Topological Properties of Stock Market Networks: A Comparison Between Chinese and US Stock Markets

Yong Tang[a,d,*], Jason Jie Xiong[b], Zi-Yang Jia[c], Yi-Cheng Zhang[d]

[a] *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 610054, China*
[b] *Department of Computer Information Systems and Supply Chain Management, Walker College of Business, Appalachian State University, Boone, NC 28608, USA*
[c] *Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA*
[d] *Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland*

Abstract

There are recent advances in applying data-driven science and network theory into the studies of social and financial systems. Financial assets and institutes are strongly connected and influence with each other. It is essential to study how the topological structures of financial networks could potentially influence market behaviors. Network analysis is a powerful method to enhance data mining and knowledge discovery in financial data. With the help of complex network theory, the network topological structures of a market can be extracted to reveal hidden information and relationships among stocks. In this study, two major markets of the most influential economies, China and the United States, are systematically studied from the perspective of financial network analysis. Results suggest that the network properties and hierarchical structures from the two stock markets are; one is developing and growing while the other is well developed and established. The patterns embedded in the price movements are revealed and shed light of the market dynamics. Financial investors and regulators can gain inspiration from these findings for applications in portfolio management, risk management, and trading.

Keywords: Financial network analysis, Quantitative finance, Computational

*Corresponding author
Email address: tangyong@uestc.edu.cn (Yong Tang)

1. Introduction

The study and visualization of networks and hierarchy structures are essential to understand complex systems like financial markets. Thanks to the great development on complex network science [1], quantitative methods and models have been applied in the studies of the network structure of financial markets. In the emerging financial network analysis, financial entities like assets, stocks, markets, companies, and institutes are modeled as vertices while their mutual relationships are abstracted as edges. This approach empowers industrial professionals and researchers to reveal hidden information embedded in the topological structures of financial networks, such as the market dynamics, trading activities, and investment sentiment. These information is essential to evaluate and monitor the financial market risks, contagions, distress propagation, as well as market mode shifts. Financial network analysis has been utilized in applications like portfolio management, trading, market regulation, stress testing, risk management, etc.

Among the global economies, the USA and China are the top two dominating economies with grand influences over the global economies. The two economies are close in market scales. However, the US economy is well established and developed while Chinese economy is emerging and still undertaking fast development. As the leading economic powerhouses, the health and stabilities of these two economies are vital for the prosperities of world economy. During the past few years, both countries suffered a series of stock market disasters, such as 2008 US subprime crisis, the 2007 and 2015 Chinese stock market bubble bursts. These dramatic market crashes brought widespread and long lasting negative impacts on economies and markets. The stock markets of both countries are also different in terms of history, regulations, maturities, and scales. Thus, it is essential to understand the properties of these two markets by utilizing the data-driven science approach. Recently, various major markets in the world

have been investigated using the financial network analysis approach. However, there is still a lack of systematical studies dedicating to compare the network structures and properties of US and Chinese stock markets using financial network analysis approach.

To understand how the two markets differ in the structures and topological properties, as well as the dynamics market properties of US and Chinese the stock markets, this research investigates the markets using a dataset spanning over 9 years. In this research, the stock markets are modeled as multiple networks including hierarchical trees, minimum spanning trees, planar maximally filtered graphs, and assets graphs are built. Meanwhile, their detailed topological properties are analyzed and systematically compared. Through quantitative analysis and network visualization, our results show the two markets are different in many ways.

This provides insights for regulators on the structures and dynamics of stock markets from the perspective of network science.

This paper is organized as follows. First, Section 2 gives a background on the theory of complex networks and financial network analysis and important complex network parameters are introduced. Then, in Section 3, the data and method used to construct networks are described. Section 4 presents the network properties. In Section 5, the detailed hierarchical structures of both stock networks are carefully investigated and compared. Finally, conclusions and discussions are presented in Section 6.

2. Literature review of financial network analysis

Still in fast developing, network science has become an innovative tool widely used in studies of complex systems in a variety of engineering and scientific domains [2–4]. The network modeling methodologies and theoretical frameworks have been found applications in many systems revealing informative and useful empirical discoveries [5]. Studying the statistical properties such as degree distribution, average length, clustering coefficient, etc., can help us to describe the

networks topologies and investigate the dynamics of network evolution. Furthermore, we are able to study the information spreading, network stability, phase changes, and hopefully to predict and control the network dynamics [6].

Using the price time series data, it's strait forward to calculate correlation matrices for a group of assets [7, 8]. From the correlation matrices, financial network analysis could be applied to construct networks for further analysis and data mining [9, 10]. In most existing literatures, assets are treated as vertices, while the interconnectivity relationships are abstracted as pairwise edges among assets. In fact, the correlation matrices are not only important for network analysis and topological visualization, but also serve as a bridge between financial network analysis and traditional finance theories like modern portfolio theory (MPT) [11, 12], for both are based on the correlation relationships among assets. Network based portfolio selection has been proposed for optimization and empirically proofed workable [13].

Since the minimum spanning tree approach is first used in study of stock market structure [14], financial network analysis has grown into an essential tool of financial big data. However, this fast growing field is still in an early stage [15]. Financial network analysis provides an unprecedented perspective shedding new insights on evaluating the market stability, market risk, shock propagation, and contagion [16]. The connectedness among assets plays critical role of market contagion phase transition [17] which is similar to other tolerance properties of other non-financial complex networks [18]. Further research reveals that intermediate level of risk diversification can enhance the market robustness [19]. The importance and risk contribution of companies can be identified through the network analysis [20, 21]. The systemic risks and stability can also be evaluated according to the topological properties of the financial network, and providing implications for market regulations [15, 22, 23]. Through investigating the clusterings of assets, portfolio optimization can be achieved with better predicted over realized risks [24]. Overlapping of portfolios is revealed by network analysis as one primary factor for market contagion [25]. In another approach, risk spillover networks are constructed to study the behaviors of financial institu-

tions [26]. Instead of single layer approach, by building multiple layer network, the banking system risk is studied and quantified [27]. Regression models can also take network structure into consideration as factors for resilience and robustness of the markets [28]. This kind of financial network approach opens more interesting new possibilities and greatly enriches regression models in finance researches. Recently there has been a thread of studies on major players in a financial networks, such as 'too interconnected to fail' institutes [29], 'too central to fail' [30], 'too big to fail' [31], etc. These researches demonstrated that financial network analysis brings new insights to finance researches and benefits to finance practices.

In the raise of quantitative trading, the causality and lead/lag relationships revealed by financial network analysis can be particularly interesting for trading strategy design [32, 33]. Many researches have revealed stylized evidences that the network structure has profound influence on the asset returns [34]. Taking risks into consideration, it has been found that investing in peripheries of financial networks might generate better returns over risks [35]. Furthermore, industry professionals would be inspired from financial network analysis to seek price movement signals for potential predictions [36].

While most existing financial network analysis literatures focus on the stock markets or specific economy sectors [37–39], there is a variety of financial systems have been studies as financial network such as global financial institutions [30, 31], world trade web [40, 41], interbank markets [42–45], exchanges [46], monetary market [47], corporate networks [48, 49], global banking [50], CDS market [51], credit market [52].

Many major individual financial markets around the world have also been studied in network approach, such as US [53, 54], China [55–57], Germany [58, 59], EU [60–62], Brazil [24, 63], Italian [44, 47], Korea [64], Russia [65], Mexico [27, 66], etc. Furthermore, there are some literatures focus on the cross-board global markets [67, 68]. Using partial data, networks of global markets are reconstructed and methods are compared [47]. Bayesian graphical models are applied to identify groups of counties which are major contributors of systemic

risks according to banking behaviors in the global banks [69]. For the European markets, the risk and contagion channels are studied and results show the EU markets are vulnerable for risks [62]. Global stock exchange network is investigated to evaluate the attractions for IPOs [70]. A recent study has demonstrated the approach which uses transfer entropy to study a selection of major individual stocks around the world and reveals that stocks are clustered according to their countries and industries [71]. By looking into the network structures of global financial network, it's possible to give new insights on the international business cycle [72]. The diversification and participation are investigated for various economies [73].

Considering the large number of assets in financial markets, the initial networks have huge number of edges. By filtering the noises of the networks, the financial networks can be significantly simplified to enable advanced analysis such as principle component analysis (PCA) [67] and random matrix theory (RMT) [74, 75], etc. to further extract hidden pattens. Hierarchical Tree (HT) [14, 59], Minimal Spanning Tree (MST) [76], Planar Maximally Filtered Graphs (PMFG) [77], Asset Graph (AG) [78, 79] are major approaches applied in filtering financial networks. Mantegna [14] first introduces the Minimum Spanning Tree method into the study of hierarchical structures in financial markets. With the network, we are capable to study the topological structure of a market or a portfolio. In this research, we adopt these frameworks to study the correlations and the corresponding networks of stock markets both in China and United States in order to systematically study how the two markets behave differently.

3. Data and Research Methods

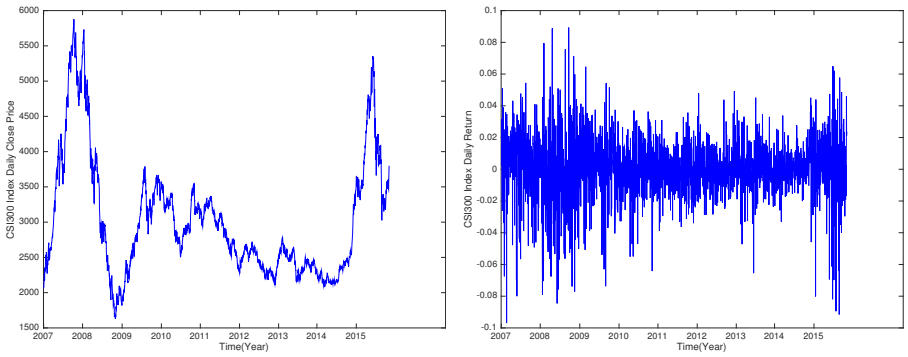
3.1. Indices of CSI300 and S&P500

We study the stock markets of China and United States, the former is a typical representative of emerging countries with fast growing GDP rate and influence on global economies, while the latter is the most established and developed economy in the world. In order to study the major stocks of each

market, we focus on the component stocks of the major indices of the two stock markets, *ie* China Securities Index 300 (CSI300) for the Chinese stock market and Standard & Poor's 500 (S&P500) for the US market. In our study, we cover a period of 9 years starts on 04/01/2007 and ends on 06/11/2015 with 2149 trading days for CSI300 and 2228 trading days for S&P500. The reason why the two markets have different numbers of trading dates is that the two markets have different trading calendars. Index and all component stocks daily price data of CSI300 are retrieved from the CSMAR Solution Database of Shenzhen GTA Education Tech. Ltd. We download the S&P500 index and component stocks daily prices data through Yahoo Finance service. Since not all stocks are traded on each trading date, so we only select those CSI300 stocks with at least 2000 trading dates and without continuous 100 non-trading dates, this selection results a final set of 163 stocks. For S&P500, we select those stocks with at least 2100 trading dates, and in results we get 468 stocks. After stocks selection, we take the prices on the available closest trading date to fill the non-trading dates. In Figs. 1(a)-1(b) we plot the daily close prices and the daily log returns for the index of CSI300 in the study period of 04/01/2007 and 06/11/2015 with 2149 trading days. In Figs. 2(a)-2(b), we plot the daily close prices and the daily log returns for the index of S&P500 in the same study period of 04/01/2007 and 06/11/2015 with 2228 trading days. From the figures, we see that the two markets show large fluctuations in the last 9 years. CSI300 experienced huge market crashes in 2008 and 2015, while S&P500 kept climbing almost continuously after the 2008 financial crisis.

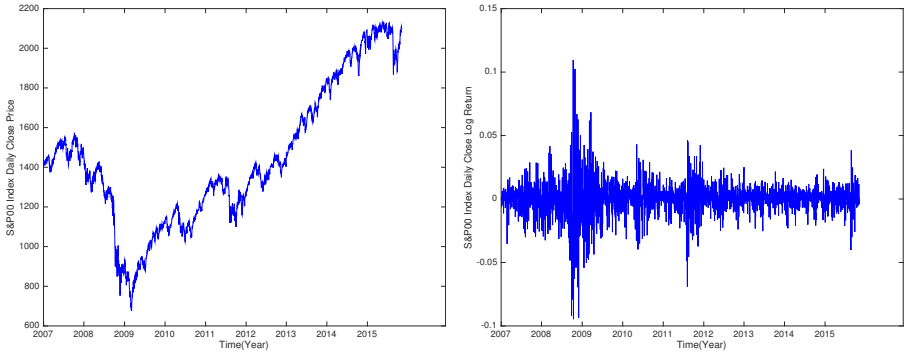
3.1.1. CSI300

China has two independent stock market exchanges, one is the Shanghai stock exchange, another is the Shenzhen stock exchange. Opened at the beginning of 1990's with only 25 years of trading history, the two markets have grown into important financial markets playing extremely important roles in China's financial markets and economy. Among the many stock market indices, the China Securities Index 300, or CSI300, was introduces by the China Securities



(a) CSI300 daily close prices (b) CSI300 daily log returns

Figure 1: CSI300 index daily close prices (a) and log returns (b) in the study period between 04/01/2007 and 06/11/2015 with 2149 trading days.



(a) S&P500 daily close prices (b) S&P500 daily log returns

Figure 2: S&P500 index daily close prices (a) and log returns (b) in the study period between 04/01/2007 and 06/11/2015 with 2228 trading days.

Index Company, Ltd. in 2005 to a base of 1000 on 31/12/2004. In CSI300, a set of 300 stocks are included as the index components, all of them have the largest market values and are actively traded in Shanghai or Shenzhen stock exchanges. CSI300 has become a widely accepted benchmark to evaluate the whole stock markets behaviors in China as well as a good basis for other derivative products. Starting from 1000 points in the early of 2004, now CSI300 has reached 3793 points as of 06/11/2015 [80]. To give an image of the Chinese stock market, we

plot the 2149 CSI300 index daily close prices and daily log returns in the study period between 04/01/2007 and 06/11/2015 in Fig. 1(a) and Fig. 1(b). In the past 9 years, CSI300 experienced two major market crashes in 2007-2008 and 2015 respectively, during which the market suffered huge losses and fluctuations. There are 163 stocks of CSI300 component stocks included in our dataset, as shown in Table 1, in which we summarize the numbers of these 163 stocks for all 20 industry sectors. As shown, all industry sectors from Agriculture to Comprehensive are represented. For convenience, we will refer to these 163 stocks as *CSI163* in the following parts.

3.1.2. S&P500

Compiled by Standard & Poor's in 1957, the S&P500 is an established American stock market index with more components, more risk-diversification, and better reflection of the overall stock market performance than both the New York Stock Exchange (NYSE) and Nasdaq. All components are large stocks in capitalization with good liquidities and diversifications in different industry sectors. The S&P500 represents major parts of the market and is considered as one of the best benchmarks for the US financial markets and economy. Starting from less than 100, after more than 50 years of development, the S&P500 reached 2099 on 06/11/2015 [81]. We plot 2228 daily close prices and log returns of the index of the S&P500 in our study period between 04/01/2007 and 06/11/2015 in Fig. 2(a) and Fig. 2(b), in which we can observe that the S&P500 index suffered a major crash in the period of 2008-2009 then recovered almost steadily with minor fluctuations. After the selection, there are 468 stocks of the S&P500 component stocks included in our dataset, as shown in Table 2. We summarize the numbers of these 468 stocks for all 10 industry sectors. As shown, all industry sectors from Energy to Utilities are represented. For convenience, from now we will refer to these 468 stocks as *S&P468* in the following parts. Table 3 gives a summary of the two datasets of both CSI163 and S&P468, CSI163 has a larger standard deviation σ_r of the log returns, indicating larger fluctuations than S&P468. We use $\langle x \rangle$ to denote the average of variable x in this thesis.

Table 1: 163 component stocks of CSI300 are included in our dataset. In this table, we list the China Securities Regulatory Commission (CSRC) industry code, sector name, and numbers of stocks for each industry sector of these 163 stocks. All 20 industry sectors are represented.

Industry code	Industry Sector	Number of Stocks
A	Agriculture	1
B	Mining	6
C0	Food & Beverage	4
C1	Textiles & Apparel	4
C3	Paper & Printing	2
C4	Petrochemicals	9
C5	Electronics	7
C6	Metals & Non-metals	20
C7	Machinery	27
C8	Pharmaceuticals	15
D	Utilities	6
E	Construction	5
F	Transportation	10
G	IT	8
H	Wholesale & retail trade	10
I	Finance and insurance	10
J	Real estate	11
K	Social Services	3
L	Communication & Cultural Industry	2
M	Comprehensive	3

3.2. Construction of stock networks

3.2.1. Price returns and correlations

From the time stamped price time series of a blanket of stocks, it is possible to calculate the correlations for any pair of stocks once a time window is given.

Table 2: 468 component stocks of S&P500 are included in the dataset. In this table, we list the Global Industry Classification Standard (GICS) code, sector name and number of stocks for each industry sector in S&P500. All 10 industry sectors are represented.

Industry code	Industry Sector	Number of Stocks
10	Energy	36
15	Materials	26
20	Industrials	63
25	Consumer Discretionary	78
30	Consumer Staples	33
35	Health Care	50
40	Financials	87
45	Information Technology	61
50	Telecommunication Services	5
55	Utilities	29

Table 3: Basic information of CSI163 and S&P468 datasets including the number of stocks, the number of trading days, and the average log returns $\langle r \rangle$, and the standard deviation σ_r of the log returns are presented. Compared to the US market, the Chinese market has higher fluctuations in our study period between 04/01/2007 and 06/11/2015.

Dataset	Stocks	Days	$\langle r \rangle$	σ_r	$\langle r_{\min} \rangle$	$\langle r_{\max} \rangle$
CSI163	163	2149	1.4795e-04	0.0340	-0.4832	0.1002
S&P468	468	2228	1.4691e-04	0.0252	-0.3413	0.2168

$P_i(t)$ is the price at time t of stock s_i. It could be one of the daily prices of open, close, high, or low. Per most literature suggests, we choose the most used daily close price. To smooth the fluctuation without loss of generality, the logarithmic return for s_i in the period of $[t - \Delta t, t]$ is defined as

$$Y_i(t) = \ln P_i(t) - \ln P_i(t - \Delta t), \tag{1}$$

and usually used instead of $P_i(t)$ itself. In most cases, daily log returns are used where $\Delta t = 1$. For stock pair of s_i and s_j, we can extract the two price time series in a time window with a length or size of L, *i.e.*, with L price values included in the window. The selection of L is expected to meet the requirement of $L/N > 1$. In a sliding window approach, we can extract subsets of prices in a serial of sliding windows: $[1, L]$, $[2, L+1]$, $\cdots$. For a given window, with the two time series of prices, it is possible to generate two log return time series using Eq. 1 for both stocks s_i and s_j. Thus, the correlation coefficient between two stocks can be calculated by using the Pearson correlation coefficient [14]

$$\rho_{ij} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{\left(\langle Y_i^2 \rangle - \langle Y_i \rangle^2\right)\left(\langle Y_j^2 \rangle - \langle Y_j \rangle^2\right)}}, \tag{2}$$

where $\langle \cdots \rangle$ stands for the average. The value of ρ_{ij} ranges between -1 and 1, where a negative value of $\rho_{ij} < 0$ indicates the two stocks fluctuate in a non-correlated manner, *i.e.*, one falls down while anther one climbs up. For a positive value of $\rho_{ij} > 0$, the two stocks fluctuate in a positively correlated way. In this case, they move in the same direction. If $\rho_{ij} \approx 0$, then they are not correlated. If $|\rho_{ij}| \approx 1$, then the two stocks are perfectly correlated or non-correlated. In a stock market, the stocks from the same industry are more likely to be correlated.

For a portfolio of N stocks $s_1, s_2, \cdots, s_N$, we can calculate all $N \times N$ pairs of correlation coefficients ρ_{ij} for any s_i and s_j. These N^2 pairs of values can be expressed as a correlation coefficient matrix C with a size of $N \times N$.

Based on the correlation matrix C, we can define the distance d_{ij} between stock pair of s_i and s_j as

$$d_{ij} = \sqrt{2\left(1 - \rho_{ij}\right)}. \tag{3}$$

The values of d_{ij} form an adjacent symmetric matrix D, in which there are $N(N-1)/2$ different elements. It is verified that this definition satisfies the three rules of Euclidean distance: (1) $d_{ij} = 0$ if and only if $i = j$; (2) $d_{ij} = d_{ji}$; (3) $d_{ij} \leqslant d_{ik} + d_{kj}$ [14]. Since $-1 \leqslant \rho_{ij} \leqslant 1$, we have $0 \leqslant d_{ij} \leqslant 2$.

With this definition, the distance for two stocks has a value of 2 when they are completely anti-correlated ($\rho_{ij} = -1$), and a small distance close to 0 when they are positively and completely correlated ($\rho_{ij} \to 1$). This makes it possible to compare the distances for any two pairs of stocks.

3.2.2. Network $N(V, E)$

With the adjacency matrix D, we can further construct the network $N(V, E)$ for the stocks, where stock s_i is represented as vertex $v_i \in V$, and $e_{ij} \in V$ represents the edge between v_i and v_j with a distance of d_{ij}. The network $N(V, E)$ is undirected in which only one edge exists between a pair of vertices. The size of the network is the number of stocks N. The possible maximum number of edges is $N(N-1)/2$ for an undirected complete network in which all vertex pairs are connected. For a portfolio with a large number of stocks N, the number of edges is a huge number, thus it is necessary to simplify the network by filtering less important edges. In a simple threshold approach, by introducing a threshold value θ, we can reduce the network by chopping those edges whose distance are greater than θ and keeping the remaining edges. In other words, we only retain the connections which are strong enough, *i.e.*, those with small distances, and all other weak edges with distances larger than the threshold θ are filtered as the following equation:

$$e_{ij} = \begin{cases} 1 & \text{if } d_{ij} < \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Or in an edge ranking approach, we only keep a certain number of top edges with the strongest relationships, say $N - 1$ edges. With this approach, the remaining edges are more likely to form loops in strongly connected vertices and are referred to as an *Asset Graph* [82].

3.2.3. Network filtering

By filtering edges in a threshold approach, we may get isolated vertices or loops in the filtered network. To avoid this, tree approaches including *Mini-*

mum Spanning Tree (MST) can be used to chop edges but still keep all vertices connected as a tree. MST is introduced to investigate the hierarchical structure of stock networks first by Mantegna [14]. Many studies also use this approach, such as Jang *et al.*, to investigate the foreign exchange market using in the periods of currency crises finding that the values of correlation coefficients decrease but the normalized tree length increase in crises [83]. Matteo *et al.* find that the dynamical *planar maximally filtered graphs* (PMFGs) can preserve same hierarchical structure as the dynamical MST, and the financial sector dominates the central role in the network [84]. As an application of network analysis in portfolio management, Onnela *et al.* suggest the assets of the classic Markowitz portfolio are always located on the outer leaves of the tree [79], and Pozzi *et al.* further suggest that even it's better to invest in the peripheries of the MST of a market [35]. In Ref. [76], MST networks extracted from real correlation data are compared with those generated from artificial random models. Results reveal that the properties of MST from real data cannot be reproduced, showing the uniqueness of real stock networks.

Based on the network $N(V, E)$, we can extract a tree connecting all vertices with $N-1$ edges with a minimum total distance also known as Minimum Spanning Tree (MST) of the stocks. By only using the $N-1$ edges out of the maximum $N(N-1)/2$ edges, the network is dramatically simplified or filtered while keeping the most important shortest edges. To extract the MST, Kruskal's algorithm was applied in three steps: (1) rank all edges according to the distances from the shortest to the longest; (2) in each round, we choose the shortest edge into the MST while avoiding loops; (3) repeat round #2 until all vertices and all $N-1$ shortest edges are added [85]. Bonanno *et al.* review the MST approach in revealing information of markets [86].

Using the MST, we can construct the *Hierarchical Tree* (HT) in which the subdominant ultra-metric distance $d_{ij}^{<}$ is defined as the maximum distance of an edge along the path between v_i and v_j. The HT satisfies the first two rules

with a stronger third one:

$$d_{ij}^{<} \leqslant \max\left(d_{ik}^{<}, d_{kj}^{<}\right), \tag{5}$$

With this ultra-metric inequality, we can construct a hierarchical tree based on a MST and present a unique topological structure of the stocks [14, 87].

By loosening the requirements of MST up to 4 vertices, but forbid crossings, as many as $3\,(N-2)$ edges containing the MST as the subgraph including all the top $N-1$ shortest edges can be gathered. This new network can be drawn on a planar surface without link crossings is called *Planar Maximally Filtered Graphs* (PMFG) [77, 84, 88–91]. This makes PMFG different from MST, which also shows richer structures of the network. In a similar construction to MST, to construct PMFG, we firstly rank the edges in ascending orders according to the distances of edge pairs. Then we add the shortest edges into the PMFG but keeping the genus $g = k$, where the g is the largest number of closed simple curves one can draw on a planar surface without separating it. For the case of $g = 0$, when all edges are considered, PMFG can be gathered [77]. It has also been proved that a MST is a subgraph of a PMFG and the number of 3- and 4-cliques in a PMFG is $3n-8$ and $3n-4$, respectively [92].

Since PMFG contains more edges and allows loops and cliques, there are more information embedded in PMFG than in MST. After the introducing of PMFG into the study of network structures of stocks, PMFG has been used in studies of many stock markets, and more recently, PMFG is applied in investment strategy design [35]. Based on PMFG, a clustering approach called *Directed bubble hierarchical tree* (DBHT) is proposed and show a good performance compared with other algorithms and also been applied to study financial data [93]. It has been reported that, in a running window approach, the PMFG shows stronger stability in a long run compared with MST [94].

4. Stock Network Topological Properties

A network $N = (V, E)$ is a graph composed by a set of vertices V and a set of edges E. In a network model, the participants are represented as the

vertices V and the relationship between any pair of two participants i and j is represented as the edge e_{ij} connecting the two vertices v_i and v_j. In this study, the following properties of financial markets are researched: (1) *Degree and Degree Distribution* which describes the connectivities of vertices; (2) *Clustering Coefficient* which is the indication of the transitivity and density of a network; (3) *Average Path Length*, which is a global property indicating how the network spans; (4) *Betweenness Centrality* which describe the global importance or centrality of vertices or edges; (5) *Components* which describe the grouping phenomena of sub-structures of the networks.

Based on how the correlations are calculated, there are two approaches, static or dynamic. In a static approach, the correlations are calculated over the whole period using all available prices, thus we get a single static correlation matrix to describe the market regardless the different market periods. While sliding windows are used in a dynamic approach, the correlation matrix based on different sliding window evolves over time generating a set of correlation matrices. The static approach, which is the most used in literature, gives a static description of the structure of the market with details of different market periods like bear markets or bull markets. However, the dynamic approach can reveal the evolution of market structures and behaviors, which are especially useful for the comparisons of calm periods and crashes.

In this part, we present the topological properties of stock networks of the two markets, CSI163 and S&P468, in a dynamic approach. Considered to meet the requirement of $L/N > 1$, we set the sliding window size $L_{CSI163} = 170$ for CSI163 and $L_{S\&P468} = 500$ for S&P468. Totally, there are 2149 windows for CSI163 and 2228 windows for S&P468. After calculating the log returns for both CSI163 and S&P468 by using Eq. 1, we calculate the correlation coefficient matrices over the period between 12/09/2007 and 06/11/2015 for CSI163, 26/12/2008 and 06/11/2015 for S&P468 using Eq. 2. Based on the correlation matrices, it is straight to get stock networks. For CSI163, we have a network of 163 vertices, and for the S&P468, we have a network of 468 vertices. The edge connecting two stocks indicates how the two stocks behave correlatively

or anti-correlatively. For a positive correlation coefficient value, the two prices move in the same direction, while for a negative value, the two prices move in opposite directions, so to normalize all correlation coefficients to positive values as edge distances, we adopt the definition of distance based on Eq. 3. Through this definition, all negative values are transformed to positive distance values and the order of values are preserved. All vertices in the networks for both markets are fixed. However, the edges vary in each sliding window as the correlation coefficients change. In the following parts, the statistical properties of both networks evolve in our study period is investigated.

4.1. Degree and degree distribution

For a network of N stocks, there are $N \times N$ edges, which is a huge number for a large N. So we normally filter the weakest edges in order to simplify the network. In the threshold approach, a threshold θ can be used to chop the edges, if $d_{ij} > \theta$. For a given network, different θ can lead to different structures with same vertices but different sets of remaining edges. Based on the correlation matrices, we first investigate the stock networks with different θ for both CSI163 and S&P468. In the sliding window approach, using daily log return time series, we first calculate the correlation matrices of 163×163 for CSI163 and 468×468 for S&P468, then we average all the correlation matrices over the study periods. After that, we get the averaged correlation matrices, with which we can apply the edge filtering process for different θ based on the Eq. 4. Based on the result, small θ closes to 0 will filter most edges while larger θ close to the maximum value 2 will keep most edges. We use an θ interval of $[0.1 - 1.5]$ with a step of 0.1. We present the basic network properties in Table 4 for the CSI163 network and Table 5 with different θ between 0.1 and 1.5. The maximum possible edges are 13203 for CSI163 and 109278 for S&P468, respectively. For different distance thresholds θ, any edges whose distances are greater than the threshold are filtered. So with smaller θ, only a few edges remain in the network and this results a smaller edge density $|e|_{density}$, smaller average degree $\langle d \rangle$, average distance $\langle d_{ij} \rangle$, minimum distance $d_{ij}^{\min}$, and maximum distance $d_{ij}^{\max}$ as

well.

Table 4: For the CSI163 network, the maximum possible number of edges of $|e|_{\max}$ for $N = 163$ vertices, the existing edge number $|e|$, the edge density $|e|_{density}$, the average degree $\langle d \rangle$, the average distance $\langle d_{ij} \rangle$, the minimum distance $d_{ij}^{\min}$, the maximum distance $d_{ij}^{\max}$ are presented for different θ from 0 to 1.5 in a step of 0.1.

θ	$\lvert e\rvert_{\max}$	$\lvert e\rvert$	$\lvert e\rvert_{density}$	$\langle d \rangle$	$\langle d_{ij} \rangle$	$d_{ij}^{\min}$	$d_{ij}^{\max}$
0.1	13203	0	0.0000	0.0000	0.0000	0.0000	0.0000
0.2	13203	0	0.0000	0.0000	0.0000	0.0000	0.0000
0.3	13203	2	0.0002	0.0245	0.0000	0.0049	0.0089
0.4	13203	30	0.0023	0.3681	0.0000	0.0010	0.0593
0.5	13203	91	0.0069	1.1166	0.0003	0.0003	0.2200
0.6	13203	276	0.0209	3.3865	0.0010	0.0003	0.3689
0.7	13203	1594	0.1207	19.5583	0.0038	0.0003	0.4291
0.8	13203	5620	0.4257	68.9571	0.0177	0.0004	0.5353
0.9	13203	10555	0.7994	129.5092	0.0666	0.0004	0.7098
1	13203	12816	0.9707	157.2515	0.1866	0.0005	0.7996
1.1	13203	13194	0.9993	161.8896	0.4052	0.0006	0.9443
1.2	13203	13203	1.0000	162.0000	0.6957	0.0902	1.0887
1.3	13203	13203	1.0000	162.0000	0.9637	0.3719	1.1910
1.4	13203	13203	1.0000	162.0000	1.0966	0.5515	1.2878
1.5	13203	13203	1.0000	162.0000	1.1141	0.5515	1.3230

In Fig. 3, we plot the edge densities of CSI163 and S&P468 for different thresholds θ from 0.1 to 1.5. In the interval of 0.1 to 0.6, the densities for both networks are close to 0, meaning all edges are filtered. While in the interval of 1 to 1.5, the densities are close to 1, meaning that all edges are preserved. Between these two intervals, we see that the two curves have a similar shape which show a slope when θ lies between 0.6 and 1. This indicates that most edges are within this interval. A similar edge density distribution is also reported in [56]. The study shows that stock networks also demonstrate a similar transforming

Table 5: For the S&P468 network, the max possible edges $|e|_{\max}$ for $N = 468$ vertices, the existing edge number $|e|$, the edge density $|e|_{density}$, the average degree $\langle d \rangle$, the average distance $\langle d_{ij} \rangle$, the minimum distance $d_{ij}^{\min}$, the maximum distance $d_{ij}^{\max}$ are presented for different θ from 0 to 1.5 in a step of 0.1.

θ	$\lvert e\rvert_{\max}$	$\lvert e\rvert$	$\lvert e\rvert_{density}$	$\langle d \rangle$	$\langle d_{ij} \rangle$	$d_{ij}^{\min}$	$d_{ij}^{\max}$
0.1	109278	0	0.0000	0.0000	0.0000	0.0000	0.0000
0.2	109278	2	0.0000	0.0085	0.0000	0.0935	0.1094
0.3	109278	2	0.0000	0.0085	0.0000	0.1959	0.2033
0.4	109278	18	0.0002	0.0769	0.0000	0.0005	0.2033
0.5	109278	137	0.0013	0.5855	0.0001	0.0003	0.3265
0.6	109278	729	0.0067	3.1154	0.0007	0.0003	0.4980
0.7	109278	3571	0.0327	15.2607	0.0038	0.0004	0.6125
0.8	109278	16433	0.1504	70.2265	0.0213	0.0004	0.7255
0.9	109278	50364	0.4609	215.2308	0.0952	0.0005	0.8122
1	109278	88680	0.8115	378.9744	0.2739	0.0005	0.9264
1.1	109278	106179	0.9716	453.7564	0.5356	0.0006	1.0334
1.2	109278	108956	0.9971	465.6239	0.7994	0.0007	1.1486
1.3	109278	109277	1.0000	466.9957	0.9906	0.0044	1.2480
1.4	109278	109278	1.0000	467.0000	1.0711	0.1959	1.3465
1.5	109278	109278	1.0000	467.0000	1.0711	0.1959	1.3465

interval.

We investigate the degree distributions of both networks with different θ. No matter if θ is too small or too large, the degree distributions are noisy, while in a narrow interval around 0.7, the distributions follow the power law. The regression fitting curve is a straight line in the plots of $\log P(k)$ against $\log(k)$, where the $\log P(k)$ is the log_{10} probability for a vertex with k degrees and the $\log(k)$ is the log_{10} degree. After running on the data, we plot the typical power law distributions in Fig. 4(a) for CSI163 and Fig. 4(b) for S&P468 respectively. For both distributions, we fit the log_{10}-log_{10} distribution and get the power law

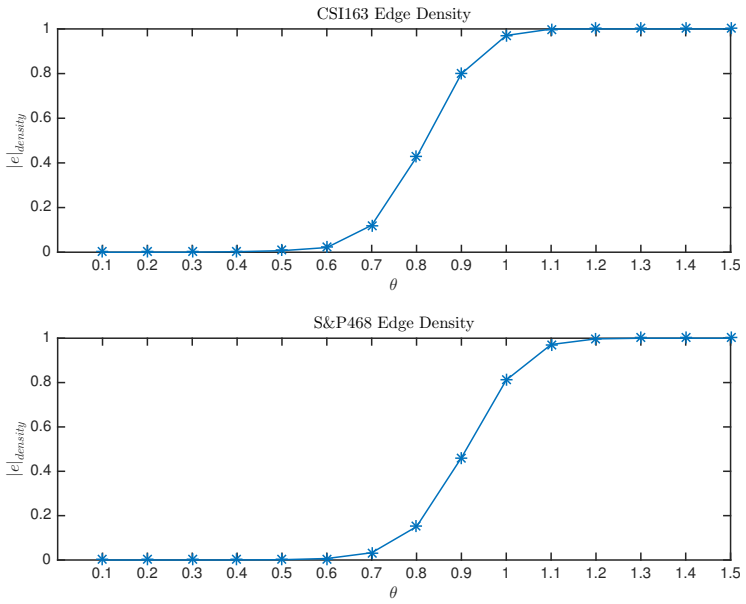
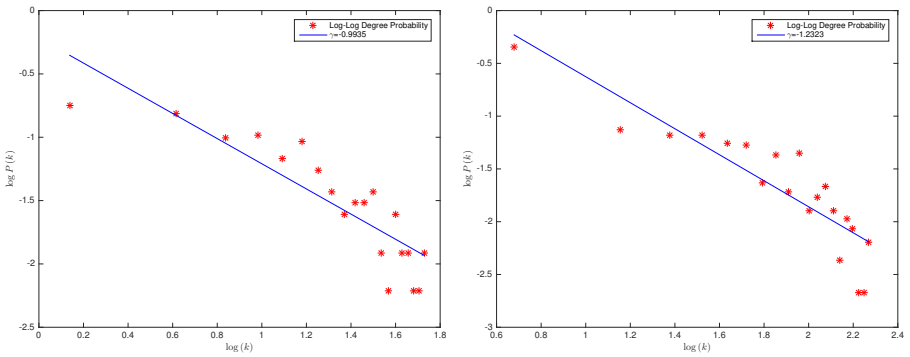


Figure 3: Edge densities of CSI163 and S&P500 for different thresholds θ from 0.1 to 1.5. It shows that the densities increase sharply from 0 to 1 in the θ interval of 0.6 and 1.

exponents $\gamma = -0.9935$ for CSI163 and $\gamma = -1.2323$ for S&P468. In the plots, we use the same $bins = 20$ to calculate the probabilities for different degrees. As shown in Fig. 4, we see that a large number of vertices have small degrees. Only a few number of vertices have large degrees. As the vertices are stocks, and the degrees are rooted the correlations between stocks, for both CSI163 and S&P468 networks, only a few stocks are the highly correlated with the most parts of rest stocks. These stocks have a wider and larger influence over the whole networks, while other stocks with relatively smaller degrees are less correlated with the rest parts of the stocks. This presents very limited influence over the network. The negative fitting slope value γ also indicates that both the CSI163 and S&P468 networks are scale free networks in which a small portion of vertices have larger degrees while a large portion of vertices have smaller degrees this agree with previous studies [95].



(a) CSI163 (b) S&P468

Figure 4: Log-Log degree distributions of CSI163 and S&P468 networks. By using different θ, we can filter out edges with larger distances. We find that not all the filtered networks can demonstrate the power law degree distributions. Only when θ falls within a narrow interval around 0.7, the filtered networks follow the power law degree distribution. In Fig. 4(a), for the CSI163 network, we use $\theta = 0.68$ and we get a fitting line with a slope of $\gamma = -0.9935$. In Fig. 4(b), for the S&P468 network, we use $\theta = 0.75$ and we get a fitting line with a slope of $\gamma = -1.2323$. This indicates that the degree distributions of both stock markets follow the power law in the form of $P(k) \sim k^{-\gamma}$, which also means the networks are scale free in which a small portion of vertices have larger degrees, while a large portion of vertices have small degrees.

4.2. Average clustering coefficient

Average clustering coefficient $\langle C \rangle$ is an average of all clustering coefficients $\langle C_i \rangle$ of all vertices. The clustering coefficient $\langle C_i \rangle$ indicates the transitivity for an individual vertex v_i, while the overall averaged clustering coefficient $\langle C \rangle$ is an indication of the transitivity and density of the whole network. In Fig. 5, we present the average clustering coefficient $\langle C \rangle$ for both CSI163 and S&P468 networks comparing with random networks. The $\langle C \rangle$ gets larger with the θ when larger θ will preserve more edges, while it remains almost unchanged with a slightly increase in both random networks. Comparing with random networks of same sizes of 163×163 for CSI163 and 468×268 for S&P468, the $\langle C \rangle$ of both two stock networks are significantly larger than that of the corresponding random networks. For CSI163, the $\langle C \rangle$ is 4.9574 times larger than that of the

random networks on average with a maximum of 11.8903 times. For S&P468 the average multiple is 5.2305 times and maximum multiple is 10.4180 times compared with the random networks. This shows that both stock networks are well connected with better transitivities. This result agrees with many other previous studies.

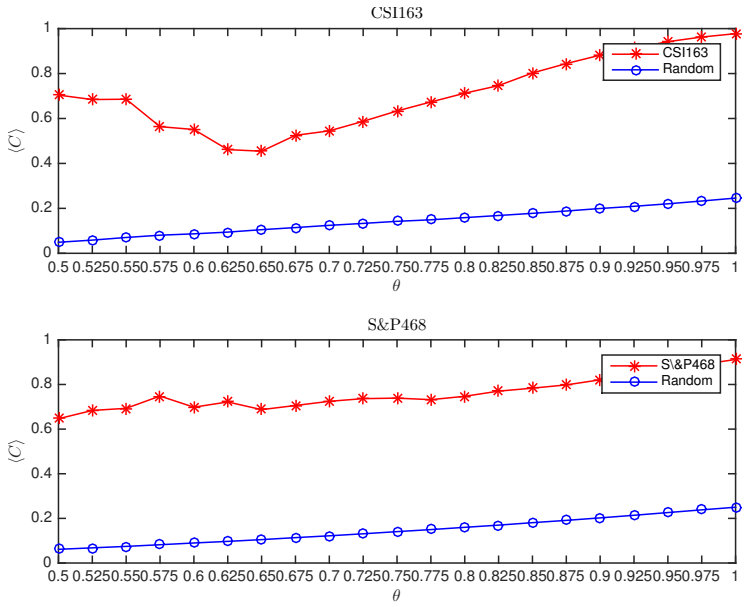


Figure 5: Average clustering coefficient $\langle C \rangle$ of CSI163 and S&P468 for different thresholds θ. It shows that the $\langle C \rangle$ gets larger with θ. To compare with random networks, we plot the corresponding average clustering coefficients under the same interval of θ. As it shown that for both CSI163 and S&P468 networks, the $\langle C \rangle$ values are significantly larger than the random networks of same size. This indicates the stock markets are far from random and the stocks are comparatively clustered.

4.3. Average path length

Unlike clustering coefficient which is a local property, for any two vertices v_i and v_j in a network, the number of edges covering the shortest route linking the two vertices are defined as the characteristic path length, l_{ij}, in [2] which is a typical global property. By averaging the lengths of all possible pairs, we can calculate the average path length $\langle L \rangle$. As an indication of how the

network is connected, many real networks have small $\langle L \rangle$ compared with random networks. In Fig. 6, we plot the average path length $\langle L \rangle$ of both CSI163 and S&P468 networks with comparisons of random networks in same sizes. The two stock networks are significantly different from the corresponding random networks with same sizes of 163×163 and 468×468. While the flat curves of $\langle L \rangle$ of random networks stay almost unchanged with the θ, this is a result of the universal homogeneous edge distribution on the whole network. There are peaks for the stock networks. On the left-hand of the peak, there is a decline of $\langle L \rangle$ with the decrease of θ, since when θ gets too small, most edges are filtered and the giant networks breaks into small parts and the $\langle L \rangle$ in small parts are decreasing significantly. But for the right-hand of the peak, the $\langle L \rangle$ gets smaller with the increase of θ due to the increasing connectivity when more and more edges are preserved. This shows the stock networks of both CSI163 and S&P468 are different from random networks.

4.4. Betweenness centrality

The betweenness b_{v_i} of vertex v_i is defined as the number of shortest paths passing v_i, which is an indication of the importance of an individual vertex in the contribution to the global connectivity. By averaging over the betweenness of all vertices, we can compare the betweenness for any two vertices. Basically, larger betweenness means greater global influence of the stock networks. This is the same to the edges. The betweenness $b_{e_{ij}}$ is the number of shortest paths passing the edge e_{ij} indicating the importance of this edge for its contribution to the global connectivity. In this study, we focus on the vertex betweenness. In the calculation of the shortest paths, we can use the original distance d_{ij} defined in Eq. 3 or simplify the network as a binary network according to

$$e_{ij} = \begin{cases} 1 & d_{ij} > 0, \\ 0 & d_{ij} = 0. \end{cases} \tag{6}$$

In the former, edges with different distances have different contributions to the paths, while for the latter, all edges of non-zero distance are normalized as 1

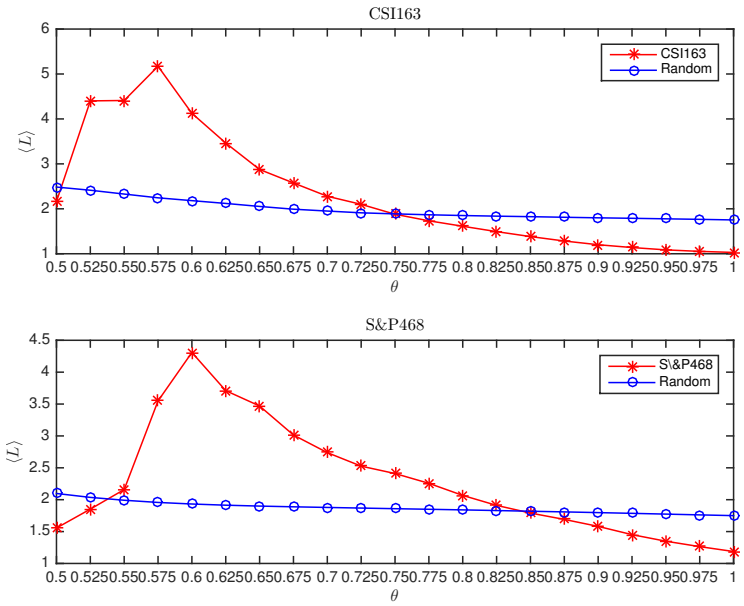
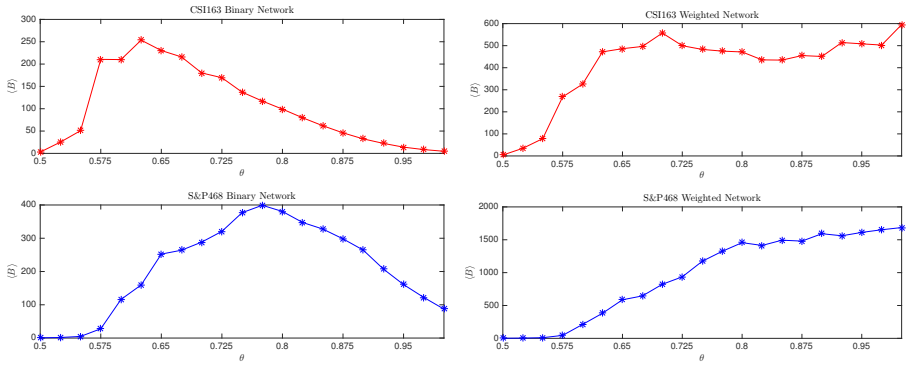


Figure 6: Average path lengths $\langle L \rangle$ for CSI163 and S&P468 under different θ compared with values for random networks in same sizes of 163×163 and 468×468. It shows that for the two networks, there are peaks of the $\langle L \rangle$ above the curve of the corresponding random networks. At first, the θ is small, most edges are filtered and the whole networks are broken into parts thus the disconnected vertices and edges are also filtered. So on the left hand, starting from the peak, with the decrease of θ, we see that $\langle L \rangle$ decreases too. While starting from the peak, with the increase of θ, we see a constantly decline of $\langle L \rangle$, for more edges are remained resulting as decreasing of $\langle L \rangle$. While for the random networks, the $\langle L \rangle$ stay almost unchanged because of the homogeneous edge distributions across the whole networks.

and treated equally with great scarifying of original distance information. As shown in Fig. 7, we plot the average betweenness $\langle B \rangle$ of CSI163 and S&P468 for both binary case and weighted case under different θ in Fig. 7(a) and Fig. 7(b) respectively. For binary network case, all edges with positive distances are normalized as unit 1, while in weighted networks, the original distances are directly used in the calculation of shortest paths. The shapes for binary and weighted networks are different. There are peaks for both stock networks in binary networks, while the $\langle B \rangle$ gets larger with the θ from almost zero to large numbers in weighted networks.



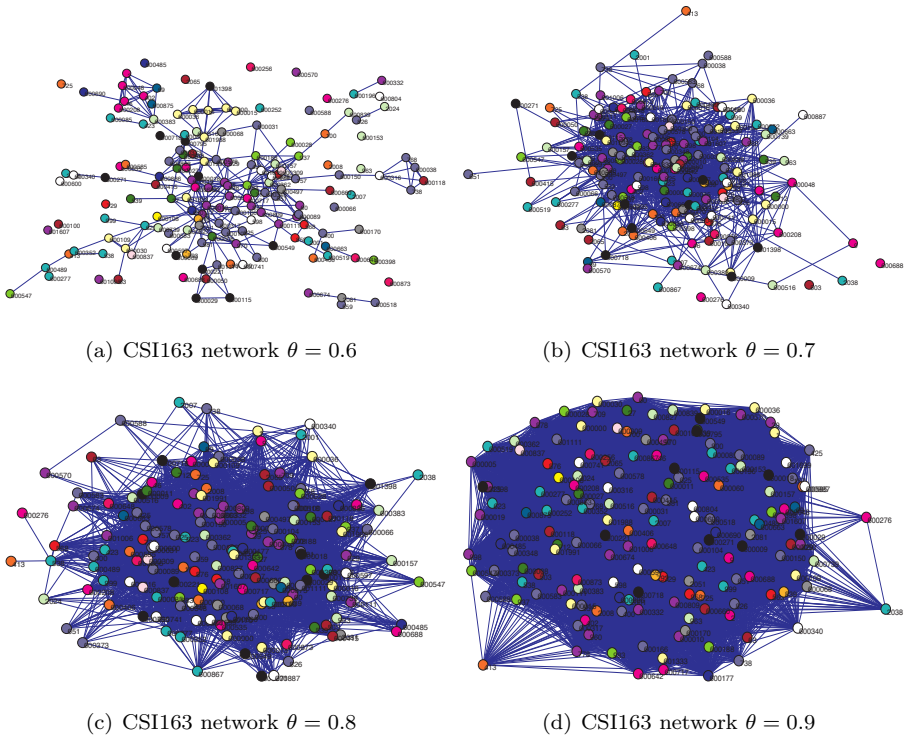
(a) Binary Network (b) Weighted Network

Figure 7: Average betweenness $\langle B \rangle$ of CSI163 and S&P468 are calculated for both the original weighted approach and binary simplification approach under different θ. The average betweenness for binary networks are different from the weighted network. For binary networks, the curves for CSI163 and S&P468 share a similar shape. On the left-hand of the peak, the $\langle B \rangle$ gets larger with the increase of θ, for more edges and more vertices are preserved and this leads to a growing number of paths. While on the right-hand of the peak, large connected network emerges leading to a small value of averaged $\langle B \rangle$. In other words, the importance of a single individual vertex or edge is weakened in well connected networks (large θ).

We visualize the stock networks of CSI163 and S&P468 with different θ of 0.6, 0.7, 0.8, 0.9 in Fig. 8(a)-Fig. 8(d) and Fig. 9(a)-Fig. 9(d), respectively. It shows that the networks can be dramatically simplified using small values of θ and the edges are preserved in larger values of θ. As listed in Table 4, the edge density of CSI163 grows dramatically from 0.0209 ($\theta = 0.6$) to 0.7994 ($\theta = 0.9$), while for S&P468 as shown in Table 5, the edge density of S&P468 grows also dramatically from 0.0067 ($\theta = 0.6$) to 0.4609 ($\theta = 0.9$). All networks in this paper are generated using the Pajek complex network software [96].

4.5. Components

A component, Com is a sub-network of the whole network with connected vertices. For a given network with a set of N vertices, the possible size of Com can range from 1 for isolated vertex to N for all connected vertices. When an individual vertex v_i is disconnected with any other vertices, v_i itself form



(a) CSI163 network $\theta = 0.6$

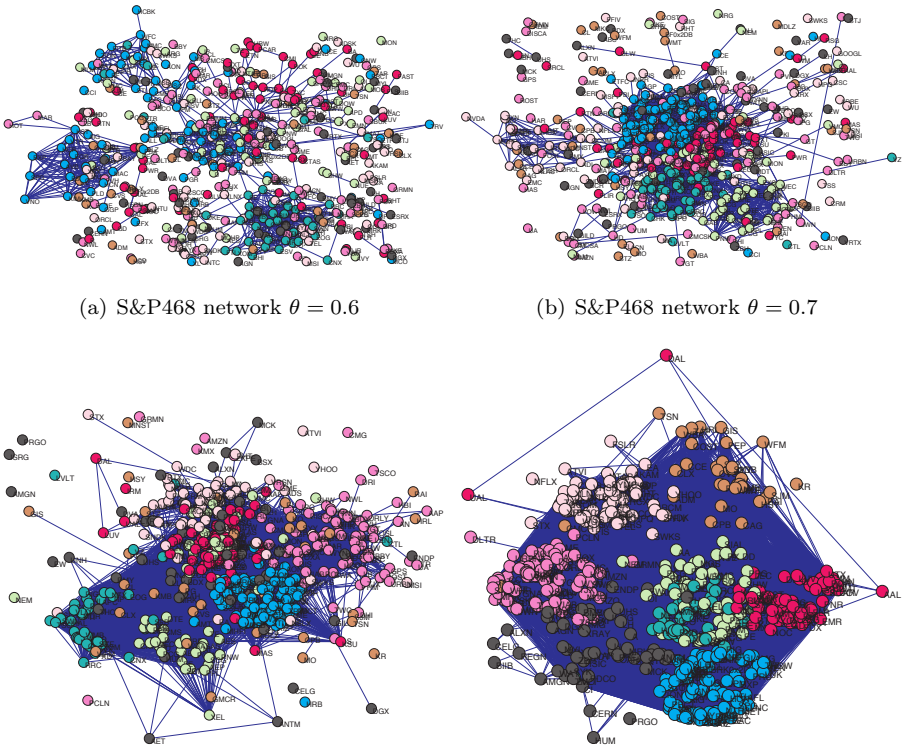
(b) CSI163 network $\theta = 0.7$

(c) CSI163 network $\theta = 0.8$

(d) CSI163 network $\theta = 0.9$

Figure 8: CSI163 networks with different θ of 0.6(a), 0.7(b), 0.8(c), 0.9(d). It shows that, the network is relatively sparser with small θ while denser with large θ, for small θ greatly simplifies the network by filtering most edges with larger distance. As a result, the edge density when $\theta = 0.9$ is about 38.25 times to that when $\theta = 0.6$. Different vertex colors indicate different industry sectors.

a smallest component with single vertex. When all the vertices are connected without any isolated vertices, the network is a single giant component. For a stock network, the stocks are correlated with each other, while stocks belonging to different components are not correlated. The component structures of stock networks have great implications on risk management of a portfolio. Since the stocks fall in the same component are correlated, so it's a bad idea to invest in most stocks from the same components. We should invest stocks from different components to diverse the risk of the whole portfolio. When the θ is small, most edges are filtered leaving many vertices isolated. As a result we see



(a) S&P468 network $\theta = 0.6$

(b) S&P468 network $\theta = 0.7$

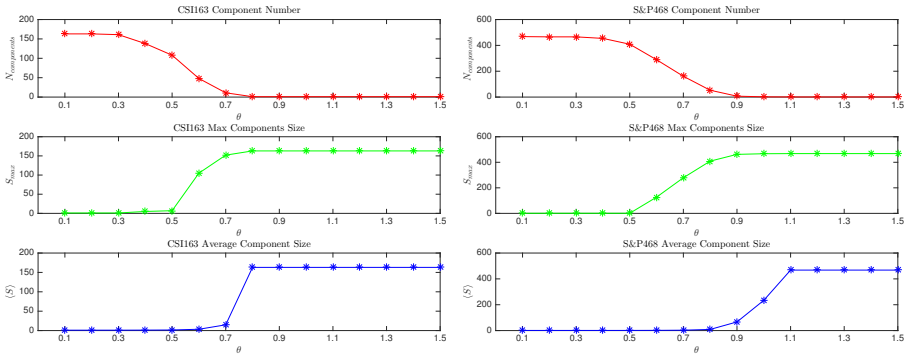
(c) S&P468 network $\theta = 0.8$

(d) S&P468 network $\theta = 0.9$

Figure 9: S&P468 networks with different θ of 0.6(a), 0.7(b), 0.8(c), 0.9(d). It also shows that the edge densities with different θ change dramatically. We find that the edge density when $\theta = 0.9$ is 68.79 times to that when $\theta = 0.6$. Different vertex colors indicate different industry sectors.

the emerging of large number of small components. While with the growth of θ, more and more edges are preserved. This allows the connectivity increase resulting in the appearing of larger components. In Fig. 10, the properties of components of CSI163 and S&P468 networks with different θ is presented. For the two networks, the number of components $N_{components}$ (red), the max component size S_{max} (green), and the average component size $\langle S \rangle$ (blue) shows similar pattern and changes with the values of θ. Critical changes are obvious for both networks in the θ interval about $[0.3 - 0.8]$ for CSI163 and in $[0.5 - 1.1]$ for S&P468, respectively. Before the transition interval, most vertices are

isolated when the whole network breaks into small components, and both of the maximum and average component sizes are small. In the transition interval, the number of components decrease with both maximum and average component sizes. After the transition interval, the three properties stay unchanged when the giant connected component appear with maximum and average size equal to the number of total vertices. The similar component properties transition under different of θ phenomena is also observed in the study of a set of Chinese stocks [56] with a reported transition critical value about $\theta = 0.17$.



(a) CSI163 Network (b) S&P468 Network

Figure 10: The component properties of the components number $N_{components}$ (red), the max component size S_{max} (green), and the average component size $\langle S \rangle$ (blue) are plotted for CSI163 and S&P468 with different thresholds θ in (a) and (b), respectively. Both networks show similar transitions when the networks transform from large number of isolated small components into a connected giant network. Before the transition interval, edges are filtered leaving isolated vertices are not correlated. After the transition interval, edges are preserved making most vertices connected to form a single giant network in which all vertices are correlated.

To investigate how industry sectors are connected in the stock network, we summarize the properties of both CSI163 and S&P468 networks with $\theta = 1.0$ listed in Table 6 and Table 7, respectively. As it shows, in the CSI163 network, the industry sectors are all most the same in average degree $\langle d \rangle$ and average clustering coefficient $\langle C \rangle$, while with significantly different values of average betweenness coefficient $\langle B \rangle$. The difference of average degree $\langle d \rangle$ and

average clustering coefficient $\langle C \rangle$ are not significant among the industry sectors. This indicates that all sectors have similar degrees and clustering coefficients. The difference of average betweenness coefficient $\langle B \rangle$ shows that the sectors contribute to the global connectivity differently. It is worth mentioning that the Finance and insurance sector has the largest average clustering coefficient $\langle C \rangle$ of 0.9829 but with a relatively small value of average betweenness coefficient $\langle B \rangle$ which is only 118.4000. For the S&P468 network, as shown in Table 7, we observe that Financials sector has the largest value of $\langle d \rangle$ of 421.7471 and the 3rd largest value of $\langle B \rangle$ of 1297.8161. While with a smaller value of average clustering coefficient $\langle C \rangle$ of 0.8975, which are very different from the CSI163 network. Furthermore, the Energy and Industrials have the largest values of $\langle d \rangle$ and $\langle B \rangle$, while Consumer Staples and Telecommunication Services have the smallest $\langle d \rangle$ and $\langle B \rangle$. From this, we also observe that for S&P468 network, the sectors with larger $\langle B \rangle$ are likely to have smaller values of $\langle C \rangle$ and vise verse. The findings indicate that the US market is dominated by Financials while the Finance and insurance in Chinese stock markets play relatively less influential roles.

By focusing on the top stocks, it is possible to look into the details of the networks. In Table 8, for the CSI163 network with $\theta = 0.8$, we present the top 10 stocks with the largest values of degree d_i and betweenness b_i ranked in descending order in the upper part and in the lower part, respectively. The stock code, company name, industry name, and values of d_i, c_i, b_i are listed. Younger Group, which is a leading fashion brand in China has the largest degree of $d_i = 133$, and HuDong Heavy Machinery, which is a major machinery manufacturer in China has the largest betweenness coefficient of $b_i = 5444$. While in the S&P468 network, as shown in Table 9, T. Rowe Price Group has the largest degree of 266, and Loews Corp. has the largest betweenness coefficient value of 13202, both stocks are in the Financials sector. For both stock networks, the two lists based on d_i and b_i are similar. In other words, top stocks with largest degree values also appear as top stocks with largest betweenness coefficients b_i. It's worth noting that the list based on the ranking of clustering coefficients c_i are

Table 6: In this table, we list the China Securities Regulatory Commission (CSRC) industry code, the sector name and the numbers of stocks, the average degree $\langle d \rangle$, the average clustering coefficient $\langle C \rangle$, and the average betweenness coefficient $\langle B \rangle$ for each industry sector of these 163 stocks. The values are calculated from the CSI163 network with $\theta = 1.0$.

Code	Industry Sector	Number	$\langle d \rangle$	$\langle C \rangle$	$\langle B \rangle$
A	Agriculture	1	162.0000	0.9703	754.0000
B	Mining	6	154.3333	0.9803	974.6667
C0	Food & Beverage	4	160.2500	0.9738	770.0000
C1	Textiles & Apparel	4	159.2500	0.9764	437.5000
C3	Paper & Printing	2	159.0000	0.9757	343.0000
C4	Petrochemicals	9	157.4444	0.9776	417.5556
C5	Electronics	7	157.1429	0.9783	176.8571
C6	Metals & Non-metals	20	159.9500	0.9756	808.0000
C7	Machinery	27	157.5926	0.9772	664.7407
C8	Pharmaceuticals	15	151.6000	0.9764	324.2667
D	Utilities	6	159.6667	0.9770	199.6667
E	Construction	5	160.2000	0.9768	945.2000
F	Transportation	10	157.9000	0.9792	426.2000
G	IT	8	156.3750	0.9775	444.0000
H	Wholesale & retail trade	10	158.8000	0.9775	341.2000
I	Finance and insurance	10	155.8000	0.9829	118.4000
J	Real estate	11	153.2727	0.9791	302.3636
K	Social Services	3	161.0000	0.9726	109.3333
L	Communication	2	160.0000	0.9775	821.0000
M	Comprehensive	3	159.6667	0.9777	582.6667

dramatically different those based on degrees or betweenness coefficients. This indicates that degree d_i and the betweenness b_i are consistent in describing the importance of an individual vertex, since the higher degree a vertex has, more likely it's on the shortest paths. As indicated in the two tables, the stocks with

Table 7: In this table, we list the industry code, the sector name, the numbers of stocks, the average degree $\langle d \rangle$, the average clustering coefficient $\langle C \rangle$, and the average betweenness coefficient $\langle B \rangle$ for each industry sector of S&P468 stocks. The values are calculated for the S&P468 network with $\theta = 1.0$.

Code	Industry Sector	Number	$\langle d \rangle$	$\langle C \rangle$	$\langle B \rangle$
10	Energy	36	419.6944	0.8948	1264.6111
15	Materials	26	404.2692	0.8948	1717.8462
20	Industrials	63	415.8571	0.8894	1937.0476
25	Consumer Discretionary	78	376.0256	0.9194	786.8205
30	Consumer Staples	33	269.0303	0.9194	37.8182
35	Health Care	50	309.6000	0.9135	213.8000
40	Financials	87	421.7471	0.8975	1297.8161
45	Information Technology	61	373.9344	0.9245	363.6721
50	Telecommunication	5	293.2000	0.9456	124.8000
55	Utilities	29	375.3448	0.9142	579.2414

codes labeled in bold appear on both top 10 lists, and in fact, the rest stocks on one list also can be found appearing in a similar ranking position on another list. We can also observe that stocks belong to Industries of Metals & Non-metals, Machinery, and Pharmaceuticals are dominant the two top 10 lists in CSI163 network. However, for S&P468 network, Financials, Industrials, and Materials are major stocks in the two lists. As an emerging market, Industrials sector has great influence in Chinese stock market, while the Financials sector has greater influence in US stock market which agree with [7]. The significant difference for the two stock markets confirms the previous studies like [97] in which with similar results indicating that Industrials is the most influential sector among all industry sectors, while the financial sector has weaker influence. This is consistent with our previous results revealed Table 6 and Table 7.

Table 8: Top stocks with highest values of degree d_i, and betweenness b_i ranked in descending order of d_i and b_i when the $\theta = 0.8$ for CSI163 network. Stock codes in bold indicate the stocks appear at both top 10 stocks.

Code	Name	Industry	$d_i \downarrow$	c_i	b_i
600177	Youngor Group	Textiles & Apparel	133	0.5416	2154
600642	Shanghai Wanye Enterprises	Real estate	131	0.5732	2362
39	China International Marine Containers (Group)	Metals & Non-metals	127	0.5929	2276
600010	Inner Mongolia Baotou Steel Union	Metals & Non-metals	125	0.5738	1112
600166	Beiqi Foton Motor	Machinery	123	0.6047	72
825	Shanxi Taigang Stainless Steel	Metals & Non-metals	122	0.6005	2644
623	Jilin Aodong Medicine Industry (Groups)	Pharmaceuticals	121	0.6058	60
600362	Beijinghualian Hypermarket	Wholesale and retail trade	121	0.6127	1552
600717	Qinhuangdao Yaohua Glass	Real estate	119	0.6095	1354
600005	Wuhan Iron And Steel	Metals & Non-metals	118	0.6009	532
Code	Name	Industry	d_i	c_i	$b_i \downarrow$
600150	HuDong Heavy Machinery	Machinery	110	0.6155	5444
898	Angang Steel	Metals & Non-metals	114	0.6367	4560
600398	Anyuan Industrial	Mining	95	0.6710	4140
2051	China CAMC Engineering	Construction	114	0.6226	4136
601607	Aluminum Corporation of China Limited	Metals & Non-metals	116	0.6304	2766
825	Shanxi Taigang Stainless Steel	Metals & Non-metals	122	0.6005	2644
600031	Sany Heavy Industry	Machinery	85	0.7300	2584
600642	Shanghai Wanye Enterprises	Real estate	131	0.5732	2362
39	China International Marine Containers (Group)	Metals & Non-metals	127	0.5929	2276
600177	Youngor Group	Textiles & Apparel	133	0.5416	2154

Table 9: Top stocks with highest values of degree d_i, and betweenness b_i ranked in descending order of d_i and b_i when the $\theta = 0.8$ for S&P468 network. Stock codes in bold indicate the stocks appear at both top 10 stocks.

Code	Name	Industry	$d_i \downarrow$	c_i	b_i
TROW	T. Rowe Price Group	Financials	266	0.3882	4536
L	Loews Corp.	Financials	263	0.4069	13202
SNA	Snap-On Inc.	Consumer Discretionary	263	0.3894	8350
IVZ	Invesco Ltd.	Financials	261	0.4016	2868
BEN	Franklin Resources	Financials	259	0.4074	7836
HON	Honeywell Int'l Inc.	Industrials	256	0.4151	1790
AMG	Affiliated Managers Group Inc	Financials	255	0.4128	2042
DD	Du Pont (E.I.)	Materials	255	0.4163	4966
SIAL	Sigma-Aldrich	Materials	254	0.4186	7020
ROP	Roper Industries	Industrials	247	0.4326	8626
Code	Name	Industry	d_i	c_i	$b_i \downarrow$
L	Loews Corp.	Financials	263	0.4069	13202
HST	Host Hotels & Resorts	Financials	223	0.4938	12510
OKE	ONEOK	Energy	174	0.5329	9978
ROP	Roper Industries	Industrials	247	0.4326	8626
SNA	Snap-On Inc.	Consumer Discretionary	263	0.3894	8350
BEN	Franklin Resources	Financials	259	0.4074	7836
FLS	Flowserve Corporation	Industrials	202	0.5271	7726
JPM	JPMorgan Chase & Co.	Financials	124	0.7409	7614
UTX	United Technologies	Industrials	217	0.4746	7590
SIAL	Sigma-Aldrich	Materials	254	0.4186	7020

5. Hierarchical structures of stock networks

Mantegna introduced the minimum spanning tree and hierarchical clustering methods into the study of financial networks [14], in which a distance matrix D is build from the correlation matrix for all stocks. Based on the distance matrix, the minimum spanning tree is extracted. Since the minimum spanning tree contains the information of edges connecting all vertices in a single spanning tree with the shortest total length, it's also possible to extract the hierarchical clustering tree from the minimum spanning tree, where the distance for vertex v_i and v_j is subdominant *ultrametric* distance $d^<(i,j)$ as the maximum value of distance along the shortest path between the two vertices v_i and v_j [98]. But the subdominant ultrametric distance approach will lose much edge distance information, for two separated vertices which are indirectly connected on the minimum spanning tree with a certain larger subdominant ultrametric distance might actually be directly connected in the original distance matrix. Vertices which should be clustered together might be separated in a hierarchical clustering tree based on ultrametric distance. In order to preserve the hierarchical structure of the minimum spanning tree as well as more information allowing loops and cliques, Planar Maximally Filtered Graph (PMFG) is proposed in [77]. Based on PMFG, the influence of different sectors of CSI300 is studied revealing Industrial sector is the dominate part of the whole market [97]. In [99], the hierarchical tree structure of multiple industry indices in China are investigated before and after a crisis showing the structure changes around the crisis period. A similar study of global financial crisis impact on stock market shows that Turkish market is less influenced [100]. Authors propose to use *Kullback-Leibler* distance for the filtering procedures in [8]. International real estate market networks in different countries are studied in [89] revealing that markets are clustered according to geographical locations. Instead of using the methods of [14], a symbolic approach is applied to extract the hierarchical structure of the German stock market in [59]. Using the industry classification information as the benchmark, authors compared the methods used to extract the clusters

in financial networks [93]. In [101], *Neighbor-Net* approach is applied in which more distance information is used in the construction of the tree compared to the hierarchical clustering.

Since a sliding window approach with a window size of L is utilized, in a study period of total P trading days, we can get a number of $P-L+1$ trading windows. We have $P_{csi163} = 2149$ for CSI163 and $P_{s\&p468} = 2228$ for S&P468 trading dates in our study period between 04/01/2007 to 06/11/2015, respectively. As we adapted in previous parts, we set the sliding window size as $L_{csi163} = 170$ for CSI163 and $L_{S\&P468} = 500$ for S&P468. We have $W_{csi163} = 1980$ windows for CSI163 and $W_{S\&P468} = 1729$ for S&P468 respectively. For each sliding window at time t, we can get the distance matrices $D_{csi163}(t)$ and $D_{S\&P468}(t)$ where $t = 1, \cdots, W$. To investigate the structures of the two markets taking the whole study period as a whole, we calculate the averaged distance matrix by averaging all elements over all sliding windows as

$$\langle D \rangle = \frac{1}{W} \sum\nolimits_t D_t. \tag{7}$$

With this averaged distance matrix, we construct the hierarchical trees, minimum spanning trees, asset graphs and study the evolvement of the properties of minimum spanning trees, asset graphs for both CSI163 and S&P468.

5.1. Hierarchical tree

In the study of the stock market or a portfolio, a set of individual stocks belonging to different economy sectors behavior correlated together. Based on the prices information, correlation matrix can be formed. Based on that a distance matrix can be derived. Using the distance matrix, clustering algorithms can be further applied to extract the clustering structures of the stocks. For the stocks falling in a same cluster, they behave similar sharing correlated price fluctuations, while for the stocks coming from different clusters, they are less similar than to the ones of same clusters. The main objective for clustering algorithms is to minimize the dissimilarity for stocks in a same cluster and maximize the dissimilarity for stocks in different clusters meanwhile. Since the

dissimilarity is naturally defined by the distance, the selection of definition of distance between clusters is important for clustering algorithms. Four distance definitions as shown in Eq. 8-Eq. 11 are used in extracting of hierarchical clustering trees. The distance of two clusters, c_m and c_n, is defined as the minimum distance for all pairs as in Eq. 8, the maximum distance for all pairs as in Eq. 9, the average distance for all pairs as in Eq. 10, and the distance between average centroids of the two clusters as in Eq. 11, respectively.

$$d_{m,n} = \min(d_{o_m^i, o_n^j}) \tag{8}$$

$$d_{m,n} = \max(d_{o_m^i, o_n^j}) \tag{9}$$

$$d_{m,n} = \frac{1}{N_m N_n} \sum\nolimits_i \sum\nolimits_j d_{o_m^i, o_n^j} \tag{10}$$

$$d_{m,n} = d(\overline{o_m}, \overline{o_n}) \tag{11}$$

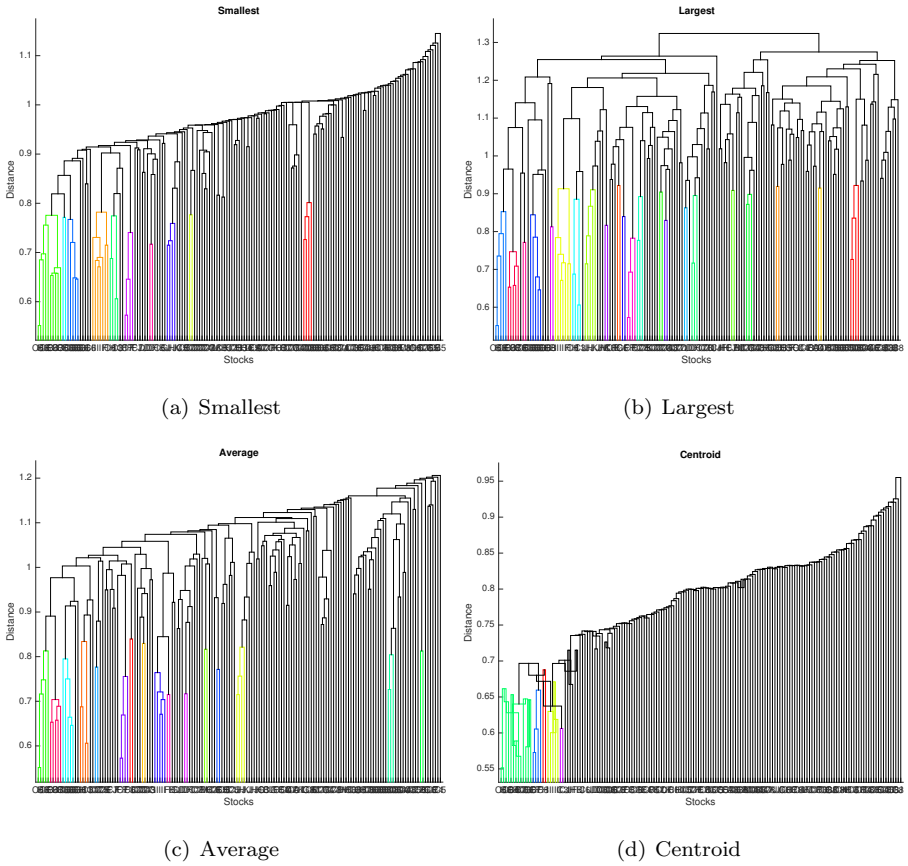
In our study, we use all these four definitions of cluster distance. For CSI163 network, we present the dendrogram hierarchical cluster trees in Fig. 11(a)-Fig. 11(d). For S&P468 network, we present the trees in Fig. 12(a)-Fig. 12(d). In these trees, the leaf nodes are individual stocks and the height for two branches merging together indicates the distance or dissimilarity between two clusters or stocks. The higher they merge, the larger the distance is. For similar clusters or stocks, they merge in a lower value of height. To color the similar stocks, a color threshold of $0.7 \times \max(d_{o_m^i, o_n^j})$ is used, thus all similar clusters or stocks are colored with same colors. With adjusting this color threshold, we can get the clusters from the dendrogram hierarchical cluster trees. As shown in the figures, using different definitions, we can get different hierarchical cluster trees and it's obvious that Fig. 11(b) and Fig. 11(c) reveal more details of the structures, in which the distance between clusters is the largest of all pairs and the average distance of all pairs, respectively. The similar effect is also observed in Fig. 12(b)

and Fig. 12(c) for S&P468 networks. These clustering results are found to agree with the classifications of stocks very well. The colored clusters are composed by stocks mostly from same economy sectors. Though there are exceptions that some stocks from different sectors are clustered together or stocks from a same sector are clustered in different clusters. It's still astonishing to see that stocks can be clustered which agree with the economy sectors classification only from the prices information. These results indicate that hierarchical cluster trees constructed from price correlation matrix can reveal economy sectors and this has potential applications in portfolio selection and risk management.

5.2. Minimum spanning tree

For a given undirected weighted network with N vertices, we can simplify the network by extracting the backbone of the network connecting all vertices but with a minimum total length of edges, this backbone is called Minimum Spanning Tree, or MST for short. Since loops or circles are not allowed to connect vertices, a MST has a topological structure of tree with $N-1$ edges which is dramatically simplified from the original network which might has a maximum of $N(N-1)/2$ edges. This brings huge advantages into the study of networks of stocks by reducing noises and simplifying the computation as well.

To construct a Minimum Spanning Tree from a given network, it's easy to be achieved by using the *Kruskal's Algorithm* [85], in which all edges are ranked in ascending order. Starting from the shortest edge on the edges ranking list, we add edges to the tree by keeping the tree in spanning form without introducing circles. After all edges are considered, we get a final minimum spanning tree comprising all connected N vertices with a minimum total length for $N-1$ edges. For a network in which all edges with distinct lengths, the extracted MST is unique. In Fig. 13, we demonstrate the process of extracting the minimum spanning tree from a 6 vertices network following the Kruskal's MST Algorithm. The edges are ranked in descending order and we start from the shortest ones and add them into the tree but omit the edges which might introduce loops, after considering all edges, we get a minimum spanning tree with a minimum



(a) Smallest

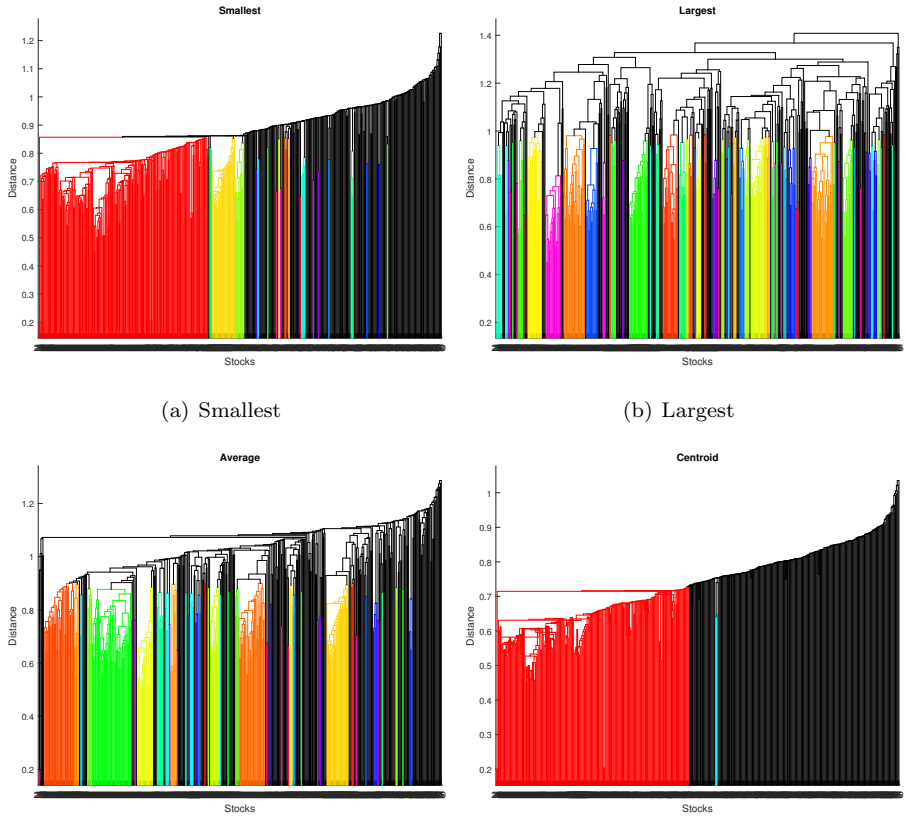
(b) Largest

(c) Average

(d) Centroid

Figure 11: CSI163 dendrogram hierarchical cluster trees extracted with different distance definitions in (a) smallest distance for stock pairs, Eq. 8; (b) largest distance for stock pairs, Eq. 9; (c) average distance for stock pairs, Eq. 10; (d) distance between centroids for clusters, Eq. 11. The color threshold is 0.7. All stocks whose linkage values are less than this threshold would be colored with a unique color. As shown in the figures, different distance definitions extract different dendrogram hierarchical cluster trees whereas the same color threshold generates different results. We see that the largest distance definition reveals more details of network.

total length. In this example, edge (3,6) and (6,4) are omitted because that $e_{3,6}$ might bring a loop of (3,6,1) and $e_{6,4}$ might bring a loop of (6,4,2,5). Another widely used algorithm is *Prim's Algorithm* [102] which begins with a starting vertex and adds the shortest one to the existing tree from all edges connected



(a) Smallest

(b) Largest

(c) Average

(d) Centroid

Figure 12: S&P468 dendrogram hierarchical cluster trees extracted with different distance definitions in (a) smallest distance for stock pairs, Eq. 8; (b) largest distance for stock pairs, Eq. 9; (c) average distance for stock pairs, Eq. 10; (d) distance between centroids for clusters, Eq. 11. The color threshold is 0.7. All stocks whose linkage values are less than this threshold would be colored with a unique color. As shown in the figures, different distance definitions extract different dendrogram hierarchical cluster trees whereas the same color threshold generates different results. Again, the largest distance definition reveals more details of network.

to the tree. By repeating this greedily, we can extract the minimum spanning tree of the given network. In this research, we apply the Kruskal's Algorithm to analyze the network structures of CSI163 and S&P468.

To extract the minimum spanning trees of the stock networks of CSI163 and

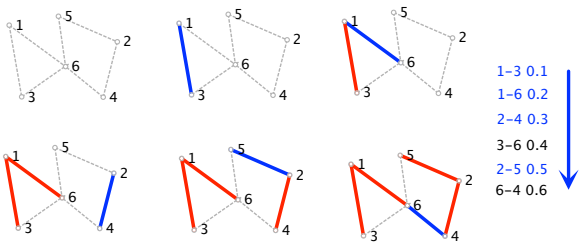


Figure 13: The extraction of a minimum spanning tree from a 6 vertices network using Kruskal's MST Algorithm. We rank all edges in descending order according to the edge lengths. Starting from the shortest edge and add the edges into the tree but avoiding loops or circles, after considering all edges, we get a final tree connecting all vertices with the minimum total edge lengths. In our example, after adding $e_{1,3}$, $e_{1,6}$, $e_{2,4}$, $e_{2,5}$, and $e_{6,4}$, we finally extract a tree of $T_{3,1,6,4,2,5}$ with a total length of 1.1.

S&P468, we average all correlation matrices over the investigated time windows, and present in Fig. 14 and Fig. 15 for CSI163 and S&P468 respectively. We see that the stocks of the same industry sectors are clustered in the MSTs for both CSI163 and S&P468, and this clustering effect is much more obvious for S&P468 in which stocks are well clustered according to the industry sectors of S&P500.

We further look into the connectivities of MSTs for both CSI163 and S&P468, we find that after the edge filtering process, some stocks are still well connected with other stocks. These stocks are the key stocks in contribution of connectivities of the MSTs, while most stocks are poorly connected with degree of only 1 or 2. In Table 10 and Table 11, we present the top 10 stocks according to their degrees in MST of CSI163 and S&P468, respectively. We find that the best connected stocks of CSI163 are diverse, while for the MST of S&P468, 3 Financials stocks appear in the top 10. This agree with other analysis that the Chinese stock market is much more divers and Financials stocks play important roles in the US market.

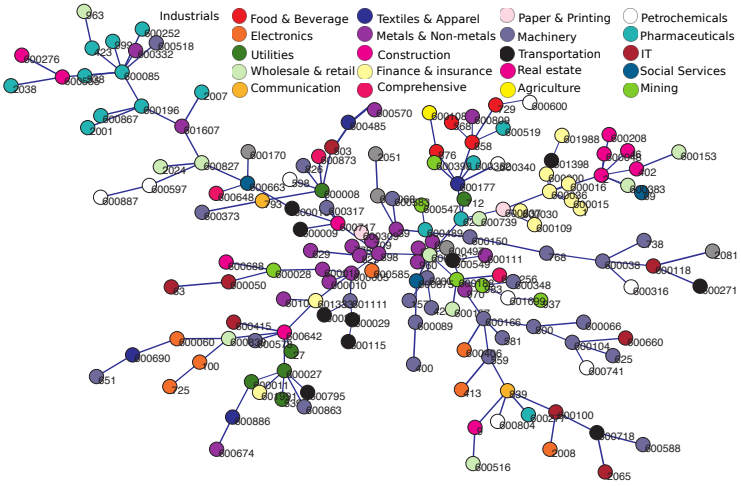


Figure 14: Minimum spanning tree of CSI163. Vertices are colored to indicate different industry sectors.

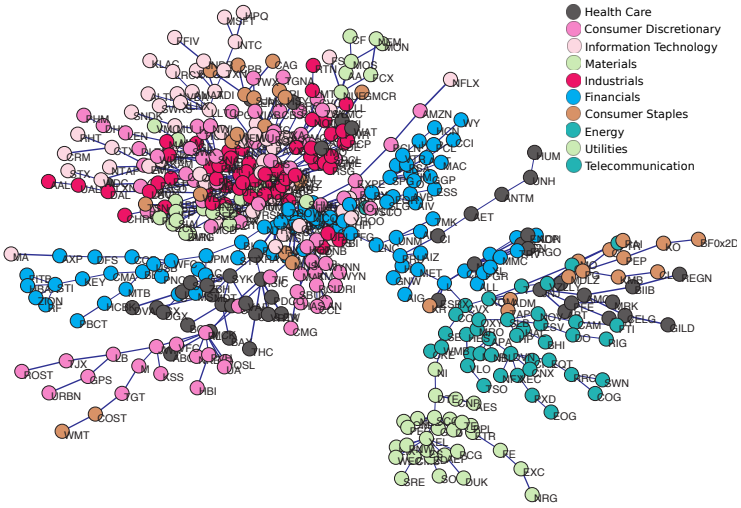


Figure 15: Minimum spanning tree of S&P468. Vertices are colored to indicate different industry sectors.

5.3. Planar maximally filtered graph

Like the minimum spanning tree (MST) approach, Planar Maximally Filtered Graphs, PMFG, is also an edge filtering method but the allowance of cliques up to 4 vertices show much richer structure information of a network

Table 10: In this table, we present the top 10 stocks with the largest degrees in the MST of CSI163. As it shown that the top 10 stocks are diverse in industry sectors with represents of 1 Wholesale & retail stock, 1 Metals & Non-metals stock, 2 Pharmaceuticals stocks, 2 Real estate stocks, 1 Finance & insurance stock, 2 Utilities stocks, and 1 Textiles & Apparel stock.

Degree	Code	Name	Industry
11	600362	Beijinghualian Hypermarket	Wholesale & retail
9	898	Angang Steel	Metals & Non-metals
8	600085	Beijing Tongrentang	Pharmaceuticals
7	2	China Vanke	Real estate
7	600036	China Merchants Bank	Finance & insurance
6	600008	Beijing Capital	Utilities
6	600027	Huadian Power	Utilities
6	600177	Youngor Group	Textiles & Apparel
6	600642	Shanghai Wanye Enterprises	Real estate
5	623	Jilin Aodong Medicine Industry	Pharmaceuticals

rather than a single tree. Based on the correlation matrix, the PMFG spans on a planar surface without crossing of edges but with loops and holes. It's believed that a PMFG might reveal more details of the networks. After the introduction of PMFG by Tumminello *et al.* in the study of 100 stocks of NYSE [77], PMFG has been applied in many studies of financial networks. In [58], the authors study the PMFG networks of DAX 30 stocks. Instead of using the correlation matrix, a p-values matrix of EngleGranger cointegration test is used to extract the PMFG for Chinese stocks in [88]. The stability and robustness of PMFG for 300 NYSE stocks are compared with MST in a running window approach and the results reveal that PMFG is stabler than MST [94]. In [84], the same authors of [94] confirm that PMFG provides stronger robustness and stability in revealing network structures of stock markets. It has also been proven that the PMFG always contains a MST for a same distance matrix [77].

The PMFGs of CSI163 and S&P468 networks are plotted in Fig. 18 and

Table 11: In this table, we present the top 10 stocks with the largest degrees in the MST of S&P468. As it shown that Honeywell Intl. is the most connected stock in the MST with a degree of 38. The top 10 stocks are composed by 2 Industrials stocks, 3 Financials stocks, 2 Utilities stocks, 1 Health Care stock, and 1 Consumer Discretionary stock.

Degree	Tick	Stock Name	Industry
38	HON	Honeywell Intl.	Industrials
18	TROW	T. Rowe Price Group	Financials
11	SCG	SCANA Corp	Utilities
10	ITW	Illinois Tool Works	Industrials
9	ADP	Automatic Data Processing	IT
9	XEL	Xcel Energy Inc	Utilities
8	JNJ	Johnson & Johnson	Health Care
8	SPG	Simon Property	Financials
8	SNA	Snap-On Inc.	Consumer Discretionary
7	AMG	Affiliated Managers	Financials

Fig. 19, respectively. We see that PMFGs have much more edges compared to MSTs. Further, we use another layout to plot the two PMFGs in Fig. 18 and Fig. 19, from which, we find that PMFGs also produce good clusters for stocks of different industry sectors.

5.4. Asset graph

In the minimum spanning tree (MST), a connected tree structure connecting all vertices with a minimum total length of edges is extracted. The selection process of adding edges in generating a MST out of a distance matrix is presented in Fig. 13, a MST is always a connected single tree without disconnected parts. By connecting the N vertices, a total of $N-1$ edges are needed, where N is the number of vertices in the original network. It's obvious that a MST does not guarantee to be with the possible minimum total lengths of the $N-1$ edges. By changing the strategy of how edges are selected and allowing disconnected

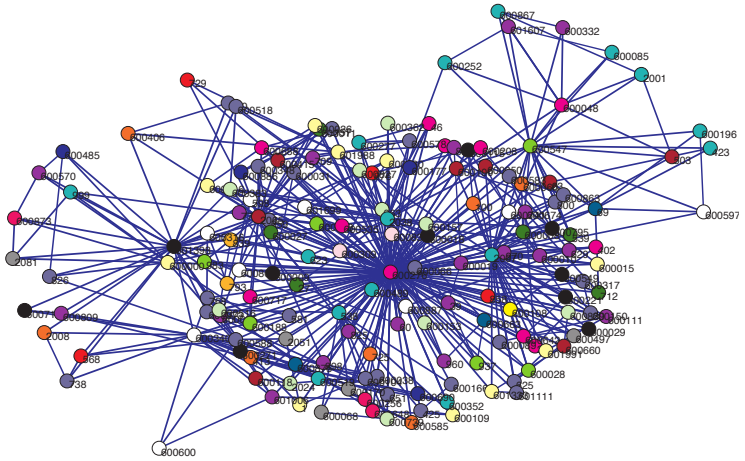


Figure 16: PMFG of CSI163.

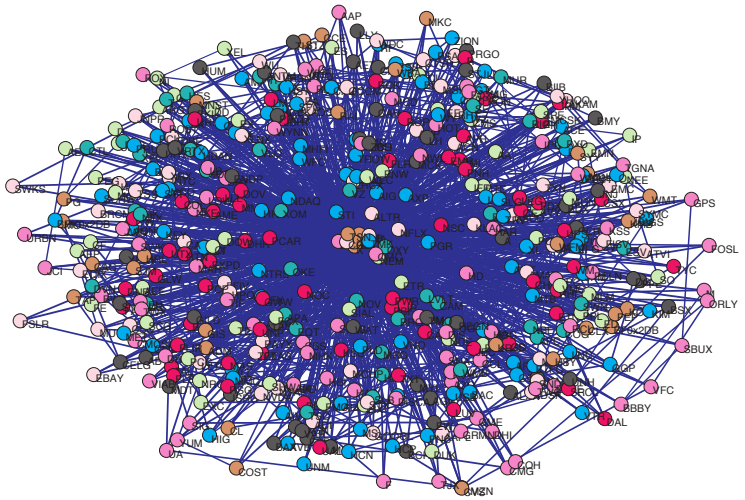
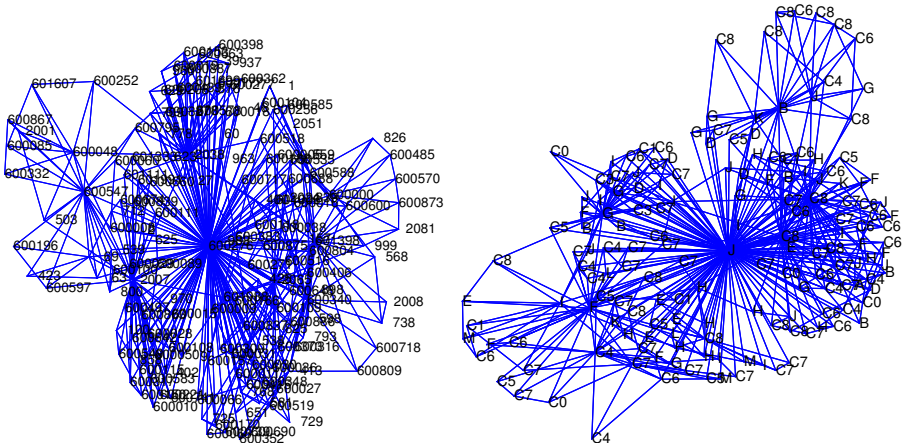


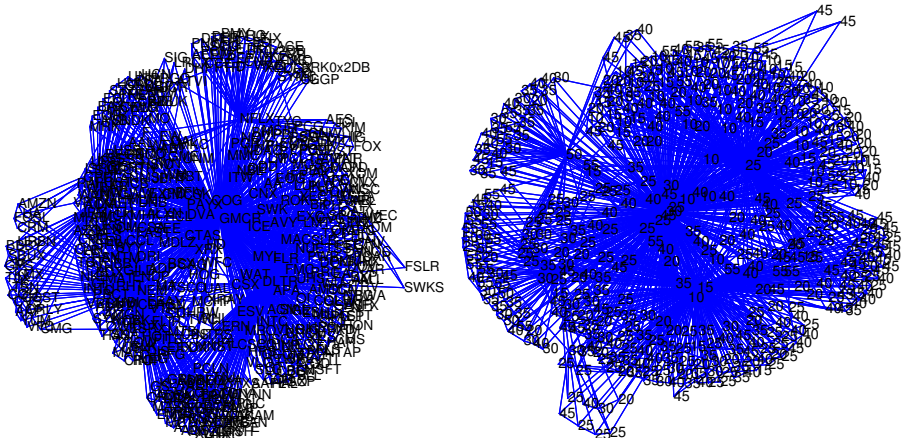
Figure 17: PMFG of S&P468.

parts, the Asset Graph (AG) approach is proposed in [78, 79]. Similar to MST, to generate an AG, we start the distance matrix containing all pairwise distances information of the network, we first rank all edges in an ascending order from the shortest to the longest. Without considering the requirement of keeping a tree connected, we simply choose the top $N-1$ edges to form a AG. It has been found that AG extracts similar structures as MST can do with smaller normalized



(a) Stock (b) Industry

Figure 18: CSI163 PMFG. We label the vertices with stock codes in (a) and industry codes in (b).



(a) Stock (b) Industry

Figure 19: S&P468 PMFG. We label the vertices with stock codes in (a) and industry codes in (b).

length and with better stable structure over time. In this section, we show the AG networks for both markets. In Fig. 20 and Fig. 21, we present AG structures for CSI163 and S&P468 networks respectively. Compare with Fig. 14 and Fig.

15 of the minimum spanning trees of CSI163 and S&P468 networks, we see that AG structures are more complex than MST and there are many isolated vertices in AG. The connected cliques in AG are the most correlated stocks connected by the shortest possible edges, in other words, by connecting the most correlated stocks, AG just omit the less correlated stocks. We also observe that many cliques emerge in AG and this reveals more information about the structures than in MST where no loops or cliques are allowed. We see that AG is a simple but effective network simplification approach in extracting the most correlated stocks. But the sacrifice is also obvious, as shown in Fig. 20 and Fig. 21, the clustering is poor in AG compared with MST for both markets.

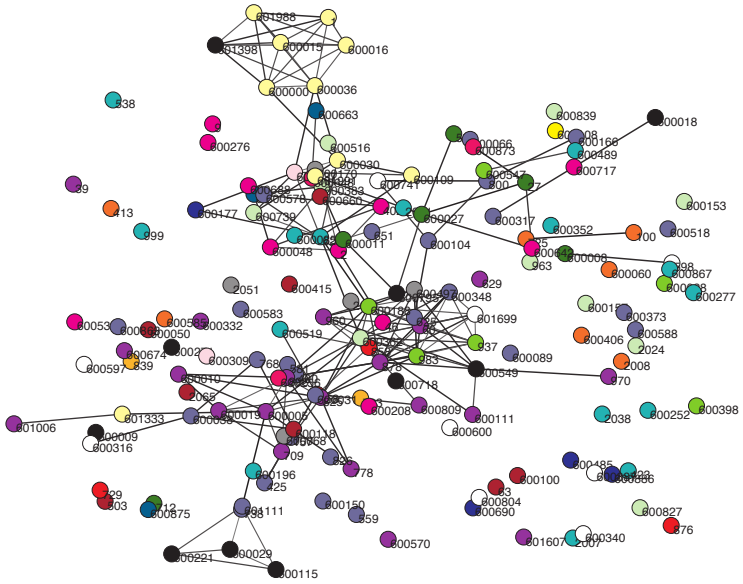


Figure 20: Asset graph of CSI163 with $N-1$ shortest edges.

As we have shown that AG allows isolated vertices and not all vertices are connected in one giant tree. To generate a AG, we can use different number of edges, with the increase of edge number, we can see that the portion of isolated vertices decline. It's interesting to investigate how the vertices are related with the edge numbers. In the original distance matrix, the maximum possible number of edges is $N(N-1)/2$. The percentage of the fraction of AG

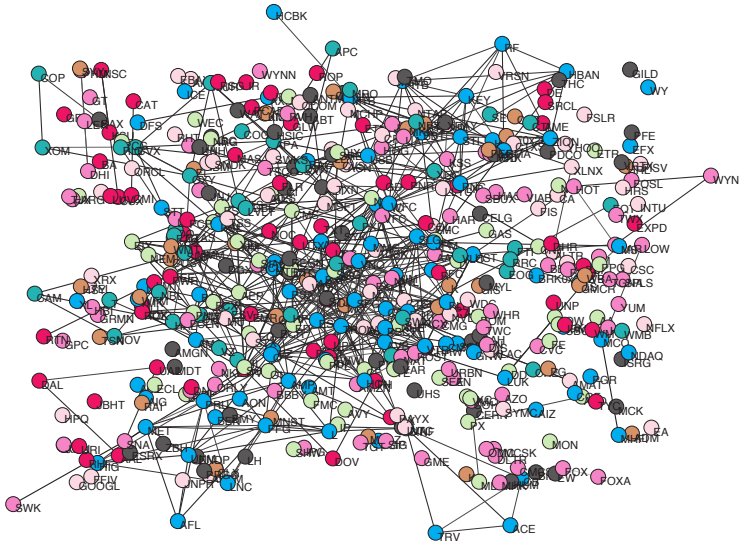
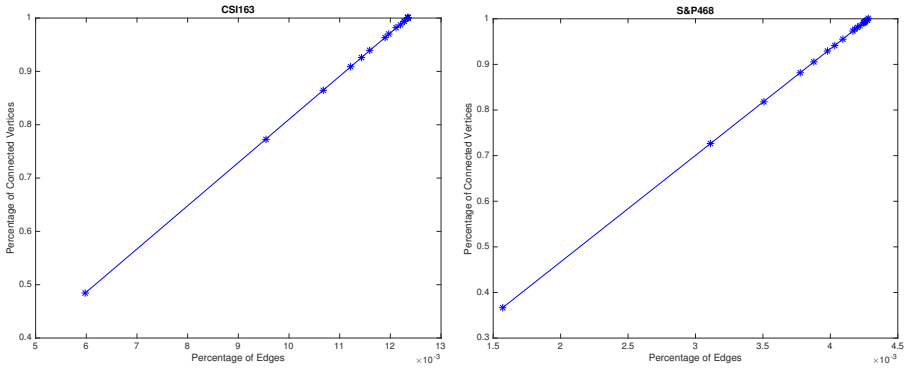


Figure 21: Asset graph of S&P468 with $N-1$ shortest edges.

is the top edges added into AG networks to the total possible edges. When we increase this edge percentage, more and more vertices are connected. We calculate the percentage of connected vertices as the number of connected to the total vertices number of N. We plot these results in Fig. 22(a) and Fig. 22(b) for CSI163 and S&p468 networks respectively. As the figures show, with a small fraction of edges are included, more and more vertices are connected, it requires only 0.0123 and 0.0043 of the total edges for all vertices to be connected in CSI163 and S&P468 respectively. This indicates that the top edges are very effective in connecting vertices for S&P468 than CSI163.

In the previous section, all structures are extracted from the average distance matrices over the whole study periods which is defined in Eq. 7 as $\langle D \rangle = \frac{1}{W}\sum_t D_t$, where W is the number of sliding windows. In this part, we investigate the dynamic structures of the filtered networks with a focus on the AG and MST. For each sliding window, at time t, we get a series of distance matrices D_t based on the returns data on the interval of $[t, t-1, \cdots, t-L+1]$ where L is the length of a sliding window. For each sliding window, using the distance matrix D_t, we construct the corresponding original network N_t, the asset graph AG_t, and the



(a) CSI163 Network (b) S&P468 Network

Figure 22: Percentages of connected vertices of AG against edge densities for CSI163 and S&P468 networks.

minimum spanning tree MST_t. Using our dataset, we get $W_{csi163} = 1980$ windows for CSI163 and $W_{S\&P468} = 1729$ for S&P468 in the study period between 04/01/2007 and 06/11/2015.

In Fig. 23, we present the distance distributions of original distance matrices N_t, asset graphs AG_t, and minimum spanning trees MST_t for both of CSI163 and S&P468 in the study period between 04/01/2007 and 06/11/2015. In the original network N_t, a number of $N(N-1)/2$ edges are considered, while for AG_t and MST_t, $N-1$ edges are considered. Since the sliding window sizes are $L_{csi163} = 170$ and $L_{S\&P468} = 500$, we should keep in mind that a slice of distribution is a result of the past L dates, *i.e.*, about half of a year for CSI163 and 2 years for S&P468. The shapes of these distributions are influenced by the lengths of L. we choose the same set of lengths by considering the requirements of random matrix theory approach which we shall discuss later. The similar plots are reported in [78] in the study of 477 stocks from NYSE which is in a similar size of our S&P468 dataset in which 468 stocks are included. We add more evidences by comparing two markets of CSI163 and S&P468. In Fig. 23(a) and Fig. 23(b), we observe obvious shifts of the distribution centers for both markets. In these shifts, positive shifts to the mean value of $\langle d_{ij} \rangle = \sqrt{2}$

roughly correspond to the normal market periods, while negative shifts to the mean value correspond the bear or collapsing market periods. This indicates the stocks behavior synchronized in bad periods and this agrees with many previous studies. This also provides a potential market measurement for investors and regulators to watch how market shift behaviors. In Fig. 23(c) and Fig. 23(d), the distributions of distances of AG for CSI163 and S&P468 are plotted. Since AG is a subgraph of the original network, and is composed by the top $N-1$ shortest edges so we expect the distributions show a left shift to the mean center of $\langle d_{ij} \rangle = \sqrt{2}$ compared to the original networks, and this is well shown in the plots for both CSI163 and S&P468, more precisely, the distributions of AG are zoom-in of the left tails of original networks. For the MST, as shown in Fig. 23(e) and Fig. 23(f), has a relatively wider distribution which is positively shifted compared to AG but negatively shifted to the original network. Also, we find that in AG and MST networks, the most parts of the distributions stay on the left hand of the center $\sqrt{2}$ which means the network is correlated on average, in other words, for periods when the mean center stays on the left hand, the network backbones of AG and MST are on average correlated, and rarely anti-correlated. A potential implication is that for the whole market, the network provides a diversified portfolio when market is normal or in bull state, but for the top edges in AG and MST, the network move together with less diversification when market falls into bear markets or crisis periods.

The distance d_{ij} indicates how the two stocks correlate with each other. Larger d_{ij} means smaller correlation and vice versa. For an original network N at time t, the total distance can be introduced as:

$$d_{total} = \sum_{i,j} d_{ij}, \tag{12}$$

and the average distance for the original network can be defined as:

$$\langle d_{ij} \rangle = \frac{1}{N(N-1)/2} \sum_{i,j} d_{ij}. \tag{13}$$

In the same way, we can calculate the total distances for AG_t and MST_t using Eq. 12, but considering the edge number for AG_t and MST_t is $N-1$, we

normalize the average distance for them as:

$$\langle d_{ij} \rangle = \frac{1}{N-1} \sum_{i,j} d_{ij}. \tag{14}$$

To investigate the tightness of the network, the total distance d_{total} and average distance $\langle d_{ij} \rangle$ for the original network N_t, AG_t, and MST_t in our study periods for both networks is investigated. We plot the results in Fig. 24 for d_{total} and Fig. 25 for $\langle d_{ij} \rangle$, respectively. For each stock market, total distance d_{total} and average distance $\langle d_{ij} \rangle$ show similar shapes. For both stock markets, the values are in this order: $N_t > MST_t > AG_t$, *i.e.*, the original networks have the largest values of d_{total} and $\langle d_{ij} \rangle$ compared to MST_t and AG_t, while AG_t has the smallest values.

By comparing the total distance d_{total} plotted in Fig. 24(a) and the average distance $\langle d_{ij} \rangle$ plotted in Fig. 25(a) for CSI163 over the study period, we find the six plots share similar shapes. The same similarities are also observed in Fig. 24(b) and Fig. 25(b) for S&P468 network. This indicates, the AG and MST are both good backbones of the whole original market networks and this tracking stays robust over time. For both networks, we also find that the curve of MST_t is above AG_t which means the total and average distances are slightly larger in MST than in AG. Our findings agree with the results reported in [78]. Since the two stock markets datasets have different stock numbers, we compare the average distance between the two markets and as it shows in Fig. 25, we see that the CSI163 is slightly sparser than S&P468, indicates that the CSI163 which is a developing market is more diversified than S&P468 which is a developed market, this also agrees with many previous researches.

In Table 12, we summary the average $\langle \langle d_{ij} \rangle \rangle$, minimum $\langle d_{ij} \rangle_{\min}$, maximum $\langle d_{ij} \rangle_{\max}$, and standard deviation $\langle d_{ij} \rangle_{\sigma}$ of CSI163 and S&P468 networks for three kinds networks: original, AG, and MST. We can see that the values are in the order of $N > MST > AG$ for average, minimum, and maximum. Also the three networks have similar standard deviations. We find that the values of $\langle \langle d_{ij} \rangle \rangle$ and minimum $\langle d_{ij} \rangle_{\min}$ for CSI163 are slightly larger than S&P468 which indicates stocks in CSI163 are less likely to correlated than in S&P468.

CSI163				S&P468			
$\langle\langle d_{ij}\rangle\rangle$	$\langle d_{ij}\rangle_{\min}$	$\langle d_{ij}\rangle_{\max}$	$\langle d_{ij}\rangle_{\sigma}$	$\langle\langle d_{ij}\rangle\rangle$	$\langle d_{ij}\rangle_{\min}$	$\langle d_{ij}\rangle_{\max}$	$\langle d_{ij}\rangle_{\sigma}$
1.1145	1.2445	0.9619	0.0650	1.0754	1.1968	0.9816	0.0731
0.7049	0.8708	0.5868	0.0578	0.6377	0.7607	0.5579	0.0623
0.8251	0.9565	0.7017	0.0540	0.7971	0.9092	0.7113	0.0615

Table 12: Average $\langle\langle d_{ij}\rangle\rangle$, minimum $\langle d_{ij}\rangle_{\min}$, maximum $\langle d_{ij}\rangle_{\max}$, and standard deviation $\langle d_{ij}\rangle_{\sigma}$ of CSI163 and S&P468 networks. The values of first row belong to the original network N, the second row belong to the AG, and the third row belong to MST.

To visualize the distributions of these three kinds of networks, we plot the probability density function (PDF) for the original network, AG, and MST for CSI163 and S&P468 in Fig. 26(a) and Fig. 26(b), respectively. We see that the distributions of all three networks share similar shapes but with different mean centers, as shown in the figures, the AG locates on the left with MST locates in the center and original locates on the right.

6. Conclusion and Discussion

In this research, we investigated the properties and models of the complex network theory and its applications from data science perspective. Using the daily close prices of two sets of stocks from CSI300 and S&P500, we constructed the correlation matrices both for the whole study periods and all sliding windows. Based on these correlation matrices, we build the networks with stocks as the vertices and correlation relationships as the edges. We systematically applied network filtering methods like hierarchical tree, minimum spanning tree, planar maximally filtered graph, and asset graph to simplify the networks. For each filtered network, the network properties are discussed. Financial markets are typical complex systems, its important to extract useful information from the noise background by applying methods like complex networks. We find that the stock markets, CSI300 and S&P500, the former is an emerging market while the latter is a mature well-developed market. They share similar properties in many

ways and also vary in many aspects. The revealed properties and robustnesses might provide sights of the structures and dynamics of the two stock markets for practitioners and regulators. Furthermore, it is interesting to develop trading strategies with the information revealed from the topological networks of stocks or indices. For instance, the pair trading [103–106] is a basic and market neutral strategy considering the movement of correlated stock pair, in which if the spread widen, then traders can short one and long another one to gain the spread. One might use the information of the networks to identify the pairs and evaluate the reliabilities. Also, considering pairs between groups of stocks rather than only two stocks, we might use the component or cluster information revealed in the networks to build the trading groups. Furthermore, the directed networks built with Granger causalities or lagged correlations might give more Lead/Lag details of stock pairs on the time factors in the asynchronous fashion. Second, with the help of network edge filtering, we can greatly simplify the networks, but most studies focus on the topological simplification without concerns of the original portfolio returns. What if we take the returns into consideration with the topology of the networks to optimize the portfolio selection? The topological structure can give us information how diverse the portfolio is but this is not enough to design the portfolio without return information. A possible way is to adjust the portfolio selection by considering measurements like the ratio of returns over total distances of a portfolio or other approaches combining both topological and return information.

References

[1] M. E. J. Newman, The structure and function of complex networks, SIAM Review 45 (2) (2003) 167–256. arXiv:http://dx.doi.org/10.1137/S003614450342480, doi:10.1137/S003614450342480.
URL http://dx.doi.org/10.1137/S003614450342480

[2] D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks,

Nature 393 (6684) (1998) 440–442.
URL http://dx.doi.org/10.1038/30918

[3] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512. arXiv:http://www.sciencemag.org/content/286/5439/509.full.pdf, doi:10.1126/science.286.5439.509.
URL http://www.sciencemag.org/content/286/5439/509.abstract

[4] B. Albert-László, Linked: the new science of networks, Perseus, 2002.

[5] L. da Fontoura Costa, O. N. O. Jr., G. Travieso, F. A. Rodrigues, P. R. V. Boas, L. Antiqueira, M. P. Viana, L. E. C. Rocha, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Advances in Physics 60 (3) (2011) 329–412. arXiv:http://dx.doi.org/10.1080/00018732.2011.572452, doi:10.1080/00018732.2011.572452.
URL http://dx.doi.org/10.1080/00018732.2011.572452

[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Physics Reports 424 (4–5) (2006) 175 – 308. doi:http://dx.doi.org/10.1016/j.physrep.2005.10.009.
URL http://www.sciencedirect.com/science/article/pii/S037015730500462X

[7] C. K. Tse, J. Liu, F. C. Lau, A network perspective of the stock market, Journal of Empirical Finance 17 (4) (2010) 659 – 667. doi:http://dx.doi.org/10.1016/j.jempfin.2010.04.008.
URL http://www.sciencedirect.com/science/article/pii/S0927539810000368

[8] M. Tumminello, F. Lillo, R. N. Mantegna, Correlation, hierarchies, and networks in financial markets, Journal of Economic Behavior & Organization 75 (1) (2010) 40 – 58, transdisciplinary Perspectives on Economic

Complexity. doi:http://dx.doi.org/10.1016/j.jebo.2010.01.004.
URL http://www.sciencedirect.com/science/article/pii/S0167268110000077

[9] V. Boginski, S. Butenko, P. M. Pardalos, Statistical analysis of financial networks, Computational Statistics & Data Analysis 48 (2) (2005) 431 – 443. doi:http://dx.doi.org/10.1016/j.csda.2004.02.004.
URL http://www.sciencedirect.com/science/article/pii/S0167947304000258

[10] V. Boginski, S. Butenko, P. M. Pardalos, Mining market data: A network approach, Computers & Operations Research 33 (11) (2006) 3171 – 3184, part Special Issue: Operations Research and Data Mining. doi:http://dx.doi.org/10.1016/j.cor.2005.01.027.
URL http://www.sciencedirect.com/science/article/pii/S0305054805000286

[11] H. Markowitz, Portfolio selection, The Journal of Finance 7 (1) (1952) 77–91. doi:10.1111/j.1540-6261.1952.tb01525.x.
URL http://dx.doi.org/10.1111/j.1540-6261.1952.tb01525.x

[12] V. Kalyagin, A. Koldanov, P. Koldanov, V. Zamaraev, Market graph and markowitz model, in: T. M. Rassias, C. A. Floudas, S. Butenko (Eds.), Optimization in Science and Engineering, Springer New York, 2014, pp. 293–306. doi:10.1007/978-1-4939-0808-0_15.
URL http://dx.doi.org/10.1007/978-1-4939-0808-0_15

[13] V. Boginski, S. Butenko, O. Shirokikh, S. Trukhanov, J. Gil Lafuente, A network-based data mining approach to portfolio selection via weighted clique relaxations, Annals of Operations Research 216 (1) (2014) 23–34. doi:10.1007/s10479-013-1395-3.
URL http://dx.doi.org/10.1007/s10479-013-1395-3

[14] R. N. Mantegna, Hierarchical structure in financial markets, Eur. Phys. J. B 11 (1999) 193–197.

[15] S. Battiston, J. D. Farmer, A. Flache, D. Garlaschelli, A. G. Haldane, H. Heesterbeek, C. Hommes, C. Jaeger, R. May, M. Scheffer, Complexity theory and financial regulation, Science 351 (6275) (2016) 818–819. arXiv:http://science.sciencemag.org/content/351/6275/818.full.pdf, doi:10.1126/science.aad0299.
URL http://science.sciencemag.org/content/351/6275/818

[16] P. Glasserman, H. P. Young, How likely is contagion in financial networks?, Journal of Banking & Finance 50 (Supplement C) (2015) 383 – 399. doi:https://doi.org/10.1016/j.jbankfin.2014.02.006.
URL http://www.sciencedirect.com/science/article/pii/S0378426614000600

[17] D. Acemoglu, A. Ozdaglar, A. Tahbaz-Salehi, Systemic risk and stability in financial networks, The American Economic Review 105 (2) (2015) 564–608. doi:doi:10.1257/aer.20130456.
URL http://www.ingentaconnect.com/content/aea/aer/2015/00000105/00000002/art00005

[18] R. Albert, H. Jeong, A.-L. Barabasi, Error and attack tolerance of complex networks, Nature 406 (6794) (2000) 378–382.
URL http://dx.doi.org/10.1038/35019019

[19] S. Battiston, D. D. Gatti, M. Gallegati, B. Greenwald, J. E. Stiglitz, Liaisons dangereuses: Increasing connectivity, risk sharing, and systemic risk, Journal of Economic Dynamics and Control 36 (8) (2012) 1121 – 1141, quantifying and Understanding Dysfunctions in Financial Markets. doi:http://dx.doi.org/10.1016/j.jedc.2012.04.001.
URL http://www.sciencedirect.com/science/article/pii/S0165188912000899

[20] N. Hautsch, J. Schaumburg, M. Schienle, Financial network systemic risk contributions, Review of Finance 19 (2) (2015) 685–738. arXiv:/oup/backfile/content_public/journal/rof/19/2/10.

1093/rof/rfu010/2/rfu010.pdf, doi:10.1093/rof/rfu010.
URL +http://dx.doi.org/10.1093/rof/rfu010

[21] G. Cimini, T. Squartini, D. Garlaschelli, A. Gabrielli, Systemic risk analysis on reconstructed economic and financial networks, Scientific Reports 5 (2015) 15758 EP –.
URL http://dx.doi.org/10.1038/srep15758

[22] E. Nier, J. Yang, T. Yorulmazer, A. Alentorn, Network models and financial stability, Journal of Economic Dynamics and Control 31 (6) (2007) 2033 – 2060, tenth Workshop on Economic Heterogeneous Interacting AgentsWEHIA 2005. doi:http://dx.doi.org/10.1016/j.jedc.2007.01.014.
URL http://www.sciencedirect.com/science/article/pii/S0165188907000097

[23] K. Anand, P. Gai, S. Kapadia, S. Brennan, M. Willison, A network model of financial system resilience, Journal of Economic Behavior & Organization 85 (Supplement C) (2013) 219 – 235, financial Sector Performance and Risk. doi:https://doi.org/10.1016/j.jebo.2012.04.006.
URL http://www.sciencedirect.com/science/article/pii/S0167268112000868

[24] B. M. Tabak, T. R. Serra, D. O. Cajueiro, Topological properties of stock market networks: The case of brazil, Physica A: Statistical Mechanics and its Applications 389 (16) (2010) 3240 – 3249. doi:http://dx.doi.org/10.1016/j.physa.2010.04.002.
URL http://www.sciencedirect.com/science/article/pii/S0378437110002992

[25] F. Caccioli, M. Shrestha, C. Moore, J. D. Farmer, Stability analysis of financial contagion due to overlapping portfolios, Journal of Banking & Finance 46 (Supplement C) (2014) 233 – 245. doi:https://doi.org/10.1016/j.jbankfin.2014.05.021.

URL http://www.sciencedirect.com/science/article/pii/S0378426614001885

[26] G.-J. Wang, C. Xie, K. He, H. E. Stanley, Extreme risk spillover network: application to financial institutions, Quantitative Finance 17 (9) (2017) 1417–1433. arXiv:http://dx.doi.org/10.1080/14697688.2016.1272762, doi:10.1080/14697688.2016.1272762.
URL http://dx.doi.org/10.1080/14697688.2016.1272762

[27] S. Poledna, J. L. Molina-Borboa, S. Martínez-Jaramillo, M. van der Leij, S. Thurner, The multi-layer network nature of systemic risk and its implications for the costs of financial crises, Journal of Financial Stability 20 (Supplement C) (2015) 70 – 81. doi:https://doi.org/10.1016/j.jfs.2015.08.001.
URL http://www.sciencedirect.com/science/article/pii/S1572308915000856

[28] T. Kauê Dal'Maso Peron, L. da Fontoura Costa, F. A. Rodrigues, The structure and resilience of financial market networks, Chaos 22 (1) (2012) –. doi:http://dx.doi.org/10.1063/1.3683467.
URL http://scitation.aip.org/content/aip/journal/chaos/22/1/10.1063/1.3683467

[29] M. Gofman, Efficiency and stability of a financial architecture with too-interconnected-to-fail institutions, Journal of Financial Economics 124 (1) (2017) 113 – 146. doi:https://doi.org/10.1016/j.jfineco.2016.12.009.
URL http://www.sciencedirect.com/science/article/pii/S0304405X16302471

[30] S. Battiston, M. Puliga, R. Kaushik, P. Tasca, G. Caldarelli, Debtrank: Too central to fail? financial networks, the fed and systemic risk, Scientific Reports 2 (2012) 541 EP –.
URL http://dx.doi.org/10.1038/srep00541

[31] C. Zhou, Are Banks Too Big to Fail? Measuring Systemic Importance of Financial Institutions, International Journal of Central Banking 6 (34) (2010) 205–250.
URL https://ideas.repec.org/a/ijc/ijcjou/y2010q4a10.html

[32] C. Brooks, A. G. Rew, S. Ritson, A trading strategy based on the lead–lag relationship between the spot index and futures contract for the ftse 100, International Journal of Forecasting 17 (1) (2001) 31–44.

[33] C. Curme, M. Tumminello, R. N. Mantegna, H. E. Stanley, D. Y. Kenett, Emergence of statistically validated financial intraday lead-lag relationships, Quantitative Finance 15 (8) (2015) 1375–1386. arXiv: http://dx.doi.org/10.1080/14697688.2015.1032545, doi:10.1080/14697688.2015.1032545.
URL http://dx.doi.org/10.1080/14697688.2015.1032545

[34] K. Chen, P. Luo, B. Sun, H. Wang, Which stocks are profitable? a network method to investigate the effects of network structure on stock returns, Physica A: Statistical Mechanics and its Applications 436 (2015) 224 – 235. doi:http://dx.doi.org/10.1016/j.physa.2015.05.047.
URL http://www.sciencedirect.com/science/article/pii/S0378437115004628

[35] F. Pozzi, T. Di Matteo, T. Aste, Spread of risk across financial markets: better to invest in the peripheries, Scientific Reports 3 (2013) 1665 EP –.
URL http://dx.doi.org/10.1038/srep01665

[36] W. Zhang, C. Li, Y. Ye, W. Li, E. Ngai, Dynamic business network analysis for correlated stock price movement prediction, Intelligent Systems, IEEE 30 (2) (2015) 26–33. doi:10.1109/MIS.2015.25.

[37] M. Billio, M. Getmansky, A. W. Lo, L. Pelizzon, Econometric measures of connectedness and systemic risk in the finance and insurance sectors, Journal of Financial Economics 104 (3) (2012) 535 – 559,

market Institutions, Financial Market Risks and Financial Crisis. doi:http://dx.doi.org/10.1016/j.jfineco.2011.12.010.
URL http://www.sciencedirect.com/science/article/pii/S0304405X11002868

[38] T. Di Matteo, F. Pozzi, T. Aste, The use of dynamical networks to detect the hierarchical organization of financial market sectors, The European Physical Journal B - Condensed Matter and Complex Systems 73 (2010) 3–11.

[39] X. F. Jiang, T. T. Chen, B. Zheng, Structure of local interactions in complex financial dynamics, Scientific Reports 4 (2014) 5321 EP –.
URL http://dx.doi.org/10.1038/srep05321

[40] L. De Benedictis, L. Tajoli, The world trade network, The World Economy 34 (8) (2011) 1417–1454. doi:10.1111/j.1467-9701.2011.01360.x.
URL http://dx.doi.org/10.1111/j.1467-9701.2011.01360.x

[41] J. He, M. W. Deem, Structure and response in the world trade network, Phys. Rev. Lett. 105 (2010) 198701. doi:10.1103/PhysRevLett.105.198701.
URL http://link.aps.org/doi/10.1103/PhysRevLett.105.198701

[42] D. in't Veld, I. van Lelyveld, Finding the core: Network structure in interbank markets, Journal of Banking & Finance 49 (2014) 27 – 40. doi:http://dx.doi.org/10.1016/j.jbankfin.2014.08.006.
URL http://www.sciencedirect.com/science/article/pii/S0378426614002738

[43] Y. Luo, J. Xiong, L. G. Dong, Y. Tang, Statistical correlation properties of the shibor interbank lending market, China Finance Review International 5 (2) (2015) 91–102. arXiv:http://dx.doi.org/10.1108/CFRI-08-2014-0036, doi:10.1108/CFRI-08-2014-0036.
URL http://dx.doi.org/10.1108/CFRI-08-2014-0036

[44] M. Affinito, A. F. Pozzolo, The interbank network across the global financial crisis: Evidence from italy, Journal of Banking & Finance 80 (Supplement C) (2017) 90 – 107. doi:https://doi.org/10.1016/j.jbankfin.2017.03.019.
URL http://www.sciencedirect.com/science/article/pii/S037842661730078X

[45] C.-P. Georg, The effect of the interbank network structure on contagion and common shocks, Journal of Banking & Finance 37 (7) (2013) 2216 – 2228. doi:https://doi.org/10.1016/j.jbankfin.2013.02.032.
URL http://www.sciencedirect.com/science/article/pii/S0378426613001295

[46] D. J. Fenn, M. A. Porter, P. J. Mucha, M. McDonald, S. Williams, N. F. Johnson, N. S. Jones, Dynamical clustering of exchange rates, Quantitative Finance 12 (10) (2012) 1493–1520. arXiv:http://dx.doi.org/10.1080/14697688.2012.668288, doi:10.1080/14697688.2012.668288.
URL http://dx.doi.org/10.1080/14697688.2012.668288

[47] G. Iori, G. D. Masi, O. V. Precup, G. Gabbi, G. Caldarelli, A network analysis of the italian overnight money market, Journal of Economic Dynamics and Control 32 (1) (2008) 259 – 278, applications of statistical physics in economics and finance. doi:http://dx.doi.org/10.1016/j.jedc.2007.01.032.
URL http://www.sciencedirect.com/science/article/pii/S0165188907000474

[48] C. Piccardi, L. Calatroni, F. Bertoni, Communities in italian corporate networks, Physica A: Statistical Mechanics and its Applications 389 (22) (2010) 5247 – 5258. doi:http://dx.doi.org/10.1016/j.physa.2010.06.038.
URL http://www.sciencedirect.com/science/article/pii/S0378437110005674

[49] S. Vitali, J. B. Glattfelder, S. Battiston, The network of global corporate control, PLoS ONE 6 (10) (2011) 1–6. doi:10.1371/journal.pone.0025995.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0025995

[50] C. Minoiu, J. A. Reyes, A network analysis of global banking: 1978–2010, Journal of Financial Stability 9 (2) (2013) 168 – 184. doi:https://doi.org/10.1016/j.jfs.2013.03.001.
URL http://www.sciencedirect.com/science/article/pii/S1572308913000193

[51] T. A. Peltonen, M. Scheicher, G. Vuillemey, The network structure of the cds market and its determinants, Journal of Financial Stability 13 (Supplement C) (2014) 118 – 133. doi:https://doi.org/10.1016/j.jfs.2014.05.004.
URL http://www.sciencedirect.com/science/article/pii/S1572308914000503

[52] L. Marotta, S. Miccichè, Y. Fujiwara, H. Iyetomi, H. Aoyama, M. Gallegati, R. N. Mantegna, Backbone of credit relationships in the japanese credit market, EPJ Data Science 5 (1) (2016) 10. doi:10.1140/epjds/s13688-016-0071-7.
URL https://doi.org/10.1140/epjds/s13688-016-0071-7

[53] F. X. Diebold, K. Ylmaz, On the network topology of variance decompositions: Measuring the connectedness of financial firms, Journal of Econometrics 182 (1) (2014) 119 – 134, causality, Prediction, and Specification Analysis: Recent Advances and Future Directions. doi:http://dx.doi.org/10.1016/j.jeconom.2014.04.012.
URL http://www.sciencedirect.com/science/article/pii/S0304407614000712

[54] S. Saavedra, L. J. Gilarranz, R. P. Rohr, M. Schnabel, B. Uzzi, J. Bascompte, Stock fluctuations are correlated and amplified across networks

of interlocking directorates, EPJ Data Science 3 (1) (2014) 30. doi:10.1140/epjds/s13688-014-0030-0.
URL https://doi.org/10.1140/epjds/s13688-014-0030-0

[55] H. Chen, Y. Mai, S.-P. Li, Analysis of network clustering behavior of the chinese stock market, Physica A: Statistical Mechanics and its Applications 414 (2014) 360 – 367. doi:http://dx.doi.org/10.1016/j.physa.2014.07.039.
URL http://www.sciencedirect.com/science/article/pii/S0378437114006116

[56] W.-Q. Huang, X.-T. Zhuang, S. Yao, A network analysis of the chinese stock market, Physica A: Statistical Mechanics and its Applications 388 (14) (2009) 2956 – 2964. doi:http://dx.doi.org/10.1016/j.physa.2009.03.028.
URL http://www.sciencedirect.com/science/article/pii/S0378437109002519

[57] F. Ren, W.-X. Zhou, Dynamic evolution of cross-correlations in the chinese stock market, PLoS ONE 9 (5) (2014) e97711. doi:10.1371/journal.pone.0097711.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0097711

[58] J. Birch, A. Pantelous, K. Soramäki, Analysis of correlation based networks representing dax 30 stock price returns, Computational Economics (2015) 1–25doi:10.1007/s10614-015-9481-z.
URL http://dx.doi.org/10.1007/s10614-015-9481-z

[59] J. G. Brida, W. A. Risso, Hierarchical structure of the german stock market, Expert Systems with Applications 37 (5) (2010) 3846 – 3852. doi:http://dx.doi.org/10.1016/j.eswa.2009.11.034.
URL http://www.sciencedirect.com/science/article/pii/S0957417409009762

[60] P. Caraiani, Characterizing emerging european stock markets through complex networks: From local properties to self-similar characteristics, Physica A: Statistical Mechanics and its Applications 391 (13) (2012) 3629 – 3637. doi:http://dx.doi.org/10.1016/j.physa.2012.02.008.
URL http://www.sciencedirect.com/science/article/pii/S0378437112001240

[61] K. Kosmidou, D. Kousenidis, A. Ladas, C. Negkakis, Determinants of risk in the banking sector during the european financial crisis, Journal of Financial Stabilitydoi:https://doi.org/10.1016/j.jfs.2017.06.006.
URL http://www.sciencedirect.com/science/article/pii/S1572308917304412

[62] N. Paltalidis, D. Gounopoulos, R. Kizys, Y. Koutelidakis, Transmission channels of systemic risk and contagion in the european financial network, Journal of Banking & Finance 61 (Supplement 1) (2015) S36 – S52, global Trends in Banking, Regulations, and Financial Markets. doi:https://doi.org/10.1016/j.jbankfin.2015.03.021.
URL http://www.sciencedirect.com/science/article/pii/S0378426615000989

[63] M. F. da Silva, É. J. de Area Leão Pereira, A. M. da Silva Filho, A. P. N. de Castro, J. G. V. Miranda, G. F. Zebende, Quantifying cross-correlation between ibovespa and brazilian blue-chips: The DCCA approach, Physica A: Statistical Mechanics and its Applications 424 (2015) 124 – 129. doi:http://dx.doi.org/10.1016/j.physa.2015.01.002.
URL http://www.sciencedirect.com/science/article/pii/S0378437115000047

[64] J. Lee, J. Youn, W. Chang, Intraday volatility and network topological properties in the korean stock market, Physica A: Statistical Mechanics and its Applications 391 (4) (2012) 1354 – 1360. doi:http://dx.doi.org/10.1016/j.physa.2011.09.016.

URL http://www.sciencedirect.com/science/article/pii/S0378437111007321

[65] A. Vizgunov, B. Goldengorin, V. Kalyagin, A. Koldanov, P. Koldanov, P. Pardalos, Network approach for the russian stock market, Computational Management Science 11 (1-2) (2014) 45–55. doi:10.1007/s10287-013-0165-7.
URL http://dx.doi.org/10.1007/s10287-013-0165-7

[66] S. Martinez-Jaramillo, B. Alexandrova-Kabadjova, B. Bravo-Benitez, J. P. Solórzano-Margain, An empirical study of the mexican banking system's network and its implications for systemic risk, Journal of Economic Dynamics and Control 40 (Supplement C) (2014) 242 – 265. doi:https://doi.org/10.1016/j.jedc.2014.01.009.
URL http://www.sciencedirect.com/science/article/pii/S0165188914000189

[67] E. Jondeau, E. Jurczenko, M. Rockinger, Moment component analysis: An illustration with international stock markets, Journal of Business & Economic Statistics (2017) 1–23.

[68] R. Kali, J. Reyes, The architecture of globalization: a network approach to international economic integration, Journal of International Business Studies 38 (4) (2007) 595–620. doi:10.1057/palgrave.jibs.8400286.
URL http://dx.doi.org/10.1057/palgrave.jibs.8400286

[69] P. Giudici, A. Spelta, Graphical network models for international financial flows, Journal of Business & Economic Statistics 34 (1) (2016) 128–138. arXiv:http://dx.doi.org/10.1080/07350015.2015.1017643, doi:10.1080/07350015.2015.1017643.
URL http://dx.doi.org/10.1080/07350015.2015.1017643

[70] N. Cetorelli, S. Peristiani, Prestigious stock exchanges: A network analysis of international financial centers, Journal of

Banking & Finance 37 (5) (2013) 1543 – 1551. doi:https://doi.org/10.1016/j.jbankfin.2012.06.011.
URL http://www.sciencedirect.com/science/article/pii/S037842661200163X

[71] L. Sandoval, Structure of a global network of financial companies based on transfer entropy, Entropy 16 (8) (2014) 4443–4482. doi:10.3390/e16084443.
URL http://www.mdpi.com/1099-4300/16/8/4443

[72] P. Caraiani, Using complex networks to characterize international business cycles, PLoS ONE 8 (3) (2013) e58109. doi:10.1371/journal.pone.0058109.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0058109

[73] T. C. Silva, S. R. S. de Souza, B. M. Tabak, Structure and dynamics of the global financial network, Chaos, Solitons & Fractals 88 (Supplement C) (2016) 218 – 234, complexity in Quantitative Finance and Economics. doi:https://doi.org/10.1016/j.chaos.2016.01.023.
URL http://www.sciencedirect.com/science/article/pii/S0960077916300145

[74] L. LALOUX, P. CIZEAU, M. POTTERS, J.-P. BOUCHAUD, Random matrix theory and financial correlations, International Journal of Theoretical and Applied Finance 03 (03) (2000) 391–397. arXiv:http://www.worldscientific.com/doi/pdf/10.1142/S0219024900000255, doi:10.1142/S0219024900000255.
URL http://www.worldscientific.com/doi/abs/10.1142/S0219024900000255

[75] I. I. Dimov, P. N. Kolm, L. Maclin, D. Y. C. Shiber, Hidden noise structure and random matrix models of stock correlations, Quantitative Finance 12 (4) (2012) 567–572.

[76] G. Bonanno, G. Caldarelli, F. Lillo, R. N. Mantegna, Topology of correlation-based minimal spanning trees in real and model markets, Phys. Rev. E 68 (2003) 046130. doi:10.1103/PhysRevE.68.046130.
URL http://link.aps.org/doi/10.1103/PhysRevE.68.046130

[77] M. Tumminello, T. Aste, T. Di Matteo, R. N. Mantegna, A tool for filtering information in complex systems, Proceedings of the National Academy of Sciences of the United States of America 102 (30) (2005) 10421–10426. arXiv:http://www.pnas.org/content/102/30/10421.full.pdf, doi:10.1073/pnas.0500298102.
URL http://www.pnas.org/content/102/30/10421.abstract

[78] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, A. Kanto, Asset trees and asset graphs in financial markets, Physica Scripta 2003 (T106) (2003) 48.
URL http://stacks.iop.org/1402-4896/2003/i=T106/a=011

[79] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, A. Kanto, Dynamics of market correlations: Taxonomy and portfolio analysis, Phys. Rev. E 68 (2003) 056110. doi:10.1103/PhysRevE.68.056110.
URL http://link.aps.org/doi/10.1103/PhysRevE.68.056110

[80] L. China Securities Index Company, China securities index 300 index, [Online; accessed 9-October-2017] (2017).
URL http://www.csindex.com.cn/sseportal_en/csiportal/zs/jbxx/report.do?code=000300

[81] S&p 500 index, [Online; accessed 9-October-2017] (2017).
URL http://us.spindices.com/indices/equity/sp-500

[82] J.-P. Onnela, et al., Complex networks in the study of financial and social systems, Ph.D. thesis, Helsinki University of Technology (2006).

[83] W. Jang, J. Lee, W. Chang, Currency crises and the evolution of foreign exchange market: Evidence from minimum spanning tree, Physica A:

Statistical Mechanics and its Applications 390 (4) (2011) 707 – 718. doi:http://dx.doi.org/10.1016/j.physa.2010.10.028.
URL http://www.sciencedirect.com/science/article/pii/S0378437110008861

[84] T. Di Matteo, F. Pozzi, T. Aste, The use of dynamical networks to detect the hierarchical organization of financial market sectors, The European Physical Journal B 73 (1) (2010) 3–11. doi:10.1140/epjb/e2009-00286-0.
URL http://dx.doi.org/10.1140/epjb/e2009-00286-0

[85] J. B. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem, Proceedings of the American Mathematical society 7 (1) (1956) 48–50.

[86] G. Bonanno, G. Caldarelli, F. Lillo, S. Micciché, N. Vandewalle, R. Mantegna, Networks of equities in financial markets, The European Physical Journal B - Condensed Matter and Complex Systems 38 (2) (2004) 363–371. doi:10.1140/epjb/e2004-00129-6.
URL http://dx.doi.org/10.1140/epjb/e2004-00129-6

[87] R. N. Mantegna, H. E. Stanley, An Introduction to Econophysics: Correlations and Complexity in Finance, Cambridge University Press, New York, NY, USA, 2000.

[88] C. Tu, Cointegration-based financial networks study in chinese stock market, Physica A: Statistical Mechanics and its Applications 402 (2014) 245 – 254. doi:http://dx.doi.org/10.1016/j.physa.2014.01.071.
URL http://www.sciencedirect.com/science/article/pii/S0378437114000983

[89] G.-J. Wang, C. Xie, Correlation structure and dynamics of international real estate securities markets: A network perspective, Physica A: Statistical Mechanics and its Applications 424 (2015) 176 – 193.

doi:http://dx.doi.org/10.1016/j.physa.2015.01.025.
URL http://www.sciencedirect.com/science/article/pii/S0378437115000278

[90] R. E. Mehmet Eryiğita, Network structure of cross-correlations among the world market indices, Physica A: Statistical Mechanics and its Applications 388 (17) (2009) 3551 – 3562. doi:http://dx.doi.org/10.1016/j.physa.2009.04.028.
URL http://www.sciencedirect.com/science/article/pii/S0378437109003318

[91] T. Di Matteo, T. Aste, Extracting the correlation structure by means of planar embedding, Proc. SPIE 6039 (2005) 60390P–1–60390P–10. doi:10.1117/12.637543.
URL http://dx.doi.org/10.1117/12.637543

[92] J. Birch, A. A. Pantelous, K. Zuev, The maximum number of 3- and 4-cliques within a planar maximally filtered graph, Physica A: Statistical Mechanics and its Applications 417 (2015) 221 – 229. doi:http://dx.doi.org/10.1016/j.physa.2014.09.011.
URL http://www.sciencedirect.com/science/article/pii/S0378437114007699

[93] N. Musmeci, T. Aste, T. Di Matteo, Relation between financial market structure and the real economy: Comparison between clustering methods, PLoS ONE 10 (3) (2015) e0116201. doi:10.1371/journal.pone.0116201.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0116201

[94] F. Pozzi, T. Aste, G. Rotundo, T. Di Matteo, Dynamical correlations in financial systems, Proc. SPIE 6802 (2007) 68021E–1–68021E–11. doi:10.1117/12.758822.
URL http://dx.doi.org/10.1117/12.758822

[95] J. Liu, C. K. Tse, K. He, Fierce stock market fluctuation disrupts scalefree distribution, Quantitative Finance 11 (6) (2011) 817–823. arXiv:http://dx.doi.org/10.1080/14697680902991627, doi:10.1080/14697680902991627.
URL http://dx.doi.org/10.1080/14697680902991627

[96] V. Batagelj, A. Mrvar, Pajek-program for large network analysis, Connections 21 (2) (1998) 47–57.

[97] Y. Mai, H. Chen, L. Meng, An analysis of the sectorial influence of csi300 stocks within the directed network, Physica A: Statistical Mechanics and its Applications 396 (2014) 235 – 241. doi:http://dx.doi.org/10.1016/j.physa.2013.11.016.
URL http://www.sciencedirect.com/science/article/pii/S0378437113010583

[98] R. Rammal, G. Toulouse, M. A. Virasoro, Ultrametricity for physicists, Rev. Mod. Phys. 58 (1986) 765–788. doi:10.1103/RevModPhys.58.765.
URL http://link.aps.org/doi/10.1103/RevModPhys.58.765

[99] R. Yang, X. Li, T. Zhang, Analysis of linkage effects among industry sectors in china's stock market before and after the financial crisis, Physica A: Statistical Mechanics and its Applications 411 (2014) 12 – 20. doi:http://dx.doi.org/10.1016/j.physa.2014.05.072.
URL http://www.sciencedirect.com/science/article/pii/S0378437114004658

[100] E. Kantar, M. Keskin, B. Deviren, Analysis of the effects of the global financial crisis on the turkish economy, using hierarchical methods, Physica A: Statistical Mechanics and its Applications 391 (7) (2012) 2342 – 2352. doi:http://dx.doi.org/10.1016/j.physa.2011.12.014.
URL http://www.sciencedirect.com/science/article/pii/S0378437111009113

[101] A. Rea, W. Rea, Visualization of a stock market correlation matrix, Physica A: Statistical Mechanics and its Applications 400 (2014) 109 – 123. doi:http://dx.doi.org/10.1016/j.physa.2014.01.017.
URL http://www.sciencedirect.com/science/article/pii/S0378437114000211

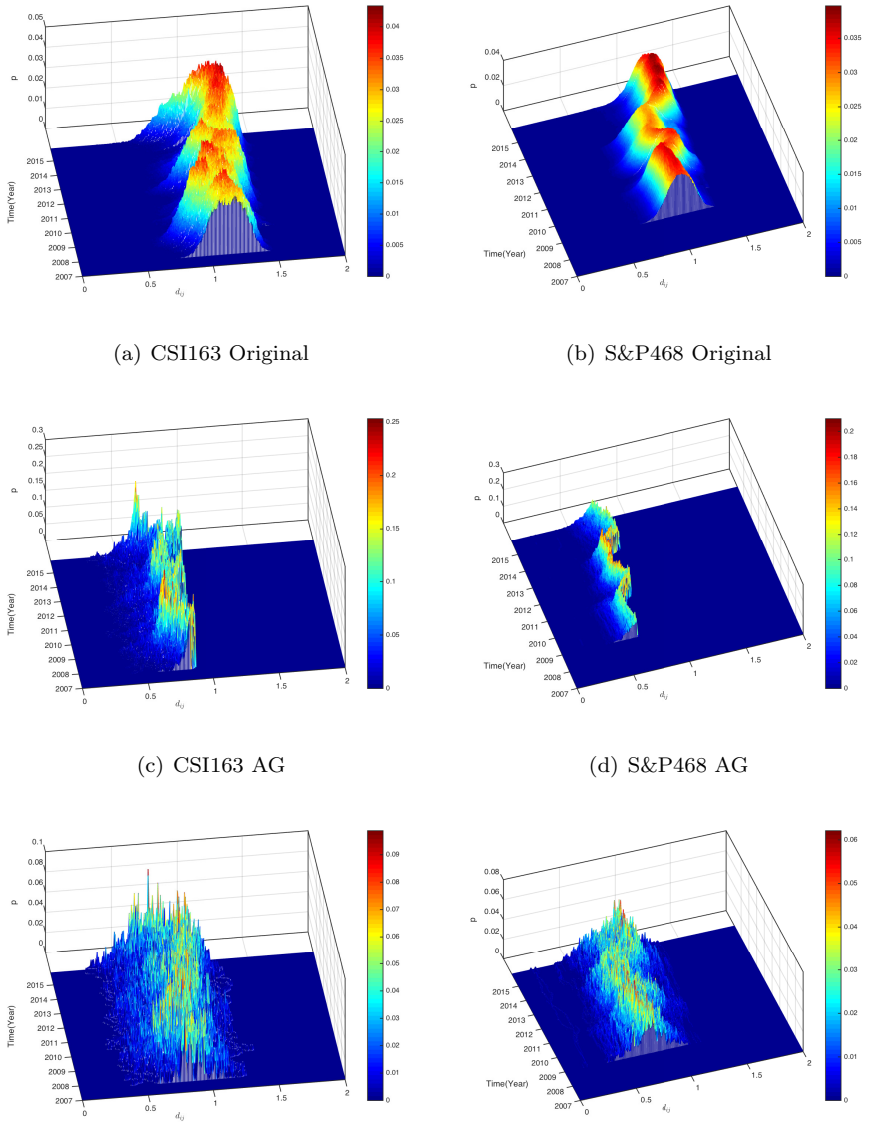
[102] R. Prim, Shortest connection networks and some generalizations, Bell System Technical Journal, The 36 (6) (1957) 1389–1401. doi:10.1002/j.1538-7305.1957.tb01515.x.

[103] R. J. Elliott, J. V. D. H. *, W. P. Malcolm, Pairs trading, Quantitative Finance 5 (3) (2005) 271–276. arXiv:http://dx.doi.org/10.1080/14697680500149370, doi:10.1080/14697680500149370.
URL http://dx.doi.org/10.1080/14697680500149370

[104] E. Gatev, W. N. Goetzmann, K. G. Rouwenhorst, Pairs trading: Performance of a relative-value arbitrage rule, Review of Financial Studies 19 (3) (2006) 797–827. arXiv:http://rfs.oxfordjournals.org/content/19/3/797.full.pdf+html, doi:10.1093/rfs/hhj020.
URL http://rfs.oxfordjournals.org/content/19/3/797.abstract

[105] S. Mudchanatongsuk, J. A. Primbs, W. Wong, Optimal pairs trading: A stochastic control approach, in: 2008 American Control Conference, 2008, pp. 1035–1039. doi:10.1109/ACC.2008.4586628.

[106] J. Wang, C. Rostoker, A. Wagner, A high performance pair trading application, in: Parallel Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on, 2009, pp. 1–8. doi:10.1109/IPDPS.2009.5161147.



(a) CSI163 Original

(b) S&P468 Original

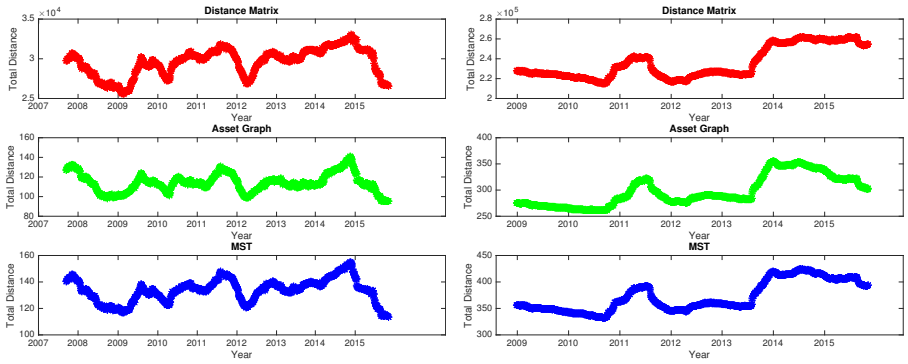
(c) CSI163 AG

(d) S&P468 AG

(e) CSI163 MST

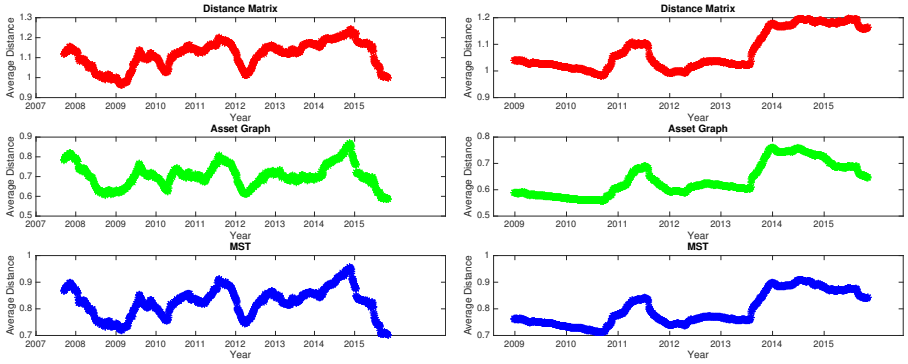
(f) S&P468 MST

Figure 23: Probability distributions of all distances d_{ij} of N_t, AG_t, and MST_t of CSI163 and S&P468 over the years in our study period between 04/01/2007 and 06/11/2015. The total number of edges is $N(N-1)/2$ for N_t, and $N-1$ for AG_t and MST_t respectively. Since the sliding window size $L_{CSI163} = 170$ and $L_{S\&P468} = 500$, so the data only starts after a period of L.



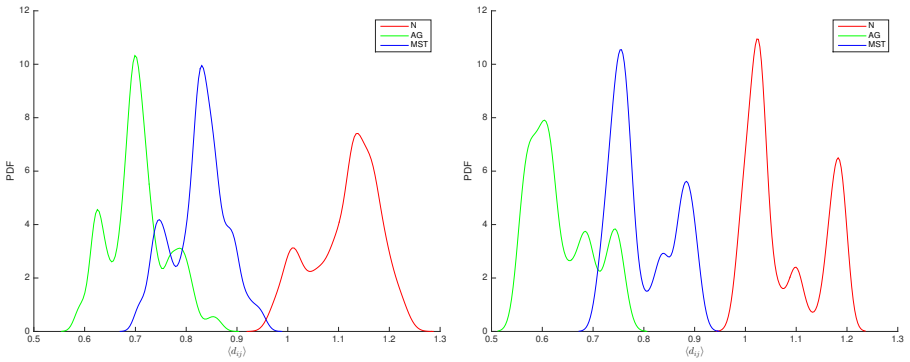
(a) CSI Network (b) S&P468 Network

Figure 24: The evolving of total distances d_{total} of original network N_t, asset graph AG_t, and minimum spanning tree MST_t for CSI163 and S&P468 over time in the study period.



(a) CSI Network (b) S&P468 Network

Figure 25: The evolving of average distances $\langle d_{ij} \rangle$ of original network N_t, asset graph AG_t, and minimum spanning tree MST_t for CSI163 and S&P468 over time in the study period.



(a) CSI163 Network

(b) S&P468 Network

Figure 26: Probability density function (PDF) of average distance $\langle d_{ij} \rangle$ of original network N, asset graph AG, and minimum spanning tree MST for CSI163 and S&P468.