Construction of Minimum Spanning Trees from Financial Returns using Rank Correlation

Tristan Millington*, Mahesan Niranjan

Department of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK

Abstract

The construction of minimum spanning trees (MSTs) from correlation matrices is an often used method to study relationships in the financial markets. However most of the work on this topic tends to use the Pearson correlation coefficient, which relies on the assumption of normality and can be brittle to the presence of outliers, neither of which is ideal for the study of financial returns. In this paper we study the inference of MSTs from daily US financial returns using Pearson and two rank correlation methods, Spearman and Kendall's τ. We find that the trees constructed using these rank methods tend to be more stable and maintain more edges over the dataset than those constructed using Pearson correlation, that there are significant differences in the agreement of the centrality of various sectors and that despite these, the trees tend to have similar topologies.

Keywords: networks, correlation, finance, minimum spanning trees

1. Introduction

Investors tend to not invest purely in single asset, but due to a desire to reduce risk and increase diversification, own portfolios made up of multiple assets. To accurately assess risk in these portfolios we must understand the dynamics of the relationships between said assets. Various methods of inferring the strength of relationships exist, for instance correlation [1] [2], partial correlation [3] [4] [5] or mutual information [6] [7], but Pearson correlation

*Corresponding Author
Email address: t.millington@soton.ac.uk (Tristan Millington)

is the most ubiquitous. From these asset to asset relationships we can use network theory to study the system as a whole.

Accurate inference of these correlation matrices from a dataset with p assets and n samples is challenging for numerous reasons. If n is not significantly larger than p, the correlation matrix will contain significant amounts of noise. Due to non-stationarities in the markets, we often want a small window of data, where we can assume that the data is stationary [8]. This often results in n being close to p, even if we have a large amount of data. Options to help remove noise in this situation include topological filtration methods (e.g. Minimum Spanning Trees [1], Planar Maximally Filtered Graphs [9]), random matrix theory approaches [10] [11] [12] or thresholding [13] [2]. In this paper we focus on the use of topological methods for this purpose, due to the simplicity of construction and interpretation. To start with, we follow Mantegna [1] and construct a distance matrix (D) from the correlation matrix (C):

$$D_{ij} = \sqrt{2(1 - C_{ij})} \tag{1}$$

This distance matrix is then used as the adjacency matrix for a new graph, the distance graph. From this we can use Kruskal's algorithm to create a minimum spanning tree (MST), which proceeds as follows

- Initialise the tree as a disconnected graph made up of the nodes in the distance graph
- Sort edges of the distance graph in ascending order and place in a list
- For each edge between i and j in this sorted list
  - If i and j are not in the same component in the tree, add this edge into the tree

Various 'stylised facts' are known about these trees, for instance

- Branches of the trees tend to contain companies in the same sector [1]
- The trees shrink and have different structures during times of market stress compared to market calm [14] [15]
- The trees tend to have a 'scale free' structure, with nodes of high degree (hubs) occurring more than would be expected from a random graph [16] [17] [14]

- Assets with large weights in Markowitz portfolios tend to be peripheral [18] [19] [20], however there is disagreement over whether to chose assets on the peripheries or center of the networks for better Sharpe ratios [21] [22] [20]
- The MSTs tend to only keep significant correlations [23]

While most of the focus has been on the US markets, MST based models have been applied to markets from other countries (e.g. Japan [24], the UK [25], Italy [26] and South Korea [27]), to cryptocurrencies [28] [29] and to networks from neuroscience [30] [31].

While the interpretability of the Pearson correlation coefficient is a big plus, it assumes normality, something which most assets return distributions do not follow [32], and is sensitive to outliers. There are of course correlation measures that do not suffer from these issues, namely those based on rank. Rank correlation methods calculate the correlation between the ranks of variables, which tends to remove the effects of outliers while still giving a measure of the degree to which two variables increase or decrease together. However to the best of our knowledge, most of the literature which studies the correlations between asset returns tends to use the Pearson correlation coefficient. Therefore in this paper we compare networks inferred from stock returns using Pearson, Spearman and Kendall's τ correlation in order to see if the robustness of these rank correlations can improve our understanding of the stock markets. Previous work [33] has briefly mentioned that MSTs constructed using Spearman correlation from volatility measures of stocks are more robust, but they did not explicitly compare the two correlation coefficients. In a paper more broadly looking at the effects of weighting observations, Pozzi et al. [34] compare Pearson correlation and Kendall's τ. They find that matrices constructed using Kendall's τ tend to contain more information than those constructed using Pearson correlation, and are affected less when they weight observations.

A paper on a more similar theme to this is written by Musmeci et al. [35], who take a multilayer network approach. Each layer is composed of a Planar Maximally Filtered Graph, constructed using a different method. Four methods are used to quantify relationships between assets, Pearson correlation, Kendall's τ, tail dependence and partial correlation. They find that these layers tend to have significant differences, with between 30% and 70% of the edges being unique to each layer. Pearson and Kendall's τ tend to be the methods that agree the most, with a correlation of around 0.7

on the degree of nodes. Interestingly they find that the level of agreement drops during times of crisis, showing that these different methods tend to pick up different signals from the markets and indicating that being mindful of multiple methods of quantifying relationships is valuable when taking a network approach to financial returns.

The final example we found is by Shirokikh et al. [36] who use a thresholding model with Spearman correlation, but they do not compare how this model differs to a Pearson based one.

The Pearson correlation between two variables r_i and r_j is defined as follows

$$C_{ij} = \frac{\sum_{i=1}^{n}(r_i(t) - \bar{r_i})(r_j(t) - \bar{r_j})}{\sqrt{\sum_{i=1}^{n}((r_i(t) - \bar{r_i})^2(r_j(t) - \bar{r_j})^2)}} \tag{2}$$

To calculate the Spearman correlation, we firstly sort the values, replace each value with its rank, and calculate the Pearson correlation between the ranks. This then measures the degree to which two variables monotonically increase or decrease together. Kendall's τ is slightly more complicated, measuring the relationship by considering the number of concordant pairs vs the number of discordant pairs. A pair of observations $(r_i(t), r_j(t)), (r_i(t+1), r_j(t+1))$ is concordant if $r_i(t) > r_j(t)$ and $r_i(t+1) > r_j(t+1)$ or if $r_i(t) < r_j(t)$ and $r_i(t+1) < r_j(t+1)$. It is discordant if $r_i(t) > r_j(t)$ and $r_i(t+1) < r_j(t+1)$ or if $r_i(t) < r_j(t)$ and $r_i(t+1) > r_j(t+1)$. The τ-a formula simply counts the number of concordant pairs vs the number of discordant pairs, divided by the total number of pairs. This does not however take into account any ties that might occur in the data, so we use the τ-b formulation, defined as

$$\tau = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{3}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, $n_0 = \frac{n(n-1)}{2}$, $n_1 = \sum t_a(t_a - 1)/2$, $n_2 = \sum u_a(u_a - 1)/2$, t_a is the number of values in the ath group of ties for variable i, u_a is the number of values in the ath group of ties for variable j. Any reference to τ in the rest of the paper refers to this τ-b formation.

2. Data and Software

The data we use is downloaded from Yahoo Finance. We use log returns from the S&P500 from 2000/03/01 to 2019/10/21. Any company missing

more than 10% of its data is removed, and any missing values are filled forwards from the first good value. If the values are missing from the start we backfill from the first good value. This results in 4790 days of return data for 229 companies. From this data we take a window of 504 days and slide along 30 days at a time, creating 142 windows to infer correlations from. Each company is tagged with a sector from the GICS classification using information from Bloomberg. This places each company into 1 of 11 sectors, Information Technology, Real Estate, Materials, Communication Services, Energy, Financials, Utilities, Industrials, Consumer Discretionary, Healthcare or Consumer Staples.

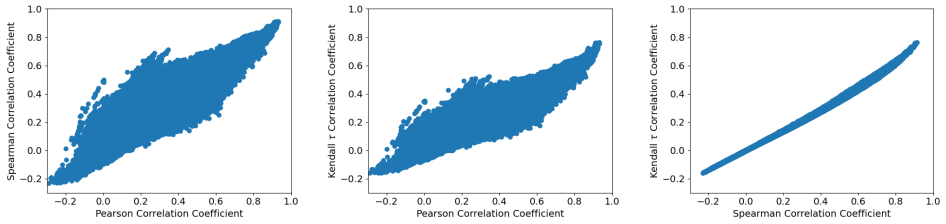
We make use of Python, NumPy and SciPy [37] for general scripting, pandas [38] for handing the data, statsmodels [39] for some of the statistical analysis, matplotlib [40] for plotting, arch [41] for the implementation of the circular bootstrap, Networkx [42] for the network analysis and gephi [43] for the MST visualization. The code and data is available at `https://github.com/shazzzm/rank_correlation_msts`.

3. Results and Analysis

3.1. Correlation Matrix Analysis

Firstly we analyze the full correlation matrices with no filtration. A starting point is to look at the correlation coefficient for the same set of values. Figure 1 shows a set of scatter plots comparing the correlations. From this we can see there is a degree of agreement between all, and generally larger correlations are more likely to be similar. However there is a 'fat' middle when comparing the rank correlations to the Pearson correlation, where there can be significant disagreement. Spearman and τ seem to be very similar, with there being a strong relationship between the two.

The largest eigenvalue of the correlation matrix is a measure of the intensity of the correlation present in the matrix, and in matrices inferred from financial returns tends to be significantly larger than the second largest [11] [12]. Generally this largest eigenvalue is larger during times of stress and smaller during times of calm [44] [11]. Firstly we study how this varies over time for each correlation measure. This is shown in Figure 2. For all of the networks there is a similar shape, with it peaking during times of market stress and dropping during times of calm. The Spearman and Pearson correlations have relatively similar values, although the Spearman has a smaller range. The τ correlation is much smaller than the other two at all times,



(a) Pearson vs Spearman (b) Pearson vs τ (c) Spearman vs τ

Figure 1: Relationship of the correlation coefficients from the entire dataset. There is obviously a degree of agreement, but the 'fat' middle is notable when comparing the Pearson and rank correlations, with it being possible to have a relatively large and significantly correlation by one coefficient, and a smaller and less significant correlation in the other measure. The rank correlations themselves are relatively similar.

and also has a smaller range. Times of stress and volatility tend to bring more outliers in returns data, which could be the cause of the difference in largest eigenvalue, but it is interesting that the eigenvalue for τ correlation is so much smaller at all times.

3.2. MST Stability

From these inferred correlation matrices we transform them into distance matrices using (1) and construct minimum spanning trees using Kruskal's algorithm. In this section we analyze and compare the stability of the trees constructed using the various coefficients. Example MSTs from the first window of data are shown in Figure 3. In all there is clear sector clustering, with branches of the trees consisting of nodes from the same sector. This has been noted before [1].

Firstly we focus on measuring the fraction of edge changes between MSTs adjacent in time, which quantifies how stable the trees are, and how well change in the market is detected. The results are plotted in Figure 4. Here we can see a large difference, with the MSTs inferred using the rank correlations showing more stability over time than those inferred using Pearson correlation. This is particularly noticeable during 2009, where the markets were very volatile due to the financial crisis. Around 50% of the edges change for the Pearson MSTs, but only 30% change for the Spearman MSTs, and 25% for the τ MSTs. It is also interesting to note that the edge difference rises at the start of the crisis and then drops during the actual crisis itself for the Pearson and Spearman MSTs, and just drops for the τ MSTs. Other

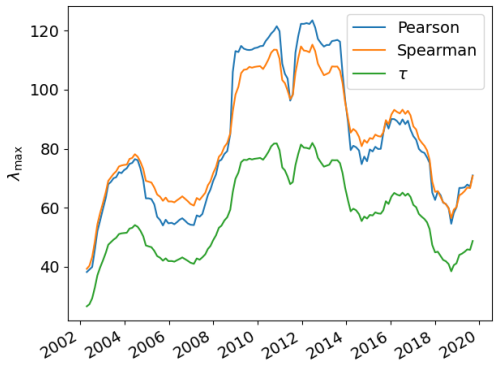


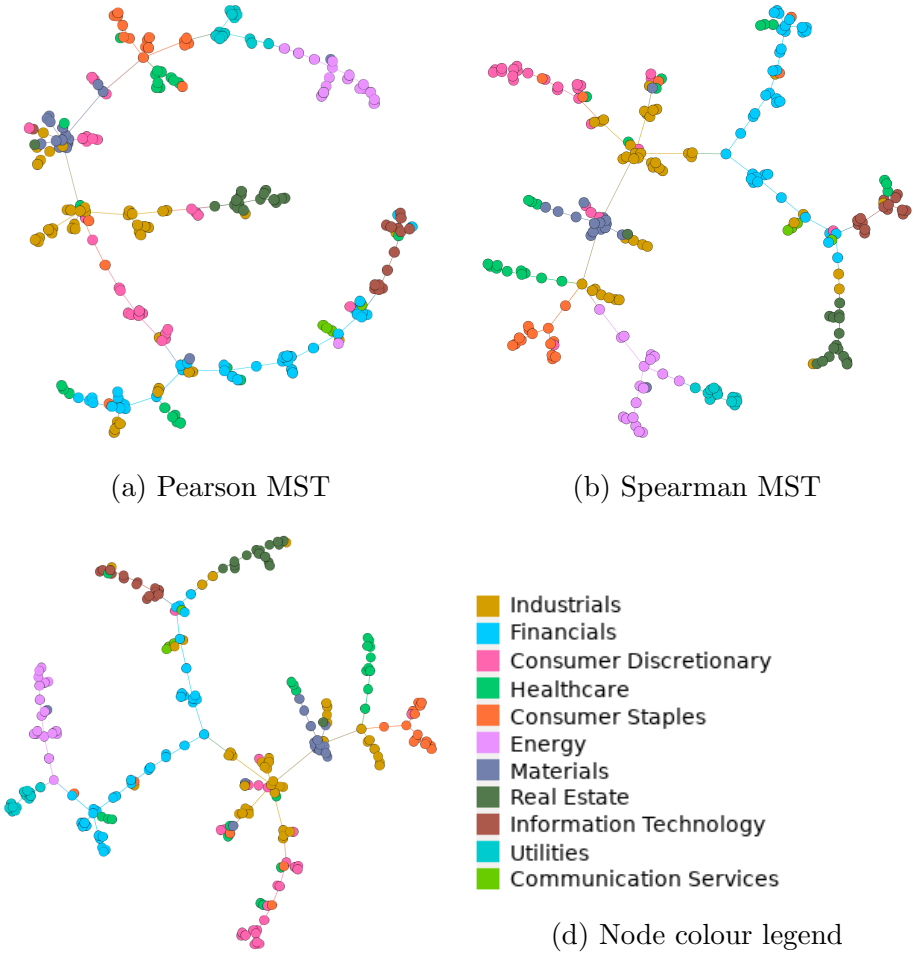
Figure 2: Largest eigenvalue ($\lambda_{\max}$) in the networks over time. From this we can see the Spearman correlation has a slightly smaller largest eigenvalue than the Pearson correlation, while the τ correlation is much smaller. The rank methods also have a smaller range than the Pearson correlation. The volatility of the markets at times of stress is likely to lead to more outliers, so the robustness of the rank correlations to these could be causing the reduced variance of the largest eigenvalue.

authors have noted that by some measures the markets could be considered more stable during these times [44] [45], but we have not found that this has been mentioned in the context of MSTs.

This therefore shows that the Spearman and τ MSTs tend to be more consistent than Pearson MSTs. This is particularly prominent at the start of the financial crisis, with the Pearson MSTs showing a large spike in difference, while the Spearman and τ MSTs show little or no change in difference. In this particular situation we would expect the heavy tails to affect the Pearson correlation between two assets more than the rank methods, and this should change the edges selected by the MST construction procedure.

Next we measure how the MSTs have changed from the first inferred tree using the fraction of edges that differ from the first tree to the current tree. This measures the life of an edge and shows us how the tree evolves. A plot of this is shown in Figure 5. From this we can see that quite rapidly the trees differ from the original, with 70% of the edges changing within 2 years. For our experiments, what is particularly interesting is that the rank MSTs maintain slightly more edges than the Pearson MSTs, but the difference between the τ and Spearman MSTs is very small.

Over the entire dataset, the Pearson MSTs maintain 4 edges, the Spearman MSTs maintain 7 edges and the τ MSTs maintain 8 edges. For the



(a) Pearson MST

(b) Spearman MST

(c) τ MST

(d) Node colour legend

Figure 3: Example MSTs constructed from the various correlation coefficients. Nodes are coloured according to their sector membership. From this colouring we can see that all coefficients show a strong degree of sector clustering, with branches of the trees tending to contain companies in the same sector, and the presence of several important hub nodes.

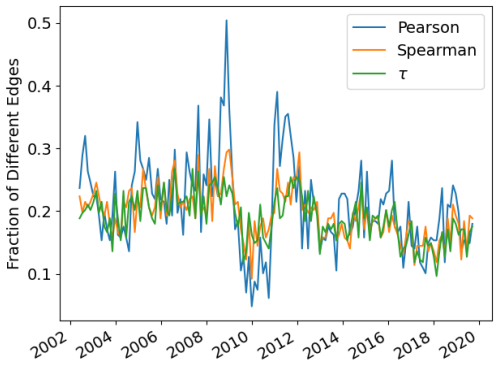


Figure 4: Edge difference between adjacent MSTs. From this we can see the MSTs inferred using rank correlation are far more stable with regards to time than those inferred using Pearson correlation. While all of the trees seem to become more similar during the financial crisis, the Pearson MSTs show a big reconfiguration as the crisis starts, while the τ MSTs shows no spike before dropping.

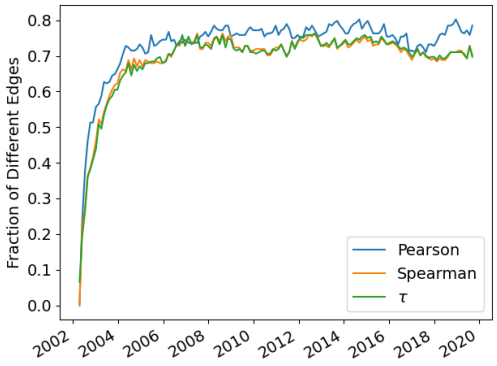


Figure 5: Fraction of edges that differ in the tree inferred at that moment in time from the first tree inferred. This gives us a measure of how long the edges persist for. Most of the edges disappear very rapidly, with around 70% changing within 2 years. The rank MSTs seem to be slightly more stable than the Pearson MSTs, maintaining more edges from the initial tree, but the difference between Spearman and Kendall's τ methods is very small.

Pearson MSTs these edges are

- T - VZ (Communication Services - Communication Services)
- CCL - RCL (Consumer Discretionary - Consumer Discretionary)
- DHI - LEN (Consumer Discretionary - Consumer Discretionary)
- HD - LOW (Consumer Discretionary - Consumer Discretionary)

In addition to these, the Spearman MSTs also maintain

- CAT - DE (Industrials - Industrials)
- FDX - UPS (Industrials - Industrials)
- GS - MS (Financials - Financials)

The τ MSTs maintain all the edges mentioned so far, plus

- HAL - SLB (Energy - Energy)

It is notable these are all intrasector edges, and that there is a large overlap between all the trees - all of the edges maintained by the Pearson MSTs are maintained by the Spearman MSTs, and all of the edges maintained for the Spearman MSTs are maintained for the τ MSTs.

There is of course the question of how the difference between the MSTs changes over time. We measure the fraction of edges that differ between the three MSTs and plot it in Figure 6. From this we can see there is a significant difference in the presence of edges between the rank MSTs and the Pearson MSTs. The difference does seem to increase during the financial crisis, with peaks occurring during 2008 and 2009. There seems to be relatively little difference between the two rank MSTs, with less than 10% of the edges being different for most of the dataset. This difference between the rank methods does not seem to be particularly affected by market conditions.

With this knowledge that the MSTs select quite different edges, we next ask which edges are selected differently. To start with we focus on the agreement between the degree of the nodes (the number of edges each node has). If we plot the Spearman correlation between the node degrees (Figure 7a) we can see that there is relatively high agreement between node degrees, aside from during 2011. This implies that nodes tend to be regarded as similarly

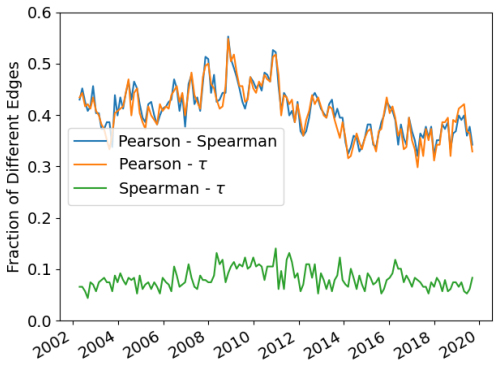


Figure 6: Fraction of the edges that differ between the various MSTs. There is a large degree of disagreement between the Pearson and rank-based correlations, even at the smallest difference around 30% of the edges differ. The difference peaks during the financial crisis, where over 50% of the edges differ at one point between the Pearson and rank MSTs. The difference between the rank MSTs themselves is small, with less than 10% of the edges differing.

important. However this does not show if this is in the core of the network or the peripheries. To measure this we set a threshold to remove high degree nodes (hubs), in this case arbitrarily defined as degree values over 4, and plot the correlation in Figure 7b. We can see that the correlation drops significantly between the Pearson MSTs and the rank MSTs, indicating that it is the hub nodes that are the greatest source of agreement. If this threshold is reversed, and the low degree nodes removed, the correlation is maintained at around 0.7.

3.3. Node and Sector Centrality

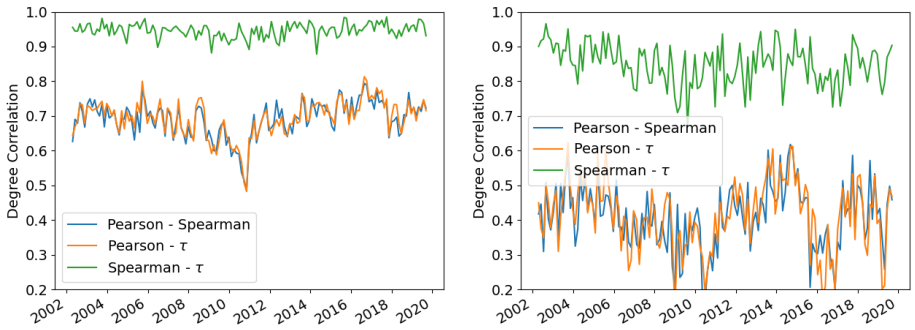
Now we look at the centrality of the nodes and economic sectors in the trees. If we sum the degree of each node for each MST we can get a measure of degree over the entire dataset, showing us which nodes are regarded as important overall. We refer to this as the total degree. The Spearman correlation between the total degrees is high (Pearson/Spearman $= 0.94$, Pearson/$\tau = 0.94$, Spearman/$\tau = 0.99$) indicating that the MSTs generally agree on the total centrality of most nodes. The 10 nodes that have the largest total degree shown in Table 1. We can see there is a significant overlap between the trees, with 5/10 shared between all, and 9/10 shared between the rank MSTs.

Pearson		
Company	Sector	Degree Sum
ITW	**Industrials**	**986**
PPG	**Materials**	**980**
PH	**Industrials**	**741**
HON	Industrials	716
ETN	**Industrials**	**712**
BLK	Financials	647
SPG	Real Estate	638
TROW	**Financials**	**629**
JPM	Financials	625
UTX	Industrials	611

Spearman		
Company	Sector	Degree Sum
PPG	**Materials**	**999**
PH	**Industrials**	**862**
MMM	*Industrials*	*721*
ITW	**Industrials**	**677**
ETN	**Industrials**	**674**
TROW	**Financials**	**658**
CAT	*Industrials*	*642*
BEN	*Financials*	*604*
AVB	*Real Estate*	*588*
JWN	Consumer Discretionary	562

τ		
Company	Sector	Degree Sum
PPG	**Materials**	**970**
PH	**Industrials**	**874**
MMM	*Industrials*	*726*
TROW	**Financials**	**700**
ITW	*Industrials*	*691*
ETN	**Industrials**	**661**
CAT	*Industrials*	*621*
BEN	*Financials*	*583*
LNC	Financials	579
AVB	*Real Estate*	*579*

Table 1: Nodes with the highest total degree in the Pearson (top), Spearman (middle) and τ (bottom) MSTs. Bold rows are shared between all, italic rows are shared between two. There is a strong overlap - 5 out of 10 are shared between all 3 indicating that generally they agree on which nodes are the most important, and 9 out of 10 are shared between the rank MSTs.



(a) All (b) Peripheries

Figure 7: Correlation between node degrees from the MSTs. Left shows the correlation between the degree of all nodes, while right shows only the correlation between the peripheries (defined as nodes with a degree of less than 5). Overall (left) we can see there is relatively high agreement on the node degrees, although as above the trees diverge during the financial crisis, while the reduction in correlation (with an average of 0.4) when removing the hubs compared to the overall correlation (with an an average of 0.7) shows most of the disagreement comes at the peripheries

On a sector note we can see the Industrials sector seems particularly central in these MSTs, with 5 of the top 10 in all being made up of companies from this sector. The next closest is the Financials sector, making up 3 of the top 10 in the Pearson and τ MSTs and 2 in the Spearman MSTs. The Materials and Real Estate have 1 entry in the top 10 for all 3 MSTs, and the Consumer Discretionary appears once in the Spearman table.

From this we can look at how the centrality of a sector varies over time. In particular we look at how the mean centrality varies by calculating the centrality of all the nodes in a sector and taking the mean. This reduces the effect of the different numbers of companies in each sector. To measure this we use both degree centrality and betweenness centrality. Betweenness centrality is calculated by looking at the fraction of shortest paths that pass through a node, and allows us to get a different perspective on which edges are regarded as important. The results are shown in Figures 8 (degree centrality) and 9 (betweenness centrality).

Focusing on the mean sector degree centrality (Figure 8) we can see that the Financials sector is important in all 3 MSTs. For the Pearson MSTs it becomes important in 2004, and particularly important in 2008. For the rank methods it becomes very important in both 2004 and 2008, but the

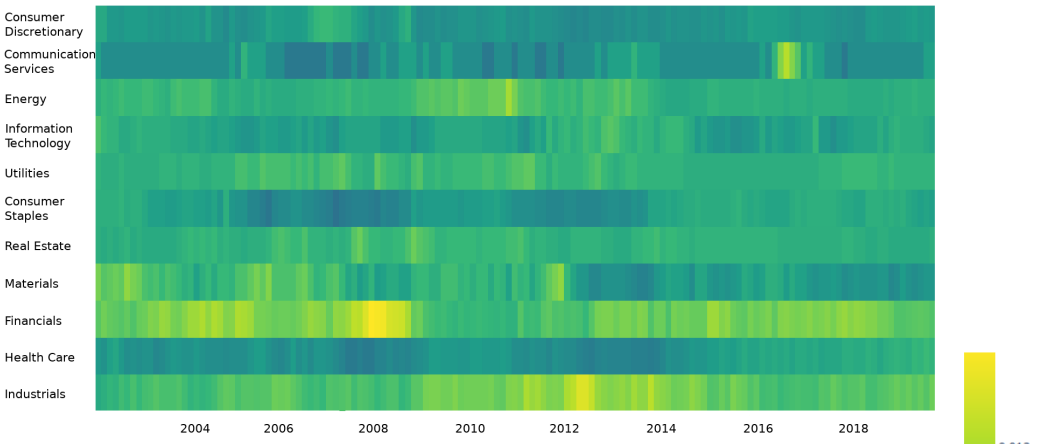
peak in 2008 seems smaller. The Industrials sector is usually important for most of the dataset, but it does seem more variable for the Pearson MSTs, with a strong peak in 2012. This is less visible in the rank trees, particularly the Spearman MSTs. The Communication Services sector is generally not central in any of the trees, with brief exceptions for 2017 in the Pearson MSTs and 2011 in the Spearman and τ MSTs. Another point of disagreement is the Information Technology sector, which does not become significantly central for the Pearson MSTs, but does for the rank trees, particularly during 2013.

Next we look at the mean sector betweenness centrality (Figure 9). Compared to the mean degree centrality there is a far larger spread of values - looking at the legend the values go from 0 - 0.18 for betweenness centrality vs 0 - 0.0018 for the degree centrality. The peaks are also much more noticeable. A particularly noteworthy peak is that of the Energy sector in the Pearson MSTs during 2010, which is not shared in either of the rank methods. All MSTs generally seem to regard the Financials sector as central, but there is a strong peak for the τ MSTs during 2002 that is less present for either the Pearson or the Spearman MSTs. There are peaks for Real Estate and Consumer Staples in the Pearson MSTs during 2018 that are not present in the rank MSTs. All three pick up a period between 2009 and 2011 where the centrality of the Materials sector increases.

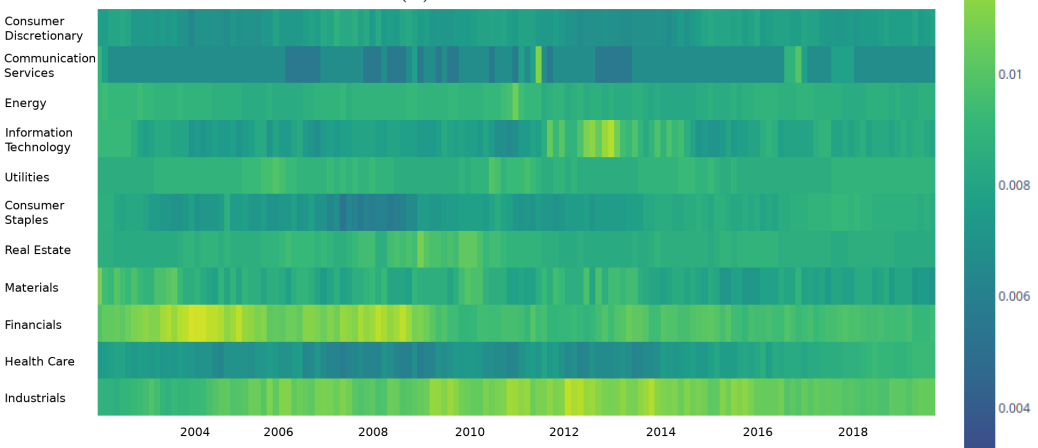
In general it seems there more agreement between the Spearman and τ MSTs than between Pearson and either one of them. There is also less agreement on mean sector centrality when using betweenness centrality vs degree centrality. To quantify this further we calculate the Spearman correlation between the mean centralities over time, and show the results in Tables 2 (degree centrality) and 3 (betweenness centrality).

For the degree centrality (Table 2) the Spearman and τ methods strongly agree on the sector centralities, with a mean Spearman correlation of 0.925 ± 0.060. This is much weaker for the Pearson and τ MSTs (0.594 ± 0.185) and the Pearson and Spearman MSTs (0.591 ± 0.196). The agreement for the Communication Services sector is notably low between the Pearson and rank MSTs, but even between the rank MSTs it is the lowest of all the sectors. However it is one of the smallest sectors, so the poor or excellent performance of a single company is less likely to be hidden by other companies in the sector. There is strong agreement in the Health Care, Consumer Staples, Real Estate and Industrials sectors between all of the MSTs.

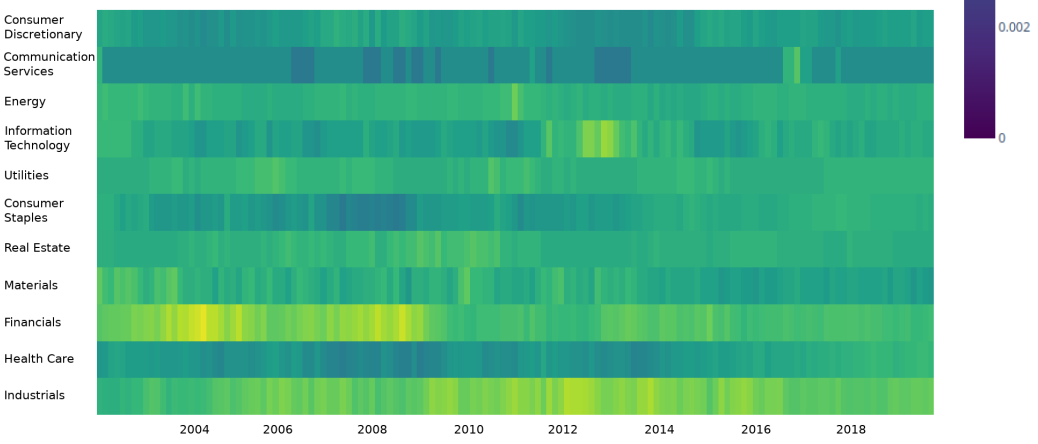
For the betweenness centrality (Table 3) again there is a generally strong agreement on the sector centralities between the Spearman and τ MSTs,



(a) Pearson

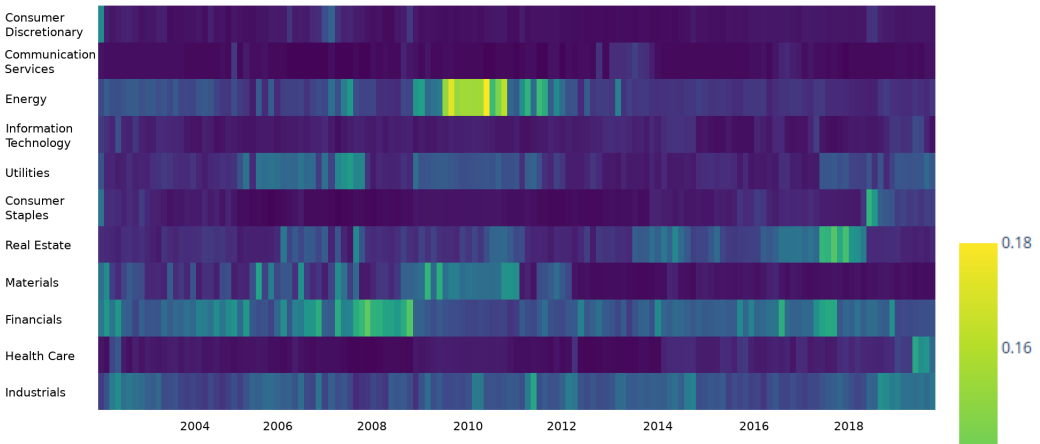


(b) Spearman

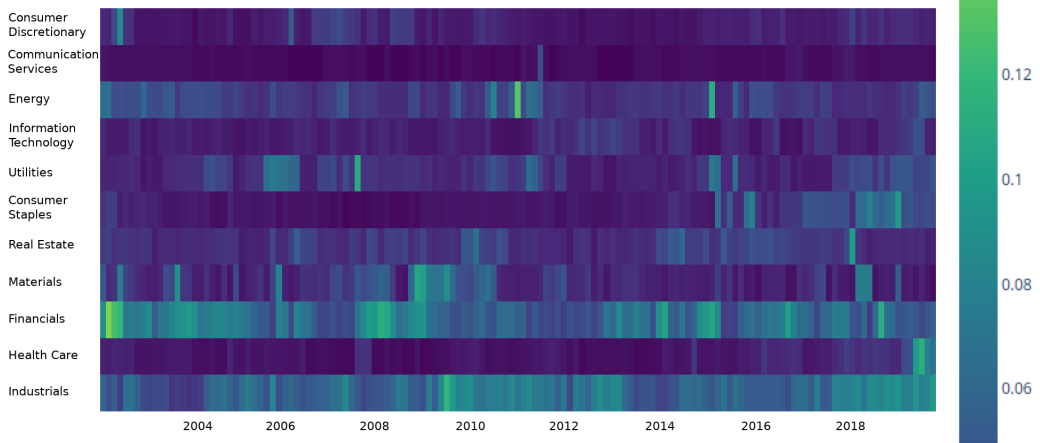


(c) τ

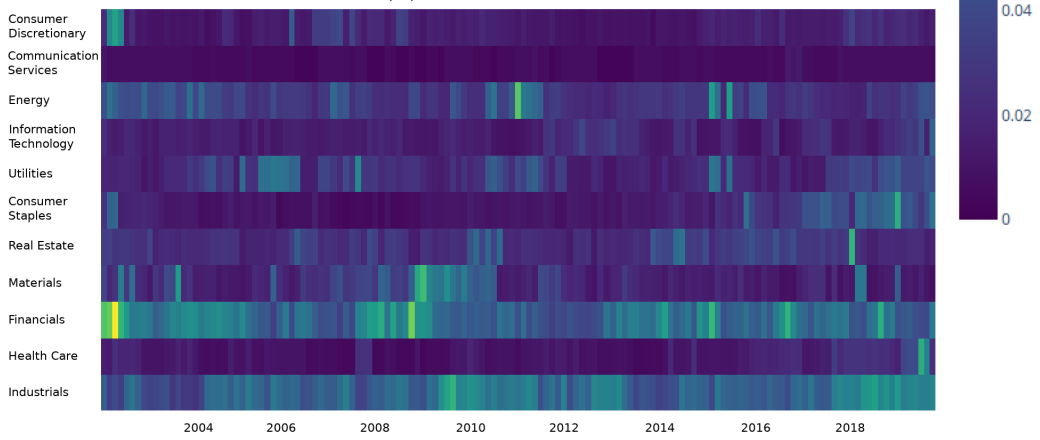
Figure 8: Mean degree centrality for the sectors over time.



(a) Pearson



(b) Spearman



(c) τ

Figure 9: Mean degree centrality for the sectors over time.

Sector	Pearson - Spearman	Pearson - τ	Spearman - τ
Industrials	0.779	0.801	0.956
Health Care	0.768	0.754	0.973
Financials	0.664	0.655	0.962
Materials	0.391	0.538	0.893
Real Estate	0.715	0.664	0.942
Consumer Staples	0.824	0.810	0.977
Utilities	0.430	0.387	0.919
Information Technology	0.657	0.654	0.946
Energy	0.543	0.551	0.911
Communication Services	0.181	0.190	0.763
Consumer Discretionary	0.545	0.526	0.936

Table 2: Correlation between the mean degree centrality for each sector between the MSTs. There seems to be reasonably broad agreement between all the MSTs on the centrality of most sectors, with the exceptions of Communication Services and Materials. The rank MSTs show a large amount of agreement, while the Pearson MSTs tend to differ more.

with the mean Spearman correlation between the sector centralities being 0.865 ± 0.058. The Pearson - τ mean correlation is 0.452 ± 0.183 and the Pearson - Spearman mean correlation is 0.442 ± 0.183, showing a much weaker relationship. Again Communication Services has the lowest agreement between all of the methods, and there is low agreement between the centrality of the Industrials, Energy and Financials sectors for the Pearson and rank MSTs. These also have a large drop in agreement in betweenness centrality compared to degree centrality. It is particularly notable as these sectors are usually regarded as important and central to the US economy, and also as there was relatively strong agreement in the mean degree centralities between the MSTs for the Industrials sector. In general there is a drop in mean correlation when using betweenness centrality as a measure when compared to degree centrality. This implies that companies tend to take different positions in the Pearson MSTs compared to the rank MSTs. There is however a much greater range in the betweenness centralities than the degree centralities, which may affect the correlation.

3.4. MST Topology

Having studied the stability of the trees over time and the importance of various sectors, we now look if the structure of the MSTs differ using

Sector	Pearson - Spearman	Pearson - τ	Spearman - τ
Industrials	0.254	0.259	0.895
Health Care	0.596	0.594	0.949
Financials	0.396	0.425	0.886
Materials	0.405	0.487	0.815
Real Estate	0.541	0.546	0.862
Consumer Staples	0.785	0.764	0.946
Utilities	0.540	0.563	0.893
Information Technology	0.392	0.395	0.778
Energy	0.316	0.379	0.803
Communication Services	0.109	0.067	0.802
Consumer Discretionary	0.527	0.492	0.889

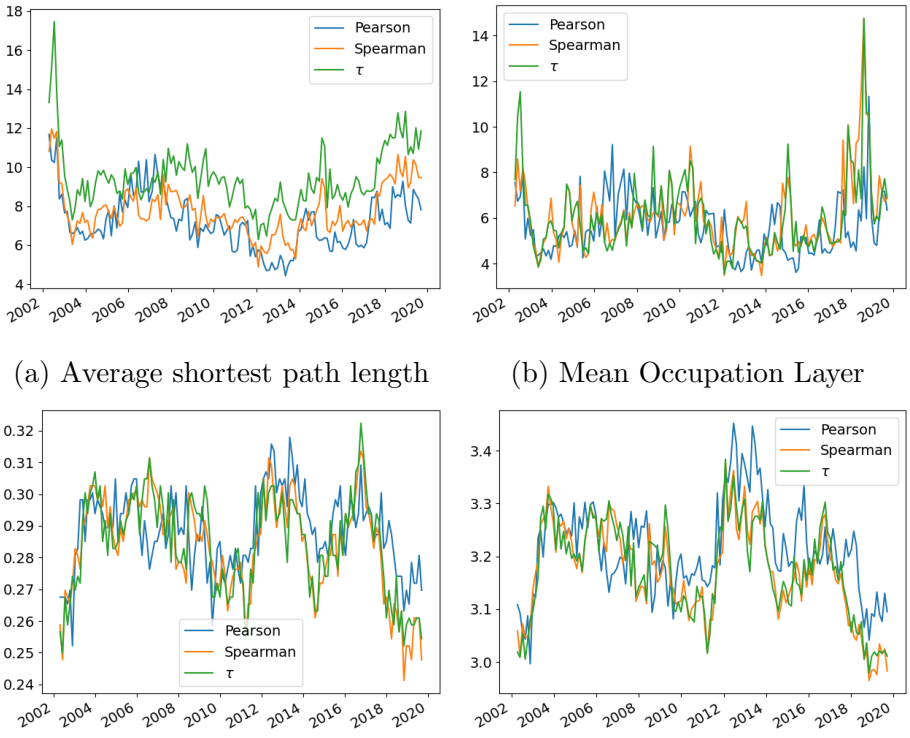
Table 3: Correlation between the mean betweenness centrality for each sector between the MSTs. There is a significant drop in the mean agreement of the betweenness sector centralities compared to the degree sector centrality, particularly for the Industrials sector. There is still generally high agreement between the rank MSTs, but the Pearson MSTs seem to have a significant difference.

some network measures. We use the leaf fraction (fraction of nodes with only 1 edge), exponent when fitting a power law to the degree distribution, the average shortest path length and mean occupation layer (the mean of all shortest paths from each node to the center of the tree). In this case we take the center of the tree as the node with the largest degree. The plots of these measures over time are shown in Figure 10.

From these we can see that irrelevant of the coefficient, the MSTs have similar structure. All of the trees have a heavy tailed degree distribution, with there being a high number of nodes with only one other edge and a small number of edges with a large degree. We can see the structures change during the financial crisis, with the leaf fraction and average shortest path length decreasing, although there is no noticeable change for the mean occupation layer.

3.5. MST Robustness

Finally we are interested in comparing the robustness of these correlation coefficients. This can be done using a bootstrap based approach, in a similar manner to Tumminello et al. [46] and Musciotto et al. [47]. Here we create 1000 pseudo-datasets using a circular bootstrap. A circular bootstrap is a



(a) Average shortest path length

(b) Mean Occupation Layer

(c) Leaf Fraction

(d) Exponent

Figure 10: Topological measures of the structure of the MSTs. From this we can see all the correlation coefficients produce MSTs with very similar structures. Macroeconomic effects are visible in the trees, with the average shortest path length and leaf fraction decreasing during the financial crisis, although the mean occupation layer seems less sensitive.

Method	Mean Difference	Correlation
Pearson	0.417 ± 0.136	0.276
Spearman	0.388 ± 0.124	0.268
τ	0.386 ± 0.121	0.268

Table 4: Mean difference between MSTs constructed from the bootstrap datasets and correlation between p-value and edge weight. The rank MSTs tend to select the same edges slightly more from the bootstrap datasets than the Pearson MSTs, but the difference is not large. None of the correlation coefficients have a particularly strong relationship between p-value and edge weight.

type of block bootstrap where we select data from a continuous stretch of time, and if the end of the dataset is reached we wrap round and start back at the beginning. This tends to be more appropriate for time series data compared to the classic bootstrap due to look ahead effects and potential autocorrelation. With these pseudo-datasets we calculate the correlations between assets and construct MSTs from these correlation matrices. Once we have this set of MSTs, we can compare the edges present in them. Ideally if there is no noise, the data is purely stationary and the methods robust all these MSTs would be the same. Of course this is not the case in real life. To run the bootstrap we take the first 1008 days of data and create 1000 bootstrapped datasets of 504 days. Firstly we measure the mean and standard deviation of the fraction of difference in edge presence across the trees. The results are shown in Table 4. From this we can see that the rank MSTs select slightly more of the same edges, but the difference is not large.

We can gain a measure of how likely an edge is to exist in these MSTs by counting the number of times an edge exists in our forest of trees and diving by the total number of bootstrap replications. This gives us a p-value for the edge. Previous work [46] has shown that there is little relationship between p-value and the edge weight in Pearson correlation based MSTs, so we explore this for the other MSTs. The results are shown in Table 4. Again there seems to be little correlation between a p-value and the edge weight for any coefficient, and all have a similar correlation between p-value and edge weight.

Companies in the same sector tend to be more correlated than those in different sectors and so we might expect those edges to have a larger p-value than the inter sector ones. To measure this we take the mean of the p-values of the edges in the tree inferred from the overall dataset for all edges,

Method	Total	Intrasector edges	Intersector edges
Pearson	0.682 ± 0.229	0.706 ± 0.222	0.572 ± 0.230
Spearman	0.705 ± 0.232	0.737 ± 0.216	0.574 ± 0.253
τ	0.705 ± 0.238	0.741 ± 0.215	0.556 ± 0.269

Table 5: Mean and standard deviation of p-values found for the edges from the MSTs inferred from the first window of data. For all correlation methods intrasector edges are more likely than intersector ones, but interestingly the rank-based correlation MSTs have a slightly higher p-value for an intrasector edge than the Pearson MSTs.

intrasector edges and intersector edges. The results are shown in Table 5. From this we can see the rank MSTs have a larger total p-value compared to Pearson MSTs, although the difference between is not very large. The rank MSTs also have a larger mean p-value for intrasector edges compared to the Pearson MSTs, although all of the differences are within the standard deviations.

4. Conclusion

In this paper we have used the Pearson, Spearman and Kendall's τ correlation coefficients to infer correlation matrices from stock returns, constructed minimum spanning trees from these matrices and compared the robustness and evolution of the trees over time. In general we have found the MSTs constructed using the rank correlations (Spearman and Kendall's τ) to be more robust than those constructed with the Pearson correlation. They tend to change less (notably during times of market stress), have edges that are maintained for a longer time period, and tend to have slightly more of the same edges selected when the reliability of the trees is tested using a circular bootstrap. The different methods do however show strong agreement on the hubs in the trees, but tend to disagree on the structure of the peripheries.

The rank MSTs in general show broad agreement on the mean centrality of each sector, but the Pearson MSTs show a significant difference in which sectors they regard as important. The agreement using degree centrality is higher than using betweenness centrality, indicating that companies tend to be found in different places on the trees in the Pearson MSTs compared to the rank ones.

Despite this, the trees tend to have a similar topology, irrelevant of coefficient and this topology tends to vary in a similar way over time for all three methods.

Finally we use a bootstrap to test the consistency of the edges selected in the MSTs. We find the rank MSTs select slightly more of the same edges than the Spearman MSTs, but the difference is not large, and that there is only a weak relationship between the strength of an edge and its likelihood of being selected. In all of the trees intrasector edges are more likely to be selected than intersector edges, and rank based trees tend to select slightly more intrasector edges, but again this difference is small.

This shows that the heavy tails of financial returns have a big effect on the construction of Pearson correlation based MSTs, and this is something to be aware of when trying to draw conclusions from these trees. It therefore may be worth also constructing MSTs using different correlation coefficients for any problem to see if the conclusions drawn are valid for both sets. Generally the Spearman and Kendall's τ correlation coefficients tend to give similar results, indicating that if computational resources are constrained then calculating the Spearman correlation is sufficient. Future work could proceed in several directions. A comparison of mutual information MSTs to these correlation MSTs to see how they differ could be interesting, or exploring different filtration models, for instance the Planar Maximally Filtered Graph. Alternatively these comparisons could be performed with returns data from other countries or assets, perhaps from data that is highly correlated and volatile, for instance for returns from cryptocurrencies or developing nations.

Funding

TM Acknowledges PhD studentship funding from the School of Electronics and Computer Science, University of Southampton;

References

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

[2] J.-P. Onnela, K. Kaski, J. Kertész, Clustering and information in correlation based financial networks, The European Physical Journal B 38 (2) (2004) 353–362.

[3] T. Millington, M. Niranjan, Quantifying influence in financial markets via partial correlation network inference, in: 2019 11th International

Symposium on Image and Signal Processing and Analysis (ISPA), 2019, pp. 306–311. doi:10.1109/ISPA.2019.8868437.

[4] D. Y. Kenett, M. Tumminello, A. Madi, G. Gur-Gershgoren, R. N. Mantegna, E. Ben-Jacob, Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market, PLOS ONE 5 (12) (2010) 1–14. doi:10.1371/journal.pone.0015032.

[5] T. Millington, M. Niranjan, Partial correlation financial networks, Applied Network Science 5 (1) (2020) 11. doi:10.1007/s41109-020-0251-z.
URL https://doi.org/10.1007/s41109-020-0251-z

[6] P. Fiedor, Information-theoretic approach to lead-lag effect on financial markets, The European Physical Journal B 87 (8) (2014) 168.

[7] X. Guo, H. Zhang, T. Tian, Development of stock correlation networks using mutual information and financial big data, PLOS ONE 13 (4) (2018) 1–16.

[8] R. S. Tsay, Analysis of financial time series, Vol. 543, John Wiley & Sons, 2005.

[9] M. Tumminello, F. Lillo, R. N. Mantegna, Correlation, hierarchies, and networks in financial markets, Journal of Economic Behavior & Organization 75 (1) (2010) 40 – 58, transdisciplinary Perspectives on Economic Complexity. doi:https://doi.org/10.1016/j.jebo.2010.01.004.
URL http://www.sciencedirect.com/science/article/pii/S0167268110000077

[10] A. Namaki, A. Shirazi, R. Raei, G. Jafari, Network analysis of a financial market based on genuine correlation and threshold method, Physica A: Statistical Mechanics and its Applications 390 (21) (2011) 3835 – 3841. doi:https://doi.org/10.1016/j.physa.2011.06.033.
URL http://www.sciencedirect.com/science/article/pii/S0378437111004808

[11] L. Laloux, P. Cizeau, M. Potters, J.-P. Bouchaud, Random matrix theory and financial correlations, International Journal of Theoretical and Applied Finance 3 (03) (2000) 391–397.

[12] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, H. E. Stanley, Random matrix approach to cross correlations in financial data, Physical Review E 65 (6) (2002) 066126.

[13] V. Boginski, S. Butenko, P. M. Pardalos, Statistical analysis of financial networks, Computational Statistics & Data Analysis 48 (2) (2005) 431 – 443. doi:10.1016/j.csda.2004.02.004.

[14] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, A. Kanto, Dynamics of market correlations: Taxonomy and portfolio analysis, Physical Review E 68 (5) (2003) 056110.

[15] Y. Zhang, G. H. T. Lee, J. C. Wong, J. L. Kok, M. Prusty, S. A. Cheong, Will the us economy recover in 2010? a minimal spanning tree study, Physica A: Statistical Mechanics and its Applications 390 (11) (2011) 2020 – 2050.

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

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

[18] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, Dynamic asset trees and black monday, Physica A: Statistical Mechanics and its Applications 324 (1) (2003) 247 – 252, proceedings of the International Econophysics Conference.

[19] A. Hüttner, J.-F. Mai, S. Mineo, Portfolio selection based on graphs: Does it align with markowitz-optimal portfolios?, Dependence Modeling 6 (1) (2018) 63–87.

[20] G. Peralta, A. Zareei, A network approach to portfolio selection, Journal of Empirical Finance 38 (2016) 157 – 180.

[21] H. Kaya, Eccentricity in asset management, Available at SSRN 2350429 (2013).

[22] F. Pozzi, T. Di Matteo, T. Aste, Spread of risk across financial markets: better to invest in the peripheries, Scientific reports 3 (2013) 1665.

[23] T. Aste, W. Shaw, T. D. Matteo, Correlation structure and dynamics in volatile markets, New Journal of Physics 12 (8) (2010) 085009. doi:10.1088/1367-2630/12/8/085009.

[24] W.-S. Jung, O. Kwon, F. Wang, T. Kaizoji, H.-T. Moon, H. E. Stanley, Group dynamics of the japanese market, Physica A: Statistical Mechanics and its Applications 387 (2) (2008) 537 – 542. doi:https://doi.org/10.1016/j.physa.2007.09.022.
URL http://www.sciencedirect.com/science/article/pii/S0378437107010114

[25] R. Coelho, S. Hutzler, P. Repetowicz, P. Richmond, Sector analysis for a ftse portfolio of stocks, Physica A: Statistical Mechanics and its Applications 373 (2007) 615 – 626. doi:https://doi.org/10.1016/j.physa.2006.02.050.
URL http://www.sciencedirect.com/science/article/pii/S0378437106006364

[26] P. Coletti, Comparing minimum spanning trees of the italian stock market using returns and volumes, Physica A: Statistical Mechanics and its Applications 463 (2016) 246 – 261. doi:https://doi.org/10.1016/j.physa.2016.07.029.
URL http://www.sciencedirect.com/science/article/pii/S0378437116304605

[27] W.-S. Jung, S. Chae, J.-S. Yang, H.-T. Moon, Characteristics of the korean stock market correlations, Physica A: Statistical Mechanics and its Applications 361 (1) (2006) 263 – 271. doi:https://doi.org/10.1016/j.physa.2005.06.081.
URL http://www.sciencedirect.com/science/article/pii/S0378437105007181

[28] D. Stosic, D. Stosic, T. B. Ludermir, T. Stosic, Collective behavior of cryptocurrency price changes, Physica A: Statistical Mechanics and its Applications 507 (2018) 499–509.

[29] J. Y. Song, W. Chang, J. W. Song, Cluster analysis on the structure of the cryptocurrency market via bitcoin–ethereum filtering, Physica A: Statistical Mechanics and its Applications 527 (2019) 121339. doi:https://doi.org/10.1016/j.physa.2019.121339.
URL http://www.sciencedirect.com/science/article/pii/S0378437119304893

[30] P. Tewarie, A. Hillebrand, M. Schoonheim, B. van Dijk, J. Geurts, F. Barkhof, C. Polman, C. Stam, Functional brain network analysis using minimum spanning trees in multiple sclerosis: An meg source-space study, NeuroImage 88 (2014) 308 – 318. doi:https://doi.org/10.1016/j.neuroimage.2013.10.022.
URL http://www.sciencedirect.com/science/article/pii/S1053811913010458

[31] P. Tewarie, E. van Dellen, A. Hillebrand, C. Stam, The minimum spanning tree: An unbiased method for brain network analysis, NeuroImage 104 (2015) 177 – 188. doi:https://doi.org/10.1016/j.neuroimage.2014.10.015.
URL http://www.sciencedirect.com/science/article/pii/S1053811914008398

[32] R. Cont, Empirical properties of asset returns: stylized facts and statistical issues, Quantitative Finance 1 (2) (2001) 223–236. arXiv:https://doi.org/10.1080/713665670, doi:10.1080/713665670.
URL https://doi.org/10.1080/713665670

[33] S. Miccichè, G. Bonanno, F. Lillo, R. N. Mantegna, Degree stability of a minimum spanning tree of price return and volatility, Physica A: Statistical Mechanics and its Applications 324 (1) (2003) 66 – 73, proceedings of the International Econophysics Conference. doi:https://doi.org/10.1016/S0378-4371(03)00002-5.
URL http://www.sciencedirect.com/science/article/pii/S0378437103000025

[34] F. Pozzi, T. Di Matteo, T. Aste, Exponential smoothing weighted correlations, The European Physical Journal B 85 (6) (2012) 175. doi:10.1140/epjb/e2012-20697-x.
URL https://doi.org/10.1140/epjb/e2012-20697-x

[35] N. Musmeci, V. Nicosia, T. Aste, T. Di Matteo, V. Latora, The multiplex dependency structure of financial markets, Complexity 2017 (2017).

[36] O. Shirokikh, G. Pastukhov, V. Boginski, S. Butenko, Computational study of the US stock market evolution: a rank correlation-based network model, Computational Management Science 10 (2) (2013) 81–103. doi:10.1007/s10287-012-0160-4.
URL https://ideas.repec.org/a/spr/comgts/v10y2013i2p81-103.html

[37] T. E. Oliphant, A guide to NumPy, Vol. 1, 2006.

[38] W. McKinney, Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference, 2010, pp. 51 – 56.

[39] S. Seabold, J. Perktold, Statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference, 2010.

[40] J. D. Hunter, Matplotlib: A 2d graphics environment, Computing in Science Engineering 9 (3) (2007) 90–95. doi:10.1109/MCSE.2007.55.

[41] K. Sheppard, S. Khrapov, G. Lipták, R. Capellini, esvhd, Hugle, JPN, X. RENE-CORAIL, M. E. Rose, jbrockmendel, bashtage/arch: Release 4.8.1 (Mar. 2019). doi:10.5281/zenodo.2613877.
URL https://doi.org/10.5281/zenodo.2613877

[42] A. A. Hagberg, D. A. Schult, P. J. Swart, Exploring network structure, dynamics, and function using networkx, in: Proceedings of the 7th Python in Science Conference, Pasadena, CA USA, 2008, pp. 11 – 15. doi:10.25080/issn.2575-9752.

[43] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: Third international AAAI conference on weblogs and social media, 2009.

[44] S. Drożdż, F. Grümmer, A. Górski, F. Ruf, J. Speth, Dynamics of competition between collectivity and noise in the stock market, Physica A: Statistical Mechanics and its Applications 287 (3) (2000) 440 – 449. doi:https://doi.org/10.1016/S0378-4371(00)00383-6.
URL http://www.sciencedirect.com/science/article/pii/S0378437100003836

[45] A. Kocheturov, M. Batsyn, P. M. Pardalos, Dynamics of cluster structures in a financial market network, Physica A: Statistical Mechanics and its Applications 413 (2014) 523 – 533. doi:https://doi.org/10.1016/j.physa.2014.06.077.
URL http://www.sciencedirect.com/science/article/pii/S0378437114005585

[46] M. Tumminello, C. Coronnello, F. Lillo, S. Micciche, R. N. Mantegna, Spanning trees and bootstrap reliability estimation in correlation-based networks, International Journal of Bifurcation and Chaos 17 (07) (2007) 2319–2329.

[47] F. Musciotto, L. Marotta, S. Miccichè, R. Mantegna, Bootstrap validation of links of a minimum spanning tree, Physica A: Statistical Mechanics and its Applications 512 (2018) 1032 – 1043. doi:https://doi.org/10.1016/j.physa.2018.08.020.
URL http://www.sciencedirect.com/science/article/pii/S0378437118309695