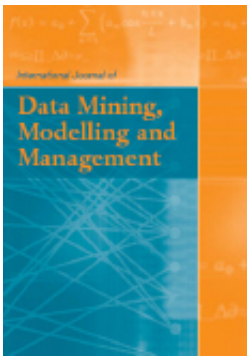


International Journal of Data Mining, Modelling and Management

ISSN online: 1759-1171 - ISSN print: 1759-1163
https://www.inderscience.com/ijdmmm

Comparative analysis of distance measures in stock network construction and cluster analysis

Serkan Alkan

DOI: 10.1504/IJDMMM.2025.10065097

Article History:

Received:	17 July 2023
Last revised:	07 February 2024
Accepted:	07 February 2024
Published online:	25 February 2025

Comparative analysis of distance measures in stock network construction and cluster analysis

Serkan Alkan

Department of Finance and Banking,
Tarsus University,
Mersin, Turkiye
Email: serkanalkan@tarsus.edu.tr

Abstract: The mutual information (MI) metric and the Pearson correlation metric are both widely used in cluster analysis and stock network construction. This paper presents a detailed comparison between the MI metric and the Pearson correlation metric. To detect nonlinear relationships, polynomial and natural cubic spline regressions are proposed as alternatives to the MI metric. The methodology for computing model-fitting indices for determining network adjacencies is explained in detail, along with a comparison of the results with the MI methodology. This study employs two data sets derived from the log returns of the daily adjusted closing prices of 402 stocks in the S&P500 index to measure the impact of a financial crisis on nonlinearity: one covering the crisis period from January 2007 to December 2009, and the other covering the non-crisis period between January 2012 and December 2015. The local and global properties of hierarchical stock networks are compared using the minimum spanning tree for each distance measure. The graph-theoretic internal cluster validity indices and external indices are also used to investigate the relationship between the performance of the community detection algorithm and the selection of metrics.

Keywords: financial networks; mutual information; Pearson correlation; regression models; community detection.

Reference to this paper should be made as follows: Alkan, S. (2025) 'Comparative analysis of distance measures in stock network construction and cluster analysis', *Int. J. Data Mining, Modelling and Management*, Vol. 17, No. 1, pp.75–102.

Biographical notes: Serkan Alkan is a Lecturer Doctor in the Department of Finance and Banking at Tarsus University in Turkiye since 2020. He earned his Bachelor's in Mathematics at Uludag University and his MS and PhD in Financial Engineering from Stevens Institute of Technology. His teaching interests include data mining, pricing and hedging, and financial risk management. His research focuses on applying complex network theory and machine learning to finance.

1 Introduction

Financial markets can be characterised as complex systems because of the interaction between heterogeneous components and existing nonlinearity (Mantegna et al., 2000). Network theory has been used to model the financial markets, in which nodes can be stocks or stock indices (Samitas et al., 2022; Zhou et al., 2023), commodities (de Oliveira Passos et al., 2020), countries (Vidya et al., 2023), banks (Wang et al., 2023), firms (Hao et al., 2023), and cryptocurrencies (Vidal-Tomás, 2021; Hong and Yoon, 2022; Hanif et al., 2023). 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) (Mantegna et al., 2000), planar maximally filtered graphs (Tumminello et al., 2005), asset trees and asset graphs (Onnela et al., 2004). 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 number of stocks. These filtering methods aim to preserve the most relevant information about pair correlation coefficients between stocks to retain the market's core structure. Stock networks have been applied to a number of problems in finance, such as portfolio optimisation (Onnela et al., 2003; Freitas and Junior, 2023), prediction of market direction (Wu et al., 2022), evaluation of systemic risk by using some properties of network structure (Khashanah and Yang, 2016; Zhang et al., 2022), measuring the stability of market (Heiberger, 2014; Zhang and Zhuang, 2019), and prediction of national economic growth (Heiberger, 2018).

Regardless of the specific use of financial networks, the initial step involves choosing measures of co-expression to establish connections among the financial assets. The most common choice is Pearson's correlation coefficient, but it captures the pairwise relationship very well when the data follows the multivariate normal distribution, and more generally for spherical and elliptical distributions (de Siqueira Santos et al., 2013). Additionally, it only works well if returns are linearly associated and fails to detect any nonlinear 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 behaviour of a normally distributed variable (Cont, 2001; Tzouras et al., 2015). Furthermore, zero correlation does not imply statistically independent. It only means linear independence and it is possible that there may be nonlinearity.

In order to account for 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 (de Siqueira Santos et al., 2013). Mutual information, introduced by Shannon (1948), provides a general measurement for dependencies, such as nonlinear 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 a co-expression measure to model a complex system such as gene regulatory networks in bioinformatics (Butte and Kohane, 1999; Shachaf et al., 2023), climate system (Donges et al., 2009), complex brain networks (Luppi and

Stamatakis, 2021), stock networks (Fiedor, 2014; Han et al., 2019), and the information flow among the cryptocurrencies (Hong and Yoon, 2022; Assaf et al., 2023).

A limited number of scholarly publications have undertaken a systematic examination of the suitability of selected similarity measures for analysing stock networks. Semenov et al. (2024) study the threshold networks of weakly correlated stock returns from the French stock market in 2021. It compares networks built using traditional and Holm processes for Pearson correlation and Kendall correlation, revealing that the Holm procedure dramatically reduces the number of edges as well as the degrees of hubs and independent sets as compared to the traditional procedure. Salgado-Hernández and Vyas (2023) discover that non-monotonic correlations constantly exist in financial markets after assessing correlations in S&P 500 market data for the period between August 2000 to August 2022 applying both the Pearson correlation coefficient and the distance correlation coefficient. Hong and Yoon (2022) examine the changes in the cryptocurrency market during the COVID-19 pandemic through network analysis using mutual information and correlation coefficient methods on data from 102 digital currencies. The results point to significant network differences post-COVID-19 with the mutual information method providing a more precise market structure reconstruction. Millington and Niranjan (2021) employ Pearson correlation together with two rank correlations, Spearman and Kendall's τ to explore the inference of minimal spanning trees from daily financial returns from the USA, the UK and Germany. The study finds that there are discrepancies between the trees due to the nonlinear relationship inherent in the dataset, and that the trees generally have comparable topologies. Hartman and Hlinka (2018) investigate the characterisation of stock market complexity through stock networks constructed from price time series using Pearson correlation coefficient and mutual information. They examine the impact of nonlinearity on network properties using a systematic multi-step approach to quantify and correct for nonlinearity effects. The study concludes that the observed nonlinearity is largely attributed to univariate non-Gaussianity and that there is strong non-stationarity in certain stocks at times, particularly during the 2008 global financial crisis.

This research article aims to provide a theoretical background for evaluating MI between stock returns and proposes a rigorous procedure to compare it with the Pearson correlation. The study constructs distinct hierarchical stock networks for each measure and compares their local and global properties to analyse their strengths and limitations. Another objective of this paper is to enhance the understanding and reliability of similarity measures in stock network analysis. The feasibility of constructing stock networks using predictive models, such as polynomial regression and natural cubic spline, is demonstrated, offering improvements over linear regression by considering nonlinear interactions.

The study also evaluates the performance of a community detection algorithm using internal and external validation indices adapted for graph clustering, comparing the algorithm's results with the Global Industry Classification Standard (GICS) classification. This investigation provides valuable insights into constructing stock networks using predictive modelling techniques and evaluating community detection algorithms, aiding in the identification of sectors and industry groups within stock networks.

The paper is structured as follows: Section 2 presents a comprehensive overview of the dataset utilised in this study. In Section 3, the discussion revolves around the similarity measures employed for constructing the stock networks. It presents a concise

definition of the tools derived from complex networks theory that are relevant to this work. Additionally, it outlines the community detection and cluster validity indices used in the paper. In Sections 4 and 5, the results are introduced, followed by a comprehensive summary of the findings in the conclusion section.

2 Data and software

The dataset analysed in this study consists of adjusted daily closure prices for 402 stocks of the S&P 500 index. The data set consists of two different periods: the crisis period, which encompasses the duration from January 2007 to December 2009, and the non-crisis period from January 2012 to December 2015. The crisis period data set consists of 756 consecutive trading days, while the non-crisis period data set consists of 1,005 days. Table 1 illustrates the distribution of 402 stocks across various sectors and industry groups according to the GICS applicable during the study periods.

Table 1 The distribution of 402 stocks across sectors and industry subgroups based on GICS classification system

Sector	*Industry group*	*Number of stocks*
Energy	Energy	32
Materials	Materials	24
Industrials	Capital goods	37
	Commercial services and supplies	8
	Transportation	11
Consumer discretionary	Automobiles and components	4
	Consumer durables and apparel	14
	Consumer services	9
	Media	8
	Retailing	25
Consumer staples	Food and staples retailing	7
	Food, beverage and tobacco	20
	Household and personal products	5
Healthcare	Healthcare equipment and services	27
	Pharmaceuticals, biotechnology and life sciences	18
Financials	Banks	15
	Diversified financials	19
	Insurance	15
	Real estate	23
Information technology	Software and services	23
	Technology hardware and equipment	14
	Semiconductors and semiconductor equipment	14
Telecommunication services	Telecommunication services	5
Utilities	Utilities	25

The daily adjusted closing prices are transformed to logarithmic returns for each stock as follows:

$$R_i(t) = \log P_i(t) - \log P_i(t-1) \tag{1}$$

where $R_i(t)$ denotes the logarithmic return of i^{th} stock at time t, $P_i(t)$ and $P_i(t-1)$ represents the closing price of the i^{th} stock at day t and $t-1$, respectively.

The augmented Dickey-Fuller (ADF) test was utilised to assess unit root characteristics, and the Jarque-Bera (JB) test was employed to evaluate normality. The findings reveal that all stock return series demonstrate stationarity, but the JB statistic rejects the null hypothesis of normality for both periods, indicating a deviation from the assumption of a Gaussian distribution in the return series.

R software for general scripting, WGCNA package for weighted correlation network analysis (Langfelder and Horvath, 2008), and igraph package for network analysis and visualisation (Csardi and Nepusz, 2006) are used in this study. The time complexities of the algorithms used in the network analysis can be found in the reference manual of the igraph package at https://igraph.org/c/doc/igraph-Structural.html

3 Methodology

3.1 Dissimilarity measures

In order to construct the hierarchical stock networks based on the minimum spanning tree (MST), we first need to compute all pairwise similarities between each stock in the portfolio, and then transform them into dissimilarity measures.

In the subsequent section, a concise overview of the similarity and dissimilarity measures is provided. It is assumed that there exist two n-dimensional vectors, denoted as X and Y, in the context of this discussion.

3.1.1 Pearson's correlation coefficient

Pearson's correlation is a measurement of the linear relationship between two vectors x and y (Galton, 1889; Pearson, 1920). Pearson sample correlation, r_p, is described as follows:

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

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 a sample mean of 0 and sample variance of 1. It can easily be shown that sample correlation is scale-invariant to linear transformations and takes values in the 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 a variety of metrics defined based on the Pearson correlation. In this work, $d_p(X,Y) = \sqrt{2(1 - r_p(X,Y))}$ is used, which is a valid metric and satisfies the metric axioms (Mantegna, 1999).

3.1.2 Mutual information

Entropy estimates the level of disorder in a dynamic system. In information theory, entropy measures the diversity of patterns in a time series to determine the amount of information contained in it. For a discrete random variable X, Shannon's entropy is expressed as follows:

$$H(X) = H(p(X)) = -\sum_{x \in X} p(x) \log p(x) = -E \log p(X) \tag{3}$$

where X is a random variable, with $p_1, \ldots, p_n$ the probabilities of occurrence of a set of events, and E is the expected value operator (Shannon, 1948).

Mutual information (MI) measures how much knowing one variable reduces uncertainty about the other. The MI has the following definition:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

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.

The MI can take only non-negative values and equals 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 decrease in uncertainty in X due to 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. Compared to correlation, mutual information has the advantage of capturing linearly and nonlinearly dependent relationships.

There are two distinct forms of MI metrics proposed by Kraskov et al. (2005). The first metric is based on the universal mutual information adjacency matrix version 1 ($AUV1$), which is defined as:

$$AUV1 = \frac{I(X,Y)}{H(X,Y)} \tag{5}$$

and $dissAUV1 = 1 - AUV1$, a universal distance function and valid metric, defines the dissimilarity (Kraskov et al., 2005; Song et al., 2012).

The second one is based on universal mutual information adjacency matrix referred to as version 2 or $AUV2$, which is formulated as follows:

$$AUV2 = \frac{I(X,Y)}{\max\left(H(X), H(Y)\right)} \tag{6}$$

and $dissAUV1 = 1 - AUV1$, a universal distance function and valid metric, also defines the dissimilarity (Kraskov et al., 2005; Song et al., 2012). $dissAUV2$ is sharper than $dissAUV1$, that is, $dissAUV2 \leq dissAUV1$ (Kraskov et al., 2005).

The mutual information and entropy measures were originally expressed for discrete or categorical variables (Cover and Thomas, 2006). Although the entropy of discrete and continuous distributions have similar properties, continuous random variables have some drawbacks: it may take a negative or infinitely large value. The integral may not exist.

Moreover, it is generally not invariant under some monotonic transformations (Cover and Thomas, 2006).

In this study, the procedure for obtaining the mutual information adjacency matrix is as follows: firstly, the stock returns are discretised using the equal-width method, with the default number of bins set as $\sqrt{n}$. Subsequently, the mutual information between each pair of discretised stock returns is computed using the Miller-Madow estimation technique. This results in the transformation of the mutual information matrix into one of the aforementioned versions of adjacency matrices or distance matrices, as defined previously.

3.1.3 Relationship between mutual information and correlation

Despite the differences between mutual information and correlation, such as mutual information depends on parameter choice and is a general measure of association not like correlation which only measures linear or monotonic relationships. In Song et al. (2012), they give a simple approximation between two measures under some assumptions. First, samples come from the bivariate normal distribution and the equal-width discretisation 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:

$$\begin{aligned} AUV2(x,y) &= \frac{I(x,y)}{\max(H(x),H(y))} \\ &\approx F^{cor-MI}(cor(x,y)) \end{aligned} \tag{7}$$

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

$$F^{cor-MI}(s) = \frac{\log(1+\epsilon-s^2)}{\log(\epsilon)}(1-w)+w \tag{8}$$

where $w = 0.43n^{-0.30}$ and $\epsilon = w^{2.2}$.

This approximation is based on if x and y are samples from bivariate normal distribution; therefore, the excess information in the two estimates is

$$AUV2_e = AUV2 - F^{cor-MI}(s) \tag{9}$$

The relationship between the $AUV2$ and F^{cor-MI} can be employed to define non-Gaussian information or extranormal information. This relationship serves as a basis for identifying cases in which there is a substantial disagreement between Pearson and mutual information measures. By leveraging this relationship, instances can be systematically identified where the two measures demonstrate significant disparities.

3.2 Network topological properties

3.2.1 Network construction

Treating stocks as network nodes, we can now construct the stock network with the minimum spanning tree method, in which network links are added according to the

distance between nodes. MST stock network generation is one of the popular methods in the literature and is introduced by Mantegna (1999). The construction process is defined as follows: we start with a random stock as the seed of a partial tree and at every step, the partial tree extends by iteratively adding an unattached stock to it by choosing the least weight until the unattached stock set is exhausted, which is known as Prim's algorithm (Prim, 1957). MST of order N has $N-1$ connections and no loops or circuits. Furthermore, MST has a strong relationship with a single linkage clustering algorithm (Jain and Dubes, 1988).

A set of well-defined tools in the literature are used to compare the different perspectives of stock network topology at the local, mesoscopic, and global levels (Newman, 2002; Boccaletti et al., 2006; Freeman, 1978). In the following subsections, we explain several commonly used measures to analyse the topological characteristics of the networks.

3.2.2 Local measures

3.2.2.1 Degree centrality

In an unweighted network, d_i is the number of vertices directly attached to the i^{th} vertex. If the network is weighted, then the node strength is the sum of link weights between node i and the other nodes incident with it. While the degree is a local feature, the average degree is a global feature.

3.2.3 Global measures

3.2.3.1 Hamming distance

The Hamming distance $H(A,B)$ of two labelled simple networks with adjacency matrices $A(i,j)$ and $B(i,j)$ computes the fraction of links that have to be added or removed to convert one network into the other provided that both networks have the identical number of nodes (Donges et al., 2009).

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

$$H(A,B) = \langle XOR(A(i,j),B(i,j))\rangle_{ij} \tag{10}$$

where

$$\text{XOR}(A(i,j),B(i,j)) = \begin{cases} 1 & \text{if } A(i,j) \neq B(i,j) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Hamming distance lays in the range $0 \leq H(A,B) \leq 1$ and computes the global probability of non-equal entries in the two adjacency matrices.

3.2.3.2 Average path length

The geodesic distance d is the least number of links required to travel between vertex v to vertex u. The average path length L of a network is the average geodesic distance between all interconnected pairs of nodes and is represented as:

$$L = \frac{1}{\binom{n}{2}} \sum_{i<j} d_{ij} \tag{12}$$

3.2.3.3 Closeness centrality

The closeness centrality computes the reciprocal average topological distance of vertex v_i to all other vertices in the network (Freeman, 1978). This measure is normalised by multiplication factor $N-1$ to $0 \leq CC_{v_i} \leq 1$, and is formulated as

$$CC_{v_i} = \frac{N-1}{\sum_j d(v_i, v_j)} \tag{13}$$

In a graph, the median node is the one with the least total distance between its neighbours.

3.2.3.4 Betweenness centrality

Betweenness centrality measures a vertex centrality in terms of its location between other pairs of vertices in a network. Typically, it is described as the number of shortest paths from all vertices to all others that traverse that vertex. BC_v is defined as

$$BC_v = \sum_{i \neq j \neq v}^{N} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \tag{14}$$

where σ_{ij} is the total number of shortest paths from i to j which includes node v (Freeman, 1978).

3.2.3.5 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.,

$$e(v_i) = \max_j \left\{ d(v_i, v_j \right\} \tag{15}$$

Eccentricity centrality is thus expressed as:

$$c(v_i) = \frac{1}{e(v_i)} = \frac{1}{\max_j \left\{ d(v_i, v_j) \right\}} \tag{16}$$

A node v_i with the smallest eccentricity is called a centre node, while the one with the greatest eccentricity is called a periphery node (Zaki and Meira, 2014).

3.2.4 Scale-free networks

One of the key features 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*:

$$f(k) = Ck^{-\alpha} \tag{17}$$

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 (Barabási and Albert, 1999) with scaling parameter α that is also called as the exponent of the power-law. Let us take the logarithm of both sides of the relation (17):

$$\log f(k) = \log (Ck^{-\alpha}) \tag{18}$$

which yields:

$$\log f(k) = -\alpha \log k + \log C \tag{19}$$

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. A least-squares fit of a line to the points $(\log k, \log f(k))$ is the standard method for checking whether a network has scale-free property. R^2 statistics is then applied to measure the straight-line relationship between the points. It takes values between 0 and 1 and explains the ratio of variance determined by the regression model. This value is called the scale-free topology fitting index, as described in Zhang and Horvath (2005).

3.3 Community detection and cluster validity indices

Cluster analysis is the process of dividing data objects into clusters based on their similarities, and it has been implemented in many areas, such as gene expression analysis (Zhang and Horvath, 2005), social networks (Newman and Girvan, 2004), image segmentation (Wu and Leahy, 1993), and finance (Nanda et al., 2010). The choice of clustering algorithms and distance functions poses significant challenges, especially when external benchmarks are unavailable.

In networks portraying real systems, connections between nodes are not random or regular at the global level, and are organised and orderly at the local level, where some nodes are tightly interconnected within themselves but few linkages with other nodes (Newman and Girvan, 2004). It is called modules or communities when these nodes are highly connected to each other or of the same type (Newman and Girvan, 2004).There have been many community detection algorithms presented to disclose these mesoscopic properties of complex networks, most of which are derived from traditional clustering methods.

In this study, the edge betweenness community detection method introduced by Newman and Girvan (2004) is chosen for the analysis. This method, available in the igraph library (Csardi and Nepusz, 2006), employs a systematic approach of removing edges based on their betweenness scores. By iteratively removing edges with high betweenness scores, a hierarchical decomposition process is performed. The edges linking different communities are expected to have high betweenness, and their removal allows for the identification of distinct communities within the network. The time complexity of this algorithm is (E^2N), where E denotes the number of links and N denotes the number of vertices in the network.

A number of internal and external indices have been proposed in the literature to assess the quality of clusters for clustering. In this study, geodesic distance is used to calculate internal indices, which allows the calculation of distances between any nodes in the network. The centroid nodes are defined as nodes that minimise the maximum distance to all other nodes within the cluster.

3.3.1 Internal indices

A brief description of the internal indices used in this study is presented in this section.

3.3.1.1 C-index

Let S_{in} be the sum of all the intracluster distances. The C-index (Hubert and Schultz, 1976) is formulated as:

$$C = \frac{S_{in} - S_{\min}}{S_{\max} - S_{\min}} \tag{20}$$

where $S_{\min}$ equals to the sum of the N_{in} smallest distances, and $S_{\max}$ is defined as the sum of the N_{in} largest distances in $\mathbf{W}$. It takes values in $[0, 1]$. Clustering is better with a smaller C-index.

3.3.1.2 Dunn index

The Dunn index measures the ratio between the minimal distance within a cluster and the maximum distance between clusters. It is defined as:

$$Dunn1 = \frac{W_{out}^{\min}}{W_{in}^{\max}} \tag{21}$$

where $W_{out}^{\min}$ indicates the least distance between two entities from two different clusters, whereas $W_{in}^{\max}$ indicates the maximum distance between two entities from the same cluster (Dunn, 1973).

3.3.1.3 Silhoutte coefficient

It is defined as follows:

$$s_i = \frac{\mu_{inter} - \mu_{intra}}{\max \mu_{inter}, \mu_{intra}} \tag{22}$$

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. s_i lies between –1 and 1. If the value of x_i is close to 1, then it is assigned to the best possible cluster, while if it is close to –1, then the point may be assigned to the incorrect cluster. The value of x_i is close to zero when it is around the boundary between two clusters. The silhouette coefficient is expressed as the average s_i values through all the points:

$$SC = \frac{1}{n} \sum_{i=1} n s_i \tag{23}$$

A value near +1 indicates good clustering. The silhouette coefficient is a metric of both cohesion and separation of clusters (Rousseeuw, 1987).

3.3.1.4 BetaCV measure

BetaCV is defined as the ratio of intra-cluster distance to inter-cluster distance. The small value indicates better clustering. It is formulated as follows:

$$BetaCV = \frac{S_{in}/N_{in}}{S_{out}/N_{out}} \tag{24}$$

3.3.2 External indices

In this study, the normalised mutual information (NMI) is utilised as an external measure (Danon et al., 2005). The NMI is given by:

$$NMI(C,T) = \frac{2I(C,T)}{H(C)+H(T)} \tag{25}$$

where

$$H(C) = \sum_{i=1}^{r} p_{C_i} \log p_{C_i} \tag{26}$$

$$H(T) = \sum_{j=1}^{k} p_{T_j} \log p_{T_j} \tag{27}$$

and $p_{C_i} = \frac{n_i}{n}$ is the probability of cluster C_i and $p_{T_j} = \frac{m_i}{n}$ is the probability of partition T_j, and

$$I(C,T) = \sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log \frac{p_{ij}}{p_{C_i} p_{T_j}} \tag{28}$$

where $p_{ij} = \frac{n_{ij}}{n}$ is the probability that an object in both C_i and T_j. $I(C,T)$ is the mutual information between C_i and T_j. If clustering and partitioning are identical, NMI equals 1, otherwise it equals 0. In this study, two types of external information are considered as ground-truth measures. The first type is at the sector level, and the second type is at the industry-group level. This allows for the evaluation of algorithm performance and dissimilarity measures in terms of their ability to detect more refined clusters.

4 Results and discussion

After the introduction of various methods for stock network construction, the comparison between the Pearson correlation and mutual information measures is conducted to assess the influence of nonlinearity in several aspects during both the crisis period and the non-crisis period.

4.1 Correlation matrix analysis for non-crisis period

The Pearson correlation and AUV2 values are calculated for all stock pairs within each portfolio. Subsequently, AUV2 is predicted based on the Pearson correlation using the approximation function F^{cor-MI}. As shown in Figure 1, the predicted AUV2 aligns closely with the actual values, demonstrating high accuracy. Furthermore, the scatter plot depicting the relationship between the Pearson correlation and AUV2 reveals a strong monotonic association (Spearman correlation of 0.9). These findings suggest that the majority of stock pairs exhibit linear relationships during the considered time period.

Figure 1 Comparison of correlation and mutual information estimates for the non-crisis period with Spearman's correlations shown at the top (see online version for colours)

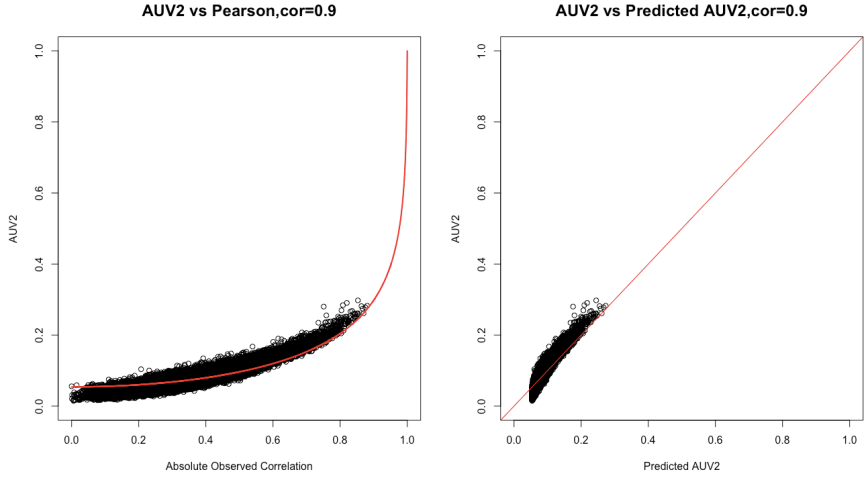
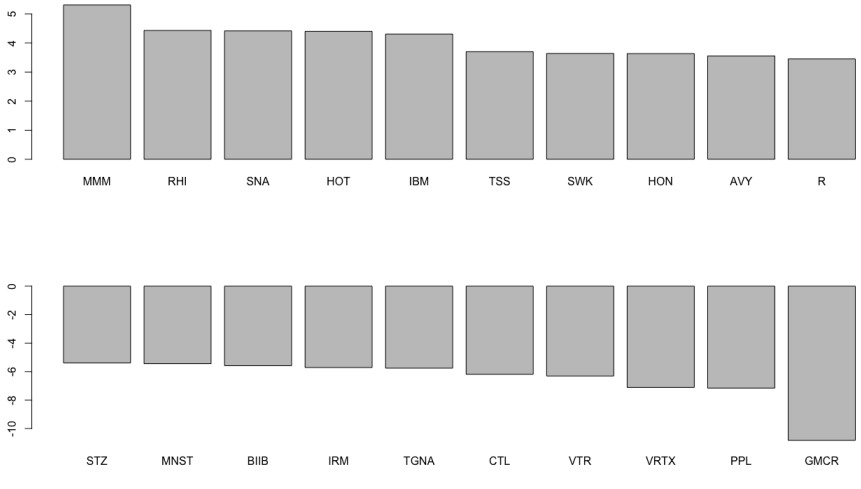


Figure 2 Extranormal information estimates for stocks, with the top ten highest positive values in the upper row and the lowest negative values in the lower row during the non-crisis period

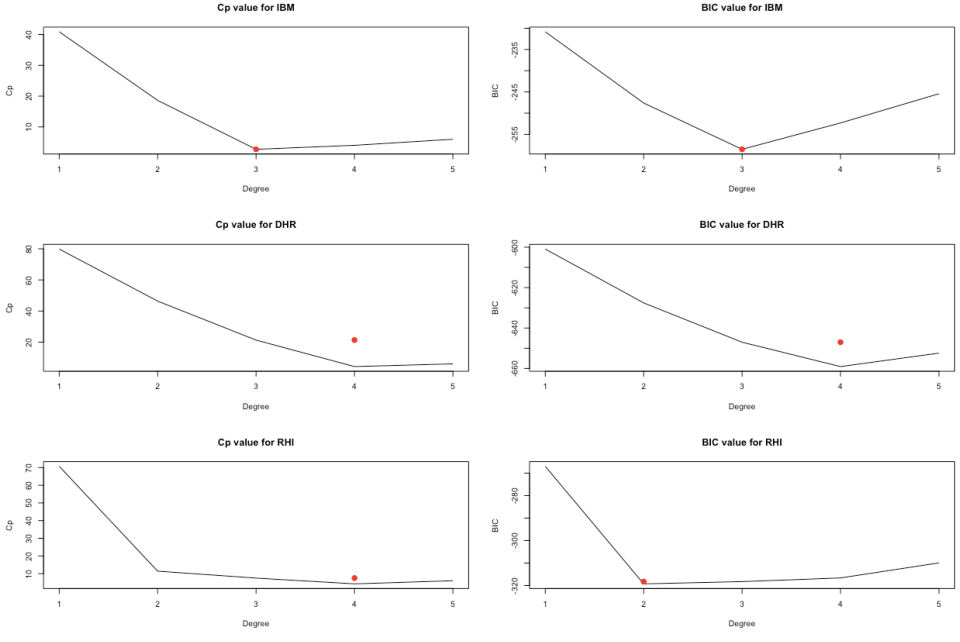


Despite the close relationship uncovered by F^{cor-MI} between the Pearson correlation and AUV2, discrepancies between the two measures are observed for specific stock pairs. To identify instances where the measures exhibit strong mismatch, the extranormal information $AUV2_e$ is computed for each stock pair. This extranormal information allows for the identification of stock pairs where AUV2 displays a higher value while F^{cor-MI} underestimates or overestimates it. Accordingly, stock pairs can be divided into two classes based on their positive and negative extranormal information.

Figure 2 illustrates the stocks with the ten highest positive extranormal information in the first row, while the bottom row displays the stocks with the lowest negative extranormal information. The extranormal information estimates for all stocks are computed, and the bars in the figure represent the sums of the rows (or columns) of the extranormal information matrix for each stock.

To demonstrate the limitations of simple linear regression in capturing the relationship for the first category of stocks, polynomial regression up to degree 5 was employed. A specific case was examined involving the stock MMM and its relationship with three other selected stocks, where a positive deviation from the approximation was observed. Figure 3 presents the results, highlighting that the BIC and Cp values confirm the inadequacy of simple linear regression in representing the relationship. Additionally, the ANOVA tables provide further support for these findings.

Figure 3 BIC and Cp values for MMM (see online version for colours)



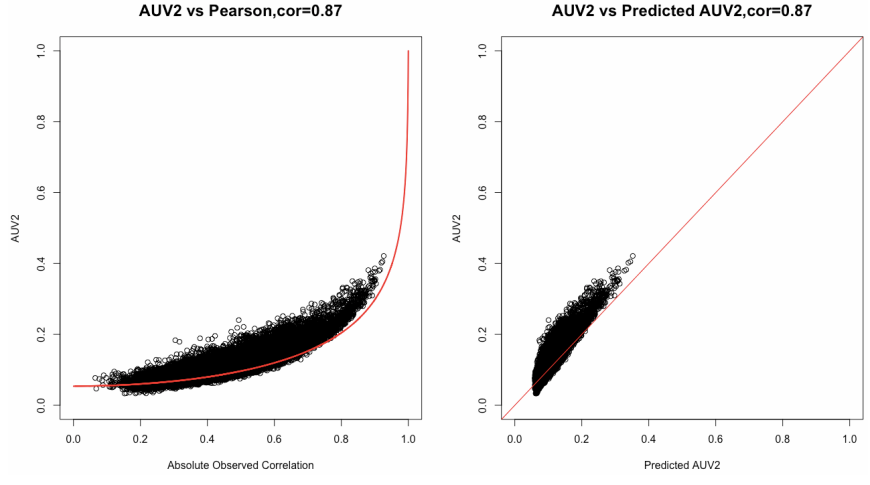
The subsequent query focuses on stocks with negative extranormal information. In this instance, stock pairs exhibit higher correlation values than AUV2 suggests. It is noteworthy that AUV2 identifies correlations as insignificant, whereas the Pearson measure fails to capture this aspect. The bar chart indicates that GMCR exhibits the lowest negative value. Upon examining its relationship, it becomes evident that GMCR

demonstrates a maximum correlation of 0.2367611, which can be characterised as relatively weak.

4.2 Correlation matrix analysis for crisis period

A similar approach is employed on a crisis period data set, using the same set of stocks, to explore nonlinearity or extranormal information among all stock pairs. Figure 4 illustrates that Spearman's correlation slightly decreases from 0.9 to 0.87. The approximation function continues to accurately predict AUV2, indicating a strong monotonic relationship between the two association measures. These findings suggest that the majority of pairwise dependencies between stock pairs remain linear. However, the utilisation of extranormal information enables the identification of stock pairs where the two measures exhibit disagreement.

Figure 4 Comparison of correlation and mutual information estimates for crisis period with the Spearman's correlations shown at the top (see online version for colours)



Upon comparing Figure 5 with Figure 2, it is evident that significantly stronger deviations from Gaussianity are observed. For instance, during the non-crisis period, MMM exhibits a maximum extranormal information value of approximately 5, while USB reaches approximately 18 in the crisis period. Additionally, it is worth noting that in the crisis period, more than half of the stocks in the portfolio, specifically 240 stocks, have extranormal information values exceeding 5. These findings highlight the pronounced departure from Gaussianity during the crisis period compared to the non-crisis period.

In a manner similar to the previous section, polynomial regression up to degree 5 is conducted for the stock USB, exploring its relationship with three other selected stocks where a positive deviation from the approximation is observed. Figure 6 presents the results, revealing that both BIC and Cp values support the notion that simple linear regression fails to capture the relationship between USB and the selected stocks. Furthermore, these findings are substantiated by the ANOVA tables, providing further

evidence of the inadequacy of linear regression in depicting the relationship between USB and the aforementioned stocks.

Figure 5 Extranormal information estimates for stocks, with the top ten highest positive values in the upper row and the lowest negative values in the lower row during the crisis period

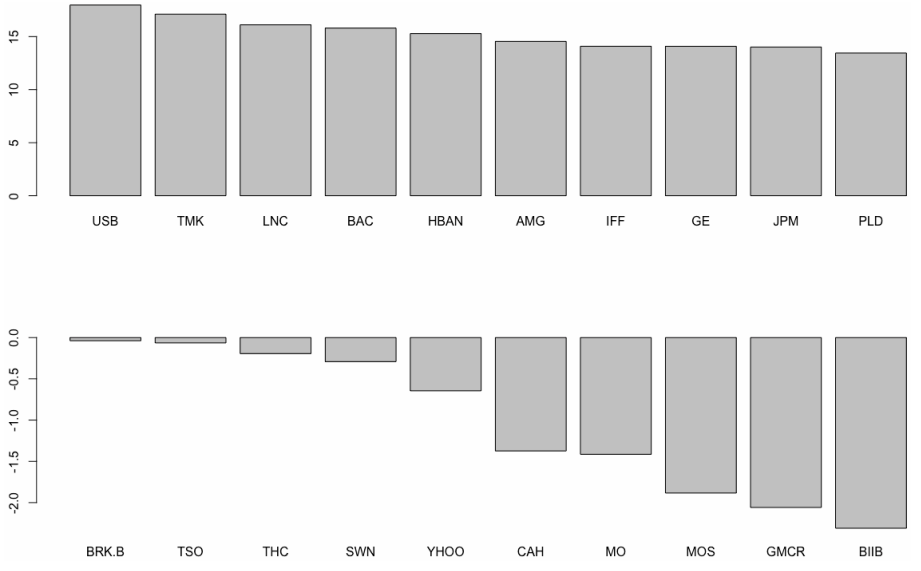
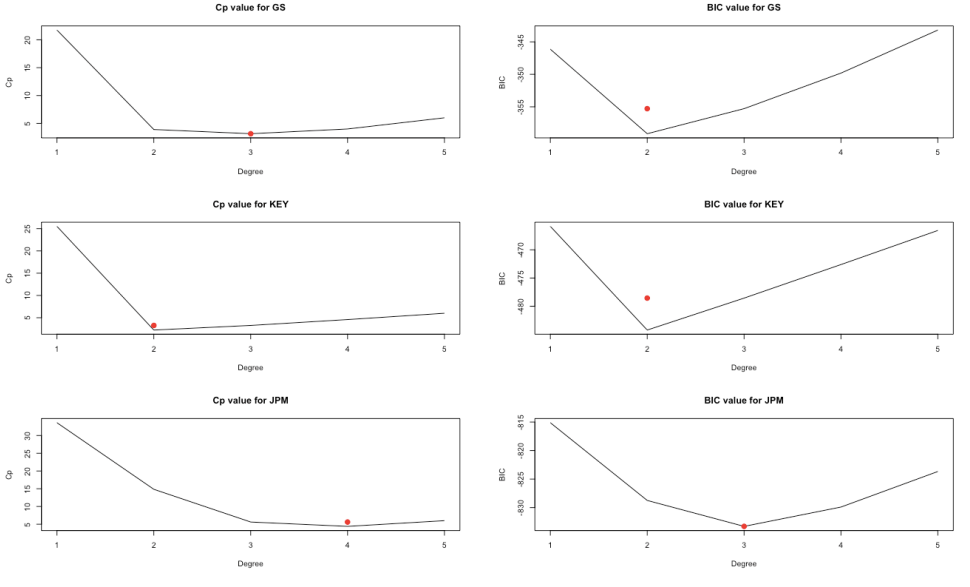


Figure 6 BIC and Cp values for USB (see online version for colours)



4.3 Polynomial and spline regression models

The incorporation of classic polynomial and spline regression models provides a means to capture nonlinear correlations between variables. These models offer simpler and faster estimations in comparison to mutual information (MI). Moreover, they facilitate the utilisation of classic statistical tests to assess the goodness of fit (Song et al., 2012).

Polynomial regression is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_d x_i^d + \epsilon_i \tag{29}$$

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 (Song et al., 2012):

$$R^2 = cor^2(y, \hat{y}) \tag{30}$$

R^2 statistics takes values between 0 and 1 and explains the proportion of variance determined by the regression model. Cubic polynomial regression is utilised to quantify the nonlinear relationship within the data. However, the resulting matrix R^2 displays asymmetry. Symmetrising non-symmetric matrices is a common practice, and in this study, a specific method is employed for this purpose:

$$S_{ij}^{ave} = \frac{S_{ij} + S_{ji}}{2} \tag{31}$$

where S is the matrix R^2.

The dissimilarities based on the R^2 statistics are computed using the following formula:

$$d(X, Y) = 1 - R^2$$

This dissimilarity measure will be referred to as $dissPolyReg$ in the subsequent analysis.

In the context of spline regression, the range of X is partitioned into K distinct regions, and a separate polynomial function is fitted for each region. These regions are defined by specific boundaries called knots, where the coefficients undergo changes. In this study, the selection of 5 knots aligns with the rule of thumb employed, which suggests using 5 knots when the sample size exceeds 100, 3 knots for sample sizes below 30, and 4 knots for intermediate sample sizes (Song et al., 2012).

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

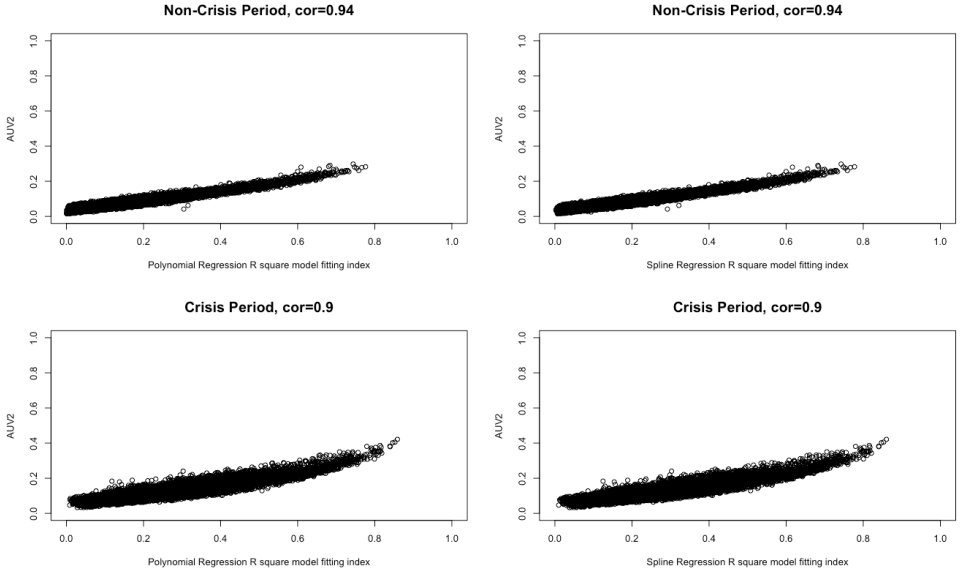
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_D x_i^D + \sum_{k=1}^{K} b_k (x_i - \xi_k)^D + \epsilon_i \tag{32}$$

where the function

$$(x - \xi)_+^D = \begin{cases} (x - \xi)^D & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

is called a truncated power basis function of degree D. One limitation of splines is the potential for high variance beyond the range of predictors, such as when X is smaller than the smallest knot or larger than the largest knot. To address this issue, a natural cubic spline enforces linearity at the boundary, resulting in more stable estimates. In this study, natural cubic splines are employed. Similarly to polynomial regression, the symmetrised R^2 can be used as a measure of similarity and dissimilarity between x and y. The dissimilarity measure based on spline regression is referred to as $dissSpline$ for conciseness. Spline regression offers increased flexibility compared to polynomial and step-wise regressions and ensures the smoothness of each observation (James et al., 2013).

Figure 7 Comparisons of regression models and mutual information with Pearson's correlation between measures are shown on the top



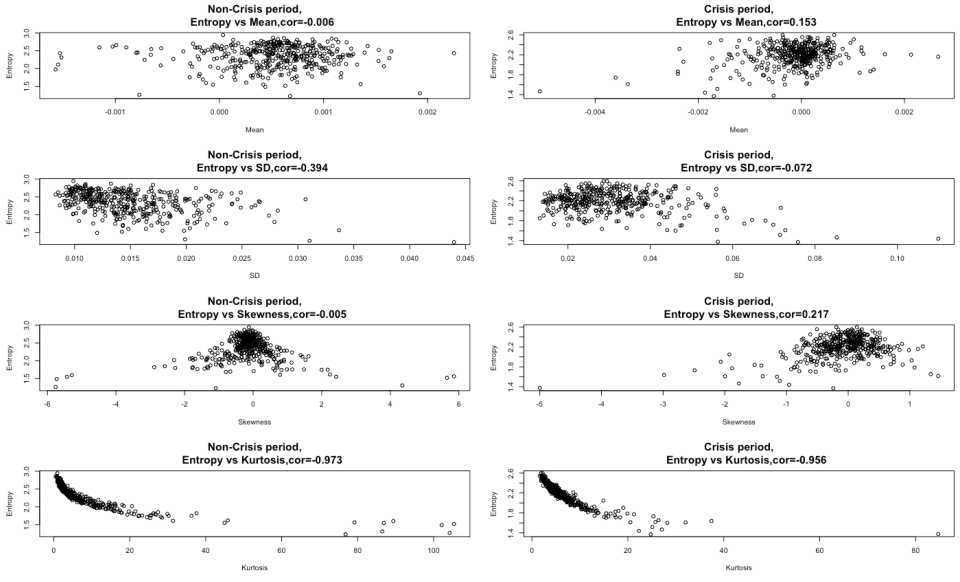
In the empirical datasets representing crisis and non-crisis periods, a notable observation depicted in Figure 7 is the high correlation between regression models and mutual information (AUV2). Furthermore, both regression models and AUV2 exhibit a stronger correlation compared to that between Pearson's correlation and AUV2. This finding provides additional evidence that AUV2 and regression models unveil nonlinear relationships among certain stock pairs, whereas Pearson's correlation fails to identify such associations. These findings are in line with the results of Semenov et al. (2024), Millington and Niranjan (2021) and McAllester and Stratos (2020). They found that there are nonlinear relationships present in financial markets at all times.

Moreover, it is observed that the regression models demonstrate a high Pearson's correlation coefficient of 0.99, indicating their identical behaviour in both time periods. Additionally, it is noteworthy that the utilisation of different symmetrisation methods (max and min) to compute R^2 also yields Pearson's correlation coefficients approaching 1. Thus, it can be inferred that any of these techniques can be safely employed.

4.4 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 Philippatos and Wilson (1972), 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 variance.

Figure 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



If the stock return distributions are not 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 (Soofi, 1997; Maasoumi, 1993; Philippatos and Wilson, 1972). Additionally, some studies suggest that entropy is a more general measure of uncertainty than variance, since it incorporates much more information about the probability distribution (Dionisio et al., 2006; Ebrahimi et al., 1999). The relationship between entropy and the first four moments (mean, standard deviation, skewness, and kurtosis) of daily log returns for all studied companies is investigated. The results are presented in a separate scatterplots for the crisis and non-crisis periods in Figure 8.

The Spearman correlation coefficient between the entropy rate and the first three moments of the daily log returns is found to be weaker compared to the Spearman correlation coefficient between the kurtosis and entropy rates. The kurtosis of a distribution quantifies the fatness of its tails. In a normal distribution, the excess kurtosis is expected to be zero, indicating no skewness and a symmetric shape. A distribution with positive excess kurtosis is referred to as leptokurtic, indicating heavier tails. In the non-crisis period, the observed maximum and minimum excess kurtosis values are

0.464 (for OMC) and 105.374 (for ARG), respectively. In contrast, during the crisis period, the minimum and maximum values are 1.641 (for CSX) and 84.704 (for STT), respectively. These findings indicate significant deviations from normality and confirm the presence of leptokurtic behaviour in stock return distributions.

It is worth noting that the average Shannon's entropy rates for all stocks are 2.346 during the non-crisis period and 2.184 during the crisis period. These values provide insights into the changes in market efficiency between the two periods. The observed differences suggest that the market experiences reduced stability and efficiency, becoming more predictable during the crisis period.

In summary, there are several reasons for the differences between mutual information and the Pearson product-moment correlation coefficient: firstly, the Pearson correlation coefficient considers only the first two moments of the data, namely the mean and variance. In contrast, mutual information incorporates entropy, which is highly correlated with higher-order moments. Notably, a striking relationship between entropy and kurtosis has been observed. Secondly, while the Pearson correlation coefficient is limited to detecting linear or monotonic relationships, mutual information has the ability to capture more general dependencies beyond linearity. Thirdly, it is important to consider that estimation errors can arise due to the finite sample size. Consequently, relying solely on linear regression may not fully describe the relationships and can lead to underestimation. Nonlinearity can emerge particularly when there is significant skewness or kurtosis in the distribution of stock returns. In conclusion, these factors demonstrate the importance of combining mutual information with Pearson correlation coefficients to gain a more comprehensive understanding of financial data relationships.

4.5 Comparing network topologies

The focus of the analysis shifts to examining the dissimilarities between networks generated by the Pearson correlation and mutual information measures in terms of their local, global, and mesoscopic properties.

4.5.1 Local and global properties

Following a similar approach as before, the networks are examined separately for each time period, enabling a comparison of their respective characteristics. Specifically, the investigation begins by considering the unweighted representation of each network and presenting a concise summary of the findings in Table 2.

In both normal and non-normal periods, the networks generated using the Pearson correlation exhibit higher hub degrees and diameters compared to those based on mutual information. Notably, during the crisis period, the mutual information-based network demonstrates a higher average path length, whereas the Pearson-based network exhibits a higher average path length in the normal period.

Assortativity in a network refers to the arrangement of connections between vertices of the same type. In this analysis, two forms of assortativity are considered: degree assortativity and nominal assortativity. Degree assortativity examines the inclination of high-degree vertices to link with vertices of similar degrees. The findings indicate that all networks exhibit disassortative mixing by degree, suggesting that vertices with high degrees tend to be linked to vertices with low degrees. Nominal assortativity focuses

on the categorisation of stocks into sectors and industry groups based on the GICS classification system. The analysis consistently reveals a pattern of assortativity across all networks. This pattern signifies a tendency for stocks within the same sector or industry group to display a higher probability of connection within the networks. In simpler terms, there is a preference for intra-sector or intra-industry connections within the networks.

Table 2 Comparative analysis of network properties in crisis and non-crisis periods

	Pearson (non-crisis)	*AUV2 (non-crisis)*	*Pearson (crisis)*	*AUV2 (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

The assortative mixing coefficients for the sector and industry group differ between mutual information and Pearson correlation in the two periods. In the non-crisis period, mutual information exhibits a higher assortative mixing coefficient for the sector, while the Pearson correlation shows a higher coefficient for the crisis period. However, both measures have the same assortative mixing coefficient for the industry group. Notably, the values for mutual information in Table 2 indicate a significant decrease during 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 the mutual information-based network in the crisis period has the highest exponent value, Pearson has the least one in the non-crisis period. However, networks show approximately scale-free behaviour. All networks have approximately similar entropy values. The average closeness and average betweenness values of each network are also approximately the same.

To summarise, significant differences between the networks are mainly observed in terms of the diameter and the hub. However, other fundamental parameters show similar values with no significant variations. The findings obtained are consistent with Millington and Niranjan (2021). They used Pearson together with two rank correlation techniques, Spearman and Kendall's τ, to study the inference of MSTs from daily financial returns from the USA, the UK and Germany. They discovered that the topologies of the trees tend to be identical, regardless of coefficient.

The diameter, being a sensitive measure that considers only the maximum distance between two nodes, can exhibit larger differences. On the other hand, the average path length provides a less biased value, which is approximately the same for the networks.

The variations in hub degrees can be explained by the construction process of the minimum spanning tree (MST) approach, which imposes geometric constraints. As a result, stocks that could be perceived as outliers or potentially having insignificant correlations may tend to connect to the hub in order to minimise the weight.

Table 3 Spearman correlation between local and global properties

	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

The following analysis focuses on quantifying the disparities between networks at both the local and global topological scales. To facilitate a quantitative comparison, the Spearman rank order correlation coefficient is computed. Local topological comparisons involve assessing the degree centrality and Hamming distance, while global topological comparisons include the evaluation of closeness centrality, betweenness centrality, and eccentric centrality. Refer to Table 3 for a concise summary of our findings.

At the local topological scale, the Pearson correlation and mutual information stock networks demonstrate a moderate similarity during the non-crisis period. However, this similarity diminishes during the crisis period, indicating more substantial distinctions between the two networks.

Shifting our focus to the global topological scale, intriguing patterns emerge, particularly during the crisis period. In the non-crisis period, there is a low-rank order correlation observed for eccentric centrality and closeness centrality among stocks. Conversely, the correlation for betweenness centrality scores is slightly higher. However, in the crisis period, both closeness centrality and eccentric centrality exhibit a significant decrease in Spearman correlation compared to the non-crisis period. Furthermore, betweenness centrality also experiences a slight decline. Overall, these results highlight notable discrepancies between the Pearson correlation and mutual information stock networks on the global topological scale during the crisis period.

Table 4 Central stocks for each network

	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

Furthermore, an examination is conducted to determine if there is an agreement between the Pearson network and the mutual information network regarding the central stocks. In network analysis, centrality indices are devised to address the question of identifying the most significant vertices in a network. Given that the notion of importance can be described in a variety of forms, numerous centrality measures have been proposed for networks.

In Table 4, it is observed that the distances concur only on the hub, represented by HON, in the non-crisis period. However, they exhibit distinct sets of central stocks for both time periods. The closeness centrality and betweenness centrality indices for the mutual information network consist of the same stocks in both time periods. In contrast, the centrality indices for the Pearson networks indicate a stark contrast in the stocks identified as central.

4.5.2 Impact of metrics on the performance of community detection

Identification of 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 (Nanda et al., 2010).

For an impartial comparison, the same community detection algorithm is employed across all networks. This ensures an unbiased assessment of the six distinct networks constructed using different distance measures. Similar to previous sections, the analysis encompasses two sets of data: the crisis period and the non-crisis period.

The impact of various distance measures on clustering is examined using two approaches. Firstly, the performance is evaluated without the availability of a ground truth cluster set. In this scenario, internal indices are employed to identify which metric yields well-separated and cohesive clusters. Geodesic distances are utilised, and the centroid vertex is selected as the vertex that minimises the maximum distance within each cluster. Secondly, the performance is assessed assuming a prior knowledge of the ground-truth class labels for stocks. Normalised mutual information (NMI) is employed to assess the performance of the metrics. This study considers two types of external information as the ground truth: sector-level and industry-group-level classifications. This allows for the evaluation of the performance of the clustering algorithm and dissimilarity measures in terms of their ability to detect more refined clusters.

Tables 5 and 6 present the performance of the community detection algorithm on six distinct distance measures. The column headings in the tables correspond to the networks generated using different metrics, while the row headings indicate the performance indices. The tables highlight the best scores achieved for each index.

Table 5 Performance of six different distance measures for non-crisis period

	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 5 reveals that mutual information networks demonstrate superior performance compared to other distance measures in terms of both internal and external indices.

While the Pearson network achieves commendable scores in C-index and BetaCV, it is relatively less successful in identifying stock sectors.

Table 6 Performance of six different distance measures for crisis period

	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 6 presents intriguing findings. The Spearman rank order network exhibits the highest score in accurately identifying stock sector and industry group labels. On the other hand, AUV1 demonstrates superior performance in terms of internal indices, indicating that the clusters it identifies are more cohesive and well-separated compared to other distance measures.

To summarise, mutual information-based networks exhibited superior performance in both external and internal indices during the non-crisis period. However, in the crisis period, the Spearman network outperformed in identifying stock sectors, while the mutual information-based networks demonstrated well-structured clusters.

5 Conclusions

This study examines two data sets of daily stock returns for 404 stocks of SP500 stocks, representing a crisis period (2007–2009) and a non-crisis period (2012–2015). During the non-crisis period, the mutual information and Pearson correlation measures show stronger relationships compared to the crisis period. Discrepancies between the measures are investigated using excess information, with the crisis period displaying more significant deviations from Gaussianity.

To explore nonlinear dependencies, polynomial and spline regression techniques are used to calculate model-fitting indices for network adjacencies. The adequacy of regression models is validated using AIC, BIC, and ANOVA tables. The study reveals that a simple regression model is insufficient in describing the relationship between stocks with notable mismatches. Entropy is found to have a strong association with kurtosis, suggesting that mutual information captures more information about probability distributions than Pearson correlation.

Hierarchical stock networks are constructed using the minimum spanning tree method for each distance measure. The local and global characteristics of these networks were found to be less similar for the crisis period, but they exhibit moderate Spearman correlation for the non-crisis period. The impact of different distance measures and community detection performance on the networks is also examined, highlighting the higher similarity of mutual information-based networks in both periods

but their diminished performance in accurately identifying stock sectors during the crisis period. Nevertheless, these networks excel in identifying coherent and distinct stock communities in both periods.

References

Assaf, A., Mokni, K. and Youssef, M. (2023) 'COVID-19 and information flow between cryptocurrencies, and conventional financial assets', *The Quarterly Review of Economics and Finance*, Vol. 89, pp.73–81.

Barabási, A-L. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science*, Vol. 286, No. 5439, pp.509–512.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D-U. (2006) 'Complex networks: structure and dynamics', *Physics Reports*, Vol. 424, Nos. 4–5, pp.175–308.

Butte, A.J. and Kohane, I.S. (1999) 'Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements', *Biocomputing 2000*, World Scientific, pp.418–429.

Cont, R. (2001) *Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues*, Taylor & Francis, Abingdon, UK.

Cover, T.M. and Thomas, J.A. (2006) *Elements of Information Theory*, 2nd ed., Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience, Hoboken, NJ, USA.

Csardi, G. and Nepusz, T. (2006) 'The igraph software package for complex network research', *InterJournal, Complex Systems*, Vol. 1695, No. 5, pp.1–9.

Danon, L., Diaz-Guilera, A., Duch, J. and Arenas, A. (2005) 'Comparing community structure identification', *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2005, No. 9, p.P09008.

de Oliveira Passos, M., Tessmann, M.S., Ely, R.A., Uhr, D. and Taveira, M.T. (2020) 'Effects of volatility among commodities in the long term: analysis of a complex network', *Annals of Financial Economics*, Vol. 15, No. 3, p.2050014.

de Siqueira Santos, S., Takahashi, D.Y., Nakata, A. and Fujita, A. (2013) 'A comparative study of statistical methods used to identify dependencies between gene expression signals', *Briefings in Bioinformatics*, Vol. 15, No. 6, pp.906–918.

Dionisio, A., Menezes, R. and Mendes, D.A. (2006) '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*, Vol. 50, Nos. 1–2, pp.161–164.

Donges, J.F., Zou, Y., Marwan, N. and Kurths, J. (2009) 'Complex networks in climate dynamics', *The European Physical Journal – Special Topics*, Vol. 174, No. 1, pp.157–179.

Dunn, J.C. (1973) *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters*, Taylor & Francis, UK.

Ebrahimi, N., Maasoumi, E. and Soofi, E.S. (1999) 'Ordering univariate distributions by entropy and variance', *Journal of Econometrics*, Vol. 90, No. 2, pp.317–336.

Fiedor, P. (2014) *Mutual Information Rate-Based Networks in Financial Markets*, arXiv preprint arXiv:1401.2548.

Freeman, L.C. (1978) 'Centrality in social networks conceptual clarification', *Social Networks*, Vol. 1, No. 3, pp.215–239.

Freitas, W.B. and Junior, J.R.B. (2023) 'Random walk through a stock network and predictive analysis for portfolio optimization', *Expert Systems with Applications*, Vol. 218, p.119597.

Galton, F. (1889) 'I. Co-relations and their measurement, chiefly from anthropometric data', *Proceedings of the Royal Society of London*, Vol. 45, Nos. 273–279, pp.135–145.

Han, D. et al. (2019) 'Network analysis of the Chinese stock market during the turbulence of 2015–2016 using log-returns, volumes and mutual information', *Physica A: Statistical Mechanics and its Applications*, Vol. 523, pp.1091–1109.

Hanif, W., Ko, H-U., Pham, L. and Kang, S.H. (2023) 'Dynamic connectedness and network in the high moments of cryptocurrency, stock, and commodity markets', *Financial Innovation*, Vol. 9, No. 1, pp.1–40.

Hao, Q., Shen, J.H. and Lee, C-C. (2023) 'Risk contagion of bank-firm loan network: evidence from China', *Eurasian Business Review*, Vol. 13, No. 2, pp.341–361.

Hartman, D. and Hlinka, J. (2018) 'Nonlinearity in stock networks', *Chaos: An Interdisciplinary Journal of Nonlinear Science*, Vol. 28, No. 8, p.083127.

Heiberger, R.H. (2014) 'Stock network stability in times of crisis', *Physica A: Statistical Mechanics and its Applications*, Vol. 393, pp.376–381.

Heiberger, R.H. (2018) 'Predicting economic growth with stock networks', *Physica A: Statistical Mechanics and its Applications*, Vol. 489, pp.102–111.

Hong, M.Y. and Yoon, J.W. (2022) 'The impact of COVID-19 on cryptocurrency markets: a network analysis based on mutual information', *PLOS One*, Vol. 17, No. 2, p.e0259869.

Hubert, L. and Schultz, J. (1976) 'Quadratic assignment as a general data analysis strategy', *British Journal of Mathematical and Statistical Psychology*, Vol. 29, No. 2, pp.190–241.

Jain, A.K. and Dubes, R.C. (1988) *Algorithms for Clustering Data*, Prentice-Hall, Inc., Englewood, Cliffs, NJ.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning*, Vol. 112, Springer, New York, NY, USA.

Khashanah, K. and Yang, H. (2016) 'Evolutionary systemic risk: Fisher information flow metric in financial network dynamics', *Physica A: Statistical Mechanics and its Applications*, Vol. 445, pp.318–327.

Kraskov, A., Stögbauer, H., Andrzejak, R.G. and Grassberger, P. (2005) 'Hierarchical clustering using mutual information', *EPL (Europhysics Letters)*, Vol. 70, No. 2, p.278.

Langfelder, P. and Horvath, S. (2008) 'WGCNA: an R package for weighted correlation network analysis', *BMC Bioinformatics*, Vol. 9, No. 1, pp.1–13.

Luppi, A.I. and Stamatakis, E.A. (2021) 'Combining network topology and information theory to construct representative brain networks', *Network Neuroscience*, Vol. 5, No. 1, pp.96–124.

Maasoumi, E. (1993) 'A compendium to information theory in economics and econometrics', *Econometric Reviews*, Vol. 12, No. 2, pp.137–181.

Mantegna, R.N. (1999) 'Hierarchical structure in financial markets', *The European Physical Journal B – Condensed Matter and Complex Systems*, Vol. 11, No. 1, pp.193–197.

Mantegna, R.N., Stanley, H.E., et al. (2000) *An Introduction to Econophysics: Correlations and Complexity in Finance*, Vol. 9, Cambridge University Press, Cambridge.

McAllester, D. and Stratos, K. (2020) 'Formal limitations on the measurement of mutual information', *International Conference on Artificial Intelligence and Statistics*, PMLR, pp.875–884.

Millington, T. and Niranjan, M. (2021) 'Construction of minimum spanning trees from financial returns using rank correlation', *Physica A: Statistical Mechanics and its Applications*, Vol. 566, p.125605.

Nanda, S., Mahanty, B. and Tiwari, M. (2010) 'Clustering Indian stock market data for portfolio management', *Expert Systems with Applications*, Vol. 37, No. 12, pp.8793–8798.

Newman, M.E. (2002) 'The structure and function of networks', *Computer Physics Communications*, Vol. 147, Nos. 1–2, pp.40–45.

Newman, M.E. and Girvan, M. (2004) 'Finding and evaluating community structure in networks', *Physical Review E*, Vol. 69, No. 2, p.026113.

Onnela, J-P., Chakraborti, A., Kaski, K., Kertesz, J. and Kanto, A. (2003) 'Dynamics of market correlations: taxonomy and portfolio analysis', *Physical Review E*, Vol. 68, No. 5, p.056110.

Onnela, J-P., Kaski, K. and Kertész, J. (2004) 'Clustering and information in correlation based financial networks', *The European Physical Journal B – Condensed Matter and Complex Systems*, Vol. 38, No. 2, pp.353–362.

Pearson, K. (1920) 'Notes on the history of correlation', *Biometrika*, Vol. 13, No. 1, pp.25–45.

Philippatos, G.C. and Wilson, C.J. (1972) 'Entropy, market risk, and the selection of efficient portfolios', *Applied Economics*, Vol. 4, No. 3, pp.209–220.

Prim, R.C. (1957) 'Shortest connection networks and some generalizations', *Bell System Technical Journal*, Vol. 36, No. 6, pp.1389–1401.

Rousseeuw, P.J. (1987) 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, Vol. 20, pp.53–65.

Salgado-Hernández, J. and Vyas, M. (2023) 'Non-linear correlation analysis in financial markets using hierarchical clustering', *Journal of Physics Communications*, Vol. 7, No. 5, p.055003.

Samitas, A., Kampouris, E. and Polyzos, S. (2022) 'COVID-19 pandemic and spillover effects in stock markets: a financial network approach', *International Review of Financial Analysis*, Vol. 80, p.102005.

Semenov, D., Koldanov, A. and Koldanov, P. (2024) 'Analysis of weakly correlated nodes in market network', *Computational Management Science*, Vol. 21, No. 1, pp.1–18.

Shachaf, L.I., Roberts, E., Cahan, P. and Xiao, J. (2023) 'Gene regulation network inference using k-nearest neighbor-based mutual information estimation: revisiting an old dream', *BMC Bioinformatics*, Vol. 24, No. 1, p.84.

Shannon, C.E. (1948) 'A mathematical theory of communication', *The Bell System Technical Journal*, Vol. 27, No. 3, pp.379–423.

Song, L., Langfelder, P. and Horvath, S. (2012) 'Comparison of co-expression measures: mutual information, correlation, and model based indices', *BMC Bioinformatics*, Vol. 13, No. 1, p.328.

Soofi, E.S. (1997) 'Information theoretic regression methods', *Applying Maximum Entropy to Econometric Problems*, pp.25–83, Emerald Group Publishing Limited.

Tumminello, M., Aste, T., Di Matteo, T. and Mantegna, R.N. (2005) 'A tool for filtering information in complex systems', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102, No. 30, pp.10421–10426.

Tzouras, S., Anagnostopoulos, C. and McCoy, E. (2015) 'Financial time series modeling using the hurst exponent', *Physica A: Statistical Mechanics and its Applications*, Vol. 425, pp.50–68.

Vidal-Tomás, D. (2021) 'Transitions in the cryptocurrency market during the COVID-19 pandemic: a network analysis', *Finance Research Letters*, Vol. 43, p.101981.

Vidya, C., Mummidi, S. and Adarsh, B. (2023) 'Effect of the COVID-19 pandemic on world trade networks and exposure to shocks: a cross-country examination', *Emerging Markets Finance and Trade*, Vol. 59, No. 3, pp.863–879.

Wang, T., Zhao, S., Wang, W. and Yang, H. (2023) 'How does exogenous shock change the structure of interbank network?: Evidence from China under COVID-19', *Emerging Markets Finance and Trade*, Vol. 59, No. 4, pp.937–958.

Wu, J., Xu, K., Chen, X., Li, S. and Zhao, J. (2022) 'Price graphs: utilizing the structural information of financial time series for stock prediction', *Information Sciences*, Vol. 588, pp.405–424.

Wu, Z. and Leahy, R. (1993) 'An optimal graph theoretic approach to data clustering: theory and its application to image segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 11, pp.1101–1113.

Zaki, M.J. and Meira, J.W. (2014) *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, USA.

Zhang, B. and Horvath, S. (2005) 'A general framework for weighted gene co-expression network analysis', *Statistical Applications in Genetics and Molecular Biology*, Vol. 4, No. 1, Article 17.

Zhang, W. and Zhuang, X. (2019) 'The stability of Chinese stock network and its mechanism', *Physica A: Statistical Mechanics and its Applications*, Vol. 515, pp.748–761.

Zhang, X., Zhang, T. and Lee, C-C. (2022) 'The path of financial risk spillover in the stock market based on the R-vine-copula model', *Physica A: Statistical Mechanics and its Applications*, Vol. 600, p.127470.

Zhou, Y., Chen, Z. and Liu, Z. (2023) 'Dynamic analysis and community recognition of stock price based on a complex network perspective', *Expert Systems with Applications*, Vol. 213, p.118944.