

TABLE OF CONTENTS viii

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xv

CHAPTERS

1 INTRODUCTION 1

1.1 Objectives in the Study 1

1.2 Contributions of the Study 2

1.3 Organization of the Thesis 3

1.4 Scope and Limitations 4

2 BACKGROUND AND RELATED WORK 5

2.1 Foundations of the Finance and Capital Markets 5

2.1.1 Brief Introduction into Finance and Economics 6

2.1.2 Capital Markets, Regulations and Disclosures 7
2.1.3 Investment and Its Relevance with Disclosure Monitoring 9
2.1.4 13F Filings, Institutional Investors and Smart Money 11
2.1.5 Bubbles, Bursts and Market Crashes 14
2.2 Data Science in Finance 16
2.2.1 Data Science in Capital Markets 16
2.2.2 Interconnectivity, Networks and Network Analysis Methods 18
2.2.3 Network Analysis in Capital Markets 20
2.2.4 Graph Analysis for Stock Bubbles and Crashes 20
3 METHODOLOGY 23
3.1 Methodology Overview 23
3.2 Tools and Implementation Details 24
3.3 Data Processing for Market Analysis and Bipartite Network Construction 25
3.4 Dynamic Network Analysis and Pattern Discovery in Stock-Investor Relationships . . . 27
4 EXPERIMENTS AND RESULTS 31
4.1 Data Processing, Analysis and Network Construction 31
4.1.1 Description of the Data 31
4.1.2 Data Transformations and Feature Extraction 34
4.1.3 Construction of Bipartite Network 41
4.1.4 Examination of Bubbles and Crashes 44
4.2 Characteristics of the Dynamic Network and its Examination 47
4.2.1 Investor Filing Delays Analysis 48
4.2.2 Significant Portfolio Changes of Investors 51

4.2.3 Network Metrics 53

4.2.4 Network Motif Frequencies 56

5 DISCUSSIONS AND CONCLUSIONS 67

5.1 Discussions 67

5.2 Conclusions 68

5.3 Future Work 70

REFERENCES 73

APPENDICES

LIST OF TABLES

Table 1 Latest updated 5 forms and their descriptions from official SEC website based on Securities Exchange Act of 1934 Stature 9

Table 2 Source data is a large dataframe of 48.719.087 rows, extracted from historical 13F filings 32

Table 3 SEC API data is used to further enrich information available in 13F filings 33

Table 4 Price Data consists of OHLCV data of available stocks 33

Table 5 S&P500 Price Data also consists of OHLCV data of overall S&P500 companies 34

Table 6 Industry Sectors and Corresponding SIC Code Ranges 38

Table 7 Multigraph is main repository of processed data with all quarters combined 42

Table 8 Summary of Major Market Crashes and Their Key Sectors 48

LIST OF FIGURES

Figure 1 Financial Markets can be classified based on various factors 8

Figure 2 Form 13F Filings are structured documents of portfolio holdings of institutional investors, publicly available in EDGAR database 12

Figure 3 Historical S&P 500 data shows noticeable declines in recent history 15

Figure 4 Data science is an umbrella term encompassing various concepts 17

Figure 5 Graph is a data structure consists of nodes and edges 19

Figure 6 High level overview of the methodology 24

Figure 7 Data processing steps depicted in diagram aims first two objectives 25

Figure 8 Output of the first objective is a dynamic bipartite graph of stocks and investors with rich features 27

Figure 9 Third and fourth objectives of the study is to examine graph structure and change of frequency of motifs over time 28

Figure 10 Motif rules has high explainability and customization 29

Figure 11 While most of filing types are 13F-HR, there are also amendmends and notice reports 32

Figure 12 S&P500 prices data at the time period of utilized 13F filings 34

Figure 13 Filing delays are populated around the 45th day since its due date for submission . 35

Figure 14 Holdings per quarter are converted into changes to be displayed as interactions . . 36

Figure 15 Share changes are categorized based on their values and prevalence in data 36

Figure 16 Investors with 5-50th percentile are inactive, 50-95 are average, and more than 95th are active classified investors 37

Figure 17 Stocks with 5-50th percentile are small, 50-95 are medium, and more than 95th are large classified stocks 38

Figure 18 Sector mentions in 13F filings are distributed as follows 39

Figure 19 Volume and Volatility Category Distributions 40

Figure 20 Quarterly aggregated volume and volatility of S&P500 index 40

Figure 21 Node and edge features of the created bipartite graph of stocks and investors 41

Figure 22 Most frequently interacted 10 investors 43

Figure 23 Most frequently interacted 10 stocks 43

Figure 24 All three crisis have increased volume and volatility, and decreased price with most significant being 2008 crisis 44

Figure 25 Popular tech companies display increase in volume and volatility during DotCom bubble 45

Figure 26 Biggest oil companies show similar hikes during oil crises 46

Figure 27 Biggest oil companies show similar hikes during oil crises 47

Figure 28 Most of the filings are submitted before 45 day delay 48

Figure 29 Filing Delay Category Counts By Industry Over Time 49

Figure 30 All industries have similar delay counts since investors submit all stocks in same filings 50

Figure 31 Hikes in percentage of late filings after market crashes are visible 51

Figure 32 Significant change (Significant Buys and Sales) counts resembles S&P500 volume and volatility graph, highlighting the impact of 2008 crisis 52

Figure 33 Large Buy and Sale counts gets closer during market crashes 52

Figure 34 Hikes in count of large sales are lagged during the Oil Crisis 53

Figure 35 Slight increases in density during crashes are visible 54

Figure 36 During DotCom crisis, software industry average degree visibly dropped 55

Figure 37 During crises, especially 2008, clustering coefficient significantly increases 55

Figure 38 This motif rule focuses on occurrences from inactive to active investors 56

Figure 39 During market downturns, motif counts for corresponding industries visibly increased 57

Figure 40 This motif also filters late submitted filings on top of the first motif 57

Figure 41 Almost all motifs occurred at the late quarters of market crashes 58

Figure 42 This motif captures overlapping sales of active investors 58

Figure 43 No significant changes, although, sales are increasing before market crashes as expected 58

Figure 44 Motif 4 is an inverse version of motif 1 with a changed flow from active to inactive investors 59

Figure 45 Not visible changes, but more counts overall compared to the first motif 59

Figure 46 Motif 5 is an inverse version of motif 2 with a changed flow from active to inactive investors 59

Figure 47 Similar to motif 2, motif 5 counts also has increased in the late periods of 2008 and oil crisis crashes 60

Figure 48 Motif 6 is similar to motif 3, although it counts significant sales of inactive investors 60

Figure 49 Limited number of motif counts because investor activities are based on activity scores, which are calculated depending on sale counts and significance 61

Figure 50 This motif captures any sale and buy edges available in the graph 61

Figure 51 No significant shifts during market downturns, but there is a visible upward trend after late 2012s 61

Figure 52 Motif 8 is similar to motif 7, however, the direction is inversed in this one 62

Figure 53 Similar with the seventh motif, with more counts and volatile increase in counts after 2012s 62

Figure 54 These motifs capture counts of flows of stocks that have high volatility in corresponding quarters 63

Figure 56 These motifs capture counts of flows of stocks that have low trading volume in corresponding quarters 64

Figure 55 Motif 9 and 10 has similar trends since overall market volatility was also increased during these periods. 64

Figure 57 Although varies in counts, motif 11 and 12 has somehow inverted display during 2008 crisis 65

LIST OF ABBREVIATIONS

SEC	Securities Exchange Commision
U.S	United States
S&P500	Standard and Poor's 500
R&D	Research and Development
IPO	Initial Public Offering
EDGAR	Electronic Data Gathering, Analysis, and Retrieval System
XBRL	Extensible Business Reporting Language
Regulation FD	Regulation Fair Disclosure
CAPM	Capital Asset Pricing Model
SMB	Small Minus Big, Market Capitalization Factor
HML	High Minus Low, Value Premium
RMW	Profitability Factor
CMA	Investment Factor
EMH	Efficient Market Hypothesis
ESG	Environmental, Social and Governance
NLP	Natural Language Processing
GraphSAGE	Graph Sample and Aggregate
LSTM	Long-term Short Memory
RNN	Recurrent Neural Network
GRU	Gated Recurring Unit
VAR	Variational Auto Encoder
GAN	Generative Adversarial Network

GCN	Graph Convolutional Networks
GAT	Graph Attention Networks
TGN	Temporal Graph Networks
JODIE	Joint Dynamic User-Item Embeddings
DYREP	Dynamic Learning Representations
CIK	Central Index Key
CUSIP	Committee on Uniform Securities Identification Procedures
API	Application Programming Interface
LLM	Large Language Models
HFT	High-Frequency Trading
PERL	Practical Extraction and Reporting Language
13F-HR	13F Holding Report
13F-NT	13F Notice Report
SIC	Standard Industrial Classification
OHLCV	Open, High, Low, Close, and Volume
NaN	Not a Number
TIAA	Teachers Insurance and Annuity Association of America
IBM	International Business Machines Corp
OPEC	Organization of the Petroleum Exporting Countries
NEC	Not Elsewhere Classified
REIT	Real Estate Investment Trusts
LPPL	Log Periodic Power Law
DJIA	Dow Jones Industrial Average

CHAPTER 1

INTRODUCTION

Due to their diverse, rich, and accessible data, combined with their intrinsic connection to the social sciences, stock markets have consistently captured the attention of investors, economists, and scholars. Ensuring transparency in these markets is critical, both for fair trading and for the protection of individual investors. Over the years, this need for transparency has led to various regulatory requirements, one of the most significant being the SEC Form 13F. Accredited investors use this form to disclose their portfolio holdings in each quarter. Even though the nature of the information presented in 13F filings has a great potential to explain market behavior, extracting meaningful insights from these filings is challenging due to complex structure of market and the natural tendency of investors to keep their strategies discreet.

In this context, this thesis presents a methodology to process and enrich historical 13F filings in order to shed light into investor behaviors within the context of major market events in U.S. stock markets. The process involves the transformation of 13F filings and related stock market data into a dynamic bipartite graph, incorporating rich edge and node attributes. Subsequently, the study delves into the analysis of specific attributes and graph metrics over time, with a concentrated focus on the timelines of market crash events and related industries in U.S. stock markets. Then, network motif discovery approach is applied to examine frequencies of certain network patterns over time, to discover changes of investor behavior during market crashes. In short, study processes and examines historical 13F filings and supplementary data to progressively examine to understand investor motivations and behaviors in relation to market crashes.

The goal of this introduction chapter is to give readers a clear understanding of the motivation behind this research. The introduction chapter starts with the Objectives of the Study section, where main goal and key objectives to reach that goal will be listed. In the Contributions of the Study section, the potential benefits of this study to finance and graph data science fields are discussed. In the Organization of the Thesis section, chosen approach to convey information and the rationale behind it is briefly explained. Finally, the Scope and Limitations section highlights specific focuses within the vast domains of finance and data science, while specifying limitations stemming from the nature of the available data.

1.1 Objectives in the Study

The study utilizes network-centric approaches to better understand the characteristics of the U.S. stock markets during market crises. Knowing the fact that stock markets are highly interconnected ecosys-

tems with numerous large players with serious impact on the overall market or even globally, study examines broad market data as well as industry perspective of stocks, and institutional investor stock holdings.

This main goal can be further divided into the following list of several objectives:

- Transforming the stock holdings data of accredited institutional investors, as disclosed in 13F filings, into a dynamic bipartite network representing stock holding exchanges of investors for each quarter.
- Examination of the overall market and specific major company stocks during stock market crashes of selected industries and sectors.

These initial objectives are expected to pave the way for the following complex objectives:

- Evaluating dynamic network attributes and metrics over time, particularly during market events, to assess changes in the global structure of the bipartite network of stock-investor relations.
- Conducting a thorough analysis of the dynamic bipartite network by examining motif count frequencies each quarter focusing on relevant industries, and comparing their variations during market crashes in U.S. markets.

1.2 Contributions of the Study

Although infrequent, financial crises are significant events with global consequences, often originating from stock market downturns rather than natural disasters or wars. From this perspective, financial crises are events examined from various perspectives.

This study centers on the notion that stock markets constitute a complex and dynamic interrelations of public companies, industries and institutional investors. As a result, the primary contribution of the study to the literature lies in offering an alternative perspective on stock market crashes by examining changes in such interrelations in context of stock market crashes.

Considering the significance of events and the magnitude of the scope, the data to be processed as the ground truth consists of SEC Form 13F filings arising from government-regulated regulations, supplemented by stock market data. In this regard, the thesis also provides alternative approaches for extracting insights from SEC 13F filings.

The study introduces a network-centric approach to the realm of financial market crashes through the analysis of 13F filings. This approach offers an alternative perspective and potentially unique features to be used further in this domain. The findings also indicate the possibility of incorporating rich features for future applications, including the monitoring of financial crises, as well as, prediction or the classification of quarterly positions of investors.

In addition, the thesis provides an in-depth literature review on various related topics such as financial markets, disclosure regulations, filings, graph data science, and stock market crashes. Through this

extensive exploration, it equips the reader with the tools and thought processes to comprehend the interconnections among these subjects.

In this regard, the study aims to provide specific insights by examining investor behaviors in the context of stock market crashes through a network-centric perspective. These insights not only provide finance professionals with an alternative viewpoint on stock market crashes, but also offer investment researchers valuable information to make informed and potentially profitable decisions in the wake of such market downturns.

1.3 Organization of the Thesis

The study commences by establishing an exhaustive background in the fields of finance and data science. Given that the intended readers may have diverse backgrounds, spanning both social science and data science, the upcoming chapter initially provides foundational knowledge on finance and capital markets. It offers general information about financial markets, their fundamentals, regulations, filings, the 13F filings used in the thesis, and background information on financial market crashes. Subsequently, the focus shifts towards the data science perspective, covering the use of data science in finance, the fundamentals of network science, background on the application of network analyses in finance, and relevant examples. Finally, the chapter concludes by presenting related works on the use of graph analysis specifically in stock market crashes.

In the methodology chapter of the study, the efforts required to achieve the objectives presented in this chapter are explained. Each section of the methodology chapter consists of block diagrams explaining steps taken to address each related objective, summaries of the methods to be used in upcoming chapter, accompanied by a rationale for their selection, with references to similar studies.

The "Experiments and Results" chapter serves as the primary section where all the steps outlined in the methodology are elaborated upon in depth, accompanied by relevant figures and commentary on findings. The chapter is divided into two sections. The first section addresses the initial two objectives, centering around the processing and exploration of data. The second section is dedicated to the examination of the processed data, addressing the last two objectives of the study.

First section of "Experiments and Results" chapter begins with an explanation of the data used in the study, presenting descriptive statistics and explanations of its columns within the context of the study. Next subsections details the steps taken to process the data into meaningful features, providing insights into the characteristics of the processed data with accompanying figures illustrating distributions. This section continues to address the first objective, explaining the construction and overall characteristics of the generated graph. Next subsection visualizes volume and volatility data of selected stocks during market crashes to highlight how diverse responses exhibited by each industry during these events.

Moving forward, second section of the chapter addresses the third objective by calculating attributes of the generated graph, such as changes in counts and filing delays, as well as, network metrics like density, average degree, and clustering coefficient for each quarter. This analysis unveils temporal changes, and enables further discussion and potential explanations on how market crashes may influence investor behaviors or vice versa. The concluding subsection of the chapter employs a similar approach to visualize changes, this time focusing on selected graph motifs. Each graph motif is

explained, illustrated and calculated for each quarter to display changes of overall market behavior. Findings are supported with potential interpretations for the observed patterns.

In the Discussions and Conclusions chapter, the discussions section delves into the interpretation of these findings, exploring the utility of the applied approach, discussing its strengths and weaknesses. In the conclusions section, the study concludes its findings by comparing obtained results with the initially outlined objectives and experimental steps taken. Lastly, in the Future Works section of the chapter, potential directions for enhancing the study's utility for regulatory bodies or investment researchers are examined for further improvements in the future.

1.4 Scope and Limitations

Primary data examined is the processed aggregate of historical SEC Form 13F filings falling within the range of 1999 to 2017 years. This limits the scope of the research to this time frame and the market crashes within it, of those occurring in the United States.

In the examining of the market crashes, the study adopts a macro and network-centric perspective. Consequently, the research provides an overall view of the S&P 500, and groupings based on selected industries. As the study is not finance-centric, it solely investigates financial metrics such as price, volume and volatility, refraining from an in-depth analysis of financial metrics. Instead, the study primarily focuses on overall graph metrics, behaviors of industry groups, and changes of investor behaviors as whole. To stay within the defined scope and prevent divergence, the study minimizes emphasis on a entity-level perspective or entity features as much as possible.

Additionally, the study refrains from making predictive assertions and is conducted with an examination approach. The primary reason for this is the low time granularity, and high sparsity of the 13F filings data processed and modeled in the study.

Covering the U.S. stock market in the time span from 1999 to 2017, the examined 13F filings encompass portfolio holdings from over ten thousand unique investors and approximately fourteen thousand distinct stocks. Reflecting the inherent nature of the stock market, both investor and stock interactions follow a long-tail distribution. This implies that a minority of investors and stocks dominate the market with substantial holdings and interactions, while the majority can be regarded as entities with minimal interactions. In addition to that, despite the study's comprehensiveness, filing submissions occur on a quarterly basis, resulting a total of 74 quarters in the time series.

The characteristics of the processed data pose a challenge for the initial modeling approach with graph neural networks, and employing techniques such as graph embedding and link predictions. The nature of the data, involving 74 time frames and over 100 million potential interaction pairs between stocks and investors, leads to suboptimal performance. Consequently, these limitations have prompted the study to adopt a rule-based motif count approach in graph data science, allowing for more explainable findings and preventing potential errors or false claims that could arise from a probabilistic predictive approach.

CHAPTER 2

BACKGROUND AND RELATED WORK

The primary objective of this chapter is to expand upon the concepts mentioned in the preceding chapter, providing comprehensive details and showcasing similar studies from the literature. This approach aims to prepare readers for the subsequent discussions on methodology and experiments.

The concepts and studies presented in this chapter draw parallels from both data science and social sciences perspectives. To ensure that the reading remains on a solid foundation, the chapter is divided into two sections. In the first section, the chapter commences by presenting the foundational principles of finance from a social sciences viewpoint. In this section, capital markets, regulatory bodies, disclosures and their effects are further examined. At the final subsection, directly related studies examining 13F filings and market crashes will be evaluated. It then shifts its focus towards the data science lens in the second section, inspecting data science and computational methods. In final subsection, related studies which specifically apply network-centric tools to analyze financial markets will be examined in depth.

The rationale behind the flow of this chapter is to guide readers through a logical progression towards the main concepts central to this thesis. Even though they are out of the scope of this work, the first section touches upon topics from economics and fundamental theories related to asset pricing, while second one delves into data science methods used in different fields of finance such as portfolio optimization and price prediction because of their prevalence in finance.

In essence, by the end of the "Background and Related Works in Literature" chapter, readers are expected to become familiarized with the concepts and similar studies highlighted in this thesis from both social science and data science perspectives.

2.1 Foundations of the Finance and Capital Markets

This section delves into the foundational concepts revolving around finance and capital markets. It begins with a broad overview of the finance domain, then examines issues related to its specific challenges and nuances. Then it highlights the significance and limitations of the primary data examined in this thesis, notably the 13F filings. It continues with providing related studies in analyzing 13F filings as well as studies addressing market crashes and bursts.

2.1.1 Brief Introduction into Finance and Economics

Finance, economics, accounting along with the law are fundamental and interrelated disciplines that shape the dynamics of our society in global scale [1]. In broad sense, economics, or macro-economics to be more exact, delves into how societies allocate scarce resources among competing uses, examining individual and collective decision-making processes that influence supply, demand, and wealth distribution. Finance, specifically addresses the ways in which individuals, institutions, and markets raise, allocate, and invest capital by determining the value of assets and liabilities. Accounting on the other hand, complements these areas by providing the structured means to record, report, and analyze financial data, ensuring that businesses operate transparently.

Finance, in its essence, concerns the art and science of money management [2]. Its segmented in three primary areas based on its target entity: personal, public, and corporate finance. Personal finance revolves around the financial undertakings of individuals, spanning activities like planning, saving, and borrowing. Public finance, conversely, deals with fiscal strategies, including taxation, budgeting, and debt management, governing how public services are financed. Corporate finance, which is the central focus of this thesis, delves into the financial operations inherent to corporations.

Before evaluating the role and purpose of a corporation, it is essential to recognize the Friedman doctrine, which posits that the primary social responsibility of a business is to increase its profits [3]. There are various paths for businesses to increase profits, from cost reduction and turnover increase to enhancing efficiency and productivity. Naturally, these efforts require certain capabilities, workforce, and material resources. This brings us to the act of 'financing,' the verb form of finance, representing the action of providing funds for an investment [2].

Financing can be viewed as a strategic tool, leveraging the time value of money to utilize anticipated future cash flows for current projects. The principle can be expressed through the formula:

$$FV = PV \times \frac{(1+r)^n - 1}{r}$$

This equation represents the future value (FV) of a present value (PV), factoring in a given rate of return (r) over a number of equal payments at regular intervals (n).

In a financial ecosystem, while some entities possess a surplus of capital, others might be in need of it. This discrepancy forms the foundation for the 'market for money', where entities with surplus seek investment opportunities and those in need of financing search to secure funds. At the heart of this exchange lies the decision between two primary methods of financing: debt and equity. Debt financing revolves around borrowing funds with the obligation of repayment with interest, while equity financing involves exchanging ownership stakes of the company for capital.

Corporations, guided by their primary responsibility of profit maximization [3], often face the pivotal decision of choosing between debt and equity financing to optimize their operations. A cornerstone in this discourse is the Modigliani-Miller theorem, which suggests that, in a world free from market imperfections, how a firm chooses to finance its operations would not matter [4]. This fundamental theorem sets a benchmark to evaluate external factors that constantly disrupt this ideal scenario.

Such disruption of an idealistic scenario is exemplified by Friedman, with an observation that a rapid rate of monetary growth can stimulate spending and influence investment decisions based on market interest rates [5]. Similarly, Minsky stated that financial stances which he categorized as hedge, speculative and ponzi profiles, shifts during economic expansions, highlights cyclical nature of financial fragility in capitalist systems [6].

The discussions on corporate financing choices, particularly the decision between debt and equity, lead naturally into a broader examination of the platforms where these financing options are occurring. Mechanisms that enable companies to raise capital with either equity or debt are generally called capital markets. Capital markets are the subset of financial markets, which facilitates the trade of equities of companies (stocks), debts of companies (bonds) and other financial instruments derived from them (derivatives).

2.1.2 Capital Markets, Regulations and Disclosures

Specifically, capital markets fundamentally arise from the financing needs of companies, whether through debt or equity. Present in nearly all developed nations, these markets are rigorously regulated to serve their primary purpose. Yet, capital markets are not merely marketplaces for corporate financing; they are intrinsically linked to a country's economic prosperity [7, 8, 9, 10]. Consequently, experts from both finance and economics have extensively studied these markets. Wurgler describes that financial markets and associated institutions improve allocation of capital by channelling investments to growing industries as a mechanism to improve real economy [8]. Greenwood and Smith suggests capital markets enhance growth due to efficient capital allocation and fostering specialization [11]. Similarly, Carlin and Mayer reveal that countries with advanced financial markets exhibit greater R&D investment in sectors reliant on external finance, which cultivates growth [12]. Beck further explored how banks and stock markets stimulates growth, while stating the crucial importance of legal frameworks to ensure transparency and protection of investors to cultivate that growth [13].

Capital markets essentially serve as platforms enabling the general public to invest in corporations, therefore this public nature lead these markets to be stringently regulated worldwide [1, 14, 15, 16, 17]. Companies aiming to raise capital are mandated to disclose their financial and other related information to the public. For financing, they can then either offer company shares through an equity initial public offering (IPO) or issue debt instruments via a debt IPO. The initial acquisition of financing from the public in these markets is termed as the primary capital market. Subsequent trading of these equities or debts among investors takes place in the secondary capital market. Common types of financial markets are depicted in Figure 1. Seamless execution of this trade depends on clear and transparent exhibition of company information, and application of effective measures to prevent fraud. Therefore its crucial to ensure fair and transparent trading both among investors, and between companies and investors.

Regulatory oversight, although varying by country, has been identified to bring significant benefits to capital markets, primarily in terms of liquidity and the protection of investor rights [15]. Hence, regulations and their implications often stand at the center of many capital market studies. A pioneering regulatory body in this realm is the Securities and Exchange Commission (SEC) of the United States, which was instituted in 1934 following the stock market crash of 1929 [16]. Besides of other regulations, SEC mandates bodies in capital markets to disclose certain forms to the public, thus, its critical to understand studies and discussions revolving around these disclosures and their effect on the capital markets.

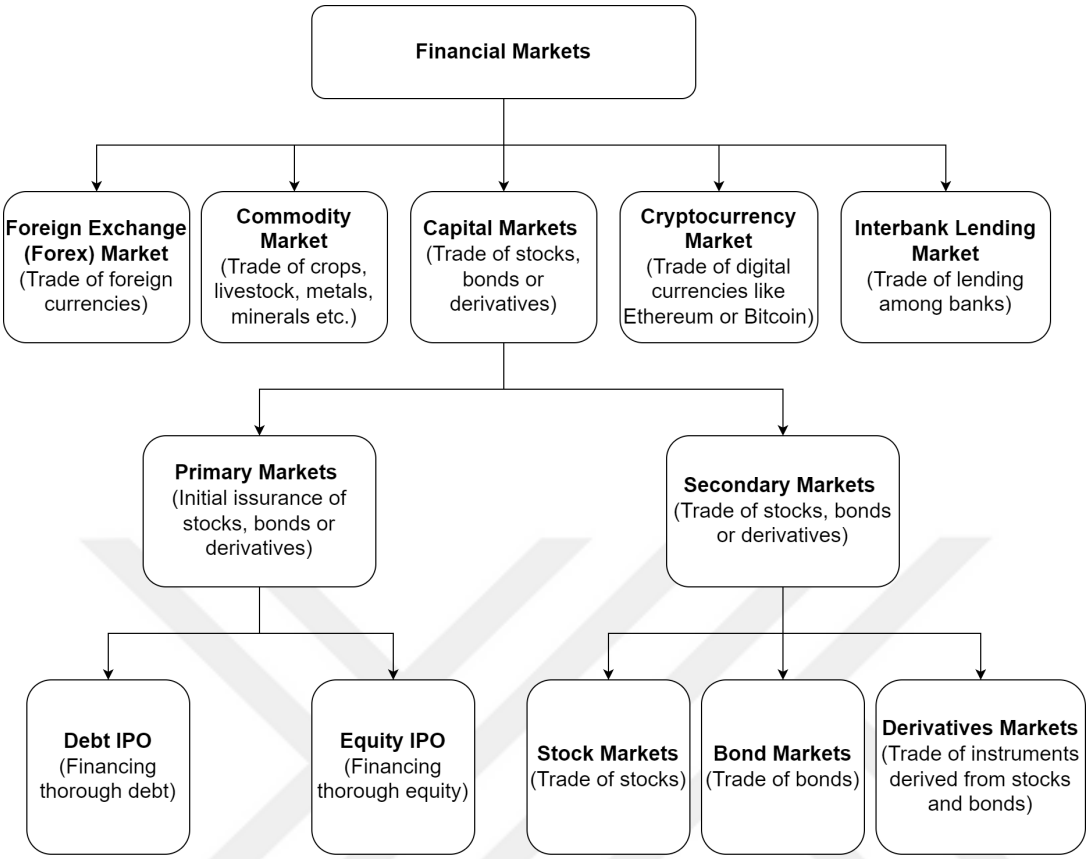


Figure 1: Financial Markets can be classified based on various factors

Among the various mandates in capital markets, public disclosures stand as the most critical component of financial market regulations [18]. The U.S. Securities and Exchange Commission (SEC) mandates the dissemination of data ranging from significant company events to financial specifics [16]. There are over 200 types of filings, some of which were recently updated are depicted in Table 1. While many capital markets globally adopt similar practices, the U.S. stands out due to its well-established systems, and pioneering initiatives such as Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, XBRL reporting standard and Regulation Fair Disclosure.

In 1992, the SEC initiated the EDGAR system, transitioning company filings from paper to digital format, with mandatory electronic submissions by 1995. This system has been notably accessed by retail investors [19]. The structured data has also fostered innovative applications to improve investment strategies among analysts. For instance, Kravet et al. observed a correlation between heightened risk disclosures in 10-K filings and subsequent stock volatility and volume [20]. Additionally, Li et al. analyzed EDGAR's log files to discern investor attention and employed post-earnings announcement drifts to achieve superior alpha returns [21].

In August 2000, the SEC implemented Regulation Fair Disclosure (Regulation FD), which mandates corporations to broadly disclose material nonpublic information instead of selective releases. It was established in response to situations where issuers provided advance warnings and other nonpublic information selectively to specific institutional investors and analysts. While Regulation FD reduced

Table 1: Latest updated 5 forms and their descriptions from official SEC website based on Securities Exchange Act of 1934 Stature

Number	Description	Last Updated	SEC Number	Topic(s)
13F	Information required of institutional investment managers pursuant to Section 13(f)	July 2023	SEC1685	Securities Exchange Act of 1934
10-Q	General form for quarterly reports under Section 13 or 15(d)	Sept. 2021	SEC1296	Securities Exchange Act of 1934, Public Companies
SD	Specialized Disclosure Report (PDF)	Sept. 2021	SEC697	Securities Exchange Act of 1934, Public Companies
20-F	Registration statement / Annual report / Transition report (PDF)	Sept. 2021	SEC1852	Securities Exchange Act of 1934, Public Companies, International
10-K	Annual report pursuant to Section 13 or 15(d) (PDF)	May 2021	SEC1673	Form 10-K, Securities Exchange Act of 1934, Public Companies

information asymmetry, enhanced market liquidity, and change market behavior particularly around earnings announcements [22], it also inadvertently increased the capital costs for smaller firms and lessened interest in them [23].

Similarly in 2009, SEC mandated XBRL, which is another initiative that aims to provide a standardized international data format for financial reporting. Kim et al. suggest that XBRL enhances capital market operations by decreasing analysis costs and time, boosting data usability, and minimizing information risks due to increased transparency [24]. Additionally, Vasarhelyi et al. emphasized the improved usefulness of financial reports by XBRL format, and also proposed future improvements such as integrating XBRL with Semantic Web using an ontology languages to combine additional information scattered around the web, and to enable machines to consume this information [25].

All these findings highlight the outcomes of the SEC's initiatives to enhance capital market regulations and the innovative approaches they have spurred in the literature. The following subsection will focus on the utilization of the structured data generated by these initiatives for more informed investment decisions.

2.1.3 Investment and Its Relevance with Disclosure Monitoring

Previous sections examined the capital markets from the companies perspectives, however, the other important actor in this ecosystem are the investors. Investors are individuals or entities aiming to

increase their wealth by investments within the complex interplay of uncertainty, risk and time. To understand the act of investment, Merton decides to focus on individual household perspective since any economic organization already encapsulate such behaviors inherently. Merton's analysis segments the investor's decision-making into two: firstly, determining the portion of wealth set aside for future consumption, and secondly, strategizing the allocation of these earmarked savings across diverse investment opportunities [26]. Markowitz inspected the second segment Merton described, as portfolio selection, which is the distribution of the savings among investment opportunities, and he further divided the portfolio selection process into two stages [27]. Initial stage involves forming beliefs based on observation and experience about future performance of securities, following the second stage, driven by these beliefs to decide the choice of a portfolio. In his work, Markowitz focuses on the latter stage, emphasizing the construction of a diversified portfolio by weighing both expected returns and associated risks, and suggesting diversifying portfolio by selecting assets with low covariance and various industries and characteristics [27].

Markowitz's foundation of the Modern Portfolio Theory has paved the way for numerous foundational hypotheses and studies in finance, including capital asset pricing model (CAPM) to determine expected return of an asset to decide inclusion of the asset to a portfolio. The CAPM formula calculates the expected return R_i of an asset based on the risk-free rate R_f, the asset's sensitivity to market returns β_i, and the expected market return R_m. Jack Treynor, one of the early pioneers of the CAPM, viewed it as a way to quantify the relationship between the risk of a security (measured by its beta) and its expected return, emphasizing that investors should be compensated for both the time value of money (risk-free rate) and the risk they take (market risk premium) [28].

$$R_i = R_f + \beta_i(R_m - R_f)$$

This initial model later further improved by Eugene Fama and Kenneth French, by also including size of the company (market capitalization-SMB) and growth potential (book-to-market ratio-HML) for another model called Fama-French three-factor model. Which explains 90% of the return instead of 70% of CAPM model [29]. The 3-factor Fama-French model calculates the expected return R_i of an asset based on the market risk premium $R_m - R_f$, the size premium represented by SMB, and the value premium represented by HML, along with the asset's specific coefficients β_m, β_s, and β_v.

$$R_i = \alpha + \beta_m(R_m - R_f) + \beta_s \times SMB + \beta_v \times HML + \epsilon$$

Then, most lately Fama and French further extended the model by adding two additional factors such as operating profitability and investment aggressiveness of the company [30]. The 5-factor Fama-French model predicts the expected return R_i of an asset based on the market risk premium $R_m - R_f$, the size premium SMB, the value premium HML, the profitability factor RMW, and the investment factor CMA, along with the asset-specific coefficients $\beta_m, \beta_s, \beta_v, \beta_r$, and β_c. These studies are some of the most fundamental frameworks to model expected return of investments to determine highest potential portfolio given certain risk preference and macro factors.

$$R_i = \alpha + \beta_m(R_m - R_f) + \beta_s \times SMB + \beta_v \times HML + \beta_r \times RMW + \beta_c \times CMA + \epsilon$$

Similarly, the Efficient Market Hypothesis (EMH) by Eugene F. Fama is also an another foundational hypothesis, which suggests that stock prices instantaneously incorporate all available information, making securities fairly priced and hard to consistently achieve unusual returns from public information [31]. While this is considered as one of the most foundational concepts of investment, it is criticized by many researchers for inadequately addressing market anomalies and irrational behaviors of investors. One of the most significant opposing view can be considered as the value investing, which states unlike EMH states, that there are market discrepancies in capital markets that can be leveraged to yield significant gains.

Graham and Dodd, pioneers of value investing, define investment as an operation that, through thorough analysis, promises safety and an adequate return. They claim that operations that do not meet these requirements are considered speculative [32]. This approach is fundamentally opposite of Fama's claim on market efficiency, and suggests it is possible to detect undervalued or overvalued stocks. In "The Intelligent Investor," Graham emphasizes the importance of recognizing market inefficiencies, such as bubbles or crises, and advocates for the fundamental analysis of companies to ascertain their true value [33]. Graham and Dodd introduced these concepts and terminologies in the wake of the 1930s Wall Street stock crash, but one of the most recognized investment guru Warren Buffett also has often cited basing his strategy on these same fundamental principles [34].

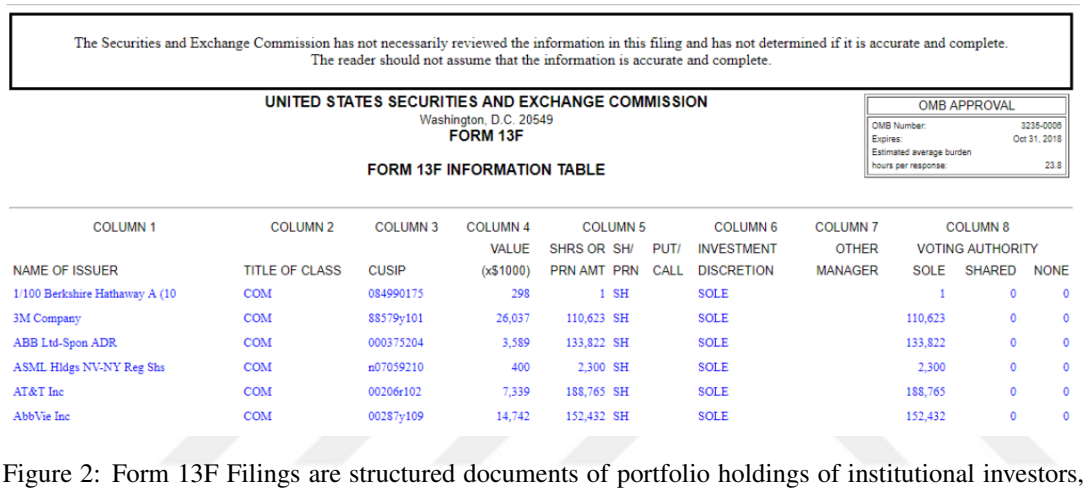
Investment aims capitalizing price discrepancies, by determining a company's intrinsic value by examining company ratios, competitive advantages, market and other core aspects of the business itself. In this context, resources such as EDGAR and its disclosures become foundational sources of information and data for this investment approach. The EDGAR database provides detailed annual investor reports through its 10-K filings, updates on quarterly developments via the 10-Q, and immediate event disclosures through the 8-K. These filings furnish investors with guidance and offer insights upon which they can base their investment decisions as suggested in previous studies and discussions.

2.1.4 13F Filings, Institutional Investors and Smart Money

While most of the filing types revolves around company information, events, financial information or board activities, 13F filings stands out by their focal entity: institutional investors. In 1975, SEC mandated 13F filings to institutional investors in the U.S. capital markets who are accredited investors with over $100 million in their portfolio [35]. These filings provide information about the institutional investor's holdings such as equity names, keys, types as well as amount of shares and voting power, displayed in Figure 2. In recent years, role of the institutional ownership grows significantly, since equities managed by institutional investors in U.S. have risen from about 8% in 1950 to 67% in 2010 [36]. Given Buffett's article mentioning numerous successful active investors capable of outperforming the market [34], contrary to the implications of the efficient market hypothesis [31], one might deduce that 13F filings present significant insights to identify investment opportunities to beat the market.

The transparency offered by 13F filings into the portfolios and activities of institutional investors has sparked numerous academic studies and discussions. Gruber first drew attention to the 'smart money' effect, to inquire the reasons why actively managed funds have grown so fast, suggesting that certain investors possess superior abilities to identify high-performing fund managers [37]. This finding popularized the term "smart money", which often refers as the capital that large institutional investors and financial professionals manage. Later studies expanded on Gruber's findings, examining whether mutual fund investors can accurately predict which funds will perform better beforehand by inspecting

the cash flow information for small funds [38, 39]. Zheng also investigated the causal relation between equity prices and aggregated trading activity presented in 13F filings, and suggested institutions buy more popular stocks after market rises [40]. Zheng also stated that, even after conditioning Fama-French factors, there is a visible relation between returns and institutional holdings, implying potential use of institutional holdings as an alternative asset pricing factor [38]. Although 13F filings were designed to increase transparency among influential institutional investors, soon its realized that these disclosures could be repurposed for evaluating investment managers as well [41]. Similarly, more recent works also provide evidence that mandatory portfolio disclosure enhances investors' ability to evaluate and select hedge fund managers effectively [42]. These findings highlight the growing interest on following the "smart money" by monitoring news, analyzing trade volumes, inspecting insider trading reports, hedge fund databases or 13F filings.



The Securities and Exchange Commission has not necessarily reviewed the information in this filing and has not determined if it is accurate and complete.
The reader should not assume that the information is accurate and complete.

UNITED STATES SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549
FORM 13F

FORM 13F INFORMATION TABLE

OMB APPROVAL	
OMB Number:	3235-0006
Expires:	Oct 31, 2018
Estimated average burden hours per response:	23.8

COLUMN 1	COLUMN 2	COLUMN 3	COLUMN 4	COLUMN 5			COLUMN 6	COLUMN 7	COLUMN 8		
			VALUE	SHRS OR	SH/	PUT/	INVESTMENT	OTHER	VOTING AUTHORITY		
NAME OF ISSUER	TITLE OF CLASS	CUSIP	(x$1000)	PRN AMT	PRN	CALL	DISCRETION	MANAGER	SOLE	SHARED	NONE
1/100 Berkshire Hathaway A (10	COM	084990175	298	1	SH		SOLE		1	0	0
3M Company	COM	88579y101	26,037	110,623	SH		SOLE		110,623	0	0
ABB Ltd-Spon ADR	COM	000375204	3,589	133,822	SH		SOLE		133,822	0	0
ASML Hldgs NV-NY Reg Shs	COM	n07059210	400	2,300	SH		SOLE		2,300	0	0
AT&T Inc	COM	00206r102	7,339	188,765	SH		SOLE		188,765	0	0
AbbVie Inc	COM	00287y109	14,742	152,432	SH		SOLE		152,432	0	0

Figure 2: Form 13F Filings are structured documents of portfolio holdings of institutional investors, publicly available in EDGAR database

Interest on institutional holdings and 13F filings are not limited to these examples. Several studies used 13F filings to examine investment preferences of institutional investors by running regressions of institutional ownership depicted in 13F filings [43, 44]. Additionally, several studies also leveraged 13F filings to document a strong positive relation between changes in institutional ownership and portfolio returns measured [38, 45, 46]. The impact of institutional trading on stock prices has also been a focus, with findings supporting the idea that price effects associated with institutional trades can significantly influence returns [47]. Expanding on these findings, Sias analyzed 13F data to explore institutional herd behavior, discovering a tendency among institutions to follow both their previous trades and the actions of other large investors. This behavior indicates a pattern that goes beyond simple momentum trading strategies and suggests a herd behavior among institutional investors [43]. Meanwhile, several studies also found that institutions trade affects stock prices, and stock returns granger-cause institutional trading, which implies institutional investors focus on momentum trading strategy [40, 48]. Some of the studies also similarly inferred strong positive correlations between institutional trading and stock returns at monthly or higher frequencies [49, 50].

Some of the recent work employs machine learning models and feature extraction techniques, and examines the usage of 13F filings for their predictive power and portfolio selection ability [51, 52, 53, 54]. For example, Fleiss used 13F filings to extract features to predict stock price movements, and to identify best performing funds, then, used these features and findings to build portfolios. He finds that both

approaches outperform 9.5% return of a popular benchmark S&P500 with 15% and 19.8% respectively, which suggests another evidence of 13F filings having significant predictive power for direction of price movements of stocks [52]. Similarly, Dipietro also claimed using 13F filings to construct quantitative models, may improve performance on long term equity investing [54]. Another work uses 13F filings to detect 'Fundamental Equity Hedge Funds'(FEHF), and identifies these funds top picks using two measures such as "conviction" (top 25% of a manager's portfolio) and "consensus" (stocks most commonly held by managers) to measure significant higher alpha from benchmarks despite the 47-day report lag [53]. Miori also used 13F filings as a source for trading signal to compute imbalances, and use this signal to detect asset price direction at the end of each quarter [51]. Another notable attempt to leverage 13F filing is to create a bipartite graph of stocks and investors, to classify both stocks and investors using recent Graph neural network models such as GraphSAGE [55]. Chouliaras provide another perspective in his work by comparing the relationship between the tone of annual reports and behavior of institutional investors as reflected in 13F filings, found that increasing pessimism in tone of annual reports also led institutional to reduce their holdings of that company [56]. Another study inclines the environmental, social, and governance (ESG) investing, which is increasingly more popular with growing concerns over global environmental risks, and examines the relation between sustainability footprints and risk-adjusted investment performance, and finds that 13F institutions with better sustainability footprints outperform [57].

While there is significant interest among researchers in 13F filings, indicating their potential, numerous studies have also underscored the risks associated with these filings due to inconsistencies and delays. As it was described earlier, the initial intent behind the SEC's 13F filings was to provide transparency, protect retail investor rights, and reduce information asymmetry [13]. These filings are mandated to be quarterly submitted, however, managers are granted with flexibility to delay disclosures up to 45 days to protect their own rights. Although occasional delays may be tolerated, penalties are typically levied when an institution demonstrates a pattern of repeatedly late or entirely absent reporting [58]. Institutions that repeatedly engage in delayed filings can face financial penalties and activity restrictions. A notable example is Quattro Global Capital, which was fined $100,000 and subjected to activity restrictions after failing to submit any 13F reports between 2002 and 2005 [58]. Additionally, while they are eventually required to release their holdings, they can request confidential treatment to conceal their certain holdings up to one year, or add new holdings, and make restatements in their existing 13F filings [59].

Various researchers study how fund managers use certain mechanisms to conceal or misreport information to safeguard their organization's interests [58, 60, 61, 62, 63]. According to Christoffersen, the primary motivations behind institutional investors strategically delaying 13F filings as a means to guard against potential front-running, which is a form of market manipulation by taking advantage of advance knowledge of pending orders, and to obscure their voting power from corporations during pivotal votes [58]. The other most popular motivation behind delaying 13F filings is preventing copycat trading. Since hedge fund strategies and ideas cannot be patented, public reveal of long positions by 13F filings left hedge funds vulnerable to copycat trading, which is mimicking of the trading actions of more experienced investors without the cost of their management fees [64]. These claims are further supported by another study that discovers 3% of all filings are incomplete as they include positions filed confidentially, and significant portion of these confidential filers are hedge funds [62]. Cao's research suggests that hedge funds frequently use restatements in their 13F filings, strategically misreporting to target copycat traders [61]. Similarly, another study focusing on the characteristics

of confidential treatment requests found that securities kept confidential during the original 13F filing period earn positive and significant abnormal returns over the post-filing confidential period [60].

The studies mentioned in previous paragraphs highlighted the multifaceted nature of 13F filings and the intricate relationships between the actors involved. Stemming from the fundamental financing needs of corporations as equities or debts, capital markets are one of the most complex structures in society. The public nature of these capital markets necessitates regulations and disclosures, which have been extensively examined in academic research. As evident from these studies, the filings mandated by the SEC have often been the preferred source for researchers to elucidate various economic, social, and financial phenomena. However, to extract meaningful insights from these filings, it is critical to understand the concerns of the institutions that create them, namely the institutional investors, and the context in which these filings exist. While many past studies have successfully derived valuable insights from 13F filings from a financial and economic perspective, numerous sources have also identified the potential for misleading information due to the concealment measures taken within these filings. These findings provide foundational knowledge about the nature of 13F filings, which will be explored in depth in subsequent chapters of the thesis, aiming to make readers aware of both the opportunities and limitations inherent in these filings.

2.1.5 Bubbles, Bursts and Market Crashes

Capital markets serve as intermediaries, facilitating the exchange of company equities, such as stocks, and debts, like bonds, among corporations, financial institutions, and retail investors, and are protected by strict regulations and government bodies as discussed earlier. However, this complex system does not always function flawlessly because significant price fluctuations can trigger panic selling, affecting the entire ecosystem to cause a phenomenon commonly referred to as a stock market crash.

A stock market crash, often characterized as a market anomaly or an outlier situation [65], typically follows periods of excessive expansion also named as bubble, speculative activities, or external economic factors, resulting in a dramatic and sudden drop in stock prices and subsequently, a market collapse. According to Johansen, crashes may have either internal (endogenous) or external (exogenous) origins [66, 67]. Endogenous origins could stem from factors like speculation, or information withholding by management, while exogenous origins might arise from external shocks, global crises, disasters and wars, significantly affecting the market dynamics [66, 67].

Understanding stock market crashes has become a prevalent area of study for academics due to their significant macroeconomic impacts. A market crash can severely damage a country's economy, leading to spikes in unemployment and a possible recession [68, 69]. This phenomenon can be seen as the other side of the coin regarding studies that discuss the positive impact of the stock market on the overall economic growth [10, 7].

However, stock market crashes can also present significant opportunities for experts who manage the crisis effectively, one case is documented in a popular movie "The Big Short", where a handful of fund managers who adeptly managed the situation in 2008 crisis were able to secure substantial returns [70]. Similarly, Benjamin Graham, the pioneer of value investing, also highlighted the importance of recognizing market inefficiencies like bubbles or crises, to determine true value of equities as well as with fundamental analysis [33].

Market crashes have a history, with roots can be traced back to the 1600s in the Netherlands, where the prices of tulip bulbs inflated to extraordinary levels before collapsing, known as a period called "Tulipmania" [71]. The United States has seen numerous stock market crises in the 21st century, notably the Dotcom Bubble burst around 2000, the market downturn post-September 11 attacks in 2001, 2008 financial crisis caused by housing market collapse and regulatory failures, and the 2015 oil crisis characterized by plummeting oil prices and industry turmoil. Globally, the European sovereign debt crisis, Chinese stock market crisis, Russian market's crash in 2022 amid the Ukraine conflict, cryptocurrency volatility, and the COVID-19 pandemic's are other significant market events and crashes in the 21st century. Many of these crises, especially Dotcom bubble, 2008 crisis and COVID-19 crisis, are visible in S&P 500 index in Figure 3.

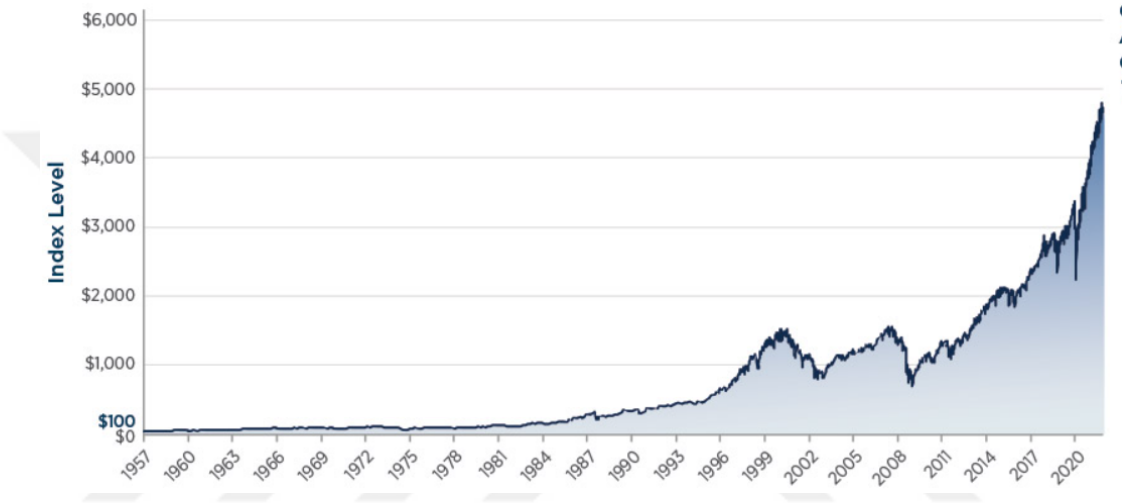


Figure 3: Historical S&P 500 data shows noticeable declines in recent history

Historically, a market crash have often been preceded by bubbles [65]. This lead numerous studies to focus on detecting bubbles and, indirectly, on predicting market crashes. Eugene F. Fama, in his interview, has expressed the need for a systematic approach to identify bubbles [1] [72]. Addressing Fama's challenge, using industry-level return data for both U.S. and other countries, Greenwood highlighted that such sharp price increases predict a substantially heightened probability of a crash [72].

Another study analyzing the S&P 500 index around the 1987 market crash found patterns characterized by dynamical critical points and log-periodic signatures similar to those observed in earthquake studies, and use these observations to suggest that the stock market functions as a self-organizing cooperative system [73]. Similar signatures also used in another work to classify crashes whether endogenous and exogenous by examining their time to revert to normal volatility levels [74]. These findings highlight previous attempts to explain market bubbles and market crashes.

Moving forward, the examples shared so far have provided a general overview of market bubbles, busts, and crashes. However, before shifting focus to the study's main aim of explaining these events through the interconnected nature of the market, it is crucial to first understand the tools and methodologies used to capture such interconnectedness.

[1] Interview with Eugene F. Fama was conducted on June 30, 2016, CBR - Economics. Source and transcript is available at: https://www.chicagobooth.edu/review/are-markets-efficient

2.2 Data Science in Finance

In the preceding section, a comprehensive examination of finance, investment, and capital markets was undertaken through a social science perspective. This investigation not only offered a panoramic view of these domains, but also highlighted the unique informational requirements and opportunities inherent to these domains.

Upon understanding the core of investment motivations and strategies, it is important to explore the computational methods and mathematical approaches. Achieving sophisticated financial analyses necessitates a synergistic approach: on one hand, understanding the data, its contextual relevance, and the broader socio-economic implications it signifies; on the other, appreciating the computational methodologies, being aware of their constraints, and grasping the foundational mathematical principles they are built upon.

Therefore, this section aims to highlight the computational and mathematical nuances of techniques, focusing mainly on machine learning and data science applications, as tools to extract insights derived from the information detailed in preceding sections. This exploration begins with a brief introduction to these techniques and terminology, and then delves into a detailed analysis of the specific techniques employed in this study.

2.2.1 Data Science in Capital Markets

In the modern age, data has evolved to become a vital strategic resource, often likened to oil due to its immense value and potential. At its essence, data consists of raw facts and figures accumulated over time. The act of examining, refining, and manipulating this data to derive meaningful insights is known as analytics. Together, the synergy between data and analytics has given rise to the interdisciplinary field of data science. This field leverages techniques from statistics, informatics, computing, communication, management, and sociology to provide actionable insights and informed decisions [75]. Figure 4 briefly summarizes the related concepts of data science.

With the proliferation of the internet, the increasing flow of data has also birthed the 'fourth science paradigm', melding together experimentation, theory, and computation, characterized by data-intensive practices [76]. This increasing accessibility of data, and paradigm shift towards data intensive research especially impacted social science studies. By tapping into vast datasets, researchers have been enabled to study human behavior with unprecedented depth and precision, rendering social studies more comprehensive and insightful than ever before [77].

This paradigm shift also visible in economics and finance, indicating numerous problems that can be addressed using data science methodologies, paving the way for novel opportunities. Significant portion of data related to capital markets is public, mainly due to the disclosure standards and regulations previously mentioned. This availability has fostered synergies between data science and capital markets. Some of the key areas often highlighted include stock market forecasting, algorithmic trading, financial sentiment analysis, portfolio allocation, asset pricing, credit risk assessment, and fraud detection [78, 79, 80, 81].

It is estimated that algorithmic trading, which is automated executions of trading in financial markets, accounted for 80% of entire transaction volume in the U.S. in 2018, up from 50% in 2011 [82]. In

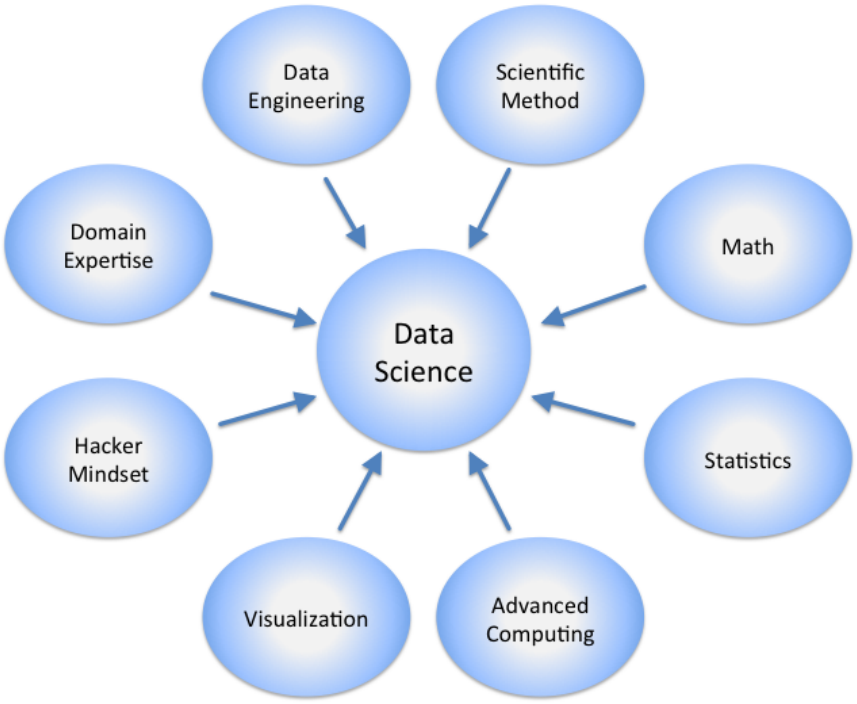


Figure 4: Data science is an umbrella term encompassing various concepts

nature, algorithmic trading provides fast and automated execution, without human biases, therefore, data science and analytics plays important role to fetch, structure, process and derive conclusions [83]. Sequential nature of time-critical data in algorithmic trading leads researchers to employ time series mining methods to predict next event, utilizing models ranging from statistical regression models to neural network based complex models such as LSTM, RNN or GRU [81, 84].

In order to further improve trading decisions, financial sentiment analysis is also used to incorporate change of public opinion in time related to the stock or asset, by capturing nuances from unstructured textual data of public social posts using natural language processing techniques [85].

Optimization of the allocation of financial assets, called portfolio optimization, is another prevalent problem addressed with data science. These set of researches incline to minimize risk and maximize return by optimizing weights mostly incorporate optimization by applying foundational mathematical models such as Markowitz model, VAR, or fuzzy approaches to neural network based techniques such as genetic algorithm, autoencoders or reinforcement learning [86, 87].

Similarly, numerous academics also studied asset pricing aspects to calculate potential returns given risk by combining linear factor based fundamental asset pricing models explained earlier to more recent deep learning based models such as generative adversarial networks (GAN) or LSTM [88, 89, 90, 91].

Banks and organizations utilize various models to classify or score clients in terms of their credit risk for their credit applications by extracting latent features from their other relevant data by utilizing deep learning and machine learning models [92, 93, 94].

Another important challenge in finance and banking domain is fraud detection, which aims to detect fraudulent activities and money transactions by analyzing large amount of historical transaction data, often leveraging both time series based and graph based models because of the interconnected nature of fraud transactions [95, 96, 97, 98, 99]. Some solutions also leverage clustering techniques to detect implicit segments of clients with similar habits, and to later analyze these groups to identify and design best potential products and services [100, 101].

Finally, another related field is actionable knowledge delivery, which aims to extract insights from various sources of financial information. This set of studies focuses on structuring various data from SEC filings, large volumes of reports and documents, and, more recently even financial charts [102, 103, 104, 105]. This enables financial research to be in natural language and improve knowledge discovery.

In the expansive field of finance, it is clear that data science techniques are employed to address a wide array of challenges. Data science itself offers a spectrum of models and methods, ranging from general statistical approaches to those that are custom-designed for specific problems. The choice of method often depends on the nature of the problem to be solved and the data at hand.

2.2.2 Interconnectivity, Networks and Network Analysis Methods

In nature, data is captured in a myriad of shapes and formats, each uniquely suited to the information it represents. While, time-series data captures variables that change over time, spatial data maps physical locations and their attributes. Similarly, certain issues, such as the dynamics of social networks or the interactions between proteins and enzymes, have an inherently interconnected nature. To effectively represent and analyze these complex relationships, graph data structures are employed.

In simple terms, graph is a way to capture and represent entities and relations between them with vertices(nodes) and edges(relationships) [106].

A graph G can be represented as a pair of sets (V, E), where:

- V represents the set of vertices or nodes, $V = \{v_1, v_2, \ldots, v_n\}$.
- E represents the set of edges or connections between these vertices, $E = \{e_1, e_2, \ldots, e_m\}$ and each edge is a pair (v_i, v_j) such that $v_i, v_j \in V$.

Graphs are efficiently represented as adjacency matrices for algebraic operations, and visually depicted through nodes and lines as displayed in Figure 5. This versatile approach is foundational across various fields, enabling the analysis, prediction, and simulation of complex, interconnected structures, from social networks to biological systems [106].

Graphs come in various types to suit different analytical needs. Undirected graphs depict bidirectional links, ideal for symmetric relationships, while directed graphs show unidirectional connections, useful for representing flow or hierarchy. Weighted graphs assign values to connections, quantifying the strength or capacity of links. Homogenous graphs contain single type of nodes and links, while heterogenous graphs may contain more than one type of nodes. Bipartite graph is a commonly used type of heterogenous graphs with two distinct set of nodes, where links occur only from one type to another.

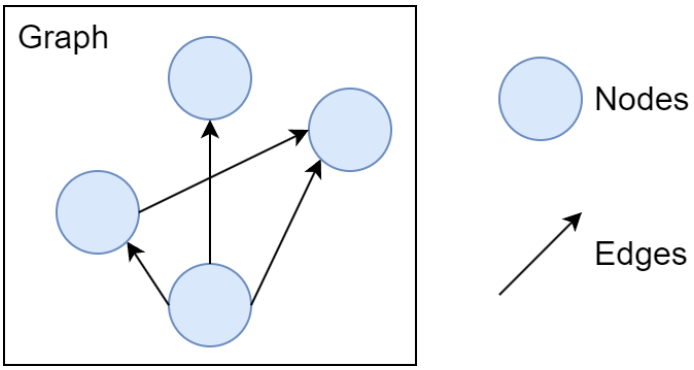


Figure 5: Graph is a data structure consists of nodes and edges

Temporal graphs capture how relationships evolve over time, adding a chronological dimension. Dynamic graphs, on the other hand, allow both nodes and edges to change over time, reflecting real-world scenarios where networks are not static [107].

Once the data is represented with either type of a graph, it opens a new realm of data science applications and techniques. For example, network topology analysis explores structural properties of networks, and enables researchers to inspect overall shape of the network, and connectivity of nodes. Community and clique detection algorithms enable identification of closely tied entity groups in a social group. Centrality measures helps to detect most influential actors in a network, while path finding algorithms enable efficiently detect shortest paths from one node to another [107]. Similarly, motif(pattern) discovery plays a crucial role to identify recurring sub-graphs to explain fundamental behavior of the network [108], while diffusion models can capture flow of information, disease or capital over time in a dynamic network [109].

Unique capabilities of graph data structures even lead them to have their own family of neural networks. Graph Neural Networks (GNNs) are emerging field in deep learning, which are neural networks tailored for graph data to efficiently encode dense node and edge representations for further tasks such as classification of nodes or predicting future links in network [110]. Evolving over time, these networks have seen advancements like Graph Convolutional Networks (GCN), which leverage graph-based convolutions [111], and Graph Attention Networks (GAT), adopting attention mechanisms proven in natural language processing [112, 113]. Innovations such as SAmple and aggreGatE (GraphSAGE) further enhance the model's accuracy by creating rich, representative embeddings by utilizing node feature information [114]. Lately, temporal GNN variants such as Temporal Graph Networks (TGN), Dynamic Learning Representations (DYREP), Joint Dynamic User-Item Embeddings (JODIE) are increasingly used to model complex, evolving relationships, offering predictions of future behaviors of entities in the network [115, 116, 117].

In summary, graphs provide a suite of analytical tools, enabling the extraction of insights in scenarios where interactions occurs among entities within a network. These tools provide applications across a broad spectrum of domains, from protein interaction analysis in biology to complex network studies in social sciences and beyond. This adaptability and depth make graph theory an indispensable resource in both academic research and practical applications, providing clarity and understanding in a world of interconnected complexities.

2.2.3 Network Analysis in Capital Markets

The financial industry deeply intertwined with a complex web of relationships among decision-makers, creates a dynamic where flow of benefits, influence and information become critical drivers of action. These decisions not only affect certain companies' paths, but also have a lasting impact on society.

This situation emphasizes the need for research and studies in finance and economics to be conducted with a keen eye on capturing these networks of interdependencies [118, 119]. Such interconnected dynamics are omnipresent, ranging from the detection of fraud in micro-transactions [96], insider trading [120] to the potential global systemic risks stemming from inter-country credit flows [121, 122, 123]. Hence, it is evident that network-based approaches are being increasingly adopted in various facets of the finance industry to address these challenges.

In studies focusing on relationships within capital markets often hone in on elements such as ownership, voting behavior, information flow and their effects on stock returns [124, 125, 126]. A notable example to use information flow to explain stock return is a research which crafted a social network based on the educational backgrounds of mutual fund managers. This study revealed that portfolio managers tend to heavily invest in stocks where they share an educational connection. Astonishingly, these "connected" stocks garnered returns up to 8.4 percent higher than their non-connected counterparts. This outperformance was attributed to the flow of information through the established network, especially when aligned with corporate news events [125]. Another research utilizing the ORBIS database, an exhaustive collection of data from 462 million companies, crafted an extensive ownership network and identified a massive central component, revealing that a limited number of companies own a significant proportion of others [127]. In the another study inspecting Chilean Stock Market, it was observed that structural changes within this network influenced the likelihood of significant variations in the volume of traded assets [126].

It is also common to use network structure to inspect portfolio overlaps and systemic risk [128, 129, 130, 131, 132]. These studies leverage graph analysis to identify systemic contagion risks, which are risks of widespread financial instability triggered by overlapping investment portfolios across various financial entities.

2.2.4 Graph Analysis for Stock Bubbles and Crashes

Stock market bubbles and crashes are not only studied by finance and economics professionals, but also by data science experts. Identifying when a specific stock is likely to reach a significant high or low can help investors make substantial profits. Various methodologies have been employed in the literature for identifying and forecasting stock market bubbles and crashes, one of the most popular being the log periodic power law (LPPL) model. LPPL is a mathematical model that fits oscillations on the historical price data of an asset, to predict the time to a likely crash occurrence, and the price level at the time of the crash [133]. This approach and its variants has been widely adopted and utilized in numerous studies aiming to predict market bubbles and crashes [134, 135, 136].

Recent research has expanded to investigate these phenomena in cryptocurrency markets as well, using various methods including natural language processing (NLP) and time-series analysis [137, 138, 136, 139]. One notable example in the literature consists of a study that applies an LPPL variant on short-time frame data of Bitcoin prices to monitor and forecast Bitcoin bubbles and crashes in real time

[136]. Considering the potential for substantial gains in the highly volatile cryptocurrency markets, it is not unexpected to have a growing interest in predicting these events. However, this study will concentrate specifically on graph-based methods applied to U.S. stock markets only.

Although various methods have been employed for market crash prediction or monitoring, graph-based approaches have been particularly effective in revealing risks arising from the interconnected nature of financial entities. These approaches allow for the examination of temporal changes in network topology, offering valuable insights into the dynamics and vulnerabilities within financial systems.

A graph-theory-based study on the 2008 financial crisis constructed a graph by filtering daily stock return correlations using a Q=0.6 threshold. Upon analyzing the degree distributions of this graph, it was observed that the financial sector had the highest degree of interconnections. This finding indicates that the crisis's initial impact on the financial sector propagated throughout the entire market, ultimately leading to a widespread recession [140].

Another relevant study focusing on the Chinese stock market devised a metric called Information Interaction among Large Shareholders (IILS) to assess its impact on crash risk. The study found a negative correlation between IILS and crash risk. IILS was calculated using a cosharing network based on the common stock holdings of investors, weighted by the number of stocks commonly held. The network's weighted clustering coefficient was then adjusted for market capitalization to mitigate its influence. Aim of this method was to evaluate how a shareholder's trading behavior for a stock is influenced by other shareholders' actions within the coshareholding network [141].

Another study introduces an innovative approach by using weighted network clustering coefficients to gauge the speed of information transmission among shareholders. This method builds upon previous studies, which relied on network density for measuring information transmission speed. Network density, while useful, overlooks the significance of connection strength, thus missing critical information conveyed by the connection weights. Furthermore, the study identifies a U-shaped relationship between the speed of information transmission and stock volatility. This finding broadens the scope of traditional research on the correlation between information transmission and stock market volatility, to explain the dynamics of shareholder communication and its impact on market behavior [142].

In a similar approach, a study constructed a network based on the portfolio overlap of financial institutions' common asset holdings to identify channels of financial contagion and systemic risks. The research revealed that the proportion of significant overlaps (links) steadily increased leading up to the 2007–2008 financial crisis, peaking at the crisis's onset. This pattern suggests that systemic risk, particularly from fire sale liquidations, was at its highest during this period [129].

Many studies have also utilized motif analysis to examine common asset holding networks [143, 144, 145]. For instance, a study focused on China's Energy Stock Markets constructed a common asset holding network, enhancing it with various node and edge types, such as weak or strong edges and investor, fund, or individual nodes. The study then analyzed the change in motif counts over time within this network and found that in the China energy stock market, individuals and funds exhibited contrasting involvements, with high shareholding similarities predominantly found among companies and funds [146].

The aforementioned studies collectively demonstrate the extensive use of graph-based approaches in understanding stock market crashes, particularly within the U.S. and Chinese stock markets. By employing techniques such as network clustering coefficients, common shareholding networks, and motif

analysis, these studies have effectively illuminated systemic risks, crash risks, risk contagion channels, communication dynamics and co-sharing within stock markets, which are all significant details to understand investor behavior.

CHAPTER 3

METHODOLOGY

This chapter outlines the methodological framework employed in this study to process and analyze 13F filings, in order to identify and analyze investor-stock interaction patterns in critical quarters. The approach consists of several steps on data acquisition, transformations, feature extraction and graph analysis techniques.

Study started with a theory that 13F filings may contain significant pivots of certain stock-investor pattern frequencies in the context of market crash events. The research is grounded in a quantitative approach, since it utilizes network motif discovery to process multi-dimensional data derived from 13F filings, and compare frequency fluctuations on significant market events.

To ensure a clear and comprehensive explanation of the methods and approaches, this chapter is divided into three sections. The first section provides an overview of the study, to give a general idea of the methodology. The second section dives into the specifics of the data, including its sources and the transformation processes applied, and conversion of this data into a dynamic bipartite graph of interactions to address the first two objectives described in the Introduction chapter. The third section explains evaluation of this graph from various perspectives, and explain methods utilized for network pattern discovery, which addresses third and fourth objectives of the study.

3.1 Methodology Overview

Before delving into the study in depth, the following block diagram has been utilized to provide a high-level explanation of the methodology and procedures. As depicted in Figure 6, the study involves the acquisition of data, its processing, and transformations to generate a dynamic directed graph with rich attributes. Additionally, time series data for corresponding quarters, annotated with flagged market event dates, is included. These representations of U.S. stock markets which depicts investor behaviors as well as overall stock market movements in selected periods will be used to address first two objectives of the study.

Experiments continue with in depth analysis of the dynamic bipartite graph of stocks and investors in terms of both counts of certain attributes as well as overall graph structure changes each quarter. These time series data, aligned with selected stock market crash periods will be examined to identify whether there are specific behaviors exhibited by investors in context of crashes. Finally, specific motifs are generated, explained and their frequencies over time are visualized in order to address the final objective of the study.

Following sections will delve deeper into each objective with specific methods and tools to be used in experiments, as well as the rationale behind them.

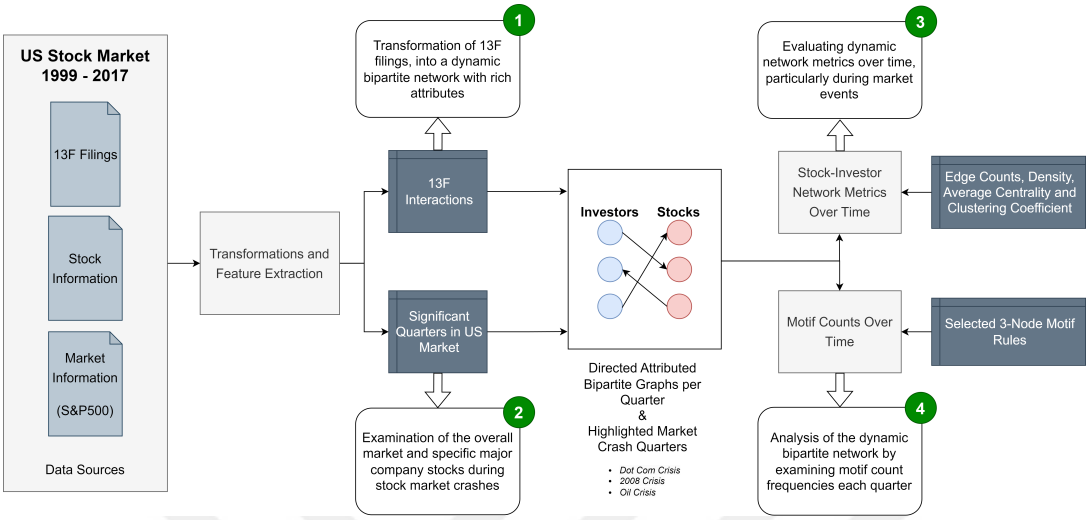


Figure 6: High level overview of the methodology

3.2 Tools and Implementation Details

The study employs numerous data processing, analysis and visualization techniques, as outlined in detail in Figure 6, to achieve its objectives. Recognizing the crucial aspect of replicability and transparency in academic research, all raw data, processed data, codes as well as commentaries are published in GitHub. This open-source approach is critical to enable fellow researchers to not only understand and test the potential findings of the study, but also improve and build upon them for further relevant studies.

The core toolkit used in this study primarily resides within the Python data science ecosystem, utilizing libraries such as Pandas, NumPy, Matplotlib, SciPy, and NetworkX alongside supplemental packages. While Pandas, Matplotlib, NumPy, and SciPy provides all the necessary tools for data exploration and processing, NetworkX plays a vital role in the manipulation and analysis of graph-structured data. For researchers without extensive coding experience, Gephi also offers a user-friendly interface for visualizing and analyzing such data. Although DotMotif package is another valuable tool for motif extraction, its processing time proved impractical for the graph used in the study with over 25 million edges. Thus, functions for motif extraction are hard coded as Python functions to reduce processing time. Other related tools for the study worth to mention are, Sec-API, a useful private API for stock information, and Yahoo Finance, a widely used publicly accessible python package for stock market pricing.

In short, to facilitate further research and improvements, all processed, raw data, as well as Jupyter notebooks containing codes and commentary are published in the public GitHub repository [1] Although

[1] Processed data, bulk data, and Jupyter notebooks for codes are all publicly available in the GitHub repository: https://github.com/ErsinDemirel94/13F-investor-graphs

minor adjustments, primarily related to file paths, might be necessary for error-free execution, the provided resources may offer a comprehensive foundation for future research.

3.3 Data Processing for Market Analysis and Bipartite Network Construction

The study stems from the disclosure requirements for public companies in the U.S., designed to enhance transparency and protect investor rights by minimizing information asymmetry, as highlighted in previous chapters. A crucial instrument for achieving this transparency is the 13F filings [35]. While several studies affirm the predictive power and benefits of these filings [46, 47, 48, 50, 49, 52, 53, 54], others have pointed out risks and the potential for investor misinformation [58, 60, 61, 62, 63]. Therefore this study consists of several data processing steps to extract critical information and structure the data, as summarized in the Figure 7.

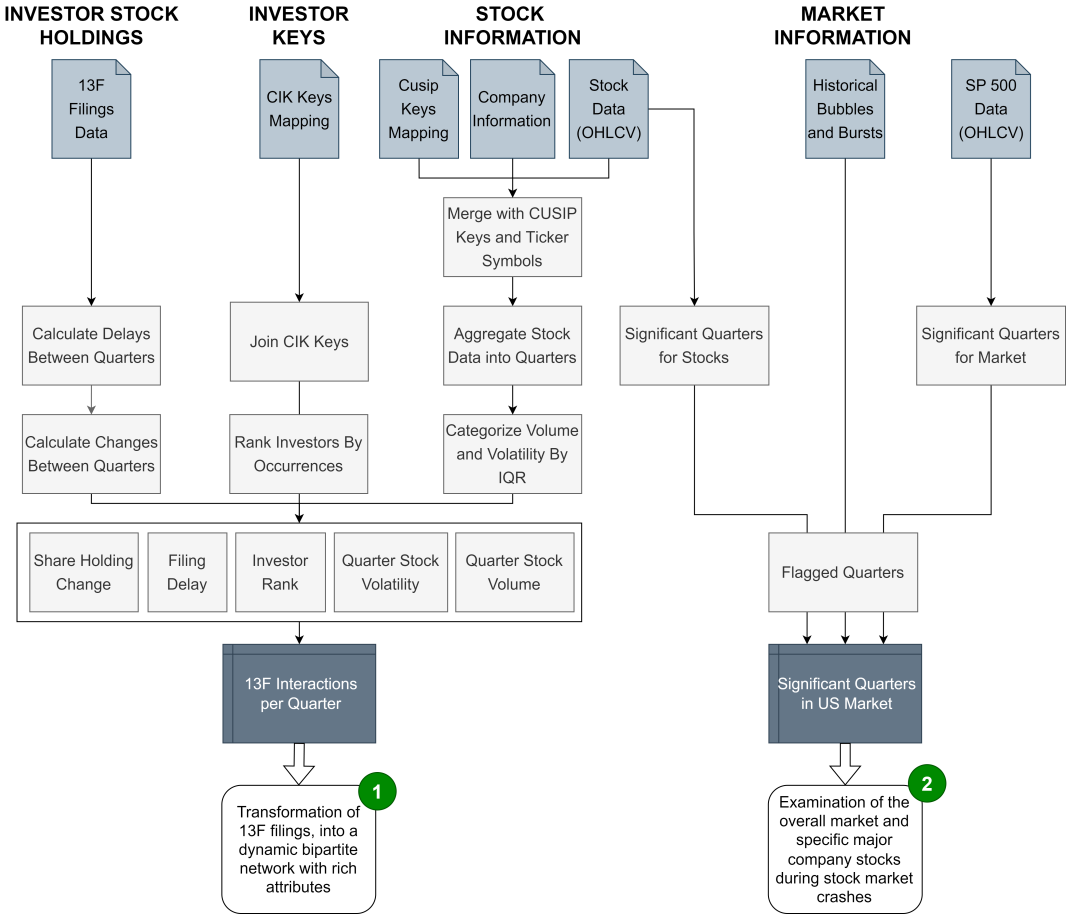


Figure 7: Data processing steps depicted in diagram aims first two objectives

This study primarily utilizes processed 13F filings from the SEC's EDGAR database, spanning fiscal years 1999 to 2017. The dataset, derived from Sinkinson's study, is a structured compilation of 13F filings scraped from the EDGAR database and processed using various text processing techniques [147]. It consists over 40 million rows, detailing CIK keys of investors, CUSIP keys of stocks, quarter and filing submission dates, type of filing, and the amount of holdings. A detailed description of

the data will follow in the next chapter, but a brief overview of its characteristics is essential for understanding the methodology and subsequent transformations.

The Central Index Key (CIK) is used in the SEC system to identify entities filing reports, while the Committee on Uniform Securities Identification Procedures (CUSIP) number uniquely identifies stocks and registered bonds in the U.S. and Canada. While CIK is straightforward, CUSIP's structure is more complex: the first six digits represent the issuer, with the remaining digits providing security type and a check digit generated by a mathematical formula. To enrich the 13F filings, base CUSIP keys are matched with company information from the SEC API, including names, industries, sectors, and stock ticker symbols, which are more commonly used in stock market contexts.

13F filings, reflecting investor holdings per quarter, are transformed into changes in holdings using a rolling window approach. This approach consists of grouping data by equity and investor, and calculating changes in share numbers. These changes are then categorized into meaningful groups such as 'large buy' or 'significant sale' based on their percentile in interquartile range of historical changes for each stock. Examination of changes of instituitional positions rather than quarterly holdings of instituitions is a popular approach to examine flow of capital in many studies [47, 40, 54, 53, 51].

As it discussed in the second chapter, filing delays are highly used mechanism by investors to hide their intentions [58, 60, 62, 63]. In order to calculate filing delays, quarter date and submission dates in 13F filing data are simply subtracted. While 13F filings are legally required to be submitted within 45 days of a quarter's end, investors may strategically delay their filings past 45 days by either providing advance notice accompanied by a request for confidential treatment, or maintaining a clear track record to avoid fines [58]. In light of these factors, and to investigate the implications of delays in 13F submissions, filing delay values are classified as either 'on-time' (submitted on the 45th day), 'early', or 'late'.

In order to group investors into meaningful categories for more explainable results, previously explained change categories, such as 'large buy' or 'significant sale' are converted into numerical values, and summed for each investor. At the end of this procedure, each investor is classified into three distinct groups, named 'active', 'average' and 'inactive' according to the interquartile range they fall into. This approach effectively segments institutional investors mentioned in 13F filings into meaningful categories, paving the way for a deeper examination of their interactions in subsequent analyses.

The study also merges stock volume and volatility of corresponding stocks for each quarter. Quarterly volume is calculated by aggregating daily volumes, while volatility is calculated standard deviation of daily returns derived from adjusted close prices. The adjusted close price is preferred since it commonly preferred variable since it accounts for dividends and stock splits to give more complete picture. Volume and volatility are then converted into categorical variables to enrich bipartite graph, and to be used as filters in motif count process, similarly with changes in equities and filing delays.

Ultimately, these processes paved the way in the creation of a dynamic bipartite network of stock and investor relations for each quarter, enriched with the attributes mentioned above, as illustrated in Figure 8, to address first objective of the study.

Gaining insights into market behavior during crash periods and defining significant quarters is the second objective of the study. To address this objective, the study not only relies on individual stock prices data but also incorporates S&P 500 data. This benchmark index data is processed into quarterly time format by averaging prices, summing volumes, and calculating volatility. Quarterly volatility,

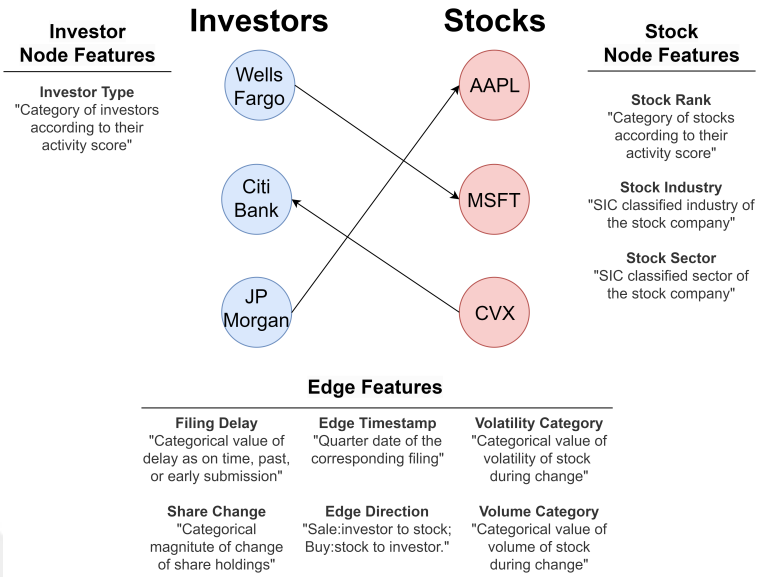


Figure 8: Output of the first objective is a dynamic bipartite graph of stocks and investors with rich features

which is a critical metric to display risk, is derived from the standard deviation of daily returns averaged over the quarter, providing a comprehensive view of market dynamics. Specific quarters are marked based on the occurrence of historical market events, such as the Dot-Com Crisis, the 2008 Financial Crisis, and the Oil Market Crisis. These events were chosen for their significance within the scope of the available data period [148, 149, 66].

Each of these events are explained briefly, highlighting key related industry groups, and displaying key stocks in these industries in terms of volume and volatility. Knowing the fact that the complex composition of the stock market employs various hedging mechanisms to mitigate sharp declines, the optimal approach involves isolating specific industries and conducting a comparative analysis during market crashes. This approach allows for a clear visualization of relative changes within these industries.

3.4 Dynamic Network Analysis and Pattern Discovery in Stock-Investor Relationships

Previous section explained steps to create dynamic stock-investor bipartite graph. In order to preserve dynamism to identify changes, for each quarter number of static graphs are extracted by keeping same nodes, but changing the edges which is a commonly used approach to model dynamic graphs [116, 126, 143, 150]. This approach can be translated into having a graph with snapshots of edges in each quarter during period between 1999 and 2017. In order to address third and fourth objectives of the study, efforts applied are shown in the Figure 9.

To address the third objective, each snapshot of the dynamic graph is analyzed using edge counts and several key graph metrics that effectively characterize network properties over time. Filing delay categories and change categories derived from previous objectives are grouped by industries and counted to highlight shifts during market crashes. In addition, selected metrics to explain global graph character-

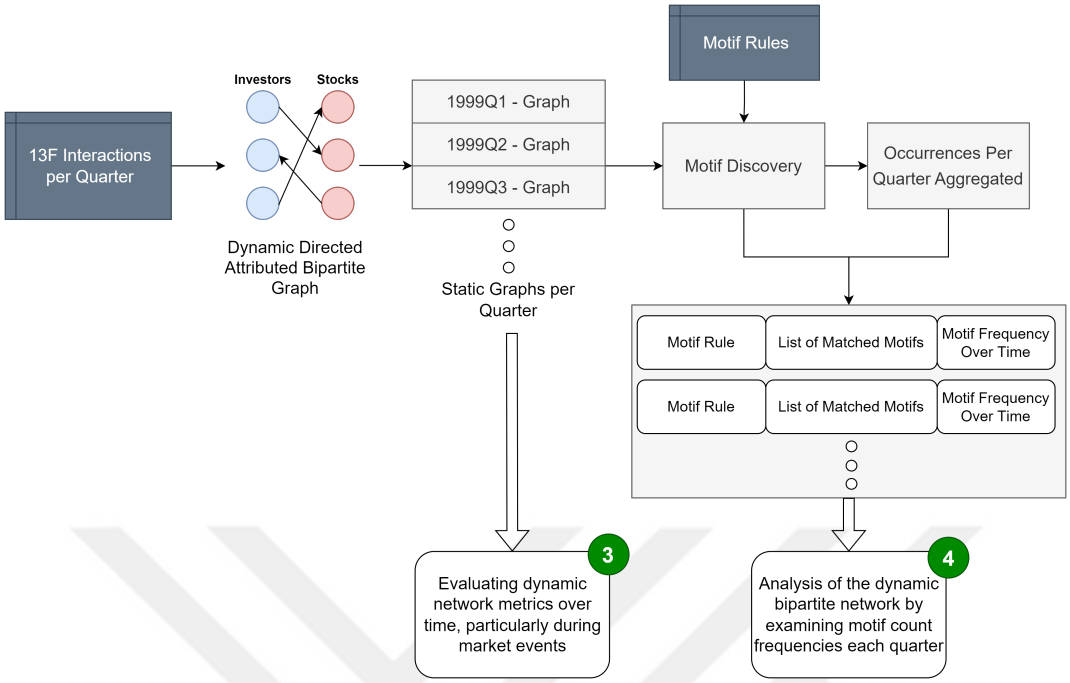


Figure 9: Third and fourth objectives of the study is to examine graph structure and change of frequency of motifs over time

istics include graph density, average degree and clustering coefficient. Graph density provides insights into the overall connectedness of the network, while the average degree reveals the typical number of connections per node. The clustering coefficient measures the tendency of nodes to cluster together, highlighting potential market segments or investor communities, as well as tendency of investors to co-invest into same set of stocks.

These metrics are crucial in this context as they illuminate the structural dynamics of the stock market network [140, 132, 129, 142]. Significant peaks or dips in these metrics over time can signal shifts in market behavior, such as the formation or dissolution of investor groups, changes in market influence of particular stocks, or shifts in investment strategies. By examining these metrics across different quarters, especially during market crises, the study aims to uncover patterns and anomalies that might be occurring to provide a deeper understanding of market dynamics and investor behavior.

In addition to standard graph metrics, this study incorporates motif discovery, a technique often employed in biological network analysis, to resolve complex networks. Motif discovery involves identifying recurring, significant patterns of interconnections among a set of nodes in a network. These motifs, or network patterns, are small subgraphs which fits under certain rules or criterias that repeatedly occur within a larger network, and are thought to signify underlying structural properties of the network. In the context of this study, motif discovery is used to extract insights beyond the global topology of the network.

Analysis of frequency of motif counts is a common method to examine dynamic graphs in various studies [143, 144, 146, 151]. In this study, similar with previous objectives, examination of motif counts particularly focuses on how these patterns change during market crash periods. Because of the

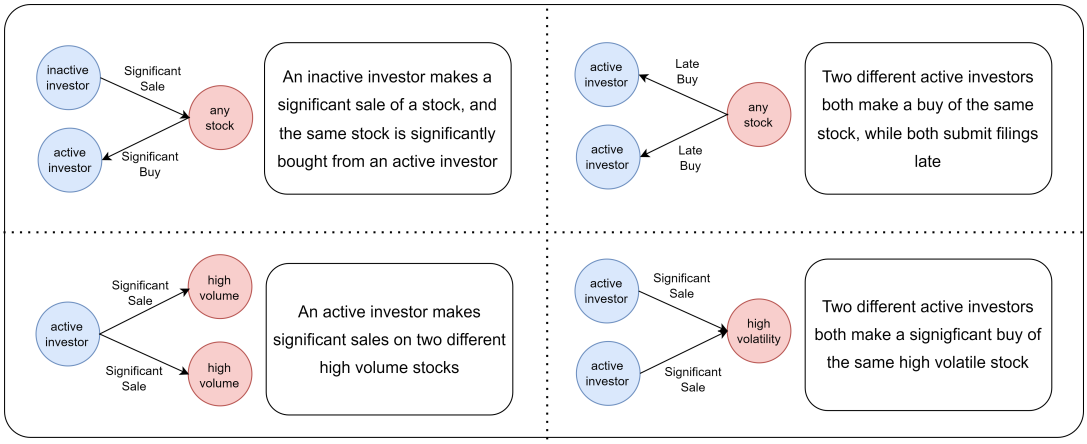


Figure 10: Motif rules has high explainability and customization

previous efforts of creating a graph with rich features, the analysis of motif counts is expected to have high explainability, as depicted in the Figure 10. Motif counts are expected to highlight collective behaviors among different investor types, or show how certain industries are preferred or avoided in different market conditions. This study adds additional layers of complexity to motif counts over time by incorporating filing delays, stock volume and volatility, industry classifications, and the magnitude of changes.

To summarize, the methodologies outlined in this section aim to comprehensively analyze the dynamic graph created from the preceding objectives, employing graph data science tools to investigate changes at both the overall graph characteristics and certain patterns within the graph. First two objectives aim to setup the scene to be inspected, by processing data from various sources to create interactions of investors, stocks in the relevant context with additional information. Second two objectives aim to capture characteristics and network oriented metrics with visualizations and interpretations.

CHAPTER 4

EXPERIMENTS AND RESULTS

This chapter presents the experiments and detailed findings in line with the defined objectives and the methodology. Starting with an in-depth look at the data, the chapter explores its characteristics, followed by the construction of the graph and an analysis of industry responses during market crashes. Calculations of key attributes and network metrics reveal temporal changes, emphasizing the impact of market downturns. The chapter concludes by examining specific graph motifs, offering visual insights accompanied by clear explanations, contributing to a comprehensive understanding of investor behaviors during stock market crashes.

4.1 Data Processing, Analysis and Network Construction

4.1.1 Description of the Data

As mentioned in the methodology, the study utilized various open data sources, with the primary one being the extensively processed and extracted 13F filings data. This dataset, spanning from 1999 to 2017, comprises approximately 40 million rows of portfolio holdings information of accredited investors. It was prepared using PERL language, sourced from EDGAR, and then processed and formatted to examine common ownership of stocks among investors [147]. This dataset was chosen due to its comprehensiveness and size, making it one of the most substantial data resources in this field. Details of 13F filings data are summarized in the Table 2.

In the dataset, investors are identified using the CIK key, while equities are identified using the CUSIP key. While the CIK key is readily matched as it's provided in the data, the CUSIP keys are more complex. CUSIP keys are alphanumeric characters used for each equity, where the first 6 characters form the base, identifying the issuer, and the remaining suffixes denote other equities offered by the issuer. CIK and CUSIP are two columns to represent equities and investors as entities in interaction.

The data reflects each investor's equity holdings, in other words, long positions in specific assets for the relevant quarter with its shares column. However, it is important to note that 13F filings have limitations, such as not including short positions, bond positions and options positions [64].

As discussed in previous chapters, 13F filings are documents required to be submitted every quarter and are notorious for their filing delays [58, 60, 61, 62, 59]. In this data, the 'rdate' column represents the quarter for which the filing is required, while the 'fdate' column indicates the actual date of submission.

Table 2: Source data is a large dataframe of 48.719.087 rows, extracted from historical 13F filings

Column	Example	Unique Count	Minimum	Maximum	Explanation
cik	354204	8170	2110	1723681	CIK key to represent the investor
cusip	00036020	17811	00030710	Y9530810	CUSIP key to represent issuer of equity
shares	164620.0	3589492	0.0	3095131904.0	Number of holding shares of investor
rdate	19990331	75	19990331	20170930	Quarter date of the filing
fdate	19990513	4133	19990401	20171221	Submission date of the filing
filetype	13F-HR	5	EX-1	13F-HR	Form 13F filing type

In the dataset, the 'filetype' column distinguishes between five types of 13F filings, primarily categorized as 13F-HR or 13F-NT shown in Figure 11. The 13F-HR, or Holdings Report, is used by institutional investors to report all securities they manage. In contrast, the 13F-NT, or Notice Report, is filed when the institutional investor does not directly manage any reported securities, indicating that these securities are managed by another entity and will be detailed in that entity's report. Moreover, an '/A' suffix in either filing type denotes an amendment, meaning updates or corrections to previously submitted information [59].

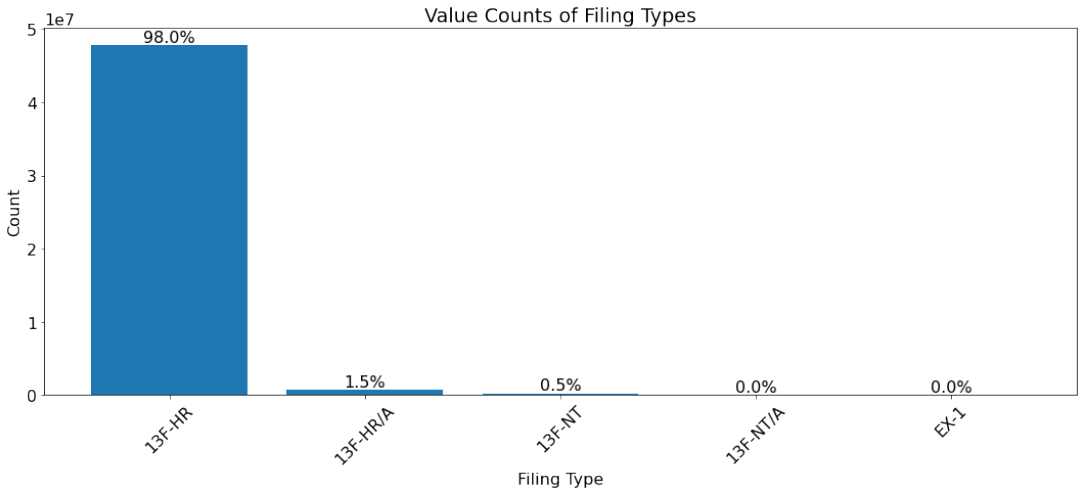


Figure 11: While most of filing types are 13F-HR, there are also amendmends and notice reports

Although the dataset is highly informative, the only detail it provides about the interacted equities is their CUSIP keys. The CUSIP, or Committee on Uniform Security Identification Procedures, is a nine-character code, either numeric or alphanumeric (e.g 037833100 for Apple), which uniquely identifies North American financial securities to aid in trade clearing and settlement. An issuer may

have multiple CUSIP keys, but the first six characters, known as the base CUSIP, are unique to each issuer or company.

However, CUSIPs alone do not provide sufficient context for understanding a stock; more detailed information is needed. To this end, data from the sec-api data source for all available markets was used. This data was then matched with the CUSIPs in the 13F filings to provide additional context about the stocks. This dataset was chosen due to its comprehensive historical records and inclusion of delisted companies, such as Lehman Brothers and Bear Stearns, which were bankrupted during the 2008 financial crisis. An example record from sec-api data is summarized in Table 3.

Table 3: SEC API data is used to further enrich information available in 13F filings

Column	Example	Unique Count	Total Count
name	ADMIRALTY BANCORP INC	23251	33308
ticker	AAAB	27798	33308
cusip	007231103	27772	33308
exchange	NASDAQ	6	33308
isDelisted	True	2	33308
category	Domestic Common Stock	24	33308
sicSector	Finance Insurance And Real Estate	11	33308
sicIndustry	State Commercial Banks	440	33308
location	Florida; U.S.A	137	33308

In this dataset, additional information such as sector, industry, category, and location was merged with the previous data for better understanding. The dataset employs the Standard Industrial Classification (SIC) system to represent company sectors and industries. According to this classification, each sector comprises various sub-industries, allowing for a detailed and structured understanding of the companies' operational fields.

Table 4: Price Data consists of OHLCV data of available stocks

	Open	Close	Adj Close	Low	High	Volume
Count	2.10e+07	2.10e+07	2.10e+07	2.10e+07	2.10e+07	2.10e+07
Mean	7.61e+01	7.61e+01	1.48e+14	7.42e+01	7.80e+01	1.23e+06
Std Dev	2.85e+03	2.87e+03	4.57e+16	2.75e+03	2.30e+03	1.32e+07
Min	4.00e-04	2.00e-04	2.28e-09	1.00e-04	4.00e-04	1.00e+00
25%	7.50e+00	7.50e+00	4.62e+00	7.36e+00	7.63e+00	2.21e+04
50%	1.55e+01	1.55e+01	1.14e+01	1.52e+01	1.57e+01	1.26e+05
75%	2.97e+01	2.97e+01	2.47e+01	2.93e+01	3.01e+01	6.07e+05
Max	2.03e+06	1.78e+06	1.89e+19	1.44e+06	2.07e+06	4.48e+09

Another dataset used in this study is historical pricing data, which includes a comprehensive record of various stocks' Open, High, Low, Close, and Volume (OHLCV) data over a specified period. This data was sourced from Yahoo Finance, a frequently used platform for stock price information. Pricing data is a frequently used tool by stock investors to understand price movements and market trends during the specified time frame. Historical pricing(OHLCV) data for each stock is described in Table 4.

Finally, to examine the impact of historical bursts and bubbles on the overall market, the study also utilized daily OHLCV data of the S&P 500 index, which is a commonly used benchmark for stock market investors. S&P500 price (OHLCV) data has same columns as the price data, which described in Table 5. Periods this study aim to investigate are visible as significant market downturns in S&P500 daily stock prices visualised in Figure 12.

Table 5: S&P500 Price Data also consists of OHLCV data of overall S&P500 companies

	Open	Close	Adj Close	Low	High	Volume
count	23323.00	23323.00	23323.00	23323.00	23323.00	23323.00
mean	486.82	486.91	486.91	483.76	489.69	769166772.71
std	726.18	726.27	726.27	721.89	730.11	1484863543.44
min	4.40	4.40	4.40	4.40	4.40	0.00
25%	23.85	23.85	23.85	23.85	23.85	1280000.00
50%	99.50	99.50	99.50	98.72	100.30	17340000.00
75%	834.03	834.70	834.70	822.88	843.74	537855000.00
max	3564.73	3580.84	3580.84	3535.22	3588.11	11456230000.00



Figure 12: S&P500 prices data at the time period of utilized 13F filings

In conclusion, raw data sources used for this study includes historical 13F filings, stock company meta information, historical stock price data and historical S&P index price data.

4.1.2 Data Transformations and Feature Extraction

To ensure the meaningfulness of the collected data, various processes were applied. These procedures involved merging sources, cleaning missing data and converting the values into categories based on their probabilistic distributions. This systematic approach was critical in structuring the data in a way that facilitates a clearer understanding of the underlying patterns and trends.

To do so, first, the 'filing delay' column was processed. This involved a subtraction to determine the filing delay value for each record, based on the difference between the filing submission date (fdate) and the relevant quarter (rdate). A notable observation from the Figure 13 is the concentration of filing

delays around a 45-day, explained by the regulatory requirement for filings to be submitted within 45 days [59]. Additionally, it was often observed that filings were submitted later than this deadline, due to investors requesting confidential treatment to delay disclosure [59]. Consequently, the filing delay was categorized into following three groups:

- <45 days of delay: early submission
- >45 days of delay: late submission
- =45 days of delay: on time submission

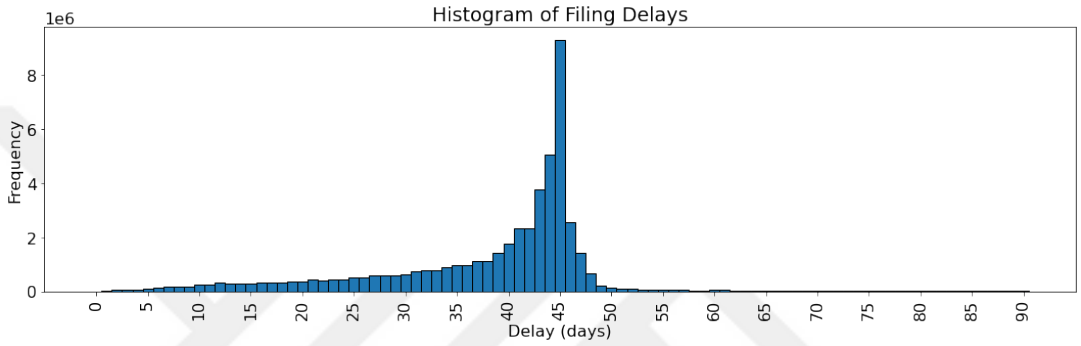


Figure 13: Filing delays are populated around the 45th day since its due date for submission

In addition, 13F holdings of investors was transformed to focus not just on quarterly positions but on the changes over time. This transformation involved the following steps:

- Data was grouped by CIK and CUSIP codes, aligning them by investor and equity.
- A new column, 'diffs', was created to reflect the change in shareholdings from one quarter to the next.
- The initial quarter's data (dated 1999-03-31) was removed, as it lacked a previous quarter for comparison.
- NaN values were considered as initial shareholdings due to the absence of prior data for comparison.
- Zero values were eliminated, indicating quarters with no change in holding.

Following these operations, the dataset was transformed into a format that facilitates interactions, forming the basis for the graphs shown in Figure 14. To further enhance this, the data was grouped by each stock ticker and ordered according to their values. Then, based on the quantile they fall into, the numerical values were converted into the following categorical values:

- Neutral: Between >2/3th of negative, and < 1/3th of positive values

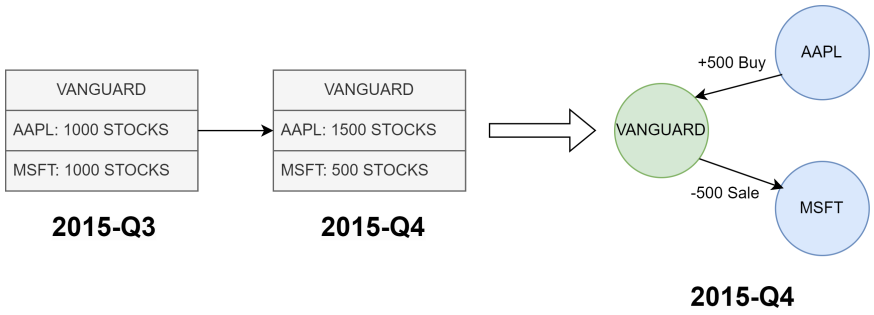


Figure 14: Holdings per quarter are converted into changes to be displayed as interactions

- Buy: Between 1/3th and 2/3th quantiles of positive values
- Sale: Between 1/3th and 2/3th quantiles of negative values
- Large Buy: Between 2/3th and 0.95th quantiles of positive values
- Large Sale: Between 0.05th and 1/3th quantiles of negative values
- Significant Buy: Highest >0.95th quantiles of positive values
- Significant Sale: Lowest <0.05th quantiles of negative values

These efforts led to convert holdings into categorical interactions of stocks and investors per quarter, displayed in Figure 15.

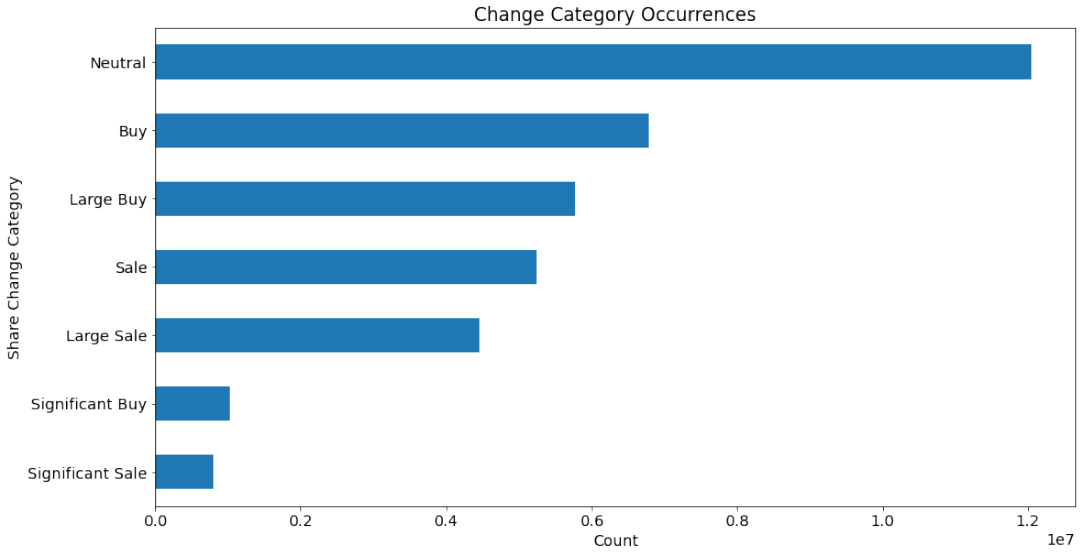


Figure 15: Share changes are categorized based on their values and prevalence in data

Subsequently, these categorical variables were utilized to categorize investors as well, based on their trading activities. To achieve a clear classification, the previous categories were assigned numerical scores: 'Significant' activities were valued at 3 points, 'Large' activities at 2 points, regular 'Buy'

and 'Sale' at 1 point, and 'Neutral' activities were assigned 0 points. The 13F interaction data was then grouped by each investor, where their activity scores were aggregated. Considering the long-tail distribution of activity scores as shown in the Figure 16, based on this cumulative scoring:

- investors with 0.05th quantiles of activity scores are removed to reduce sparsity,
- investors in between 0.05th and 0.50th (median), are flagged as inactive investor,
- investors in between 0.50th (median) and 0.95th quantiles are flagged as average investor,
- investors in equal or higher than 0.95th quantile are flagged as active investor.

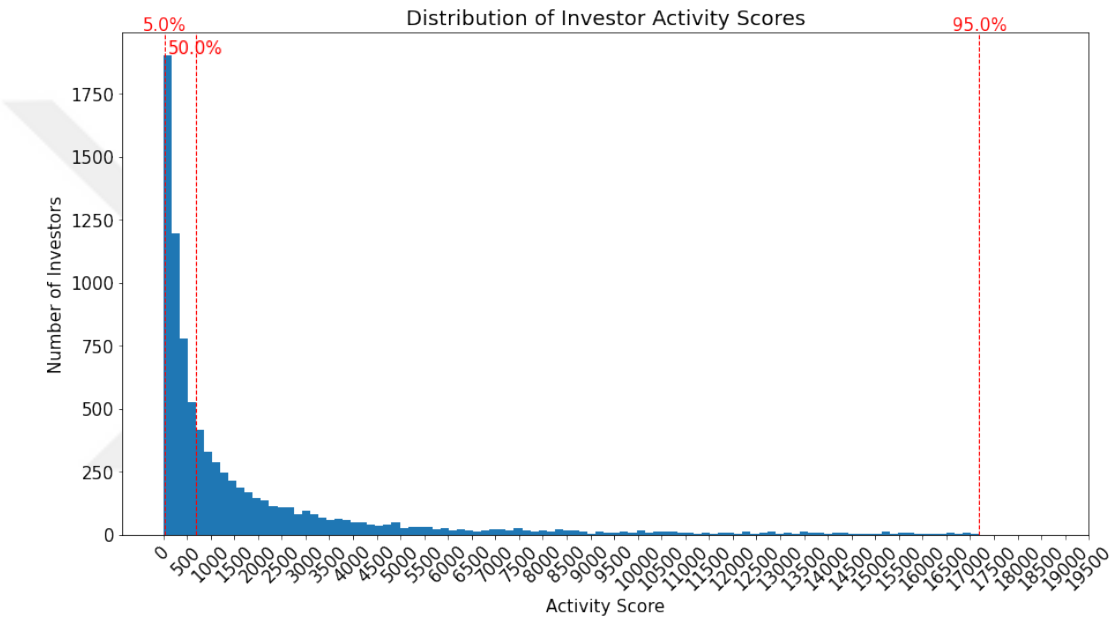


Figure 16: Investors with 5-50th percentile are inactive, 50-95 are average, and more than 95th are active classified investors

Similar approach is also applied to stocks as well given the fact that they also have a long tail distribution, shown in Figure 17:

- 0.05th quantiles of stocks are removed to reduce sparsity.
- Between 0.05th and median, are flagged as small stocks.
- Between median and 0.95th quantiles are flagged as medium stocks.
- Higher than 0.95th range are flagged as large stocks.

It is important to note that these two approaches for classifying investors and stocks are solely based on 13F filing mentions of these entities. Therefore, they are not absolute, but rather assumptions derived from 13F filings. These variables should be interpreted in this way in subsequent analyses.

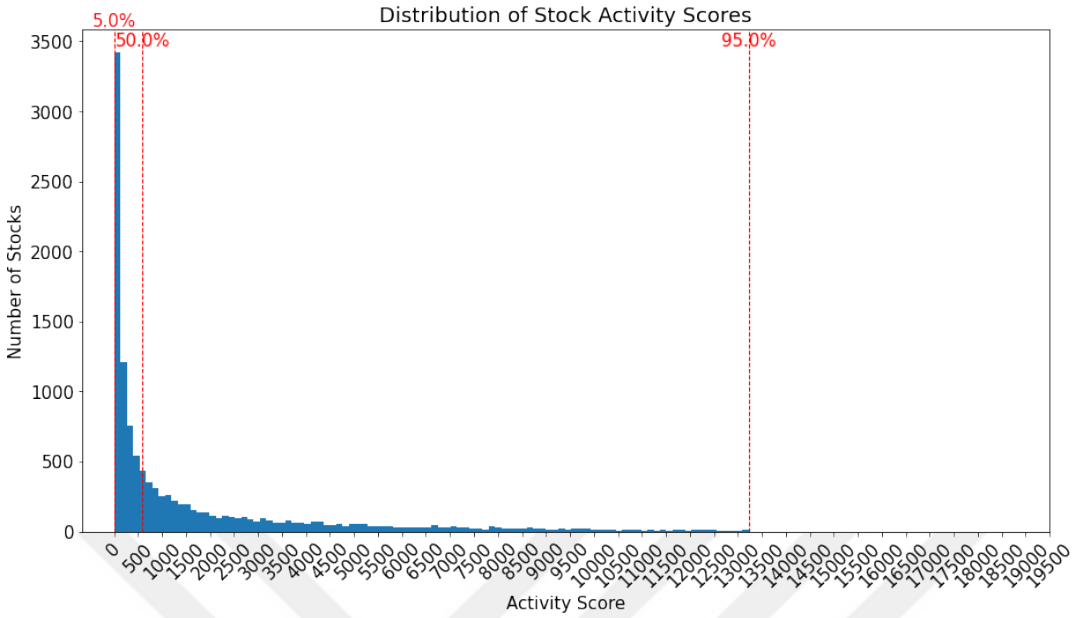


Figure 17: Stocks with 5-50th percentile are small, 50-95 are medium, and more than 95th are large classified stocks

To enrich the stocks with additional attributions, the sec-api dataset was merged with the 13F filings by their base CUSIP keys. This enabled to attribute each stock by its SIC industry and SIC sector as shown in Figure 6 with distributions of occurrences in 13F filings shown in Figure 18. Additionally, merging stock tickers enabled joining price data of corresponding stocks, for more specific analyses. Given the importance of industry and sector information, records that did not successfully merge were removed from the dataset and analysis.

Table 6: Industry Sectors and Corresponding SIC Code Ranges

SIC Code Range	**Sector**
0100-0999	Agriculture, Forestry and Fishing
1000-1499	Mining
1500-1799	Construction
1800-1999	not used
2000-3999	Manufacturing
4000-4999	Transportation, Communications, Electric, Gas and Sanitary Service
5000-5199	Wholesale Trade
5200-5999	Retail Trade
6000-6799	Finance, Insurance and Real Estate
7000-8999	Services
9100-9729	Public Administration

To make the price data compatible for analysis, it was necessary to aggregate values on a quarterly basis. For this purpose, the price data was grouped by each ticker, with days matched and grouped according to their corresponding quarters. For aggregation of values, following approaches are applied:

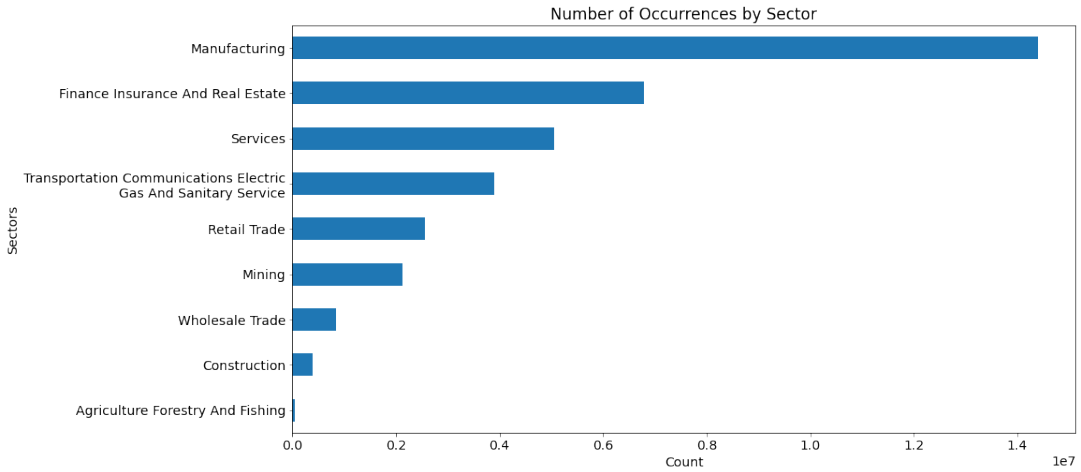


Figure 18: Sector mentions in 13F filings are distributed as follows

- Quarterly Volume: The total volume for each quarter was calculated by summing up the daily volumes within that quarter.
- Quarterly Volatility: First daily returns are calculated as percentage change of adjusted close values. The data was then grouped into quarters, and the standard deviation of the daily returns within each quarter was calculated as quarterly volatility.

Then, these numerical values, grouped by each stock and quarter, are categorized based on their position within the data's quantile distribution, as visible in distribution of values in Figure 19. This classification is as follows:

- Very Low: Values below the 0.05 quantile in volume or volatility, indicating minimal trading activity (volume) or price fluctuation (volatility).
- Low: Values between the 0.05 and 0.50 quantiles, representing below-average activity or fluctuation.
- High: Values above the median but below the 0.95 quantile, showing above-average activity or fluctuation.
- Very High: Values above the 0.95 quantile, marking extremely high activity or significant price changes.
- NaN: For stocks and quarters lacking sufficient data, volume or volatility categorization is marked as 'NaN'.

Finally, to inspect the overall market performance for each quarter, S&P 500 data was similarly collected for the relevant date range and aggregated on a quarterly basis. This approach enabled the calculation of quarterly volume and volatility values for the S&P 500 index as depicted in Figure 20, providing a comprehensive overview of the market's trends during each quarter.

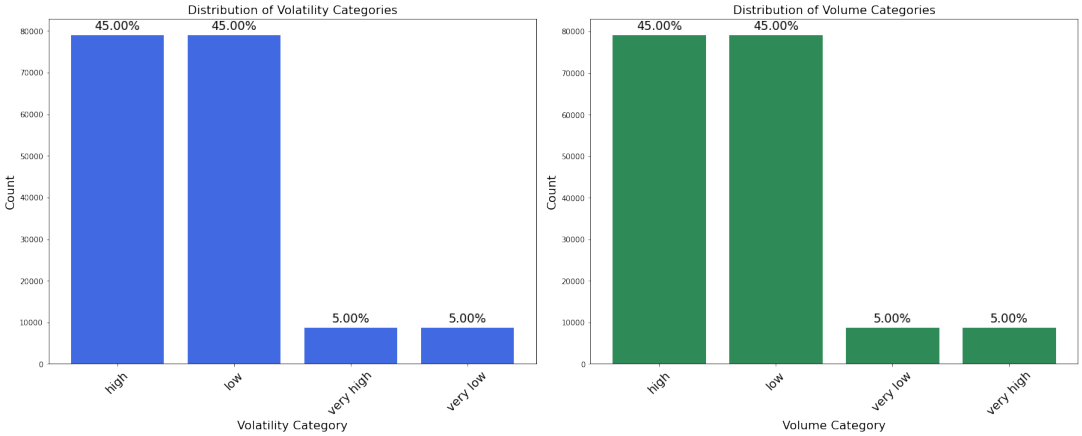


Figure 19: Volume and Volatility Category Distributions

The S&P 500 is a popular benchmark for the entire stock market as it tracks the performance of the 500 largest companies, weighted by their market capitalization. Therefore, the figure depicting volatility and volume of the S&P 500 index per quarter, as illustrated in Figure 20, can highlight uncertainty and heightened market activity throughout different quarters. Notably, the increasing trend in volume since the mid-2004s may suggest a growing popularity of the stock market. Additionally, the figure highlights peaks in market volatility during the late 2002s, mid-2009s, 2012s, and 2016s, which may signify periods of market uncertainty.

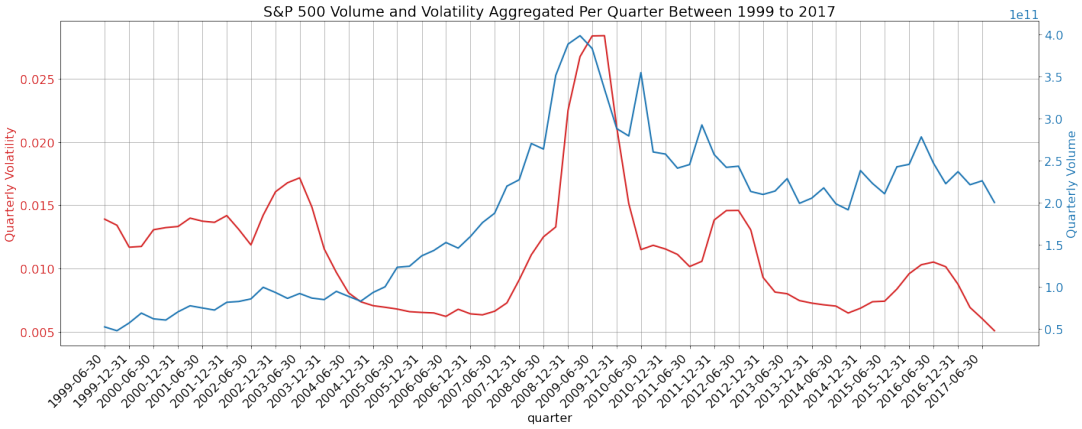


Figure 20: Quarterly aggregated volume and volatility of S&P500 index

Through these efforts, the 13F filings, price data, and stock information data have been merged, cleaned and appropriately transformed into categorical variables based on specific contexts. The data obtained at the end of this process will be used in the following sections for constructing stock-investor graphs, and examining specific quarters during bubble and crash periods.

4.1.3 Construction of Bipartite Network

In the preceding subsection, the process of transforming 13F filings into interactions was detailed. While graphs are common tools for contexts that require an examination of interconnectedness, the literature demonstrates a variety of approaches to graph creation. Numerous other studies have similarly utilized 13F filings to trace the flow of stocks or funds, adopting comparable methodologies [124, 126, 43, 53]. Other studies also processed different data sources to create information networks [43, 53, 125, 127].

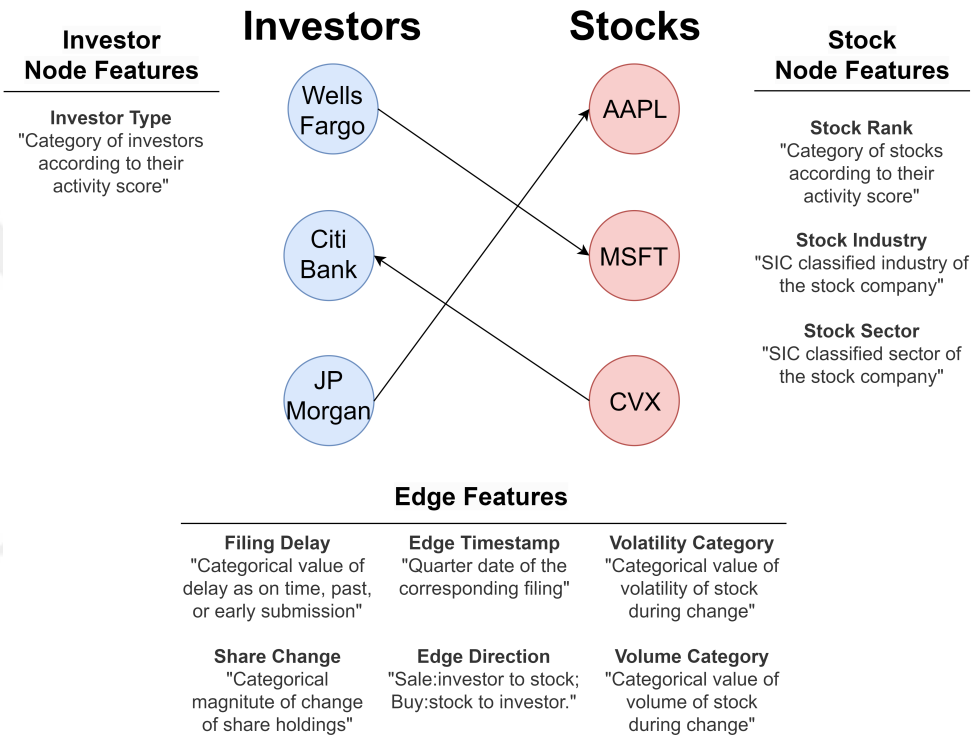


Figure 21: Node and edge features of the created bipartite graph of stocks and investors

The objective of this thesis is to explore not only the topological patterns, but also the rich features like the amount of change, filing delays, and stock market conditions. Therefore, in this study, interactions are directly utilized to form a directed bipartite multigraph with edge attributes, as depicted in Figure 21. Since the study focuses on patterns involving edge attributes, it is not feasible to simply aggregate or sum these edge features, thus, projection techniques are not utilized.

- **Bipartite:** The graph is bipartite, since it consists of two distinct sets of nodes: investors and stocks. While there are overlaps, as some investor companies are also stocks, the nodes are clearly separated into these two categories with suffixes to only allow edges between them.

- **Dynamic:** The graph's dynamism stems from the quarterly reporting of holdings. Although it is not granular in time, the interactions are dynamic, with edges marked by corresponding quarters.

- **Attributed Edges:** Each edge in the graph is attributed with features derived from various data sources. This provides a richer context for understanding how the interactions occur and evolve over time.
- **Multigraph Structure:** The foundational graph is a multigraph, as it contains multiple links between the same pair of nodes. However, for analysis, this graph splitted into 74 distinct graphs, each representing a different quarter. These separate graphs maintains the same set of nodes but allows for the examination of how edges change over time.

The multigraph created in this study represents notable changes in equity holdings by investors from 1999 to 2017, thus, insignificant amount of changes in equities are not included in the graph. In other words, previously neutral, buy and sale classified changes in holdings mentioned in 13F filings are discarded. The resulted multigraph is summarized in Table 7.

Table 7: Multigraph is main repository of processed data with all quarters combined

Variable	Value	Explanation
Nodes	20391	Number of unique entities in graph
Stock	8836	Number of unique stocks in graph
Investor	11555	Number of unique investors in graph
Edges	24596386	Number of unique interactions in graph
Density	0.1183	Ratio of edges in graph by maximum possible edges
Timestamp	74	Number of available timestamps in graph
Edge Features	9	ChangeCat, DelayCat, VolatilityCat, VolumeCat, Direction, Sector, Industry, Quarter, EdgeId

Finally, the original dataset, comprising over 48 million rows, is transformed into a multigraph containing 24.5 million edges. This reduction results from converting holdings into quarterly changes, unmatched CUSIP keys of stocks, unmatched industry and sector information of stocks, removed inactive investors and stocks to reduce sparsity, and primarily, excluding insignificant changes (buy, sale, neutral classified changes) in equities.

To quickly identify key stocks and investors, a common method involves assessing node centrality, particularly the degree of each node. The degree of each node, reflects the total number of transactions each node (stock or investor) has made over all the years covered by the study, highlights the most active entities in the market.

As shown in Figure 22, most active 10 investor nodes include major banks such as State Street, Deutsche Bank, Morgan Stanley, Citigroup, and Bank of America, as well as key financial institutions like Northern Trust, Goldman Sachs, Vanguard, and TIAA (Teachers Insurance and Annuity Association of America).

Examining the stock nodes with the highest degree centrality in the network in Figure 23, Comcast emerges prominently, reflecting its global conglomerate status in telecommunications and its position as the largest cable TV, internet, and home phone service provider. Technology giants like Microsoft, Cisco, General Electric, Intel, and IBM (International Business Machines Corp) also feature prominently. Additionally, companies such as Johnson & Johnson and Procter & Gamble, which operate in

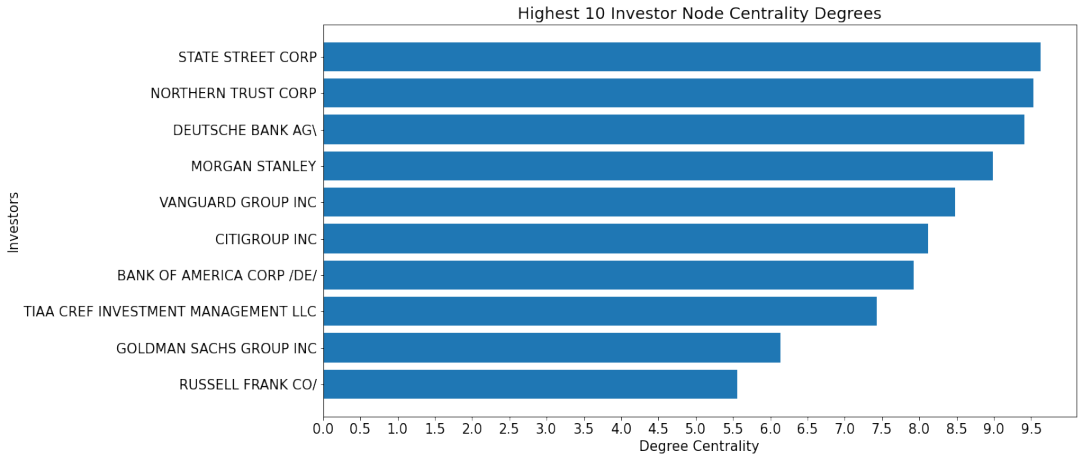


Figure 22: Most frequently interacted 10 investors

various consumer goods sectors, are notably active. Exxon, a major player in the oil industry, is also among the stocks with the highest centrality in the network.

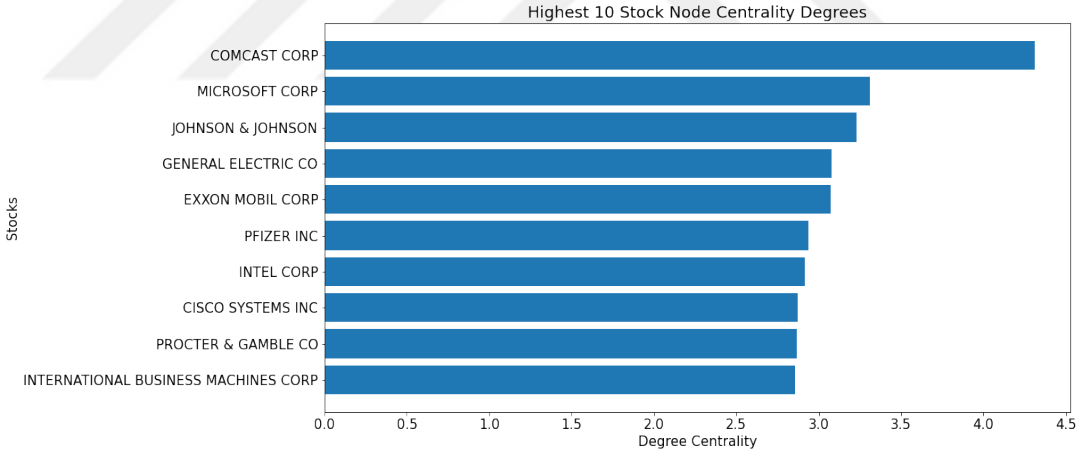


Figure 23: Most frequently interacted 10 stocks

In conclusion, the sections discussed so far have outlined the transformation of datasets into a network structure, primarily focusing on constructing a multigraph as a fundamental data repository which is summarized in Table 7. The forthcoming sections of this study will delve into the analytical phase, partitioning this multigraph into a series of static graphs based on quarterly timestamps. This approach will enable an analysis of the evolution of interactions within the financial market across different time periods, providing a clearer view of the dynamic nature of these interactions and their impact over time.

4.1.4 Examination of Bubbles and Crashes

Before delving into network analysis, defining boundaries of stock market bubbles and crashes within the study's scope is crucial. While this study does not aim to predict these events, it aims to examine investor networks and market conditions to identify potential precursors and characteristics of crashes. Given that the timeframes are based on quarterly data, this section aims to effectively identify specific quarters and provide a context for each identified period.

The study primarily concentrates on three pivotal events:

- **DotCom Crisis:** From Q2 2000 to Q4 2002
- **Financial Crisis:** From Q4 2007 to Q4 2009
- **Oil Crisis:** From Q2 2014 to Q1 2016

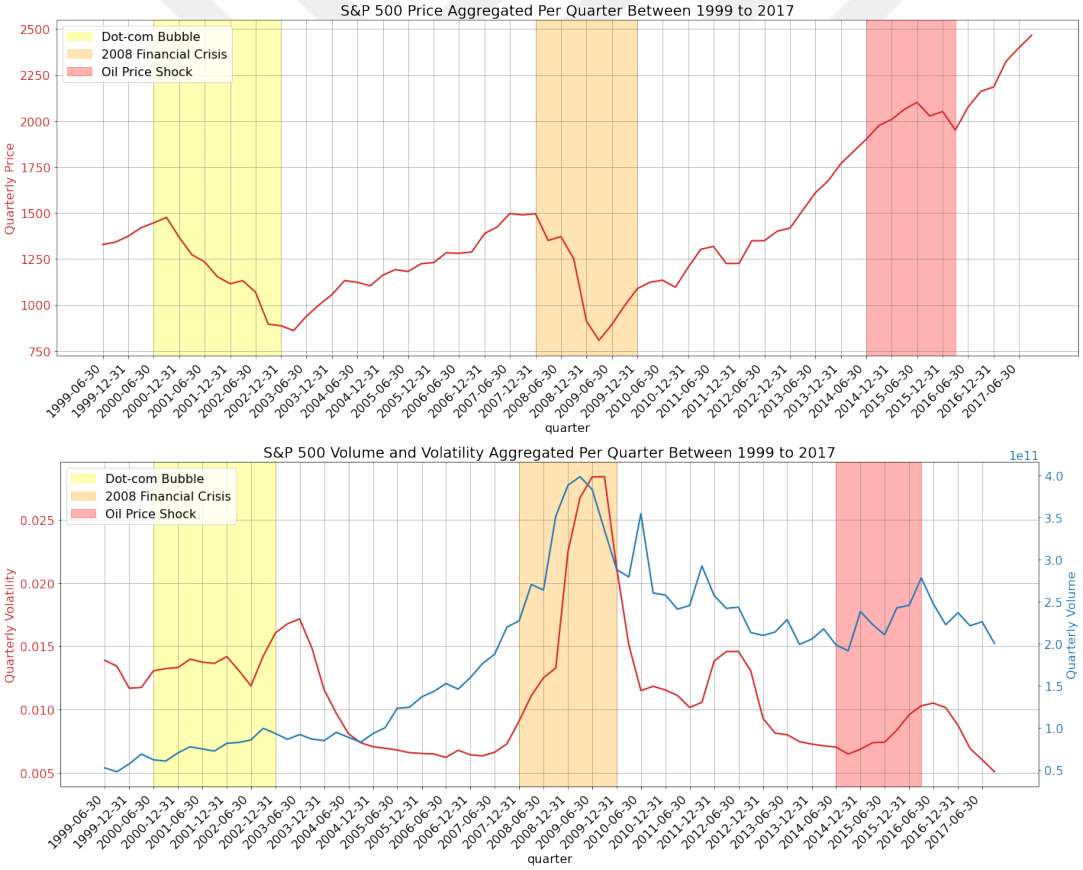


Figure 24: All three crisis have increased volume and volatility, and decreased price with most significant being 2008 crisis

These events were chosen for their distinct causes and impactful characteristics, as well as their occurrence within the study's timeline. While other events such as the 9/11 attacks and the 2011 U.S. credit

rating downgrade also affected the markets, they were relatively short-lived and had less profound impacts compared to the selected three. Additionally, significant events such as the 2018 cryptocurrency crash and the 2020 recession triggered by COVID-19 lockdowns are also noteworthy but fall outside the study's timeframe.

All these market crashes are visible in aggregated quarterly price of S&P500 index, as well as volume and volatility in Figure 24. It is also visible that during the 2008 crisis, S&P volume and volatility values are higher than the other two market crashes, highlighting the significance and impact of 2008 crisis.

DotCom Crisis: 2000s

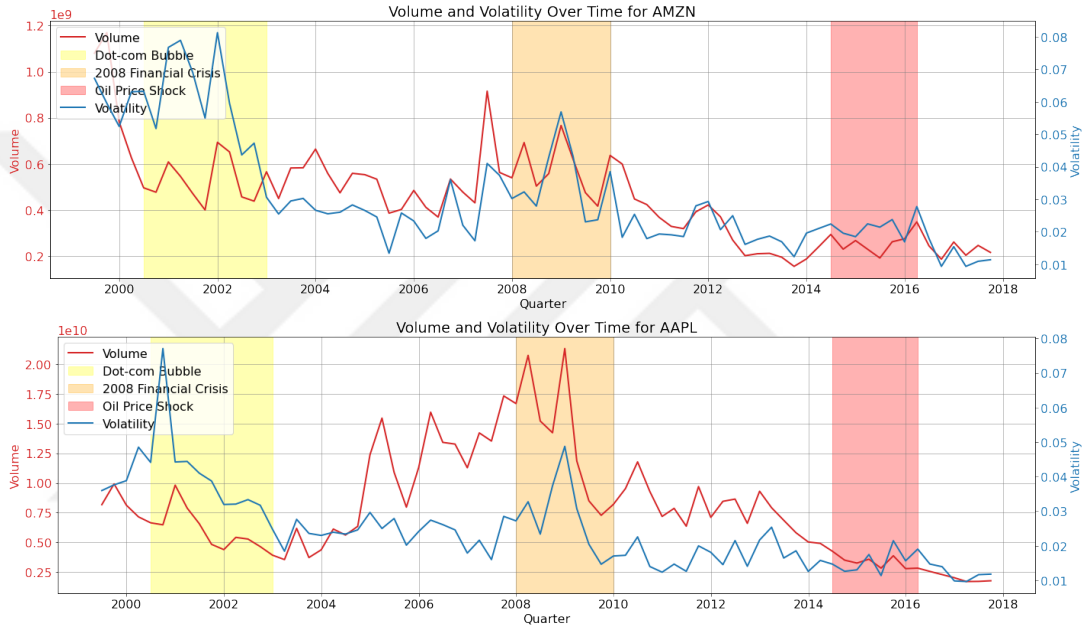


Figure 25: Popular tech companies display increase in volume and volatility during DotCom bubble

The DotCom bubble, a significant market event, began in the 1990s, fueled by the rapid rise of internet and tech startups. This era, catalyzed by the advent of the World Wide Web in the early 90s, saw numerous ventures in the U.S. markets, characterized by grand promises but often underpinned by weak business plans. These companies quickly gained value and frequently went public through Initial Public Offerings (IPOs), creating a frenzied investment environment that significantly inflated their market valuations. The burst of the DotCom bubble, typically recognized in March 2000, led to the closure of numerous tech companies and severe market disruptions. Even established firms faced significant challenges, with notable increases in volatility and trading volume, as illustrated in Figure 25.

Financial Crisis: 2008s

The 2008 financial crisis, considered one of the most severe economic downturns of the 21st century, had far-reaching global impacts. It originated from the collapse of the housing bubble in the United States, fueled by years of low-interest rates and easy credit. Financial institutions had heavily invested in mortgage-backed securities, which plummeted in value when the housing market crashed. This led

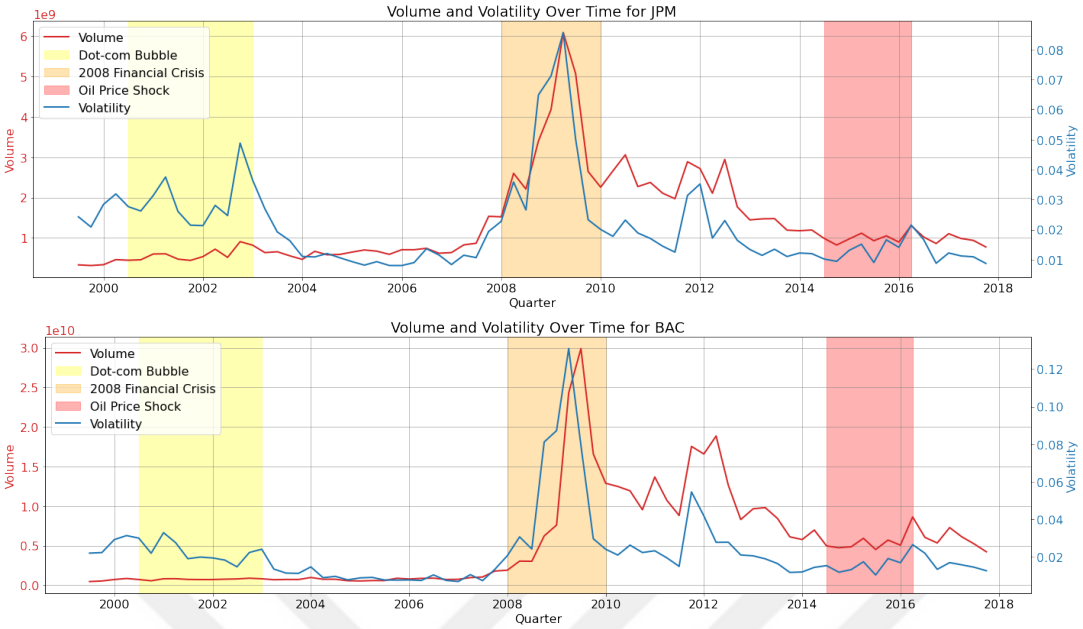


Figure 26: Biggest oil companies show similar hikes during oil crises

to the failure of key investment banks, notably Lehman Brothers and Bear Stearns, triggering a panic that spread to the global financial system. The crisis peaked in September 2008, resulting in significant government interventions worldwide, including bailouts of banks and other financial institutions. It is visible that not only financial instutitions, tech and oil companies were also affected by this crisis as shown in Figures 25, 26 and 27. It is also visible that large banks such as JP Morgan Chase and Bank of America stocks are more affected by 2011 U.S. credit rating downgrade than DotCom bubble or oil crisis as shown in Figure 26.

Oil Price Crisis: 2015s

Final event focused in the study is the 2014-2016 world oil price crisis, which was a significant economic event characterized by a sharp decline in oil prices globally. The crisis began in mid-2014 when a combination of factors, including increased oil production in the U.S, and OPEC's decision to maintain production levels despite the surplus, led to an oversupply in the market. This surplus was further exacerbated by slowing demand, particularly from emerging economies. The impact was a dramatic drop in oil prices, from over $100 per barrel to below $30 by early 2016. This plunge severely affected oil-dependent economies and industries, leading to budget deficits in oil-producing countries and forcing a reevaluation of energy policies worldwide. The crisis eventually stabilized by the end of 2016, as production cuts were implemented and global demand began to recover, helping to gradually increase prices and restore a degree of market equilibrium. Large corporations like Chevron and Conoco Philips stocks have responded to the event with hikes in volume and volatility as shown in Figure 27.

During experimentation, in order to get a cleaner picture, stocks within certain industries are grouped according to their relevance to each identified market event. The grouping methodology and the rationale for selecting these industries in relation to each event are elaborated in the Table 8.

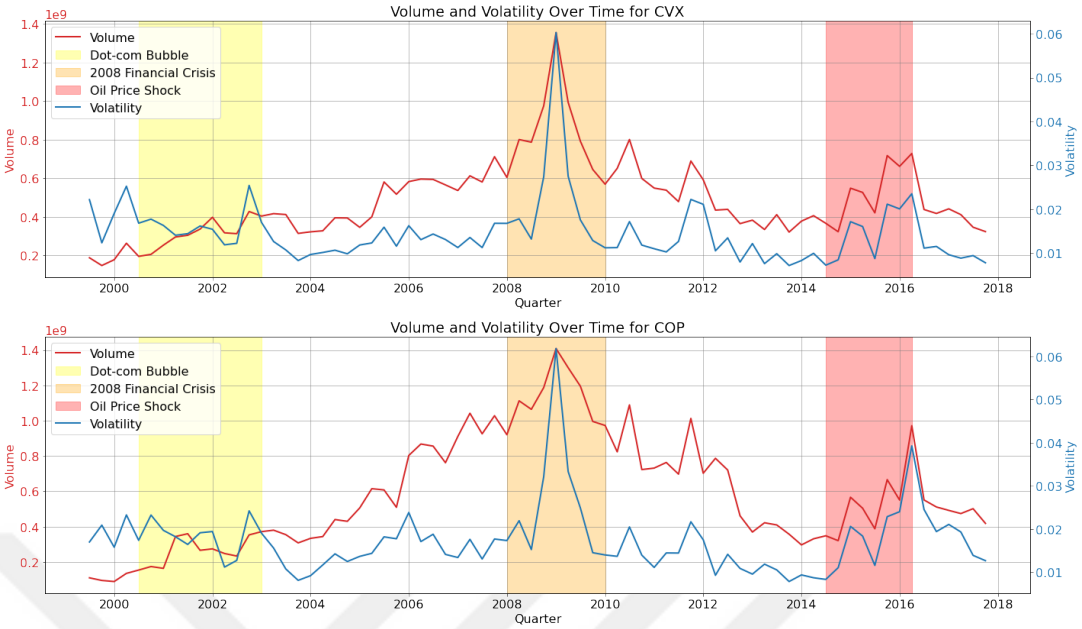


Figure 27: Biggest oil companies show similar hikes during oil crises

This phase of the study focused on assessing the impacts of selected major market events, defining the time boundaries of these events, and supporting claims with volume and volatility data of largest companies in each industry. It is visible that each of the market crashes are differing in magnitude, duration, and impact, highlighted distinct sectoral vulnerabilities. The 2008 financial crisis had widespread effects but predominantly impacted the financial services sector. In contrast, the DotCom crisis primarily affected technology companies, with minimal impact on the oil and financial sector. Conversely, the 2014-2016 oil price crisis significantly impacted oil companies but had a negligible effect on technology and financial institutions. These observations guide the study's direction and emphasize the importance of sectoral analysis in understanding the dynamics of market crashes, which will be further explored in the subsequent experiments.

4.2 Characteristics of the Dynamic Network and its Examination

The preceding sections detailed the creation of a dynamic bipartite network enriched with comprehensive features, alongside an analysis of the three largest market events. This section shifts focus to inspecting investor actions, calculating key network metrics of overall market, within specific sectors, and examining their temporal evolution, particularly during the market crashes previously discussed. The analysis aims to uncover how these interactions and network characteristics vary across different time periods, offering insights into sectoral responses during significant market disruptions.

Table 8: Summary of Major Market Crashes and Their Key Sectors

Event	**DotCom Crisis**	**Financial Crisis**	**Oil Crisis**
Period	2000 Q2 - 2002 Q4	2007 Q4 - 2009 Q4	2014 Q2 - 2016 Q1
Sector	Services	Finance, Insurance & Real Estate	Transportation, Communications, Electric Gas & Sanitary Service
Description	Characterized by a speculative bubble in technology and internet startups, this crisis led to inflated valuations and a sharp market correction in the early 2000s	Originated from the collapse of housing market and risky mortgage lending practices, severely impacted finance sector and lead to a global recession	Drastic drop in oil prices (around %70) due to oversupply and reduced demand, affected energy sector, particularly companies in oil extraction and distribution
Industries	Services-Computer Programming, Prepackaged Software, Computer Integrated Systems Design, Computer Processing & Data Preparation, Computer Programming, Data Processing, Etc.	Mortgage Bankers & Loan Correspondents, Real Estate, Real Estate Operators (No Developers) & Lessors, Operators of Non-residential Buildings, Apartment Buildings, Lessors of Real Property, NEC, Real Estate Agents & Managers, Real Estate Dealers, Land Subdividers & Developers, Real Estate Investment Trusts	Crude Petroleum & Natural Gas, Drilling Oil & Gas Wells, Oil & Gas Field Exploration Services, Oil & Gas Field Services, NEC

4.2.1 Investor Filing Delays Analysis

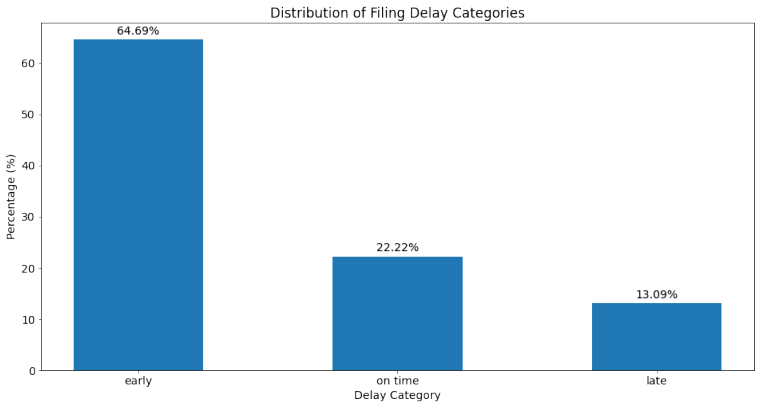


Figure 28: Most of the filings are submitted before 45 day delay

As it was explained in previous sections, filing delays of 13F filings are calculated and grouped based on the requirement for filing to be submitted within 45 days [59]. The distribution of these filing delays, as depicted in Figure 28, shows filings are mostly submitted either early or on time. Around 13% of filings are submitted late, possibly due to requests for confidential treatment delaying disclosure, although the specific reasons for these delays are not explicitly known [59].

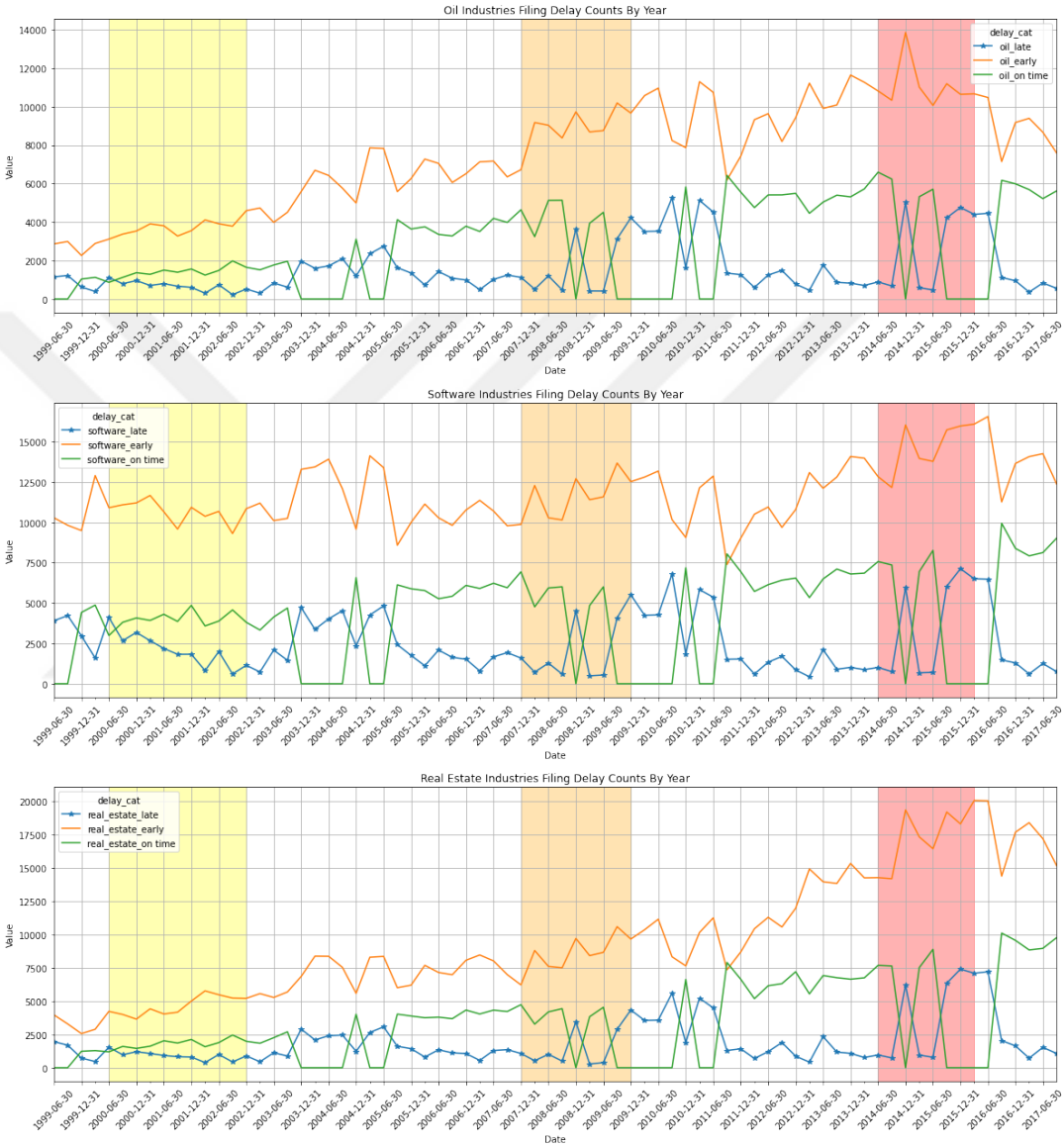


Figure 29: Filing Delay Category Counts By Industry Over Time

In examining the distribution of filing delays across industries, Figure 29 reveals similar trends for each industry. This similarity arises because 13F filings represent an investor's entire portfolio, submitted as a collective unit. Consequently, the filing delays tend to be uniform across different stocks within the same filing. However, notable exceptions are observed during specific market events. For instance, during the DotCom crisis, the software industry exhibits a subtle decrease in late filing submissions,

a trend distinct from other industries. Similarly, during the oil crisis, there is a noticeable decline in early filings for the oil industry, differing from the patterns seen in the other two industry groups.

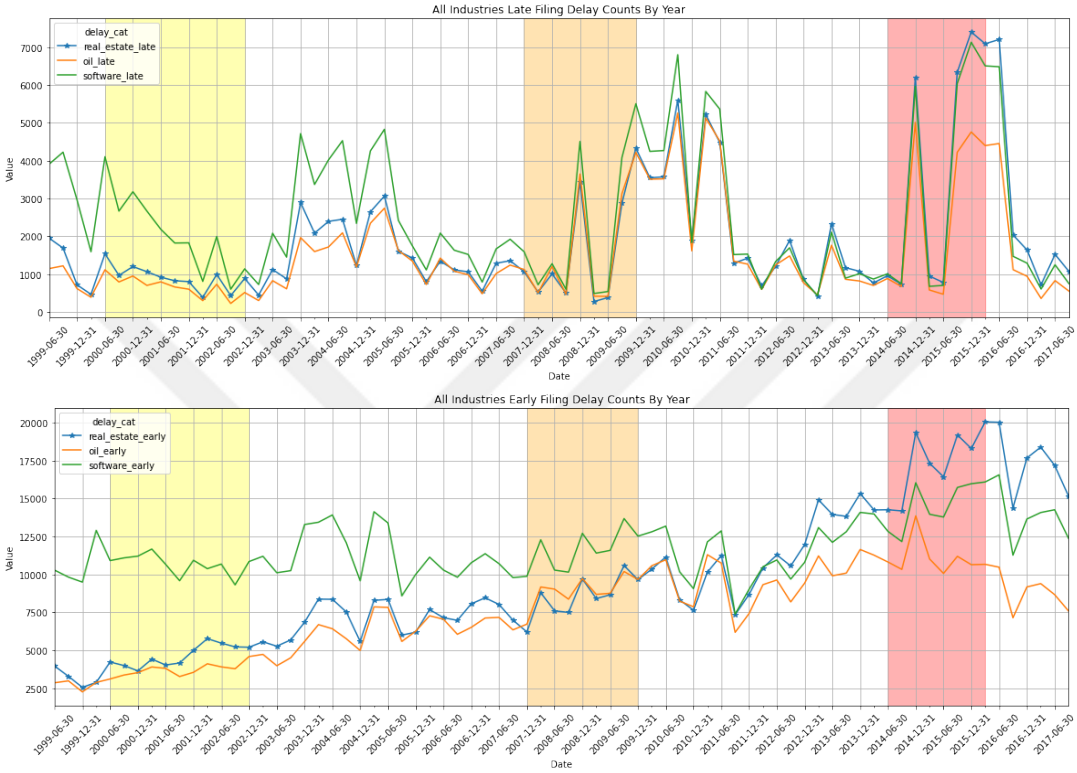


Figure 30: All industries have similar delay counts since investors submit all stocks in same filings

Although the count varies, Figure 30 illustrates that all three sectors exhibit remarkably similar behaviors, highlighting the uniformity across different industries as explained earlier. This observation underscores the subtle yet crucial differences specifically observed during the DotCom and oil crises, emphasizing their impact on industry-specific filing behaviors.

Figure 31, examining the percentage distribution of filing delays offers insights into the overall patterns during market crises. For each of the three major crises, a notable trend emerges: in the aftermath of the crisis, there is a significant drop in early filings and a spike in late filings. Conversely, the period leading up to a crisis often shows a decrease in late filings compared to other times. Considering that late filings constitute only about 13% of the total delay categories, as shown in Figure 28, this shift in behavior around market events suggests that late filings can serve as an indicator of market turbulence or investor caution during these periods.

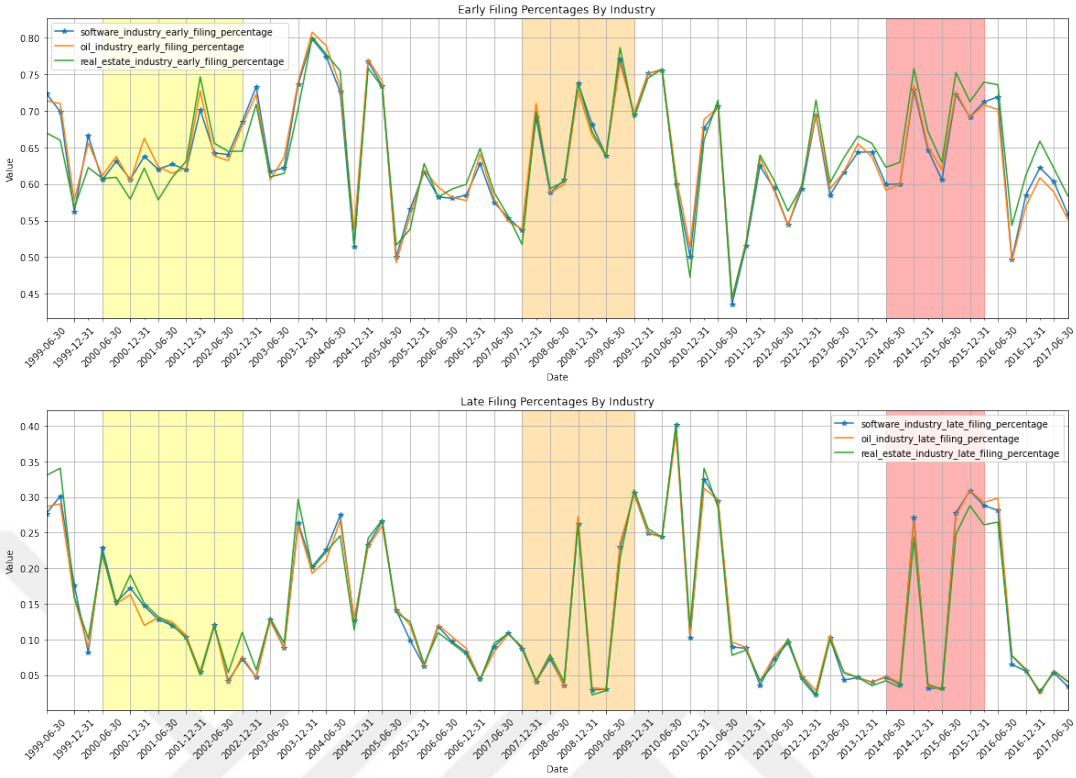


Figure 31: Hikes in percentage of late filings after market crashes are visible

The findings indicate that during the initial quarters of the DotCom and oil crises, a divergence in interaction begins to emerge between investors with corresponding industry groups. This finding could suggest that investors respond to such events by submitting 13F filings as late as possible, pointing the potential of using late classified filing delays as definitive features in analysis.

4.2.2 Significant Portfolio Changes of Investors

When the overall interactions within the network are filtered to focus specifically on significant sales and buys by investors, the resulting trends displayed in Figure 32 exhibit similarities with the volume and volatility graph previously referenced in Figure 20. These observations suggest a correlation between the investment behaviors of significant sales and buys and the overall market dynamics captured in the volume and volatility data.

Upon examining the major sales and purchases for each industry, distinct trends are observed in Figure 33 around the major market events. During the Dotcom crisis, there is a notable decrease in significant buys preceding the crisis, accompanied by an increase in significant sales, particularly in the software industry where sales sharply rise. In the 2008 financial crisis, a similar pattern emerges with a significant increase in sales in the software industry. Intriguingly, despite being central to the crisis, the real estate sector shows an increase in purchases. In contrast, during the Oil crisis, unlike the other events, the sales in all three sectors, which had been stable previously, start to increase, albeit with a slight lag, indicating a delayed response to the crisis conditions.

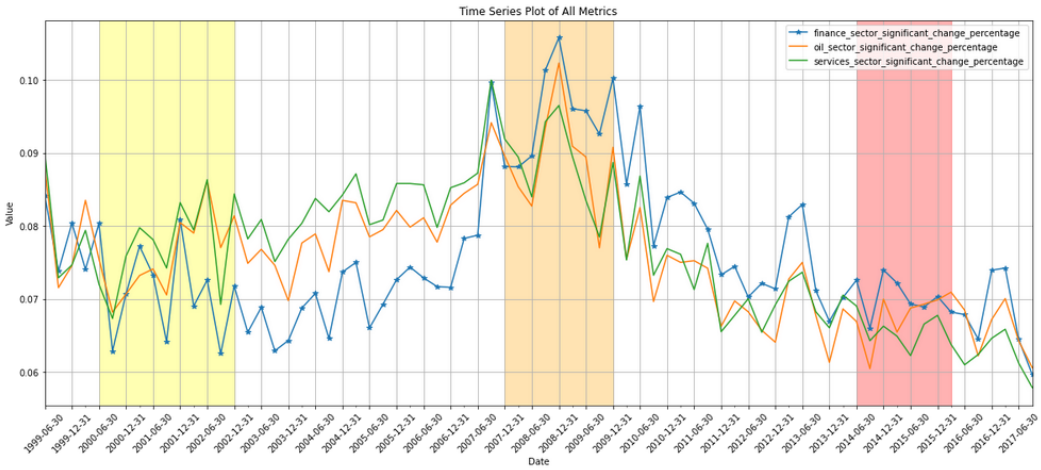


Figure 32: Significant change (Significant Buys and Sales) counts resembles S&P500 volume and volatility graph, highlighting the impact of 2008 crisis

When the counts of large (and significant) buy and sale transactions are analyzed individually, as illustrated in Figure 34, each market crash becomes discernible through a distinctive pattern characterized by a decrease in the number of notable buys and a concurrent rise in the number of notable sale transactions.

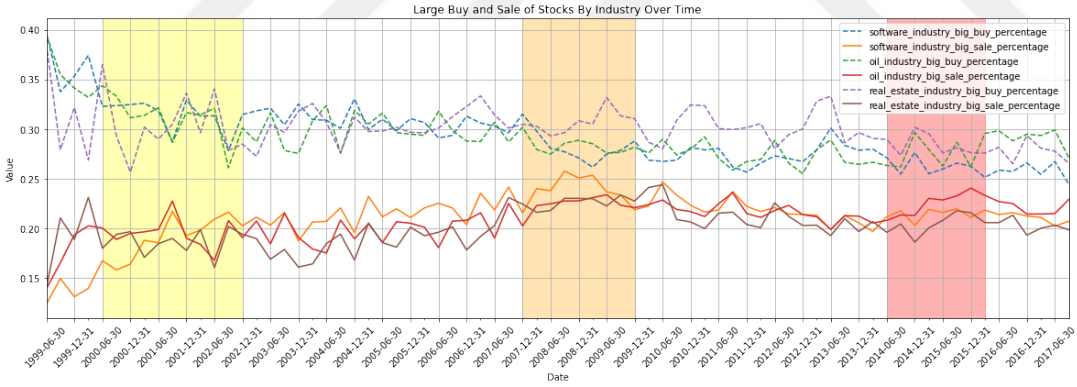


Figure 33: Large Buy and Sale counts gets closer during market crashes

While the count of large buys for each related industry and market crash does not have a visible pattern, large sales have all increased before and during of crashes. Notably, in the cases of the DotCom crisis and the 2008 financial crisis, the sales count demonstrated a gradual buildup almost a year before the respective crashes. In contrast, the Oil crisis displayed a distinct behavior, with a lag in the increase, becoming evident after a two-quarter period. This observation suggests that while the DotCom crisis and the 2008 crisis may have had internal causes, the Oil crisis appears to be more influenced by global supply and demand dynamics, indicating external roots. This may also imply that investors are more adept at taking precautions against a market bubble or potential crisis when it originates from internal factors, rather than being driven by global political events.

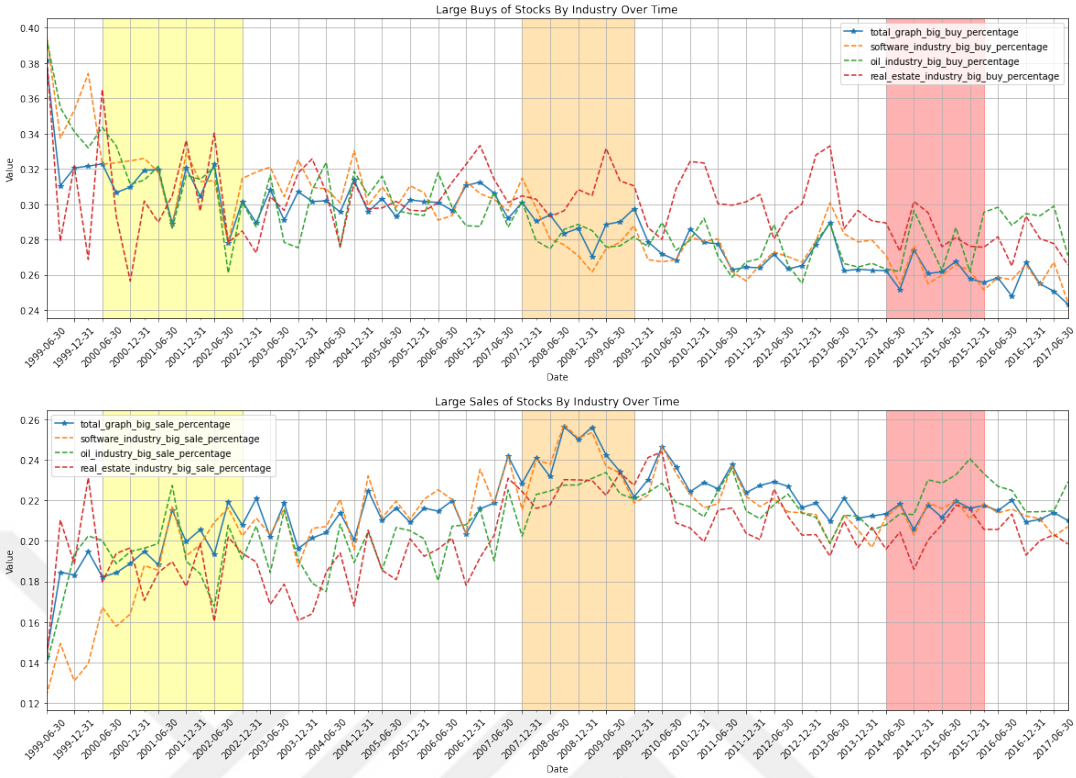


Figure 34: Hikes in count of large sales are lagged during the Oil Crisis

The findings from the analysis suggest that the processed data correlates to a certain extent with market events, yet it does not significantly indicate or highlight distinct changes. Given the high-level approach of the study that focuses on overall market trends, even subtle differences can provide meaningful evidence. These observations imply that features such as filing delays and interaction counts could potentially serve as useful indicators for monitoring market crashes.

4.2.3 Network Metrics

In this subsection, various network metrics are evaluated for each quarter, similar to the previous sections. This approach, focusing on collective behavior rather than individual interactions, is beneficial for understanding the changes of the overall structure of the network.

Network Density

It is a measure of how closely knit a network is. It is calculated as the ratio of the number of existing edges to the maximum possible number of edges in the network, using the formula:

$$\text{Bipartite Density} = \frac{|E|}{|V_1||V_2|}$$

where $|E|$ represents the number of edges in the graph, and $|V_1|$ and $|V_2|$ are number of vertices for the first and second partition, which are stocks and investors in this case.

High network density suggests a tightly-knit network with many connections among nodes, indicative of strong interconnectivity. Conversely, low density implies a sparser network, where nodes are less interconnected.

In context of the bipartite interaction graph, generated for this study, network density simply indicates variability in market activity. Figure 35 displays a drop of density, followed by continuous increase in all of these three market crashes.

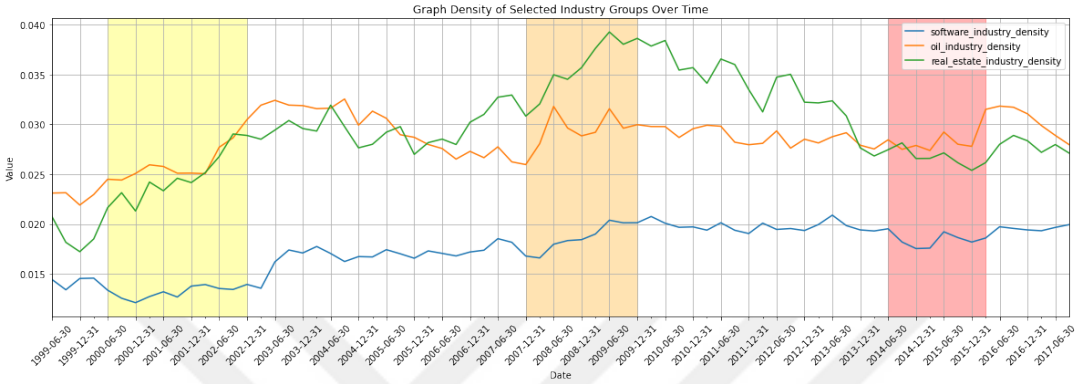


Figure 35: Slight increases in density during crashes are visible

Average Degree

It represents the average number of connections each node has. Calculated with the formula:

$$\text{Average Degree for Investors} = \frac{\sum_{i=1}^{N_{\text{investors}}} \text{Degree}(i)}{N_{\text{investors}}}$$

where i is an investor node, $Degree(i)$ is the total number of edges each i has, and $N_{\text{investors}}$ is the number of investor nodes in graph.

In the context of a stock-investor network, degree indicates, on average, how many stocks each investor buy or sell. A higher average degree can signify diversified market with widespread investment across various stocks, while a lower average degree might point to a more concentrated market with investors focusing on fewer stocks.

Figure 36 shows that during the DotCom crisis, a notable and continuous decline in the average degree was observed in the software industry, indicating that investors possibly focused on a few large companies, avoiding risk. The oil industry showed a similar pattern of decline during the oil crisis, differing from other industries. In contrast, during the 2008 financial crisis, all three industries demonstrated no visible changes in average degree. Specifically, the real estate sector continued its general upward trend, while others remained stable.

Clustering Coefficient

The clustering coefficient in a bipartite graph measures the degree to which nodes in one set (e.g., investors) tend to cluster around common nodes in the other set (e.g., stocks).

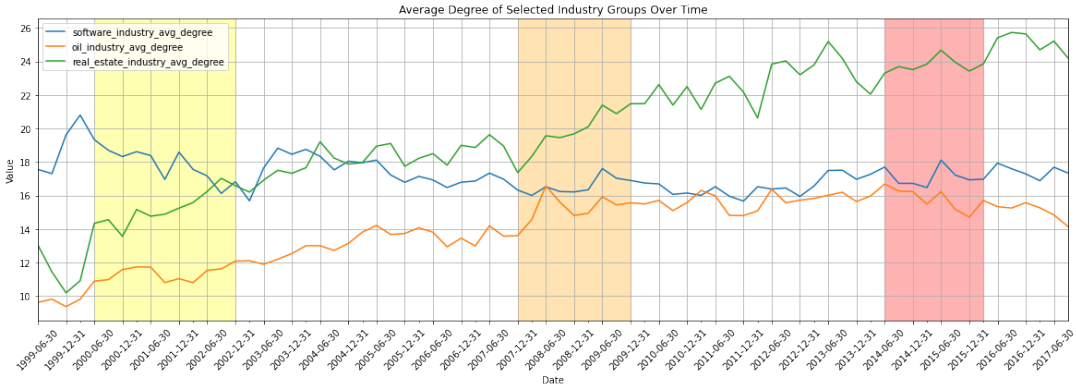


Figure 36: During DotCom crisis, software industry average degree visibly dropped

The bipartite clustering coefficient C_u of node u in a bipartite network, measuring the tightness of connections between its neighbors, is defined as:

$$C_u = \frac{\sum_{v \in N(N(u))} c_{uv}}{|N(N(u))|}$$

where $N(N(u))$ represents the set of second-order neighbors (neighbors of u's neighbors, excluding u itself), c_{uv} is the pairwise clustering coefficient between u and v, and $|N(N(u))|$ is the total number of second-order neighbors. Averaging clustering coefficient across all nodes in graph provides general sense of clustering in the graph.

In the context of study, it can indicate the tendency of investors to share similar investment profiles or the prevalence of certain stocks in multiple portfolios. A higher clustering coefficient suggests a market with common investment patterns or popular stocks, while a lower value indicates more diverse investment behaviors.

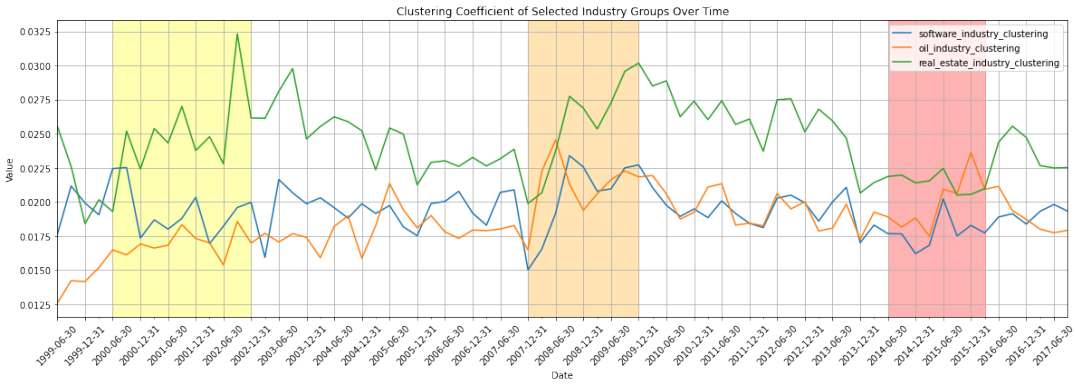


Figure 37: During crises, especially 2008, clustering coefficient significantly increases

In Figure 37, it is visible that the clustering coefficient in the network experienced an increase during highlighted market crashes. Most notably, right after the start of 2008 crisis, there is a sharp increase in clustering coefficient, for each industry group, especially real estate. Similarly, during the oil crisis,

there is a delayed but notable uptick in the clustering coefficient, aligning with similar trends to other previously observed metrics, with oil industry having higher increase compared to other two industry groups. Additionally, in contrast, during the DotCom crisis, the software industry displays a decrease in clustering coefficient. This indicates that during DotCom crisis, investors may not have a consensus on stocks, thus have more diverse focuses in software industry.

This subsection presents that also graph metrics can be used to explain investor behaviors during market crashes other than large buy-sell counts or filing delays. Three common metrics: density, degree, and clustering coefficient, are utilized to identify significant changes in investor behavior during market crashes. As also another similar study implied, clustering coefficient highlights more meaningful changes especially during market crashes, than density, although it was applied on an information network [142].

Findings highlighted shifts in metrics in context of market crashes, with most visual is being shifts of clustering coefficient in the beginning of 2008 crisis. The study also found that interactions with the software industry tend to behave differently from interactions with other industries. Also, across all three metrics, every investor demonstrates similar behavior during the 2008 crisis.

These findings highlight the potential of using network metrics to understand investor behaviors during crashes as intended. Although it was expected to have significant shifts in metrics, comparison of industries also reveals distinct characteristics of crashes.

4.2.4 Network Motif Frequencies

The examinations conducted so far have demonstrated that certain network-centric features and edge attributes over time, particularly the clustering coefficient, can serve as useful metrics to understand shifts in investor behaviors. However, in this study, to derive more fine-grained insights, exploring network motifs was deemed more logical and reasonable. As explained in the Methodology chapter, the rationale behind this decision lies in the fact that interactions, supported by rich features, can be examined through motifs, which act as queries, offering highly explainable findings.

To achieve this, each designed motif are expressed with a diagram and an explanatory paragraph. Similarly, the counts for each motif, for each related industry groups defined in previous sections, are visualized and interpreted.

Motif 1

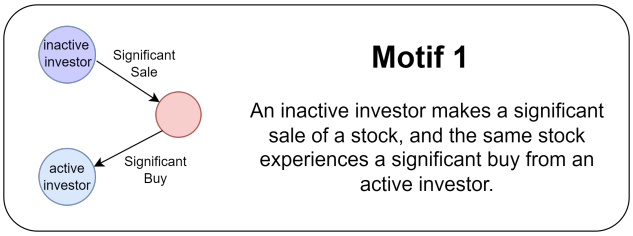


Figure 38: This motif rule focuses on occurrences from inactive to active investors

First motif aims to capture occurrences where a stock flows thorough an inactive classified investor to active classified investor as depicted in the Figure 38. In the Figure 39, it is visible that certain periods has increase in this pattern, especially right before and during DotCom crisis for software industry stock, during 2008 crisis for real estate industry. General trend is in increase after 2012, most significant being oil industry.

This finding highlights a visible correlation between this motif and market crash periods, can be interpreted as, less active or small investors tend to increase significant sale of stocks while active ones buying, potential shift of ownership to larger or more active investors during market crashes. It is worth to note that investor classifications in these motifs are based solely on 13F filing activities, as detailed in the methodology chapter.

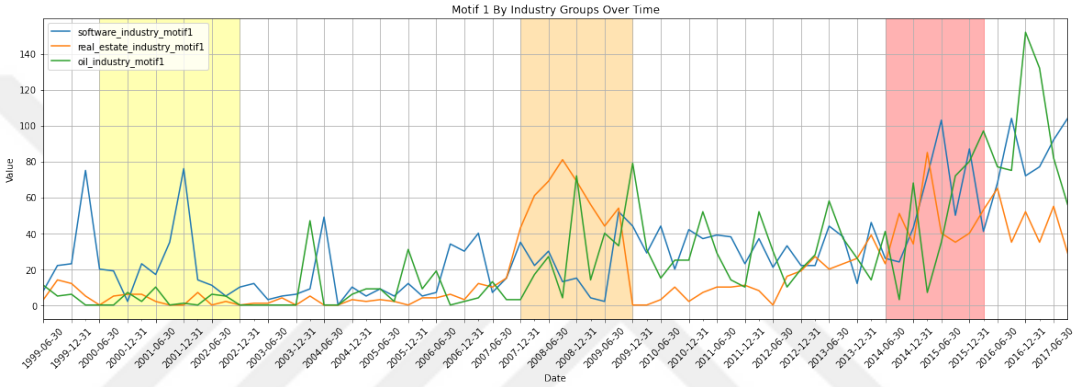


Figure 39: During market downturns, motif counts for corresponding industries visibly increased

Motif 2

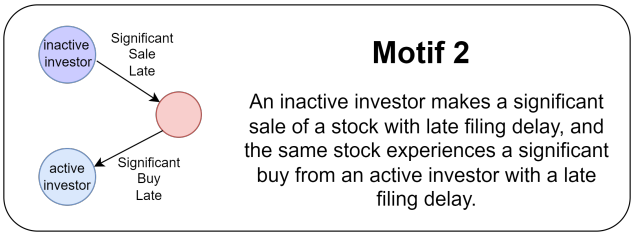


Figure 40: This motif also filters late submitted filings on top of the first motif

Second motif is similar to the first one, although it also considers delay category as late, shown in the Figure 40. This also highlights the finding even more visibly, by displaying almost all motif counts occurred at the late quarters of each of the market crash in the Figure 41, indicating a potential metric to highlight the end of the market downturn.

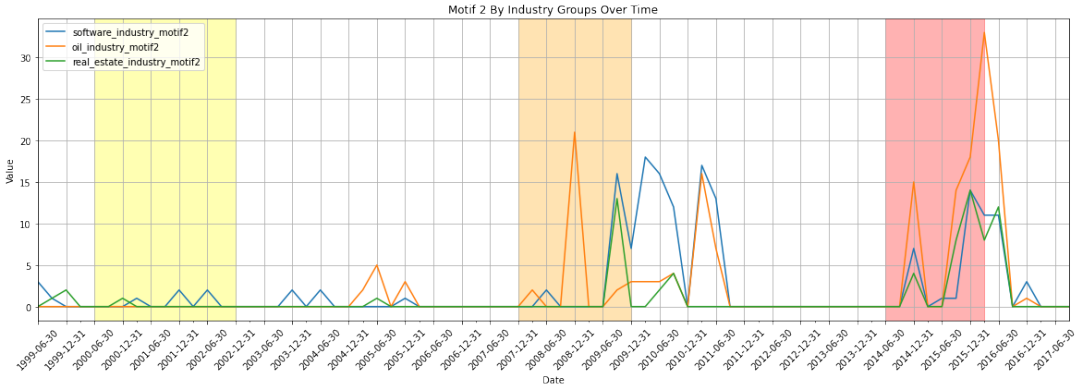


Figure 41: Almost all motifs occurred at the late quarters of market crashes

Motif 3

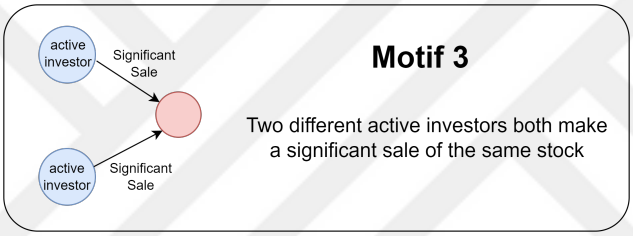


Figure 42: This motif captures overlapping sales of active investors

Third motif has a more simple pattern as explained in the Figure 42, which captures significant sales of two different active investors of the same stock within a quarter. As it shown in the Figure 43, it has not have any visible significance. However, each industry has increasing motif count at the beginning of each market crash. This may indicate that at the early stages of crashes, active investors quickly restructure their portfolio by selling same of their holdings.

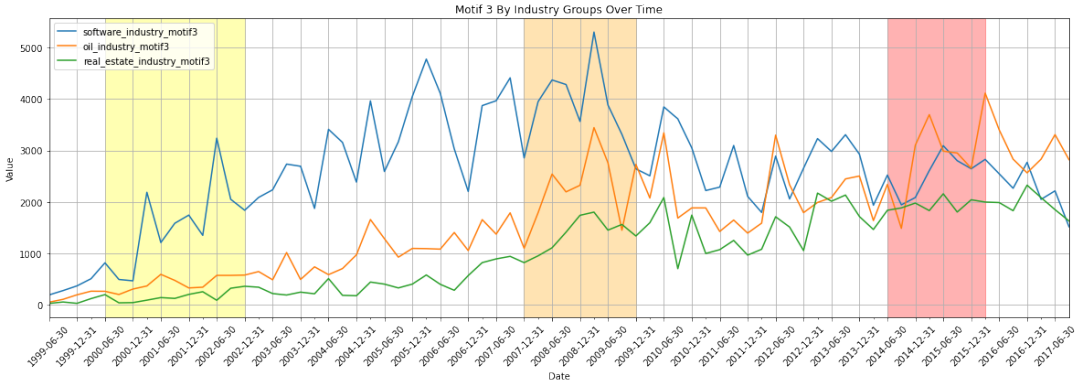


Figure 43: No significant changes, although, sales are increasing before market crashes as expected

Motif 4

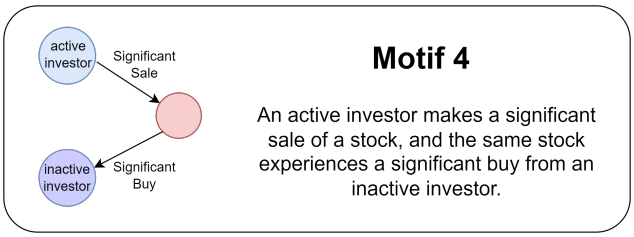


Figure 44: Motif 4 is an inverse version of motif 1 with a changed flow from active to inactive investors

Fourth motif is an inverse version of the first motif, to capture flow from active investor to inactive investors, as depicted in the Figure 44. Although, each related industry has increased number of motif counts in corresponding crashes, considering the first motif, this motif has less profound changes during market crashes as depicted in the Figure 45. However, it is important to note that counts for each quarter almost doubles in this motif, compared to the first motif.

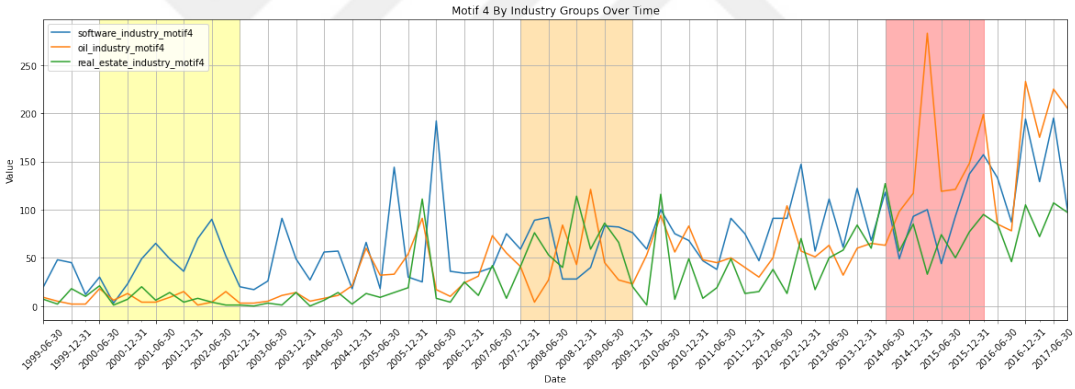


Figure 45: Not visible changes, but more counts overall compared to the first motif

Motif 5

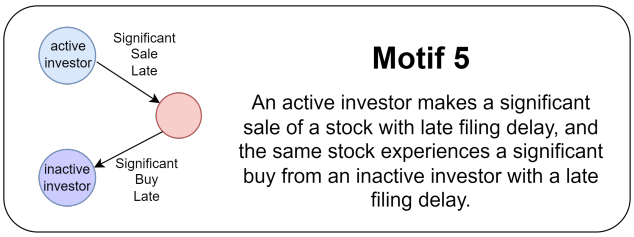


Figure 46: Motif 5 is an inverse version of motif 2 with a changed flow from active to inactive investors

Fifth motif is an inverse version of second motif, to capture flow from active investor to inactive investors, as depicted in the Figure 46. Similar with the second motif, this one also highlights increased activity between these two groups, especially using late filings as depicted in the Figure 47. It is also worth to note that both motif counts were almost zero during DotCom bubble crash periods.

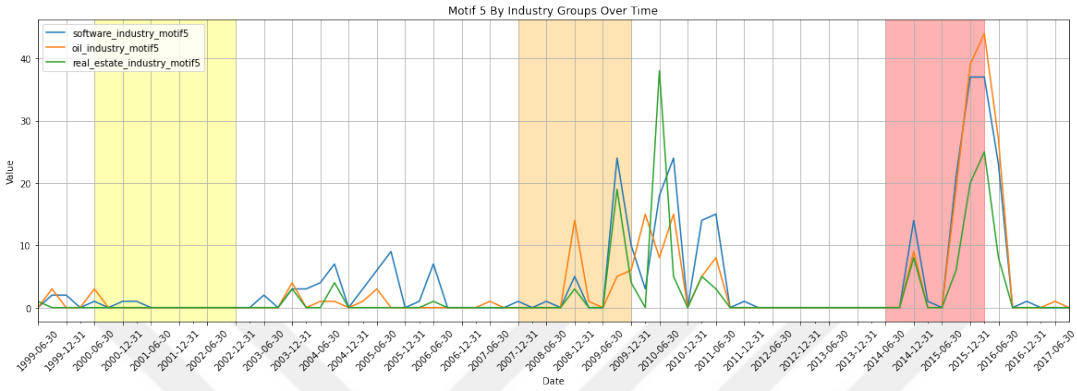


Figure 47: Similar to motif 2, motif 5 counts also has increased in the late periods of 2008 and oil crisis crashes

Motif 6

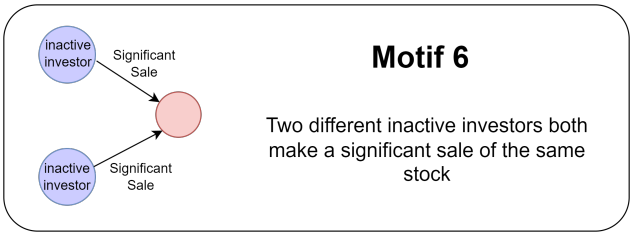


Figure 48: Motif 6 is similar to motif 3, although it counts significant sales of inactive investors

Sixth motif has same logic with the third motif, although this time, it counts significant sales of a same stock by two inactive classified investors, as depicted in the Figure 48. Compared to the third motif, this motif has significantly less number of occurrences over time. Although both of the motifs does not reveal any visible shifts over time, it emphasized that active investors might have a tendency to sell similar stocks, while inactive ones do not, displayed in the Figure 49. However, more in-depth analysis is required specifically to this subject, since significant sale or buy counts with investor activity scores are highly correlated. It is important to state that these motif counts and methodology is exploratory, to highlight potential findings for further analyses.

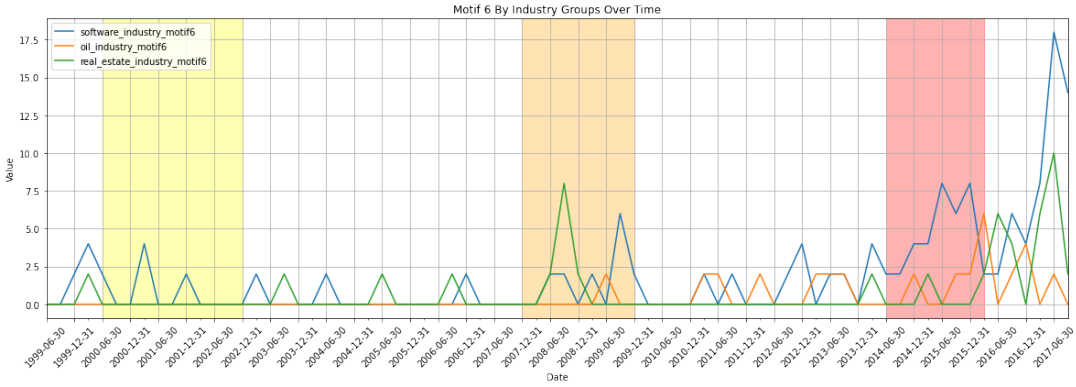


Figure 49: Limited number of motif counts because investor activities are based on activity scores, which are calculated depending on sale counts and significance

Motif 7

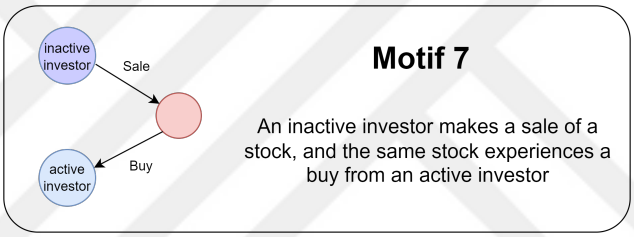


Figure 50: This motif captures any sale and buy edges available in the graph

Seventh motif counts patterns where an inactive investor sells a stock, and the same stock is bought by an active investor, as depicted in Figure 50. Although not significant changes during market crashes, motif counts increased significantly after early 2012s, may indicate increased market activities as plotted in the Figure 51.

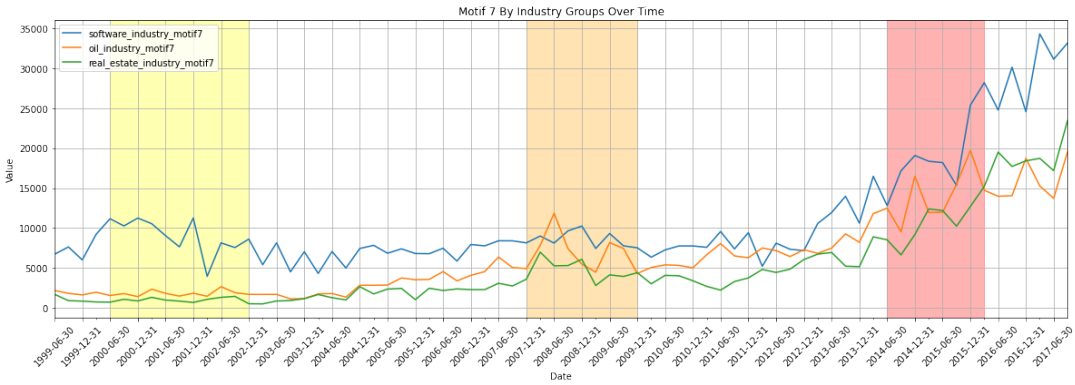


Figure 51: No significant shifts during market downturns, but there is a visible upward trend after late 2012s

Motif 8

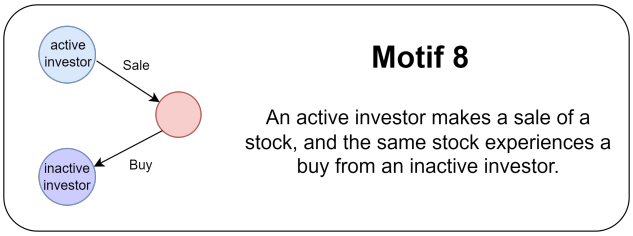


Figure 52: Motif 8 is similar to motif 7, however, the direction is inversed in this one

Similar to the seventh motif, eighth motif counts any sale of a stock by an active investor, while the same stock is bought by an inactive investor, which is illustrated in Figure 52. Similarly, this motif also has an increasing trend after early 2012s, with more counts compared to seventh motif as displayed in Figure 53.

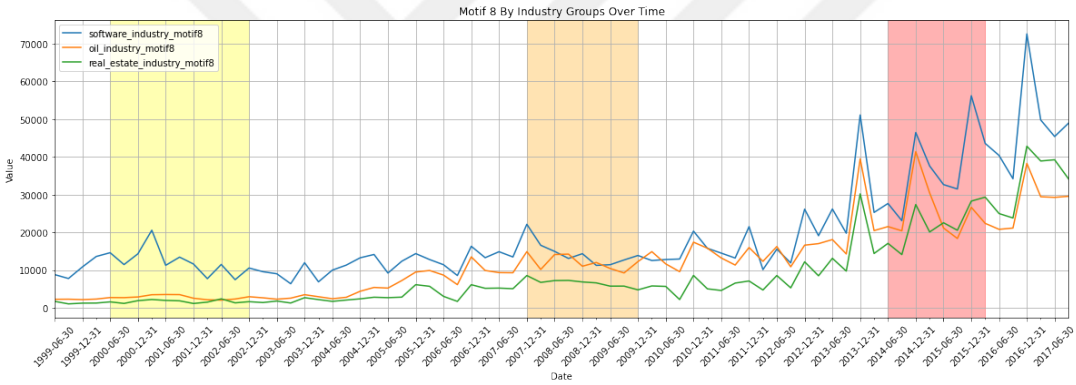


Figure 53: Similar with the seventh motif, with more counts and volatile increase in counts after 2012s

Motifs examined so far had different characteristics, counts and trends over time as expected. Since some of the motifs are inverse version of others, with slight changes in either direction or node type, comparison of their counts might provide additional insights on market behavior. Considering the motif counts over highlighted market crash periods, motif 1,2 and 5 has more visible changes during market crash periods as opposed to other motifs. While motif 1 counts flow of equities from inactive to active investors, motif 2 filters changes in stock holdings which are disclosed in public late. Almost all occurrences in motif 2 can be found in late periods of crashes, which might suggest these periods might be used by financial institutions to restructure their holdings without publicly disclosing their details, despite having a risk of paying fines. Similarly, motif 5, an inverse version of motif 2 with changed direction, highlights similar pattern as well. These motif counts over time highlights certain rules in motifs has high fluctuations or visible patterns during market crashes, indicating their potential for further applications for this domain.

Following stage of the study further examines different motifs, focusing stock characteristics of corresponding quarters. Considering previous efforts on combining stock price and volume data, these

values were aggregated and classified into categories of stock volume and volatility of corresponding quarters.

Motif 9 and 10

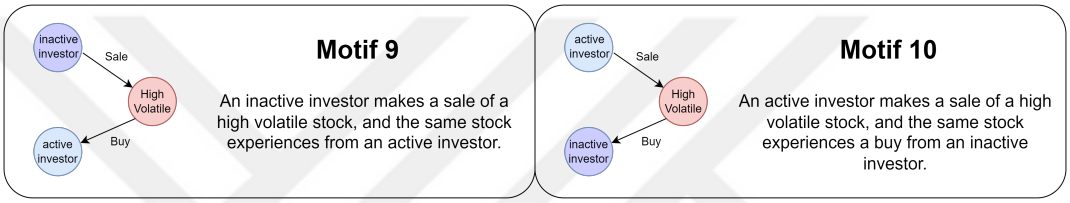


Figure 54: These motifs capture counts of flows of stocks that have high volatility in corresponding quarters

Motif 9 and 10 captures occurrences in the graph that are based on potential flow of funds of stocks that has high volatility in corresponding quarters, as summarized in Figure 54. These motifs identify any buying and selling activity between two investor types for highly volatile stocks within a quarter. However, it is crucial to recall that the graph does not account for insignificant changes between quarters. Figure 55 reveals that motif counts spiked during all market crashes, including the 2011 U.S. credit downgrade crisis, even though it was not specifically highlighted. In addition, each motif count peaked for its corresponding industry groups.

While it was observed in previous sections that during the market downturns, stocks tend to have higher volatility, the methodology employed in this study provided a decomposed view, allowing for the inspection of individual occurrences. It is important to note that this study served as an examination, and therefore, these occurrences were only analyzed in terms of counts, without delving into individual cases.

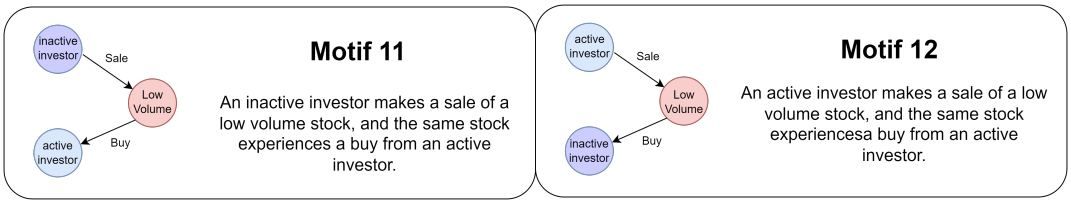


Figure 56: These motifs capture counts of flows of stocks that have low trading volume in corresponding quarters

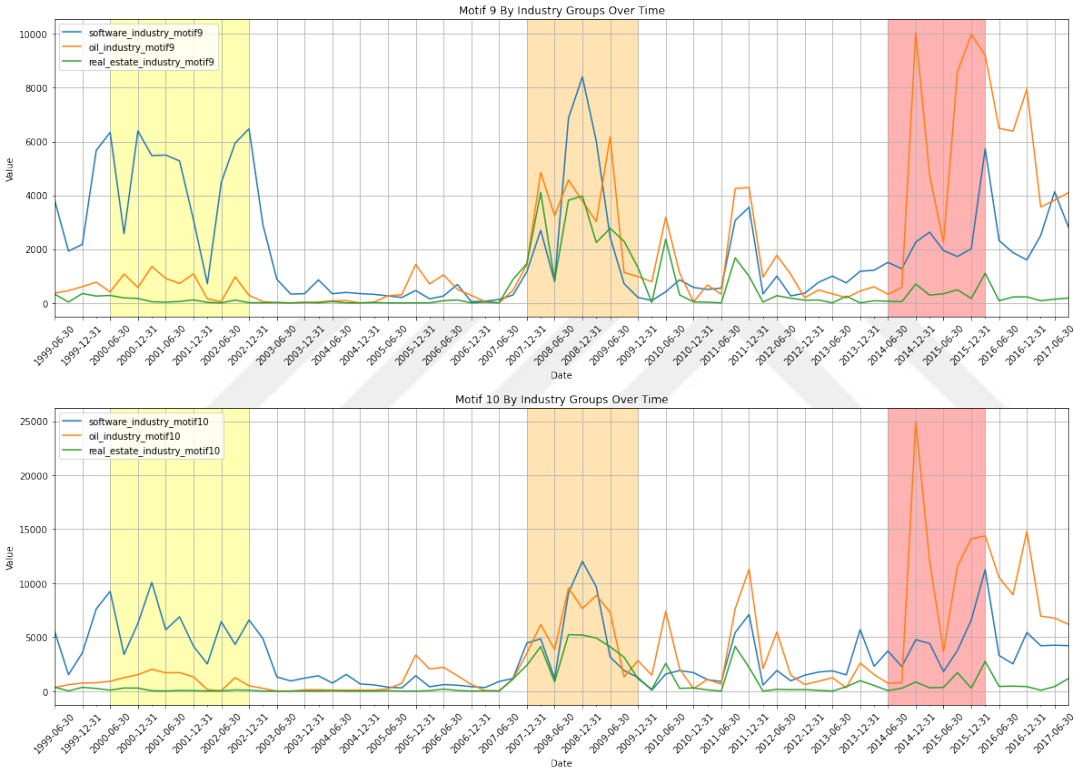


Figure 55: Motif 9 and 10 has similar trends since overall market volatility was also increased during these periods.

Motif 11 and 12

Motif 11 and 12 captures occurrences in graph that are based on potential flow of funds of stocks that has low trading volume in corresponding quarters, as summarized in Figure 56. Similar with previous motifs, although there are no visible pattern other than somehow inverted display during 2008 crisis, there is an increasing trend of counts after 2012s, as depicted in Figure 57, which was also highlighted in overall market trading volume in Figure 20.

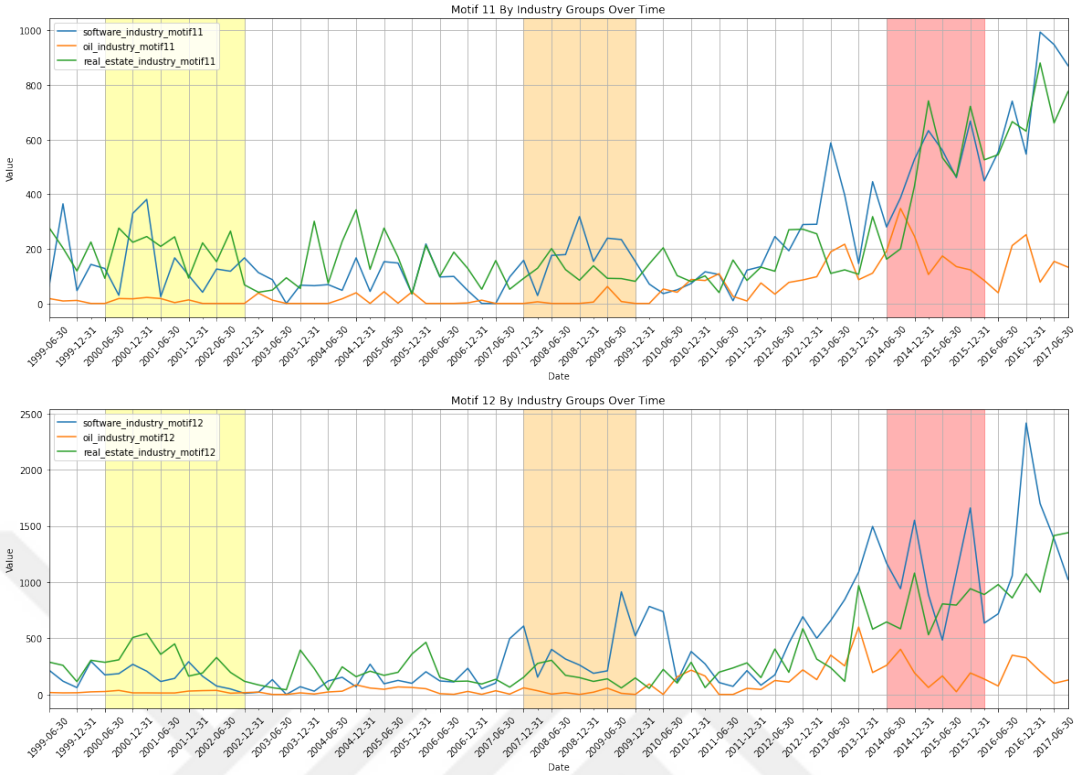


Figure 57: Although varies in counts, motif 11 and 12 has somehow inverted display during 2008 crisis

In the final part of the experiments, the activities of grouped investors were visualized and discussed using motif counts. The findings clearly indicate that specific occurrences are particularly observed during market crash periods. However, this approach has two major limitations. Firstly, since investor types are derived from market activities, it is highly dependent on the sale and buy counts and magnitudes, weakening the strength of the findings. Secondly, when stocks are filtered based on volume and volatility, i.e., 9, 10, 11, and 12 motifs, the graphs resemble the S&P 500 market volume and volatility graph. This also limits the significance of findings, since volume and volatility are already widely used indicators.

Nevertheless, the strength of the motif count approach lies in its ability to decompose the market from the desired perspective by visualizing the frequencies of occurrences of explainable specific cases. This was the main emphasis of the study. With this methodology applied in the study, an expert in the field of economics or finance can create and interpret motif rules with highly specific contexts for the phenomena they want to examine. Furthermore, since the methodology captures individual cases alongside motif counts, a more in-depth analysis on captured motifs may identify entities and conditions that contribute to these trends. However, since the study focused on high-level industry-wise behaviors during market crashes, captured individual motifs are not examined, and only temporal changes in motif counts were analyzed.

CHAPTER 5

DISCUSSIONS AND CONCLUSIONS

In this final chapter, the study summarizes its findings on investor behaviors in the U.S. stock markets during major events, discusses limitations and suggests potential future directions. In the Discussion section, the findings interpreted, and the applied methodology's strengths and weaknesses are discussed. In the Conclusions section, aligning with the initial goals, the chapter concludes the steps taken and results achieved for each objective. And finally, the Future Works section outlines potential directions for improving the study's benefits for regulatory bodies and investment researchers.

5.1 Discussions

There are various approaches in literature to create networks of actors in the stock market and apply graph algorithms. These may include price correlation networks of stocks, co-investment networks of investors, money flow, or even the educational background of board members. In this study, edges were not constructed based on assumptions or certain rules, but rather directly constructed based on the information extracted from the 13F filings. This approach minimizes information loss and highlights that the findings were directly derived from the official documents.

The inherent structure of the graph derived from 13F filings is bipartite, involving two key entities for each row: the investor (CIK) and the stock (CUSIP). Bipartite networks are effective at representing binary relationships, but they have limitations in capturing diverse topological patterns. Networks with a single node type, on the other hand, can be used for a wider variety of analyses. Bipartite projection is a common method for converting a bipartite graph into two separate networks, one for each node type. This approach can be useful for certain analyses, but it can also lead to a significant information loss on the overlapped edges during projection. In this study, the focus was on the rich node and edge attributes that were created from the 13F filings along with supplementary data. Therefore, bipartite projection was not used, as it would have resulted in potential misinterpretations.

The second objective, examining market crashes during the data period, centered on three significant occurrences. To identify relevant industry groups for each of these occurrences, literature and public sources are examined. While this approach was sufficient given the magnitude of the events, a more precise, accurate, and measurable selection of industry groups could have been achieved by conducting analyses based on price, volume, and volatility for each group of stocks.

Regarding the third objective, buy and sell counts behaved as anticipated, offering limited novel insights. An alternative approach could involve inverting the perspective when constructing buy and

sell edges. This would be attainable by focusing on stocks that investors retained during downturns, potentially revealing "safe haven" assets.

Filing delays, on the other hand, yielded more informative findings, particularly concerning investor behavior during and after crashes. However, the utility of this attribute was hindered by overlaps, since filings encompass all stocks held within a quarter. This systemic correlation underlying the filing delay and other variables in graph weakened the value of findings and usability of the filing delays as a feature for potential future use cases.

In line with findings from related literature, the clustering coefficient demonstrated more pronounced fluctuations during market crashes compared to other graph metrics. This observation highlighted how investor preferences visibly shift during such events, and displayed the potential of the clustering coefficient as a valuable indicator to be used for market monitoring systems.

The motif count over time approach, which had been used in similar cases with different data and graphs, was relatively innovative in this context. Due to the bipartite nature, the examination focused on attribute-oriented 3-node motifs, rather than topological or formation-based motifs. Although, this approach provided an advantage by allowing the counting of highly interpretable motifs, it also prevented the use of topology oriented one-mode graph motifs that may explain various phenomena specific to social networks.

In essence, the study's focus on 13F filings narrowed down the available data format, processing and analysis options. Furthermore, the study aimed to minimize information loss and directly utilize the information extracted from filings as much as possible. Although there are various ways and purposes for creating graphs in the context of the stock market, the chosen method focused on official documents, to ensure that the obtained findings are interpretable and reliable.

5.2 Conclusions

This analysis points towards a promising direction for predicting, analyzing and interpreting stock market crashes. By examining investor behavior within the network during critical periods, study provided valuable insights into the factors preceding and characterizing crashes, as well as how investors react to such events. This knowledge is crucial for both risk mitigation of policymakers and investment strategies of financial institutions. As it experienced in the 2008 crisis, identifying stock market bubbles and crashes presents lucrative opportunities and significant returns for informed investors. Similarly, anticipating the bubble beforehand would prevent millions of job losses, bankruptcies and losses.

This study provided two contributions: First, it illuminated how investor behavior transforms in the context of market crashes. Second, it tested the effectiveness of a network-oriented approach in this specific domain. To achieve these contributions, the study focused on four key objectives:

The first objective centered on transforming raw 13F filings and other related data into a dynamic bipartite graph that captures the interactions between public company stocks and investors. To achieve this initial objective, the study applied a set of rule-based techniques and statistical measures, with descriptions and explanations on findings. These steps combine and transform raw data into a meaningful bipartite graph with a slight loss of information.

For the second objective, the study aimed to provide context on market crashes that occurred between 1999 and 2017. This involved analyzing price, volume, and volatility of the popular market index, S&P 500, and selected stocks during this period. The analysis focused on three stock market crashes: the DotCom crisis in 2000s, the 2008 financial crisis, and the 2016 oil crisis. Each of these crises are examined by plotting volume and volatility of notable stocks in related industries, and by giving explanations about the origins of each downturn. Following these efforts, findings were summarized into a table, which outlines the respective periods and SIC classified industry groups associated with each market crash.

For the third objective, the study integrates the findings of the first two objectives. The output from the first objective, the stock-investor bipartite graph divided for each quarter, is combined with the results from the second objective, which highlighted crisis boundaries and identified industry groups. Consequently, edges of the created bipartite graph were grouped for each relevant industry, and evaluated within the context of the respective market crashes. For this objective, the edges of the graph for each quarter were examined and interpreted in terms of filing delays, significant portfolio changes, as well as network density, average degree, and clustering coefficient metrics. During the examination, it was observed that early filing submissions increased during the crisis periods, while late submissions peaked towards the following quarters after the end of crises. Additionally, significant portfolio changes over quarters revealed similarities with the S&P 500 volume graph, indicating an increase in significant sales and a decrease in buys in the relevant industries during downturn periods. Regarding network metrics, while network density and average degree did not convey any significant fluctuation, the clustering coefficient exhibited notable fluctuations, particularly at the onset of the 2008 crisis, across all three industries. All these findings highlighted the alignment between the created graph and the examined market crashes, suggesting that certain metrics could be valuable in understanding market downturns.

The fourth and final objective aims to offer more detailed insights by employing a network motif count approach for each quarterly stock-investor interaction graph. Similar to the third objective, the study counted and plotted occurrences of selected three-node motifs with specific edge and node attributes for each quarter. During the study, twelve distinct three-node motif rules are defined, to capture and count their occurrences in the graph over time. Each motif rule captured various aspects of the generated graph, including the magnitude of change, filing delay, investor type, as well as the stock's volume and volatility for each industry group over time. Notably, motifs that captured significant sales and purchases of the same stock, with both filings submitted late, only occurred towards the late stages of the 2008 and 2016 crises. While some motifs displayed an increasing trend without significant changes, Motifs 9 and 10, which captured significant sales and purchases of highly volatile stocks, peaked for each industry during the corresponding market crashes. Although this outcome was anticipated given the increased volatility of stocks during these periods, it is noteworthy that this approach managed to decompose the analysis for each related industry, period or even investor group.

As highlighted in the background section of the study, stock market crashes are rare events with serious consequences. Since these events may have various causes and are almost impossible to predict, the study applied rule based methods such as graph motifs, instead of a GNN based prediction model. This approach allowed the study to capture occurrences that happened before, and visualize patterns of previous events, without any indication of future occurrences. The study highlighted that almost all occurrences of motif 2 and 5, which captures "significant flows between different type of investors with late submitted filings", were found at late stages of crashes. Although without making predictions, this finding highlights an anticipation of such flows at late stages of the next possible market

crash. Considering the size of the market and the total asset under management of certain investors, a subtle edge on potential investor behavior is critical for both retail and institutional investors in making better investment decisions. Further analysis on these captured motif occurrences might also reveal stock, investor, or industry-level findings that could improve risk management or portfolio selection addressing late stages of market crashes.

In the related work section of the study, similar studies have been mentioned wherein different approaches to graph construction targeting investors were employed to derive clustering coefficients directly or in various forms, and these coefficients were utilized in the context of volatility and crash risk [142, 141]. Although this study adopted a distinct graph construction approach, it corroborated the use of clustering coefficients in measuring market risk. Similarly, another set of studies examined the portfolio overlap of financial institutions and observed an increase in overlap leading up to the 2008 crisis [129, 140]. However, since this study focused on portfolio changes rather than portfolios themselves, it failed to capture the buildup of this systemic risk. Additionally, brief mentions were made of recent studies that utilized motif count over time approach on constructed graphs for similar cases. However, these studies employed Chinese stock market co-shareholding networks using investors as the sole node type [143], global financial market relationships based on global stock market indices [144], or high order graphs based on DJIA stock price time-series [151]. Although this study employed fundamentally different approach for both construction of graph and examination of motifs, it can be associated with others as a work supporting the use of motif discovery and analysis for financial crisis and market risk.

In conclusion, the study examined investor behaviors in the stock market through diverse metrics with a focus on network analysis. By utilizing 13F filings with supplementary data, the research yielded explainable patterns of investor behaviors during market crashes.

5.3 Future Work

The conducted study has demonstrated that certain metrics of a graph constructed from 13F filings, such as significant buy and sale counts, filing delays, clustering coefficient and motif counts, can provide valuable insights about stock crashes. As a result, the study has the potential to evolve in various directions, to offer practical results.

One potential direction for future research involves extending the exploration of graphs beyond the created bipartite graph, incorporating the construction of graphs in multiple dimensions. This approach would enable the investigation of more intricate motifs, thereby enhancing the depth of the analysis. While the study focused solely on capturing quarter-over-quarter ownership changes, alternative networks such as information networks, stock price correlation networks, and industry similarity networks exist in the literature. By adding these connections to the graph, various new edge types can be introduced, allowing for the identification and measurement of more specific motifs.

Similarly, although the study did consider node and edge attributes to a certain extent, there was potential for extracting more advanced attributes related to stocks and investors. Enriching the analysis of node and edge attributes could have been achieved by integrating additional web-based information. One promising approach would be to maintain the current strategy, and include other filings, such as 10-K, 10-Q, or 8-Ks, which contain crucial details about corporate events, risk factors, financials,

and management issues. Exploring potential connections between institutional investor positions and corporate events might also offer valuable insights into how capital markets operate.

Another possible direction for future research in the study involves refining the selection of motif rules. In this study, the examined motif counts serve as functions similar as complex graph queries, meaning that the results are linked to posing insightful questions. Moreover, with the diversification of node and edge attributes, the potential combinations grow exponentially, making it challenging to decide which motifs to explore without the guidance of a domain expert. Therefore, a valuable direction for the study could involve experts defining motif rules in natural language to calculate temporal changes on-demand. Advancements in Natural Language Processing (NLP), especially in Large Language Models (LLM) today make this potential enhancement feasible, transforming it into a practical decision support system.

Although study emphasizes the necessity for tools and mechanisms to identify and understand these rare events, bubbles and crashes are much more common in cryptocurrencies due to their lack of regulation, dependence on social sentiment, and susceptibility to whale movements. Cryptocurrencies exhibit both market-wide and individual coin-wise bubbles and crashes, often following fraudulent schemes. Since the transparent nature of blockchain allows visibility of transactions, methods applied in the study can provide similar or even more in depth insights for crypto markets as well. However, off-chain transactions may complicate interpretation of findings, and reduce value of the extracted insights.

Additionally, in the literature, deep learning-based motif discovery methods have been observed. If motif discovery were linked to an end goal, specific events, or anomalies, it would be possible to train a model to detect motifs that are associated with certain events. This could offer a more comprehensive and beneficial approach, particularly through the exploration of Graph Neural Network (GNN) based methods. Such an approach could evolve into a solution to detect, classify or forecast actors or events in capital markets.

In summary, leveraging 13F filings and network analysis, this study provided insights into investor behavior during crashes, by visualizing market dynamics from various viewpoints, such as changes of holdings, filing delays, graph metrics and motif counts over time. These findings not only contribute to the literature, but also pave the way for future applications of network-based monitoring systems that could enhance regulatory oversight by identifying early warning signs of future crashes or help investors recognize lucrative investment opportunities during downturns.

REFERENCES

[1] C. Leuz and P. D. Wysocki, "Economic consequences of financial reporting and disclosure regulation: A review and suggestions for future research," *SSRN Electronic Journal*, 2008.

[2] J. Black, N. Hashimzade, and G. Myles, *A dictionary of economics*. Oxford University Press, USA, 2012.

[3] M. Friedman, "The social responsibility of business is to increase its profits," in *Corporate ethics and corporate governance*, pp. 173–178, Springer, 2007.

[4] F. Modigliani and M. H. Miller, "The cost of capital, corporation finance and the theory of investment," *The American economic review*, vol. 48, no. 3, pp. 261–297, 1958.

[5] M. Friedman, *The role of monetary policy*. Springer, 1995.

[6] H. P. Minsky and H. Kaufman, *Stabilizing an unstable economy*, vol. 1. McGraw-Hill New York, 2008.

[7] B. Cornell, "Economic growth and equity investing," *Financial Analysts Journal*, vol. 66, no. 1, pp. 54–64, 2010.

[8] J. Wurgler, "Financial markets and the allocation of capital," *Journal of financial economics*, vol. 58, no. 1-2, pp. 187–214, 2000.

[9] M. Pagano, "Financial markets and growth: An overview," *European economic review*, vol. 37, no. 2-3, pp. 613–622, 1993.

[10] A. Demirgüç-Kunt and R. Levine, "Stock markets, corporate finance, and economic growth: an overview," *The World Bank Economic Review*, vol. 10, no. 2, pp. 223–239, 1996.

[11] J. Greenwood and B. D. Smith, "Financial markets in development, and the development of financial markets," *Journal of Economic dynamics and control*, vol. 21, no. 1, pp. 145–181, 1997.

[12] W. Carlin and C. Mayer, "Finance, investment, and growth," *Journal of financial Economics*, vol. 69, no. 1, pp. 191–226, 2003.

[13] T. Beck, "Stock markets, banks, and economic development: theory and evidence," *EIB papers*, vol. 8, no. 1, pp. 37–54, 2003.

[14] C. Shi, K. Pukthuanthong, and T. Walker, "Does disclosure regulation work? evidence from international ipo markets," *Contemporary Accounting Research*, vol. 30, no. 1, pp. 356–387, 2013.

[15] H. B. Christensen, L. Hail, and C. Leuz, "Capital-market effects of securities regulation: Prior conditions, implementation, and enforcement," *The Review of Financial Studies*, vol. 29, no. 11, pp. 2885–2924, 2016.

[16] G. J. Benston, "Required disclosure and the stock market: An evaluation of the securities exchange act of 1934," *The American Economic Review*, vol. 63, no. 1, pp. 132–155, 1973.

[17] S. J. Choi and A. T. Guzman, "National laws, international money: Regulation in a global capital market," *Fordham L. Rev.*, vol. 65, p. 1855, 1996.

[18] I. Goldstein and L. Yang, "Information disclosure in financial markets," *Annual Review of Financial Economics*, vol. 9, pp. 101–125, 2017.

[19] S. Chi and D. M. Shanthikumar, "Do retail investors use sec filings? evidence from edgar search," *Evidence from EDGAR Search (October 25, 2018)*, 2018.

[20] T. Kravet and V. Muslu, "Textual risk disclosures and investors' risk perceptions," *Review of Accounting Studies*, vol. 18, pp. 1088–1122, 2013.

[21] R. Li, X. Wang, Z. Yan, and Y. Zhao, "Sophisticated investor attention and market reaction to earnings announcements: Evidence from the sec's edgar log files," *Journal of Behavioral Finance*, vol. 20, no. 4, pp. 490–503, 2019.

[22] C. N. Chiyachantana, C. X. Jiang, N. Taechapiroontong, and R. A. Wood, "The impact of regulation fair disclosure on information asymmetry and trading: An intraday analysis," *Financial Review*, vol. 39, no. 4, pp. 549–577, 2004.

[23] A. Gomes, G. Gorton, and L. Madureira, "Sec regulation fair disclosure, information, and the cost of capital," *Journal of Corporate Finance*, vol. 13, no. 2-3, pp. 300–334, 2007.

[24] J. W. Kim, J.-H. Lim, and W. G. No, "The effect of first wave mandatory xbrl reporting across the financial information environment," *Journal of Information Systems*, vol. 26, no. 1, pp. 127–153, 2012.

[25] M. A. Vasarhelyi, D. Y. Chan, and J. P. Krahel, "Consequences of xbrl standardization on financial statement data," *Journal of Information Systems*, vol. 26, no. 1, pp. 155–167, 2012.

[26] R. Merton, "Capital market theory and the pricing of financial securities," *Handbook of monetary economics*, vol. 1, pp. 497–581, 1990.

[27] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, pp. 77–91, 1952.

[28] J. L. Treynor, "Market value, time, and risk," *Time, and Risk (August 8, 1961)*, 1961.

[29] E. F. Fama and K. R. French, "The cross-section of expected stock returns," *the Journal of Finance*, vol. 47, no. 2, pp. 427–465, 1992.

[30] E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of financial economics*, vol. 116, no. 1, pp. 1–22, 2015.

[31] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[32] B. Graham and D. Dodd, "Security analysis: Foreword by warren buffett (security analysis prior editions)," 2008.

[33] B. Graham and B. McGowan, *The intelligent investor*. Harper Collins New York, 2005.

[34] W. E. Buffett, "The superinvestors of graham-and-doddsville," *Hermes*, pp. 4–15, 1984.

[35] U. Securities, E. Commission, *et al.*, "Form 13f—reports filed by institutional investment managers," 2011.

[36] M. E. Blume and D. B. Keim, "Institutional investors and stock market liquidity: trends and relationships," *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*, 2012.

[37] M. J. Gruber, "Another puzzle: The growth in actively managed mutual funds," *The journal of finance*, vol. 51, no. 3, pp. 783–810, 1996.

[38] L. Zheng, "Is money smart? a study of mutual fund investors' fund selection ability," *the Journal of Finance*, vol. 54, no. 3, pp. 901–933, 1999.

[39] T. Sapp and A. Tiwari, "Does stock return momentum explain the "smart money" effect?," *The Journal of Finance*, vol. 59, no. 6, pp. 2605–2622, 2004.

[40] F. Cai and L. Zheng, "Institutional trading and stock returns," *Finance Research Letters*, vol. 1, no. 3, pp. 178–189, 2004.

[41] T. P. Lemke and G. T. Lins, "Disclosure of equity holdings by institutional investment managers: An analysis of section 13 (f) of the securities exchange act of 1934," *Bus. Law.*, vol. 43, p. 93, 1987.

[42] B. U. Kang, A. Sinclair, and S. Xenomorph, "Does portfolio disclosure make money smarter?," in *Proceedings of Paris December 2021 Finance Meeting EUROFIDAI-ESSEC*, 2021.

[43] J. A. Bennett, R. W. Sias, and L. T. Starks, "Greener pastures and the impact of dynamic institutional preferences," *The Review of Financial Studies*, vol. 16, no. 4, pp. 1203–1238, 2003.

[44] P. A. Gompers and A. Metrick, "Institutional investors and equity prices," *The quarterly journal of Economics*, vol. 116, no. 1, pp. 229–259, 2001.

[45] S. Jones, D. Lee, and E. Weis, "Herding and feedback trading by different types of institutions and the effects on stock prices," *Unpublished working paper, Indiana University, Indianapolis Kennesaw State University, Kennesaw GA, and Merrill Lynch and Company, New York*, 1999.

[46] M. Grinblatt, S. Titman, and R. Wermers, "Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior," *The American economic review*, pp. 1088–1105, 1995.

[47] R. W. Sias, L. T. Starks, and S. Titman, "The price impact of institutional trading," *Available at SSRN 283779*, 2001.

[48] S. G. Badrinath and S. Wahal, "Momentum trading by institutions," *The Journal of Finance*, vol. 57, no. 6, pp. 2449–2478, 2002.

[49] J. Y. Campbell, T. Ramadorai, and A. Schwartz, "Caught on tape: Institutional trading, stock returns, and earnings announcements," *Journal of Financial Economics*, vol. 92, no. 1, pp. 66–91, 2009.

[50] R. W. Sias, L. T. Starks, and S. Titman, "Changes in institutional ownership and stock returns: Assessment and methodology," *The Journal of Business*, vol. 79, no. 6, pp. 2869–2910, 2006.

[51] D. Miori and M. Cucuringu, "Sec form 13f-hr: Statistical investigation of trading imbalances and profitability analysis," *arXiv preprint arXiv:2209.08825*, 2022.

[52] A. Fleiss, H. Cui, S. Stoikov, and D. M. DiPietro, "Constructing equity portfolios from sec 13f data using feature extraction and machine learning," *The Journal of Financial Data Science*, 2019.

[53] L. Angelini, M. Iqbal, and F. Jivraj, "Systematic 13f hedge fund alpha," *Available at SSRN 3459526*, 2019.

[54] D. M. DiPietro, "Alpha cloning: Using quantitative techniques and sec 13f data for equity portfolio optimization and generation," *The Journal of Financial Data Science*, 2019.

[55] G. Bescos, A. Moyal, and P. Veyrat, "Classification of assets and investors in a financial bipartite graph," 2019.

[56] A. Chouliaras, "Institutional investors, analysts' recommendations, annual reports, textual analysis and stock returns: Evidence from sec edgar 10-k and 13-f forms," *Analysts' Recommendations, Annual Reports, Textual Analysis and Stock Returns: Evidence from SEC EDGAR*, 2015.

[57] R. Gibson, P. Krueger, and S. F. Mitali, "The sustainability footprint of institutional investors: Esg driven price pressure and performance," *Swiss Finance Institute Research Paper*, no. 17-05, 2020.

[58] S. K. Christoffersen, E. Danesh, and D. K. Musto, "Why do institutions delay reporting their shareholdings? evidence from form 13f," *Rotman School of Management Working Paper*, no. 2661535, 2015.

[59] Securities and E. Commission, "Information required of institutional investment managers pursuant to section 13(f) of the securities exchange act of 1934 and rules thereunder," *SEC DOCKET*, 2022.

[60] G. O. Aragon, M. Hertzel, and Z. Shi, "Why do hedge funds avoid disclosure? evidence from confidential 13f filings," *Journal of Financial and Quantitative Analysis*, vol. 48, no. 5, pp. 1499–1518, 2013.

[61] S. Cao, Z. Da, D. Jiang, and B. Yang, "Do hedge funds strategically misreport their holdings? evidence from 13f restatements," *Evidence from 13F Restatements (June 15, 2022)*, 2022.

[62] V. Agarwal, W. Jiang, Y. Tang, and B. Yang, "Do institutional investors have an ace up their sleeves? evidence from confidential filings of portfolio holdings," *Evidence from Confidential Filings of Portfolio Holdings (March 16, 2010)*, 2010.

[63] A. M. Anderson and P. Brockman, "Form 13f (mis) filings," *Available at SSRN 2809128*, 2016.

[64] S. J. Brown and C. Schwarz, "Do market participants care about portfolio disclosure? evidence from hedge funds' 13f filings," *Evidence from Hedge Funds' 13F Filings (December 17, 2020)*, 2020.

[65] D. Sornette and A. Johansen, "Significance of log-periodic precursors to financial crashes," *Quantitative Finance*, vol. 1, no. 4, p. 452, 2001.

[66] A. Johansen, D. Sornette, *et al.*, "Shocks, crashes and bubbles in financial markets," *Brussels Economic Review*, vol. 53, no. 2, pp. 201–253, 2010.

[67] D. Sornette, Y. Malevergne, and J.-F. Muzy, "Volatility fingerprints of large shocks: endogenous versus exogenous," in *The Application of Econophysics: Proceedings of the Second Nikkei Econophysics Symposium*, pp. 91–102, Springer, 2004.

[68] R. E. Farmer, "The stock market crash of 2008 caused the great recession: Theory and evidence," *Journal of Economic Dynamics and Control*, vol. 36, no. 5, pp. 693–707, 2012.

[69] R. J. Barro and J. F. Ursúa, "Stock-market crashes and depressions," *Research in Economics*, vol. 71, no. 3, pp. 384–398, 2017.

[70] M. Torres, "The big short: Inside the doomsday machine," *Journal of Markets & Morality*, vol. 14, no. 1, pp. 216–220, 2011.

[71] E. A. Thompson, "The tulipmania: Fact or artifact?," *Public Choice*, vol. 130, pp. 99–114, 2007.

[72] R. Greenwood, A. Shleifer, and Y. You, "Bubbles for fama," *Journal of Financial Economics*, vol. 131, no. 1, pp. 20–43, 2019.

[73] D. Sornette, A. Johansen, and J.-P. Bouchaud, "Stock market crashes, precursors and replicas," *Journal de Physique I*, vol. 6, no. 1, pp. 167–175, 1996.

[74] A. Johansen and D. Sornette, "Stock market crashes are outliers," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 1, pp. 141–143, 1998.

[75] L. Cao, "Data science: a comprehensive overview," *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, pp. 1–42, 2017.

[76] K. M. Tolle, D. S. W. Tansley, and A. J. Hey, "The fourth paradigm: Data-intensive scientific discovery [point of view]," *Proceedings of the IEEE*, vol. 99, no. 8, pp. 1334–1337, 2011.

[77] V. Dhar, "Data science and prediction," *Association for Computing Machinery*, vol. 56, no. 12, 2013.

[78] L. Cao, "Ai in finance: challenges, techniques, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–38, 2022.

[79] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, "Deep learning for financial applications: A survey," *Applied Soft Computing*, vol. 93, p. 106384, 2020.

[80] A. Bahrammirzaee, "A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems," *Neural Computing and Applications*, vol. 19, no. 8, pp. 1165–1195, 2010.

[81] D. Zhang and L. Zhou, "Discovering golden nuggets: data mining in financial application," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 4, pp. 513–522, 2004.

[82] N. Burgess, "An introduction to algorithmic trading: Opportunities & challenges within the systematic trading industry," *Available at SSRN 3466213*, 2019.

[83] L. Zhang, T. Wu, S. Lahrichi, C.-G. Salas-Flores, and J. Li, "A data science pipeline for algorithmic trading: A comparative study of applications for finance and cryptoeconomics," in *2022 IEEE International Conference on Blockchain (Blockchain)*, pp. 298–303, IEEE, 2022.

[84] A. Chatterjee, H. Bhowmick, and J. Sen, "Stock price prediction using time series, econometric, machine learning, and deep learning models," in *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*, pp. 289–296, IEEE, 2021.

[85] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, "Big data: Deep learning for financial sentiment analysis," *Journal of Big Data*, vol. 5, no. 1, pp. 1–25, 2018.

[86] M. Zanjirdar, "Overview of portfolio optimization models," *Advances in mathematical finance and applications*, vol. 5, no. 4, pp. 419–435, 2020.

[87] J. Sen and S. Dasgupta, "Portfolio optimization: A comparative study," *arXiv preprint arXiv:2307.05048*, 2023.

[88] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, 2020.

[89] L. Chen, M. Pelger, and J. Zhu, "Deep learning in asset pricing," *Management Science*, 2023.

[90] G. Feng, J. He, and N. G. Polson, "Deep learning for predicting asset returns," *arXiv preprint arXiv:1804.09314*, 2018.

[91] G. Feng, J. He, N. G. Polson, and J. Xu, "Deep learning in characteristics-sorted factor models," *arXiv preprint arXiv:1805.01104*, 2018.

[92] A. Gupta, P. Gulati, and S. P. Chakrabarty, "Classification based credit risk analysis: The case of lending club," *arXiv preprint arXiv:2210.05136*, 2022.

[93] J. M. Clements, D. Xu, N. Yousefi, and D. Efimov, "Sequential deep learning for credit risk monitoring with tabular financial data," *arXiv preprint arXiv:2012.15330*, 2020.

[94] D. Zakrzewska, "On integrating unsupervised and supervised classification for credit risk evaluation," *Information technology and control*, vol. 36, no. 1, 2007.

[95] P. Tiwari, S. Mehta, N. Sakhuja, J. Kumar, and A. K. Singh, "Credit card fraud detection using machine learning: a study," *arXiv preprint arXiv:2108.10005*, 2021.

[96] S. X. Rao, C. Lanfranchi, S. Zhang, Z. Han, Z. Zhang, W. Min, M. Cheng, Y. Shan, Y. Zhao, and C. Zhang, "Modelling graph dynamics in fraud detection with" attention"," *arXiv preprint arXiv:2204.10614*, 2022.

[97] D. Wang, J. Lin, P. Cui, Q. Jia, Z. Wang, Y. Fang, Q. Yu, J. Zhou, S. Yang, and Y. Qi, "A semi-supervised graph attentive network for financial fraud detection," in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 598–607, IEEE, 2019.

[98] Z. Liu, Y. Dou, P. S. Yu, Y. Deng, and H. Peng, "Alleviating the inconsistency problem of applying graph neural network to fraud detection," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 1569–1572, 2020.

[99] Z. Liu, C. Chen, X. Yang, J. Zhou, X. Li, and L. Song, "Heterogeneous graph neural networks for malicious account detection," in *Proceedings of the 27th ACM international conference on information and knowledge management*, pp. 2077–2085, 2018.

[100] D. Zakrzewska and J. Murlewski, "Clustering algorithms for bank customer segmentation," in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pp. 197–202, IEEE, 2005.

[101] M. Aryuni, E. D. Madyatmadja, and E. Miranda, "Customer segmentation in xyz bank using k-means and k-medoids clustering," in *2018 International conference on information management and technology (ICIMTech)*, pp. 412–416, IEEE, 2018.

[102] D. Burdick, M. Hernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. C. Popa, I. Stanoi, S. Vaithyanathan, and S. R. Das, "Extracting, linking and integrating data from public sources: A financial case study," *Available at SSRN 2666384*, 2015.

[103] Z. Wang, Y. Li, J. Wu, J. Soon, and X. Zhang, "Finvis-gpt: A multimodal large language model for financial chart analysis," *arXiv preprint arXiv:2308.01430*, 2023.

[104] Y. Yang, Y. Tang, and K. Y. Tam, "Investlm: A large language model for investment using financial domain instruction tuning," *arXiv preprint arXiv:2309.13064*, 2023.

[105] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[106] I. Robinson, J. Webber, and E. Eifrem, *Graph databases: new opportunities for connected data*. " O'Reilly Media, Inc.", 2015.

[107] T. Hegeman and A. Iosup, "Survey of graph analysis applications," *arXiv preprint arXiv:1807.00382*, 2018.

[108] F. Xia, H. Wei, S. Yu, D. Zhang, and B. Xu, "A survey of measures for network motifs," *IEEE Access*, vol. 7, pp. 106576–106587, 2019.

[109] M. Li, X. Wang, K. Gao, and S. Zhang, "A survey on information diffusion in online social networks: Models and methods," *Information*, vol. 8, no. 4, p. 118, 2017.

[110] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[111] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[112] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.

[113] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[114] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[115] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein, "Temporal graph networks for deep learning on dynamic graphs," *arXiv preprint arXiv:2006.10637*, 2020.

[116] R. Trivedi, M. Farajtabar, P. Biswal, and H. Zha, "Dyrep: Learning representations over dynamic graphs," in *International conference on learning representations*, 2019.

[117] S. Kumar, X. Zhang, and J. Leskovec, "Predicting dynamic embedding trajectory in temporal interaction networks," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1269–1278, 2019.

[118] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White, "Economic networks: The new challenges," *science*, vol. 325, no. 5939, pp. 422–425, 2009.

[119] G. Kou, X. Chao, Y. Peng, and F. Wang, "Network resilience in the financial sectors: Advances, key elements, applications, and challenges for financial stability regulation," *Technological and Economic Development of Economy*, vol. 28, no. 2, pp. 531–558, 2022.

[120] A. Kulkarni, P. Mani, and C. Domeniconi, "Network-based anomaly detection for insider trading," *arXiv preprint arXiv:1702.05809*, 2017.

[121] I. Aldasoro and I. Alves, "Multiplex interbank networks and systemic importance: An application to european data," *Journal of Financial Stability*, vol. 35, pp. 17–37, 2018.

[122] X. Freixas, B. M. Parigi, and J.-C. Rochet, "Systemic risk, interbank relations, and liquidity provision by the central bank," *Journal of money, credit and banking*, pp. 611–638, 2000.

[123] C. León and J. Miguélez, "Interbank relationship lending: A network perspective," *Physica A: Statistical Mechanics and its Applications*, vol. 573, p. 125922, 2021.

[124] L. Enriques and A. Romano, "Institutional investor voting behavior: A network theory perspective," *U. Ill. L. Rev.*, p. 223, 2019.

[125] L. Cohen, A. Frazzini, and C. Malloy, "The small world of investing: Board connections and mutual fund returns," *Journal of Political Economy*, vol. 116, no. 5, pp. 951–979, 2008.

[126] J. Eberhard, J. F. Lavin, A. Montecinos-Pearce, *et al.*, "A network-based dynamic analysis in an equity stock market," *Complexity*, vol. 2017, 2017.

[127] J. Engel, M. Nardo, and M. Rancan, "Network analysis for economics and finance: an application to firm ownership," in *Data Science for Economics and Finance: Methodologies and Applications*, pp. 331–355, Springer International Publishing Cham, 2021.

[128] A. Gupta, R. Wang, and Y. Lu, "Addressing systemic risk using contingent convertible debt–a network analysis," *European Journal of Operational Research*, vol. 290, no. 1, pp. 263–277, 2021.

[129] S. Gualdi, G. Cimini, K. Primicerio, R. Di Clemente, and D. Challet, "Statistically validated network of portfolio overlaps and systemic risk," *Scientific reports*, vol. 6, no. 1, p. 39467, 2016.

[130] F. Franch, L. Nocciola, and A. Vouldis, "Temporal networks in the analysis of financial contagion," 2022.

[131] S. Giansante, "Recent advances in modelling systemic risk using network analysis-ecb," *The CDS Network*, p. 11, 2010.

[132] X. Wan, Z. Zhang, C. Zhang, and Q. Meng, "Stock market temporal complex networks construction, robustness analysis, and systematic risk identification: a case of csi 300 index," *Complexity*, vol. 2020, pp. 1–19, 2020.

[133] P. Geraskin and D. Fantazzini, "Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask," *New Facets of Economic Complexity in Modern Financial Markets*, pp. 30–55, 2020.

[134] Z.-Q. Jiang, W.-X. Zhou, D. Sornette, R. Woodard, K. Bastiaensen, and P. Cauwels, "Bubble diagnosis and prediction of the 2005–2007 and 2008–2009 chinese stock market bubbles," *Journal of economic behavior & organization*, vol. 74, no. 3, pp. 149–162, 2010.

[135] T. Kaizoji and D. Sornette, "Market bubbles and crashes," *arXiv preprint arXiv:0812.2449*, 2008.

[136] M. Shu and W. Zhu, "Real-time prediction of bitcoin bubble crashes," *Physica A: Statistical Mechanics and its Applications*, vol. 548, p. 124477, 2020.

[137] R. C. Phillips and D. Gorse, "Predicting cryptocurrency price bubbles using social media data and epidemic modelling," in *2017 IEEE symposium series on computational intelligence (SSCI)*, pp. 1–7, IEEE, 2017.

[138] R. Sawhney, S. Agarwal, V. Mittal, P. Rosso, V. Nanda, and S. Chava, "Cryptocurrency bubble detection: a new stock market dataset, financial task & hyperbolic models," *arXiv preprint arXiv:2206.06320*, 2022.

[139] M. Shu, R. Song, and W. Zhu, "The 2021 bitcoin bubbles and crashes—detection and classification," *Stats*, vol. 4, no. 4, pp. 950–970, 2021.

[140] J. R. Abrams, J. Celaya-Alcalá, D. Baldwin, R. Gonda, and Z. Chen, "Analysis of equity markets: A graph theory approach," *Society for Industrial and Applied Mathematics. doi*, vol. 10, 2016.

[141] J. Li, L. Wang, Z.-Q. Zhou, and Y. Zhang, "Monitoring or tunneling? information interaction among large shareholders and the crash risk of the stock price," *Pacific-Basin Finance Journal*, vol. 65, p. 101469, 2021.

[142] J. Li, Y. Zhang, and L. Wang, "Information transmission between large shareholders and stock volatility," *The North American Journal of Economics and Finance*, vol. 58, p. 101551, 2021.

[143] J. Li, Y. Zhang, and L. Wang, "Information transmission among multiple investors: a micro-perspective revealed by motifs," *Nonlinear Dynamics*, vol. 108, no. 3, pp. 2833–2850, 2022.

[144] W.-J. Xie, Y. Yong, N. Wei, P. Yue, and W.-X. Zhou, "Identifying states of global financial market based on information flow network motifs," *The North American Journal of Economics and Finance*, vol. 58, p. 101459, 2021.

[145] Z.-Q. Jiang, W.-J. Xie, X. Xiong, W. Zhang, Y.-J. Zhang, and W.-X. Zhou, "Trading networks, abnormal motifs and stock manipulation," *Quantitative Finance Letters*, vol. 1, no. 1, pp. 1–8, 2013.

[146] Q. Guan, H. An, N. Liu, F. An, and M. Jiang, "Information connections among multiple investors: evolutionary local patterns revealed by motifs," *Scientific Reports*, vol. 7, no. 1, p. 14034, 2017.

[147] M. Backus, C. Conlon, and M. Sinkinson, "Common ownership in america: 1980–2017," *American Economic Journal: Microeconomics*, vol. 13, no. 3, pp. 273–308, 2021.

[148] D. Sornette, *Why stock markets crash: critical events in complex financial systems*. Princeton university press, 2009.

[149] A. Beattie, "Market crashes: what are crashes and bubble?," 2016.

[150] L. Cai, Z. Chen, C. Luo, J. Gui, J. Ni, D. Li, and H. Chen, "Structural temporal graph neural networks for anomaly detection in dynamic graphs," in *Proceedings of the 30th ACM international conference on Information & Knowledge Management*, pp. 3747–3756, 2021.

[151] A. O. Bielinskyi, V. N. Soloviev, S. V. Hushko, A. E. Kiv, and A. V. Matviychuk, "High-order network analysis for financial crash identification," in *CEUR Workshop Proceedings*, pp. 132–149, 2023.