mathematics

MDPI

Article

Distance Correlation Market Graph: The Case of S&P500 Stocks

Samuel Ugwu [1,*], Pierre Miasnikof [1,2] and Yuri Lawryshyn [1,3]

[1] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON M5S 3G8, Canada
[2] Data Science Institute, University of Toronto, Toronto, ON M5G 1Z5, Canada
[3] Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON M5S 3E5, Canada
* Correspondence: samuel.ugwu@utoronto.ca

Abstract: This study investigates the use of a novel market graph model for equity markets. Our graph model is built on distance correlation instead of the traditional Pearson correlation. We apply it to the study of S&P500 stocks from January 2015 to December 2022. We also compare our market graphs to the traditional market graphs in the literature, those built using Pearson correlation. To further the comparison, we also build graphs using Spearman rank correlation. Our comparisons reveal that non-linear relationships in stock returns are not captured by either Pearson correlation or Spearman rank correlation. We observe that distance correlation is a robust measure for detecting complex relationships in S&P500 stock returns. Networks built on distance correlation networks, are shown to be more responsive to market conditions during turbulent periods such as the COVID crash period.

Keywords: distance correlation; complex networks; market graph; stock market network; non-linearity; investment science

MSC: 68R10

check for updates

Citation: Ugwu, S.; Miasnikof, P.; Lawryshyn, Y. Distance Correlation Market Graph: The Case of S&P500 Stocks. *Mathematics* **2023**, *11*, 3832. https://doi.org/10.3390/math11183832

Academic Editors: Nazim Choudhury and Matloob Khushi

Received: 8 August 2023
Revised: 4 September 2023
Accepted: 5 September 2023
Published: 7 September 2023

1. Introduction

The goal of this work is to test a relatively novel correlation measure, distance correlation, in market graph modeling applications. Traditionally, market graph models are built using the Pearson correlation coefficient [1]. Our graph models are built using the more flexible and more robust distance correlation [2]. We use the constituents of the S&P500 index to illustrate our model. Our empirical findings suggest that distance correlation more effectively captures non-linear relationships and interconnectedness between asset returns compared to traditional correlation measures.

Correlations between asset returns are vital in portfolio optimization, risk assessment and hedging strategies [3]. Typically, stock correlations are measured using the Pearson correlation coefficient, which estimates the linear relationship between two random variables [4]. Unfortunately, Pearson correlation, while popular, has several limitations [5,6]. First, it only measures linear relationships. Second, it is sensitive to outliers, which can result in misleading conclusions [7]. Third, a Pearson correlation coefficient of zero does not imply independence between the variables. It only indicates the absence of a linear relationship, which does not rule out the presence of non-linear relationships [8]. However, it has been increasingly recognized that the relationships between financial assets are not always linear and can often be influenced by market conditions [9], investor behavior [10], company-specific factors [11] or economic factors [12]. To overcome these limitations, we use distance correlation, a non-parametric approach to correlation analysis [2]. Distance correlation captures both linear and non-linear relationships. Additionally, a distance correlation of zero implies complete independence between variables.

In parallel, complex network analysis has emerged as a tool for studying large systems comprising numerous interconnected elements [13]. Hence, complex network analysis is often used to study stock markets. Networks (graphs) are ideally suited to capture interactions in large systems with interacting components [14]. In a market graph, each node represents an asset (typically a stock) and the edges between nodes indicate the correlations between their returns [3]. These networks have been instrumental in analyzing market dynamics and predicting future prices [15]. Various algorithms such as the Minimum Spanning Tree (MST), Planar Maximally Filtered Graph (PMFG) [16] and Correlation Coefficient Threshold Method have been used to construct stock networks [17–20]. Most of these networks are constructed using Pearson correlation, which, as described earlier, lacks robustness and does not capture non-linear relationships. Our work seeks to address these shortcomings, through the use of distance correlation [2].

While traditional methods often assume financial returns to be independent and identically distributed (i.i.d.) [21], we recognize the complexity of real-world financial data, where returns may exhibit non-linear and interdependent relationships. By employing distance correlation, we aim to capture these intricate dynamics, offering a more nuanced understanding of financial networks and asset correlations, even in non i.i.d. scenarios [2]. Using distance correlation, we build our market graph, a network of stock index constituents. We then analyze its topological properties, to gain insight into market dynamics. In this article, our analysis is applied to the constituents of the Standard and Poor's S&P500 index for the period between January 2015 and December 2022. The remainder of this article is organized as follows. Section 2 contains a review of the relevant literature. Section 3 introduces distance correlation and outlines the methodology for constructing our market graph. Section 4 presents the results of our graph-based analysis to the S&P500 index. We include an examination of the topological properties of the resulting networks. Section 5 concludes this article with a summary and discussion of the implications of our findings.

2. Literature Review

Graphs have been used in the study of financial markets in the past. Indeed, graphs, mathematical models of networks, have been shown to be useful tools to model complex relationships between assets. Early work by Mantegna, R. N. (1999); Onnela et al. (2003); Boginski et al. (2004, 2006); Tse et al. (2010) [1,22–25] pioneered the use of graph models in finance. Mantegna, R. N. (1999)'s [1] work on the hierarchical structure in financial markets using the Minimum Spanning Tree (MST) is considered seminal. Further works by Shirokikh et al. (2013); Faizliev et al. (2019); Semenov et al. (2023) [26–28] analyze stock markets using graphs built using Pearson, Spearman rank and Kendall correlation, respectively. Millington and Niranjan's (2021) [29] work compares networks built using these different correlations. In doing so, they highlight the sensitivity of Pearson correlation. More recently, research has focused on the uncertainty surrounding various network structures, such as maximum cliques, maximum independent sets, maximum spanning trees and planar maximally filtered graphs [30]. Authors in the field have also introduced a novel measure of similarity based on the probability of the coincidence of signs of stock returns, a distribution-free statistical procedure for threshold graph identification and offered an analysis of the evolution of market graph structures over time [31–33].

As described earlier, financial networks were typically constructed using Pearson correlation. However, Guo et al. (2018) [12] critically evaluated Pearson correlation in the specific context of financial returns. They highlighted its inability to capture non-linear relationships, especially during turbulent market conditions. To address this limitation, these authors proposed using mutual information coefficients for stock correlation networks. While networks built using mutual information address the limitations of Pearson correlation and capture non-linear relationships between stocks, the method is constrained by the necessity of modeling the probability distributions of the variables, a challenging task [34].

In addition to these developments, the literature has seen a growing interest in understanding the interconnectedness and systemic risk within financial institutions. Wang et al. (2017) [35] proposed an extreme risk spillover network using the CAViaR technique and Granger causality risk test, identifying the real estate and bank sectors as net senders of extreme risk spillovers. Diebold and Yılmaz (2014) [36] introduced connectedness measures derived from variance decomposition, offering a more nuanced understanding of connectedness in the financial context. Corsi et al. (2018) [37] further contributed to the understanding of financial distress propagation through Granger-causality tail risk networks, focusing on systemically important banks and sovereign bonds. This approach allowed for the identification of flight-to-quality dynamics and the propagation of financial distress through interconnected financial institutions. Billio et al. (2012) [38] emphasized the increased interrelation among hedge funds, banks, broker/dealers, and insurance companies over the past decade, proposing econometric measures of connectedness based on principal components analysis and Granger-causality networks. These measures were found to be useful out-of-sample indicators of systemic risk, highlighting the multi factorial nature of systemic risk and the importance of understanding the connections and interactions among financial institutions. Guowei Song et al. (2023) [15] introduced the Multi-relational Graph Attention Ranking (MGAR) network, which dynamically captures stock relationships to predict return rankings. While our work emphasizes distance correlation to detect complex relationships in equity markets, the MGAR network offers a distinct approach by focusing on stock ranking prediction through adaptive learning mechanisms.

Distance correlation is an alternative correlation measure that does not require prior knowledge or assumptions of probability distributions. It also captures linear and nonlinear relationships [39]. A study by Hou et al. (2022) [8] demonstrated the superiority of distance correlation over Pearson and Spearman rank correlations in measuring complex relationships between gene profiles. In a similar application, Liu et al. (2021) [40] employed distance correlation in constructing gene co-expression networks. These authors successfully captured the complex relationships without any assumptions regarding distributions.

Distance correlation has also been applied to the study of the stock market. For example, ref. [41] used this measure to analyze the dependencies in stock returns. Their study showed that distance correlation effectively detected serial dependence without requiring data transformations. The superiority of distance correlation in the context of financial markets has been highlighted again in the very recent literature [42]. This article seeks to fill the literature gap by exploring the application of distance correlation to financial networks and its potential to provide deeper insight into the complex interactions between assets in the stock market.

3. Methodology

Using a dataset comprising the adjusted daily closing prices of the S&P500 index constituents from January 2015 to December 2022, we study the inter-dependencies between stocks. We use distance correlation, a statistical measure of dependence that overcomes the limitations of traditional correlation measures and captures non-linear relationships. The methodology outlined in this section describes distance matrix construction and distance covariance estimation.

We construct a fully connected weighted graph to represent the relationships between stocks, utilizing the distance correlation coefficient of stock pairs. This graph illustrates how stocks are interconnected based on their similarities. To identify the most significant connections while minimizing dissimilarity, we apply Prim's algorithm [43] to obtain the Minimum Spanning Tree (MST) of the graph. This process reveals the hierarchical organization of the stock market, highlighting the influence of individual stocks within the broader market context. All experiments were carried out using Python 3.7.13, leveraging its extensive libraries and tools for data analysis. The dataset employed in this analysis is widely recognized and reflective of broad market trends, offering a robust foundation for future exploration of stock market dynamics.

3.1. Data

We compiled the adjusted daily closing prices of the S&P500 index constituents for the period between January 2015 and December 2022. Adjusted market prices are stock prices modified to account for corporate actions such as stock splits or dividend payments. These data were obtained from Bloomberg. They consist of 500 assets and 2015 observations. We then compute log returns on the adjusted market prices of each stock. Our analysis is based on this returns series of 2014 observations. The log return r_t of a stock for a given day t is calculated as

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right),$$

where P_t is the adjusted market price on day t and P_{t-1} is the adjusted market price on the previous day.

Missing Values. The dataset was scrutinized for missing values, which are common in financial time-series data. To maintain data integrity and continuity, stocks with missing values were removed from our sample. About 12% of stocks in our dataset had instances of missing values, a phenomenon that can be attributed to several factors inherent to the dynamic nature of financial markets. Among these factors, changes in index composition and delisting play significant roles, which is why we only use 466/500 constituents of the S&P500 index constituents. As highlighted in Table 1, the aggregate descriptive statistics underscore a mean return close to zero, with slight negative skewness, indicating potential downside outliers and a high kurtosis suggesting the presence of heavy tails or extreme values. The removal of stocks with missing values was essential to maintain data integrity and continuity, ensuring that the analysis was conducted on a complete and consistent dataset. This decision helped avoid potential biases and inaccuracies that could arise from filling in missing values, preserving the true dynamics of the stock market in our analysis.

Table 1. Aggregate descriptive statistics of data.

Statistic	**Value**
Mean	0.000369
Median	0.000770
Standard Deviation	0.019838
Skewness	−0.546825
Kurtosis	14.922314
Minimum	−0.190508
Maximum	0.156763
Q1	−0.008492
Q3	0.009884

3.2. Distance Correlation

Distance correlation, introduced by Sźekely et al. (2007) [2], is a non-parametric measure of dependence that captures non-linear relationships between variables. The first step in obtaining the distance correlation between two stocks is to compute the (returns) distance matrices A and B. Each matrix A and B has dimensions $n \times n$, where n represents the number of observations (time points) in the log return data (one matrix per stock). The elements $A_{k\ell}$ and $B_{k\ell}$ are the Euclidean distances between the log returns at time points k and ℓ:

$$A_{k\ell} = \|X_k - X_\ell\|, \quad B_{k\ell} = \|Y_k - Y_\ell\|, \quad \forall k,\ell \in \{1,\ldots,n\}, \tag{1}$$

where X_k and Y_k represent the log returns of stock X and stock Y at time k, respectively.

3.2.1. Calculation of Distance Covariance

The distance covariance measures the dependence between two stocks with (returns) distance matrices A and B. To obtain the distance covariance, we center the distance matrices A and B to obtain matrices C and D, respectively. The elements $C_{k\ell}$ and $D_{k\ell}$ of matrices C and D are computed as follows:

$$C_{k\ell} = A_{k\ell} - \overline{A}_k - \overline{A}_\ell + \overline{A}, \quad D_{k\ell} = B_{k\ell} - \overline{B}_k - \overline{B}_\ell + \overline{B}\,. \tag{2}$$

Here, $\overline{A}_k$, $\overline{B}_k$, $\overline{A}_\ell$, $\overline{B}_\ell$ are the row means for rows k and ℓ, respectively. Meanwhile, $\overline{A}$, and $\overline{B}$ are the overall means of the elements in matrices A and B.

With the centered distance matrices C and D, we compute the distance covariance $V_n^2(X,Y)$ as

$$V_n^2(X,Y) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{\ell=1}^{n} C_{k\ell}D_{k\ell}\,. \tag{3}$$

3.2.2. Calculation of Distance Variance and Distance Correlation

The distance variance $V_n^2(X,X)$ captures the self-dependence of the stock with distance matrix A. It is computed as follows:

$$V_n^2(X,X) = \frac{1}{n^2}\sum_{k,\ell=1}^{n} C_{k\ell}^2 \tag{4}$$

The distance correlation $R_n(X,Y)$ between the two stocks is the square root of the following equation: (i.e., $R_n(X,Y) = \sqrt{R_n^2(X,Y)}$)

$$R_n^2(X,Y) = \frac{V_n^2(X,Y)}{\sqrt{V_n^2(X,X) \times V_n^2(Y,Y)}}\,, \tag{5}$$

$V_n^2(X,X)$ = 0 if every observation in a stock is identical. The distance correlation method outlined above, including the computation of distance variance and distance correlation, is implemented using the `distance_correlation` function in Python's `statsmodels` package [44].

3.3. Market Graph Construction Using Correlations

In constructing the market graph, we begin with a complete weighted undirected graph. Our set of assets is represented by a graph $G = (V,E)$, where $V = \{v_1,\ldots,v_n\}$ is the set of vertices and $E = \{e_{i,j},\ldots,e_{n-1,n}\}$ is the set of edges. Each stock is represented by a vertex. Edge weights are the dissimilarity measure based on the correlation between stocks i and j.

One of the stated objectives of this work is to provide the reader with a comparison of graph models based on distance and Pearson correlations. For this comparison, we build two market graph models of our universe of stocks (S&P500 2015–2022) described earlier. One is built using Pearson correlation, the other with distance correlation.

3.3.1. Pearson Correlation Dissimilarity

The Pearson correlation coefficient for two random variables x and y, denoted as ρ_{xy}, measures the linear relationship between them. To obtain the corresponding dissimilarity measure $D(x,y)$, we apply the following transformation:

$$D(x,y) = \sqrt{2(1-\rho_{xy})}. \tag{6}$$

The resulting $D(x,y)$ ranges from 0 to 2, where 0 indicates a perfect positive linear relationship, and 2 indicates a perfect negative linear relationship.

Edge weights in our Pearson correlation graph correspond to this quantity. Here, each stock is a random variable represented by a vertex. Edge weights correspond to the dissimilarity between each vertex pair.

3.3.2. Distance Correlation Dissimilarity

As described earlier, distance correlation, denoted as $R_n(x,y)$, is a non-parametric measure of dependence that captures both linear and non-linear relationships between variables. To obtain the dissimilarity measure $D(x,y)$, we apply the following transformation:

$$D(x,y) = 1 - R_n(x,y). \tag{7}$$

The resulting $D(x, y)$ ranges from 0 to 1, where 0 indicates a perfect association between variables, and 1 indicates complete dissimilarity. This transformation is an extension of the Pearson dissimilarity metric [1] above to distance correlation.

Here again, we build a (complete weighted) graph where each stock is a random variable represented by a vertex. Edge weights correspond to the dissimilarity between each vertex pair.

3.3.3. Minimum Spanning Tree (MST) Construction

Once the dissimilarity measure $D(v_x, v_y)$ is obtained for all stock pairs, we use Prim's algorithm [43] to obtain the MST, which connects all nodes in the network with the minimum sum of edge weights. The resulting MST captures the hierarchical organization of the stock market and provides valuable insight into the relationships between stocks.

Prim's algorithm starts with an arbitrary node and iteratively adds the edge with the shortest distance (smallest dissimilarity measure) connecting the current node to an unreached node. The process is repeated until all nodes are included in the MST.

The market graph, obtained by constructing the MST based on either Pearson correlation or distance correlation, can be visualized using Python's `NetworkX` library. By comparing the MSTs constructed using these two correlation methods, we gain insight into the limitations of linearity constraints in studies of the stock market. Full details are presented in the next section.

3.4. Market Graph Comparisons and Centrality Measures

In our analysis, we compare the MSTs constructed from distance correlation graphs to those obtained from Pearson correlation graphs. This comparison is focused on two (global) key graph characteristics, average node-node distance and longest node–node distance. At a more microscopic level, our comparisons also consider stock-level features. We use node degree to measure the significance of individual nodes (stocks) within each network.

Finally, we compare sector centrality, as Millington and Niranjan (2021) [29] did in their recent work. Like these authors, we evaluate betweenness and degree centrality at the sector level. Here, we use the Morgan Stanley Capital International-Standard & Poors Global Industry Classification Standard (GICS). GICS is a widely used classification of companies into economic sectors, based on their primary business activities.

In summary, our comparison of distance and Pearson correlation MSTs provides an assessment of their ability to accurately model markets. These comparisons offer valuable insight into each model's ability to capture market cohesion, interconnection, and the significance of individual sectors. It also offers perspective into the influence of nonlinear relationships and hidden dependencies on market dynamics.

3.4.1. Average Distance (AD) and Longest Distance (LD)

The AD of the MST, the mean distance separating vertices on the tree, provides crucial insight into the market's overall cohesion [45]. Indeed, in our market graph model, edge weights represent return dissimilarity. Therefore, a larger AD indicates weaker correlations between stocks, on average [22,45]. It indicates a more scattered and diversified market. In such a scenario, individual stock movements are less influenced by overall market trends than by idiosyncratic factors. Overarching market trends are not as obvious. Conversely, a smaller AD signals a more interconnected and cohesive market, with higher average correlations between stocks. In this case, overall market movements have a more pronounced effect on individual stock behavior. Market trends are also easier to infer in such cases.

The LD represents the maximum distance between any two stocks (vertices) on the MST. It measures the maximum possible dissimilarity between a pair of assets. A larger LD value suggests a market with greater dispersion and heterogeneity among its constituents. It means idiosyncratic factors affect returns more strongly than in instances with a shorter LD. This dispersion may also indicate the presence of subsets of stocks that are weakly correlated to the broader population (index in this case) Birch et al. (2016)'s article [46]. Identifying these subsets is useful for portfolio diversification. On the contrary, a smaller LD in the MST indicates a more interconnected and cohesive market, where the distance between individual stocks is relatively limited. Such a tightly knit market structure implies that the majority of stocks are influenced by similar market forces and tend to move in tandem with overall market trends. This cohesion can make it easier to identify overarching market trends and predict the behavior of individual stocks based on broader market movements.

3.4.2. Centrality Measures: Betweenness Centrality (BC) and Degree Centrality (DC)

As mentioned earlier, our analysis also extends to market sectors. We calculate two centrality measures for each market sector, Betweenness Centrality (BC) and Degree Centrality (DC). To obtain these quantities, we begin by aggregating stocks into their respective GICS sectors. We then compute our aggregate centrality measures for each sector, as follows.

Betweenness Centrality (BC): Betweenness Centrality of a node (stock) measures the number of shortest paths between all pairs of nodes that pass through that node [47]. For a sector, the aggregate BC of its constituent nodes quantifies its importance. A high BC indicates that a sector is highly connected to other sectors and critical to maintaining connections between sectors in the market [45]. It signals that the sector acts as a bridge between other sectors and has a significant influence on broader market dynamics and information flow. Betweenness Centrality (BC) for a sector in the MST is computed as follows:

$$BC(\text{sector}) = \frac{\text{sum of shortest paths passing through the sector}}{\text{total number of shortest paths on the tree} - 1}\,.$$

Degree Centrality (DC): Degree Centrality of a node (stock) represents the number of edges (connections) that a node has in the network [29]. For sectors, it measures the aggregate number of connections a sector has with other sectors. A high degree centrality for a sector implies that it is strongly connected to other sectors and plays a critical role in shaping market relationships. Degree Centrality (DC) for a sector is computed as follows:

$$DC(\text{sector}) = \text{sum of degree centrality of nodes in the sector}\,.$$

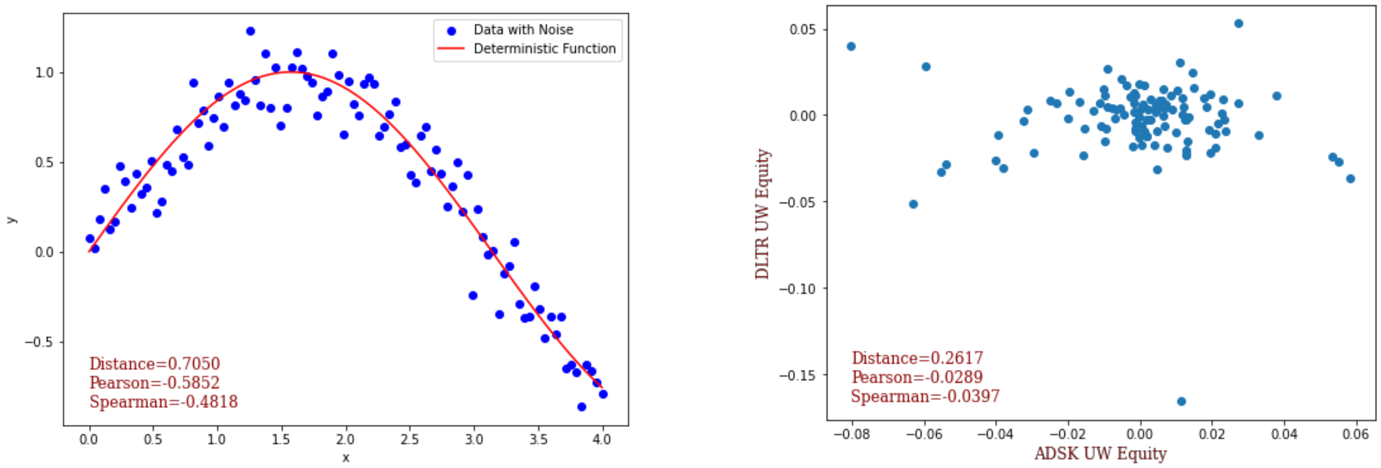
4. Results

We begin our empirical comparisons with an examination of three correlation measures: Pearson correlation, Spearman rank correlation, and distance correlation, using synthetic data. We generate a uniform random variable X on the interval $(0, 4)$. We then generate Y, another random variable that is a non-linear transformation of X. The objective of this initial comparison is to illustrate the ability (or inability) of each correlation coefficient to capture non-linear relationships between random variables. Results are reported in Table 2. They show that distance correlation outperforms Pearson and Spearman correlations in capturing non-linear associations. Notably, the correlation coefficients obtained using distance correlation were consistently higher for all non-linear scenarios. For instance, in the case of $Y = X^2$, distance correlation yielded a correlation coefficient of 0.54, while Pearson and Spearman correlations produced much lower values of 0.11 and 0.03, respectively. Similarly, for $Y = \sin(X)$ and $Y = \cos(X)$, distance correlation achieved significantly higher coefficients than the other correlation measures.

Figure 1a,b further highlight distance correlation's superiority in cases of non-linear relationships between variables. Even in the presence of noise, as in the example shown in Figure 1a, distance correlation (DCor) clearly outperforms the other correlations. Meanwhile, Figure 1b shows a very noisy real-world example of a non-linear relationship between stock returns. It displays the relationship between the returns of Autodesk and Dollar Tree, for the period of January 2020–June 2020. We observe that Pearson and Spearman rank correlation coefficients are close to zero, indicating the absence of a relationship. In contrast, distance correlation is significant.

Table 2. Correlation coefficients ρ_{XY} for random variables X and Y (sample size n = 1000).

Relationship	Pearson	Spearman	Distance
$Y = X^2$	0.1106	0.0256	0.5450
$Y = \sin(X)$	0.9083	0.9755	0.9679
$Y = \cos(X)$	−0.0947	−0.0257	0.5537
$Y = e^X$	0.7442	1	0.9170



(**a**) Generated example Y = sin(x) + noise

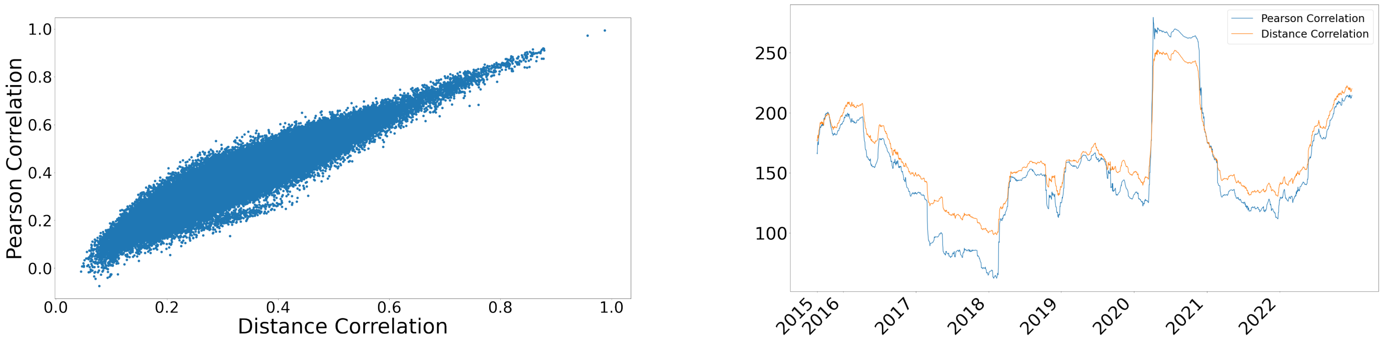
(**b**) Real World Example Autodesk vs. DollarTree

Figure 1. Scatter plots of non-linear relationships.

4.1. Correlation Matrix Analysis

This analysis consists of a comparison of correlation matrices containing Pearson and distance correlations. These coefficients are computed using the log returns of S&P500 index constituents for the period between January 2015 and December 2022. Figure 2a highlights the divergence between Pearson correlation and distance correlation. While both the coefficients are positively related, a substantial divergence is observed. The correlations observed in the analysis are typically positive, providing a coherent basis for making a meaningful comparison between distance and Pearson correlation methods.

Next, we analyze the maximum eigenvalue of correlation matrices using a sliding window of six months. The maximum eigenvalue is a measure of the intensity of the correlation. In matrices inferred from financial returns, the maximum eigenvalue tends to be significantly larger than the second maximum [29]. We study the evolution over time of the maximum eigenvalue of both matrices. Results are presented in Figure 2b. Both correlations follow a similar trend, with the maximum eigenvalue peaking during times of market downturn and dropping during times of low market volatility, which concurs with the findings of Drożdż et al. (2000) [48]. For the given period 2015–2020, the maximum eigenvalue has a lower range for distance correlation than Pearson correlation. During periods of market crashes and volatility, returns data often exhibit a higher occurrence of outliers. This increase in outlier presence could potentially account for the observed disparity in the maximum eigenvalue. However, distance correlation demonstrates a more stable eigenvalue range than Pearson correlation for all periods, as shown in Table 3.



(**a**) Distance Correlation vs. Pearson Correlation (**b**) Maximum Eigenvalue per period

Figure 2. Correlation coefficient properties.

Table 3. Range of eigenvalues (Max–Min).

Year	Pearson Correlation	Distance Correlation
2015	34.82	23.42
2016	69.36	61.90
2017	69.18	48.68
2018	91.05	61.51
2019	41.80	36.98
2020	157.13	112.88
2021	65.47	47.30
2022	89.81	83.12

4.2. Market Graph Analysis

The characteristics of networks based on distance correlation (DCor) and Pearson correlation (PCor) are presented in Table 4. Three categories of stocks were identified based on their degrees, as described in Guo et al. (2018) [12]. Pivotal nodes with degrees exceeding seven act as influential hubs, facilitating information, resource and influence flow. Stocks with degrees ranging from three to seven serve as conduits for propagating market information. Stocks with degrees of one to two have limited impact on network dynamics and do not act as significant hubs. The majority of nodes in these MST networks have a degree of one to two.

Table 4. Network topology properties.

Network	Degree Distribution			Topology	
	1–2	3–7	$\geq$8	AD	LD
PCor 1 (January 2015–December 2022)	79.81%	17.68%	2.51%	13	35
DCor 1 (January 2015–December 2022)	79.59%	18.31%	2.1%	14	38
PCor 2 (January 2020–June 2020)	75.51%	22.22%	2.27%	15	36
DCor 2 (January 2020–June 2020)	78%	19.71%	2.29%	10	22

Distance correlation networks exhibit distinct properties and dynamics. They differ especially from Pearson correlation networks during periods of market turbulence, such as the COVID crash in 2020. These networks capture non-linear dependencies. This dependency detection translates into changes in network structure. For example, during the period between January and June 2020, which corresponds to the COVID crash, the distance correlation network displayed a more compact and efficient connectivity pattern than the Pearson correlation network. Specifically, the AD and LD between nodes, shown in Table 4, were lower for the DCor network. The AD for the DCor network is 10, while the AD for the PCor network is 15. Similarly, the LD for the DCor network is 22, whereas the LD for the PCor network is 36. These observations suggest that the distance correlation network captures interconnectedness and non-linearity between nodes, during the turbulence of the COVID crash. These results are consistent with the findings of Onnela et al. (2003) [22]. These authors described market graph shrinking during market crashes. These authors documented shrinking average distance (AD) and longest distance (LD) during periods of turbulence.

Tables 5 and 6 show network characteristics for S&P500 sectors during the periods from January 2020 to June 2020 and January 2015 to December 2022 using DCor and PCor. Sector-specific dynamics within the networks are also evident. The Information Technology sector ranks highly in distance correlation LD during the crisis, as shown in Table 5, highlighting its heightened importance and influence during the pandemic. Conversely, the Consumer Staples sector consistently maintains a high rank across both periods, indicating the essential nature of its products and services. These sector-specific patterns provide valuable insight into the interconnections and behavior of stocks.

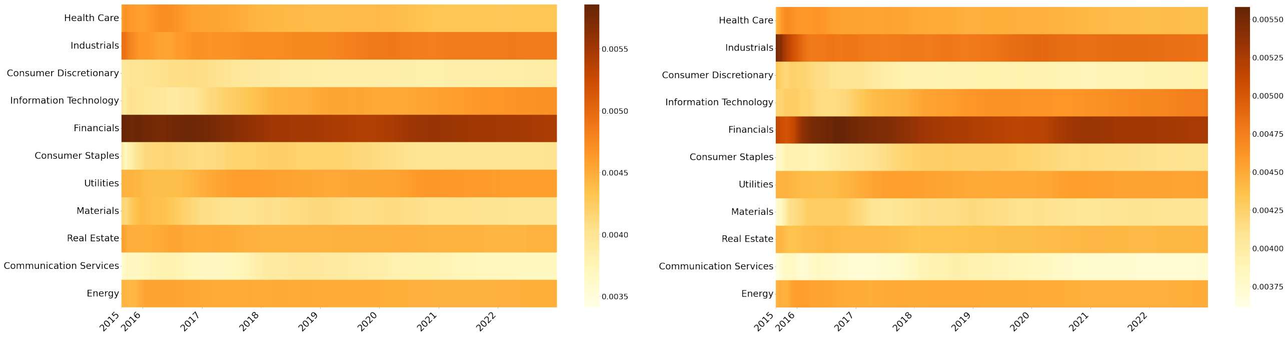
Table 5. Network topology properties (longest distance) per sector analysis.

	January 2020–June 2020		**January 2015–December 2022**	
Sector	**DCor**	**PCor**	**DCor**	**PCor**
Information Technology	21	32	17	15
Materials	17	30	11	11
Utilities	9	11	7	10
Health Care	14	33	25	22
Industrials	18	30	31	22
Real Estate	15	28	11	22
Consumer Discretionary	19	26	29	29
Consumer Staples	18	32	32	30
Energy	14	14	10	12
Financials	16	31	26	22
Communication Services	18	22	27	26

Table 6. Network topology properties (average distance) per sector analysis.

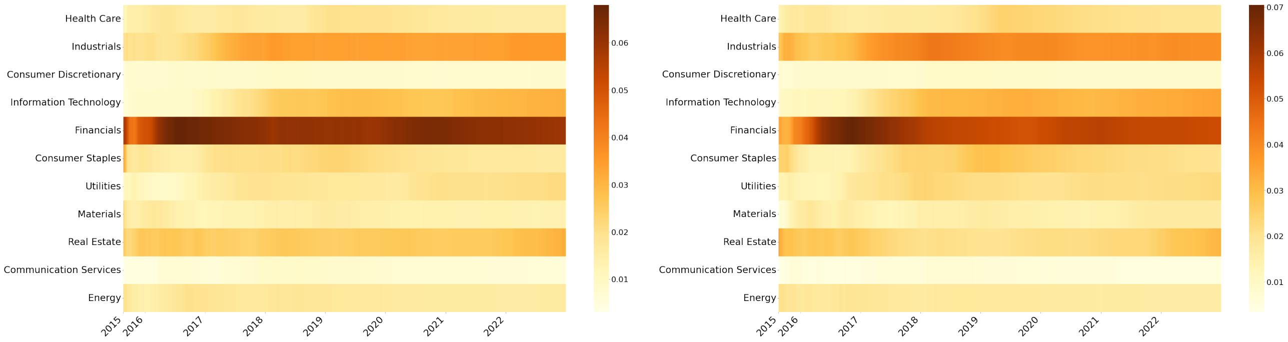
	January 2020–June 2020		**January 2015–December 2022**	
Sector	**DCor**	**PCor**	**DCor**	**PCor**
Information Technology	9.91	10.14	6.14	5.66
Materials	7.22	11.5	4.46	5.29
Utilities	3.89	4.98	3.26	4.35
Health Care	7.26	12.51	10.27	10.47
Industrials	7.92	12.8	8.76	7.21
Real Estate	6.84	12.61	4.76	6.64
Consumer Discretionary	9.73	13.35	12.41	12.78
Consumer Staples	10.35	16.25	14.3	14.35
Energy	5.31	6.6	4.16	4.72
Financials	6.91	10	12.12	7.98
Communication Services	8.85	12.6	14.29	10.56

From 2015 to 2022, we conduct a comparative analysis of sector degree centrality using Pearson correlation and distance correlation methods on a sliding window of six months. Figure 3a,b shows Financial and Industrial sectors consistently exhibiting significance in both methods throughout the entire dataset. Additionally, the analysis highlights a more significant rise in the importance of the Technology sector for both networks from 2018 to 2022 . In early 2015, the influence of the Industrial and Consumer Discretionary sectors appeared more pronounced in the distance correlation-based trees than in the Pearson correlation-based trees. In the analysis of betweenness centrality from 2015 to 2020 shown in Figure 4a,b, similar observations were made using both Pearson correlation and distance correlation methods. The Financials and Industrials sectors consistently exhibited high betweenness centrality, indicating their crucial role in connecting various sectors within the market. These sectors served as key "bridges", facilitating information flow and influencing market dynamics.



(**a**) Pearson Correlation (**b**) Distance Correlation

Figure 3. Degree centrality.



(**a**) Pearson Correlation (**b**) Distance Correlation

Figure 4. Betweeness centrality.

5. Conclusions

In summary, this paper focuses on the construction of market graphs using distance correlation as the underlying measure of correlation. The study compares distance correlation with traditional measures such as Pearson and Spearman rank correlation, using both real-world data from the S&P500 index and synthetic data. These findings from the study clearly indicate that distance correlation offers a more comprehensive understanding of the non-linear relationships inherent in financial markets, as compared to traditional linear correlation measures. This underscores the relevance and significance of accounting for non-linearity in market graph construction and analysis. Constructing market graphs based on distance correlation provides additional insight into the structure and dynamics of the stock market. The analysis reveals higher degree distribution and shorter average distances in distance correlation networks, suggesting an interconnected market structure during specific periods, including market crashes. These findings are entirely consistent with the broader literature, beyond market graphs. The utilization of distance correlation in market graph construction underscores the necessity of refining our analytical methodologies to more accurately capture non-linear relationships in financial markets, especially when compared to traditional correlation measures, suggesting a direction for enhanced precision in future financial analyses.

It is important to acknowledge the limitations of the study. This work was only applied to the S&P500 index, which represents a subset of the overall stock market. Future research should expand the analysis to include other market indices and time periods to validate and extend the findings. Additionally, comparing distance correlation-based networks with other non-linear dependency-based networks would contribute to a comprehensive understanding of network construction approaches in capturing complex relationships in financial data.

Author Contributions: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by S.U. and P.M. The first draft of the manuscript was written by S.U. and all authors commented on previous versions of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Center for Management of Technology & Entrepreneurship (CMTE).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Bloomberg and are available with the permission of Bloomberg.

Acknowledgments: The authors thank John Lukovich, Harshitha Janakiraman, Shreya Patki, Tushar Verma and Christoph Lohrmann for their careful review and advice.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mantegna, R.N. Hierarchical structure in financial markets. *Eur. Phys. J. B* **1999**, *11*, 193–197. [CrossRef]
2. Székely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794. [CrossRef]
3. Czasonis, M.; Kritzman, M.; Turkington, D. The Stock-Bond Correlation. *J. Portf. Manag.* **2021**, *47*, 67–76. [CrossRef]
4. Dunlap, H.F. An Empirical Determination of the Distribution of Means, Standard Deviations and Correlation Coefficients Drawn from Rectangular Populations. *Ann. Math. Stat.* **1931**, *2*, 66–81. [CrossRef]
5. Coppack, S.W. Correspondence Section: Limitations of the Pearson Product-Moment Correlation. *Clin. Sci.* **1990**, *79*, 287. [CrossRef]
6. Richard, A. Should Pearson's correlation coefficient be avoided? *Ophthalmic Physiol. Opt.* **2019**, *39*, 316–327. [CrossRef]
7. Pernet, C.; Wilcox, R.; Rousselet, G. Robust Correlation Analyses: False Positive and Power Validation Using a New Open Source Matlab Toolbox. *Front. Psychol.* **2012**, *3*, 606. [CrossRef]
8. Hou, J.; Ye, X.; Feng, W.; Zhang, Q.; Han, Y.; Liu, Y.; Li, Y.; Wei, Y. Distance correlation application to gene co-expression network analysis. *BMC Bioinform.* **2022**, *23*, 81. [CrossRef]
9. Campbell, J.Y.; Lo, A.W.; MacKinlay, A. *The Econometrics of Financial Markets*; Princeton University Press: Princeton, NJ, USA, 1997.
10. Barberis, N.; Shleifer, A.; Vishny, R. A Model of Investor Sentiment. *J. Financ. Econ.* **1998**, *49*, 307–343. [CrossRef]
11. Sprenger, T.; Welpe, I. News or Noise? The Stock Market Reaction to Different Types of Company-Specific News Events. *SSRN Electron. J.* **2011**. [CrossRef]
12. Guo, X.; Zhang, H.; Tian, T. Development of stock correlation networks using mutual information and financial big data. *PLoS ONE* **2018**, *13*, e0195941. [CrossRef]
13. Zhang, M.; Huang, T.; Guo, Z.; He, Z. Complex-network-based traffic network analysis and dynamics: A comprehensive review. *Phys. Stat. Mech. Its Appl.* **2022**, *607*, 128063. [CrossRef]
14. Su, Q.; Tu, L.; Wang, X.; Rong, H. Construction and robustness of directed-weighted financial stock networks via meso-scales. *Phys. A Stat. Mech. Its Appl.* **2022**, *605*, 127955. [CrossRef]
15. Song, G.; Zhao, T.; Wang, S.; Wang, H.; Li, X. Stock ranking prediction using a graph aggregation network based on stock price and stock relationship information. *Inf. Sci.* **2023**, *643*, 119236. [CrossRef]
16. Tumminello, M.; Aste, T.; Matteo, T.D.; Mantegna, R.N. A tool for filtering information in complex systems. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10421–10426. [CrossRef] [PubMed]
17. Djauhari, M.A.; Gan, S.L. Minimal spanning tree problem in stock networks analysis: An efficient algorithm. *Phys. A Stat. Mech. Its Appl.* **2013**, *392*, 2226–2234. [CrossRef]
18. Schober, P.; Boer, C.; Schwarte, L. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [CrossRef]
19. Sharma, C.; Habib, A. Mutual information based stock networks and portfolio selection for intraday traders using high frequency data: An Indian market case study. *PLoS ONE* **2019**, *14*, e0221910. [CrossRef]
20. Wang, X.; Li, S.; Hou, C.; Zhang, G. Minimum Spanning Tree Method for Sparse Graphs. *Math. Probl. Eng.* **2023**, *2023*, 8591115. [CrossRef]
21. Groenewold, N.; Fraser, P. Tests of asset-pricing models: How important is the iid-normal assumption? *J. Empir. Financ.* **2001**, *8*, 427–449. [CrossRef]

22. Onnela, J.P.; Chakraborti, A.; Kaski, K.; Kertész, J.; Kanto, A. Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E* **2003**, *68*, 056110. [CrossRef] [PubMed]
23. Boginski, V.; Butenko, S.; Pardalos, P. Network models of massive datasets. *Comput. Sci. Inf. Syst.* **2004**, *1*, 75–89. [CrossRef]
24. Boginski, V.; Butenko, S.; Pardalos, P.M. Mining market data: A network approach. *Comput. Oper. Res.* **2006**, *33*, 3171–3184. [CrossRef]
25. Tse, C.K.; Liu, J.; Lau, F.C. A network perspective of the stock market. *J. Empir. Financ.* **2010**, *17*, 659–667. [CrossRef]
26. Shirokikh, O.; Pastukhov, G.; Boginski, V.; Butenko, S. Computational study of the US stock market evolution: A rank correlation-based network model. *Comput. Manag. Sci.* **2013**, *10*, 81–103. [CrossRef]
27. Faizliev, A.; Balash, V.; Vlasov, A.; Tryapkina, T.; Mironov, S.; Androsov, I.; Petrov, V. Analysis of the Dynamics of Market Graph Characteristics. In Proceedings of the Third Workshop on Computer Modelling in Decision Making (CMDM 2018), Saratov, Russia, 14–17 November 2018; Atlantis Press: Amsterdam, The Netherlands, 2019; pp. 13–19. [CrossRef]
28. Semenov, D.; Koldanov, A.; Koldanov, P. Analysis of weakly correlated nodes in market network. *Res. Sq.* **2023**, *preprint*. [CrossRef]
29. Millington, T.; Niranjan, M. Construction of minimum spanning trees from financial returns using rank correlation. *Phys. A Stat. Mech. Its Appl.* **2021**, *566*, 125605. . [CrossRef]
30. Kalyagin, V.; Koldanov, A.P.; Koldanov, P.A.; Pardalos, P.M. *Statistical Analysis of Graph Structures in Random Variable Networks*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020. [CrossRef]
31. Bautin, G.; Kalyagin, V.; Koldanov, A.; Koldanov, P.; Pardalos, P. Simple measure of similarity for the market graph construction. *Comput. Manag. Sci.* **2013**, *10*, 105–124. [CrossRef]
32. Majapa, M.; Gossel, S.J. Topology of the South African stock market network across the 2008 financial crisis. *Phys. A Stat. Mech. Its Appl.* **2016**, *445*, 35–47. [CrossRef]
33. Kalyagin, V.A.; Koldanov, A.P.; Koldanov, P.A. Robust identification in random variable networks. *J. Stat. Plan. Inference* **2017**, *181*, 30–40. [CrossRef]
34. McAllester, D.; Stratos, K. Formal Limitations on the Measurement of Mutual Information. *arXiv* **2020**, arXiv:1811.04251.
35. Wang, G.J.; Xie, C.; He, K.; Stanley, H. Extreme risk spillover network: Application to financial institutions. *Quant. Financ.* **2017**, *17*, 1417–1433. [CrossRef]
36. Diebold, F.X.; Yılmaz, K. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *J. Econom.* **2014**, *182*, 119–134. [CrossRef]
37. Corsi, F.; Lillo, F.; Pirino, D.; Trapin, L. Measuring the propagation of financial distress with Granger-causality tail risk networks. *J. Financ. Stab.* **2018**, *38*, 18–36. [CrossRef]
38. Billio, M.; Getmansky, M.; Lo, A.W.; Pelizzon, L. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *J. Financ. Econ.* **2012**, *104*, 535–559. . [CrossRef]
39. Monti, F.; Stewart, D.; Surendra, A.; Alecu, I.; Nguyen-Tran, T.; Bennett, S.A.L.; Čuperlović Culf, M. Signed Distance Correlation (SiDCo): An online implementation of distance correlation and partial distance correlation for data-driven network analysis. *Bioinformatics* **2023**, *39*, btad210. [CrossRef]
40. Liu, K.; Liu, H.; Sun, D.; Zhang, L. Network Inference from Gene Expression Data with Distance Correlation and Network Topology Centrality. *Algorithms* **2021**, *14*, 61. [CrossRef]
41. Davis, R.; Matsui, M.; Mikosch, T.; Wan, P. Applications of Distance Correlation to Time Series. *Bernoulli* **2016**, *24*, 3087–3116. [CrossRef]
42. Hernández, J.E.S.; Vyas, M. Non-linear correlation analysis in financial markets using hierarchical clustering. *arXiv* **2023**, arXiv:2301.05080.
43. Rosen, K.H. *Discrete Mathematics and Its Applications*; The McGraw Hill Companies: New York, NY, USA, 2007.
44. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
45. Rakib, M.I.; Nobi, A.; Lee, J.W. Structure and dynamics of financial networks by feature ranking method. *Sci. Rep.* **2021**, *11*, 17618. [CrossRef] [PubMed]
46. Birch, J.; Pantelous, A.A.; Soramäki, K. Analysis of Correlation Based Networks Representing DAX 30 Stock Price Returns. *Comput. Econ.* **2016**, *47*, 501–525. [CrossRef]
47. Perez, C.; Germon, R. Chapter 7—Graph Creation and Analysis for Linking Actors: Application to Social Data. In *Automating Open Source Intelligence*; Layton, R., Watters, P.A., Eds.; Syngress: Boston, MA, USA, 2016; pp. 103–129. [CrossRef]
48. Drożdż, S.; Grümmer, F.; Górski, A.; Ruf, F.; Speth, J. Dynamics of competition between collectivity and noise in the stock market. *Phys. A Stat. Mech. Its Appl.* **2000**, *287*, 440–449. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.