COMPLEX MUTUAL INFORMATION-THEORETIC STOCK NETWORKS

by

Serkan Alkan

A DISSERTATION

Submitted to the Faculty of the Stevens Institute of Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

_____

Serkan Alkan, Candidate

ADVISORY COMMITTEE

_____

Dr. Khaldoun Khashanah, Chairman Date

_____

Dr. Charles Suffel Date

_____

Dr. Ionut Florescu Date

_____

Dr. Rupak Chatterjee Date

STEVENS INSTITUTE OF TECHNOLOGY
Castle Point on Hudson
Hoboken, NJ 07030
2019

COMPLEX MUTUAL INFORMATION-THEORETIC STOCK NETWORKS

ABSTRACT

Financial markets can be characterized as complex systems because of interactions between heterogeneous components and existing nonlinearities. Network theory can be used to model the financial market in which nodes can be stocks, commodities or currencies. The current state of the art is to measure distances between nodes using Pearson correlations. In this thesis, mutual information (MI) is applied to describe similarities between assets. MI is a general measure of dependency and can detect linear and non-linear relationships whereas Pearson correlation fails to detect the nonlinear relationships.

This dissertation consists of three essays. In essay 1, we propose a comprehensive comparison approach between the mutual information (MI) metric and the Pearson correlation metric. We find that networks constructed by these two measures become very different during depending on market regimes and network topological structure. We pay special attention to the cases where two measures are strongly mismatched and examine the reasons. Relationship between the entropy and the moments of daily stock return distributions are investigated and we find very strong relationship between entropy and the kurtosis. We compare the performance of MI and correlation techniques in terms of identification of coherent and well-separated stock communities and results show that MI approaches perform better in crisis and non-crisis periods.

In essay 2, we analyzed how local, mesoscopic and global topological properties of mutual information based stock networks evolve annually. To detect and quantify the impact of a major crisis on the market network structure, we propose to use

the information-theoretic quantifiers. We observe that the entropy of the system relatively reduces during crisis regimes. We propose to use classic internal cluster indices and information-theoretic quantifiers to capture the topological evolution of sectors by treating them as fixed communities, i.e. labels of stocks do not change over time in the system. We find that the structural changes of the financial sector during the subprime crisis and it has the strongest correlation with markets in terms of structural uncertainty.

In essay 3, we analyze how the homogeneity in each aggregation level of Global Industry Classification System (GICS) scheme changes over time and identify the industries which have more homogeneous structure than others in terms of stock returns comovement. We propose a mesoscopic approach in terms of network-theoretic framework to describe the relationship between industry groups. We investigate the time evolution of the interaction structure of the industry groups and identify the important ones in the market. Our analysis reveals that during a crisis period, banks, as an industry group, become more important in the market and the local interaction structure of industry groups becomes much simpler or a star network topology.

In conclusion, we find that some asset pairs have non-linear relationships and MI provides a better alternative measure to define the links in financial networks. MI network and Pearson network become very different at local and global topological scales during various regimes. Community detection algorithms yield well-separated stock communities with MI networks comparing Pearson ones. Besides the classic network measures, information-theoretic measures provide an information advantage in revealing nonlinear dependencies to quantify and detect financial crises and regime switching.

Author: Serkan Alkan

Advisor: Dr. Khaldoun Khashanah

Date: March 25, 2019

Department: Financial Engineering

Degree: Doctor of Philosophy

Acknowledgments

I would like to express my sincere appreciation to my thesis supervisor, Dr. Khaldoun Khashanah for all his support and encouragement. He provided me with wise guidance and reliable counsel, and challenging me into learning and opening new research questions. Without his constructive criticism, inspiration, and support, this dissertation would have been impossible.

I want to thank the other members of my committee Dr. Charles Suffel, Dr. Ionut Florescu and Dr. Rupak Chatterjee for your valuable suggestions and insights.

My special thanks go to my colleague Baris Morkan with whom I have been able to engage in intellectual discussions and supportive talks over coffee.

I would like to acknowledge Abdullah Ibrahim Khanfor and Saud AlMahdi for their moral support and motivation, which encourage me to give my best.

I would also like to to especially acknowledge and thank my friend Pacita for her continuing support, infinite patience, friendship and encouragement. I am grateful for her friendship.

Last but not least, my deepest gratitude goes to my family for their unwavering love and support. It is only because of their constant encouragement, countless sacrifices, and unconditional support that any of my aspirations have been realized.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1 Background and Motivation

Many complex systems have been analyzed by complex networks such as the World Wide Web [4], Internet [35], social networks [77], food web [44], scientific citations [90], sexual contacts among individuals [62] and financial systems [56]. In each network model, nodes and links represent different meanings according to the system considered. For example, in social networks, nodes represent people or groups and links show the existence relationship or friendship between individuals.

Financial markets can be characterized as complex systems because of interaction between heterogeneous components and existing nonlinearity [65]. Network theory can be used to model the financial market, in which nodes can be stocks [5, 80], commodities [94, 1, 94], currency [71, 66, 69], interest rates as in [28], and banks [14]. The links can represent the correlation between financial assets, or display the bilateral exposure between any two banks in the system.

Network theory has been employed in the stock markets not only to filter the correlation matrix but also to extract the communities (clusters, modules, or sectors) from it. The most widely used methods are Minimal Spanning Tree (MST) [65], Planar Maximally Filtered Graphs [99], asset trees and asset graphs [82]. They aim to find out the subgraphs of the complete graph with $\frac{n(n-1)}{2}$ links created from a correlation matrix, where n represents the stock. Links aim to preserve the most relevant information about pair correlation coefficient between stocks in order to retain the market's core structure. Stock networks have been applied into a number of

problems in finance, such as portfolio optimization [81], prediction of market direction [58], evaluation of systemic risk by using some properties of network structure [57, 9], to measure the stability of market [47], and prediction of national economic growth [48].

No matter what the application of financial networks, the first step requires the selection of co-expression measure to define networks among stocks. The most common choice is the Pearson's correlation coefficient, but it captures the pairwise relationship very well when the data follows multivariate normal distribution, and more generally for spherical and elliptical distributions [27]. Also, it only works well if returns are linearly associated and fails to detect any non-linear relationships. However, empirical research in finance shows that the (unconditional) distribution of returns displays a heavy tail with positive excess kurtosis, which is in contrast to the behavior of a normally distributed variable [22, 100]. Furthermore, zero correlation does not imply statistically independent. It only means linear independence and it is possible that there may be some non-linearity.

In order to account both linear and nonlinear associations between nodes, one needs to describe the nodal inter-dependency in a more general sense than correlations. There are several statistical association measures have been proposed based on ranks or information theory [27]. Mutual information, introduced by [93], provides a general measurement for dependencies, such as non-linear or non-functional relationships and is a measure of how much information two systems exchange or two data sets share. Furthermore, by using the definition of statistical independence between two random variables, it can be shown that mutual information $I(X;Y) = 0$ if and only if X and Y are independent random variables. Mutual information has been used as co-expression measure to model a complex system such as gene regulatory networks in bioinformatics [17], climate system [31], complex brain networks [6], and

recently for stock networks [37].

One of the main motivations of the research is to establish the theoretical background for defining weighted network among stocks by using mutual information as co-expression measure. Therefore, a deeper assessment of the existence of nonlinearity in stock return time series and its impact on stock network properties are conducted to find out the advantages and disadvantages of both measures empirically. The second motivation is investigating the interplay between dependence structure and financial crisis by analyzing the annual mutual information stock networks and illustrate the capability of the information-theoretic measures to predict and quantify a crisis. Finally, an alternative simplification method for a financial market is proposed at mesoscopic scales in order to analyze how the roles of industries and their interactions change over time especially which industries become dominant in crisis periods.

1.2 Research Objectives and Questions

One of the purposes of the study is to clarify the essential mathematics behind Pearson's correlation coefficient and mutual information used to model stock networks. Furthermore, we investigate the differences between the two measures and the impact of the nonlinearity on network measures. In addition, this research aims to explore how the market and its components are interrelated and what kind of collective behaviors emerge differently during crisis periods, which could be used as early-warning indicators of an impending crisis. Finally, this work presents a new network-information-theoretic approach to exhibit the market at the mesoscopic level (i.e., sector or industry groups). We analyze how the homogeneity of each level of classification schemes change over time regarding stock return comovements and whether

there is a big difference between each other. The primary objective of this research is to answer the following questions:

Research Question 1:

What associative inter-relations can mutual information measure uncover that Pearson correlation or regression models cannot discover?

We compare the Pearson correlation and mutual information in order to characterize the nonlinearity in stock returns and how substantially the networks constructed by two measures are different from on local and global topological scales. Mutual information networks and Pearson correlation networks are very different and in crisis period this difference significantly increases. Performance of community detection algorithms is tested on the networks constructed by six different distance measures and we find that communities identified on mutual information networks has more compact and well-separated structures than other distance measures. We find that kurtosis and entropy have very high negative Spearman correlation in crisis and non-crisis periods. That indicates one of the main reason of mismatch measures between mutual information and Pearson correlation. Spline regression and polynomial regression based stock networks are illustrated and they also confirm the existence of nonlinear dependency between stock returns.

Research Question 2:

Can the measure of connectedness of stock networks identify and quantify a financial crisis period?

We find that stock networks significantly change at the local, global, and mesoscopic topological scales during a crisis. In addition to the significant shrinkage at global level as detected by the classic network quantifiers, entropy of the system indicates lower values in crisis periods than non-crisis periods, i.e market structure becomes less random. Furthermore, degree distribution of stock networks presented

scale-free structure for the non-crisis periods; however, during crisis periods, exponent of scaling α had values slightly less than 2.

Research Question 3:

How do each GICS aggregation perform in terms of forming groups of stocks whose price movements are related? Which industries have more homogeneous structure? How do local interaction structures of industry groups change during crisis periods?

We propose a technique to construct a network whose nodes are the industry groups in order to investigate how the local industry group interactions change over time. Investigating time evolution of the industry group network structure can help us identify which industry groups that experience important changes over various market regimes. We find that during the crisis industry group network structure becomes very simple and only one single industry group-4010 Banks becomes hub, i.e dominates the market.

The current work contributes in the following areas and applications: First, we provide the theoretical background for mutual information and furthermore illustrate how to construct stock networks based on mutual information metrics. We propose a procedure not only to evaluate the difference between mutual information and the Pearson correlation but also to quantify the impact of the nonlinearity on some stock network properties by comparing the networks constructed based on each association measure. Second, besides the classic network quantifiers, we propose information-theoretic quantifiers to identify and quantify the financial crisis periods. Third, we evaluate the homogeneity and separation scores of each industry classification schemes. We propose novel complex network-theoretic methodology to construct networks whose nodes are industry groups in order to evaluate the role of industries in the market and their local interactions over time.

1.3 Key Artifacts

This section presents a summary of the publications involved in the dissertation work. These artifacts serve as key milestones during the preparation of this dissertation, and all publications support the research work.

- "Structural Evolution of Stock Networks." Signal-Image Technology & Internet-Based Systems (SITIS), 2015 11$^{\text{th}}$ International Conference on. IEEE, 2015.
- Comparing the Quality Functions for Community Detection *22nd Asia-Pacific Conference on Global Business, Economics, Finance & Social Sciences, 2019*
- Linear and Nonlinear Hierarchical Stock Network Methods, *submitted*
- Dynamic Evolution of Complex Mutual Information Theoretic Stock Networks *working paper.*
- Industry Classifications and Identification of Important Industry Groups *working paper.*

1.4 Structure of the Thesis

This dissertation consists of three essays and each essay is written as self-contained.

In Chapter 2, we describe the mutual information for defining weighted networks among stocks. Several approaches based on prediction methods, such as polynomial regression and spline regression models are explained to construct stock networks. Basic terminology and network concepts are introduced. Network statistics is applied to describe the topological properties of a single stock network and for comparing two or more stock networks constructed by different metrics. The relationship between correlation-based measures and mutual information, and also the impact of

the nonlinearity on stock networks are investigated. The performance of community detection algorithms on stock networks constructed by different metrics are evaluated regarding to identification of stocks' true sectors and homogeneity of clusters.

In chapter 3, the structural evolution of mutual information stock networks are investigated and the drastic topological changes are observed at local, global, and mesoscopic levels in crisis periods. Besides shrinkage of market, the entropy of the system becomes lower in crisis period than usual periods. Furthermore, new statistical network concepts are introduced to characterize the mesoscopic structure of the market (e.g. sector) and their time evolution within the system.

In chapter 4, we evaluate the homogeneity and separation scores of each industry classification schemes and propose complex network-theoretic methodology to construct networks whose nodes are industry groups in order to identify the dominant industries over time.

Finally, chapter 5 summarizes the key findings of this dissertation.

Chapter 2

Essay 1. Linear and Nonlinear Hierarchical Stock Network Methods

Abstract In this work, the role of distance measures in cluster analysis and in a stock network construction is investigated. We propose a comprehensive comparison approach between the mutual information (MI) metric and the Pearson correlation metric. Polynomial and natural cubic spline regressions are used to detect non-linear relationships between stock returns and compared to MI as an alternative. In order to measure the impact of the recent financial crisis on nonlinearity, we use two data sets composed of log-returns of daily adjusted close prices of 402 stocks of the S&P500. While the first one is in the time period from January 2007 to December 2009 as crisis period, the second one from January 2012 to December 2015 representing the non-crisis period. For each distance measure, we construct the hierarchical stock networks using the minimum spanning tree and compare their local and global properties. Furthermore, we examine the relationship between the performance of community detection algorithm and selection of metric as compared to graph-theoretic internal cluster validity indices and external indices.

2.1 Introduction

Complex networks have been one of a very popular tool in various scientific domains, such as physics, biology, social science, computer networks, and financial networks [74]. In these systems, nodes represent the entities and links show the interaction between them. There are several related issues involved in the analysis of such complex networks and those issues include in the selection of a clustering algorithm, the selection of a metric, and network construction.

Clustering or cluster analysis is an unsupervised machine learning method and is the process of classifying objects in data into clusters or groups such that all objects in the same clusters are more similar to each other than objects in other clusters and the proposed clusters are maximally separated. Cluster analysis has been implemented in many areas, such as gene expression analysis [107], social networks [76], image segmentation [103] and finance [70]. Due to its application to different areas, it has been identified with its application outcome, for example, segmentation, partitioning, and community detection.

Clustering algorithms can be grouped into four different groups [98]: hierarchical, partitional, exclusive, overlapping, fuzzy, and complete or partial. Most of the very well-known classic algorithms, such as k-means or self-organizing maps, are unable to discover the true cluster structure in data which does not come from Gaussian distribution or spherical distribution [39]. Graph clustering algorithms can handle these issues and are successful in finding out the clusters of different sizes and shapes in data provided that clusters are well-separated [54]. In graph clustering algorithms, the aim is to find the cluster of the objects in a graph, in which objects are represented as node and edges indicate similarities or dissimilarities between objects.

Besides the selection of clustering algorithms, one of the harder tasks in cluster analysis is deciding what distance function should be used for the data. The problem becomes much more challenging if external information (no benchmark) is not available, which is the typical case in most clustering analysis. The distance measure plays a key role in identifying clusters and relationships between objects in a network.

Many financial systems can be efficiently modeled using a network structure where the system entities can be stocks, commodities, countries, banks, or firms and the relations between the entities are the network links. These links could be weighted or unweighted and directed or undirected. For example, the interbank lending mar-

ket, which is considered as a network where banks are the vertices and the claims and liabilities between banks define the weight of links is an example of directed and weighted networks. Moreover, the weights between banks or entities are defined through the bilateral exposure between any two in the system. This type of connectivity is sometimes called *structural connectivity*. However, in some case, the weights need to be inferred from the data as a first step to construct the network. This type of connectivity is sometimes called *functional connectivity* [45].

Co-expression measures are often used to define networks among entities in the system, such as stock data and gene expression data. The definition of the distance measure is a key factor for successful identification of the relationships between stocks and networks. Different similarity measures are likely to result in different networks and clusters, although based on the same expression data.

Pearson correlation based similarity or distance is one of the most common metrics used to compute dependence between stock returns to construct stock networks, however; it only works well if returns are linearly associated and fails to detect non-linear relations. Furthermore, zero correlation does not imply statistical independence. Uncorrelated returns only means linear independence while it is possible that there may be some non-linear dependence. On the other hand, zero mutual information (MI) means stock returns are independent and furthermore MI provides a general measurement for dependencies, such as non-linear relationships. Furthermore, several studies have detected the non-linearity in finance: stock returns [15, 88, 79, 57], market index returns [2, 41, 3] and currency exchange rates [15, 49, 67, 16, 89].

Within the large body of research on stock networks, there are few publications that systematically explore the appropriateness of chosen similarity measures. In this essay, we provide a theoretical background to evaluate the MI between stock returns and furthermore we propose a procedure compare the mutual information (MI) metric

versus Pearson correlation metric. We construct a separate minimum spanning tree hierarchical stock networks based on each metric and compare their local and global properties. We conduct a systematic approach to explain the reasons for the cases where two measures strongly disagree. We illustrate that it is possible to construct stock networks through predictive models, such as polynomial regression and natural cubic spline, which are an improvement of classic linear regression in order to take into account nonlinear interactions between objects. Furthermore, we also compare the performance of the community detection algorithm by using internal validation indices from classic cluster analysis to adapted the graph clustering and external indices. We choose the Global Industry Classification Standard (GICS) classification of stocks as the benchmark, i.e, the ground truth cluster sets, in terms of sector level and industry group level in order to compare the performance of the community detection algorithms.

The paper is organized as follows. In section 2.2, the data set is described. In section 2.3, we discuss the similarity measures and constructing the stock networks. After that, we briefly define the tools from complex networks theory used in this work and then explain the community detection algorithms. In sections 2.5 and 2.6, we show that how classic internal validation is adapted to graph clustering. In sections 2.7 and 2.8, we introduce our results and in the last section, we summarize our results.

2.2 Data

In this study, the data set is composed of log-returns of daily adjusted close prices of 402 stocks of SP500. We consider two separate time periods. First one consists of the period from the beginning of 2007 to the end of 2009 and we call it as the crisis period. Second data set extends from the beginning of 2012 to the end of 2015

and it is called as the non-crisis period. In both data, we remove the days if it has incomplete data. While the length of crisis period data set is 756 consecutive trading days, the length of non-crisis data is 1005 days.

2.3 Dissimilarity Measures and Network Construction

In order to construct the hierarchical stock networks based on minimum spanning tree (MST), we first need to compute all pairwise similarity between each stock in the portfolio, then transform them into dissimilarity. A dissimilarity measure on a vector space X is called a metric if it is a mapping $d : X \times X \to \mathcal{R}$ satisfying the metric axioms:

M1) $d(x, y) = 0$ if and only if x=y;
M2 $d(x, y) = d(y, x)$;
M3) $d(x, z) \leq d(x, y) + d(y, z)$

In clustering, dissimilarity measures are not required to be a valid metric for practical reasons and some algorithms, such as partition around medoids (PAM) [55], can accommodate the dissimilarity properties.

MST stock network generation is one of the popular methods in the literature and is introduced by Mantegna [64]. The construction process is defined as follows: we start with an arbitrary stock as the root of a partial tree and at every step, the partial tree grows by iteratively adding an unconnected stock to it by choosing the lowest weight, until the unconnected stock set is exhausted, which is known as Prims algorithm [87]. MST of order N has exactly N-1 links and no loops or circuits. Furthermore, MST has a strong relationship with the single linkage clustering

algorithm [51].

In the next section, we briefly describe the similarity and dissimilarity measures to generate MST-based stock networks. We assume that we have two numeric vectors X and Y of size n.

2.3.1 Pearsons product-moment correlation coefficient

The Pearsons product-moment correlation is a measure of linear dependence between two vectors X and Y [43, 84]. Pearson sample correlation, r_p, is defined as follows:

$$r_p(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.3.1}$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ are sample means of the vector X and Y, respectively.

It can also be defined as the cosine correlation between two scaled vectors, of which each has sample mean 0 and sample variance 1. It can easily be shown that sample correlation is scale-invariant with respect to linear transformations and takes values in interval $-1 \leq r_p(X,Y) \leq 1$. In case $r_p(X,Y) = 0$, it does not mean random variables X and Y are independent, there may be some nonlinear pattern between them. There are variety of metrics defined based on the Pearson correlation. In this work, we use $d_p(X,Y) = \sqrt{2(1 - r_p(X,Y))}$, which is a valid metric and satisfies the metric axioms [64].

2.3.2 Mutual information

Entropy measures the uncertainty of a random variable. The Shannons entropy for discrete distributions is defined by [93] as follows:

$$H(X) = H(p(X)) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -E \log p(X) \tag{2.3.2}$$

where $x \in \mathcal{X}$ is a random variable, with $p_1, , p_n$ the probabilities of occurrence of a set of events, and E is the expected value operator. If the base of log is set to 2, entropy is measured in bits. If the base is e, then it is expressed in nats. For continuous distributions, the entropy is defined as:

$$H(X) = -\int_{x \in \mathcal{X}} p_x(x) \log p_x(x) dx \tag{2.3.3}$$

The mutual information $I(X, Y)$ is a measure of shared information between two discrete random variables X and Y with joint distribution $p(x, y)$, $x \in \mathcal{X}, y \in \mathcal{Y}$ can be defined as:

$$I(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{2.3.4}$$

where $p(x, y)$ is the joint probability mass function of X and Y and $p(x)$ and $p(y)$ are marginal probability mass functions [93]. For continuous distributions, the mutual information is defined as:

$$I(X, Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \, dx \, dy \tag{2.3.5}$$

where $p(x, y)$ is the joint probability density function of X and Y and $p(x)$ and $p(y)$ are their marginal density functions, respectively [23]. The mutual information

is the relative entropy between the joint distribution of X and Y and the product of their marginal distributions. Mutual information can also be equivalently defined as follows:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2.3.6}$$

where H(X,Y) is the joint entropy and can be computed from the joint distribution of X and Y.

The mutual information can take only non-negative values and equals to zero if and only if X and Y are statistically independent and $I(X,X) = H(X)$. The high value of $I(X,Y)$ means a large reduction in uncertainty in X due to the knowledge of Y while low $I(X,Y)$ means that knowing something about Y reveals little knowledge about X. Since $I(X,Y) = I(Y,X)$, the same statements apply to knowledge about Y from observing X. The main advantage of mutual information is that it can capture linear and non-linear dependence relationships while the correlation measures the linear or monotonic relationships.

There exists two different versions of mutual information metrics defined in [59]. The first one is based on universal mutual information adjacency matrix referred to as version 1 or $AUV1$, which is defined as follows:

$$AUV1 = \frac{I(X,Y)}{H(X,Y)} \tag{2.3.7}$$

and the dissimilarity is defined as $dissAUV1 = 1 - AUV1$, which is a universal distance function. It is a valid metric and satisfies the triangle inequality [59, 95].

The second one is based universal mutual information adjacency matrix re-

ferred to as version 2 or $AUV2$, which is defined as follows:

$$AUV2 = \frac{I(X,Y)}{max\{H(X), H(Y)\}} \tag{2.3.8}$$

and similarly, the dissimilarity defined as $dissAUV2 = 1 - AUV2$ is also a universal metric [59, 95]. $dissAUV2$ is sharper than $dissAUV1$, that is, $dissAUV2 \leq dissAUV1$ [59].

The mutual information and entropy measures were originally expressed for discrete or categorical variables [23]. Although the properties of entropy of discrete and continuous distributions are similar, the entropy of continuous random variable has some drawbacks: It may take infinitely large value, or the integral may not exist. Moreover, it is generally not invariant under some monotonic transformations [23]. It is mentioned that continuous distributions, the differential entropy has some drawbacks as it may become negative in [23].

There are basically three different approaches to estimate $I(X,Y)$: histogram estimators, kernel-based estimators, and parametric methods [30]. Furthermore, the histogram methods can be divided into categories based on discretization methods. There are two widely used methods to discretize the continuous random variable: the first depends on equal-width discretization (uniform width partitions) and the second uses equal frequencies (uniform frequency partition). Discretizing data is also referred to as *binning* data–a very common approach to identify levels in a given vector of the categorical data analysis.

Equal-width is the simplest approach to divide data into subintervals and the main idea is to partition the interval $[x_{min}, x_{max}]$ into equal-widths. The interval can

be calculated as follows:

$$w = \frac{x_{max} - x_{min}}{k} \tag{2.3.9}$$

where k is the only parameter in this method. Thus, X is divided into k intervals and their boundaries can be computed from

$$a_i = x_{min} + iw, \quad i = 1, ..., k-1 \tag{2.3.10}$$

and it yields the partition $[x_{min}, a_1], (a_1, a_2], , (a_{k-1}, x_{max}]$ with k typically chosen as $\sqrt{n}$ where n is the sample size.

In equal frequency binning the range of X is divided into intervals of which each have approximately the same number of data points. Due to repeated values, some bins can have different number of points. The width of each interval can be different. These intervals can be calculated through empirical quantiles.

After random vector X is divided into subintervals, through either the equal-width or equal-frequency discretization methods, observed relative frequencies of each bin can be found simply by

$$p_i = \frac{O_i}{n} \tag{2.3.11}$$

where O_i is the observed number of elements in the ith interval or bin and n is the size of X and p_i denotes the frequency of the ith bin; thus, we obtain a vector of relative frequencies $p = (p_1, ...p_i)$.

These relative frequencies are used to compute the entropy and we simply plug in these probabilities in the formula below, which is called **the empirical estimator** (or also known as naive, plug-in or maximum likelihood estimator) of discretized

entropy of X; see [95].

$$\hat{H}^{emp}(X) = -\sum_{i=1}^{m} p_i \log p_i, \tag{2.3.12}$$

where m is the number of bins. The empirical entropy estimators are biased and have lower value than true entropy value of discritezed X. To overcome this problem, there are some alternative estimators proposed in the literature. The **Miller-Madow estimator** is defined as follows [68]:

$$\hat{H}^{MM}(X) = \hat{H}^{emp}(X) + \frac{m-1}{2n} \tag{2.3.13}$$

where m is default number of bins as $\sqrt{n}$. There are other widely used estimators such as **shrinkage estimator** [46], in which relative frequencies are estimated by combining two different estimator, one with low variance and one with low bias and **Schurmann-Grasberger estimator** [92], in which Dirichlet distribution is utilized to calculate the entropy of a discrete random variable .

In this work, in order to obtain mutual information adjacency matrix, we first discretized the stock returns through equal-width method with default number of bins as $\sqrt{n}$, and secondly, the mutual information between each pair of discretized stock returns is computed according to the Miller-Madow estimation technique. Thereby the mutual information matrix is transformed to one of the versions of adjacency matrices or distance matrices as defined above.

2.3.3 Relationship between Pearsons correlation coefficient and AUV2

Despite the differences between mutual information and correlation, such as mutual information depends on parameters choice and is a general measure of association not like correlation which only measures linear or monotonic relationships. In [95], they

give a simple approximation between two measures under some assumptions. First, samples come from a bivariate normal distribution and the equal-width discretization method with $\sqrt{n}$ chosen as the number of bins. Under these assumptions, they show that $AUV2$ can be accurately predicted by the Pearson correlation as input in the following approximation:

$$\begin{aligned} AUV2(x,y) &= \frac{I(x,y)}{max(H(x),H(y))} \\ &\approx F^{cor-MI}(cor(x,y)) \end{aligned} \tag{2.3.14}$$

where the F^{cor-MI} function is given as follows:

$$F^{cor-MI}(s) = \frac{\log(1+\epsilon-s^2)}{\log(\epsilon)}(1-w)+w \tag{2.3.15}$$

w and ϵ given as:

$$\begin{aligned} w &= 0.43n^{-0.30} \\ \epsilon &= w^{2.2} \end{aligned} \tag{2.3.16}$$

This approximation is based on if x and y are samples from a bivariate normal distribution; therefore, the excess information in the two estimates can be defined as follows:

$$AUV2_e = AUV2 - F^{cor-MI}(s) \tag{2.3.17}$$

We can use this relationship between $AUV2$ and F^{cor-MI} to define the non-Gaussian information or extranormal information and utilize it to find out cases where

the Pearson and mutual information measures strongly disagree.

2.4 Elements of Complex Network Theory

Formally, a graph or network $G = (V, E)$ is a mathematical structure composed of a finite nonempty set V of vertices or nodes, and a set $E \subseteq V \times V$ of edges or links composed of unordered pairs of vertices. The number of nodes $|V| = n$ is called the order of G and the number of edges $|E| = m$ is called size of G.

The binary adjacency matrix $\mathbf{A}$ of G is the matrix with elements A_{ij} defined as:

$$\mathbf{A}(i,j) = \begin{cases} 1 & \text{if v}_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

If the graph is weighted, then $n \times n$ weighted adjacency matrix $\mathbf{A}$ is defined as follows:

$$\mathbf{A}(i,j) = \begin{cases} w_{ij} & \text{if v}_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

where w_{ij} is the weight on edge $(v_i, v_j) \in E$.

In order to compare the different perspective of stock networks topology on local, mesoscopic, and global scales, we choose variety of measures which are well-defined in the literature [72, 12, 42]. For example, degree centrality uses only local information on the connection of a node v; therefore it is a local measure. On the other hand, the closeness, betweenness, and eccentric centralities are defined based on shortest paths between pairs of nodes in the network; thereby they are considered

as global measures [31].

2.4.1 Local measures

Degree centrality

The connectivity (also known as degree) of the node v_i is defined by

$$d_i = \sum_j \mathbf{A}(i,j) \tag{2.4.1}$$

If the network G is unweighted, d_i equals the number of nodes that are directly linked to the i^{th} node. In weighted networks, the connectivity equals the sum of connection weights or strengths between node i and the other nodes incident with it. While degree is a local attribute, average degree is global attribute and defined as

$$\mu_d = \frac{\sum_i d_i}{n} \tag{2.4.2}$$

2.4.2 Global measures

Hamming distance

The Hamming distance $d_H(A,B)$ of two labeled simple networks, with adjacency matrices $A(i,j)$ and $B(i,j)$ with the same dimensions, computes the fraction of edges that have to be inserted or deleted to transform one graph into the other.

$d_H(A,B)$ is defined as follows:

$$d_H(A,B) = \Big\langle XOR(A(i,j),B(i,j)) \Big\rangle_{ij} \tag{2.4.3}$$

where

$$\mathrm{XOR}(A(i,j),B(i,j)) = \begin{cases} 1 & \text{if A(i,j)}\neq\text{B(i,j)} \\ 0 & \text{otherwise} \end{cases}$$

The range of Hamming distance takes values over the interval $[0,1]$ and computes the global probability of non-equal entries in the two adjacency matrices.

Average path length

Let d_{ij} denote the geodesic distance between the vertices $v_i, v_j \in V$, which is the minimum number of edges that have to be crossed to travel from vertex v_i to vertex v_j. The average path length L of a graph is the average geodesic distance between all connected pairs of vertices and defined as:

$$L = \frac{1}{\binom{n}{2}} \sum_{i<j} d_{ij} \tag{2.4.4}$$

Closeness centrality

The closeness centrality measures the reciprocal average topological distance of vertex v_i to all other vertices in the network [42]. This measure is normalized by multiplication factor $N-1$ to $0 \leq CC_{v_i} \leq 1$, and is formulated as

$$CC_{v_i} = \frac{N-1}{\sum_j d(v_i, v_j)} \tag{2.4.5}$$

A node v_i with the smallest total distance is called the *median node*.

Betweenness centrality

Betweenness centrality measures a vertex centrality in terms of its location between other pairs of vertices in graph. It is most commonly calculated as the number of shortest paths from all vertices to all others that pass through that node v. BC_v is defined as

$$BC_v = \sum_{i,j \neq v}^{N} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \tag{2.4.6}$$

where σ_{ij} is the total number of shortest paths from i to j which includes node v [42].

Eccentric centrality

The eccentricity of a node v_i is defined as the maximum distance from v_i to any other node in the network, i.e.

$$e(v_i) = \max_j \left\{ d(v_i, v_j \right\} \tag{2.4.7}$$

Eccentricity centrality is thus expressed as:

$\mathrm{c}(\mathrm{v}_i) = \frac{1}{e(v_i)} = \frac{1}{\max_j \left\{ d(v_i, v_j) \right\}}$

A node v_i with the smallest eccentricity is called a center node, while the one with the greatest eccentricity is called a periphery node [105].

2.4.3 Network entropy

Shannon entropy [93] measures the randomness of the network when we apply it to networks degree distribution. The higher the value of entropy, the more random is the network. Let $H(G)$ be the expected number of nats in a network $G = (V, E)$ and

$p(k)$ the probability that randomly selected node has exactly k edges, i.e., $p(k) = \frac{k}{N}$ where N is the total number of nodes.

$$H(G) = -\sum_{k}^{n} p(k)\log p(k) \tag{2.4.8}$$

$H(G)$ is used to calculate entropy of the degree distribution $p(k)$. Degree distribution informs us about the shape of a network, while entropy measures the regularity in the shape of a network [61]. The maximum value of graph entropy is $H_{max}(G) = \log N$ for $p(i) = 1/N \ \forall i = 1, ..., N$ and the minimum value of graph entropy is $H_{min}(G) = 0$. Therefore, if $H(G)$ is the entropy of the network, then the normalized entropy of a given network is defined as:

$$H(G)_{norm} = \frac{H(G)]}{\log N} \tag{2.4.9}$$

2.4.4 Scale-Free networks

One of the most important properties of a network is its degree distribution. In many real-world networks, it has been found that the empirical degree distribution $f(k)$ follows a *power-law*:

$$f(k) = Ck^{-\alpha} \tag{2.4.10}$$

where C and α denote positive real numbers. The networks whose degree distribution follows power-law degree distributions are said to present scale-free topology [7] with scaling parameter α that is also called as the exponent of the power-law. Let

us take the logarithm of the both sides of the relation (2.4.10):

$$\log f(k) = \log\left(Ck^{-\alpha}\right) \tag{2.4.11}$$

which yields:

$$\log f(k) = -\alpha \log k + \log C \tag{2.4.12}$$

Thus scale-free topology implies a straight line relationship in the $log - log$ plot of k versus $f(k)$, with $-\alpha$ giving the slope of the line. The standard way to check if a network has scale-free behavior is to apply a least-square fit of the points $(\log k, \log f(k))$ to a line.

To measure the straight line relationship between $(\log k, \log f(k))$, we apply R^2 statistics that takes values between 0 and 1 and explains the proportion of variance explained by the least-square regression line. We call this value as the scale-free topology fitting index as in [107].

A power law indicates that most nodes have very small degrees, whereas there are a few nodes (hubs) that have substantially higher degrees, i.e, they connect with these nodes of very small degrees and connect them to the system. Therefore, scale-free networks are substantially heterogeneous and their topology is controlled by these hubs [107].

2.5 Community Detection and Mesoscopic Structure

Networks representing the real systems are not regular or random at the global level and have some level of order and organization at locally. Thus, some nodes are highly interconnected and few connections with the rest of the network. The group

of nodes with highly connected to each other or same type is called modules or communities [76]. Many community detection algorithms have been proposed to reveal these mesoscopic properties of complex networks and are mostly based on traditional clustering algorithms [40].

The modularity measure was introduced by Newman and Girvan [76] to compute the quality of a network partition. It calculates the density of links inside communities and compares it to links between communities. It is defined for weighted networks as follows:

$$Q = \frac{1}{2m} \sum_{ij} \Big[A_{ij} - \frac{k_i k_j}{2m} \Big] \delta(c_i, c_j) \tag{2.5.1}$$

where A_{ij} is the weighted adjacency matrix of the edges between i and j, k_i is the sum of connection weights between node i and other nodes and $2m = \sum_{ij} A_{ij}$ [11]. The $\delta(c_i, c_j)$ is the Kronecker delta and $\delta(c_i, c_j) = 1$ when $i = j$ and 0 otherwise.

The modularity measures the difference between the observed and expected number of links in each partition and summed over all partitions. It takes values between -1 and strictly less than 1 and can be approached to its upper limit in case the communities have been perfectly identified. If the links represent the distance between each node in the network, the smaller value of the modularity measure means better clustering.

The modularity measure is used by many community detection algorithms to assess the quality of partitions. It is considered as the most credible objective function for network partition. Furthermore, its normalized version is also used to calculate whether the network is assortative or not–known as assortative mixing coefficient [73].

In this work, we choose edge betweenness community detection method, which is available in igraph [24]. This algorithm introduced by [76] is a hierarchical de-

composition process in which edges are removed in the decreasing order of their edge betweenness scores. The edges connecting different communities are expected to have high edge betweenness; therefore, after removal of these links, communities in the network can be found. This algorithms time complexity is (m^2n).

2.6 Cluster Validity Indices

One of the major problems with clustering algorithms is that the algorithm will identify clusters even if the data does not have any inherent clusters. Because of this, cluster validation indices are developed to quantify the result of clusters. There are several validity indices and statistics that have been proposed to assess the quality of clusters. For this work, we have used the internal indices and external indices.

Internal indices assume that there are no pre-assigned labels for points in the data and they are based on the nxn distance matrix (or proximity matrix) of all pairwise distances among the n points and denoted by $\mathbf{W}$. Since the goal of the clustering algorithms is to partition the data into groups of objects by optimizing the average distance within cluster to be as low as possible and average distance between clusters to be as large as possible, they measure the compactness, that is how close the objects inside the same cluster and separation, that is, how far the clusters from each other. Therefore, a clustering algorithm is an optimization problem of the minimization-maximization type (a min-max problem).

Internal indices are often applied to find out the number of clusters and choose the proper clustering algorithm in case no external labels are available in the data set. They are originally defined for the data points of vectors of attributes and require distance measure between objects. We have used the geodesic distance between any two nodes on the network to apply the internal indices which are defined for vector

clustering, in which the distance can be calculated between any two vectors. The centroid of nodes is chosen as the node that minimizes the largest distance from any other nodes in the cluster.

2.6.1 Internal indices

We define the validation statistics used in this work and explain how we apply them. The internal measures are relied on the distance matrix of all pairwise distances among the n nodes in the graph.

C-index

Let S_{in} be the sum of all the intracluster distances. The C-index [50] is defined as:

$$C = \frac{S_{in} - S_{min}}{S_{max} - S_{min}} \tag{2.6.1}$$

where S_{min} is the sum of the N_{in} smallest distances and S_{max} is the sum of the N_{in} largest distances in $\mathbf{W}$. The C-index lies in the rage $[0, 1]$. The smaller the C-index, the better clustering.

Dunn index

It is defined as

$$Dunn1 = \frac{W_{out}^{min}}{W_{in}^{max}} \tag{2.6.2}$$

where $W_{out}^{min} = \min_{i,j>i}\{w_{ab}|x_a \in C_i, x_b \in C_j\}$ (i.e, the minimum intercluster) and $W_{in}^{max} = \max_i\{w_{ab}|x_a, x_b \in C_i\}$ (i.e,the maximum intracluster distance), which is called the diameter of the cluster [32].

In this work, we also consider the another version of Dunn index (Dunn2), which is simply defined as the ratio between minimum average dissimilarity between two clusters and maximum average within cluster dissimilarity.

Silhoutte coefficient

For each node $x_i \in V$ in a graph $G(V, E)$, the Silhoutte coefficient of x_i is defined as:

$$s_i = \frac{\mu_{inter} - \mu_{intra}}{\max\{\mu_{inter}, \mu_{intra}\}} \tag{2.6.3}$$

where μ_{inter} is the mean distance from x_i and every other node that is not in the same cluster as x_i and μ_{intra} is the mean of the distances from x_i to points in its own cluster set.

The range of s_i is between -1 and 1. A close value to 1 indicates that x_i is assigned to best possible cluster whereas -1 indicates that x_i is much closer to another cluster than its own cluster meaning that the point may be assigned in the wrong cluster. A value close to zero indicates that x_i is near the boundary between two clusters.

The silhouette coefficient is defined as the mean s_i value across all the points:

$$SC = \frac{1}{n} \sum_{i=1}^{n} s_i \tag{2.6.4}$$

A value close to +1 indicates a good clustering. The Silhouette coefficient is a measure of both cohesion and separation of clusters [91].

BetaCV measure

It is simply the ratio of average distance within cluster to the average distance between clusters and formulated as follows:

$$BetaCV = \frac{S_{in}/N_{in}}{S_{out}/N_{out}} \tag{2.6.5}$$

A small value of BetaCV indicates a better clustering [106].

2.6.2 External indices

External indices assume that ground-truth class labels for each point is known a priori and aims at comparing the identified clusters to them. They can be used to choose the proper clustering algorithm for the data because the number of clusters and labels are known in advance.

Let $C = C_1, C_2, ..., C_r$ denote the clustering of the dataset into r clusters and let $T = T_1, T_2, , T_k$ denote the ground-truth cluster membership for the same dataset. Further, let $n_i = |C_i|$ be the number of points in cluster C_i and $m_j = |T_j|$ the number of points in the ground-truth partition T_j.

The external indices are based on the $m \times k$ contingency table (matrix) $N = (n_{ij}), i = 1, ..., m; j = 1, ..., k$ and is defined as follows:

$$n_{ij} = |C_i \cap T_j| \tag{2.6.6}$$

where n_{ij} represents the number of shared points between C_i and T_j.

There are several external measures that have been proposed. In this work, we

use the normalized mutual information (NMI) [25] and it is defined as follows:

$$NMI(C,T) = \frac{2I(C,T)}{H(C)+H(T)} \tag{2.6.7}$$

where

$$H(C) = -\sum_{i=1}^{r} p_{C_i} \log p_{C_i} \tag{2.6.8}$$

$$H(T) = -\sum_{j=1}^{k} p_{T_j} \log p_{T_j} \tag{2.6.9}$$

and $p_{C_i} = \frac{n_i}{n}$ is the probability that a point in cluster C_i and $p_{T_j} = \frac{m_i}{n}$ is the probability that a point in partition T_j, and

$$I(C,T) = \sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log \frac{p_{ij}}{p_{C_i} p_{T_j}} \tag{2.6.10}$$

where $p_{ij} = \frac{n_{ij}}{n}$ is the probability that a point in both C_i and T_j. I(C,T) is the mutual information between C_i and T_j. The NMI equals 1 if the clustering is identical to partitions and equals 0 if they are independent.

2.7 Results

After having introduced a collection of measures for stock network construction, we turn to the methodology in the sense of comparing the Pearson correlation and the mutual information measures to quantify the impact of nonlinearity in several aspects in a crisis period and non-crisis period.

As a first step, we look for the cases where the Pearson correlation and the mutual information measures are strongly disagreed. Then, we investigate the reasons

for the mismatch stock pairs by utilizing the predictive models and investigating the relationship between entropy and the first four moments of stock distributions.

As a final step, we compare the properties of the complex networks generated using the Pearson correlation and the mutual information on local, mesoscopic and global topological scales in order to measure the impact of the nonlinearity. For a quantitative comparison, we compute the Spearman rank correlation coefficient of the interested network property of mutual information and the Pearson correlation networks, such as a degree.

Empirical results for non-crisis period

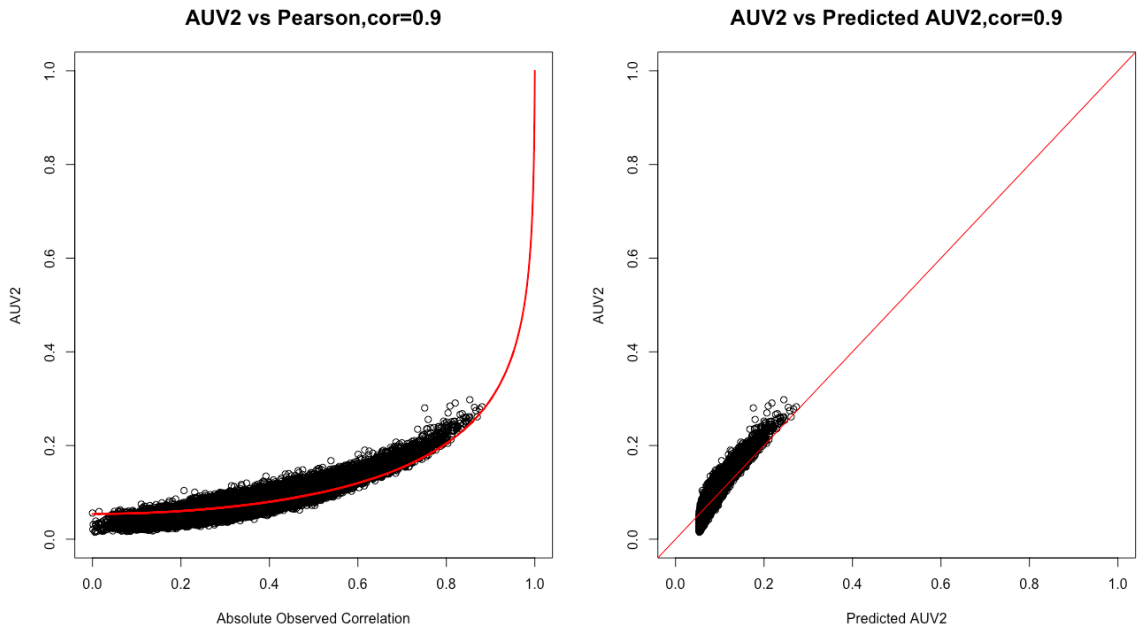

Figure 2.1: Comparison of correlation and mutual information estimates for non-crisis period with the Spearman's correlation shown at the top.

We first calculate the Pearson correlation and AUV2 for all stock pairs in each portfolio. After that, we predict AUV2 from the Pearson based on F^{cor-MI}. In the Figure 2.1, given approximation function predicts $AUV2$ highly accurately. Moreover,

inspecting the scatter plot between the Pearson and $AUV2$, both measures show a strong monotonic relationship (Spearman correlation is 0.9). These results indicate that most of the stock pairs satisfy linear relationships for considering the time period.

Even though F^{cor-MI} uncovers a close relationship between the Pearson and $AUV2$, there are some stock pairs where the two measures disagree. In order to find out these stock pairs where two measures are strongly mismatched, we compute the extranormal information $AUV2_e$ for each pair of stocks. This extranormal information can reveal the stock pairs where $AUV2$ has higher value while F^{cor-MI} under(over)estimates; therefore we have two categories of stock pairs with positive and negative extranormal information.

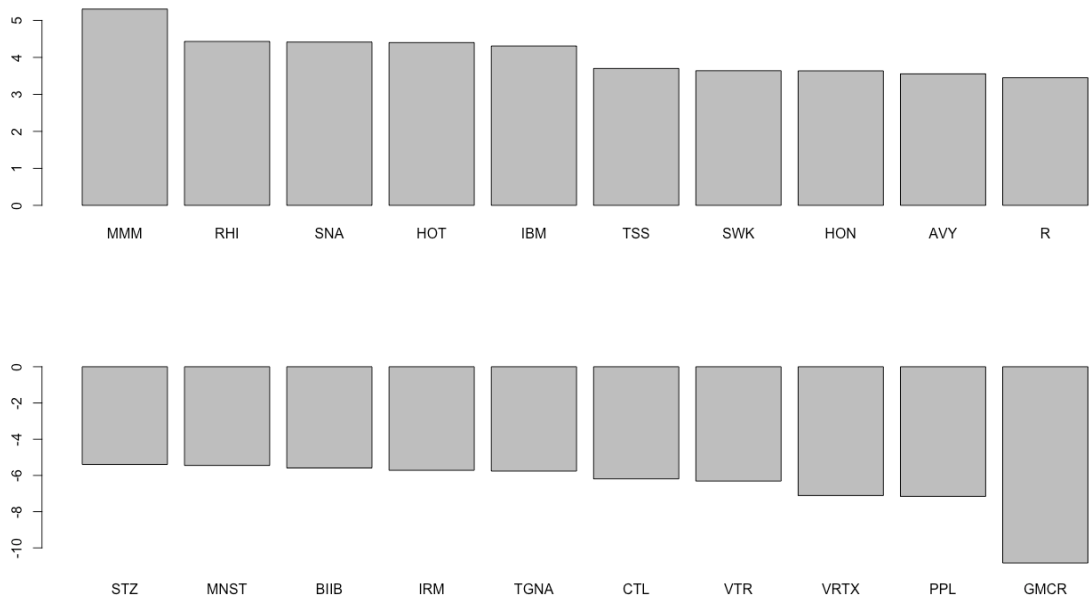

Figure 2.2: Extranormal information estimates for some stocks

In Figure 2.2, the first row shows the stocks with ten highest positive extranormal information and the bottom row shows the stocks with the lowest negative extranormal information. Extranormal information estimates for all stocks are com-

puted and bars show the sums of the rows (or columns) of extranormal information matrix for each stock.

In order to show that simple linear regression is not enough to capture the relationship for the first category of stocks, we run polynomial regression up to degree 5. As an illustration, we selected the MMM and its relation with the other three stocks where it has the positive deviation from the approximation. In the next Figure 2.3, we see that BIC and C_p values confirm that simple linear regression cannot depict the relationship between MMM and selected stocks. Furthermore, our results are confirmed by ANOVA tables.

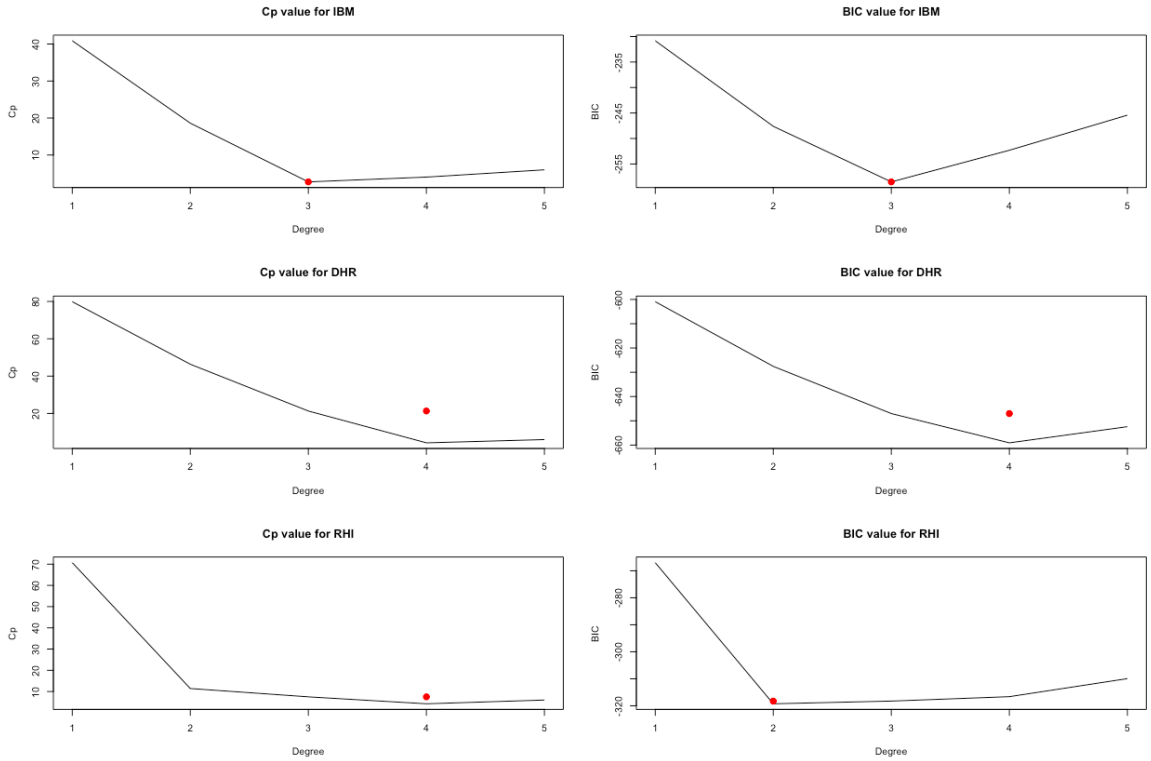

Figure 2.3: BIC and Cp values for MMM

The next question concerns with the stocks which have negative extranormal information. In this case, we have stock pairs whose correlation is higher than $AUV2$. We observe that $AUV2$ has found insignificant correlations while the Pearson could

not find it. In our bar chart, we see the stock GMCR has the lowest negative value. When we investigate its relationship, we find that it has a maximum correlation of 0.2367611, which is not very strong.

Empirical results for crisis period

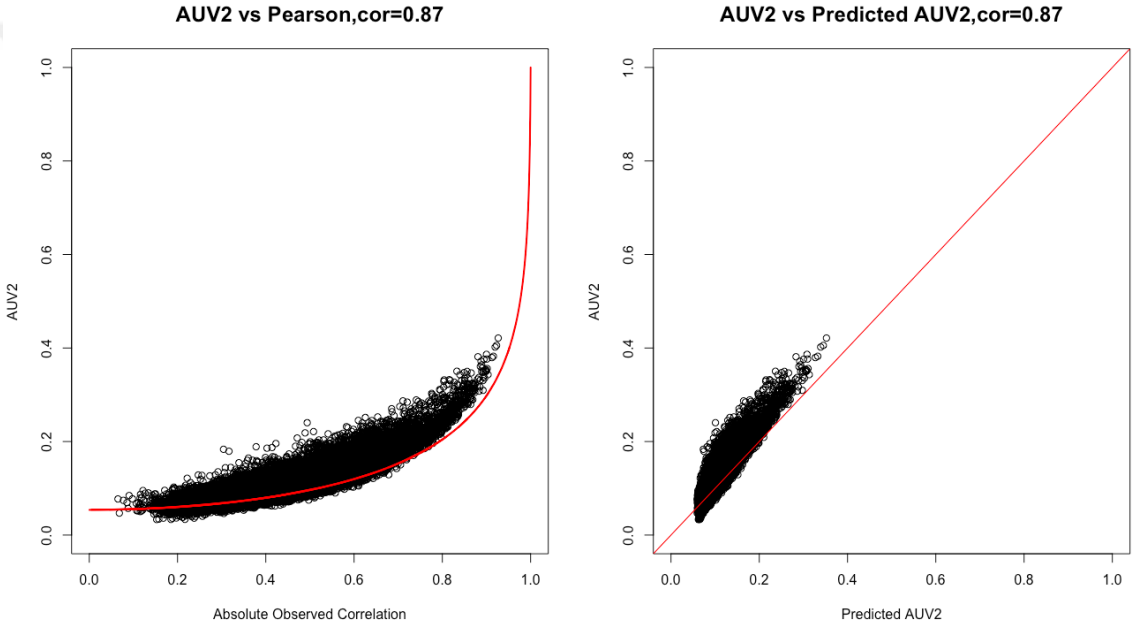

Figure 2.4: Comparison of correlation and mutual information estimates for crisis period with the Spearman's correlation shown at the top.

We conduct a similar approach on the data set sampled from the crisis period for the same set of stocks in order to capture nonlinearity or extra-normal information between all pairs of stocks. In Figure 2.4, we observe that the Spearman's correlation has slightly dropped from 0.9 to 0.87. Given the approximation function still predicts $AUV2$ highly accurately and strong monotonic relationship between two association measures exists, i.e. the majority of pairwise dependency between stock pairs are linear. Using the extranormal information, we are interested in identifying the stock pairs where two measures disagree.

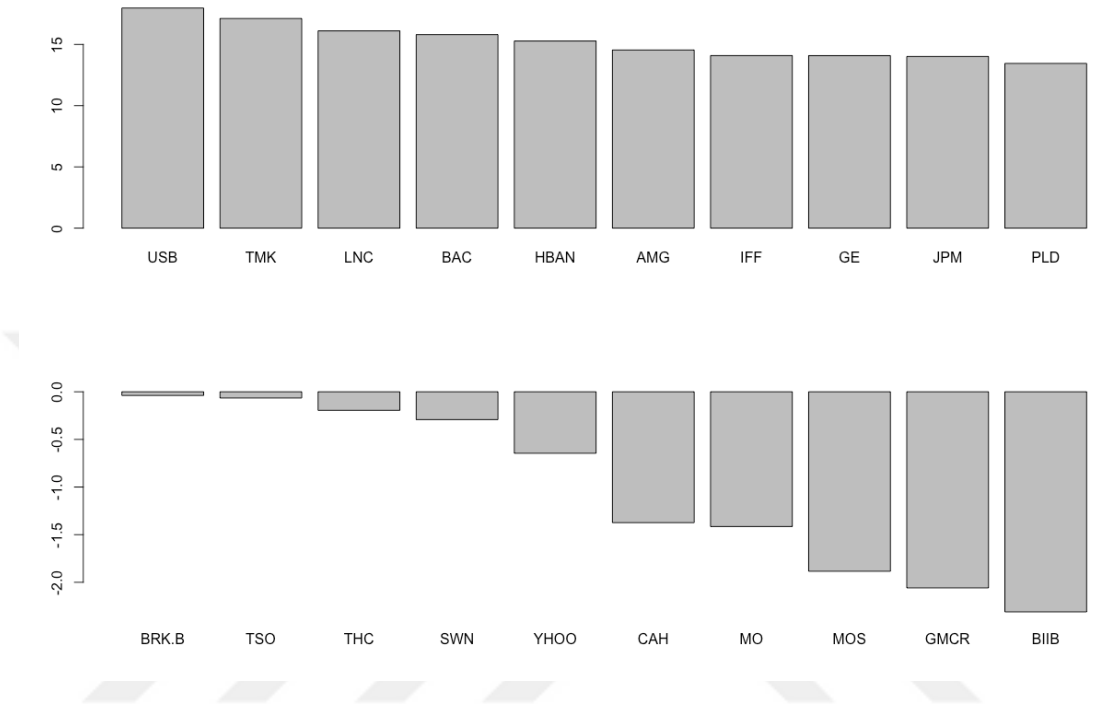

Figure 2.5: Extranormal information estimates of 10 highest and 10 lowest for the crisis period

Comparing Figure 2.5 with Figure 2.2, we observe that much stronger deviations from Gaussianity, for example, while in non-crisis period MMM has the maximum extranormal information value nearly 5, UBS has approximately 18 in crisis-period. Moreover, we find that 240 stocks have extranormal information greater 5 in the crisis period, which is more than half of the stocks in the portfolio.

Similar to the previous section, for the stock USB, we run polynomial regression up to degree 5 and its relation with the other three stocks where it has the positive deviation from the approximation. In Figure 2.6, we see that BIC and C_p values confirm that simple linear regression cannot depict the relationship between USB and selected stocks. Furthermore, our results are confirmed by ANOVA tables.

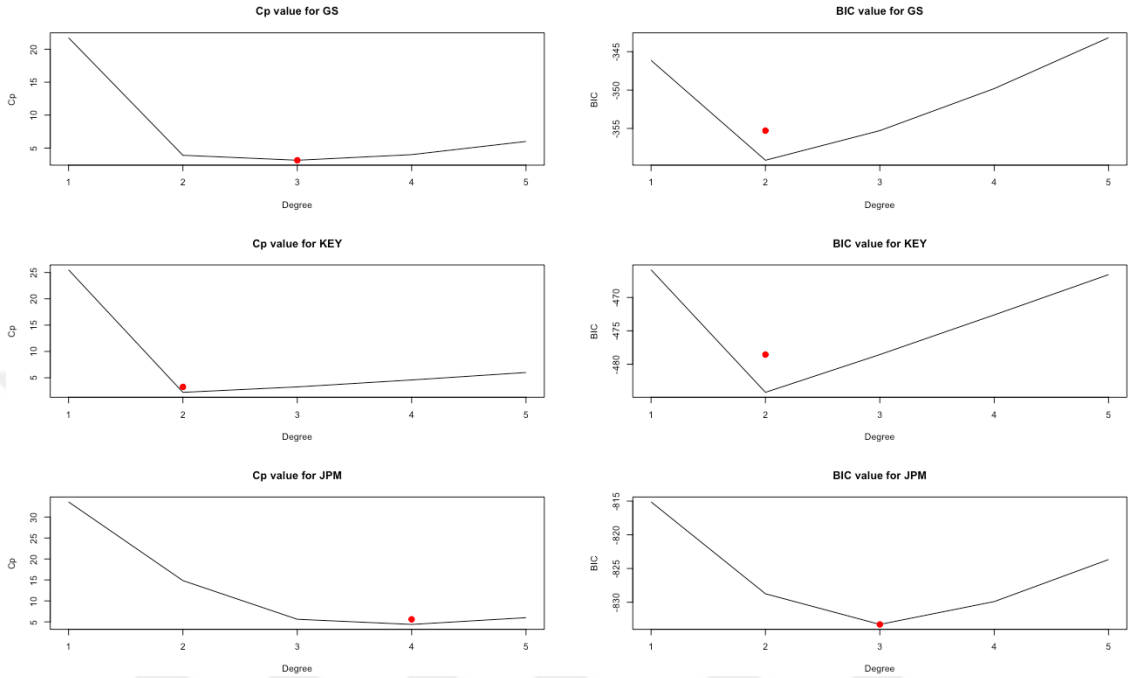

Figure 2.6: BIC and Cp values for USB

2.7.1 Polynomial and spline regression models

The classic polynomial and spline regression models can also capture the nonlinear correlation between variables. Comparing to the MI, their estimations are simpler and faster; moreover, the classic statistical tests are available to check if the model fits well or not [95]. Next, we define model fitting index measures based on the polynomial regression and spline regression models.

Polynomial regression is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_d x_i^d + \epsilon_i \tag{2.7.1}$$

where ϵ_i is the error term. The coefficients $\hat{\beta}$ can be easily estimated by using the least squares linear regression. Given $\hat{\beta}$, we can compute the fitting index R^2 as

follows [95]:

$$R_2 = cor^2(y, \hat{y}) \tag{2.7.2}$$

R^2 statistics takes values between 0 and 1 and explains the proportion of variance explained by the model. We use cubic polynomial regression to quantify the nonlinear relationship in the data. However, the matrix R^2 produced is not symmetric. There are various ways to symmetrize the non-symmetric matrices. In this work, the method we use as follows:

$$S_{ij}^{ave} = \frac{S_{ij} + S_{ji}}{2} \tag{2.7.3}$$

where S is the matrix R^2. In order to calculate the dissimilarities based on R^2 statistics, we use the following:

$d(X, Y) = 1 - R^2$. We call it as $dissPolyReg$ for short in the followings

Spline Regression requires dividing the range of X into K distinct regions and fit a separate polynomial function for each one. Boundaries of each subinterval, where the coefficients are changed, is called knots. We have used 5 knots by following the rule of thumb, which says if sample size more than 100 use 5 knots, if it is less than 30 use 3 knots, otherwise use 4 knots [95].

A spline of degree D with K knots is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_D x_i^D + \sum_{k=1}^{K} b_k (x_i - \xi_k)^D + \epsilon_i \tag{2.7.4}$$

where the function

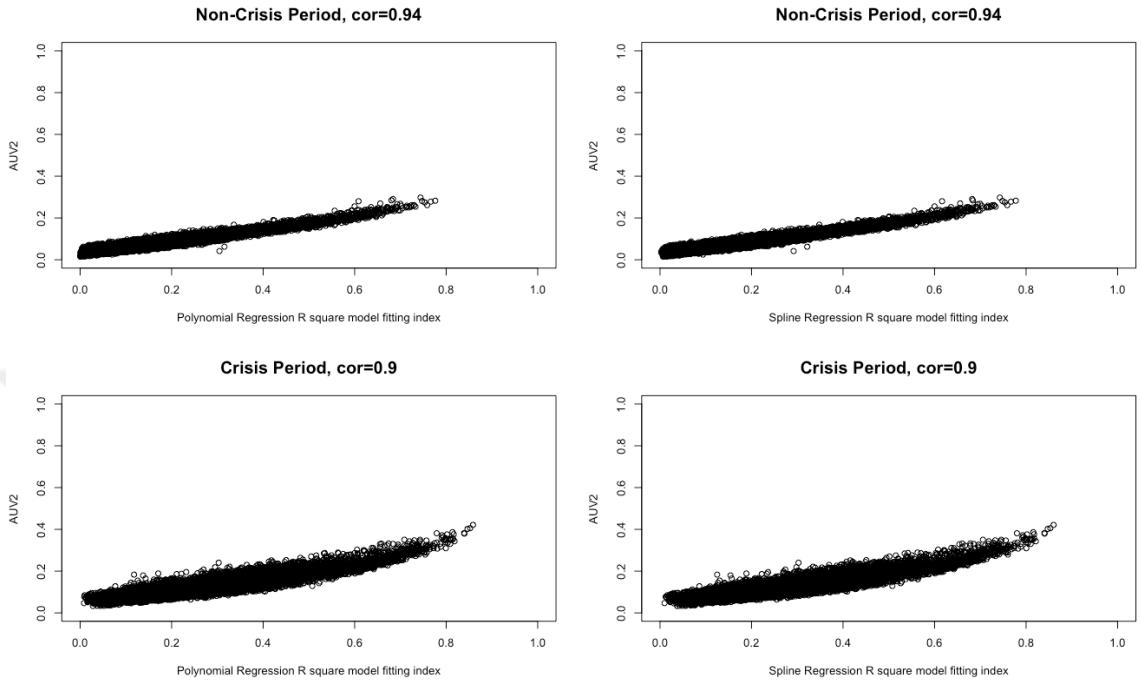

Figure 2.7: Comparisons of regression models and mutual information with Pearson's correlation between measures are shown on the top.

$$\text{(x-}\xi)_+^D = \begin{cases} (x-\xi)^D & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases}$$

is called a truncated power basis function of degree D. One of the issue with splines is they can have high variance at the outside of the range the predictors, that is, X is smaller than the smallest knot, or larger than the largest knot. A natural cubic spline requires the function to be linear at the boundary, so it can generate more stable estimates. We use the natural cubic splines. Like polynomial regression, R^2 can be used as the similarity and dissimilarity measure between x and y after symmetrized. We call the dissimilarity based on the spline regression as *dissSpline* for short in the followings. Spline regression is more flexible than polynomial and

stepwise regressions and asserts the smoothness of each observation [52].

In our empirical data sets for crisis and non-crisis periods, we observe in Figure 2.7 that regression models and mutual information ($AUV2$) have high correlation and they both have a stronger relationship than between Pearson's correlation and $AUV2$. This result also supports that $AUV2$ and regression models reveal some stock pairs' non-linear relations, whereas Pearson's correlation cannot discover them.

Furthermore, we find that both regression models have Pearson's correlation of 0.99, i.e they are identical in both time periods. We also note that different symmetrization methods (max and min) are also used to compute R^2 and we find that they all have Pearson's correlation almost 1; therefore, it is safe to use any of these techniques.

2.7.2 Relationship between entropy and the first four moments of the daily stock returns

The entropy as a measure of uncertainty in finance is firstly introduced by [85], in which they present the mean-entropy efficient portfolios and compare it against the mean-variance approach. Furthermore, they conclude that entropy is more general and better fitted for the selection of portfolios than the variance.

If the stock return distributions are not a normal distribution, additional moments or another measure of uncertainty is needed because variance cannot work well in specific situations, such as the existence of non-symmetry and fat-tails or extreme events in probability distributions [96, 63, 85]. Furthermore, some authors [29, 33] conclude that entropy is a more general uncertainty measure than the variance since it uses more information about the probability distribution.

Now let us investigate the relationship between entropy and first four moments of daily log returns for all studied companies: mean, standard deviation, skewness,

and excess kurtosis and results are presented as a scatterplot In Figure 2.8 for crisis and non-crisis period.

and the first three moments of the daily log returns are weaker than the Spearman correlation coefficient between the kurtosis and entropy rates. Kurtosis measures the degree of a distribution expressed as fat tails. For a normal distribution, an excess kurtosis has to be equal to 0 and no skewness (symmetric). A distribution with positive excess kurtosis is called leptokurtic. While the observed maximum and minimum excess kurtosis values are 0.464 (for OMC) and 105.374 (for ARG) in the non-crisis period, respectively, we observe 1.641 (for CSX) as minimum and 84.704 (for STT) as the maximum in the crisis period. This results indicate the strong deviations from normality and confirms the leptokurtic behavior of stock return distributions.

We further note that the average Shannon's entropy rates for all stocks are equal to 2.346 for the non-crisis period and 2.184 for the crisis period. This allows us to see how different the market efficiency changes in two periods. These results indicate that the market becomes less stable and less efficient (more predictable) during the crisis.

In summary, differences between mutual information and Pearson product-moment correlation coefficient arise for the following reasons: First, as the name indicates Pearson correlation uses only first two moments mean and variance. However, mutual information uses entropy which is highly correlated to a higher order of moments. For example, we have found a striking relationship between entropy and kurtosis. Secondly, Pearson's correlation is limited to identify linear or monotonic relationship while mutual information can measure general dependence beyond the linearity. Third due to the size of a finite sample there may be estimation errors. In short, nonlinearity exists between stock returns and simple linear regression cannot be enough to fully describe the relationships and therefore may underestimate in some

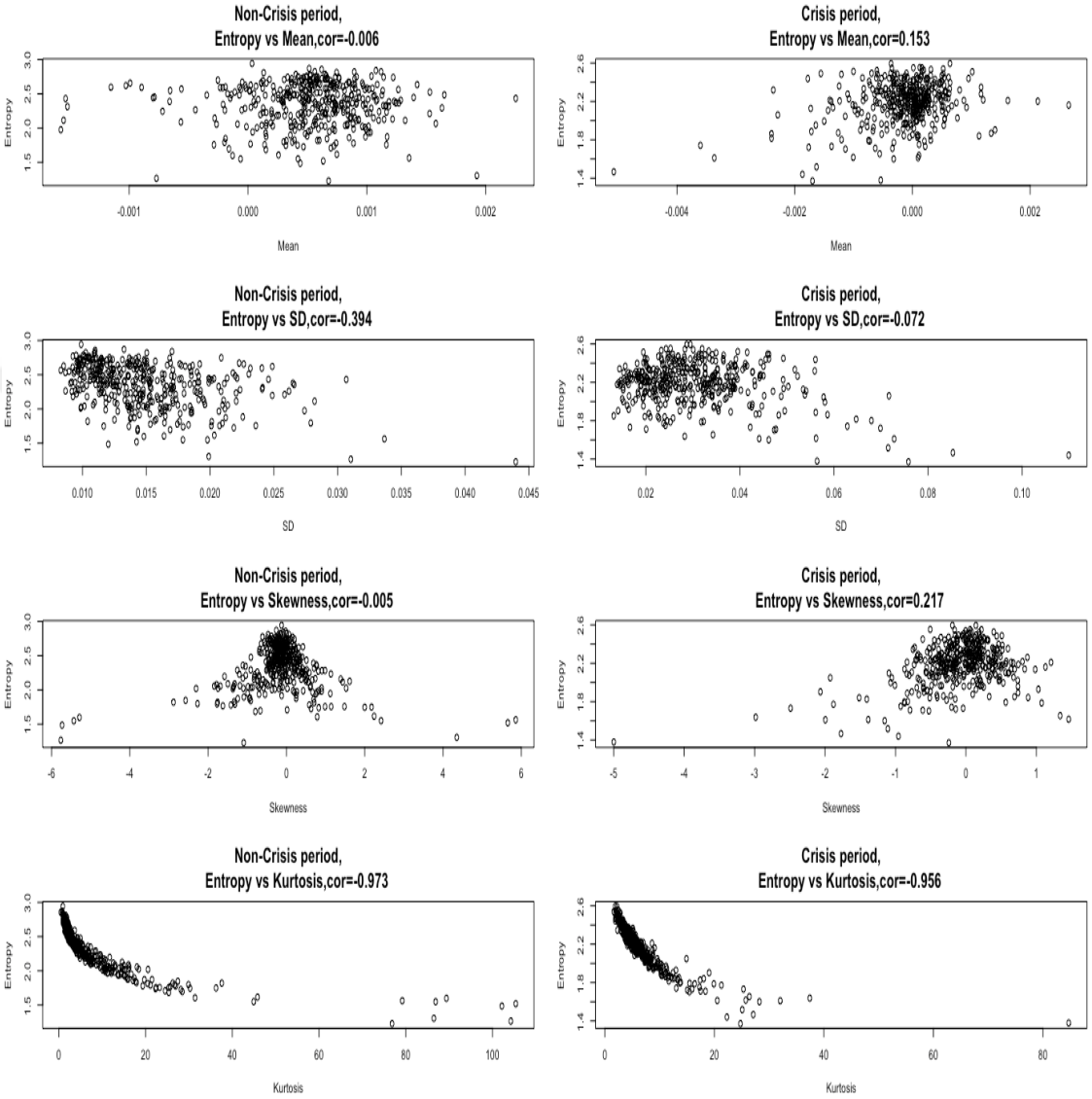

Figure 2.8: Relationship between entropy and the first four moments of daily stock returns for non-crisis and crisis periods. Spearman's correlation between measures are shown on the top.

cases. Nonlinearity may exist due to high skewness or kurtosis of stock returns.

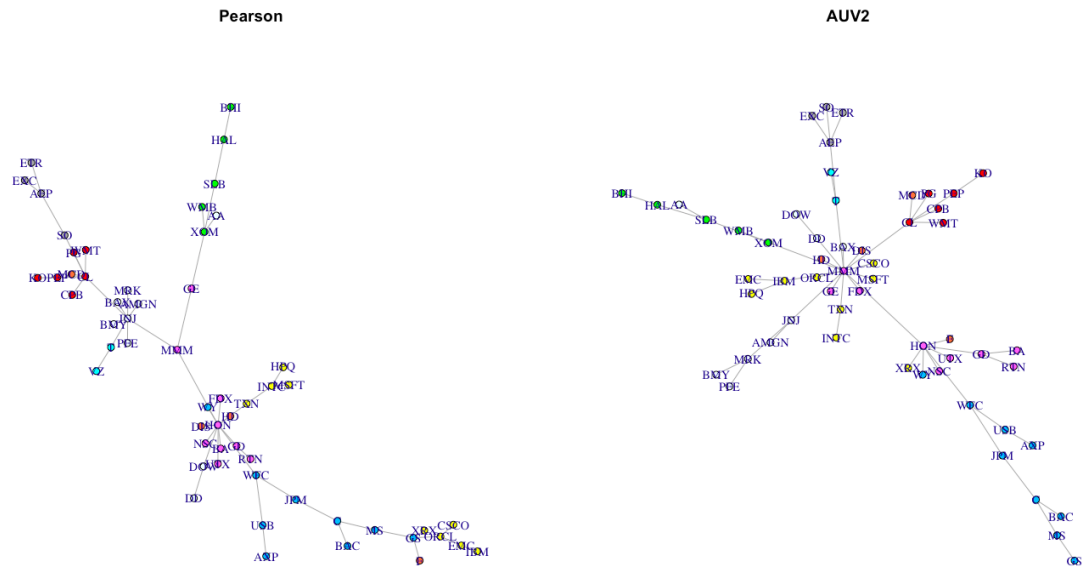

Figure 2.9: Example of networks constructed by Pearson and AUV2 for 57 stocks in the time period between 2011 and 2015.

	Pearson (Non-Crisis)	NMI (Non-Crisis)	Pearson(Crisis)	NMI(Crisis)
Hub Degree	23	20	20	14
Diameter	33	31	28	26
Avg Path Length	11.42	10.69	11.24	12.31
Assortativity By Degree	-0.18	-0.11	-0.18	-0.20
Assortativity By Sector	0.81	0.84	0.77	0.73
Assortativity By Industry.Group	0.74	0.75	0.65	0.65
Avg.Closeness	0.09	0.10	0.09	0.08
Avg.Betweenness	0.03	0.02	0.03	0.03
Entropy	1.31	1.33	1.32	1.36
SF-Fitting Index	0.94	0.95	0.90	0.95
Exponent of SF	2.01	1.96	2.06	2.38

Table 2.1: Comparison of Some Properties of the Networks (n=402, m=401)

2.8 Comparing Network Properties

2.8.1 Local and global comparison

Now we are interested in differences between networks created by these two measures in terms of local, global and mesoscopic levels. Our analysis is similar to the previ-

ous section, such as we consider two periods separately and conduct the comparison between network properties. First, let us consider each network unweighted and look at some fundamental global properties of the networks. In Table 2.1, we summarized the results for the Pearson and mutual information based networks.

The Pearson based networks have higher hub degrees and diameters than mutual information based one for the normal and non-normal period. Interestingly, mutual information based network has higher average path length in the crisis period while Pearson has a higher one in the normal period.

The network is considered as assortative if a significant fraction of links in the networks stays between vertices of the same type. We consider here two types of assortativity. Assortativity degree values show that all networks are disassortative mixing by degree, indicating that high degree nodes tend to be linked to low-degree ones. Stocks in the GICS classification system are categorized into four hierarchical categories: sector, an industry group, industry, and sub-industry. In assortativity nominal, we measure the tendency of stocks for sector and industry group. It is clear that each network is assortative for both cases.

While mutual information has a higher assortative mixing coefficient for the sector in the non-crisis period, Pearson has a higher one for the crisis period. They both have the same assortative mixing coefficient for the industry group. For mutual information, In Table-2.1 it can be seen that these values significantly drops in the crisis period.

Scale-free fitting indices of each network show that they have very high fitting values for their degree distributions. For each network, the exponent of the scale-free distribution has values approximately close to or higher than 2. While mutual information based network in crisis period has the highest exponent value, the Pearson has the least one in the non-crisis period. However, networks show approximately

scale-free behavior.

Degree distribution informs us about the shape of a network, while entropy measures the regularity in the shape of a network [61]. All networks have approximately similar entropy values. Average closeness and average betweenness values of each network are also approximately same.

In summary, major differences between networks are observed in the diameter and the hub. Other considered fundamental parameters are approximately the same and no big difference between them. The diameter is a very sensitive measure since it only considers the extreme distance between two nodes, but average path length indicates less biased value than it. When we consider the average path length of networks, they are approximately the same. The differences between hub degrees can be explained as follows: since MST approach has geometric constraints in the construction process, the stocks that could be considered as outliers or possibly insignificant correlation could prefer to attach to the hub in order to obtain the minimum weight.

	Non-Crisis	Crisis
Degree Centrality	0.61	0.55
Hamming Distance	0.00452	0.00553
Closeness Centrality	0.52	0.25
Betweenness Centrality	0.61	0.55
Eccentricity Centrality	0.45	0.14

Table 2.2: Spearman correlation between local and global properties

Next, we quantify the difference between networks on the local and global topological scale. For a quantitative comparison, we calculated the Spearman rank order correlation coefficient. We consider the degree centrality and Hamming distance between the networks on the local topological scale, whereas on the global scale we compare the closeness centrality, betweenness centrality, and eccentric centrality. In the Table-2.2, we summarized our results. On the local topological scale, we find

that the Pearson correlation and mutual information stock networks are moderately similar in non- crisis period, whereas in the crisis period they became less similar. On the global topological scale, we find more interesting results, especially for the crisis period. In the non-crisis period, we observe low-rank order correlation for eccentric centrality and closeness centrality between stocks while betweenness centrality scores for stocks have a little higher correlation. In the crisis period, while the Spearman correlation for closeness centrality and eccentric centrality becomes extremely lower than the non-crisis period, betweenness centrality also becomes slightly less than the non-crisis period. In summary, these results indicate that the Pearson correlation and mutual information stock networks are very different on the global topological scale in the crisis period.

	Hub	**Eccentricity**	**Betweenness**	**Closeness**
Pearson (Non-Crisis)	HON	LNC	BRK.B	BK
NMI (Non-Crisis)	HON	BRK.B	MMM	MMM
Pearson (Crisis)	CSCO	CAT	DD	PPG
NMI (Crisis)	FOXA	LNC	TMK	TMK

Table 2.3: Central Stocks for each network

Furthermore, we investigate whether the Pearson network and mutual information network agree on the central stocks. In network analysis, centrality indices are developed to answer the question which are the most important vertices in a network?. Since importance has a variety of meanings or definitions, there is a large number of centrality measures proposed for networks.

Many of the centrality measures were originally introduced in social network analysis and many of the terms used to measure centrality indicate their sociological origin. These methods are also used in a large number of different disciplines or areas outside of the social network analysis, such as biology, computer science, urban networks, physics, finance, the Internet, and super-spreaders of disease.

In Table 2.3, both distances agree on only the hub in the non-crisis period, which is HON and they have a completely different set of central stocks in both time period. Closeness centrality and betweenness centrality indices for Mutual information network are same stocks in two time periods, whereas centrality indices indicate very different stocks for Pearson networks.

2.8.2 Impact of metrics on the performance of community detection

Financial systems as complex systems display hierarchical structures, in which clusters can be further partitioned into smaller clusters at certain level [64]. For example, the stock classification systems are based on this hierarchical taxonomy, such as the GICS classification system. The GICS methodology developed by Standard and Poors and MSCI/Barra in 1999 has been commonly accepted as an industry analysis framework for investment research, portfolio management, and asset allocation. In this classification scheme, companies are classified based on their principal business activity. As well as earnings and market perceptions are considered. Revenues play a significant role in this classification.

The GICS has a four-level hierarchical industry classification scheme. It consists of 10 sectors, 24 industry groups, 67 industries, and 147 sub-industries [1]. 8-digit code with text explanation is assigned to each company. For example, sector: Materials (GICS code: 15), industry group: Materials (GICS code: 1510), industry: Chemicals (GICS code: 151010 and sub-industry: Commodity chemicals (GICS code: 15101010).

GICS is the official SP industry classification system. GICS enables market

[1] As of September 1, 2016, S&P Dow Jones Indices and MSCI moved Equity Real Estate Investment Trusts and Real Estate Management & Development companies from the Financials Sector of their Global Industry Classification Standard (GICS) to a new Real Estate Sector. Please visit the website https://www.msci.com/gics.

participants to identify if stock movements are generally according to local or are part of a broader global trend. GICS classification is used also to test the clustering algorithm performance as a ground-truth set. It is the mainly used system in complex stock networks.

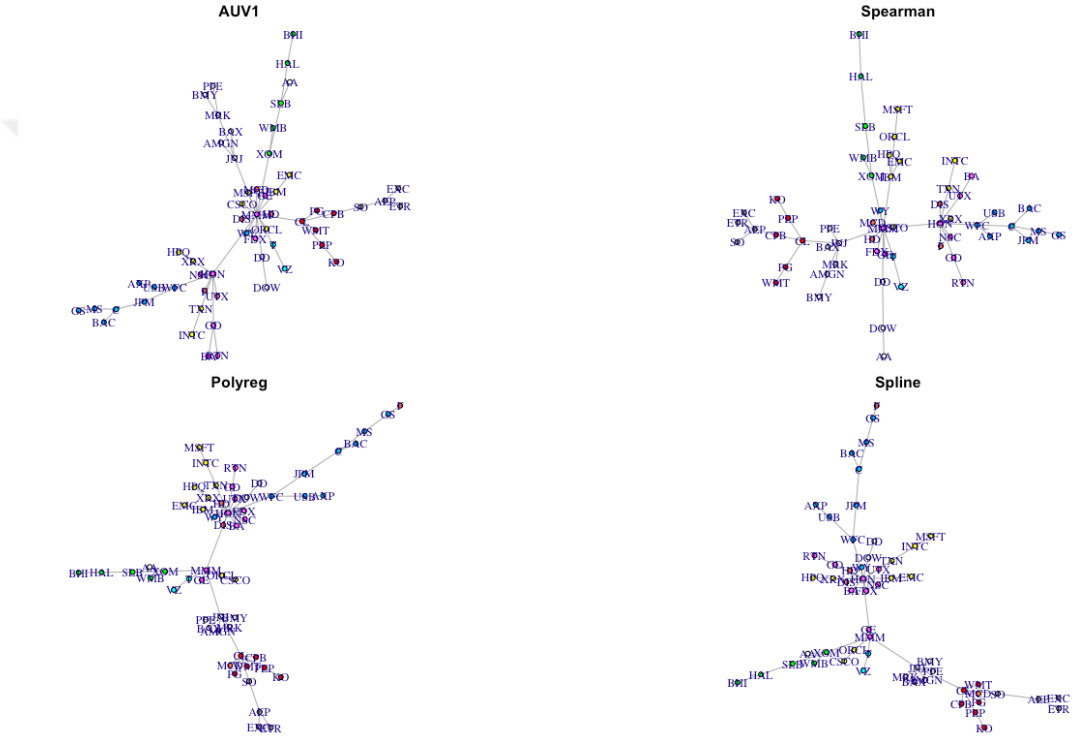

Figure 2.10: Example of networks constructed by Spearman, AUV1, Polynomial Regression and Spline Regression for 57 stocks in the time period between 2011 and 2015.

Identification of these clusters or communities based on stock returns has a variety of applications in finance such as creating efficient portfolios and risk management. For example, well-diversified portfolios can be created by choosing sample stocks from different clusters [70].

In order to provide an unbiased comparison, we use the same community detection algorithm for all networks. The separate networks constructed by six different

distance measures are considered and furthermore similar to previous sections two set of data set considered-crisis and non-crisis period.

We also consider the Spearman correlation between X and Y. It is simply defined as Pearsons correlation of the ranks of X and Y as follows:

$$r_s(X,Y) = cor(rank(X), rank(Y)) \tag{2.8.1}$$

,where elements of each vector are replaced with their ranks [97]. Therefore, r_s can detect monotonic relationships, i.e as the value of X increases, so does the value Y; or as the value of X increases, the value of Y decreases. In contrast to Pearsons correlation, the Spearman correlation does not require assumptions of linearity in the relationship between variables. However, similar to Pearsons correlation coefficient, $r_s(X,Y) = 0$ does not imply that random variables X and Y are independent, but only it means they are monotonically independent. In order to calculate the dissimilarities based on Spearman-rank correlation, we use the following: $d_s(X,Y) = 1 - r_s(X,Y)$. We call it *dissSpearman* for short in the followings. Please see the Figure-2.10 for illustration of networks for AUV1, Spearman, polynomial regression and spline regression.

Impact of the distance measures on clustering is assessed in two ways. First, the performances are evaluated by the assumption that no ground truth cluster set is available. In this case, we apply the internal indices to find out which metric produces well-separated and cohesive clusters. As mentioned previously, we apply geodesic distances and use the centroid of nodes chosen as the node that minimizes the largest distance from any other nodes in the cluster. Secondly, the performances are evaluated by the assumption that the ground-truth class labels for stocks are known a priori. We use normalized mutual information in order to compute the performance

of metrics. The NMI equals 1 if the clustering is identical to partitions and equals 0 if they are independent. In our study, we consider two types of external information as ground-truth. The first one is sector level and the second one is industry-group level, so we can evaluate the performance of algorithms and dissimilarity measures in terms of their ability to detect finer clusters.

For the performance of the community detection algorithm on six different distance measures, we created Tables-[2.4-2.5], in which column names represent the networks based on different metrics, and row names indicate the performance indices. The best scores for each index are highlighted in the tables.

	Pearson	Spearman	Polyreg	Spline	AUV1	AUV2
Number of Communities	23	19	21	22	19	18
C-Index	**0.0429**	0.0457	0.0592	0.0546	0.0528	0.0466
BetaCV	**0.2956**	0.3168	0.3598	0.3479	0.3552	0.3292
Dunn1	0.0728	0.0648	0.0662	0.0663	**0.0867**	0.0853
Dunn2	0.9	0.8719	0.7149	0.7825	**0.9351**	0.9296
Silhoutte	0.2957	0.3221	0.2860	0.2883	**0.3376**	0.3245
Entropy	3.0613	2.9020	2.9397	2.9988	2.8797	2.8490
NMI1	0.6070	0.6285	0.6153	0.6168	0.6464	**0.6574**
NMI2	0.7057	0.7270	0.6934	0.7160	**0.7393**	0.7211

Table 2.4: Performance of six different distance measures for non-crisis period

In Table 2.4, we observe that mutual information networks clearly outperforms all other distance measures in terms of internal indices and external indices. Pearson network has good scores in $C - Index$ and $BetaCV$ but is less successful for finding stock sectors.

In Table 2.5, we find very interesting results. Spearman rank order network has the best score in terms identification of each stock sector and industry group label. However, one of mutual information network has better scores for internal indices indicating that clusters found are coherent and well-separated comparing to other distance measures.

	Pearson	Spearman	Polyreg	Spline	AUV1	AUV2
Number of Communities	22	20	24	24	22	22
C-Index	**0.0468**	0.0450	0.0528	0.0612	0.0507	0.0492
BetaCV	0.3091	**0.2887**	0.3112	0.3377	0.3154	0.2966
Dunn1	0.0596	0.0489	0.0290	0.0271	**0.0908**	0.0804
Dunn2	1.0659	0.8419	0.7766	0.7702	**1.0684**	0.9619
Silhoutte	0.3082	0.3033	0.2722	0.2782	**0.3317**	0.3133
Entropy	3.0512	2.9694	3.1341	3.1432	3.0486	3.0385
NMI1	0.5510	**0.5930**	0.5351	0.5369	0.4883	0.4944
NMI2	0.6339	**0.6573**	0.6348	0.6353	0.6030	0.6129

Table 2.5: Performance of six different distance measures for crisis period

In summary, mutual information based networks outperformed other networks in terms of external indices and internal indices for the non-crisis period. In the crisis period, the Spearman network outperforms for finding stock sectors, whereas mutual information based networks have well-structured clusters.

2.9 Conclusion

In this study, we use two data sets composed of log-returns of daily adjusted close prices of 404 stocks of SP500. While the first one is in the time period from January 2007 to December 2009 called as crisis period, the second one is in the time period from January 2012 to December 2015 called as non-crisis period.

We have found that the mutual information and the Pearson correlation measures have a stronger monotonic relationship than crisis period. Furthermore, we investigate the cases where two measures are strongly disagreed by using extranormal information for each pair of stocks. We find that in the crisis period these values are stronger than the non-crisis period indicating strong deviations from Gaussianity. To assess the nonlinear dependence as an alternative to mutual information, we illustrate how to apply polynomial and spline regression to compute the model fitting indices

for defining the network adjacencies. Furthermore, for the strongly mismatched stock pairs, we conduct the polynomial regression and compare the results with a simple regression model. AIC, C_p and ANOVA tables confirm that simple regression is not enough to describe the relationship between these stocks. The relationship between entropy and the first four moments of daily stock return distributions are examined and we find a very sharp relationship between entropy and kurtosis. We list the reasons for the mismatches. We conclude that the main reason is that mutual information uses more information about the probability distribution than the Pearson correlation, which only uses the first two moments. And these mismatches could be due to having finite data samples.

As a second step, for each distance measure, we constructed separate minimum spanning tree based hierarchical stock networks and compared their local and global properties. We have found that in crisis period similarity between two networks significantly drops; however, in the non-crisis period, they have moderate Spearman correlation on local and global topological scales.

Finally, we also investigated the role of selection distance measures and the performance of community detection on each network obtained by using them. We found that two different versions of MI-based networks show high similarity with each other in both periods and their performance gets weaker in a crisis period in terms of identifying true sectors for each stock. However, if the criteria changes as finding coherent and well-separated stock communities, by far they were the most successful ones in two periods.

Chapter 3

Essay 2. Dynamic Evolution of Complex Mutual Information Theoretic Stock Networks

Abstract In this work, we analyzed how the local, mesoscopic and global topological properties of mutual information stock networks evolve annually in the time period from 2000 to 2015 in order to quantify the impact of the major crisis on the network structure. In addition to the classic network quantifiers, such as diameter and average path length, we also to use the information theoretic quantifiers. System entropy and normalized entropy of the market have shown decreased values during crash periods. The classic quantifiers indicate the shrinkage during those periods and, furthermore, we observed the hub emergence. Moreover, we check the scale free properties of networks. For mesoscopic or sector analysis, we propose some metrics adapted from classic cluster analysis in order to capture topological evolution of sectors within the system and Sectoral Entropy Index (SEI) is introduced to compute their structural changes over time. Sectors' structural changes are compared with each other and with the market. We list the important stocks for each year identified by four different centrality measures and compared with each other.

3.1 Introduction

Recent global financial crisis considered as the worst economic disaster since the Great Depression of 1929 by many economists started in 2007 with subprime mortgage crisis, which was triggered by a drastic fall in housing prices. It peaked with the collapse of Lehman Brothers and led to international banking crisis with huge impact on global financial system. Since then, many tools and methods have been developed to analyze

the relationship between financial market structure and economic crisis to describe the characters of financial dynamics [10].

Mantegna was one of first who applied MST into portfolio of stocks [64]. He used the data set composed of daily price return of S&P500 index in the time period from July 1989 to October 1995. He detected the hierarchical structure in the financial market in which stocks belong the same industry sector was homogeneously clustered. Bonanno et al. observed how the topology of MST changes in a portfolio of stocks sampled from different time horizons [13]. It is found that the topology MST is changing dramatically when considering time horizon gets shorter. They found that Epps effect exists between assets in the market, which says that the shorter the time period correlation diminishes. When the time horizon getting shorter, MST starts becoming more like a star shape network and GE is again the hub and in the center. Vandewalle et al. found that the minimum spanning tree constructed by the cross correlations of daily fluctuations for N = 6358 US stock prices for the year 1999 had scale-free topology the exponent $\alpha = 2.2$ [101]. Onnela et al. used the portfolio composed of 116 stocks of the S&P500 index and time period is from 1982 to 2000 [80]. It's been found that the asset tree shrinks during the market crisis and the value of the power-law exponent of the scale-free degree distribution changed during the Black-Monday period. Khashanah and Miao investigated the structural evolution of the US financial systems by applying MST method and PCA. They used the data set composed of the S&P 500 (SPX), VIX, three-month treasury bill (three-month T-bill), US dollar index (USDX), gold, and oil in time period from June 2006 to May 2009 [56]. They found that VIX was a dominator factor in the financial system and analyzing different snapshots of MSTs they identified that market became more integrated during the crisis. Zhao et al. studied daily correlations of 322 stocks of S&P500 by using heterogeneous time scales, they found global expansion and local

clustering market behaviour and while the links between sectors decrease, the links within the sector increase during the crisis [108].Heiberger analyzed the stability of US stock market by May-Wigner theorem from complex complex ecosystems and found the violation of the stability during the dot com bubble and subprime mortgage crisis [47]. Nobi et al. studied the effects of the 2008 global financial crisis on a local Korean market by constructing thresholds networks before, during and after the crisis. The networks are observed to be fatter during crisis than other periods and they have scale free structure within restricted range of threshold values [78].

In this study, we use the data set composed of log-returns of daily adjusted close prices of 404 stocks of SP500 during 4032 consecutive trading days in the time period from January 2000 to December 2015 after the removal of a few days because of the incomplete data. We study the evolution of the local, mesoscopic and global structural properties of the mutual information-based hierarchical stock networks over time and draw conclusions regarding the dynamics of the stock market based on the effect of major financial crisis.

The annual change of the first four moments of non-diagonal elements of mutual information matrix (MIM) are investigated and increase in the mean value during the crisis is observed. We compute the classic quantifiers from the network theory analysis, such as average path length and diameter and it is observed that they both shrink during the crisis indicating market behaves like one. The relationship between centrality measures, and annual evolution of their means are studied. We found the hub emergence during the crisis and they are almost always disagree on the identified important stocks except for the year 2015.

Degree distribution of networks is analyzed to determine whether they are scale-free. In addition to the classical quantifiers, we propose to apply information-theoretic quantifiers, such as the entropy and normalized entropy to investigate the

system entropy for various market levels. We discover that a decrease in the financial network entropy and in the normalized network entropy is associated with a crisis condition. Furthermore, we study the time evolution of the sector structure of market by assortativity mixing coefficients and we propose some metrics adapted from cluster analysis to quantify this evolution– sector diameter and sector average path length are utilized in this paper. Stock Sectoral Entropy (SSE) is proposed to measure structural changes of each sector. We investigate the structural similarities between sectors and markets.

The essay is organized as follows. Next section we introduce the methodology for how to compute mutual information between each pair of stocks to construct hierarchical stock networks in our study. In Section 3. 3, we investigate the first four moments of the non-diagonal elements of mutual information matrix. In section 3.4, annual evolution of the hub of the system and mean of other centrality measures are studied and important stocks are listed for each year according to each measure. In section 3.5, we show how the degree distribution of the networks change. In section 3.6, entropy of the system is illustrated together with normalized entropy. In section 3.7, we analyze how the mesoscopic structure of the market evolve over time in terms of sector structure and topological evolution of sectors is also studied. Finally, the last section is the summary of the work.

3.2 Methodology

In order to construct the hierarchical stock networks based on minimum spanning tree (MST), we first need to compute all pairwise similarity between each stock in the portfolio, then transform this similarity or proximity matrix into dissimilarity matrix.

In this work, in order to obtain mutual information based adjacency matrix, stock returns are firstly discretized by the equal-width approach. We choose the number of bins as $\sqrt{n}$. Then, pairwise mutual information between each stock is computed by using the Miller-Madow estimation [68]. Finally, we estimate the universal distance metric version 2, i.e $dissAUV2$, from the obtained adjacency matrix. Please see Section-2.3.2 for more details. After that, we apply Prim's algorithm to identify the MST structure in the portfolio [87].

3.3 Characterization of Market

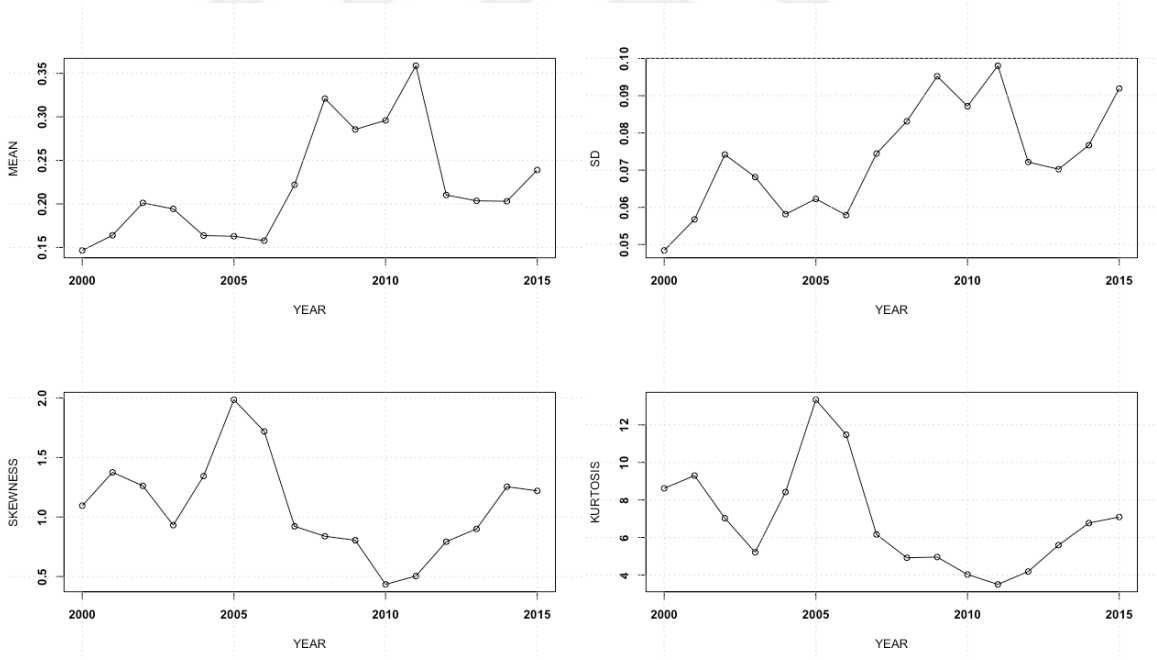

Figure 3.1: The mean, standard deviation, skewness, and excess kurtosis of the pairwise mutual information annually

Let us investigate the annual change of the first four moments of the non-diagonal ($i \neq j$) elements I_{ij}^t of mutual information matrix and their relation to each other.

The first moment is the mean defined as

$$\bar{I}(t) = \frac{1}{N(N-1)/2} \sum_{I_{ij}^t \in I^t} I_{ij}^t \tag{3.3.1}$$

Other higher order normalized moments are computed as well.

Variance is :

$$\lambda_2(t) = \frac{1}{N(N-1)/2} \sum_{I_{ij}^t \in I^t} (I_{ij}^t - \bar{I}^t)^2 \tag{3.3.2}$$

the skewness is :

$$\lambda_3(t) = \frac{1}{N(N-1)/2} \sum_{I_{ij}^t \in I^t} \frac{(I_{ij}^t - \bar{I}^t)^3}{\lambda_2(t)^{\frac{3}{2}}} \tag{3.3.3}$$

and the kurtosis is

$$\lambda_4(t) = \frac{1}{N(N-1)/2} \sum_{I_{ij}^t \in I^t} \frac{(I_{ij}^t - \bar{I}^t)^4}{\lambda_2(t)^2} \tag{3.3.4}$$

The annual change of the mean, standard deviation, skewness, and excess kurtosis of the mutual information measures are plotted in Figure 3.1. The effect of Dot-com bubble during 200-2002, the financial crisis of 2007-2008, and European sovereign debt crisis starting from the end of 2009 are clearly visible in all these quantities, such as an increase in mean mutual information values during those periods. In [81], similar results are observed for all quantities for the effect of Black-Monday, such as the mean correlation coefficient is seen higher than average on that time interval.

In Figure 3.1, it is clear that the first two moments and the last two moments show similar patterns. We computed the Pearson's and Spearman's correlation coef-

ficients between mean and standard deviation, which are 0.9 and 0.92 in order, and between skewness and kurtosis, which are 0.96 and 0.95 in order. Results show that first two and last two are strongly correlated to each other.

Normalized tree is length is defined by [81] as follows:

$$L(t) = \frac{1}{N-1} \sum_{d_{ij}^t \in T^t} d_{ij}^t \tag{3.3.5}$$

The mean, standard deviation, skewness, and excess kurtosis of normalized tree lengths are shown in Figure 3.2

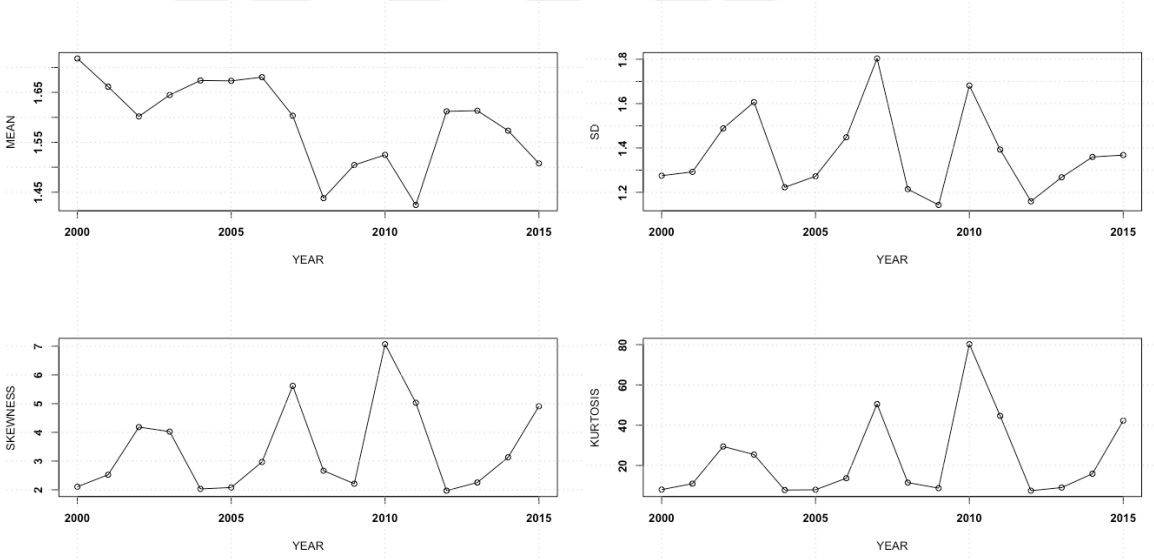

Figure 3.2: The mean, standard deviation, skewness, and excess kurtosis of the annual normalized tree length

The Pearson's and Spearman's correlation between mean mutual information and mean of normalized tree length are -0.96 and -0.94 in order as expected. That is, they have very strong anti-correlation because the way we compute distances from the mutual information matrix. Furthermore, we compute the correlation between corresponding moments of normalized tree length and mutual information matrix. The

Pearson's and Spearman's correlation between the standard deviation of the mutual information and standard deviation of the normalized tree length are -0.06 and -0.09. For skewness, -0.47 and -0.27 and for kurtosis, -0.42 and -0.33 are found respectively. Thus, these found correlations show the essential differences between mutual information and the normalized tree length for higher moments. Furthermore, in [81], authors found very striking anti-correlation between skewness of the correlation matrix and skewness of the normalized tree length in their considered time period from 1980 and 1999. This is one of the main difference changing the metric from Pearson's correlation to mutual information.

Increase in the mean of the mutual information and decrease of the normalized tree length indicate that how the market moving together with strong correlation causing to the very strong shrinkage in the average of the pairwise distances between stocks.

Besides the normalized tree length, other important concepts related to the network topology are average path length and diameter of the network. The diameter of the network is the simply the length of the longest geodesic path between any nodes in the network. The diameter is usually less useful than the average path length of the network in real networks because it only considers the extreme distance between two nodes and therefore it is very sensitive to outliers [72]. That is, a simple change to one node or couple nodes can have a huge impact on the diameter. A more robust indicator of the network behavior as a whole is the average path length (i.e the mean distance or the characteristic path length), which is computed as the average of pairwise geodesic distances between all nodes in the network. Average path length and diameter of the stock network as a function of time is plotted in Figure 3.3. The Pearson's correlation and Spearman's correlation between two are found as 0.93 and 0.96 in order. Due to the tree structure, they have a very strong correlation. Again,

major financial crises are clearly visible in these figures and average path length and diameter indicate that during crashes the market is moving together tightly and it shrinks.

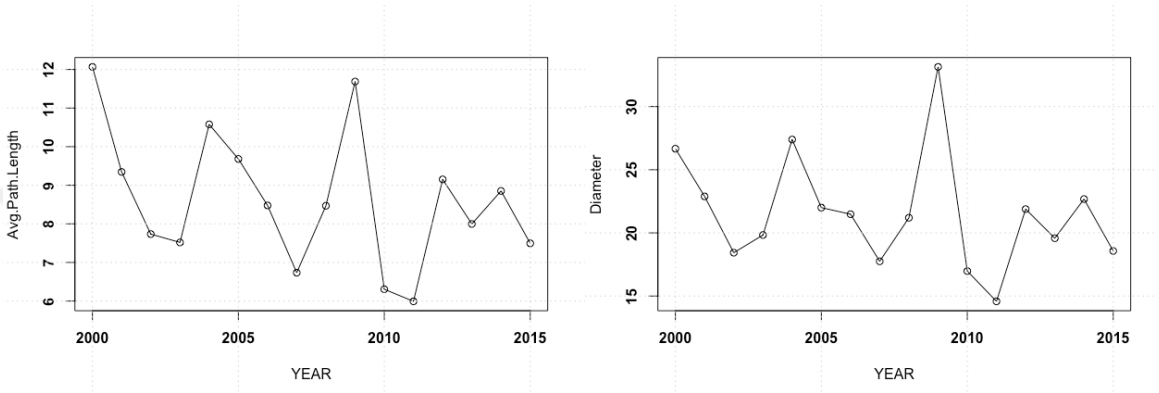

Figure 3.3: Average path length and diameter

3.4 Time Evolution of Centrality Measures

In network analysis, centrality indices are developed to answer the question which are the most important vertices in a network?. Since importance has a variety of meanings or definitions, there is a large number of centrality measures proposed for networks.

Most of the centrality measures were originally introduced in social network analysis and many of the terms used to measure centrality indicate their sociological origin. These methods are also used in a large number of different disciplines and areas outside of the social network analysis, such as biology, computer science, urban networks, physics, finance, the Internet, and super-spreaders of disease.

In financial networks, centrality measures are widely applied in the interbank lending market, which is considered as a network where banks are the vertices and the claims and liabilities between banks define the weight of links. The goal is typically to

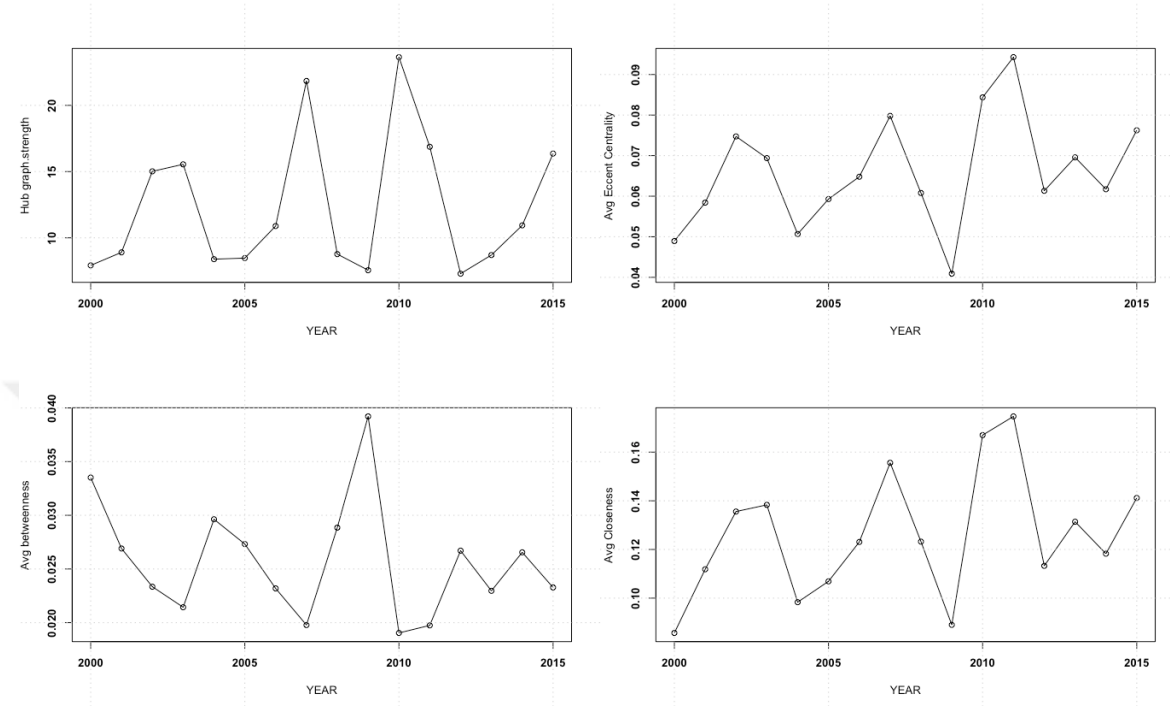

Figure 3.4: Mean of Centrality Measures

identify the important financial institutions and how their elimination in the network affect the stability of network structure and thus the stability of the banking system [14].

Comparing the interbank networks, the centrality measures are not explored too much yet in the stock networks; however, there are some examples, such as [80] and [21]. [80] illustrated that the stocks of the optimal Markowitz portfolio always located on the outskirts of the tree with respect to the central nodes or root of the tree. In [21], the top 10% most highly connected stocks are selected from the threshold network to compute new indexes based on degree connectivity. The performance of new indexes are tested against existing major indexes by calculating the Spearman's correlation and find statistically significance.

More formally, the centrality can be defined as a function $f : V \rightarrow R$, which induces a total order on V. Node v_i is said as central as v_j if $f(v_i) \leq f(v_j)$ [106]. Next

we give the definitions for the centrality measures which we used in this study.

Degree centrality

The adjacency matrix A of $G = (V, E)$ is the matrix with elements A_{ij} defined as:

$$\mathbf{A}(i,j) = \begin{cases} 1 & \text{if } \mathrm{v}_i \text{ is adjacent to } \ v_j \\ 0 & \text{otherwise} \end{cases}$$

Thus, the connectivity (also known as degree) of the node v_i is defined by

$$d_i = \sum_j \mathbf{A}(i,j) \tag{3.4.1}$$

If the network G is unweighted d_i equals the number of nodes that are directly linked to it. In weighted networks, the connectivity equals the sum of connection weights or strengths between node i and the other nodes incident with it. While degree is a local attribute, average degree is global attribute and defined as

$$\mu_d = \frac{\sum_i d_i}{n} \tag{3.4.2}$$

Eccentric centrality

The eccentricity of a node v_i is define as the maximum distance from v_i to any other node in the network, i.e.

$$e(v_i) = \max_j \left\{ d(v_i, v_j \right\} \tag{3.4.3}$$

Eccentricity centrality is thus expressed as:

$\mathrm{c}(\mathrm{v}_i) = \frac{1}{e(v_i)} = \frac{1}{\max_j \left\{ d(v_i, v_j) \right\}}$

A node v_i with smallest eccentricity is called a center node, while the one with greatest eccentricity is called a periphery node.

Closeness centrality

The closeness centrality measures the reciprocal average topological distance of vertex v_i to all other vertices in the network. This measure is normalized by multiplication factor N-1 to $0 \leq CC_{v_i} \leq 1$, and is formulated as

$$CC_{v_i} = \frac{N-1}{\sum_j d(v_i, v_j)} \tag{3.4.4}$$

A node v_i with the smallest total distance is called the *median node*.

Betweenness centrality

Betweenness centrality measures a vertex centrality in terms of its location between other pairs of vertices in graph. It is most commonly calculated as the number of shortest paths from all vertices to all others that pass through that node.BC_v is defined as

$$BC_v = \sum_{i,j \neq v}^{N} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \tag{3.4.5}$$

where σ_{ij} is the total number of shortest paths from i to j which includes node v.

We plot the annual change of maximum node weight, i.e. hub, and the mean of eccentric centrality, the mean of closeness centrality and the mean of betweenness centrality in Figure 3.4. In all figures, the major crisis can be detected clearly. One of the interesting phenomena of time-varying stock networks is the hub emergence during the crisis. In the year 2010, the stock SNA (Snap-on Incorporated) is observed

to have an all-time highest node degree of 31 among all other stocks. SNA's sector is Industrials and its industry is Machinery. We also list the tick of the hub stocks for each year in Table 3.2 along with other central stocks found to be important according to other centrality definitions.

In Figure 3.4, the impact of topological shrinkage in networks is clearly seen on the centrality measures, as well. While the average eccentric centrality and the average closeness centrality increase during crashes, the average betweenness centrality decreases. Moreover, we computed Spearman's correlation between the centrality measures and the result is shown in Table 3.1. While betweenness centrality shows strong anticorrelation with others, the hub, the closeness, and the eccentricity show a strong correlation with each other. The closeness and eccentricity have an almost perfect correlation.

Table 3.1: Spearman's Correlation between mean of the centrality measures

Centrality Measure	**Eccentricity**	**Betweenness**	**Closeness**
Hub	0.86	-0.84	0.89
Eccentricity		-0.94	0.97
Betweenness			-0.91

Let us now investigate how these correlation and anticorrelation have an impact on the identified important stocks. In Table 3.2, we also included stocks of the same maximum degrees and marked the stocks with maximum weight. It is clearly seen that all these four centrality measures are not agreed with each other in terms of important stocks for some years. For example, Eccentric centrality and closeness centrality seem to be very related to each other; however, they indicate different stocks as important for the year 2014. They all agree upon the same important stock only for the year 2015, which is MMC (Marsh & McLennan Companies, Inc.). It is evident that there is no specific stock constantly considered as important over

time; however, when each stock's sector is considered, the Finance Sector is found to be more important according to all four measures. From 2000-2015, the stocks belong this sector became 56.25% as hub and 62.5% as important according to other centrality measures. After the finance sector, stocks belong to the Industrials sector became important 31.25% of all considered time period.

Year	**Hub**	**Eccentricity**	**Betweenness**	**Closeness**
2000	GPC*,CMA	ORCL	ORCL	ORCL
2001	C*,BBT	GS	MS	MS
2002	RF	BBT	BBT	BBT
2003	JPM	C	JPM	JPM
2004	DD	ETN	ETN	ETN
2005	PX	MTB	BBT	BBT
2006	BEN	CMI	PH	CMI
2007	LNC	USB	USB	CMA
2008	HST	MAC	VNO	VNO
2009	DOV	VNO	DD	DD
2010	SNA	DOV	PCAR	DOV
2011	TROW	IVZ	TROW	TROW
2012	JEC	CAT	JEC	JEC
2013	TROW	IVZ	TROW	TROW
2014	SNA	HON	TROW	AME
2015	MMC	MMC	MMC	MMC

Table 3.2: Central Stocks Annually

3.5 Degree Distribution of Stock Networks

One of the most important properties of a network is its degree distribution. In many real-world networks, it has been found that the empirical degree distribution $f(k)$ follows a *power-law*:

$$f(k) = Ck^{-\alpha} \tag{3.5.1}$$

where C and α denote positive real numbers. The network whose degree distri-

bution follows power-law degree distributions are said to present scale-free topology [7] with scaling parameter α, that is also called as the exponent of the power-law.

Let us take the logarithm of the both sides of equation (3.5.1):

$$\log f(k) = \log\left(Ck^{-\alpha}\right) \tag{3.5.2}$$

which yields:

$$\log f(k) = -\alpha \log k + \log C \tag{3.5.3}$$

Thus scale-free topology implies a straight line relationship in the log-log plot of k versus $f(k)$, with $-\alpha$ giving the slope of the line. The standard strategy to check if a network has scale-free behavior is to apply a least-square fit of the points $(\log k, \log f(k))$ to a line [75].

To measure the straight line relationship between $(\log k, \log f(k))$, we apply R^2 statistics that takes values between 0 and 1 and explains the proportion of variance explained by the least-square regression line. We call this value as the scale-free topology fitting index as similar in [107].

The power law indicates that the most of nodes in the network have very small degrees, whereas there are a few nodes (hubs) that have substantially higher degrees, i.e, they connect with these nodes of very small degrees and connect them to the system [7]. Therefore, scale-free networks are substantially heterogeneous and their topology is controlled by these hubs [107].

Vandewalle et al. studied the minimum spanning tree constructed by the cross-correlations of daily fluctuations for 6358 US stock prices for the year 1999 and found

that scale-free topology for the network with the exponent $\alpha = 2.2$ [101]. Similarly, authors in [81] also found that the asset tree generally had scale-free properties with exponent $\alpha = 2.1$ for the outside of crash periods; however, for the Black Monday period, they had observed a low value of exponent as $\alpha = 1.8$.

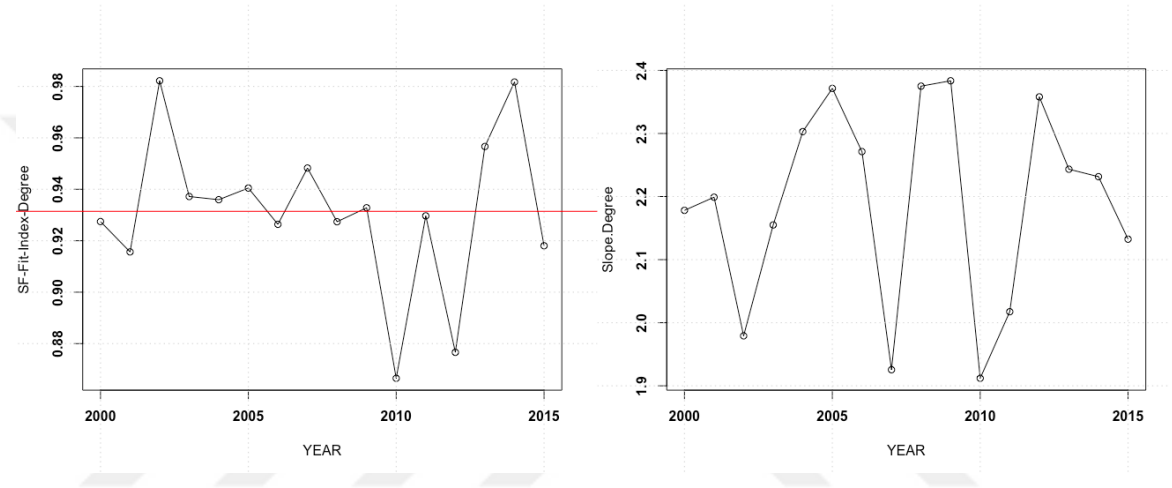

Figure 3.5: Annual Scale Free Fit Index and Slope

We now investigate how this exponent α changes over time for our mutual information based stock networks and if they hold the scale-free properties. We plot the results in Figure 3.5. The value of exponent α typically fluctuates between 2.13 and 2.38 for the normal topology and we have $\alpha = 1.98$ for 2002,$\alpha = 1.93$ for 2007,$\alpha = 1.91$ for 2010 and $\alpha = 2.02$ for 2011. Clearly, we observe the similar results for the non-normal topology of stock networks that we have exponent α values slightly less than or equal to 2.

In Figure 3.5, we also showed how R^2 values change annually and found that its average value of 0.93 including crisis periods. We see the minimum R^2 value as 0.86 in the year 2010, indicating that annual values of R^2 are always higher than 0.85.

In Figure 3.6, we present the log-log degree distribution of stock networks for each year. When we look at closely the years, in which the exponent α less than 2, we

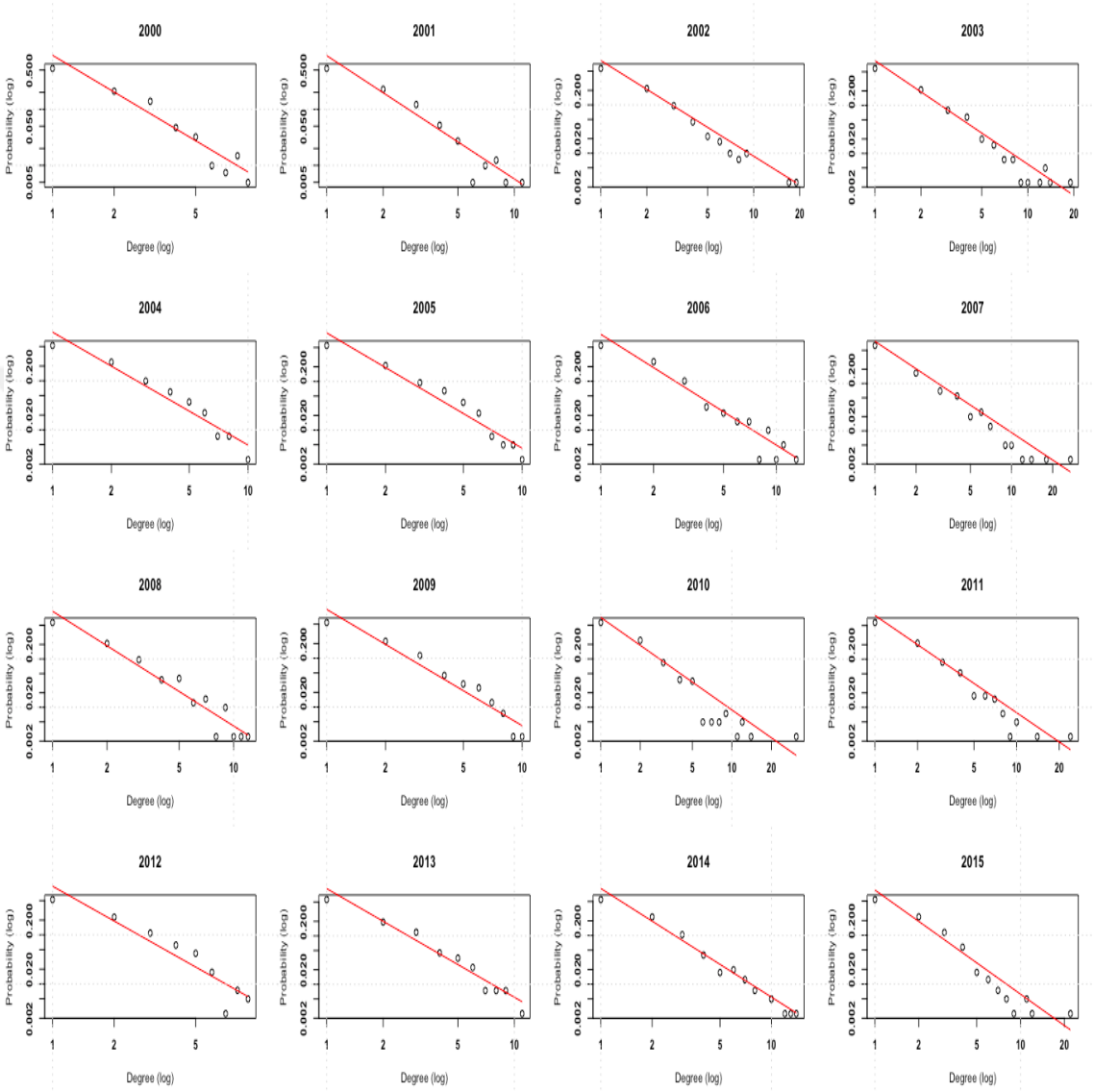

Figure 3.6: Log-Log degree distribution

see that some nodes have a huge number of connections. In Figure 3.4, the emergence of hubs with a large degree of connections also was observed during non-usual business time periods. However, we still included these stocks for the calculation of R^2 and did not treat them as outliers.

In summary, the most of time the mutual information based stock networks had scale-free properties and high scale-free topology fitting index values even including

crash periods; however, exponent α values approximately less than or equal to 2 for three major crisis periods.

3.6 Entropy of the System

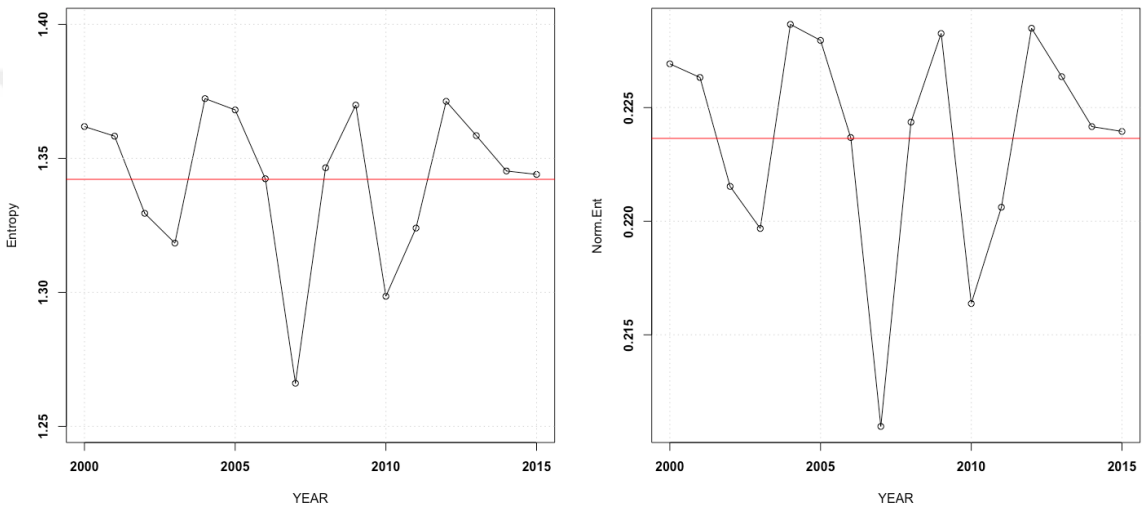

Figure 3.7: System Entropy

Another approach to exploring changes in the dynamic interaction between system entities is by the use of entropy developed in Information Theory. Shannon entropy [93] measures the randomness of the network when we apply it to networks degree distribution. The higher the value of entropy, the more random is the network. Let $H(G)$ be the expected number of nats in a network $G = (V, E)$ and $p(k)$ the probability that randomly selected node has exactly k edges, i.e., $p(k) = \frac{k}{N}$ where N is the total number of nodes.

$$H(G) = -\sum_{k}^{n} p(k) \log p(k) \tag{3.6.1}$$

$H(G)$ is used to calculate entropy of the degree distribution $p(k)$. Degree distribution informs us about the shape of a network, while entropy measures the

regularity in the shape of a network [61]. The maximum value of graph entropy is $H_{max}(G) = \log N$ for $p(i) = 1/N \ \forall i = 1, ..., N$ and the minimum value of graph entropy is $H_{min}(G) = 0$. Therefore, if $H(G)$ is the entropy of the network, then the normalized entropy of a given network is defined as:

$$H(G)_{norm} = \frac{H(G)]}{\log N} \tag{3.6.2}$$

We show how the entropy and the normalized entropy of stock networks changes for each year in Figure 3.7. While the minimum entropy value is observed as 1.266 in the year 2007, the maximum entropy value is found as 1.372 in the year 2004. Moreover,the year 2012 actually has very close entropy value to 2004, which is observed as 1.371. The entropy values indicate that during crisis market have lower entropy values than 1.342-the average entropy value of the market and it is colored as red in Figure 3.7. The normalized entropy values show similar patterns as the unnormalized entropy values.

Therefore, low entropy and normalized entropy values indicate that the market tends to act like a unified way or one during the major crisis periods. In another words, stock networks becomes more structured than random, because its entropy is lower than non-crisis periods.

Moreover, we compare the average entropy and normalized entropy of stock networks with very well known network structures: Barabarasi-Albert (BA) Scale-free Network [7], Erdos-Renyi (ER) random network [34], 2-regular network, and Watts-Strogatz (WS Small-world network [102] models. For each network structure, we fix the total number nodes equal as the number of stocks in our portfolio and we compute the average entropy and normalized entropy of 1000 simulations of each network structure. In Table 3.3, we see that on average stock networks are less random

than ER networks and more random than BA networks. Conversely, they are less structured than WS, 2-regular, and BA networks.

Network Models	Entropy	Norm. Entropy
BA	1.296	0.216
ER	1.691	0.282
WS	0.19	0.032
2-Regular	0	0
Stock Network	1.342	0.224

Table 3.3: The Entropy of Different Network Structures

3.7 Mesoscopic Analysis of the Market

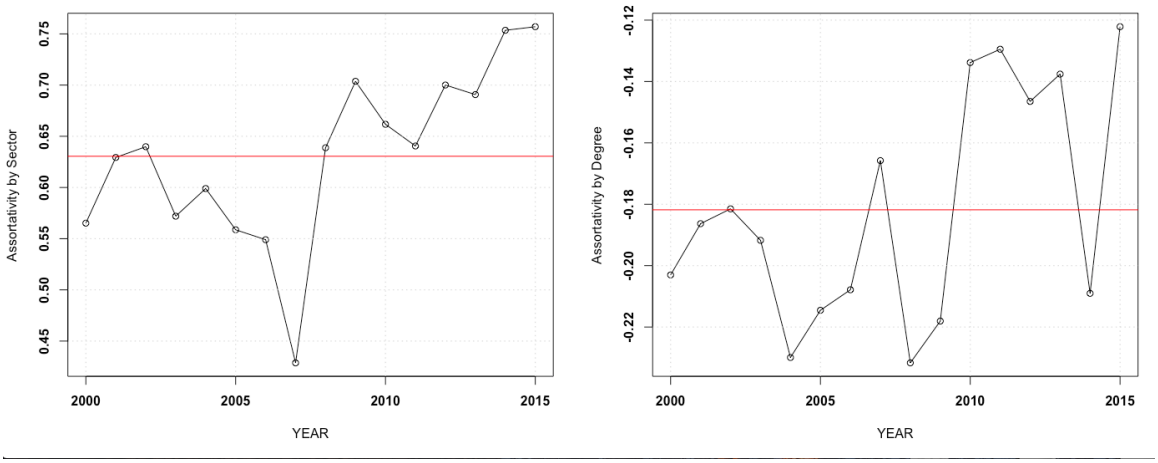

Figure 3.8: Annual Assortavity by Sector and Degree

Networks representing the real systems are not regular or random at the global level and have some level of order and organization at mesoscopic level, in which some nodes are highly connected among themselves and few connections with rest of the network [76]. These groups of nodes densely interconnected is called modules or communities [72].

Financial systems as complex systems display hierarchical structures, in which clusters can be further partitioned into smaller clusters at certain level [64]. For example, the stock classification systems are based on hierarchical taxonomy, such as GICS (Global Industry Classification Standard) classification system.

The GICS methodology developed by Standard and Poors and MSCI/Barra in 1999 has been commonly accepted as an industry analysis framework for investment research, portfolio management, and asset allocation. In this classification, companies are classified based on their principal business activity. As well as earnings and market perceptions are considered. Revenues play a significant role in this classification.

The GICS has a four level hierarchical industry classification scheme. It consists of 10 sectors, 24 industry groups, 67 industries and 147 sub-industries [1]. 8-digit code with text explanation is assigned to each company. For example, sector: Materials (GICS code: 15), industry group: Materials (GICS code: 1510), industry: Chemicals (GICS code: 151010 and sub-industry: Commodity chemicals (GICS code: 15101010).

GICS is the official S&P industry classification system. GICS enables market participants to identify if stock movements are generally according to local or are part of a broader global trend. GICS is the one of the mostly assumed to be the ground-truth clusters in order to investigate the performance of clustering algorithms in stock cluster analysis and furthermore in stock network studies, in which nodes are mostly colored according to this classification scheme.

Fluctuations of stock returns are not independent and have strong correlations within the sector or industry which they belong [21]. The assortativity of network

[1] As of September 1, 2016, S&P Dow Jones Indices and MSCI moved Equity Real Estate Investment Trusts and Real Estate Management & Development companies from the Financials Sector of their Global Industry Classification Standard (GICS) to a new Real Estate Sector. Please visit the website https://www.msci.com/gics.

measures the similarity of connections in a network, for example, social networks are often assortative by race, language or ethnicity. In our case, the assortative mixing by sector means that stocks prefer to have connections with the stocks from the same sector. Moreover, the assortative mixing by degree means that high-degree stocks prefer to be connected to other high-degree stocks, or low ones to low ones.

We plot the annual assortativity mixing by the sector and by the degree in Figure 3.8 and the red line indicates the all-time average. In the figure, the stock networks before the subprime crisis have high values of assortativity indicating a strong sector structure and a clear sector division of the network. The stock market has an identifiable sector structure even during the dot-com crisis. The assortativity by sector decreases in the year 2007 and starts increasing after the crisis which illustrates that the links between stocks are not randomly distributed and have a clear sector structure.

In Figure 3.8, the assortativity mixing by degree shows that the stock networks have a disassortative mixing behavior indicating that stocks with high-degree links tend to be connected to the stock with low-degree. However, we observe clear pattern changes during the crisis and the reconfiguration of stock connections became highly mixed, such as the tendency of the stocks with many connections prefer to have connections with other high degree stocks.

Beside the assortative mixing, there are two other fundamental concepts that can be used to quantify the sector structure or the sector division of the stock networks. First one is related to the mean distance between stocks of each sector. To quantitatively characterize the interaction structure of the stock sectors in a network, we propose to use the average topological (i.e. geodesic) distance within the business

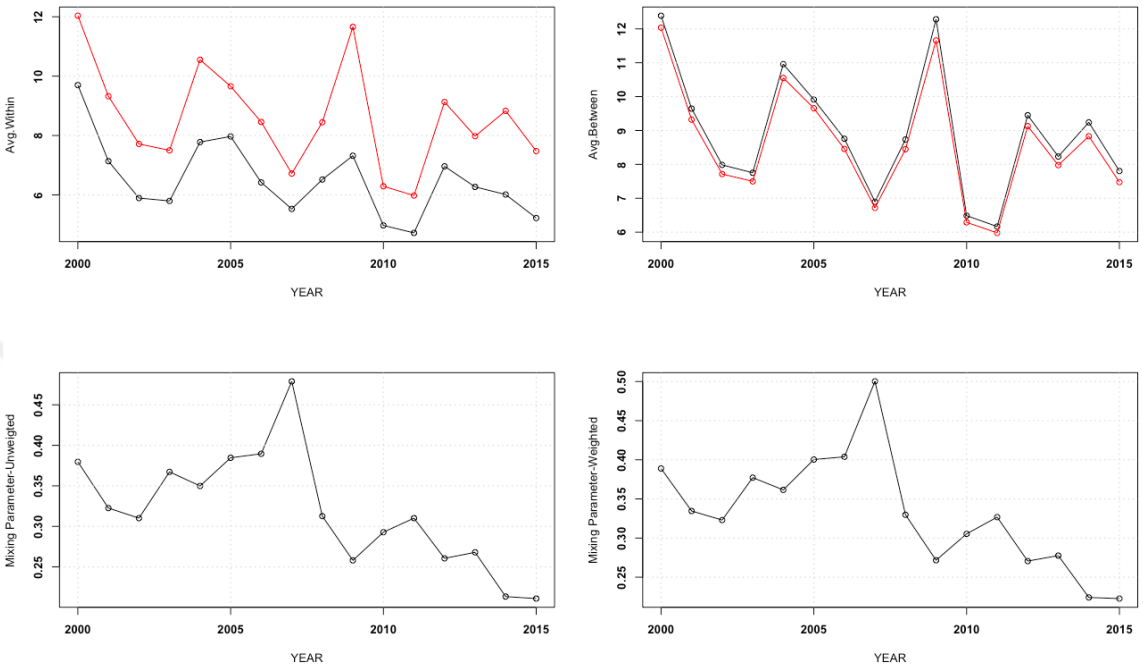

Figure 3.9: Mixing Parameter

sector as

$$\bar{d}_{ij}^{in} = \frac{1}{N_{in}} \sum_{c} \sum_{i \neq j} d_{ij}(c) \tag{3.7.1}$$

where $d_{ij}(c)$ is the geodesic distance between stock i and stock j belonging to same sector and N_{in} is the number of distinct intrasector distances. Furthermore, we define the average topological intersector distance as

$$\bar{d}_{ij}^{be} = \frac{1}{N_{out}} \sum_{c \neq q} \sum_{i \neq j} d_{ij}(cq) \tag{3.7.2}$$

where $d_{ij}(c)$ is the geodesic distance between stock i and stock j belonging to different sectors and N_{out} is the number of distinct intersector distances.

We plot the results of the average topological intrasector and intersector in Figure 3.9 together with the average path length colored red in order to compare

each other. By the definition of the communities or clusters, the average distance inside a community should be less than that between two communities. The figure clearly shows that average path length has been always higher than average intra-sector (i.e., within the sector) and very close to the intersector (between sectors). Shrinkage of the market during the crisis is again clearly visible in both top figures in the plot. Furthermore, we can see in the figure that both the average of intrasector distances and the average of intersector distances most of the time increase and decrease together.

And the last concept known as *mixing parameter* is defined as

$$\mu = \frac{\sum_i k_i^{ext}}{\sum_i k_i^{tot}} \tag{3.7.3}$$

where k_i^{ext} is short for the external degree of vertex i representing the total number of links connecting it to other vertices from different communities and k_i^{tot} is the total degree of vertex i [60]. Based on the definition of community in a strong sense, each vertex should have more links within their community than with the rest of the network. Thus, for $\mu > 1/2$ communities cannot be not easily detected and they disappear [104]. The mixing parameter is one of the most important parameter for in the LFR (Lancichinetti, Fortunato & Radicchi) benchmark networks [83], which is computer-generated networks with a well defined community structure (i.e,a known ground truth) in order to test a community detection algorithm.

We compute the mixing parameter considering the stock networks as weighted and as unweighted. In Figure 3.9, bottom left and right shows the results and they both have the clear anti-correlation between the assortative mixing by sector. In the year 2007, disruption of sector division in the network is again clearly visible.

In summary, we quantify the time evolution of the sector structure or sector

division of the stock networks in different ways. The mixing parameter and assortative mixing by sector pointed the year 2007 indicating that stocks from the same sectors prefer to have connections with stocks from different sectors and leaving the stock connections with the highly mixed environment. While during the subprime crisis the sector structure dissolves, in non-crisis periods stocks tend to have connections with stocks from the same sector.

3.7.1 Structural evolution of sectors within the system

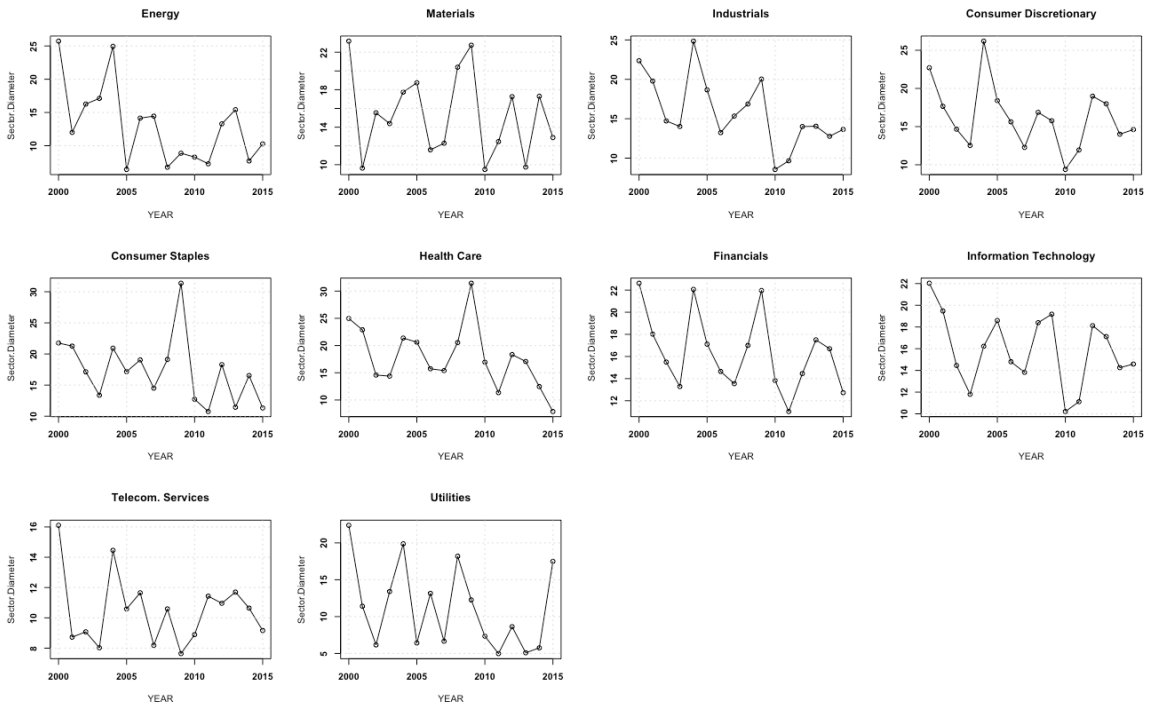

Figure 3.10: Sectors' diameter

The most of the studies in graph clustering focus on the static version of the problem because the problem itself does not have a generally accepted definition; therefore, the analysis of dynamic communities is not mature yet [40]. So far we have quantified sector division and correlation between their interactions in a network, we propose to use classic network quantifiers and information theoretic quantifiers in

order to capture the topological evolution of each sector by treating them as fixed communities, in which stock labels (sectors) are not changed over time.

The topological properties we are interested in are the diameter and the average path length for each sector and how they change annually. We use the geodesic distances again for each calculation. In Figure 3.10 we plot how annually the diameter of each sector change and in Figure 3.11 we plot the annual change of the average path length for each sector. In both plots, we can clearly see the impact of shrinkage of the market on each sector during the crisis, such as, information technology sector during the dot-com bubble and Financial sector during the subprime crisis.

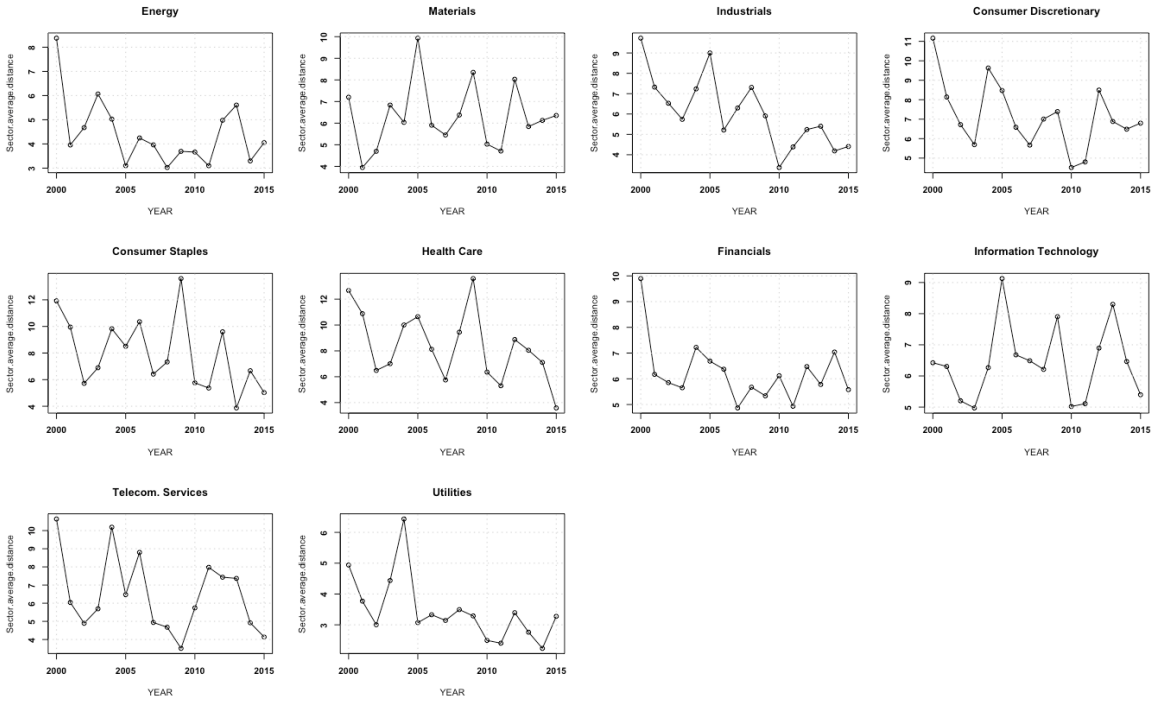

Figure 3.11: Sectors' avg path length

Let us investigate the relationship between each sector's diameter with its own average path length if any strong correlation exists as the one exists between the market's diameter and market average path length. In Table **??**, we show the computed Spearman's correlation values. While it is clear that some sectors' diameter

and average path length have a very strong correlation, such as Energy and Consumer Discretionary sectors, and some do not, such as Information Technology and Financials sectors.

Sector	ρ	**Sector**	ρ
Energy	0.94	Health Care	0.92
Materials	0.74	Financials	0.57
Industrials	0.90	Infor.Tech	0.59
Cons.Disc.	0.97	Tel.Serv	0.74
Cons.Staples	0.92	Utilities	0.87

Table 3.4: Correlation between Sector Diameter and Average Path Length

Now let us find out which sector or sectors show similar behavior with the market and with each other. We calculated the Spearman's correlation between each sectors' diameter and average path length including the market and illustrated the results in Table-3.5.

Finance sector's diameter and Consumer Staples' diameter have a very strong correlation with the market; furthermore, both sectors have also very high correlation Health Care sector. We also observe that the Information Technology have a strong correlation with the Health care sector and Consumer Discretionary sector. Materials and Energy sectors have a weak anti-correlation with each other.

As we already mentioned before, the diameter is sensitive the outliers and measure the extreme distance between two nodes in a network. When we investigate Table 3.5, Health Care sector has the highest correlation with the market and the second sector with the highest correlation with the market is the Consumer Discretionary sector. Two sectors have also a high correlation between each other, but the Health care sector has the highest correlation with the Consumer Staples sector. Telecommunication Services Sector and Materials Sector have a weak anti-correlation with each other.

	Market	10	15	20	25	30	35	40	45	50	55
Market	1.000	0.15	0.60	0.68	0.73	0.84	0.73	0.81	0.76	0.17	0.48
10		1.00	-0.02	0.35	0.30	0.18	0.12	0.23	0.06	0.22	0.36
15			1.00	0.56	0.50	0.56	0.41	0.48	0.51	0.17	0.41
20				1.00	0.71	0.75	0.79	0.79	0.75	0.02	0.49
25					1.00	0.66	0.73	0.78	0.83	0.52	0.41
30						1.00	0.84	0.79	0.78	0.05	0.56
35							1.00	0.86	0.82	0.09	0.41
40								1.00	0.79	0.30	0.32
45									1.00	0.20	0.41
50										1.00	0.09
55											1.00

	Market	10	15	20	25	30	35	40	45	50	55
Market	1.000	0.22	0.54	0.61	0.87	0.81	0.93	0.66	0.57	0.25	0.54
10		1.00	0.08	0.18	0.32	0.12	0.11	0.27	0.02	0.40	0.50
15			1.00	0.23	0.50	0.44	0.47	0.23	0.43	-0.06	0.33
20				1.00	0.70	0.49	0.67	0.27	0.25	0.14	0.61
25					1.00	0.60	0.80	0.58	0.48	0.31	0.63
30						1.00	0.84	0.45	0.35	0.26	0.66
35							1.00	0.52	0.53	0.23	0.56
40								1.00	0.25	0.56	0.28
45									1.00	0.15	-0.04
50										1.00	0.27
55											1.00

	Market	10	15	20	25	30	35	40	45	50	55
Market	1.000	0.26	-0.27	0.30	-0.20	-0.27	-0.17	0.83	0.10	-0.07	0.53
10		1.00	-0.35	0.05	0.42	-0.23	-0.15	0.47	-0.41	0.36	0.08
15			1.00	0.30	-0.32	-0.06	-0.43	-0.51	0.23	0.22	-0.52
20				1.00	0.10	-0.28	0.02	0.05	-0.03	0.12	-0.14
25					1.00	0.13	0.52	-0.10	-0.47	0.12	-0.05
30						1.00	0.53	-0.21	0.13	0.04	-0.10
35							1.00	-0.13	0.10	0.09	0.18
40								1.00	-0.03	-0.20	0.44
45									1.00	0.14	-0.21
50										1.00	-0.01
55											1.00

Table 3.5: Upper table: Correlation between diameter Middle table:Correlation between average Path length Lower table:Correlation between entropy

3.7.2 Sectors with cohesive structure in the market

The Silhouette coefficient is a measure of both cohesion and separation of clusters. It measures the difference between the average distance of a point to the closest cluster and the average distance of a point to its own cluster. It is one of the most popular index in order not only to identify the number of clusters in the data but also to use as an internal index to quantify the quality of the final clusters or compare with different clustering algorithms. We propose to use the Silhouette coefficient for our case to find out the sectors whose structures are more cohesive and more separable than the other sectors in the hierarchical network.

For each stock x_i we compute its Silhouette coefficient s_i as follows

$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}} \tag{3.7.4}$$

where $\mu_{in}(x_i)$ is the mean distance from x_i to stocks in its own sector $\bar{y}_i$ and $\mu_{out}^{min}(x_i)$ is the mean of the distances from x_i to stocks in the closest sector.

The s_i value of a stock ranges in the interval $[-1, +1]$. A value close $+1$ means that x_i is much closer to stocks in its own sector and is far from the neighboring sector. A value close to zero means that x_i is close to the boundary between its own sector and the nearest sector. A value close to -1 means that x_i is much closer to the neighboring sector than its own sector, and therefore, the stock may be considered as misclustered. Furthermore, the Silhouette coefficient of a sector is defined as the mean s_i value across all the stocks in the same sector and the global Silhouette coefficient is defined as the mean s_i value across all the points:

$$SC = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{3.7.5}$$

A value close to +1 means a good clustering.

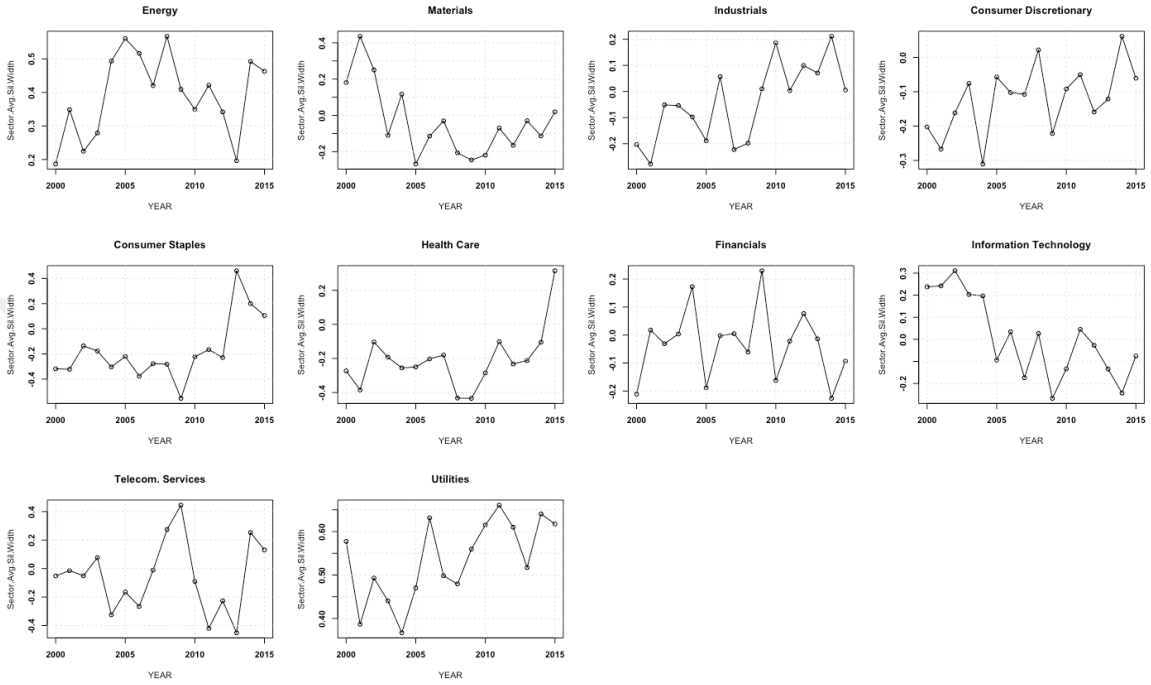

Figure 3.12: The annual Silhouette coefficients of each sector

We plot the Silhouette coefficient of each sector and show how they change annually in Figure 3.12. Energy sector and Utilities sector show clear cohesive and separable structures comparing the other sectors all the time. According to the classic clustering analysis, it means these sectors have well-defined cluster structures and points in these clusters are closer to its own clusters than nearest clusters.

3.7.3 Stock sectoral entropy (SSE)

We propose the sectoral entropy in order to measure the structural uncertainty of sectors in terms of their degree distributions and to track the sectors as for how their topology evolve over time. We compute the entropy of each sector through the edge-induced subgraph approach. A graph G' is called a *subgraph* of a graph G if $V(G') \subseteq V(G)$ and $E(G') \subseteq E(G)$, where the sets V and E are the vertex set and

edge set of G in order. It is denoted as $G' \subseteq G$. Let S be a subset of edges of $E(G)$ and non-empty, then the subgraph of G induced by the set S is called *edge-induced subgraph* of G and denoted as $G[S]$. This induced subgraph has the edge set S and has all vertices which are incident with at least one edge in S [20].

In our case, S is the set of edges incident with stocks from the same sector in the MST. Since all stocks that are incident with at least one edge in the set S are required to be in the induced subgraph, we have stocks from different sectors. Then, we simply calculate the entropy of each sector. We plot the results in figure 3.13.

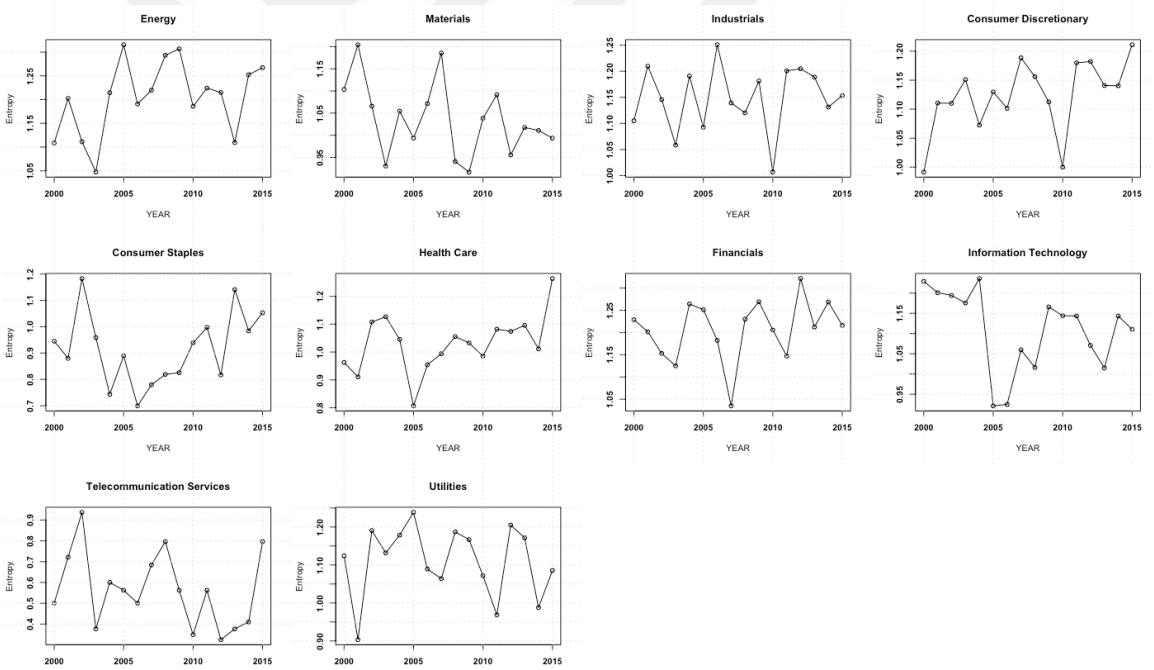

Figure 3.13: Sectoral Entropy

We can see the structural change of the Financials sector during the subprime crisis. Furthermore, in the table 3.5, Financials Sector has the strongest correlation with the market in terms of structural uncertainty and we see some sectors have a weak anti-correlation with the market, such as Materials and Health Care.

3.8 Conclusion

In this work, we analyzed how topological properties of mutual information based stock networks evolve annually from 2000 to 2015 in order to quantify the impact of the major crisis on the network structure by using classic network quantifiers and information theoretic quantifiers. The classic network quantifiers showed similar results found in earlier researches conducted with Pearson' linear correlation, such as shrinkage in the market and increase in the correlation during the crisis. We also observed hub emergence for those periods meaning that some stocks had a substantially high number of connections. Even though the time evolution of the mean of centrality measures show a strong correlation and anti-correlation with each other, they were generally disagreed on the important stocks except for the year 2015. We found that there were no specific stocks stay as important all the time and it always changed.

System Entropy and normalized entropy are introduced as information theoretic quantifiers for market analysis. We observed low values during the dot-com bubble, subprime crisis, and the European debt crisis indicating that the collective behavior of the stocks. Furthermore, we introduced metrics adapted from classic cluster analysis to quantify the structural changes of sectors within the system. We observed that almost all sectors showed decreased in diameter and average path length during the subprime crisis. The impact of the dot-com bubble on Information Technology was much clearly visible with this quantifiers. Average Silhouette width showed that Energy and Utilities are well structure all the time comparing to other sectors. Sectoral Entropy Index (SEI) is proposed to capture the structural randomness of sectors in the system. We compare the sectors with each other and with the market in terms of classic average path length, diameter, and the entropy. Some sectors, such as Finance and Consumer Staples showed very similar behavior with the market; however,

in terms of annual entropy change, we find a strong relationship between the Financial Sector and the market. Degree distribution of stock networks showed the scale-free structure for each year and during the crisis exponent of scaling α had values slightly less than 2.

Chapter 4

Essay 3. Industry Classifications and Identification of Important Industry Groups

Abstract

Researchers in finance, managers, and people interested in the stock market believe that stock prices generally move together. Financial analysts and academic researchers have been using different techniques to construct homogeneous stock groupings; however, one of the most popular among them is the Global Industry Classification System (GICS) as the standard approach. We analyze how the homogeneity in each aggregation level of GICS scheme changes over time and identify the industries which have more homogeneous structure than others in terms of stock returns metric. We propose techniques from complex network theory to illustrate how to construct networks whose nodes are financial industry groups and after that, we investigate the time evolution of the interaction structure of the financial industry groups. Our findings indicate that major local stock networks undergo changes in interaction structure on a regular basis and that such structural changes intensify during a financial crisis. We also show which industry groups dominate the market at particular times and identify which industry groups that experience important changes over various market regimes.

4.1 Introduction

Academic researches and portfolio managers apply a different number of strategies to create homogeneous stock groupings. However, defining which stocks belong to which industry does not have a unique answer. Therefore, a variety of classification

schemes are developed. They can be roughly divided into two types of schemes based on their approach: First is the purely statistical approach with given procedures, such as applying clustering algorithms aiming to partition stocks into similar sets based on predefined properties (return, market capitalization, or operating performance)[19]. The second approach attempts to group economically similar stocks by considering their industry affiliation. Thus, one can view the stock typology itself as feature-based.

In past, a number of industry classification schemes have been introduced. Standard Industrial Classification Codes (SIC) assigns firms into an industry by considering their final products or their production techniques. The Fama and French (1997) classification is based on reorganizing the firms' 4-digit SIC codes into 48 industry groupings. The Global Industry Classification System (GICS) methodology developed by Standard and Poors (S&P) and MSCI/Barra in 1999 has been commonly accepted as an industry analysis framework for investment research, portfolio management, and asset allocation. In this classification, companies are classified based on their principal business activity. Earnings, Revenues, and market perceptions are considered in the classification [86].

The GICS has a four-level hierarchical industry classification scheme. It consists of 10 sectors, 24 industry groups, 67 industries and 147 sub-industries [1]. 8-digit code with text explanation is assigned to each company. For example, sector: Materials (GICS code: 15), industry group: Materials (GICS code: 1510), industry: Chemicals (GICS code: 151010 and sub-industry: Commodity chemicals (GICS code: 15101010).

[1] As of September 1, 2016, S&P Dow Jones Indices and MSCI moved Equity Real Estate Investment Trusts and Real Estate Management & Development companies from the Financials Sector of their Global Industry Classification Standard (GICS) to a new Real Estate Sector. Please visit the website https://www.msci.com/gics.

GICS is the official S&P industry classification system. GICS enables market participants to identify whether stock movements are due to local trends or part of a broader global trend or due to joint effects. Portfolio managers, financial analysts, and academic researchers have been using the (GICS) as a standard benchmark [19]. GICS classification is used also to test the clustering algorithm performance as a ground-truth set. It is also one of the most common reference systems in the application of complex stock networks.

Our first goal in this paper is how 2-digit and 4-digit GICS performs creating homogeneous groups in terms of stock returns. For this goal, we are interested in which aggregated level performs better and within each level which sector and industry group have more homogeneous structure comparing to others. Similar questions have been addressed by [19, 8] by utilizing the Pearson correlation. However, it captures the pairwise relationship very well when the data follows a multivariate normal distribution, and more generally for spherical and elliptical distributions [27]. Also, it only works well if returns are linearly associated and fails to detect any non-linear relationships. However, empirical research in finance shows that the (unconditional) distribution of returns displays a heavy tail with positive excess kurtosis, which is in contrast to the behavior of a normally distributed variable [22, 100]. Furthermore, zero correlation does not imply statistically independent. It only means linear independence and it is possible that there may be some non-linearity.

In order to account both linear and nonlinear associations between stock returns, one needs to describe the inter-dependency in a more general sense than the Pearson correlation. There are several statistical association measures have been proposed based on ranks or information theory [27]. Mutual information, introduced by [93], provides a general measurement for dependencies, such as non-linear or non-functional relationships and is a measure of how much information two systems ex-

change or two data sets share. Furthermore, by using the definition of statistical independence between two random variables, it can be shown that mutual information $I(X;Y) = 0$ if and only if X and Y are independent random variables. Mutual information has been used as co-expression measure to model a complex system such as gene regulatory networks in bioinformatics [17], climate system [31], complex brain networks [6], and recently for stock networks [37].

Our approach is different in several ways comparing to existing literature. First, we use mutual information and analyze the compactness and separation of each industry according to this metric. Furthermore, we identify 4-digits which have a different pattern of behavior regarding annual homogeneity structure in the same sector or under the same 2-digit GICS code. Besides performance of GICS classification, our second objective in this paper is to identify the important industries and how their local interaction changes as time evolves. In literature, this has received less coverage (except [53]). We propose a tool from complex network theory to show the local interaction of industries.

Many complex systems have been analyzed by graph theory. World Wide Web [4], Internet [35], social networks [77], food web [44], scientific citations [90], sexual contacts among individuals [62] and financial systems [56] provide some examples.

Financial markets can be characterized as complex systems because of interaction between heterogeneous components and existing nonlinearity [65]. Network theory can be used to model the financial market in which stocks are represented by nodes [5, 57], commodities [94], currency [36], or banks [14, 26] and links can represent the similarity between stocks, or display the bilateral exposure between any two banks in the system.

In network analysis, centrality indices are developed to answer to the question which are the most important vertices in a network?. Since importance has a variety

of meanings or definitions, there are large number of centrality measures proposed for networks.

Many of the centrality measures were originally introduced in social network analysis and many of the terms used to measure centrality indicate their sociological origin. These methods are also used in large number of different disciplines or areas outside of the social network analysis, such as biology, computer science, urban networks, physics, finance, the Internet, and super-spreaders of disease.

In financial networks, centrality measures are widely applied in the interbank lending market, which is considered as a network where banks are the vertices and the claims and liabilities between banks define the weight of links. The goal is typically to identify the important financial institutions and how their elimination in the network affect the stability of network structure and thus the stability of the banking system [14].

More formally, the centrality can be defined as a function $f : V \to R$, which induces a total order on V. Node v_i is said as central as v_j if $f(v_i) \leq f(v_j)$ [106]. In this work, we use the node degree or the hub as a centrality measure to identify the important industry groups.

Identification of important stocks or central stocks have been done in the literature by using some centrality measures, such as node degree or closeness; however, identifying important industry groups is not easy and have been covered less in the literature. Fiedor proposed to use aggregated Markov centrality in [38]; however, this approach can't help us to understand how the important industries connect with others and how their connection patterns are changing in time.

Our proposed approach is an alternative simplification method of a correlation matrix. Typically, a minimum spanning tree approach is used to reduce the complexity of connection pattern from $n(n-1)/2$ to $n-1$ for n stocks. Constructing

a network whose nodes industries can shed light on the properties of the market at mesoscopic level (known as community or cluster [72, 40, 74, 11]), in which each node represents an industry composed of variety number of stocks.

The remainder of this paper is organized as follows. In Section 4.2, we discuss the data and methodology. In section 4.3.1, we compare the classification systems and find out which industries have more homogeneous group in terms of stock returns. In section 4.3.2, we analyze how the homogeneity of the sectors change in time and compare each other. In section 4.3.3, similar to sector, how the homogeneity of industry groups change in time and compare them to their peers under same sector. In section 4.3.4 we investigate evolution of local interaction of industry groups and identify the important industries over time. In Section 4.3.5 we summarize our results and propose some future studies.

4.2 Materials and Methodology

4.2.1 Data

In this study, we have used the data set composed of log-returns of daily adjusted close prices of 404 stocks of SP500 during 4032 consecutive trading days in the time period from January 2000 to December 2015 after the removal of a few days because of the incomplete data.

4.2.2 Dissimilarity measures

The stocks in the same group are expected to have similar comovement in their returns and less similarity with the stocks outside of their sector. To quantify the homogeneity of industry classifications, we utilize the notions from internal cluster validity indices. Internal indices assume that there are no preassigned labels for points in the data

and they are based on the $n \times n$ distance matrix, also known as the proximity matrix, denoted by $\mathbf{W}$, of all pairwise distances among the n points.

In this work, in order to obtain $\mathbf{W}$ from mutual information based adjacency matrix, we follow similar steps described in previous essays. The equal-width approach is applied to discretize the stock returns with the default number of bins selected as $\sqrt{n}$. After that, we use the Miller-Madow estimation technique to compute the pairwise similarity between stocks. Finally, the similarity matrix is transformed into a distance matrix by using universal distance metric version 2, i.e $dissAUV2$. Please see Section-2.3.2 for more details.

4.2.3 Internal indices

Internal indices are often applied to find out the number of clusters and choose the proper clustering algorithm in case no external labels are available in the dataset. Since the goal of the clustering algorithms is to partition the data into groups of objects by optimizing the average distance within cluster to be as low as possible and average distance between clusters to be as large as possible, they measure the compactness, that is how close the objects inside the same cluster and separation, that is, how far the clusters from each other. Therefore, a clustering algorithm is an optimization problem of the minimization-maximization type (a min-max problem). There are several validity indices and statistics that have been proposed to assess the quality of clusters.

Let us consider the proximity matrix $\mathbf{W}$ as the adjacency matrix of the weighted complete graph G over the n objects, in which the nodes are stocks and edges are defined as pairwise distance between them. Therefore, there is close connection between graph clustering and internal indices [105]. Let us assume that we are given as industry classification system (such as 2-digit codes) $I = \{I_1, ..., I_k\}$ with industry I_i

containing $n_i = |I_i|$ stocks. Industry classification scheme can be considered as k-way cut in G because $I_i \neq \emptyset$ for all i, $I_i \cap I_j = \emptyset$ for all i, j and $\cup_i I_i = V$. For given any subsets $K, L \subset V$, we can compute sum of the weights on all links between this two sets as follows:

$$W(K, L) = \sum_{x_i \in K} \sum_{x_j \in L} w_{ij} \tag{4.2.1}$$

Let us denote $\overline{K}$ for given $K \subseteq V$ as the complementary set of nodes, i.e $\overline{K} = V - K$. We can define the sum of all the intraindustry and interindustry weights as follows:

$$W_{in} = \frac{1}{2} \sum_{i=1}^{k} W(I_i, I_i) \tag{4.2.2}$$

$$W_{out} = \frac{1}{2} \sum_{i=1}^{k} W(I_i, \overline{I_i}) = \sum_{i=1}^{k-1} \sum_{j>i} W(I_i, I_i) \tag{4.2.3}$$

Therefore, the average of all the intraindustry and interindustry weights is given as:

$$\overline{W_{in}} = \frac{W_{in}}{N_{in}} \tag{4.2.4}$$

$$\overline{W_{out}} = \frac{W_{out}}{N_{out}} \tag{4.2.5}$$

, where N_{in} is the number of different intraindustry links and N_{out} is the number of different interindustry links. They are defined as follows:

$$N_{in} = \sum_{i=1}^{k} \binom{n_i}{2} \tag{4.2.6}$$

$$N_{out} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_i n_j \tag{4.2.7}$$

It is clear that the total number of links in G equal to $N = N_{in} + N_{out} = \binom{n}{2}$

By comparing the average of all the intraindustry with respect to the average of all interindustry weights, we can measure the compactness of industries, that is how similar the stocks inside the same industry and separation, that is, how far the industries from each other. Therefore, the larger the difference between the average of all the interindustry and intraindustry weights means that industry has well-separated homogeneous structure.

4.2.4 Networks whose nodes are sectors or industry groups

Here we outline how to build a network among sectors and industry groups, i.e. each vertex in the network corresponds to a sector or industry groups. However, this method can be used to construct a network among modules (clusters) as well. Let us denote I_{q_1} the set of n^{q_1} stocks inside sector q_1. The distance between stocks of two sectors can be defined by an $n^{q_1} x n^{q_2}$ as submatrix $W^(q_1, q_2)$ of the distance matrix. In order to calculate distance between sectors, we define the matrix W^{q_1,q_2} by a number between 0 and 1:

$$W^{avg}_{q_1,q_2} = mean(W^{(q_1,q_2)}) = \frac{\sum_{i \in S_{q_1}} \sum_{j \in S_{q_2}} W_{ij}}{n^{(q_1)} n^{(q2)}} \tag{4.2.8}$$

$$W^{max}_{q_1,q_2} = max(W^{(q_1,q_2)}) = \max_{i \in S_{q_1}, j \in S_{q_2}} W_{ij} \tag{4.2.9}$$

$$W^{min}_{q_1,q_2} = min(W^{(q_1,q_2)}) = \min_{i \in S_{q_1}, j \in S_{q_2}} W_{ij} \tag{4.2.10}$$

These measures are used in the computation of the proximity between two clusters in the various agglomerative hierarchical techniques. In our case we use W^{avg}, since it is statistically more robust than other two; however, they can also be used in applications. The average distance measures between sectors defined above can be used to construct a network between sectors, e.g.,

$$W_{q_1,q_2} = \begin{cases} W^{avg}_{q_1,q_2} & \text{if } q_1 \neq q_2 \\ 0 & \text{if } q_1 = q_2 \end{cases} \tag{4.2.11}$$

Let us denote $W_{sectors}$ as the QxQ dimensional symmetric matrix whose q_1, q_2 element is given by $W^{avg}_{q_1,q_2}$ which measures the distance between two sectors. In order to identify the important sector (an industry group), we filter the complete weighted networks defined by $W_{sectors}$ through minimum spanning tree (MST) in order to extract the most relevant information from the distance matrix between sectors. Depending on the application, a complete network can also be used and important sector(node) can be defined simply taking row(or column) weight of it. However, since we are interested in sectors local interactions between each other, MST approach better fits in our filtration task. MST stock network generation is one of the popular methods in financial network literature and is introduced by Mantegna [64]. The construction process is defined as follows: we start with an arbitrary stock as the root of a partial tree and at every step, the partial tree grows by iteratively adding an unconnected stock to it by choosing the lowest weight, until the unconnected stock set is exhausted, which is known as Prims algorithm [87]. MST of order N has exactly N-1 links and no loops or circuits. Furthermore, MST has a strong relationship with a single linkage clustering algorithm [51].

In summary, by using defined average distance between sectors and industry

groups, we generate annual MST industry groups networks in order to identify central nodes defined as a hub in each network and investigate the change of their time-dependent-connection pattern structure over time.

4.3 Results

4.3.1 Average distances

Sector Number	Inside the industry		Between the Industry		Difference		No of Firms
	GICS4	GICS2	GICS4	GICS2	GICS4	GICS2	
Energy (10)	0.827	0.827	0.905	0.905	0.078	0.078	32
Materials (15)	0.869	0.869	0.894	0.894	0.025	0.025	24
Industrials (20)	0.861	0.867	0.892	0.892	0.031	0.025	56
Consumer disc (25)	0.877	0.891	0.898	0.901	0.021	0.01	60
Consumer staples (30)	0.888	0.904	0.909	0.911	0.021	0.007	32
Health Care (35)	0.896	0.9	0.907	0.908	0.012	0.008	45
Financials (40)	0.819	0.856	0.89	0.894	0.071	0.038	72
Information tech (45)	0.866	0.88	0.901	0.903	0.035	0.023	51
Telecom. services (50)	0.886	0.886	0.908	0.908	0.022	0.022	5
Utilities (55)	0.825	0.825	0.908	0.908	0.083	0.083	27
Average	0.861	0.871	0.901	0.902	0.04	0.032	total=404

Table 4.1: Average pairwise distances between each stock return and stocks inside the industry and outside the industry

After splitting our data set into non-overlapping one year period, we assign each stock to a sector and an industry group. For each period, the intraindustry and interindustry weights are computed according to the given scheme, and the results are then averaged over 16 periods. Table 4.1 displays the average distances inside an industry and between the industries. We report our results in the table according to 2-digits GICS scheme only in order to have clear table formatting. At the bottom of the table, the simple average over 16 sub-periods is presented for each classification level.

By definition of the clustering, the average distance between objects in the same cluster should be less than average distances with the objects outside the cluster. Therefore, it is clear that 2-digits GICS and 4-digits GICS scheme are successful in creating industries of homogeneous stocks. From the table, it is evident that there is a slight difference between the two coding schemes, the average differences are 0.032 for 2-digit GICS codes and 0.04 for 4-digit GICS codes.

For 2-digit GICS codes, to identify the more homogeneous sectors, we investigate the average difference for each sector. Utilities, Energy and Financial sectors have more homogeneous structure than other sectors and their average differences are 0.083, 0.078 and 0.038, respectively. Therefore, for the stocks in these sectors, lines of business are well-defined and uniform [19]. For 4-digit codes, the same results hold; however, we see some sectors are getting more homogeneous. For example, Financial sectors under 2-digit codes have average difference 0.038 but under the 4-digit code, it has 0.071 with a big jump.

In Table 4.2, we show the average distances inside an industry and outside the industry for each sub-periods from 2000 to 2015. For each year, 4-digits GICS groups have slightly had more homogeneous industries, which supports our earlier result in Table 4.1. It is clear that there a little improvement from going to 2-digit GICS codes to 4-digit GICS codes in terms of average annual differences. Furthermore, in [18, 19], they found some evidence that the average stock's correlation with other stocks decreased in time. However, our findings are inconsistent with them. We have found that the average distances between industries have decreased, i.e, meaning that average correlation outside of the industry has increased over time. For example, in the year 2000, the average distance between industries is 0.935 in the year 2000 for 2-digit GICS codes and for same scheme, it decreases 0.893. It is clear that similar results hold for 4-digit GICS codes as well.

Sample Period	Inside the industry		Between the Industry		Difference	
	GICS4	GICS2	GICS4	GICS2	GICS4	GICS2
2000	0.912	0.92	0.934	0.935	0.022	0.015
2001	0.892	0.903	0.925	0.926	0.033	0.023
2002	0.871	0.883	0.909	0.911	0.038	0.027
2003	0.885	0.892	0.913	0.914	0.028	0.022
2004	0.891	0.906	0.927	0.928	0.036	0.022
2005	0.886	0.905	0.927	0.928	0.041	0.023
2006	0.892	0.908	0.929	0.929	0.036	0.021
2007	0.861	0.874	0.899	0.901	0.038	0.027
2008	0.787	0.804	0.845	0.847	0.058	0.043
2009	0.819	0.833	0.872	0.874	0.053	0.041
2010	0.826	0.84	0.868	0.87	0.042	0.03
2011	0.784	0.797	0.833	0.835	0.049	0.038
2012	0.867	0.879	0.905	0.907	0.038	0.028
2013	0.869	0.884	0.907	0.908	0.038	0.024
2014	0.852	0.873	0.907	0.908	0.055	0.036
2015	0.831	0.852	0.891	0.893	0.06	0.042

Table 4.2: Annual average pairwise distances between each stock return and stocks inside the industry and outside the industry

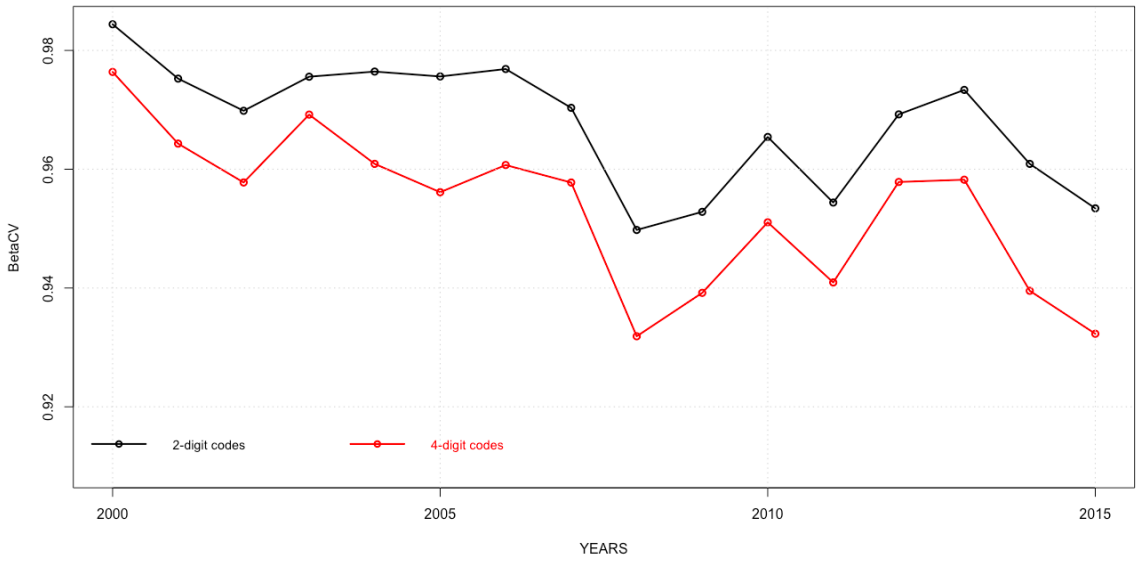

Figure 4.1: Annual BetaCV values for sectors and industry groups

To get better insight of how the average distances inside and outside industries evolve in time, we propose to use the measure known as BetaCV, which is simply the ratio of average distance within industries to the average distance between industries and formulated as follows:

$$BetaCV = \frac{W_{in}/N_{in}}{W_{out}/N_{out}} \tag{4.3.1}$$

The small value indicates the better clustering [105]. In Figure 4.1, it is evident again that 4-digits GICS groups have slightly have more homogeneous industries than 2-digits GICS groups. Furthermore, we can see the clear impact of sub-prime mortgage crisis on the homogeneity structure on two classification schemes. It also displays the ratio of average distance within industries to the average distance between industries has declined over time for two schemes.

4.3.2 Sector

Now, we are interested in how the average differences between interweights and intraweights change annually for each sector and which ones successfully maintain to keep this difference large in time.

Figure 4.2 reports how the average differences of 2-digits GICS code change annually and average over 16-periods are shown at the top. Firms in the same business lines are expected to have higher correlations than the firms outside of their business lines. Therefore, sectors consisting of well-defined firms should have a higher average difference, i.e. more homogeneous groups. Furthermore, well-defined sector structure is relevant for managers to avoid tracking error [19].

Figure 4.2 supports our previous results in Table 4.1. Utilities, Energy and Financial sectors maintain more homogeneous groups than other sectors; however, it

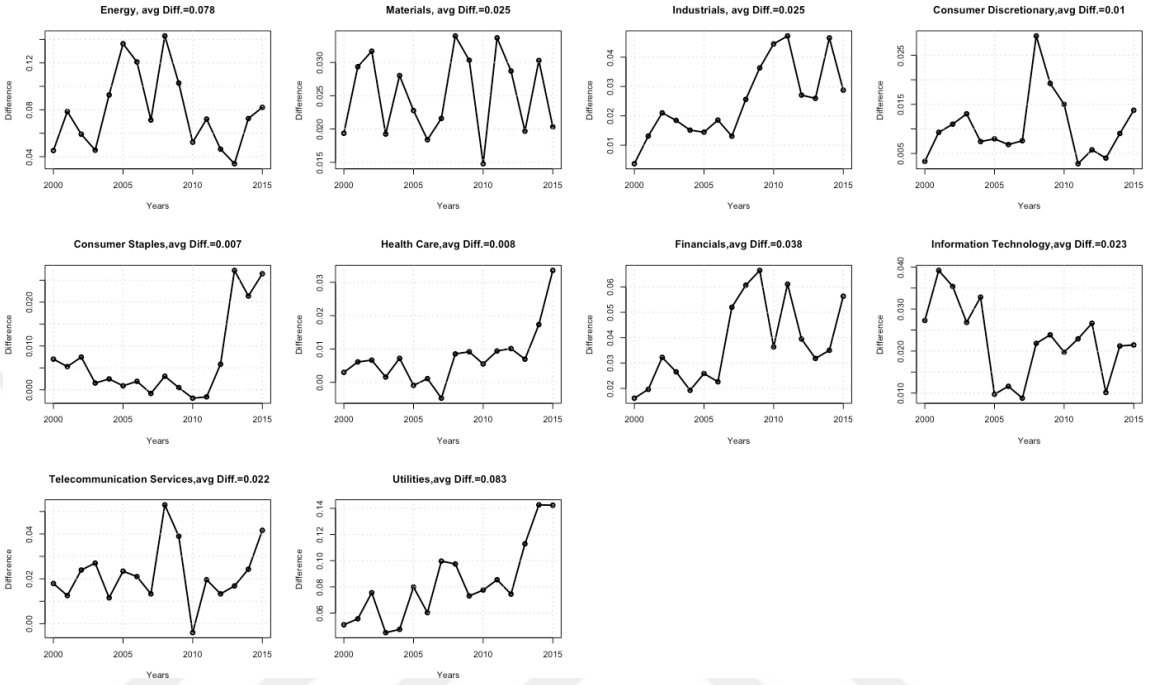

Figure 4.2: Annual average differences for 2-digits

is evident that it changes by time, i.e., some sectors have higher average differences in some period but less in other periods. For example, Consumer Discretionary (25) have a less homogeneous structure on average among all sectors but in the year 2008 average difference jumps to 0.029.

4.3.3 Industry groups

In previous sections, we have found that 4-digits GICS codes have slightly more homogeneous industries than 2-digits GICS codes. Similar to the previous section, we are interested in how this homogeneous structure for 4-digits GICS groups evolve annually and if these industry groups under the same sectors show a similar pattern of change in time. Average differences for each industry groups are computed by taking difference average distance outside and average distance inside the same group as usual. Note that some sectors do not have separate 4-digits GICS groups,

such as Energy (10), Materials (15), Telecommunication Services (50) and Utilities (55); therefore their results are identical with the Figure-4.2 in the previous section. Furthermore, for space concern, we only illustrate the results for average differences, dropping the figures for the annual change of average distance inside and outside for industries.

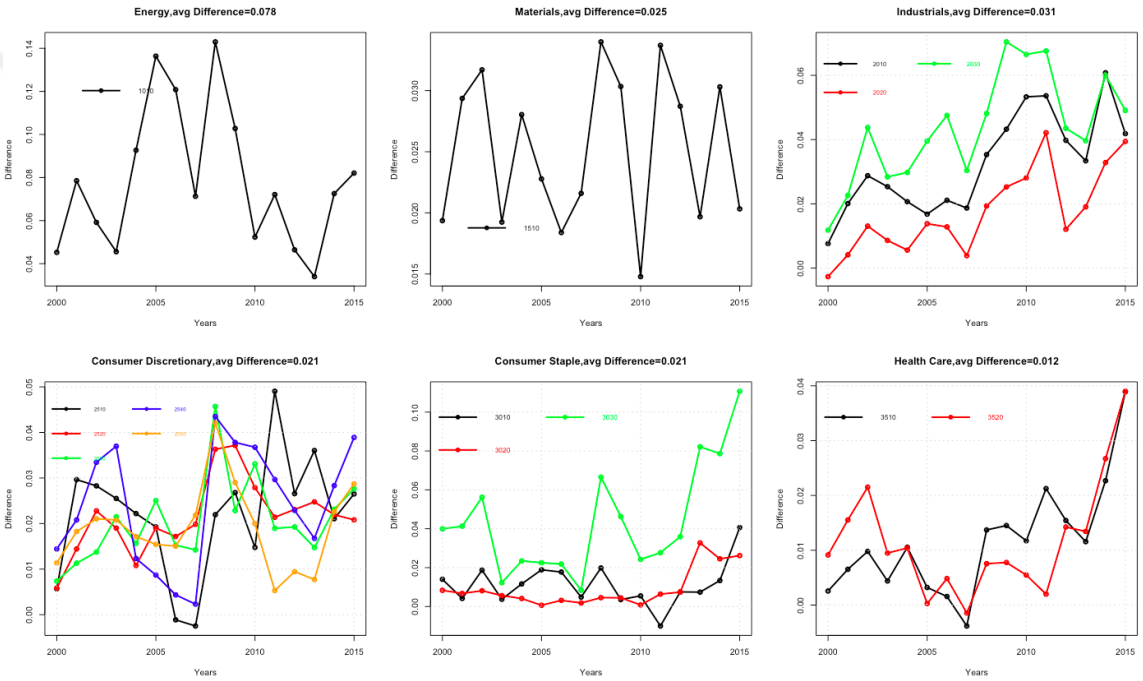

Figure 4.3: Annual average differences of industry groups for Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, and Health Care sectors.

Figure-4.3 and Figure-4.4 displays the results for each industry groups and results are shown together under their sectors which they belong in order to make the comparison clearly.

From the figures, we can clearly see that some of the 4-digits GICS groups are a little more homogeneous than other industry groups under their same sector and furthermore some display very different pattern of annual change comparing to their peers. For example, Semiconductors & Semiconductor Equipment (4530) does not

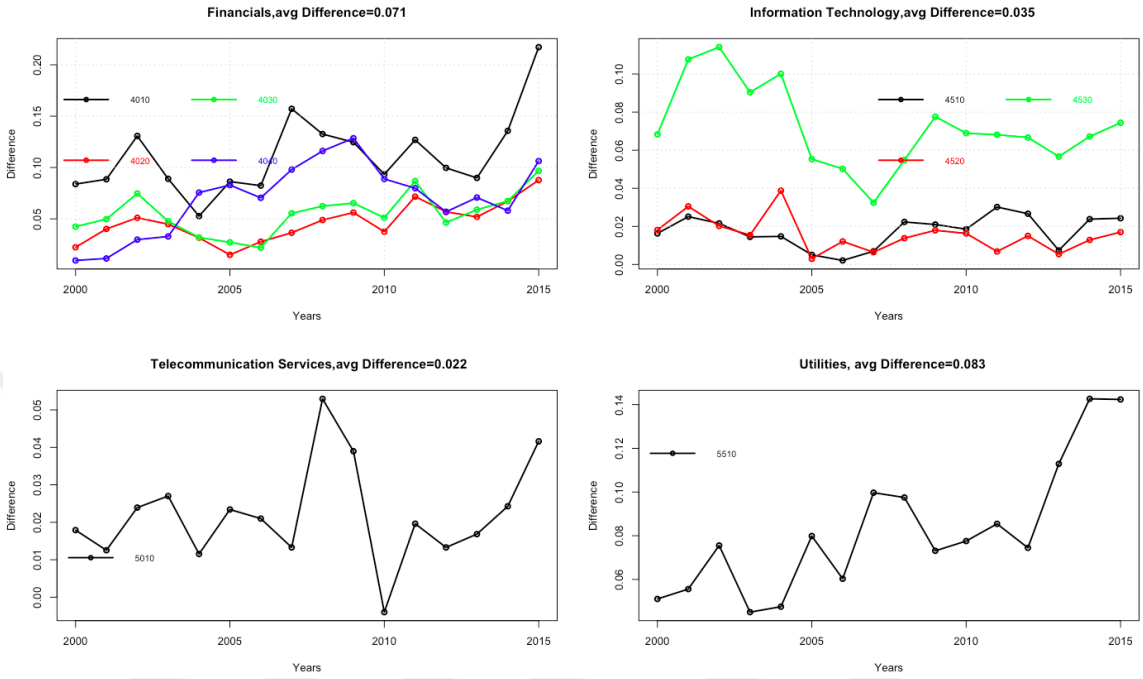

Figure 4.4: Annual average differences of industry groups for Financials, Information Technology, Telecommunication Services, and Utilities sectors.

only have more coherent structure but also exhibits little different pattern change over time comparing to Software & Services (4510) and Technology Hardware & Equipment (4520), which both have similar patterns.

In Financials sector, Banks (4010) and Real Estate (4040) are different from diversified Financials (4020) and Insurance (4030) industry groups. Real Estate (4040) have an interesting pattern: starting from the early 2000s it starts getting increasingly more coherent structure till the end of 2009 and after that, it starts decreasing till 2014. That shows the impact of the most recent subprime crisis on the Real Estate market. Furthermore, 4010 Banking shows also different coherent structure. It also starts dropping from 2007 to 2010. In Consumer Staples sector Household& Personal Products (3030) have more homogeneous groupings.

4.3.4 Time evolution of interaction structure of industry groups

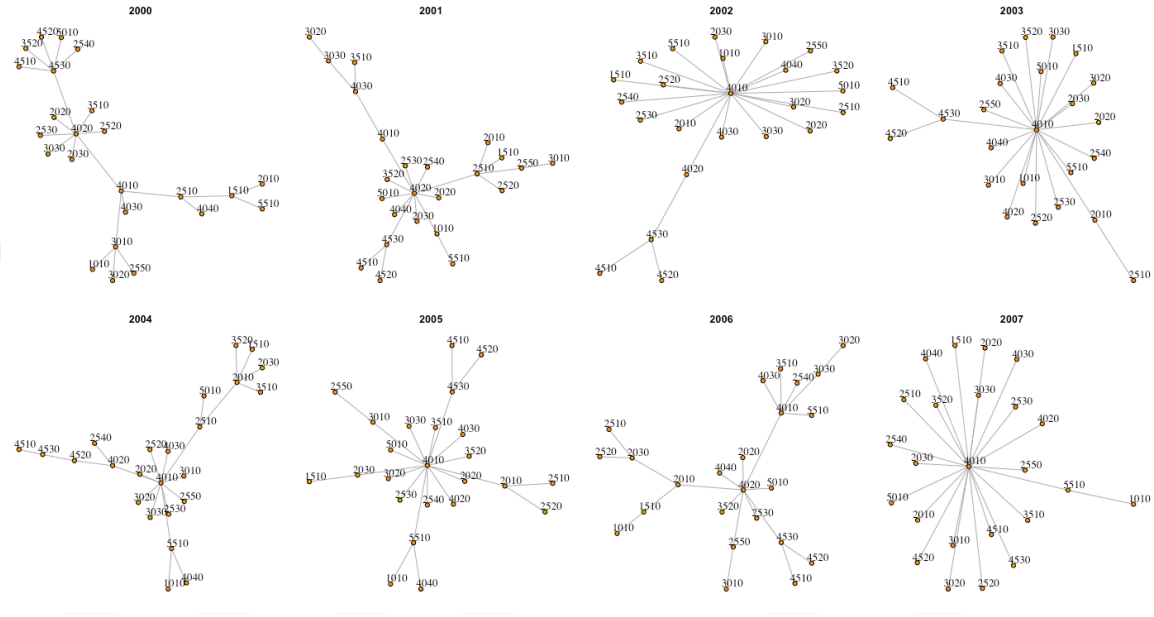

Figure 4.5: Annual evolution of interaction structure of industry groups between 2000 and 2007

We now investigate how the industry group interaction structure evolves over time by using the MST approach constructed from the distance matrix of the average pairwise distances between industry groups. For the SP500 index, the time period of daily data of the 404 stocks is divided into 16 annual time windows from January 2000 till the end of December 2015 and for each time window, we created MST of industry groups to investigate their local interaction and identify the important industry groups in the market. Nodes in the network are labeled by the 4-digit GICS codes for readability. Please see the Table-6.1 for 4-digit GICS codes and their associated industry groups.

The local interaction structure of the industry groups is displayed for the annual time windows in Figure-4.5 and in Figure-4.6. It is clear from the figures that these local interaction structures change with the time and especially before, during

and after the financial crisis. This result is agreed with the result of [108]. During the early 2000s recession (period 2002 and 2003) and during the subprime mortgage crisis industry group network structure becomes very simple star shape network comparing the other time periods.

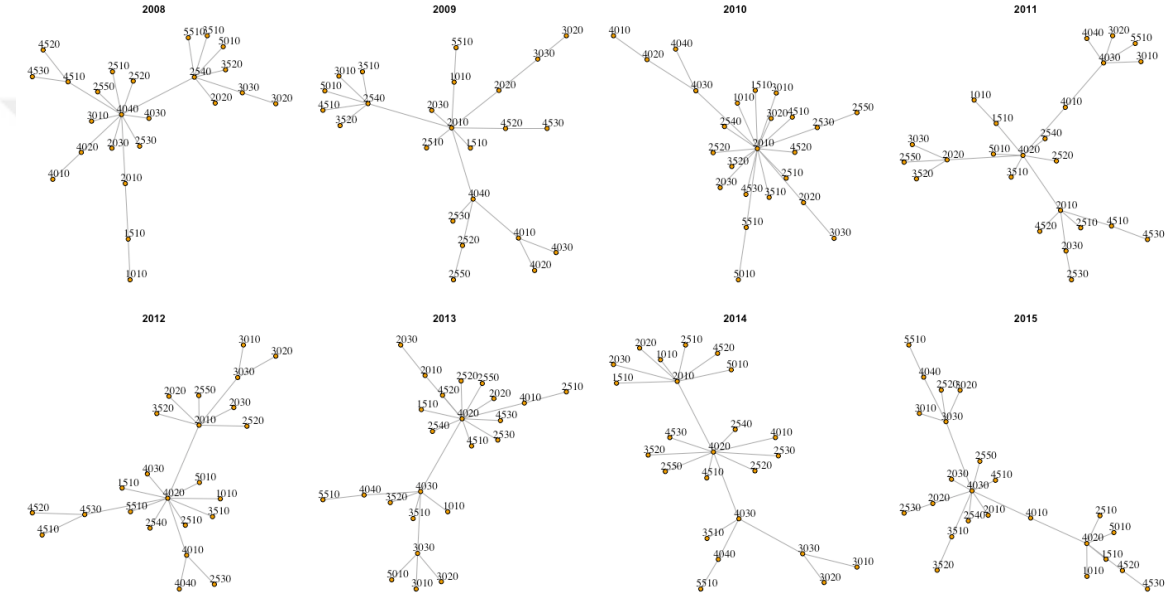

Figure 4.6: Annual evolution of interaction structure of industry groups between 2008 and 2015

Next, we are interested in which are the important industry groups (hubs) in the market and how these dominant industry groups change in time. From the figures, it is obvious that during the early 2000s recession (period 2002 and 2003) and during the subprime mortgage crisis, 4010-Banks is the dominant industry group in the simple structured networks. During the Dotcom bubble, we observe besides the financial industry groups (4010 and 4020), one of the IT industry group (4510-Semiconductors & Semiconductor Equipment) becomes dominant and plays the role as a bridge between IT industry groups and rest of the market. After the crisis it does not become as important as it was in the crisis period; however, most of the

time it still stays as a bridge between the IT sector and market.

Overall, the market is dominated mostly by financial industry groups. In the year 2007, the beginning of the subprime crisis, we observe 4010 becomes the most dominant industry group. However, during 2008 and 2009, 4040-Real estate industry group becomes more important in the market. In this time period, we observe new emerging important industry groups, such as 2540-Media and 2010-Capital Goods, which becomes the most dominant industry group in the year 2010 and maintains it for some periods. Interestingly, we do not observe the industry group 4010 as one of the central industry groups after the subprime crisis; however, 4020 still stays important for the following years and 4030 emerges as important. Until the subprime crisis, the market is mostly dominated by two industry groups 4010 and 4020, during the crisis, in the year 2007 it becomes only 4010 and in the years of 2008 and 2009, we see 4040 as well. Even though after the crisis we do not see 4010 anymore, the market still gets dominated by financial industry groups 4020 and 4030 and plus industry group 2010. Starting from the year 2010, we observe 4030-Insurance industry group becomes one of the important industry group. IT industry groups mostly clustered together for considered time periods except for the year 2007.

In summary, the local interaction structure between industry groups is highly time-dependent and changes dramatically in the crisis periods. Specifically, the structure becomes simple or star shape and industry groups associated with the crisis becomes the most dominant industry group in the market.

4.4 Conclusion

In this paper, we find that Utilities, Energy and Financial sectors have stronger spread between within-industry correlations and outside-industry correlations comparing to

other sectors. Furthermore, we show that finer levels of industry partitioning have a slightly higher average difference, i.e moving from 4-digit GICS codes to 2-digit GICS code has little benefit in terms of grouping the common return movements of stocks inside an industry compared to the common return movements of stocks outside the industry. Homogeneity of industry groups is compared with their peer industry groups under their main sector. Some industry groups are found to have more uniform groupings, for example, Semiconductors & Semiconductor Equipment (4530) in IT (45) sector.

Furthermore, we analyze how the local interaction structure of the industry groups change over time. We find that the market has been dominated by a couple of industry groups. We also show which industry groups dominate the market at particular times. We identify the role of the particular industry in the market during the crisis and find that the industry groups under stress became more central nodes. Moreover, the structure becomes star network during the crisis.

Chapter 5

Conclusion and Future Work

5.1 Summary

In the first essay, we provide a theoretical background for mutual information and furthermore illustrate how to construct stock networks by using mutual information metrics. Furthermore, we compare the Pearson correlation and mutual information in order to characterize the nonlinearity in stock returns and how substantially the networks constructed by two measures are different from each other on local and global topological scales. We find that these networks are very different and in crisis period this difference significantly increases. Performance of edge-betweenness community detection algorithms is tested on the networks constructed by six different distance measures and we find that communities identified on mutual information networks have more compact and well-separated structures than other distance measures. The relationship between entropy and the first four moments of stock return distributions are analyzed and we find that kurtosis and entropy have very high negative Spearman correlation in crisis and non-crisis periods. That indicates one of the main reason of mismatch measures between mutual information and Pearson correlation. We further investigate the cases for the stocks with extra-normal information and find that higher order of polynomial regression performs better than the simple regression for their relationship with certain stocks. Spline regression and polynomial regression based stock networks are illustrated and they confirm also the existence of nonlinear dependency between stock returns.

In the second essay, we studied how the local, global, and mesoscopic topolog-

ical properties of mutual information based stock networks evolve annually between 2000 to 2015 in order to quantify the impact of the major crisis on the network structure by using classic network quantifiers and information theoretic quantifiers. Our findings related to classic quantifiers are similar to earlier findings; such as such as shrinkage in the market and an increase in the correlation during the crisis. Moreover, we observe lower entropy values in crisis periods indicating that market structure becomes less random than normal periods. Degree distribution of stock networks presented scale-free structure for the non-crisis periods; however, during the crisis exponent of scaling α had values slightly less than 2. Some metrics related market mesoscopic structure is proposed, such as Sectoral Entropy Index (SEI) to capture the structural randomness of sectors in the system. Further, structural change of the sectors within the system over time is quantified and similarity between each other and market are measured. We find a strong relationship between the Financial Sector and the market.

In the third essay, we have a couple aims to search for. First, we show which sectors and industry groups have more homogeneous grouping in terms of stock returns and how they change over time. Second, moving from 4-digit GICS codes to 2-digit GICS code has little benefit in terms of grouping the common return movements of stocks inside an industry compared to the common return movements of stocks outside the industry. In addition to the homogeneity, we illustrate a technique to construct a network whose nodes are sector or industry groups in order to investigate how their local interaction change over time and which industry groups dominate the market at particular times. We find that during the crisis only one single industry group becomes a hub and market network structure becomes very simple.

5.2 Future Research

Our work is limited to constructing the undirected and weighted networks for daily stock returns. The work could be extended in many directions and applications. Firstly, we could use general prediction and machine learning methods for measuring non-linear relationships. Secondly, our findings can be tested on different financial markets, such as the commodity markets and the Forex market whether there is strong disagreement between the Pearson correlation and mutual information. Furthermore, the role of the frequency of the data and nonlinearity can be investigated. Finally, we look for the other co-expression measures for nonlinearity, which does not require the discretization.

Chapter 6

Appendix

	Sector		Industry.Group	
1	10	Energy	1010	Energy
2	15	Materials	1510	Materials
3	20	Industrials	2010	Capital Goods
4			2020	Commercial Services & Supplies
5			2030	Transportation
6	25	Consumer Discretionary	2510	Automobiles & Components
7			2520	Consumer Durables & Apparel
8			2530	Consumer Services
9			2540	Media
10			2550	Retailing
11	30	Consumer Staples	3010	Food & Staples Retailing
12			3020	Food,Beverage & Tobacco
13			3030	Household & Personal Products
14	35	Health Care	3510	Health care Equipment&Services
15			3520	Pharmaceuticals, Biotechnology & Life Sciences
16	40	Financials	4010	Banks
17			4020	Diversified Financials
18			4030	Insurance
19			4040	Real Estate
20	45	Information Technology	4510	Software & Services
21			4520	Technology Hardware & Equipment
22			4530	Semiconductors & Semiconductor Equipment
23	50	Telecommunication Services	5010	Telecommunication Services
24	55	Utilities	5510	Utilities

Table 6.1: The GICS structure

Bibliography

[1] Frédéric Abergel, Bikas K Chakrabarti, Anirban Chakraborti, and Asim Ghosh. *Econophysics of systemic risk and network dynamics*. Springer, 2013.

[2] Abhay Abhyankar, Laurence Sidney Copeland, and Woon Wong. Nonlinear dynamics in real-time equity market indices: evidence from the united kingdom. *The Economic Journal*, pages 864–880, 1995.

[3] Abhay Abhyankar, Laurence S Copeland, and Woon Wong. Uncovering nonlinear structure in real-time stock-market indexes: the s&p 500, the dax, the nikkei 225, and the ftse-100. *Journal of Business & Economic Statistics*, 15(1): 1–14, 1997.

[4] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *nature*, 401(6749):130–131, 1999.

[5] Serkan Alkan and Khaldoun Khashanah. Structural evolution of stock networks. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2015 11th International Conference on*, pages 406–412. IEEE, 2015.

[6] Chris G Antonopoulos, Shambhavi Srivastava, Sandro E de S Pinto, and Murilo S Baptista. Do brain networks evolve by maximizing their information flow capacity? *PLOS computational biology*, 11(8):e1004372, 2015.

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

[8] Sanjeev Bhojraj, Charles MC Lee, and Derek K Oler. What's my line? a comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, 41(5):745–774, 2003.

[9] Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, 104(3):535–559, 2012.

[10] Dimitrios Bisias, Mark Flood, Andrew W Lo, and Stavros Valavanis. A survey of systemic risk analytics. *Annu. Rev. Financ. Econ.*, 4(1):255–296, 2012.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[12] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424 (4-5):175–308, 2006.

[13] Giovanni Bonanno, Fabrizio Lillo, and Rosario N Mantegna. High-frequency cross-correlation in a set of stocks. 2001.

[14] Michael Boss, Helmut Elsinger, Martin Summer, and Stefan Thurner 4. Network topology of the interbank market. *Quantitative Finance*, 4(6):677–684, 2004.

[15] William A Brock and W Davis Dechert. Non-linear dynamical systems: instability and chaos in economics. *Handbook of mathematical economics*, 4:2209–2235,

1991.

[16] Christopher Brooks. Testing for non-linearity in daily sterling exchange rates. *Applied Financial Economics*, 6(4):307–317, 1996.

[17] Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific, 1999.

[18] John Y Campbell, Martin Lettau, Burton G Malkiel, and Yexiao Xu. Have individual stocks become more volatile? an empirical exploration of idiosyncratic risk. *The Journal of Finance*, 56(1):1–43, 2001.

[19] Louis KC Chan, Josef Lakonishok, and Bhaskaran Swaminathan. Industry classifications and return comovement. *Financial Analysts Journal*, 63(6):56–70, 2007.

[20] Gary Chartrand and Ping Zhang. *A first course in graph theory*. Courier Corporation, 2013.

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

[22] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. 2001.

[23] Thomas M Cover and Joy A Thomas. Elements of information theory 2nd edition (wiley series in telecommunications and signal processing). 2006.

[24] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

[25] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[26] Giulia De Masi, Yoshi Fujiwara, Mauro Gallegati, Bruce Greenwald, and Joseph E Stiglitz. An analysis of the japanese credit network. *Evolutionary and Institutional Economics Review*, 7(2):209–232, 2011.

[27] Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, Asuka Nakata, and André Fujita. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in bioinformatics*, 15 (6):906–918, 2013.

[28] Tiziana Di Matteo, Tomaso Aste, and Rosario Nunzio Mantegna. An interest rates cluster analysis. *Physica A: Statistical Mechanics and its Applications*, 339(1):181–188, 2004.

[29] Andreia Dionisio, Rui Menezes, and Diana A Mendes. An econophysics approach to analyse uncertainty in financial markets: an application to the portuguese stock market. *The European Physical Journal B-Condensed Matter and Complex Systems*, 50(1-2):161–164, 2006.

[30] Andreia Dionisio, Rui Menezes, and Diana A Mendes. Entropy and uncertainty analysis in financial markets. *arXiv preprint arXiv:0709.0668*, 2007.

[31] Jonathan F Donges, Yong Zou, Norbert Marwan, and Jürgen Kurths. Complex

networks in climate dynamics. *The European Physical Journal-Special Topics*, 174(1):157–179, 2009.

[32] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

[33] Nader Ebrahimi, Esfandiar Maasoumi, and Ehsan S Soofi. Ordering univariate distributions by entropy and variance. *Journal of Econometrics*, 90(2):317–336, 1999.

[34] P ERDdS and A R&WI. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[35] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM, 1999.

[36] Daniel J Fenn, Mason A Porter, Peter J Mucha, Mark McDonald, Stacy Williams, Neil F Johnson, and Nick S Jones. Dynamical clustering of exchange rates. *Quantitative Finance*, 12(10):1493–1520, 2012.

[37] Paweł Fiedor. Mutual information rate-based networks in financial markets. *arXiv preprint arXiv:1401.2548*, 2014.

[38] Paweł Fiedor. Sector strength and efficiency on developed and emerging financial markets. *Physica A: Statistical Mechanics and its Applications*, 413: 180–188, 2014.

[39] Pasquale Foggia, Gennaro Percannella, Carlo Sansone, and Mario Vento. Benchmarking graph-based clustering algorithms. *Image and Vision Computing*, 27 (7):979–988, 2009.

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

[41] Philip Hans Franses and Dick Van Dijk. Forecasting stock market volatility using (non-linear) garch models. *Journal of Forecasting*, 15(3):229–235, 1996.

[42] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[43] Francis Galton. I. co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45(273-279):135–145, 1889.

[44] Diego Garlaschelli, Guido Caldarelli, and Luciano Pietronero. Universal scaling relations in food webs. *Nature*, 423(6936):165–168, 2003.

[45] David Hartman and Jaroslav Hlinka. Nonlinearity in stock networks. *arXiv preprint arXiv:1804.10264*, 2018.

[46] Jean Hausser et al. *Improving entropy estimation and the inference of genetic regulatory networks*. PhD thesis, Citeseer, 2006.

[47] Raphael H Heiberger. Stock network stability in times of crisis. *Physica A: Statistical Mechanics and its Applications*, 393:376–381, 2014.

[48] Raphael H Heiberger. Predicting economic growth with stock networks. *Physica A: Statistical Mechanics and its Applications*, 489:102–111, 2018.

[49] David A Hsieh. Testing for nonlinear dependence in daily foreign exchange rates. *Journal of Business*, pages 339–368, 1989.

[50] Lawrence Hubert and James Schultz. Quadratic assignment as a general data analysis strategy. *British journal of mathematical and statistical psychology*, 29 (2):190–241, 1976.

[51] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[52] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[53] XF Jiang, TT Chen, and B Zheng. Structure of local interactions in complex financial dynamics. *Scientific reports*, 4:5321, 2014.

[54] Piotr Juszczak. Learning to recognise: A study on one-class classification and active learning. 2006.

[55] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[56] Khaldoun Khashanah and Linyan Miao. Dynamic structure of the us financial systems. *Studies in Economics and Finance*, 28(4):321–339, 2011.

[57] Khaldoun Khashanah and Hanchao Yang. Evolutionary systemic risk: Fisher information flow metric in financial network dynamics. *Physica A: Statistical Mechanics and its Applications*, 445:318–327, 2016.

[58] Minjun Kim and Hiroki Sayama. Predicting stock market movements using network science: an information theoretic approach. *Applied Network Science*, 2(1):35, 2017.

[59] Alexander Kraskov, Harald Stögbauer, Ralph G Andrzejak, and Peter Grassberger. Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, 70(2):278, 2005.

[60] Andrea Lancichinetti, Mikko Kivelä, Jari Saramäki, and Santo Fortunato. Characterizing the community structure of complex networks. *PloS one*, 5(8):e11976, 2010.

[61] Ted G Lewis. *Network science: Theory and applications*. John Wiley & Sons, 2011.

[62] Fredrik Liljeros, Christofer R Edling, Luis A Nunes Amaral, H Eugene Stanley, and Yvonne Åberg. The web of human sexual contacts. *Nature*, 411(6840): 907–908, 2001.

[63] Esfandiar Maasoumi 1. A compendium to information theory in economics and econometrics. *Econometric reviews*, 12(2):137–181, 1993.

[64] Rosario N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.

[65] Rosario Nunzio Mantegna, Harry Eugene Stanley, et al. *An introduction to econophysics: correlations and complexity in finance*, volume 9. Cambridge university press Cambridge, 2000.

[66] Mark McDonald, Omer Suleman, Stacy Williams, Sam Howison, and Neil F Johnson. Detecting a currencys dominance or dependence using foreign exchange network trees. *Physical Review E*, 72(4):046106, 2005.

[67] Richard A Meese and Andrew K Rose. An empirical assessment of nonlinearities in models of exchange rate determination. *The Review of Economic Studies*, 58(3):603–619, 1991.

[68] George A Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 2(95):100, 1955.

[69] Takayuki Mizuno, Hideki Takayasu, and Misako Takayasu. Correlation networks among currencies. *Physica A: Statistical Mechanics and its Applications*, 364: 336–342, 2006.

[70] SR Nanda, Biswajit Mahanty, and MK Tiwari. Clustering indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12):8793–8798, 2010.

[71] Michael J Naylor, Lawrence C Rose, and Brendan J Moyle. Topology of foreign exchange markets using hierarchical structure methods. *Physica A: Statistical Mechanics and its Applications*, 382(1):199–208, 2007.

[72] Mark EJ Newman. The structure and function of networks. *Computer Physics Communications*, 147(1-2):40–45, 2002.

[73] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2): 026126, 2003.

[74] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[75] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[76] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[77] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99 (suppl 1):2566–2572, 2002.

[78] Ashadun Nobi, Seong Eun Maeng, Gyeong Gyun Ha, and Jae Woo Lee. Effects of global financial crisis on network structure in a local stock market. *Physica A: Statistical Mechanics and its Applications*, 407:135–143, 2014.

[79] Kyong Joo Oh and Kyoung-jae Kim. Analyzing stock market tick data using piecewise nonlinear model. *Expert Systems with Applications*, 22(3):249–255, 2002.

[80] J-P Onnela, A Chakraborti, K Kaski, and J Kertiész. Dynamic asset trees and portfolio analysis. *The European Physical Journal B-Condensed Matter and Complex Systems*, 30(3):285–288, 2002.

[81] J-P Onnela, Anirban Chakraborti, Kimmo Kaski, Janos Kertesz, and Antti Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 68(5):056110, 2003.

[82] J-P Onnela, Kimmo Kaski, and Janos Kertész. Clustering and information in correlation based financial networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):353–362, 2004.

[83] Günce Orman and Vincent Labatut. A comparison of community detection algorithms on artificial networks. In *Discovery science*, pages 242–256. Springer, 2009.

[84] Karl Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, 1920.

[85] George C Philippatos and Charles J Wilson. Entropy, market risk, and the selection of efficient portfolios. *Applied Economics*, 4(3):209–220, 1972.

[86] Standard Poors and MSCI/Barra. Global industry classification standard. URL `https://www.unm.edu/~maj/Security%20Analysis/GICS.pdf`.

[87] Robert Clay Prim. Shortest connection networks and some generalizations. *Bell system technical journal*, 36(6):1389–1401, 1957.

[88] Min Qi. Nonlinear predictability of stock returns using financial and economic variables. *Journal of Business & Economic Statistics*, 17(4):419–429, 1999.

[89] Min Qi and Yangru Wu. Nonlinear prediction of exchange rates with monetary fundamentals. *Journal of Empirical Finance*, 10(5):623–640, 2003.

[90] Sidney Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.

[91] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[92] Thomas Schürmann and Peter Grassberger. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427, 1996.

[93] CE Shannon. A mathematical theory of communication, bell system technical journal, vol. 27, 379-423 & 623-656, july & october. 1948.

[94] Paweł Sieczka and Janusz A Hołyst. Correlations in commodity markets. *Physica A: Statistical Mechanics and its Applications*, 388(8):1621–1630, 2009.

[95] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13(1):328, 2012.

[96] Ehsan S Soofi. Information theoretic regression methods. In *Applying Maximum Entropy to Econometric Problems*, pages 25–83. Emerald Group Publishing Limited, 1997.

[97] Charles Spearman. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

[98] Pang-Ning Tan et al. *Introduction to data mining*. Pearson Education India, 2006.

[99] Michele Tumminello, Tomaso Aste, Tiziana Di Matteo, and Rosario N Man-

tegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10421–10426, 2005.

[100] Spilios Tzouras, Christoforos Anagnostopoulos, and Emma McCoy. Financial time series modeling using the hurst exponent. *Physica A: Statistical Mechanics and its Applications*, 425:50–68, 2015.

[101] N Vandewalle, F Brisbois, X Tordoir, et al. Non-random topology of stock markets. *Quantitative Finance*, 1(3):372–374, 2001.

[102] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440, 1998.

[103] Zhenyu Wu and Richard Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 15(11):1101–1113, 1993.

[104] Zhao Yang, René Algesheimer, and Claudio J Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, 6: 30750, 2016.

[105] Mohammed J. Zaki and Jr. Wagner Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, May 2014. ISBN 9780521766333.

[106] Mohammed J Zaki, Wagner Meira Jr, and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[107] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[108] Longfeng Zhao, Wei Li, and Xu Cai. Structure and dynamics of stock market in times of crisis. *Physics Letters A*, 380(5):654–666, 2016.

Vita

Serkan Alkan

Personal

PLACE OF BIRTH: Mersin, Turkey

ADDRESS: 272 Knickerbocker Rd, Dumont, NJ, 07628

PHONE: +1-201-702-5093

EMAIL: serkanalkan83@gmail.com

Education

- Doctoral Candidate in Financial Engineering, Stevens Institute of Technology, Hoboken, NJ, expected graduation May 2019.
- M.Sc. Financial Engineering, Stevens Institute of Technology, Hoboken, NJ, May 2011.
- B.S. Mathematics, Uludag University, Bursa, Turkey, June 2006.

Employment

- Stevens Institute of Technology, Department of Financial Engineering, Hoboken, NJ, USA Full time Teaching Assistant, Fall 2014-Spring 2016
- Stevens Institute of Technology, Department of Mathematical Sciences Hoboken, NJ, USA Part time Teaching Assistant, Spring 2012-Spring 2013
- Trabzon Cagdas Cozum Dergisi Dershanesi (private high school) Trabzon, Turkey Mathematics Teacher, Fall 2006-Spring 2008

Publications

Journal Articles

- Linear and Nonlinear Hierarchical Stock Network Methods *working paper*
- Dynamic Evolution of Complex Mutual Information Theoretic Stock Networks *working paper.*
- Industry Classifications and Identification of Important Industry Groups *working paper.*

Proceedings

- Structural Evolution of Stock Networks, Signal-Image Technology & Internet-Based Systems (SITIS), 2015 11$^{\text{th}}$ International Conference on IEEE, 2015.
- Comparing the Quality Functions for Community Detection,*22nd Asia-Pacific Conference on Global Business, Economics, Finance & Social Sciences, 2019*

Research Interests

Complex Networks Applications to Finance, Cluster Analysis and Community Detection Algorithms, Multivariate Data Analysis, Information Theory, Machine Learning Applications to Finance, Portfolio Theory.

Computer Skills

MATLAB, R, C++, LATEX, Microsoft VBA, Microsoft Excel, Microsoft Word.

Awarded Scholarships And Honors

- Scholarship from Ministry of National Education of The Republic of Turkey, to study MS and PhD in Financial Engineering in the USA
- 2001-2005, Scholarship from MMCI-Mersin Chamber of Commerce and Industry, Mersin, Turkey
- 2001-2005, Scholarship from ATAS Refinery, Mersin, Turkey

Languages

- Turkish (Native), English (fluent)