Application of network filtering techniques in finding hidden structures on the Johannesburg Stock Exchange

by

Yashin Gopi

Submitted in partial fulfilment of the requirements for the degree

MSc in Financial Engineering

in the

Department of Mathematics and Applied Mathematics
Faculty of Natural and Agricultural Sciences

UNIVERSITY OF PRETORIA

January 2023

SUMMARY

Application of network filtering techniques in finding hidden structures on the Johannesburg Stock Exchange

by

Yashin Gopi

Supervisor: Professor Eben Maré

Department: Department of Mathematics and Applied Mathematics

University: University of Pretoria

Degree: MSc in Financial Engineering

Keywords: Econophysics, Correlation-based Network, Network Filter, Minimal Spanning Tree (MST), Planar Maximally Filtered Graph (PMFG), Hierarchical Cluster Analysis, Directed Bubble Hierarchical Tree (DBHT), Network Topology Measures

Researchers from the field of econophysics have favoured the idea that financial markets are a complex adaptive system, consisting of entities that behave and interact in a diverse manner, leading to non-linear, emergent behaviour of the system. In the last twenty years, there has been an increasing focus on modelling complex adaptive systems using network theory. Correlation-based networks, where stocks are represented as entities in the network, and the relationships amongst the stocks are based on the strength of the co-movements of the stocks, have been widely studied. Network filtering tools, such as the Minimal Spanning Tree (MST), and the Planar Maximally Filtered Graph (PMFG), have been useful to attenuate the impact of noise in these networks, thereby allowing important macroscopic and mesoscopic structures to emerge. One of the main benefits of the PMFG is that it is accompanied by a hierarchical clustering algorithm called the Directed Bubble Hierarchical Tree (DBHT). This method has the benefit of being fully unsupervised in that it does not require the user to decide a priori on the number of clusters that the data should be split into.

These techniques have been applied here to analyse the complex interactions amongst stocks on the Johannesburg Stock Exchange. A structure emerged in which shares from similar ICB sectors tended to cluster together. However, the so-called Rand Hedge shares, and shares which exhibited low liquidity, tended to override the sector effect and clustered together. From a dynamic perspective, the MST and PMFG seemed to shrink during market crashes, while the Basic Materials sector was typically the most important or central sector over time. Over the long-term, the DBHT divided the stocks in the South African stock market into six clusters. This technique was compared to other popular hierarchical clustering algorithms, and the amount of economic information that each method extracted was quantified. The most recent PMFG and DBHT showed a changed structure as compared to the long-term data, highlighting that the way that market participants view South African shares can change over time.

DECLARATION

I declare that the dissertation/thesis, which I hereby submit for the degree MSc in Financial Engineering at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

ETHICS STATEMENT

The author, whose name appears on the title page of this dissertation/thesis, has obtained, for the research described in this work, the applicable research ethics approval.

The author declares that s/he has observed the ethical standards required in terms of the University of Pretoria's Code of Ethics for Researchers and the policy guidelines for responsible research.

Yashin Gopi

January 2023

ACKNOWLEDGEMENTS

This project could not have been completed without the guidance of my friend and supervisor Professor Eben Maré. Thank you for convincing me to sign up for my MSc, and for pushing me to get it over the line. Your constant support and mentorship has been nothing short of phenomenal.

To Professor Dave Bradfield and the quantitative research team at Cadiz Securities – thank you for taking a chance on me all those years ago. Much of this work began under your supervision, and I am glad to have finally gotten it into a proper academic setting.

To my colleagues at Sentio Capital Management – thank you for all the support and encouragement. It is truly a fantastic working, learning, and fun environment.

I would like to thank my parents, Kishore, Ashara, Ish, and Janey for instilling a love of learning in me, and for reminding me that anything is possible if you put your mind to it.

To my little ones, Tiggr, Bella, Sachin, and Viradh. Sorry for ignoring you for so long. I promise to spend more time with you now! But hopefully this dissertation serves as a reminder that we never stop learning, no matter how old we are.

And finally, to Yuri. I know that spending late nights and weekends working on this dissertation, has meant less time for us to be together, laughing, reading, watching tv, drinking tea, and just enjoying each other’s company. I also know that you have had to shoulder a greater burden during this time. Thank you so much for supporting me and allowing me the freedom to do this. I could not have asked for a better life partner…

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

Abbreviation	Meaning
ALCA	Average Linkage Cluster Analysis
ALSI	FTSE/JSE South African All Share Index
APT	Arbitrage Pricing Theory
ARI	Adjusted Rand Index
BC	Betweenness Centrality
CAPM	Capital Asset Pricing Model
CC	Closeness Centrality
COVID-19	Coronavirus Disease 2019
DBHT	Directed Bubble Hierarchical Tree
DC	Degree Centrality
EC	Eccentricity Centrality
EVC	Eigenvector Centrality
FTSE	Financial Times Stock Exchange
GARP	Growth at a Reasonable Price
GFC	Global Financial Crisis (2008)
GICS	Global Industry Classification Standard
ICB	Industry Classification Benchmark
INDI25	FTSE/JSE South African Industrial 25 Index
JSE	Johannesburg Securities Exchange
MDS	Multi-Dimensional Scaling

Abbreviation	**Meaning**
MSCI	Morgan Stanley Capital International
MST	Minimal Spanning Tree
NTL	Normalised Tree Length
NYSE	New York Stock Exchange
PCA	Principal Component Analysis
PMFG	Planar Maximally Filtered Graph
QIS	Quadratic-Inverse (covariance) Shrinkage
RMT	Random Matrix Theory
SLCA	Single Linkage Cluster Analysis
SOM	Self-Organising Maps
SVG	Scalable Vector Graphics
SWIX	FTSE/JSE South African Shareholder Weighted All Share Index
TMFG	Triangulated Maximally Filtered Graph
UK	United Kingdom
USA	United States of America

Important Definitions and Concepts

Concept	Definition
Econophysics	A portmanteau of economics and physics. The application of methods from physics to match observed financial markets data, including financial market anomalies, followed by the proposition of more general theoretical frameworks.
Network	A network is a representation of objects/entities in a system and the interactions/relationships between these entities. The entities are referred to as nodes or vertices, and the relationships between the entities are referred to as links or edges.
Correlation-based Network	Typically used to model a network in a stock market by using the strength of the relationships (which are captured by the edges/links) amongst the stocks in the market (which are represented by the nodes/vertices). The strength of the relationship is usually defined by determining the similarity of stocks based on the correlation of share price movements.
Network Filter	A process to convert a dense network (such as a correlation matrix in a stock market) into a sparse network, by filtering out noise, while still retaining the maximum amount of useful information. The structure of the filtered network is then analysed to discover important insights into the collective properties of the underlying system.
Denoising	Correlation matrices are estimated from noisy time series data, and typically the number of historical observations that are used in the estimation process is limited. This results in the estimated correlation matrix including some amount of noise. Techniques from Random Matrix Theory (RMT) and the so-called shrinkage methods have been used to reduce noise and enhance any signal that is inherent in an empirical correlation matrix, prior to using any network filtering techniques.
Market Mode	The largest common factor affecting all the shares in a market. The impact of this factor is usually significantly larger as compared to any other factor, and consequently, it tends to swamp any other interesting effects or hidden structures in the underlying data. It is therefore typically removed prior to performing any analysis.

Concept	Definition
Minimal Spanning Tree (MST)	A popular network filter that satisfies the following criteria: any two vertices are connected to each other by exactly one unique path (i.e. a tree structure), all vertices are connected (i.e. they are spanning), and the total distance across all the edges in the tree is minimised (i.e. the distance is minimal).
Planar Maximally Filtered Graph (PMFG)	A network filter that can be thought of as a less constrained version of the MST, allowing the filtered network to retain more links (and therefore more information). In graph theory, a planar graph is a graph that can be embedded in the plane, i.e. it can be drawn on a flat surface such that no edges cross each other.
Hierarchical Cluster Analysis	A method for grouping entities into a tree-like structure, or taxonomy, based on how similar they are. It has been shown to have deep links with the MST and PMFG.
Directed Bubble Hierarchical Tree (DBHT)	A hierarchical clustering algorithm that emerges from the Planar Maximally Filtered Graph. An important benefit of the DBHT algorithm is that the user of such a technique does not have to specify a priori how many clusters to group the data into.
Bootstrapping	Introduced by Efron (1979), it is one of the common statistical ways of assessing reliability or confidence for empirical statistics that do not have readily available statistical tests.
Network Topology Measures	Metrics that can be used to describe the structure of a network, such as the overall length of the network, or the importance/centrality of specific entities/vertices.

Chapter 1 Introduction

1.1 Motivation for the Study - Stock Markets as Systems

The job of an equity fund manager is a difficult one. They must digest large quantities of information from varying sources, and then determine what information is relevant to the universe of stocks that they can invest in. Once this information has been synthesised into views on the stocks, a portfolio must be constructed, while considering all the relevant information about the joint behaviour of these stocks. For many decades, practitioners have turned to traditional financial market theories to help navigate this difficult task. However, many practitioners have found that these theories have struggled to match the complicated reality of financial markets. This is borne out by the proliferation of many empirical anomalies that are observed in real financial market data but are at odds with the traditional theories of finance.

These contradictions are often due to the assumptions upon which these theories are built, namely that financial markets are efficient; that stock price returns are normally distributed; that investors behave rationally; and that risk and return are linearly related. Over time, the focus of academic researchers has shifted to try and explain these anomalies by relaxing some of the assumptions. The field of behavioural finance is one such attempt in which researchers focus on how psychological biases can lead to investors behaving irrationally, thereby explaining some of the anomalies that are observed in financial markets.

An alternate track was followed by engineers, mathematicians, and physicists, who began joining financial markets in earnest in the late 1980s. These researchers applied methods from physics to match the observed financial market data, including the so-called financial market anomalies, followed by proposing more general theoretical frameworks to explain the results. This approach was counter to traditional financial markets theory in which simplified theories were given precedence (for the sake of mathematical tractability), despite disagreement with empirical data. This field was named *econophysics*, a portmanteau of economics and physics. Many of the physicists came from the sub-field of statistical mechanics which provides a mathematical description of the relation between large quantities of microscopic entities and the macroscopic behaviour that emerges from such a system (i.e., how does the collective behaviour of the small entities impact the large scale behaviour of the system). Such systems are called *complex adaptive systems* if they consist of entities, or agents, that behave and interact in a diverse manner, which can then lead to non-linear emergent behaviour of the system.

Mauboussin (2012) describes complex adaptive systems as follows:

> You can think of a complex adaptive system in three parts... First, there is a group of heterogeneous agents. These agents can be neurons in your brain, bees in a hive, investors in a market, or people in a city. Heterogeneity means each agent has different and evolving decision rules that both reflect the environment and attempt to anticipate change in it. Second, these agents interact with one another, and their interactions create structure— scientists often call this emergence. Finally, the structure that emerges behaves like a higher-level system and has properties and characteristics that are distinct from those of the underlying agents themselves.

One can see how stock markets can be viewed as complex adaptive systems. The stock market consists of a variety of agents (such as retail investors, institutional investors, hedge funds, banks, regulatory institutions, etc.), each with various incentives, and therefore behaving in a heterogeneous manner. Given that these agents observe the market and then adapt their behaviour (often irrationally so), they create feedback loops as the output from one phase of the market becomes the input for the next phase of the market, therefore changing the structure of the market. This collective behaviour cannot be predicted by analysing and observing the individual market participants alone. Furthermore, the complex adaptive behaviour of market participants can lead to the complex adaptive behaviour of the underlying stocks in the market.

In the last twenty years, there has been an increasing focus on modelling complex adaptive systems using network theory. A network is a representation of entities or agents in a system and the interactions or relationships between these entities. Network theory has been used to model a variety of such systems. For example, in biological sciences, network theory has been used to model human diseases (Barabási, Gulbahce and Loscalzo, 2011) as well as the spread of infectious diseases (Brockmann and Helbing, 2013). Another example is social network analysis in which social structures are investigated using networks. This analysis extends from physical, real-life networks, such as students in classrooms (Grunspan, Wiggins and Goodreau, 2014), to online social networks (Grandjean, 2016). Recent literature from the econophysics field has focused on the modelling of financial and economic systems using such a network-based approach (see Section 2.8 for more information).

1.2 Research Aim

In this dissertation, we aim to apply techniques from network theory to analyse the complex interactions amongst stocks in the South African stock market. We study correlation-based networks where stocks are represented as entities in the network, and the relationships amongst the stocks are based on the strength of the co-movements of the stocks (as measured by the standard Pearson correlation metric).

In particular, network filtering tools are used to prune less relevant information, or noise, in these networks, thereby allowing the important macroscopic and mesoscopic (i.e., on a scale between macro and micro) structures to emerge.

These filtered networks are also accompanied by a visual representation that allows the user to easily unearth meaningful information about the complex market dynamics and its emergent structure. Furthermore, one can extract useful metrics from such networks that describes their structure or topology, highlighting which shares are important or central in the network. The analysis of the temporal evolution of networks can also assist in understanding the underlying trends in the structure of the market.

Lastly, many of these techniques have also been shown to have deep relationships with traditional hierarchical cluster analysis, allowing the user to draw on the rich knowledge base from this field.

1.3 Research Objectives

We focus on the application of network filtering tools to financial markets, by introducing the Minimal Spanning Tree (MST), and the Planar Maximally Filtered Graph (PMFG). The PMFG is an interesting technique that has a hierarchical clustering representation called the Directed Bubble Hierarchical Tree (DBHT). This method is novel in that it is fully unsupervised and does not require the user, or the use of an external technique, to choose or validate the number of groups or clusters that the stocks are separated into.

While the MST has been applied to the South African stock market, to the best of our knowledge this is the first application of the PMFG and the DBHT in such a setting.

1.4 Dissertation Structure

This dissertation is organised as follows. In Chapter 2 we introduce the *basic concepts* of network theory, followed by real-life examples of networks from a variety of fields. We then address the *application of network filtering tools to financial markets*. We begin by discussing how the relationships between stocks have been modelled in the academic literature, followed by the introduction of simple filters such as Asset Graphs, and

Threshold Networks, leading towards the more advanced methods such as MSTs, and PMFGs. We also highlight the deep relationship between these network filters and hierarchical cluster analysis, paying particular attention to the DBHT. We then provide a detailed examination of network topology measures, highlighting the important metrics that measure importance or centrality in networks.

In Chapter 3 we discuss various *modelling, pre-processing, and sample/feature selection techniques and considerations.* The choices associated with many of these concepts can have a significant impact on the outcome of any analysis, and so we discuss each of them in detail, referring to published literature or practitioner insights to guide the choices that are made. We then provide a thorough insight into the methodology that has been followed in the analysis.

In Chapter 4 we present the *results* of our analysis of the South African stock market. We consider both a *long-term* (or static) analysis, followed by a *dynamic* (or temporal) analysis. The dynamic analysis is highly relevant to investment professionals as it allows us to determine the impact that varying market conditions have on the nature of the structure of the South African stock market.

In Chapter 5 we *discuss* the results and propose *future directions* of research.

Appendix A contains the full list of the shares that were considered for the analysis, as well as their full names, and economic classifications. It also contains the filtering that was applied at each step to reduce the universe of stocks from 136 down to 72. Appendix B contains information relating to the source code that was used to conduct the analysis.

Chapter 2 Network Analysis

2.1 What is a Network?

A network is a representation of *objects*/*entities* in a system and the *interactions*/*relationships* between these entities. The entities are referred to as *nodes* or *vertices*, and the relationships between the entities are referred to as *links* or *edges*. In mathematics, networks are often referred to as graphs, and the area of mathematics concerning the study of graphs is called graph theory. In this dissertation, we will use the terms networks and graphs interchangeably.

The earliest academic work on networks and graph theory is thought to have stemmed from the work by the mathematician Leonard Euler in 1736 in relation to a riddle called the Seven Bridges of Königsberg (Newman, Barabási and Watts, 2006). In the city of Königsberg, there existed seven bridges that connected various land masses and a popular brainteaser was to devise a walk through the city that would cross each of those bridges once and only once. Euler proved the impossibility of such a path by making use of a graph. He abstracted away all details of the original problem except for the connectivity, leaving four vertices which represented the four land masses and the seven edges joining the vertices in the pattern of the seven bridges, i.e. leaving a graph.

In the subsequent sections of this chapter, we introduce the following basic concepts of graph theory that will be used to describe and analyse networks in this dissertation:

- The definitions of the components of a network (or graph)
- The types of edges (or links) in a network
- Vertex (or node) attributes
- The topology or structure of a network
- Algorithms for the layout of a network

2.2 Basic Definitions and notation

A network can be represented either (i) graphically, (ii) using set notation, or (iii) using matrix notation. We illustrate these representations using a hypothetical example of a simple network of friendships amongst a group of five people (taken from the website 'Graph Theory – The Network Pages', no date) Each person in the network is represented by a node or vertex, while the existence of a relationship between two people is represented by a link or edge between the two relevant nodes.

A *graphical* representation of this simple network can be seen in Figure 2.1, with the original figure on the left, and a software-generated graph on the right.

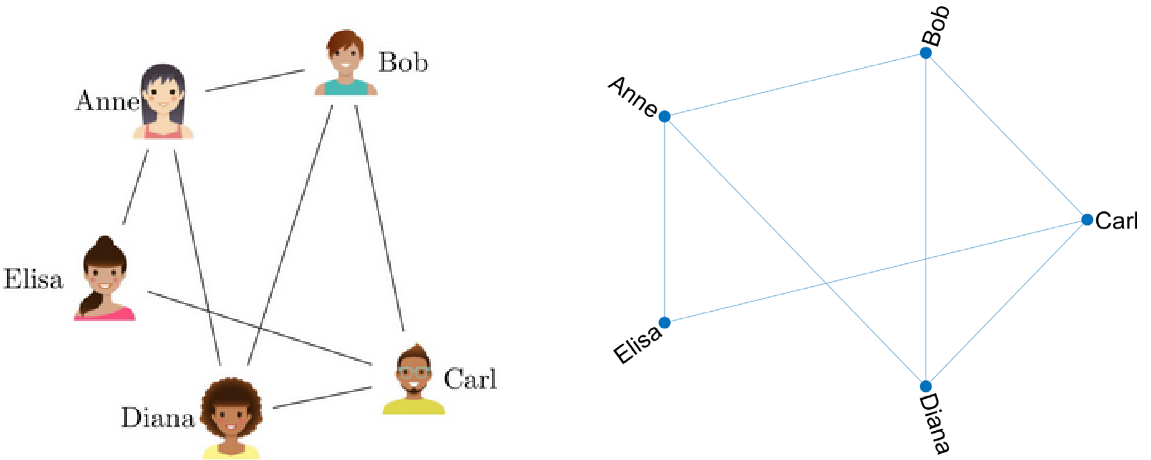


Figure 2.1 Simple Friendship Network: Illustration (left) and Abstracted Graph (right)

One can immediately see the various relationships among these people. For example, Diana is friends with three people (Anne, Carl, and Bob), while Elisa is friends with Carl and Anne.

Secondly, using *set notation*, a graph (G) can also be characterised by a set of nodes/vertices (V), which consists of a list of all of the nodes/vertices (V_i) in the graph, and a set of edges (E) which consists of a list (E_i) of all of the connected vertices in the graph. Using this notation, the friendship graph from Figure 2.1 can also be represented as follows:

$$V = \{Anne, Elisa, Diana, Carl, Bob\}.$$

$$E = \{(Anne, Bob), (Anne, Elisa), (Anne, Diana), (Elisa, Carl),$$
$$(Diana, Carl), (Diana, Bob), (Carl, Bob)\}\ .$$

Finally, a network or graph can be represented by an *adjacency matrix.* This is a square matrix A where:

$$A_{ij} = \begin{cases} 1, & \mathit{if\ there\ is\ an\ edge\ from\ vertex\ i\ to\ vertex\ j} \\ 0, & \mathit{otherwise.} \end{cases}$$

Using this notation, the adjacency matrix of the friendship graph from Figure 2.1 can be represented as follows:

Table 2.1 Simple Friendship Network: Adjacency Matrix

	Anne	Bob	Carl	Diana	Elisa
Anne	0	1	0	1	1
Bob	1	0	1	1	0
Carl	0	1	0	1	1
Diana	1	1	1	0	0
Elisa	1	0	1	0	0

2.3 The Types of Network Edges

The edges of a network can be embedded with additional information about the relationships between each of the entities (which are represented by the vertices). This additional information gives rise to the following types of edges in a network:

- **Undirected edges**: There is a simple connection between two vertices with no implication of any directionality or flow in the relationship (e.g. a simple friendship between two people).
- **Directed edges**: There is directionality from one vertex to the other (e.g. unrequited love between two people in a friendship network).
- **Weighted edges**: An edge can contain quantitative information about the strength of the link between two vertices (e.g. the strength of the relationship in a friendship network).

Weighted edges in a network can be represented by a *weighted adjacency matrix*. If there is an edge between vertices i and j then

$$A_{ij} = \begin{cases} \mathit{weight\ of\ edge}\ (i,j), & \mathit{if\ there\ is\ an\ edge\ from\ vertex\ i\ to\ vertex\ j} \\ 0, & \mathit{otherwise.} \end{cases}$$

Returning to the hypothetical example of the friendship network from Figure 2.1, let us assume that varying levels of friendship are represented on a scale of zero to four (with a value of zero indicating no relationship and a value of four indicating the strongest level of friendship), leads to the weighted adjacency matrix in Table 2.2.

Table 2.2 Simple Friendship Network: Weighted Adjacency Matrix

	Anne	Bob	Carl	Diana	Elisa
Anne	0	2	0	1	4
Bob	2	0	2	1	0
Carl	0	2	0	1	4
Diana	1	1	1	0	0
Elisa	4	0	4	0	0

In this example, while Diana has three friendships and Elisa has two, the strength of Elisa's two friendships is much stronger (with a higher weight of four), as compared to Diana's friendships (each with a lower weighting of one). These edge weights can also be encoded in the visual graph by varying the thickness, the line style, or the colour of each edge, in proportion to the value of each weight. An example of this is shown in Figure 2.2, with the thickest edges originating from Elisa, and the thinnest from Diana.

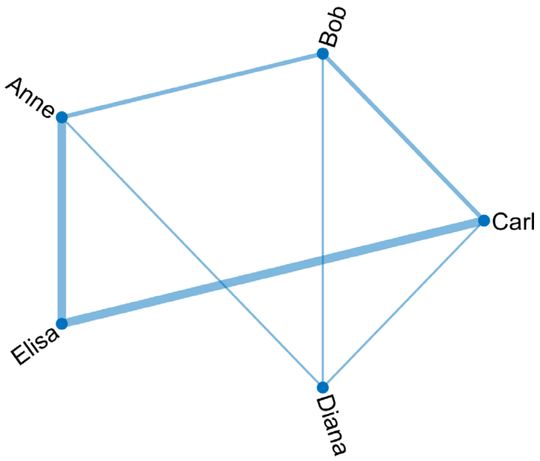


Figure 2.2 Simple Friendship Network: Graph Representation of the Weighted Adjacency Matrix

2.4 Vertex (or Node) Attributes

Like the edges in a network, the vertices of a network may also have additional information attached to them. This information may be a qualitative attribute of each vertex (perhaps some sort of categorical data, or data on an ordinal scale), or quantitative data. This data can again be encoded on the graph by varying the colour or the size of each vertex in accordance with each vertex's attribute. For example, in the simple friendship network, we can perhaps colour each vertex according to the school that each person goes to. If the three strongest friends (Anne, Elisa, and Carl) go to the same school, then the graph may look like the one in Figure 2.3.

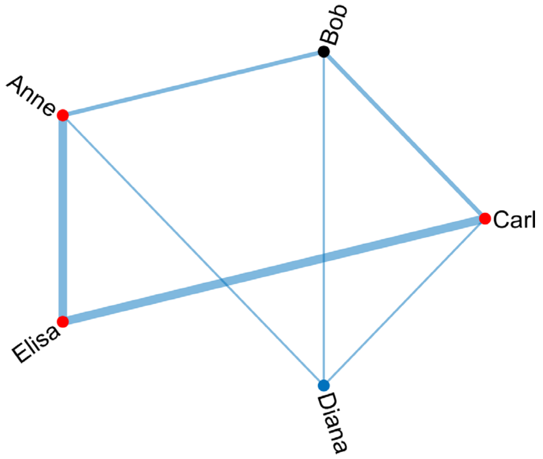


Figure 2.3 Simple Friendship Network: Graph Representation of Node Attributes

2.5 An Introduction to Network Topology

Many real-life networks are large, contain many vertices, and have complicated structures. However, they also have properties which can be extracted to glean various insights from the network. *The way that the vertices and edges in a network are arranged is referred to as the topology of the network*. These topological properties may refer to the overall network, or the vertices or edges. For example, the *normalised length* can be used to calculate the overall strength of the relationships amongst all of the entities in the network. However, the various *centrality metrics* are used to describe which vertices and edges are important in a network. The centrality metric that one chooses, depends on the definition of importance. For example, the *degree* of a vertex is a metric which defines importance based on the number of connections (or edges) emanating from a vertex. Vertices with many edges are deemed to be important (and likened to a person having many friends in the network from Figure 2.1). If importance was defined as the ability to connect other vertices, then the so-called *betweenness centrality* metric would be useful in identifying important vertices. Alternatively, the *closeness centrality* metric defines importance as a vertex that is close to other vertices.

Figure 2.4 from Oldham et al. (2019) shows how various vertices can be deemed to be important depending on the topology of the network and the definition of importance. Panel A shows an example of a star network. The red vertex in the middle has the greatest number of connections, the highest closeness centrality, and the highest betweenness centrality. For this network, the three centrality measures agree. However, Panel B shows a network in which vertex importance, depends on the metric that is being used. The red node has the highest betweenness and closeness centrality, but it also has the lowest number of connections.

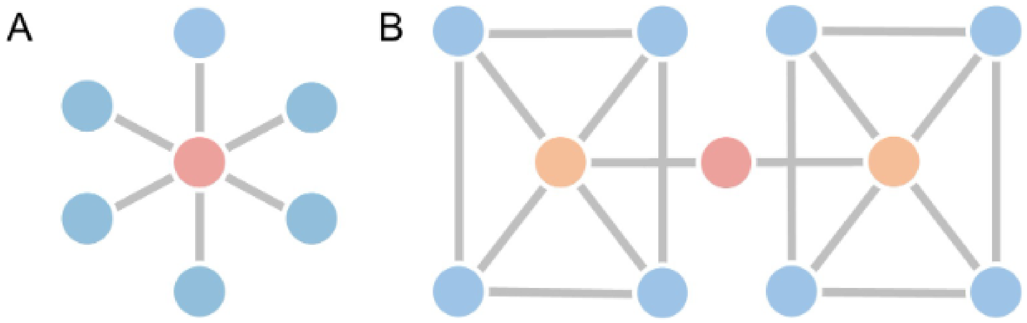


Figure 2.4 An Example of Network Topologies and Centrality Metrics (Oldham *et al.*, 2019)

2.6 Graph Layout Algorithms

Humans are visual creatures by nature. The theory of *multimedia learning* (Mayer, 1997) states that we can learn more deeply from words and pictures together than we can from just words alone, i.e. it is often easier for people to retain and understand information if there is an accompanying visual component. Edward Tufte, who is considered to be a pioneer in the field of data visualisation, said the following in his book "*The Visual Display of Quantitative Information"* (2001):

> Modern data graphics can do much more than simply substitute for small statistical tables. At their best, graphics are instruments for reasoning about quantitative information. Often the most effective way to describe, explore, and summarize a set of numbers - even a very large set - is to look at pictures of those numbers. Furthermore, of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful.

Therefore, using a graph to visually represent a network, as opposed to examining an adjacency matrix, is often preferred. For high dimensional networks, a graph representation can often reveal patterns, trends, outliers, and connections that may be especially difficult or impossible to find in any other way.

However, graphs of larger networks must be created using dedicated visualisation software or toolboxes. The following list is a sample of such software:

- Gephi (Bastian, Heymann and Jacomy, 2009)
- Pajek (Batagelj and Mrvar, 2004)
- MATLAB (MATLAB, 2022)
- The Networx toolbox (Hagberg, Swart and S Chult, 2008) in Python (Van Rossum and Drake, 2009).
- The Statnet toolbox (Pavel N. Krivitsky *et al.*, 2003) in R (R Core Team, 2016).

Drawing a graph is not a simple process. This is because the method that is used to draw the graph can have a significant impact on the interpretation of the graph. For example, if the vertices of two entities are plotted near each other, the user may infer that the two entities have a significant relationship even if they do not have any significant edges or paths that link them together. Therefore, the objective of any graph layout algorithm is to reveal important relationships without misleading the user.

Gibson, Faith and Vickers (2013), Hu (2011), and Bajramovic et al. (2011) provide surveys of commonly used methods for visualising networks. These methods typically fall into one of three categories: force-directed methods, dimension reduction methods, and multi-level methods. For simple networks, such as the Friends network in Figure 2.1, the circular layout (which is the layout used in that figure) can be a useful starting point.

The force-directed methods are the most widely used algorithms and are typically integrated into many network visualisation tools. These methods aim to create graphs that target certain aesthetic properties such as minimising edge crossing; enforcing symmetry; imposing uniform lengths of edges and distribution of vertices; separating vertices that are not linked; and preventing vertices from overlapping. Maximising these aesthetic principles should intuitively improve the readability of a graph. However, for large graphs, these aesthetics can have a long runtime and may not lead to a globally optimal layout (violating the targeted aesthetic principles). Popular force-directed methods are the Fruchterman-Reingold layout (Fruchterman & Reingold 1991), the Kamada-Kawai layout (Kamada & Kawai 1989), and the ForceAtlas method which was developed by Jacomy et al. (2014) for use in Gephi.

The dimension reduction techniques are used to take high dimensional data and project them down onto a lower dimensional space while retaining as much of the information from the high dimensional space as possible. Popular dimension reduction techniques are Multi-Dimensional Scaling (MDS) methods such as Pivot MDS by Brandes and Pich (2007), linear dimension reduction methods such as High-Dimensional Embedding from Harel and Koren (2004) and self-organising maps (SOM).

Multi-level (or multi-scale, or multi-dimensional) techniques that are used for large graphs, start with force-directed algorithms and make them more efficient. They do this by recursively *coarsening* the graph and then refining it using layout refinements. Examples of this technique are the Yifan Hu Multilevel method (Hu, 2005) and the OpenOrd method (Martin *et al.*, 2011) which is available in Gephi.

Table 2 in Gibson, Faith and Vickers (2013) contains a useful summary of the various methods, their performance, the key differences, and size limitations. In this dissertation, given that the number of entities that are being graphed is not onerous (less than one hundred), the Kamada and Kawai method (in the Pajek graphing software) is used to layout the graph.

2.7 Examples of Networks

The hypothetical network shown in Figure 2.1 is a simple example, which was used to explain the basic concepts of a network. In this section, we provide examples of network graphs that were generated from actual data, followed by a brief commentary on the insights that can be gleaned from each one.

2.7.1 Character Interactions in Game of Thrones

Beveridge and Shan (2016) generated a network based on the Game of Thrones series of books, and the television series, by George R. R. Martin. The franchise is well known for its complicated plotlines, with a vast number of characters and groups of characters, that are spread over multiple geographic locations. Therefore, these books provide the perfect environment to showcase the benefits of network theory. The authors generated a network based on the interactions between the characters in the third book of the series, "*A Storm of Swords*" (Martin, 2002). This network can be seen in Figure 2.5. The authors noted the following about the structure of the network:

> The complex structure of our network reflects the interweaving plotlines of the story. Notably, we observe two characteristics found in many real-world networks. First, the network contains multiple denser subnetworks, held together by a sparser global web of edges. Second, it is organized around a subset of highly influential people, both locally and globally.

They also mentioned the following conclusion regarding important characters in the book:

> In our network, three characters stand out consistently: Tyrion, Jon, and Sansa. Acting as the Hand of the King, Tyrion is thrust into the center of the political machinations of the capitol city. Our analysis suggests that he is the true protagonist of the book. Meanwhile, Jon Snow is uniquely positioned in the network, with connections to highborn lords, the Night's Watch militia, and the savage wildlings beyond the Wall. The real surprise may be the prominence of Sansa Stark, a de facto captive in King's Landing. However, other players are aware of her value as a Stark heir and they repeatedly use her as a pawn in their plays for power. If she can develop her cunning, then she can capitalize on her network importance to dramatic effect.

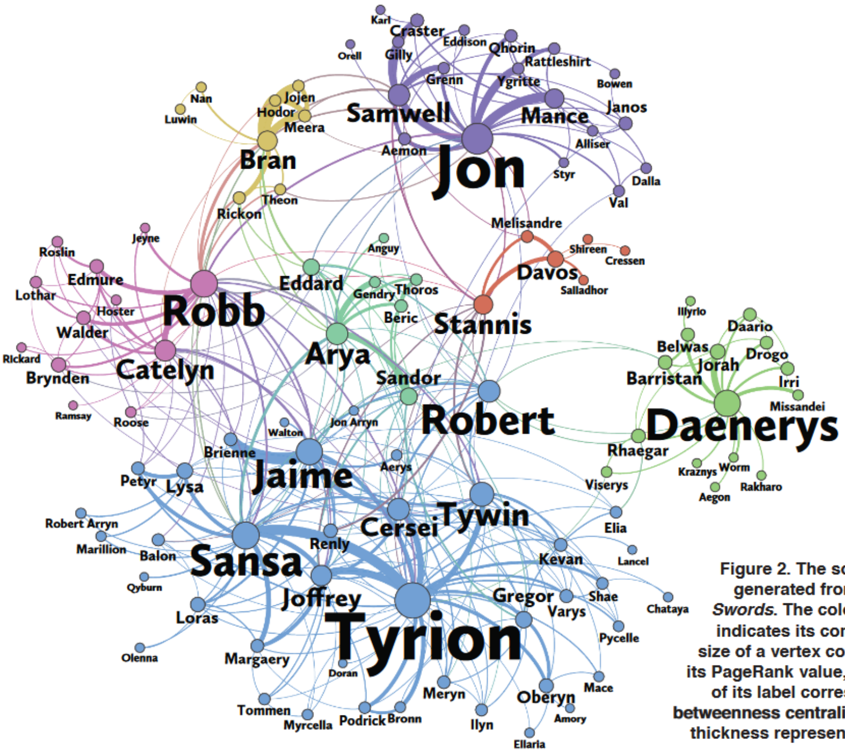


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.

Figure 2.5 A Social Network Generated from the Book, A Storm of Swords

(Beveridge and Shan, 2016)

So, using techniques from network theory, the authors were able to take a complex storyline, and synthesise important information from it. They were also able to provide conclusions regarding the importance of certain characters that may not be obvious upon cursory reading.

2.7.2 Passing Networks in Football

Advances in technology have led to the rapid growth in the quantity and type of data that is available for modern football teams to analyse. By using wearable devices, video tracking systems, and manual data capture processes, large quantities of data can be obtained. This data can be used for a variety of purposes, such as player scouting, opposition analysis, individual player improvement, and tactical/positional analysis. The way that players interact with each other on the football pitch (by passing the ball to each other) has led to the creation of passing networks. In a passing network, the vertices are the football players and edges represent the number of passes between any two players of the team. Buldú et al. (2019) used techniques from network science to analyse the passing signature of the historic Futbol Club (FC) Barcelona team that was managed by Pep Guardiola from the period 2008 to 2012. Figure 2.6 shows the passing network for F.C. Barcelona during a match played against Real Madrid, during the season 2009/2010. The position of each vertex/player reflects their average position on the field, and the width of the edges/links is proportional to the number of passes between the connected players.

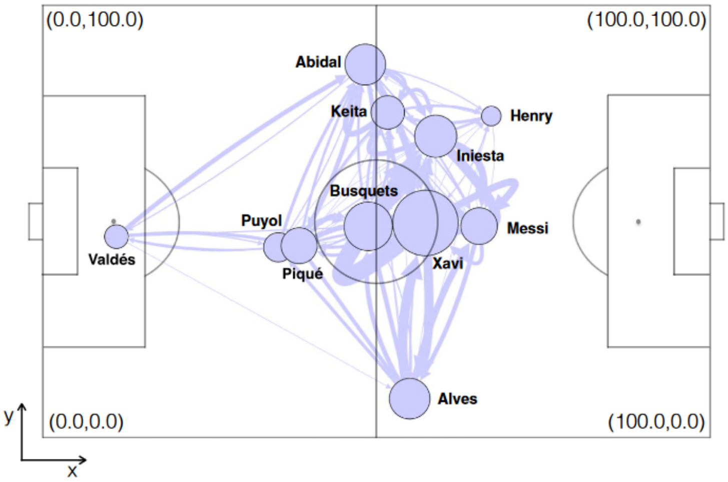


Figure 1. Schematic illustration of a football passing network. In the plot, players are represented by circular nodes, whose size is proportional to their eigenvector centrality, a mesure of importance in the network structure. The position of each player is given by the average of the positions of all passes made by the player along the match. The width of the links is proportional to their weights, which account for the number of passes between players. Note that links are unidirectional. In this example, we plot the average passing network of the match between F.C. Barcelona and Real Madrid, played during the season 2009/2010 at Santiago Bernabeu Stadium. Datasets leading to the passing network were provided by Opta.

Figure 2.6 A Football Passing Network for F.C. Barcelona (Buldú *et al.*, 2019)

The large size of the vertex for the player Xavi, indicated that he was important in the passing network (and many football fans who have him play would agree with this result), while the width of the edge from the players Iniesta to Messi, indicated that Iniesta provided the bulk of the passes forward to Messi (who scored a total of 47 goals in all competitions that season, winning his first Ballon d'Or trophy).

2.7.3 A Network of Illegal Ivory Trade

The illegal hunting of wild animals is a global problem that has led to the decline of many species of wildlife. In Africa, the period from 1979 to 1989 saw the population of elephants decline by 50% as they were hunted for their ivory. Huang, Wang and Wei (2020) constructed a country-level ivory trading network to illustrate the smuggling routes and the volume of ivory that was smuggled between countries. The analysis aimed to determine which countries were the key hubs in this network, and what were the most significant smuggling routes. Figure 2.7 shows this network laid out on to a map of the world. Countries from the different continents are marked with different colours and the size of a vertex illustrates the largest trading routes. The thickness of the edges represents the trafficking intensity of ivory between countries (with a thicker edge implying a higher intensity).

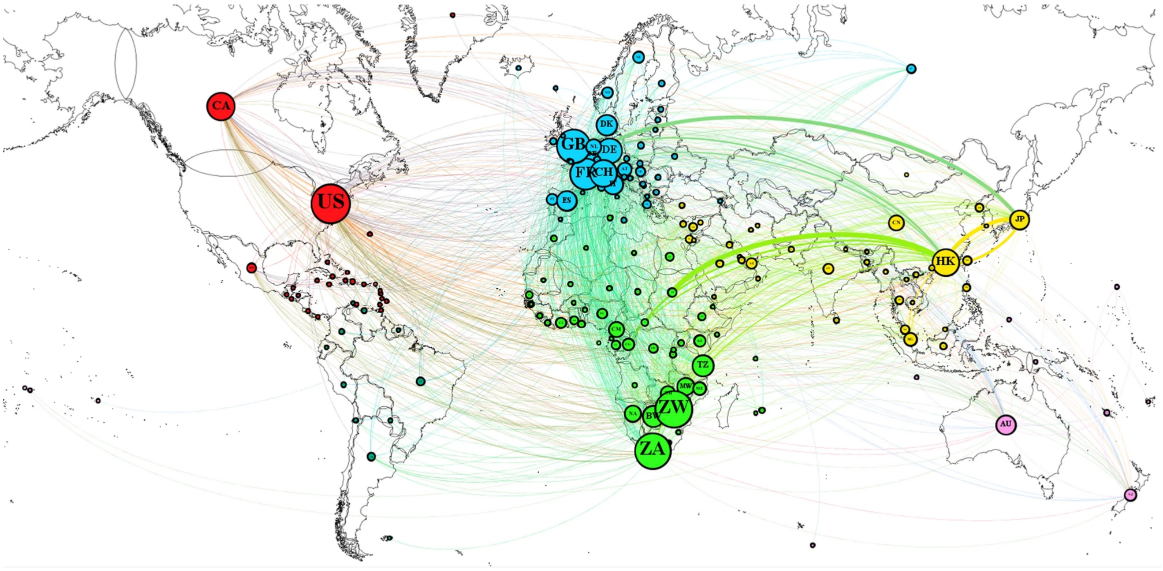


Figure 2.7 Illegal Ivory Trading Network (Huang, Wang and Wei, 2020)

Based on the analysis, Huang, Wang and Wei (2020) provided various insights, as well as targeted strategies to effectively control the ivory trading network. They noted that the most important hubs in the worldwide ivory trade were identified as the USA, the UK, Zimbabwe, South Africa, China, Japan, Sudan, Belgium and Hong Kong. They suggested that customs should strengthen inspections of vessels coming and going between these countries. They also noted stated that three significant ivory trafficking routes will be of more concern in the future and should be closely monitored. These routes were from African countries to Asian countries, from Belgium to Asian countries, and between Japan and Hong Kong.

2.8 Networks in Stock Markets

A network in a stock market can be modelled using the strength of the relationships (which are captured by the edges/links) amongst the various stocks in the market (which are represented by the nodes/vertices). The strength of the relationship is usually defined by determining the similarity of stocks based on the Pearson correlation of share price movements (or returns). This results in a network that is based on how market participants view stocks and trade them accordingly.

While return-based Pearson correlation networks have been studied extensively, one can define similarity in terms of a variety of metrics to highlight features of stock price co-movements that are also believed to be important (such as the non-linear share price co-movements or the lead-lag relationships amongst stocks). Similarity measures that focus on non-linear relationships amongst stock price movements, and a sample of academic studies that address these measures, are shown in Table 2.3.

Table 2.3 Studies of Stock Market Networks that use Non-linear Metrics

Metric	Reference
Copula methods	Wang et al. (2014)
Kendall’s tau	Millington and Niranjan (2021) Musmeci et al. (2016)
Mutual information	Barbi and Prataviera (2019) Goh, Hasim and Antonopoulos (2018) Guo, Zhang and Tian (2018)
Spearman’s rank correlation	Millington and Niranjan (2021) Musmeci et al. (2016)
Tail dependence	Denkowska and Wanat (2020) Lohre, Rother and Schäfer (2020) Musmeci et al. (2016)

One can also focus on how the movements in one stock’s price affects the prices of other stocks, such as partial correlations, lead-lag relationships, and Granger causality. Studies that focus on these measures, can be seen in Table 2.4.

Table 2.4 Studies of Stock Market Networks that use Causal-Type Metrics

Metric	Reference
Granger causality	Billio et al. (2012)
Lead-lag relationships	Bennett, Cucuringu and Reinert (2022)
Partial correlations	Kenett et al. (2010) Millington and Niranjan (2020)

Instead of focusing exclusively on share price movements (which tells us how market participants view each stock), one could analyse the similarity amongst other features, such as volatility, trading volumes, accounting fundamentals, risk factor exposures, or the language/commentary in annual reports or on earnings calls.

Table 2.5 Studies of Stock Market Networks that use Alternative Features

Feature	Reference
Accounting fundamentals	Fodor, Jorgensen and Stowe (2021) Henningsen (2019) Knudsen, Kold and Plenborg (2017)
Language/commentary in annual reports or earnings calls	Winton (2018)
Risk factor exposures	Heywood, Marsland and Morrison (2003)
Share price volatility	Miccichè et al. (2003)
Trading volumes	Brida and Risso (2008)

Returning to the commonly used correlation-based network, ρ_{ij} can be defined as an entry in a $N \times N$ correlation matrix representing the correlation between any two shares, i and j, and can be calculated in the usual manner as follows:

Let N = the number of stocks in the market.

Let P_t^i = the price of stock i at time t.

Then $r_t^i = \frac{P_t^i}{P_{t-1}^i} - 1$ = the return of stock i at time t.

Let T = the number of return observations.

Let $\bar{r}^i = \frac{1}{T} \sum_{i=1}^{T} r_t^i$ = the average return of stock i.

$$\text{Then } \rho_{ij} = \frac{\sum_{t=1}^{T} (r_t^i - \bar{r}^i)(r_t^j - \bar{r}^j)}{\sqrt{\sum_{t=1}^{T} (r_t^i - \bar{r}^i)^2 \sum_{t=1}^{T} (r_t^j - \bar{r}^j)^2}} . \qquad \textbf{(1)}$$

Table 2.6 contains a small sample example of a correlation matrix[1] consisting of ten stocks that were listed on the Johannesburg Securities Exchange (JSE) in South Africa. We use the JSE ticker code to represent each share, while the full names and other details of these shares can be found in Appendix A1. The sample of stocks emanated from a variety of economic sectors such as general mining (AGL, and BHG), gold mining (GFI, and ANG), banking (FSR, SBK, and ABG), and retail (MRP, TFG, and TRU) sectors.

Table 2.6 Example of a Ten Stock Correlation Matrix

	AGL	BHG	GFI	ANG	FSR	SBK	ABG	MRP	TFG	TRU
AGL	1.00	0.63	-0.13	-0.10	-0.23	-0.28	-0.07	-0.29	-0.19	-0.51
BHG	0.63	1.00	-0.16	0.00	-0.29	-0.25	-0.27	-0.22	-0.12	-0.36
GFI	-0.13	-0.16	1.00	0.78	-0.56	-0.55	-0.65	-0.34	-0.47	-0.23
ANG	-0.10	0.00	0.78	1.00	-0.51	-0.62	-0.75	-0.36	-0.38	-0.31
FSR	-0.23	-0.29	-0.56	-0.51	1.00	0.64	0.63	0.02	0.11	0.10
SBK	-0.28	-0.25	-0.55	-0.62	0.64	1.00	0.72	0.08	0.08	0.11
ABG	-0.07	-0.27	-0.65	-0.75	0.63	0.72	1.00	0.12	0.06	0.16
MRP	-0.29	-0.22	-0.34	-0.36	0.02	0.08	0.12	1.00	0.34	0.29
TFG	-0.19	-0.12	-0.47	-0.38	0.11	0.08	0.06	0.34	1.00	0.33
TRU	-0.51	-0.36	-0.23	-0.31	0.10	0.11	0.16	0.29	0.33	1.00

Even in this simple correlation matrix, one can that there was some structure, where stocks within the same sector had a positive correlation to each other, and stocks across sectors had a low or negative correlation. This indicated that even in this simple example, the sector classification employed by the JSE did agree with the manner in which stocks actually traded. We will discuss this phenomenon in more detail in later sections.

However, to apply the techniques of network theory to a correlation matrix, one cannot simply use the correlation metric in its raw form. This is because many of the techniques from network theory require that the weights of the edges in a network must satisfy the three axioms of a *metric space*. One must therefore define a *distance metric* $d(i,j)$ such that:

$$\begin{array}{ll} I. & d(i,j) \geq 0 \ \ \textit{and}\ d(i,j) = 0 \ \textit{if and only if}\ i = j. \\ II. & d(i,j) = d(j,i). \\ III. & d(i,j) \leq d(i,k) + d(k,j). \end{array} \qquad \textbf{(2)}$$

[1] Note that these correlations were calculated after having subtracted the average return across the ten stocks from each stock. This is one of the methods for removing the *market mode* and creates a *relative* correlation matrix, which makes it easier to identify any emergent structure. Section 3.1.3 contains more information in this regard.

Axiom I is referred to as *positivity*, Axiom II as *symmetry*, and Axiom III as the *triangle inequality*. One can see why a network that is constructed purely from a correlation matrix would not satisfy the above axioms. For example, correlations can vary in value from -1 to 1, violating the first axiom. Also, the correlation of a stock with itself is equal to 1, and not 0 as required by the first axiom. The third axiom could also be violated in cases where you have two stocks that have a high correlation with each other, but each stock has a low correlation with a third stock (Birch, 2016).

Following Mantegna (1999) and Gower and Ross (1969) one can transform the correlation coefficient between stocks i and j into a valid distance metric as follows:

$$d(i,j) \;=\; \sqrt{2(1-\rho_{ij})}. \qquad \textbf{(3)}$$

The relationship between the distance metric and the correlation coefficient can be seen in Figure 2.8. The relationship is negative and non-linear, with a correlation of -1 implying a distance of 2, a correlation of 0 implying a distance of $\sqrt{2}$ (≈ 1.41), and a correlation of 1 implying a distance of 0. So, highly correlated stocks will have a distance close to 0, and in the extreme case, the distance of a stock to itself is 0. Therefore, this metric satisfies the axioms of a distance metric and can be used to create correlation-based networks where the smaller the weight of an edge (or the shorter the length) between any two vertices, the higher the correlation between them.

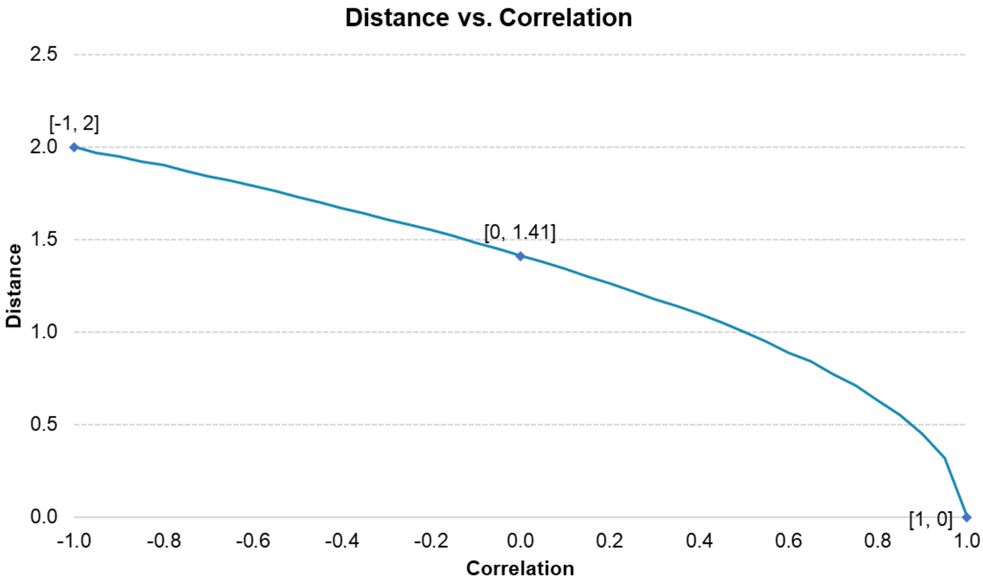


Figure 2.8 Inverse Relationship Between Distance and Correlation

Using the distance metric from Equation (3) one can construct a distance matrix associated with the correlation matrix from Table 2.6. This distance matrix is shown in Table 2.7.

Table 2.7 Example of a Ten Stock Distance Matrix

	AGL	BHG	GFI	ANG	FSR	SBK	ABG	MRP	TFG	TRU
AGL	0.00	0.86	1.50	1.48	1.57	1.60	1.46	1.61	1.54	1.74
BHG	0.86	0.00	1.52	1.41	1.60	1.58	1.59	1.56	1.50	1.65
GFI	1.50	1.52	0.00	0.66	1.77	1.76	1.82	1.64	1.71	1.57
ANG	1.48	1.41	0.66	0.00	1.74	1.80	1.87	1.65	1.66	1.62
FSR	1.57	1.60	1.77	1.74	0.00	0.85	0.86	1.40	1.33	1.34
SBK	1.60	1.58	1.76	1.80	0.85	0.00	0.75	1.36	1.36	1.33
ABG	1.46	1.59	1.82	1.87	0.86	0.75	0.00	1.33	1.37	1.29
MRP	1.61	1.56	1.64	1.65	1.40	1.36	1.33	0.00	1.15	1.19
TFG	1.54	1.50	1.71	1.66	1.33	1.36	1.37	1.15	0.00	1.16
TRU	1.74	1.65	1.57	1.62	1.34	1.33	1.29	1.19	1.16	0.00

As discussed, in Section 2.1 one can also represent such a network as a graph. This can be seen in Figure 2.9. Note that in this graph, the thickness of each edge is proportional to the correlation between the two connected or adjacent shares (vertices). Negative correlations between any two shares are depicted by edges that are coloured red. We have used a basic circular layout for this graph.

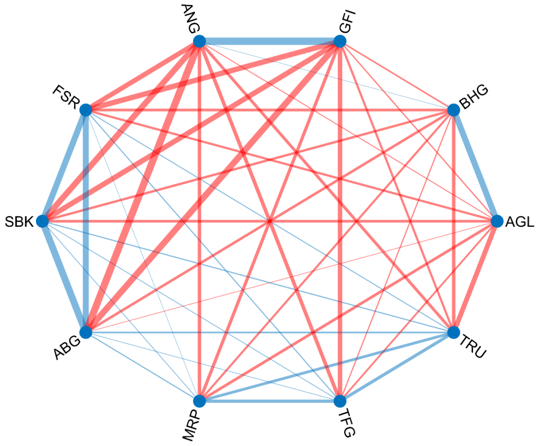


Figure 2.9 Graph Representation of the Ten Stock Correlation Matrix

From Figure 2.9 one can see that even in the case of this small network, inferring any meaningful insights from the graph is difficult because the network is fully connected (i.e. all vertices are connected). So, although a correlation matrix can be used to represent a network in a stock market, one needs to use techniques to filter out the less relevant edges from the network. This would leave us with the most informative connections and assist in revealing the hidden insights in the network. We discuss these filtering processes in the following section.

2.9 Network Filtering Techniques

Most analysis of stock markets tends to suffer from the curse of dimensionality. This refers to the fact that an analysis of a large number of stocks, with a limited length of time series data that is being used to model the relationships amongst the stocks, can often lead to spurious or noisy results. Therefore, a process that can filter out the noise, while still retaining the maximum amount of useful information (or the underlying signal in the data), is vital.

Furthermore, although stock markets consist of a large universe of stocks, the behaviour of these stocks tends to be to be driven by a smaller number of factors. In his well-known paper, Sharpe (1963) introduced the single index model, in which the returns of stocks are influenced (in varying degrees) by a general market factor. This theory was expanded upon by Ross (1976) with the concept of the Arbitrage Pricing Theory (APT) which posited the idea that the returns on stocks are driven by multiple factors from a variety of settings (e.g. equity markets, equity sectors, interest rates, currencies, economic factors, etc.). King (1966) put forward the idea that the share prices of certain groups of stocks tend to behave similarly, due to market participants viewing them as homogenous groups. This was one of the earliest known academic works that used the concept of *cluster analysis* to identify these homogenous groups. So, although stock markets can be viewed through one lens as a high dimensional setting, one can use techniques to reduce the complexity by grouping together stocks that exhibit similarity in the selected set of features.

To this end, *network filtering techniques* have been shown to be effective tools that can convert a dense network (such as a correlation matrix in a stock market) into a sparse network, and one can then analyse the structure of such a network to discover important insights on the *collective properties* of the underlying system. Many applications of network filtering in financial markets have been proposed. Mantegna (1999), is thought to be the pioneer of the utilisation of network filtering techniques in financial markets, proposing the use of a Minimal Spanning Tree (MST) to uncover structure (i.e. whether groups of stocks behaved similarly) amongst stocks in the Dow Jones Industrial Average and S&P500 indices. Other popular techniques include Asset Graphs, Threshold Networks, and Planar Maximally Filtered Graphs (PMFG). We discuss these network filtering techniques next.

2.10 Threshold Networks and Asset Graphs

Perhaps the simplest method to filter a correlation-based network is to focus on the strongest correlations (and therefore the shortest distances in the network) and to discard the rest of the connections (effectively setting the weights of these edges to a value of zero).

The *Asset Graph*, which was first proposed in Onnela, Chakraborti, Kaski, Kertesz, et al. (2003a), is a method in which pairwise correlations (which represent the weights of the edges in the network) are ranked from the strongest (i.e. the most correlated stocks) to the weakest and then most anti-correlated (or is terms of distances, the pairs are ranked from the shorted distance to the longest). Only a certain number of the top correlated pairs of stock are retained. In a stock market of *N* stocks (which results in a correlation matrix with $\frac{1}{2}N(N-1)$ unique entries), this number is usually set to *N-1*, but can be varied to create a denser network (by retaining more edges), or a sparse one by retaining fewer of the ranked edges.

The *Threshold Network* was introduced by Boginski, Butenko and Pardalos (2005). In this method, the fully connected stock market network is filtered to a less complex one by only including an edge between two stocks if their correlation is larger than a set threshold value. The complexity/density of the resulting network can be determined by varying this threshold value. The threshold may be absolute in nature (keeping the largest positive and negative correlations), or it may focus on only large positive correlations, in which case the network would favour stocks that are highly correlated to each other and ignore other information. For example, applying an absolute correlation threshold of |0.25| to the ten stock correlation matrix in Table 2.6 leads to the filtered correlation matrix in Table 2.8

Table 2.8 Ten Stock Correlation Matrix: Threshold Filtered = |0.25|

	AGL	BHG	GFI	ANG	FSR	SBK	ABG	MRP	TFG	TRU
AGL	1.00	0.63				-0.28		-0.29		-0.51
BHG	0.63	1.00			-0.29		-0.27			-0.36
GFI			1.00	0.78	-0.56	-0.55	-0.65	-0.34	-0.47	
ANG			0.78	1.00	-0.51	-0.62	-0.75	-0.36	-0.38	-0.31
FSR		-0.29	-0.56	-0.51	1.00	0.64	0.63			
SBK	-0.28		-0.55	-0.62	0.64	1.00	0.72			
ABG		-0.27	-0.65	-0.75	0.63	0.72	1.00			
MRP	-0.29		-0.34	-0.36				1.00	0.34	0.29
TFG			-0.47	-0.38				0.34	1.00	0.33
TRU	-0.51	-0.36		-0.31				0.29	0.33	1.00

Similar to Figure 2.9, one can represent the filtered correlation network from Table 2.8 as a graph. This can be seen in Figure 2.10. We show the original, fully connected network on the left and the filtered network on the right. One can see how the network has been "pruned" to keep only the largest (in magnitude) connections.

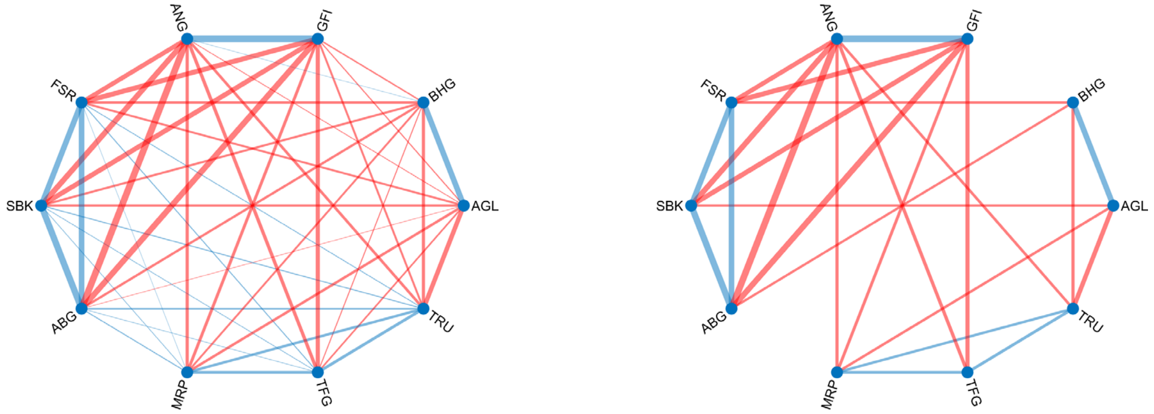


Figure 2.10 Graph of the Filtered Correlation Matrix (right-hand side) with Threshold = |0.25|

According to MacMahon and Garlaschelli (2015), because these methods discard the weakest correlations, and these correlations tend to be the noisiest, these methods tend to be more robust to noise. However, the choice of the threshold (in the case of the Threshold Network), or the number of edges to include (in the case of the Asset Graph) is arbitrary. One usually investigates how the properties of the filtered network changes as the threshold or the number of edges is varied.

But more importantly, the Asset Graph (and the Threshold Network if one only focuses on positive correlations, or the shortest distances), may also ignore an important feature of complex systems. In complex systems, such as stock markets, there are important relationships that exist on a local or microscopic scale (i.e. amongst the most closely linked vertices or stocks), but also on a global or macroscopic scale, with a hierarchical structure that may exist in the data. As mentioned by Song, Di Matteo and Aste (2012):

> We are therefore facing the problem of catching simultaneously two complementary aspects: on one side there is the need to reduce the complexity and the dimensionality of the data by identifying clusters which are associated with local features; but, on the other side, there is a need of keeping the information about the emerging global organisation that is responsible for cross-scale activity. It is therefore essential to detect clusters together with the different hierarchical gatherings above and below the cluster levels.

The Asset Graph and the Threshold Network methods only consider the local, pairwise correlation structure amongst stocks, and typically ignore any global or hierarchical correlation structure. This is because the use of a global correlation threshold (i.e. the same threshold or ranking cut-off across the entire network) prevents the identification of clusters where the correlation of the stocks within a cluster (i.e. the intra-cluster correlations) is lower than the threshold (and these connections will therefore be excluded from the network), but this correlation is still stronger than the correlation that these stocks have with other clusters (i.e. the inter-cluster correlations). Therefore, although the Asset Graph and Threshold Networks are valuable filtering techniques, they tend to discard a significant amount of information and are not best suited to detect any emergent clustering structure that may occur amongst stocks.

2.11 Minimal Spanning Trees (MST)

2.11.1 An Introduction to MSTs

One of the most popular methods of filtering a correlation-based network is to construct a Minimal Spanning Tree (MST) from the distance matrix of the network. In graph theory, a tree is a graph (with undirected edges) in which any two vertices are connected to each other by exactly one unique path (i.e. there are no loops or alternative paths that join one vertex to another). The MST is a tree in which all vertices are connected, which is what the term *spanning* refers to. Furthermore, the total distance across all the edges of the tree is minimised, hence the reference to *minimal*.

The MST is one of the most important and well-known filtering techniques emanating from graph theory. This is because there are algorithms to calculate the MST that are efficient and suitable for large networks. As such, it is a tool that has been used in many settings, such as the optimisation of computer and telecommunication networks, electrical grids, transportation routes, and water supply networks (Graham and Hell, 1985). Although the most popular and efficient algorithms that are used to create an MST are attributed to Kruskal (1956) and Prim (1957), it was thought to be Borůvka (1926), who presented the first algorithm to create an MST from a network. Note that all algorithms result in the same tree (unless there are non-unique distances between all edges), however, they can vary in the time taken, and the computational power required, to arrive at the solution. For the interested reader, Graham and Hell (1985), and Nešetřil (1997), provided a detailed history of the MST.

In finance, since the seminal work of Mantegna (1999), who applied the MST technique to uncover structure amongst stocks in the Dow Jones Industrial Average and S&P500 indices, there have been numerous papers in which the authors apply the MST technique to various other financial settings. Given the extended list of these papers, we list them in tabular format in Table 2.9, with the setting in which MSTs have been applied (on the left), followed by the reference (on the right).

Table 2.9 Studies of MSTs Applied to Financial Markets

Setting	Reference
Stocks: New York Stock Exchange (NYSE)	Bonanno et al. (2003) Onnela, Chakraborti, Kaski, Kertesz, et al. (2003b)

Setting	Reference
	Coronnello et al. (2007)
Stocks: United Kingdom (UK)	Coelho et al. (2007) Coronnello et al. (2005)
Stocks: Brazil	Tabak, Serra and Cajueiro (2010b)
Stocks: Germany	Birch, Pantelous and Soramäki (2016)
Stocks: Korea	Jung et al. (2006)
Stocks: Greece	Garas and Argyrakis (2007)
Stocks: Vietnam	Nguyen, Nguyen and Nguyen (2019)
Stocks: Italy	Coletti (2016)
Global Equity Markets	Bonanno, Vandewalle and Mantegna (2000) Roy and Sarkar (2011) Aslam et al. (2020)
Fixed Income	Lucey (2010) Dias (2012)
Global Listed Real Estate	Wang and Xie (2015)
Currency Markets	Kwapien et al. (2009) McDonald et al. (2005) Jang, Lee and Chang (2011) Wang et al. (2012) Keskin, Deviren and Kocakaplan (2011) Wang and Xie (2016) Wang et al. (2013)
Commodities	Sieczka and Hołyst (2009) Ji and Fan (2016)

Setting	Reference
	Tabak, Serra and Cajueiro (2010a)
Hedge Funds	Miceli and Susinno (2004)
Cryptocurrencies	Briola and Aste (2022) Nguyen et al. (2022) Song, Chang and Song (2019) Giudici and Polinesi (2021) Hong and Yoon (2022)
Asset Classes	Výrost, Lyócsa and Baumöhl (2019)
Portfolio Selection	Peralta and Zareei (2016) Pozzi, Di Matteo and Aste (2013) López de Prado (2016)

In their aptly titled paper "A Review Of Two Decades Of Correlations, Hierarchies, Networks And Clustering In Financial Markets", Marti et al. (2021) provide a comprehensive review of the various settings, metrics, advances, and problems with MSTs.

2.11.2 MSTs in the South African setting

We now discuss applications of MSTs in settings that include South African assets. The South African Rand has been featured in analyses of global currencies markets, such as Kwapien et al. (2009), Jang, Lee and Chang (2011), Keskin, Deviren and Kocakaplan (2011), Wang et al. (2012), Wang et al. (2013), and Wang and Xie (2016), while South African stock market indices have featured in analysis by Bonanno, Vandewalle and Mantegna (2000), and Aslam et al. (2020).

A long-term analysis of stocks on the South African stock market, the JSE, was performed by Gopi (2008), Gopi (2010), Gopi (2012b), and Gopi (2014). The author found that the grouping of shares on the MST did somewhat agree with the Industry Classification Benchmark (ICB) economic sector classification. However, the overlap was not perfect, especially with the so-called *Rand Hedge* shares. These shares typically emanate from the Financial or Industrial ICB sectors and derive a large portion of their revenue from offshore markets. Therefore, the earnings of these companies tend to benefit from a weakening in the Rand Dollar exchange rate and consequently, during these periods, market participants tend to bid up the prices of these shares (we discuss this phenomenon in more detail in Section 4.2.2). Particular attention was also paid to enhancing the visual nature of the tree by allowing the edges in the tree to vary by colour and size. This was used to highlight factor exposures or the correlation/distance between the two adjacent shares on an edge/link. Furthermore, the vertices/nodes in the tree were varied in colour and size to highlight various attributes. For example, the size of the vertex was used to indicate the market capitalisation of a share, or the weight of a share in a portfolio. This can be useful in identifying portions of a market or a portfolio that are under-diversified. The size and colour of the vertices were also used to highlight the exposure of stocks to particular economic factors (such as commodity prices, interest rates, or the Rand Dollar exchange rate), showing the overlap between the economic factor exposures and the positioning of shares on the tree. This same technique was also used to show the quantitative style characteristics of stocks such as momentum, value, growth, and growth at a reasonable price (GARP). This allowed the author to highlight pockets of style opportunities on the tree.

Majapa and Gossel (2016) performed a long-term analysis of the South African stock market, and also found substantial clustering and homogeneity among stocks. Furthermore, the authors performed a sub-sample analysis and found that the tree shrank before and during the 2008 financial crisis, and slowly expanded afterwards. Mbatha and Alovokpinhou (2022) investigated the topology of the South African stock market network during the COVID-19 hard lockdown[2]. The results showed an expansion of MST during the (strictest) level 5 lockdown and shrinkage of the MST after the level 5 lockdown. After the level 5 lockdown period, stocks in the Health Care Equipment & Services sector formed a small cluster that did not exist before the lockdown period. Both papers, therefore, highlighted the dynamic nature of the MST and the changing structure of the South African stock market.

2.11.3 An Example of an MST in the South African Setting

We return to the simple ten stock correlation matrix from Table 2.6 and its associated distance matrix from Table 2.7. The MST filtered correlation matrix for this example is shown in Table 2.10, and in Figure 2.11 one can see the MST using the circular layout (on the left-hand side), while the MST using the Fruchterman and Reingold (1991) force-directed layout (which was discussed in Section 2.6) can be seen on the right.

[2] A lockdown was a set of measures that were aimed at reducing the transmission of COVID-19 and was typically mandatory. These measures included stay-at-home orders, curfews, quarantines, and other social restrictions.

Table 2.10 Ten Stock Correlation Matrix: MST Filter

	AGL	BHG	GFI	ANG	FSR	SBK	ABG	MRP	TFG	TRU
AGL	1.00	0.63					-0.07			
BHG	0.63	1.00		0.00						
GFI			1.00	0.78						
ANG		0.00	0.78	1.00						
FSR					1.00	0.64				
SBK					0.64	1.00	0.72			
ABG	-0.07					0.72	1.00			0.16
MRP								1.00	0.34	
TFG								0.34	1.00	0.33
TRU							0.16		0.33	1.00

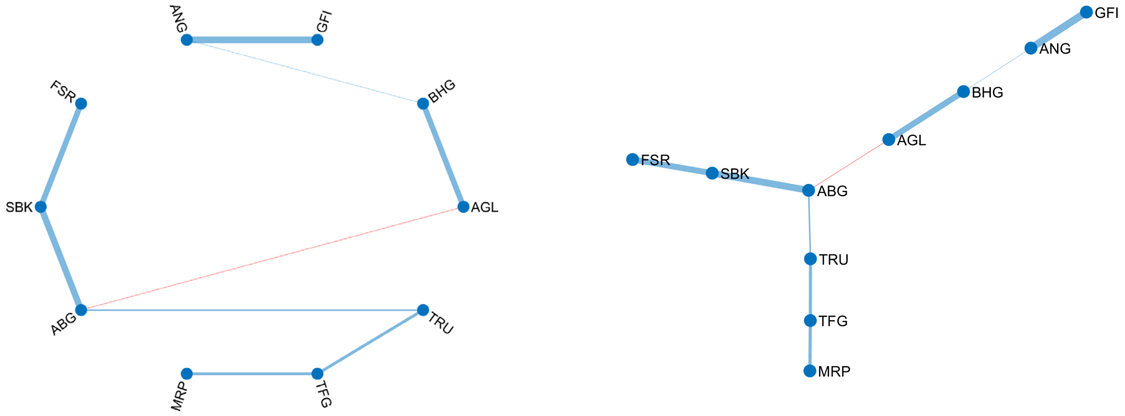


Figure 2.11 Ten Stock Correlation Matrix: MST Filtered - Circular vs Force-Directed Layout

From Figure 2.11 it is apparent that the force-directed layout does a better job of representing the relationships in this example (as it uses the underlying data in the network to determine the layout), than the arbitrary circular layout.

2.11.4 Linking MSTs to Hierarchical Cluster Analysis

In his seminal paper, Mantegna (1999), highlighted an important property of MSTs – the fact that the filtered distance matrix from an MST satisfies the properties of a *subdominant ultrametric distance*. An *ultrametric distance,* $u(p,q)$, must satisfy the following axioms (Birch, 2016):

$$\begin{array}{rl} I. & u(p,q) \geq 0 \ \ \textit{and}\ u(p,q) = 0 \ \textit{if and only if}\ p = q. \\ II. & u(p,q) = u(q,p). \\ III. & u(p,q) \leq max[u(p,r), u(r,q)]. \end{array} \qquad \textbf{(4)}$$

Condition *III* is a stronger version of the *triangle* from Equation (2) and is known as the *ultrametric inequality*. A *subdominant* ultrametric is a unique ultrametric distance that satisfies the above axioms in addition to $u(p,q) \leq d(q,p)$.

While seemingly esoteric, ultrametricity is a useful concept because it is directly linked to the concept of hierarchy or taxonomy (Onnela, 2002). In the case of the MST, the fact that it satisfies the properties of a subdominant ultrametric distance implies that the MST has an equivalent hierarchical representation or taxonomy. Even though Mantegna (1999) was the first author to highlight the link between the MST and hierarchical clustering in a stock market setting, it was Gower and Ross (1969) who provided the first published evidence of this link (in a generic setting), showing that an MST has all the information required to construct a specific hierarchical representation that can be created by a hierarchical clustering technique called Single Linkage Cluster Analysis (SLCA).

As alluded to in Section 2.9, King (1966), put forward the idea that although stock markets can be viewed through one lens as a high dimensional setting, one can use techniques, such as cluster analysis, to reduce dimensionality by grouping together stocks that exhibit similarity in selected features. In a network setting, clustering can be used to reveal communities, or clusters, of entities in any system. Entities belonging to the same community share more information (i.e. they are highly similar) than entities belonging to different communities (i.e. they are less dissimilar). In hierarchical clustering, entities are sharing information according to the communities that they belong to, and communities are organised in a nested structure or taxonomy.

This is intuitive in stock markets, which already have a hierarchical structure as defined by companies such as FTSE (who created the Industry Classification Benchmark or ICB), and MSCI and Standard & Poor's (who created the Global Industry Classification Standard or GICS). The ICB classifies companies into the following hierarchy (from the top down): industries, super-sectors, sectors, and subsectors. Figure 2.12 shows an example (Vass, 2019) of the hierarchical structure (or taxonomy) of the Consumer Staples Industry. For example, one can see that the Sugar subsector is contained in the Food Producers sector, which is contained in the Food, Beverage, and Tobacco super-sector which is finally contained in the Consumer Staples Industry. However, the construction of these structures was typically performed using qualitative information (e.g. the business segment that companies operate in), or fundamental accounting information (e.g. where they derive their revenue from), and may not mirror the way that stock prices actually move together. To this end, it is useful to have a technique that can determine such a hierarchical structure, without using any pre-specified qualitative or accounting information (or to use modern machine learning jargon, the process would be unsupervised).

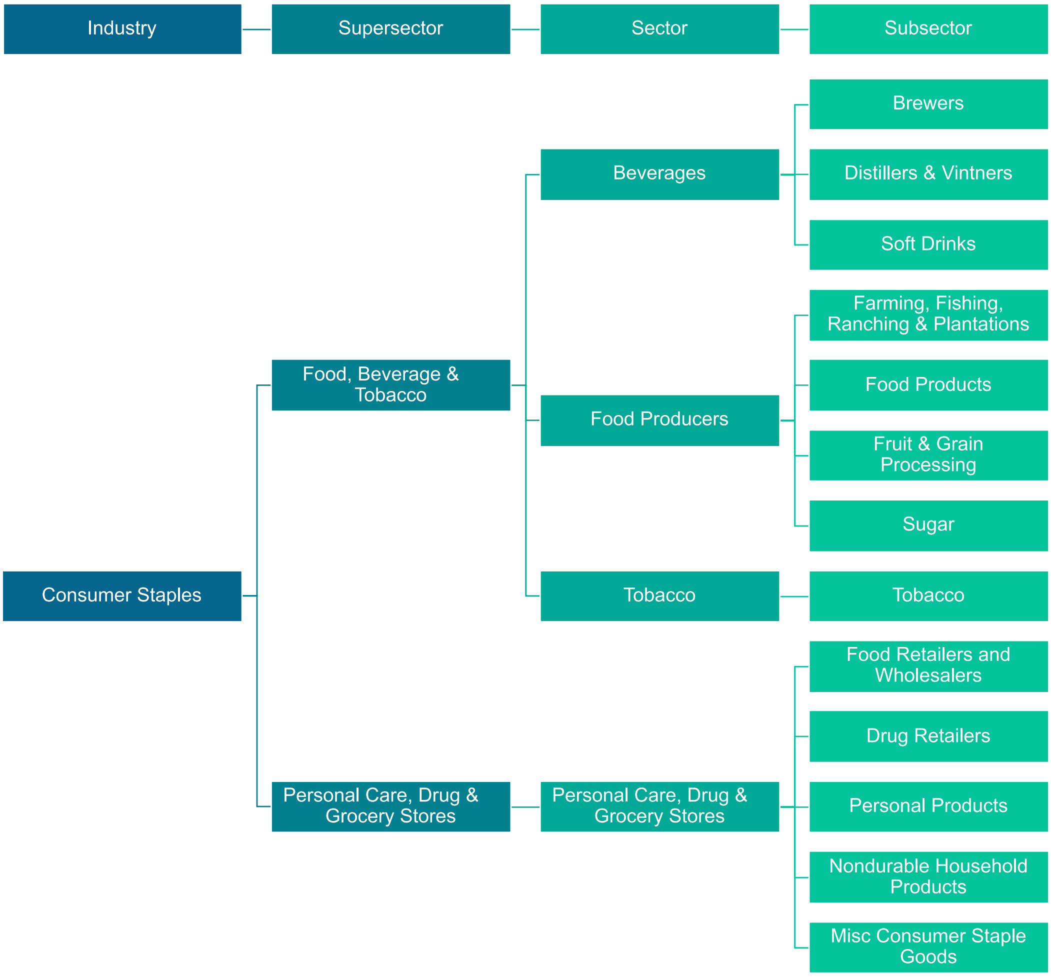


Figure 2.12 Example of the ICB Taxonomy for the Consumer Staples Industry

The most popular method of hierarchical clustering is referred to as *agglomerative clustering* in that each object starts as an individual cluster, and at each step, the closest pair of clusters is merged (or agglomerated) to form a new cluster. This process continues until all objects reside in a single cluster. The example below from Rhys (2020) illustrates this process well. The clusters that are closest to each other at each iteration are merged, with ellipses indicating the formation of clusters at each iteration, going from top left to bottom right.

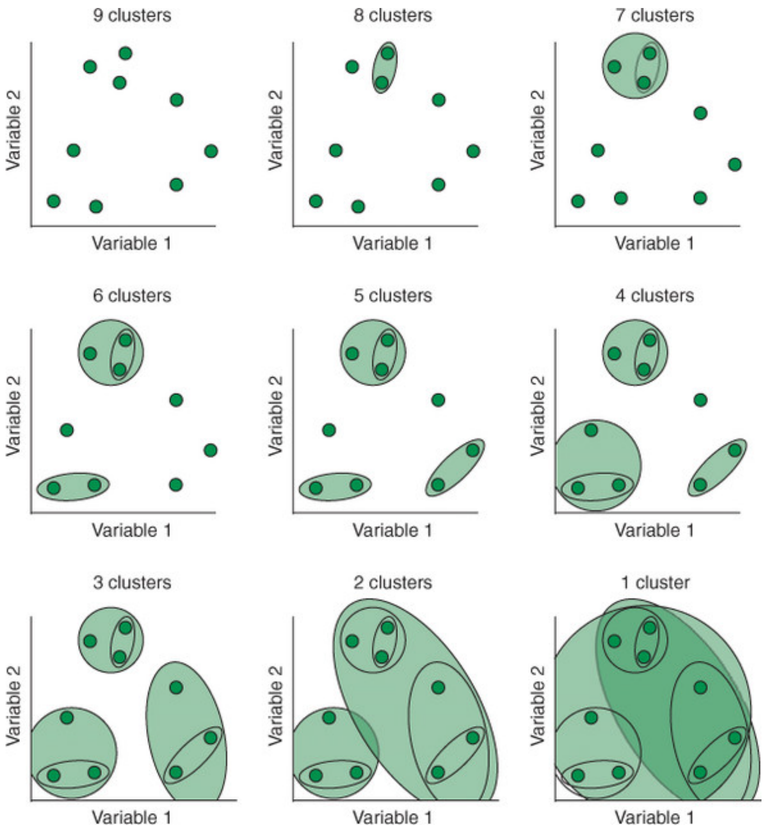


Figure 2.13 Illustration of Agglomerative Clustering from Rhys (2020)

One of the requirements for a hierarchical clustering algorithm is a method to determine the distance between clusters so the closest two clusters can be merged. Calculating the distance between two individual objects is relatively simple, but how does one calculate the distance between two clusters that may contain multiple objects? This requires the use of a *linkage* function. Each type of linkage function uses a different criterion to define the notion of closeness between clusters. These linkage functions are summarised in Table 2.11, while Figure 2.14 from Rhys (2020) depicts the various linkages methods visually. As mentioned previously, it was Gower and Ross (1969) who showed the link between the MST and SLCA, i.e. hierarchical clustering using the single linkage function.

Table 2.11 Summary of Popular Linkage Methods

Method	Description
Centroid linkage	Minimises the distance between each cluster's centroid.
Single linkage (SLCA)	Minimises the smallest distance between objects in the two clusters (also referred to as *nearest neighbour*).
Complete linkage	Minimises the largest distance between objects in the two clusters (also referred to as *farthest neighbour*).
Average linkage (ALCA)	Minimises the average distance between all pairs of objects in the two clusters.
Ward's method	Minimises the incremental variance, i.e. the increase in the total within-cluster variance as a result of merging two clusters.

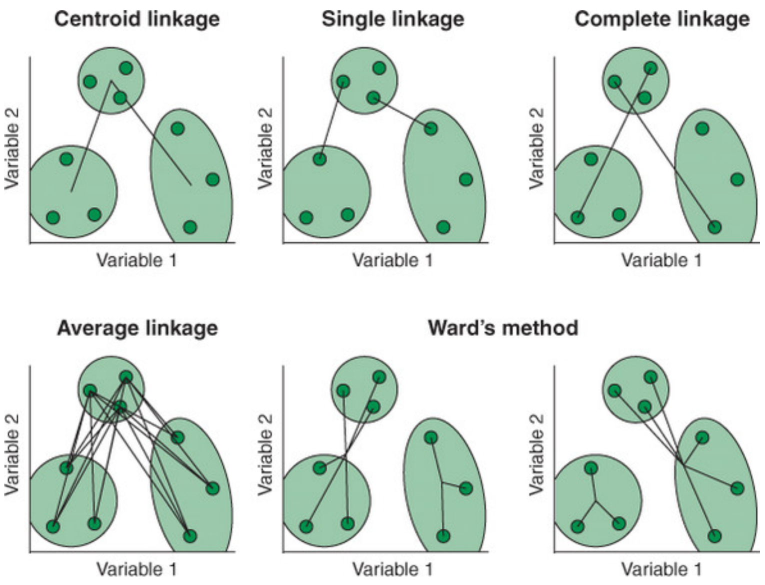


Figure 2.14 Various Linkage Methods from Rhys (2020)

The result of hierarchical clustering is usually depicted visually using a dendrogram. This is a tree-like structure that shows how clusters merge. The ICB sector classification example, in Figure 2.12 is an example of a dendrogram. In Figure 2.15 we show a dendrogram of the small sample correlation matrix from Table 2.6 using SLCA.

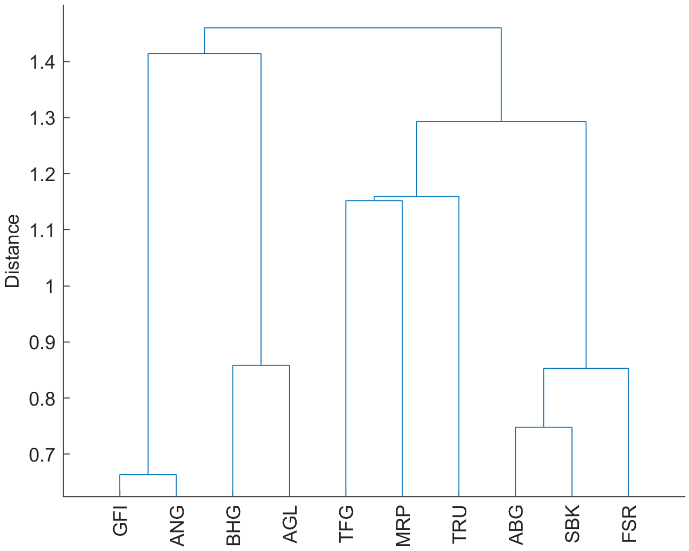


Figure 2.15 Ten Stock Correlation Matrix: Dendrogram using SLCA

It is interesting to note how even in this simple example, the clustering of the stocks aligns closely with the ICB sector classification of the shares - gold mining (GFI, and ANG), general mining (AGL, and BHG), retail (MRP, TFG, and TRU), and banking (FSR, SBK, and ABG). It is also interesting to note how GFI, ANG, BHG, and AGL, which form part of the Basic Resources ICB super-sector, also form a cluster together at a high level on the dendrogram.

2.11.5 Analysing the topology of an MST Using Network Metrics

As mentioned in Section 2.5, networks can be analysed using various metrics to summarise important properties about their structure. Metrics such as tree length, degree, or centrality can be used to glean insights from an MST. For example, Onnela, Chakraborti, Kaski and Kertész (2003) showed how the tree length of an MST of stocks on the NYSE, shortens during times of market crashes, while Majapa and Gossel (2016) show a similar result for the South African stock market. Tabak, Serra and Cajueiro (2010a) used network metrics to show that agricultural commodities are very important in a commodity network, followed by metals and energy. Di Matteo, Pozzi and Aste (2010) introduced measures, which differentiate between well-connected stocks and central stocks. They showed that stocks could be well-connected in the network, but at the same time they could be peripheral, that stocks could be poorly connected but also central, and that stocks could be both poorly connected and peripheral.

2.11.6 A Summary of MSTs

The MST offers a variety of benefits when being used to analyse financial markets. It has been widely researched, and therefore there are many theoretical results, and analyses relating to it. It offers a customisable visual representation, whereby one can highlight valuable information, or metadata, on the MST. A cluster structure emerges on the MST without the use of any prior information, and one can observe how the various clusters are connected. The hierarchical clustering representation also provides intuitive information, which aligns closely with the standard ways of classifying stocks using qualitative information. Furthermore, network metrics can be used to analyse MSTs and they can be calculated dynamically over time.

However, MSTs are not without their disadvantages. Firstly, the clusters that form, as well as the edges of the MST, can be unstable. Small changes in the input data, noise, and even various choices in the pre-processing of data can have a large impact on the outcome of the analysis. The instability of the MST is thought to be related to the sensitivity of SLCA to outliers, as well as the chaining effect in SLCA in which clusters that are produced are elongated and difficult to interpret. Due to these problems, Marti et al. (2021) emphasise that although there is broad agreement in the conclusions of the various empirical studies, there are also various contradictory claims among them. The fact that there are no widely utilised benchmarks to compare the various methods of implementation is also problematic.

Finally, with many types of cluster analysis (and the MST/SLCA is no different), the user must specify up front how many clusters the data should be grouped into. This is difficult to do unless the user has prior knowledge of the data set and the expected number of clusters in the data. While there are many different methods of determining the optimal number of clusters in a dataset, there is no consensus on the best method. Often the best method depends on the dataset being analysed. Ullmann, Hennig and Boulesteix (2022) and Akhanli and Hennig (2020) contain interesting comments in this regard. The network filtering approach that is outlined in the next section offers a solution to this problem (i.e. it has a built-in community detection algorithm), as well as embedding additional information in the network, and therefore potentially reducing the sensitivity of the filtered network to noise and outliers.

2.12 Planar Maximally Filtered Graph (PMFG)

2.12.1 An Introduction to PMFGs

Although the MST has been shown to have appealing properties, especially in a financial market setting, there were certain concerns relating to the stability and robustness of the MST. Furthermore, the filtering of a network down to a tree structure (i.e. an undirected graph in which any two vertices are connected by exactly one unique path), may perhaps be too strict a constraint, resulting in the loss of potentially useful information.

So although the MST is not as strict as the Asset Graph or Threshold Network, and it does allow some global correlations to filter into the network, it may be overly punitive as a network filter. Aste, Shaw and Matteo (2010) offer the following example:

> In particular, the condition that the extracted network should be a tree is a strong constraint. Indeed, let us, for instance, consider the case where three companies are involved in similar activities and therefore have strongly correlated behaviors in the dynamics of their stock prices. In the MST construction, unavoidably only two of these companies can be directly connected with an edge in the filtered graph because the connection with an extra edge of the third company will form a triangular cycle, a 3-clique, which is not allowed in a tree. Ideally, one would like to be able to maintain the same powerful filtering properties of the MST but also allow the presence of extra links, cycles and cliques in a controlled manner.

In graph theory, a *clique* is a group of vertices in which every vertex is connected (via undirected edges) with every other vertex in the clique (or more mathematically, a clique is a graph, or a sub-graph, such that every two distinct vertices are adjacent). Typically, a clique looks like a triangle or a structure composed of triangles. Figure 2.16 shows examples of 2-cliques (top-left), 3-cliques (top-right), 4-cliques (bottom-left), and 5-cliques (bottom-right). Tumminello et al. (2005) introduced a network filter that allows 3 and 4-cliques, thereby enriching the information retained in the filtered network. They named this method a Planar Maximally Filtered Graph (PMFG). In graph theory, a planar graph is a graph that can be embedded in the plane, i.e. it can be drawn on a flat surface such that no edges cross each other.

The PMFG can be thought of as a less constrained version of the MST, allowing the filtered network to retain more links. In fact, Tumminello et al. (2005) prove the MST is always a sub-graph of the PMFG. While the MST has $N-1$ edges, the PMFG has $3 \times (N-2)$ edges, and the number of 3-cliques in the PMFG will be more than $2 \times (N-4)$. The construction algorithm and the topological constraints on the PMFG force each element to participate in at least one 3-clique. While an MST can be constructed from a fully connected network using the methods of Kruskal (1956) or Prim (1957), the PMFG can be constructed using the algorithm described in Briola and Aste (2022).

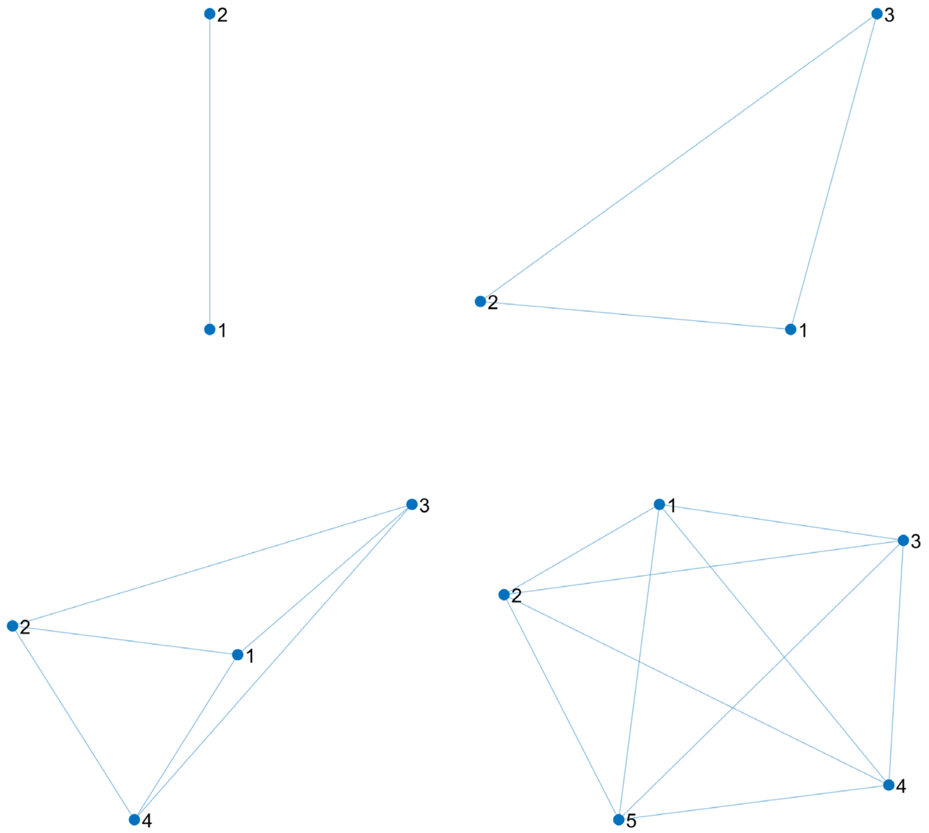


Figure 2.16 Examples of 2-Cliques, 3-Cliques, 4-Cliques, and 5-Cliques

(Clockwise from the top-left)

Even though the PMFG was introduced in 2005, it was initially not applied as widely to financial markets as the MST. However, in recent years the application of this filtering tool has increased. Table 2.12 contains a sample of various studies that have used the PMFG to analyse financial markets. Note that there is some overlap between the studies in the table below and the studies on the financial market applications of the MST in Table 2.9.

Table 2.12 Studies of PMFGs Applied to Financial Markets

Setting	**Reference**
Stocks: New York Stock Exchange (NYSE)	Tumminello et al. (2005), Aste, Shaw and Matteo (2010) Kenett et al. (2010) Yan, Xie and Wang (2015) Zhao, Li and Cai (2016)

Setting	Reference
	Musmeci, Aste and Matteo (2015) Kukreti et al. (2020) Coronnello et al. (2007)
Stocks: United Kingdom (UK)	Coronnello et al. (2005)
Stocks: Australia	Yan et al. (2020) Pozzi, Di Matteo and Aste (2013)
Stocks: China	Guo et al. (2022)
Stocks: Germany	Birch, Pantelous and Soramäki (2016)
Global Equity Markets	Wen, Yang and Zhou (2019) Song et al. (2011) Eryiğit and Eryiğit (2009)
Fixed Income	Aste et al. (2005)
Currency Markets	Wang and Xie (2016)
Cryptocurrencies	Briola and Aste (2022) Giudici and Polinesi (2021) Hong and Yoon (2022)
Asset Classes	Výrost, Lyócsa and Baumöhl (2019)
Portfolio Selection	Pozzi, Di Matteo and Aste (2013)

2.12.2 An Example of a PMFG in the South African Setting

We return to the simple ten stock correlation matrix from Table 2.6 and its associated distance matrix from Table 2.7. The PMFG filtered correlation matrix for this example is shown in Table 2.13. One can see that a larger number of entries in the matrix was retained here as compared to the correlation matrix that was filtered using the MST (see Table 2.10).

On the right-hand side of Figure 2.17, one can see the PMFG filtered network using the Fruchterman and Reingold (1991) force-directed layout. The MST, which is a sub-graph in the PMFG, is also visible on the left of Figure 2.17 and is also depicted in the PMFG (on the right) by highlighting the edges that are common between the MST and PMFG in red. One can see that the PMFG is a more complex version of the MST, but hopefully, the added complexity provides additional benefits in the form of retaining additional relevant information.

Table 2.13 Ten Stock Correlation Matrix: PMFG Filter

	AGL	BHG	GFI	ANG	FSR	SBK	ABG	MRP	TFG	TRU
AGL	1.00	0.63	-0.13	-0.10			-0.07		-0.19	
BHG	0.63	1.00	-0.16	0.00			-0.27	-0.22	-0.12	
GFI	-0.13	-0.16	1.00	0.78						
ANG	-0.10	0.00	0.78	1.00					-0.38	
FSR					1.00	0.64	0.63		0.11	0.10
SBK					0.64	1.00	0.72			0.11
ABG	-0.07	-0.27			0.63	0.72	1.00	0.12	0.06	0.16
MRP		-0.22					0.12	1.00	0.34	0.29
TFG	-0.19	-0.12		-0.38	0.11		0.06	0.34	1.00	0.33
TRU					0.10	0.11	0.16	0.29	0.33	1.00

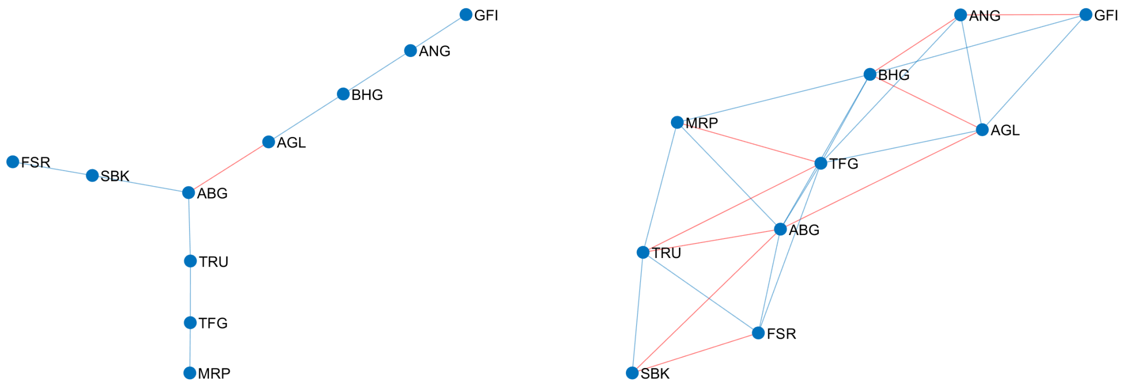


Figure 2.17 Ten Stock Correlation Matrix: MST (left) vs PMFG (right)

2.12.3 Linking PMFGs to Hierarchical Cluster Analysis

As highlighted in the prior sections, the MST shares a deep connection to a hierarchical clustering method, the SLCA. Interestingly, Song, Di Matteo and Aste (2011) explored a method to characterise the hierarchical structure of a PMFG and proposed a framework to define communities on these graphs and to extract their hierarchical structure.

As mentioned previously, the PMFG has a property of typically being constructed of 3-cliques, and these building blocks also define a class of larger sub-graphs, that Song, Di Matteo and Aste (2011) name *bubbles*. They showed that a hierarchical relationship emerges naturally in a planar graph, and this relationship is directly associated with the system of 3-cliques and the bubble structure. Song, Di Matteo and Aste (2012) formally introduce an algorithm that exploits this property and named this technique the Directed Bubble Hierarchical Tree (DBHT). The DBHT exploits the distinction between separating and non-separating 3-cliques to identify clustering partitions of all the nodes in the PMFG. This structure can then be depicted in the traditional dendrogram visualisation.

An important benefit of the DBHT algorithm is that the user of such a technique does not have to specify a priori how many clusters to group the data into. This is in contrast to typical hierarchical clustering techniques, in which objects/clusters are iteratively agglomerated based on the similarity of the clusters into a hierarchical structure, and the user must then decide how many clusters to split the data into. Alternatively, one must use cluster validation techniques to determine the optimal number of clusters. In the case of the DBHT, the objects are split upfront into the optimal number of clusters, and the hierarchy is then inferred from the inter-cluster and intra-cluster similarities. The algorithm to construct the DBHT can be found in Song, Di Matteo and Aste (2012), and a modified version written using the MATLAB programming language can be found in Aste (2014).

Note that the DBHT technique is not the first to have automatic cluster/community detection built in. Omran, Salman and Engelbrecht (2005) discuss a variety of clustering techniques that do not require an a priori specification of the number of clusters. In the econophysics setting, Giada and Marsili (2002) propose an unsupervised, parameter-free approach to finding clusters, based on the maximum likelihood principle. The authors applied this technique to stocks on the NYSE, while authors such as Mbambiso (2008), and Hendricks, Wilcox and Gebbie (2016) have applied the same technique to stocks on the JSE. However, a comparison of this technique to the DBHT is beyond the scope of this dissertation.

The DBHT technique has not been widely applied to financial market data. Musmeci, Aste and Matteo (2015) used the DBHT technique for the first time in financial markets and compared this technique to other popular clustering algorithms. Using the ICB classification as a benchmark, they showed that the DBHT could outperform other methods, being able to retrieve more information with fewer clusters. Raffinot (2017) utilised the DBHT to construct a hierarchical clustering-based asset allocation method. He found that the DBHT-based portfolios attained marginally superior risk-adjusted returns across a variety of settings.

2.12.4 PMFGs and DBHTs in the South African setting

The application of the PMFG and the DBHT to South African instruments has been limited. Song et al. (2011) analysed global equity markets and the South African Industrial 25 index (INDI25) formed part of that analysis, while Wang and Xie (2016) analysed the global currency market, which included the South African Rand. To the best of our knowledge, PMFGs and DBHTs have not been used to analyse the South African stock market.

2.12.5 A Summary of PMFGs and DBHTs

The PMFG was introduced by Tumminello et al. (2005) as a technique that can maintain the powerful filtering properties of the MST but also encodes a larger amount of information into its internal structure, in a controlled manner. This should lead to a more informative graph, but at the expense of a more complex graphical depiction. The PMFG does, however, have the MST embedded in it. Furthermore, the PMFG also has a hierarchical representation (similar to the MST and SLCA) called the DBHT. The DBHT has the benefit of having automatic cluster/community detection built-in, and therefore the user does not need to specify the required number of clusters/communities a priori. To the best of our knowledge, these techniques have not been used to analyse the South African stock market.

2.13 A Detailed Examination of Network Topology Measures

Many real-life networks are large, contain many vertices, and have complicated structures. However, they also have properties, which can be extracted to glean various insights from the network. As discussed in Section 2.5, the *way that the vertices and edges in a network are arranged is referred to as the topology of the network*. These topological properties can refer to the overall network, or specific vertices or edges. In this section, we introduce various metrics that can be used to describe a network's topology (while leaving the detailed formulae for Section 3.2.11). It should be noted that this list is not exhaustive, but is a list of commonly utilised metrics for analysing networks in financial markets. Newman (2010), and Fornito, Zalesky and Bullmore (2016) contain general introductions to a variety of network measures and metrics, while Pozzi, Di Matteo and Aste (2013), and Samal et al. (2021) contain information relating to network metrics that have been applied to financial markets.

2.13.1 Overall Topology

The *normalised tree length (NTL)* was introduced in Onnela, Chakraborti, Kaski, Kertesz, et al. (2003b), as a method to calculate the overall strength of the relationships amongst the entities in the network. A larger NTL would indicate larger distances between vertices and therefore weaker relationships in the network. A smaller NTL would indicate smaller distances between vertices and therefore stronger relationships in the network. Note that since we will be calculating distances from a correlation matrix using Equation (3), the NTL will be inversely related to the correlation matrix.

2.13.2 Measures of Centrality

Researchers analysing a graph, or a network, often want to identify the roles that different vertices/nodes play, determining which vertices are important or central, and which are peripheral. However, there are many possible definitions of importance, and correspondingly many centrality measures for networks. The centrality metric that one chooses, depends on what the researcher deems to be important.

The simplest centrality metric is the *degree* or *degree centrality* of a vertex, and it is measured (in an unweighted network) by counting the number of edges that are connected to that vertex. The logic is easy to understand – important vertices have many connections. A variation of the vertex degree is the *normalised degree centrality* in which the degree of each vertex is normalised by taking into account the number of edges in the graph. This allows comparisons across various types of networks. Although simple, degree centrality can be a highly effective measure of the importance of a vertex. For example, in many social networks, people with more connections tend to be more important. However, one limitation of degree centrality is that all connections are treated equally.

Eigenvector centrality can be thought of as an extension of degree centrality, in which we take into account not only how many neighbours a vertex has, but also how central or important those neighbours themselves are. Therefore, it caters not only for the quantity of connections, but also for the quality of the connections. In a social network, connections to people who are themselves influential will offer a person more influence than connections to less influential people. Interestingly, a variant of eigenvector centrality is PageRank centrality (Brin and Page, 1998) which is one of the algorithms used by Google to rank websites in the results of their search engine.

Closeness centrality is a metric that is based on the concept of a network path. A path in a network is the sequence of vertices that is travelled along by following edges from one vertex to another. Closeness centrality measures the importance of a vertex by first calculating the shortest paths from that vertex to all other vertices, and then calculating the average of those path lengths. The closer a vertex is to the other vertices (on average), the more important it is.

Related to closeness centrality, is the concept of *eccentricity centrality*. Similarly, one calculates the shortest paths from a vertex to all other vertices, but instead of finding the average path length, one uses the maximum, or longest path. The closer a vertex is to the other vertices (based on the longest path that it has to other vertices), the more important it is.

Betweenness centrality is also a metric that is also built upon the notion of a network path. It defines importance as being able to connect vertices to each other. This is done by measuring the number of times a vertex appears in a path between other vertices. The more often a vertex appears in paths, the more important it is in the network. Betweenness centrality can be thought of as a measure of the control that a vertex exerts over the flow of information between other vertices in the network. In the context of a social network, a vertex with high betweenness centrality will not necessarily exert influence by being highly connected, but by connecting other vertices to each other.

Figure 2.18 is an example from Pike (2015) that shows the difference between closeness, betweenness, and eigenvector centrality, as highlighted by the vertices coloured red. One can see that closeness centrality favours vertices that are close to other groups of vertices, while betweenness centrality favours vertices that connect other vertices. Eigenvector centrality favours vertices that are connected to other important vertices.

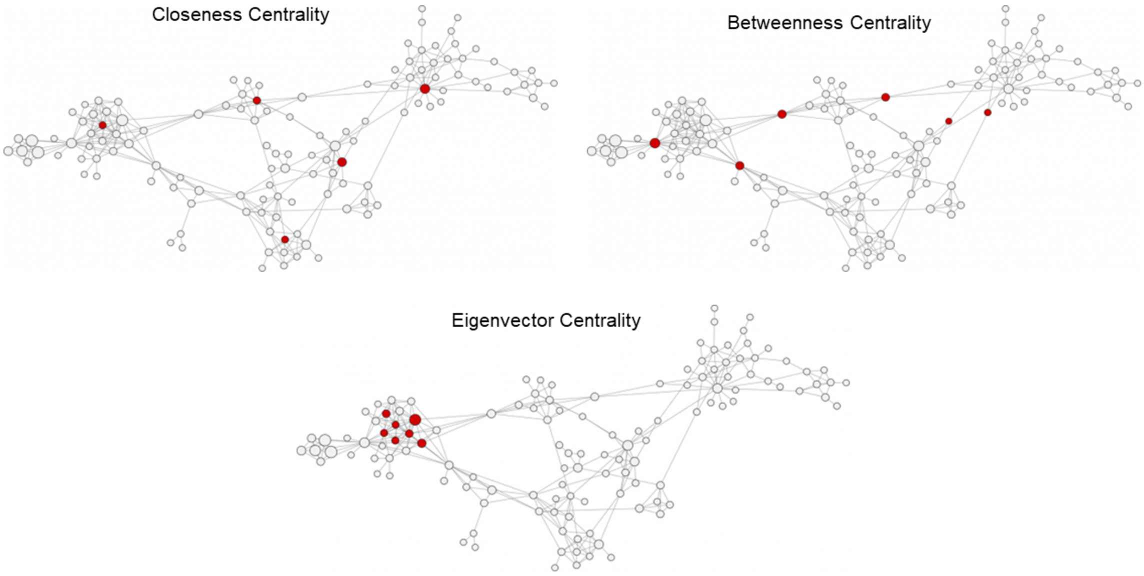


Figure 2.18 Examples of Three Types of Centrality Metrics (Pike, 2015)

Several centrality measures have been proposed in the literature (many of which have not been discussed here), and given that they can reflect different criteria for assessing importance, it is not unusual that a vertex can be considered central for one measure and peripheral for another. Pozzi, Di Matteo and Aste (2013) therefore proposed the concept of *hybrid centrality*. They introduced two measures X and Y, whereby X is an average of the ranks of degree centrality and betweenness centrality (both weighted and unweighted metrics, and so four metrics in total), while Y is an average of the ranks of eccentricity centrality, closeness centrality, and eigenvector centrality (all weighted and unweighted, and so six metrics in total). Given these two measures, the authors assert the following guidelines for assessing the importance of each vertex:

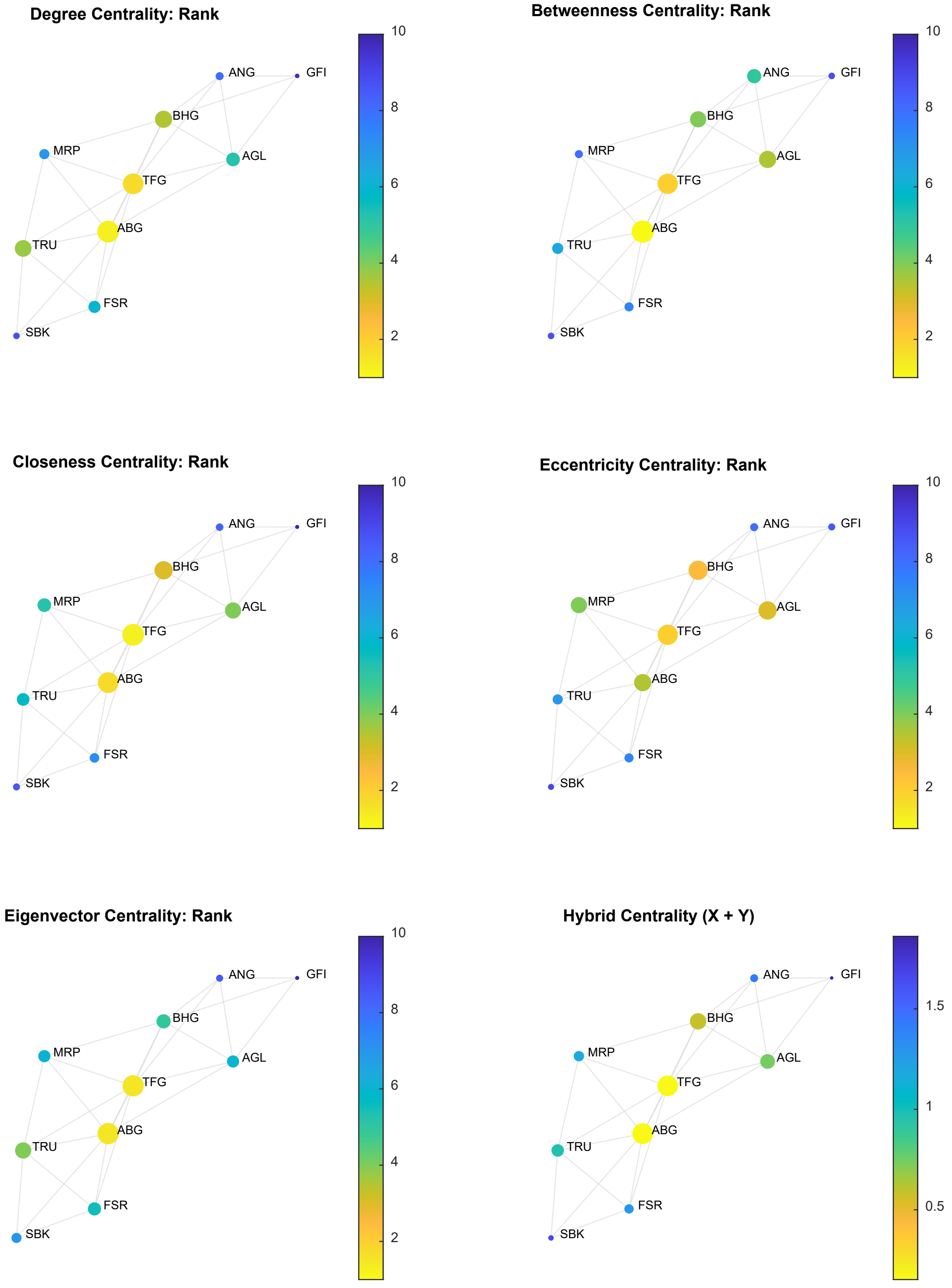
- **Small X, small Y** – a highly connected vertex, connected to other highly connected vertices.
- **Small X, large Y** – a highly connected vertex, connected to scarcely connected vertices.
- **Large X, small Y** – a scarcely connected vertex, connected to highly connected vertices.
- **Large X, large Y** – a scarcely connected vertex, connected to scarcely connected vertices.

Therefore, the quantity (X + Y) is small for central vertices and large for peripheral vertices, while the quantity (X – Y) is large if the vertex has few important connections and it is small if it has many unimportant connections. In this dissertation, we utilise the hybrid centrality quantity (X + Y) when we analyse networks of the South African stock market.

In Figure 2.19 we plot the ranks of the five centrality measures of interest (degree, betweenness, closeness, eccentricity, and eigenvector centrality), as well as the hybrid centrality (X+Y) quantity from Pozzi, Di Matteo and Aste (2013), for the PMFG of the small sample correlation matrix from Table 2.6 and its associated distance matrix from Table 2.7. Note that these metrics are ranked in ascending order, so stocks which rank better in terms of these centrality metrics (i.e. they would be considered to be more central) would have lower ranks (closer to a value of one). The vertices have been enhanced in terms of colour and size, with better ranked stocks having a larger vertex, and being shaded closer to a yellow colour. Stocks which rank poorer in terms of these metrics (i.e. they are peripheral) would have smaller vertices and vertices that are shaded closer to a blue colour. Figure 2.20 shows the ranks of the various centrality measures in a bar chart, while Figure 2.21 shows a bar chart of the final (X + Y) hybrid centrality metric.

In this simple network, one can see that TFG, and ABG consistently rank highly (with lower values) across the various metrics. These stocks would be considered to be highly central in the network. Interestingly, the eccentricity centrality for ABG is relatively worse as compared to its other metrics (with AGL, and BHG having better outcomes for this metric). This highlights the motivation to use various centrality metrics and combine them into a hybrid centrality score, as done by Pozzi, Di Matteo and Aste (2013).

Of the other stocks, BHG, and AGL also rank highly. GFI, SBK, ANG, FSR, MRP, and TRU show more peripheral behaviour in the network.



Degree Centrality: Rank

Betweenness Centrality: Rank

Closeness Centrality: Rank

Eccentricity Centrality: Rank

Eigenvector Centrality: Rank

Hybrid Centrality (X + Y)

Figure 2.19 Ten Stock Correlation Matrix: PMFG Centrality Measures Overlaid onto the Graph

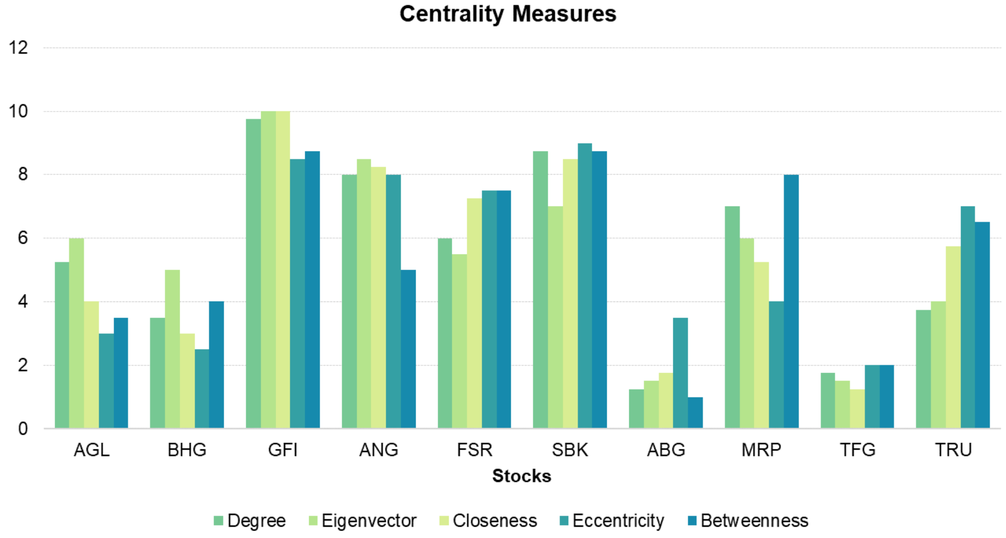


Figure 2.20 Ten Stock Correlation Matrix: PMFG Centrality Measures

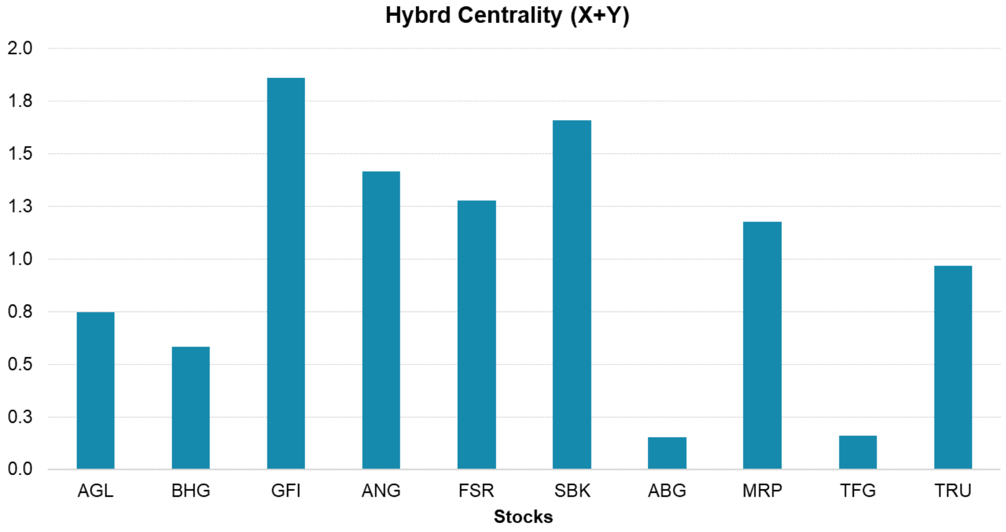


Figure 2.21 Ten Stock Correlation Matrix: PMFG Hybrid Centrality (X+Y)

2.14 Practitioner's Viewpoint

We conclude this section by addressing the relevance of this analysis, both network filtering as well as cluster analysis techniques, to financial market practitioners. As discussed in Section 1.1, network filtering tools are useful to prune noise in stock market networks, thereby allowing important macroscopic and mesoscopic structures to emerge. These filtered networks are also accompanied by a visual representation that allows the user to easily unearth meaningful information about the complex market dynamics and its emergent structure. Furthermore, one can extract useful metrics from such networks that describes their structure or topology, highlighting which shares are important or central in the network. The analysis of the temporal evolution of networks can also assist in understanding the underlying trends in the structure of the market.

On a more granular level, Marti et al. (2021) provided a detailed list of academic studies in which these techniques have been used. They broke down the use cases into four main groups: portfolio design, trading strategies, risk management, and financial policy making. In the following paragraphs, we highlight various papers that fall into the first three categories, which are the most relevant categories for investment professionals.

From a *portfolio design* point-of-view, clustering techniques are valuable tools to reduce the complexity of analysing many shares, into a smaller, more manageable, subset of clusters. One could even use shares that are within the same cluster as proxies for each other, if there is a constraint on investing in one of the shares. There have also been several studies in which portfolios have been constructed based on the results from network filters or cluster analysis. For instance, Pozzi, Di Matteo and Aste (2013) stated that one gets better diversification benefits by investing in shares that are on the periphery of a network, while López de Prado (2016) introduced a portfolio diversification technique called hierarchical risk parity in which one allocates an equal risk budget to hierarchical clusters.

In terms of *trading strategies*, one may find that when searching for mean-reverting trading opportunities such as pair trades, stocks that are found within the same cluster may be better candidates for such strategies Han, He and Toh (2022) and Sarmento and Horta (2020) provided strategies in which clustering techniques were used to construct profitable pair trades, while Qu et al. (2016) apply these techniques to mean-reverting strategies from long-short basket trades (where the baskets are based on statistical clusters).

And finally for the *risk management* category, monitoring the contribution to risk from clusters may provide a better understanding of the risk concentration in a portfolio (as opposed to using an economic sector classification). As done in Gopi (2012b), the ability to overlay portfolio weights, factor exposures, and risk contributions onto an MST or PMFG (by varying the size and colour of the vertices) also provides an intuitive method for portfolio managers to easily visualise the risks inherent in their funds. Seabrook, Caccioli and Aste (2021) utilised network filtering techniques in a stress testing framework, and lastly, Cook, Soramäki and Laubsch (2016) made use of network-based methods to visually identify systemic risks.

Chapter 3 Methodology, Modelling Considerations, and Data

In this section we discuss various modelling, pre-processing, and sample/feature selection techniques and considerations. The choices associated with many of these concepts often have a significant impact on the outcome of any analysis, so we discuss each of them in detail here, referring to published literature or practitioner insights to guide the choices that are made. Section 3.1 contains a general discussion regarding modelling considerations, while in Section 3.2 we discuss the actual data that was used and the methodology that we followed.

3.1 Modelling Considerations

3.1.1 Universe of Stocks, Thin-trading Concerns, and Frequency of Data

In this dissertation, the relationships among stocks listed on a South African stock exchange, the JSE, are analysed. The focus is on stocks that are listed on the main board of the JSE, and more specifically, stocks that form part of the main equity index, namely the All Share Index, or ALSI. Note that while the SWIX and Capped SWIX are also popular indices, all three indices are constructed using the same set of shares, with only a difference in the weighting scheme in each index. This was done to ensure that the analysis is relevant for practitioners, since shares that are not listed on the main board, or do not form part of the ALSI, are often not investable for many investment professionals. Furthermore, these shares do not trade regularly and when they do trade, they often trade in small lot sizes. Stocks that do not trade frequently or in a realistic lot size often seem to be less volatile than stocks that trade frequently (i.e. they have a lower volatility). However, this apparent lower volatility is misleading, as the lack of proper market liquidity tends to artificially bias down the volatility of these stocks making them seem less volatile than they are. Additionally, any relationships that are modelled with other stocks (e.g. calculating a correlation between these stocks) will tend to also show an artificially weaker relationship. This effect has been modelled and catered for when estimating stock market betas by using various techniques to adjust the beta estimates (see for example Bradfield, 2003).

However, applying these techniques to bivariate (and higher dimensional modelling) like correlation estimation is difficult. One can use techniques that adjust for the lack of synchronicity between stock prices (i.e. thinly traded stocks do not trade at the same time, which artificially lowers the strength of the correlation between them), such as those employed by Hayashi and Yoshida (2005), Clayton (2018), or Münnix, Schäfer and Guhr (2010). We choose to employ the simpler method of screening out stocks that trade infrequently or in lower volumes. The trade-off for the use of this method is that our universe of stocks becomes smaller, and we would potentially miss interesting outcomes that may have occurred by including these shares in the analysis.

Furthermore, the use of weekly data to estimate the cross-correlations amongst the shares would also lessen the thin-trading/asynchronous effect as compared to using daily data. This is because the effect of asynchronous prices is averaged out over the course of each week. This was the approach followed by Bonanno, Vandewalle and Mantegna (2000) when estimating correlations amongst global stock market indices that trade at different times, and is also the approach that is followed here (i.e. the use of weekly data as opposed to daily).

3.1.2 Dealing with Noisy Correlation Matrices

Given that correlation matrices are estimated from noisy time series data, and that typically the number of historical observations that are used in the estimation process is limited (as compared to the number of stocks that are involved in the estimation process), the estimate of the correlation matrix will include some amount of noise. This noise may negatively impact the quality of the analysis that is performed using the correlation matrix, and in extreme cases may render the analysis to be spurious. The goal of this section is to discuss various procedures for reducing the noise and enhancing any signal that is inherent in an empirical correlation matrix, prior to performing any network filtering techniques on it. These procedures have generally followed two paths, the *Random Matrix Theory* (RMT) path, and the *shrinkage* path.

Researchers in the econophysics field pioneered the application of the concept of RMT, which dates back to work by Marčenko and Pastur (1967) and Wigner (1955), to remove noise from correlation matrices. Marčenko and Pastur (1967) showed that the eigenvalues of a correlation matrix constructed from N completely random time series of length T has a specific distribution known as the Marčenko-Pastur distribution. In particular, the eigenvalues, λ_i, for these random matrices fall within the range of $\lambda_\pm = \left(1 \pm \sqrt{\frac{N}{T}}\right)^2$, which is also called the *noise band*. So, in theory, one can calculate the eigenvalues of an empirical correlation matrix and compare these to the Marčenko-Pastur distribution. If there are eigenvalues that do fall within the noise band, they are deemed to be noise, and should be filtered out or attenuated to reduce their impact on the overall correlation matrix. Typically, eigenvalues of the correlation matrix that are greater than the upper bound of the Marčenko-

Pastur distribution (λ_+), are treated as having violated the random matrix hypothesis and are therefore regarded as containing *true information*. From the equation, one can see that the upper bound of the noise band increases as the ratio of N to T gets bigger, i.e. as the number of stocks increases in comparison to the number of data points. Laloux et al. (1999) were amongst the first authors to use RMT techniques to analyse empirical correlation matrices on the NYSE, while Laloux et al. (2000) suggested a process in which one replaces the eigenvalues that fall within the noise band with a constant value, thereby removing the bias associated with them. Since those early applications, there have been improvements suggested in the literature, with a comprehensive review found in Bun, Bouchaud and Potters (2017).

Independently of the econophysics field, the application of the shrinkage path to financial markets was pioneered Olivier Ledoit and Michael Wolf. These techniques drew inspiration from Stein (1956), and James and Stein (1961), who proposed that when estimating a mean vector in a multivariate setting, a better estimator than the sample mean can be constructed by using a linear combination of the sample mean and a target vector, i.e. by shrinking the sample mean to a target vector. Ledoit and Wolf (2003) suggested that linearly blending the sample covariance (or correlation) matrix with a structured pre-defined target matrix such as the identity matrix (Ledoit and Wolf, 2003), one defined by a factor model (Ledoit and Wolf, 2004a), or one with a constant correlation (Ledoit and Wolf, 2004b), results in a better estimate of the covariance matrix, as compared to each matrix on its own. The target matrix generally requires a smaller number of parameters to be estimated (because it is highly structured) and therefore has little estimation error. However, this comes at the expense of having a high bias due to the simplistic nature of these matrices. This is in contrast to the sample covariance matrix which, is an unbiased estimator but contains a large number of free parameters. Therefore, optimally combining these two matrices results in better outcomes as compared to using either of them on their own.

Later, Ledoit and Wolf (2012), the authors introduced the concept of non-linear shrinkage by noting that instead of shrinking the empirical covariance matrix by a global factor towards a predefined target, one can improve on this by shrinking individual eigenvalues of the empirical covariance matrix by varying amounts, resulting in sizeable improvements over linear shrinkage (Ledoit and Wolf, 2017). Finally, in Ledoit and Wolf (2020) the authors draw inspiration from Stein (1975), whereby non-linear shrinkage was used to robustify the covariance matrix against noise. They introduced the concept of Quadratic Inverse Shrinkage (QIS) in which sample eigenvalues are shrunk to close neighbours on either side, with the level of shrinkage decaying with the distance away from these neighbouring eigenvalues. In terms of accuracy, speed, and scalability, QIS performed extremely well compared to other state-of-the-art covariance matrix estimators, but it was also significantly less complex than many of these alternatives. It is for these reasons that we choose to use the QIS approach to remove the noise inherent in the estimated correlation matrix, prior to performing any analysis. For the interested reader, Ledoit and Wolf (2022) provided a comprehensive review of shrinkage estimation, both linear and nonlinear (although this review does not include the QIS method).

3.1.3 The Market Mode/Factor

If one of the aims of our analysis is to determine if any natural clustering structure emerges from the data, then one should not want to distinguish between two stocks that differ only in terms of overall market exposure. In simpler terms, one does not want shares to group together only due to the fact that they have a high beta, or they have a low beta, but rather due to sensitivities to various other economic factors. Consequently, much of the literature advocates removing the largest common factor affecting all of the shares in the market, prior to performing any network or cluster analysis. This largest common factor is often referred to as the *market mode*, as all stocks are affected by this factor in the same direction, i.e. a positive return on this factor implies a positive return on all shares (but in varying magnitudes depending on each share's exposure to this factor) and vice versa. The impact of this factor is usually significantly larger as compared to any other factor, and consequently, it tends to swamp any other interesting effects or hidden structures in the underlying data. MacMahon and Garlaschelli (2015) provide a useful analogy, comparing the market mode to the manner in which the tide affects all boats in a harbour. They state that:

> … all boats in a harbor will rise and fall with the tide. In order to clearly see which 'boats' are rising and falling relative to one another, one must subtract out the common 'tide'.

Interestingly, Borghesi, Marsili and Miccichè (2007) show that although removing the market mode lowers the average correlation amongst stocks, it also makes any cluster structure more evident. They also find that the cluster structure is less sensitive to the periodicity of the data used (at least within the intraday time horizon), when the market mode is removed. Furthermore, Musmeci, Aste and Matteo (2015) show that removing the market mode increases the amount of economic information that weaker hierarchical clustering methods, such as SLCA and ALCA, can extract from a return-based correlation matrix.

3.1.4 The Dynamic (Temporal) Analysis

A long-term analysis, over the full historical period of data, can provide a useful long-term view of the South African stock market. However, analysing the evolution of networks through time will provide valuable insights into the dynamic structure of the stock market, over varying market environments. This analysis mirrors the work done by Majapa and Gossel (2016) and Mbatha and Alovokpinhou (2022) in the South African setting, and Musmeci, Aste and Matteo (2015) on the NYSE.

To ensure that changing market conditions are captured in a timeous manner, and to prevent observations that are rolling out of the data set at each point in time from having an outsized effect on the analysis, exponentially weighted estimation techniques are used. Using this technique, more recent data points are upweighted, while older data points are downweighted. These techniques are applied to both the process in which the market mode is removed and for the estimation of the correlation matrix. This is similar to the analysis done by Pozzi, Di Matteo and Aste (2012) and Musmeci, Aste and Matteo (2015).

3.1.5 Assessing the Robustness of the Analysis

For any statistical analysis, it is worthwhile to determine how robust the results are to small changes in the data set. This type of sensitivity analysis can assist in highlighting the reliability (or lack therefore) of the various insights that we aim to extract from the filtered networks. One of the common statistical ways of assessing reliability or confidence in statistical inference is the use of the concept of *bootstrapping*, which was introduced by Efron (1979). The idea behind bootstrapping is simple, choose random samples with replacement from a data set, perform the analysis on each sample, and recalculate any metrics of interest. One can then look at the distribution of the metrics across all of the samples to determine if they are robust (i.e. small variability across samples would indicate highly robust metrics). This idea was used by Tumminello et al. (2007), Gopi (2008), and Musciotto et al. (2018) to determine the reliability of the links in a filtered network such as the MST and PMFG.

The bootstrap is used here to assess the reliability of the links in the long-term MST and PMFG, and to assess the reliability of the number of clusters that emerges unsupervised from the long-term DBHT.

3.2 Methodology and Data

In this section, we discuss the actual data that was used in the analysis and the methodologies that were followed.

3.2.1 Period of Analysis

The period of the analysis that was considered was approximately 20 years (over 1000 points of weekly data), from 2003 to 2020. This was considered to provide a good balance between having a sufficient amount of historical data over which to conduct the analysis (also including a variety of market conditions such as bull markets, bear markets, and crashes), and retaining a sufficiently large number of stocks that have the data over this period.

3.2.2 Universe of Stocks

The ALSI, as at the end of June 2022, was selected as the starting point to determine the universe of stocks for the analysis. The index consisted of 136 unique stocks at this date. There were 58 stocks that were removed due to insufficient historical data over the period of analysis (2003-2022). Stocks that had thin-trading concerns were also removed. This list consisted of 6 stocks that traded in less than 95% of the total trading days, as well as stocks that had wide bid-ask spreads. Taking the above into consideration, the universe of stocks that was eventually used in the analysis was filtered as shown in Figure 3.1. Appendix A2 contains a full list of stocks that were considered for inclusion, reasons for exclusions, as well as the final list of stocks that were used for the analysis.

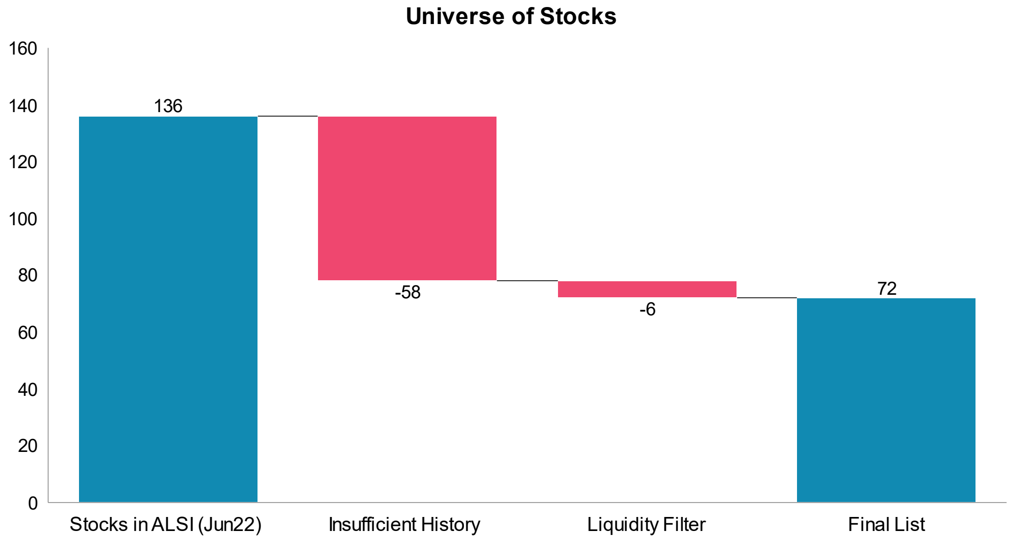


Figure 3.1 Filtering Process for the Universe of Stocks

3.2.3 Data Frequency/Periodicity

With regards to the periodicity of the return data that was used to estimate the correlation matrices, one may argue that the use of daily data may be subjected to high levels of market noise, but also that this noise should be averaged out over longer time horizons (such as weekly, or monthly). Given this assumption, monthly data should contain the least amount of noise. However, using monthly data will also result in a smaller number of historical observations to employ in the estimation of the correlation matrix. As discussed in Section 3.1.2, RMT indicates that the reduced number of data points may result in statistically noisier estimates of the correlation matrix (especially if the number of stocks is fixed). Therefore, weekly data was chosen as a compromise between daily data and monthly. Furthermore, the use of weekly data should mitigate some of the thin-trading effects seen in less liquid shares (as mentioned in Section 3.1.1).

Note that although we have not focused on methods of quantitatively choosing the optimal length of the period of analysis, or the optimal frequency of the data, Marti et al. (2015), and Marti et al. (2016) provided a framework to tackle these questions, that the interested reader may find useful.

3.2.4 Removing the Market Mode

Borghesi, Marsili and Miccichè (2007) discussed various methods of removing the market mode. These methods vary according to the definition of the market mode and how it is calculated. Some of the definitions included the use of the average return across all stocks, the use of the return of a market index, or the use of the largest factor from a Principal Component Analysis (PCA). There are also a variety of methods in which the market mode can be removed, i.e. by simple subtraction, or by orthogonalising the return of each stock to the market mode using a regression framework.

The method employed here is as follows:

1. Standardise (z-score) the return data to remove the impact of stock volatilities. This effectively takes the stock returns into correlation space.
2. Perform a Principal Component Analysis (PCA) on the standardised stock return data.
3. Check that the first principal component is indeed the market mode. This is reasonable if all or most of the entries in the first eigenvector have the same sign.
4. Find the first principal component scores, which represent the returns of the market mode.
5. For each stock, perform a regression of the stock returns against the first principal component scores.
6. Replace the stock returns with the residual returns from this regression analysis.

According to Borghesi, Marsili and Miccichè (2007), the use of this method resulted in increased stability of clusters across various periodicities, and improved information gain relative to the economic sectors. Note that these steps were performed before any denoising processes were applied.

3.2.5 The Choice of the Similarity Measure

As discussed in Section 2.8 a variety of similarity measures have been employed in financial market networks, such as Spearman's rank correlations, Kendall's tau, mutual information, tail dependence, and copula methods.

In this dissertation, the well-understood Pearson correlation estimate was used to create the correlation-based networks using the formula in Equation (1). Note that each correlation estimate in a correlation matrix was converted to a distance metric before any analysis was performed, as discussed in Section 2.8, using Equation (3). Note that for the long-term analysis no exponential weighting of time points was applied.

3.2.6 De-noising the Correlation Matrices

As discussed in Section 3.1.2 we removed noise inherent in the correlation matrix prior to performing any filtering techniques. We employed the Quadratic-Inverse Shrinkage (QIS) method using MATLAB code from Ledoit (2022).

3.2.7 Quantifying the Amount of Economic Information Retrieved

As done in Musmeci, Aste and Matteo (2015), we quantified and compared the degree of economic information that was extracted from the various clustering methods using the Adjusted Rand Index (ARI). The ARI was introduced by Hubert and Arabie (1985), and allows one to compare the outcomes from differing clustering methods on the same set of data. Given that the ICB classification can be used to partition the stocks into various groups, we used the ARI to determine the overlap/similarity between the clusters and the various ICB classifications. The ARI then served as a proxy for the economic information that was extracted from each clustering method.

Assume that we are given an n object set $S = \{O_1, \dots, O_n\}$, and suppose $U = \{U_1, \dots, U_R\}$ and $V = \{V_1, \dots, V_C\}$ represent two different partitions (e.g. a clustering or an ICB classification) of S. Letting n_{ij} denote the number of objects that are common to classes u_i and v_j, the information on class overlap between the two partitions U and V can be written in the form of the contingency table below.

Table 3.1 Contingency Table

	V_1	V_2	$\cdots$	V_C	*sums*
U_1	n_{11}	n_{12}	$\cdots$	n_{1C}	a_1
U_2	n_{21}	n_{22}	$\cdots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	a_R
Sums	b_1	b_2	$\cdots$	b_C	

The Adjusted Rand Index (ARI) is then calculated as below:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}. \tag{5}$$

Where

$$\binom{n}{k} = \frac{n \times (n-1) \times \ldots \times (n-k-1)}{k \times (k-1) \times \ldots \times 1}.$$

The ARI typically takes on values between zero and one (with a value of one indicating full agreement between the clustering method and the ICB classification). However, negative values are possible if there is a negative correlation between the ICB classification and the clustering method.

Another method that was used to help identify the link between each ICB industry or super-sector was to use a Sankey chart. Sankey charts are used to depict the *flow* of data from one (or more) groups, clusters, or partitions to another set of groups, clusters, or partitions. Marti et al. (2015) used this method to visualise how a variety of perturbations in the clustering inputs, led to changes in the resultant clusters.

3.2.8 Network Filtering and Clustering Techniques

The MSTs, PMFGs, and DBHTs were all created using MATLAB code. The MSTs were created using the built-in MATLAB function, the PMFGs were built using code from Aste (2012), and the DBHTs were constructed using code from Aste (2014). We also compared the DBHT clustering to clusters that were extracted from hierarchical clustering using SLCA, ALCA, and Ward's linkage functions (which were discussed in Section 2.11.4).

For the network visualisations we used the Pajek (Batagelj and Mrvar, 2004) graphing software to plot the basic layout of the networks (using the force-directed method of Kamada and Kawai method for the layout, as discussed in Section 2.6). These figures are then exported as SVG files and imported into the vector graphics editor program, Inkscape (Inkscape Project, 2022) to edit and improve the overall quality of the visualisation.

Figure 3.2 depicts the methodological flow that is followed for the long-term (full period) analysis.

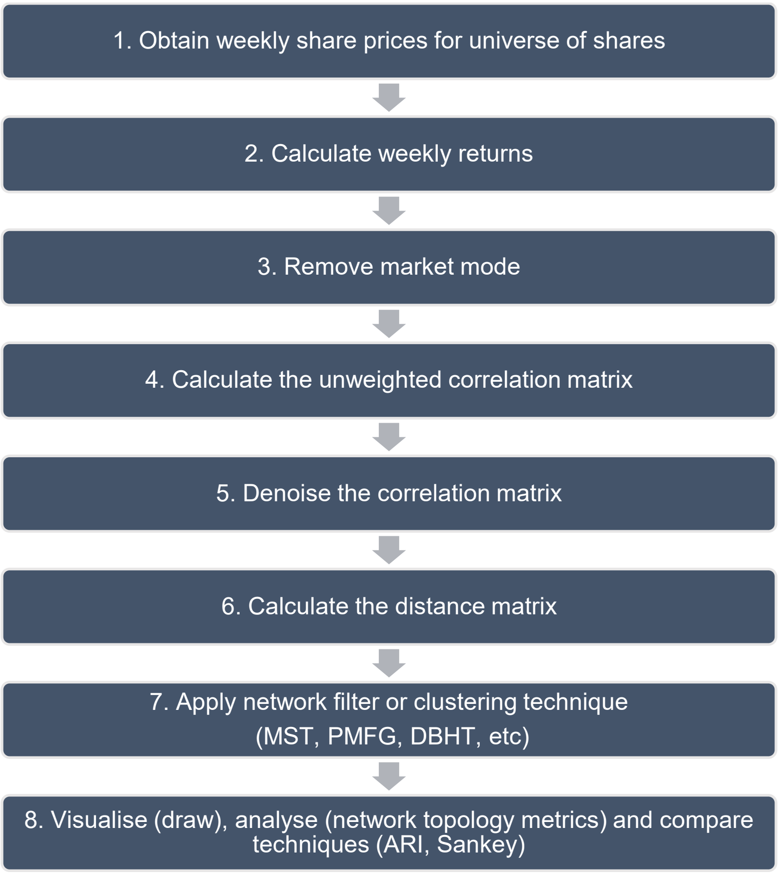

1. Obtain weekly share prices for universe of shares
2. Calculate weekly returns
3. Remove market mode
4. Calculate the unweighted correlation matrix
5. Denoise the correlation matrix
6. Calculate the distance matrix
7. Apply network filter or clustering technique (MST, PMFG, DBHT, etc)
8. Visualise (draw), analyse (network topology metrics) and compare techniques (ARI, Sankey)

Figure 3.2 Process Flow for the Long-Term Analysis

3.2.9 The Dynamic (Temporal) Analysis

For the dynamic analysis, steps 2 – 8 from Figure 3.2 were calculated using a rolling 3-year (156-week) window, stepping forward by 4 weeks (approximately 1 month) each time. Note that we did not redetermine the universe of stocks at each point in time, but we used the overall universe of stocks as described in Section 3.2.2. These were the stocks that traded continuously over the full period of analysis. While this may have resulted in some survivorship bias, given that we were not using these techniques to predict any future outcomes, but rather to analyse the structure of the market at each point in time, we do not believe that this was a stumbling block.

As discussed in Section 3.1.4, exponentially weighted estimation techniques were used to ensure that the changing market conditions were captured in a timeous manner, and to prevent observations that are rolling out of the data set at each point in time from having an outsized effect on the analysis. The following formula was used to determine the weight applied to each data point t amongst the $T = 156$ weekly observations used for each rolling window:

$$w_t \ = w_0 \, e^{\alpha(t-T)}, \ \forall \, t \in \{1,2,\ldots,T\}. \qquad \textbf{(6)}$$

The α parameter is referred to as the *exponential decay factor*, and controls the amount of weight being applied to more recent observations, i.e. the larger the α parameter, the larger the weight being applied to more recent observations. The constant value, w_0, is solved for by ensuring that the weights all sum up to a value of one. In our analysis, the exponential decay factor was set to a value of 0.005. This resulted in the most recent year in each rolling window receiving a weight of 42%, the second most recent year receiving 33%, and the earliest of the three years receiving a weight of 25%. This provided a good balance between responsiveness to market changes and robustness of the results (i.e. less susceptibility to market noise). Figure 3.3 shows the weight applied to each time point in a 156-week estimation period using a normal/equal time-weighting scheme, as compared to an exponentially-weighted scheme (with a decay factor of 0.005). One can see how the more recent observations have an increasingly larger weight, as compared to older observations.

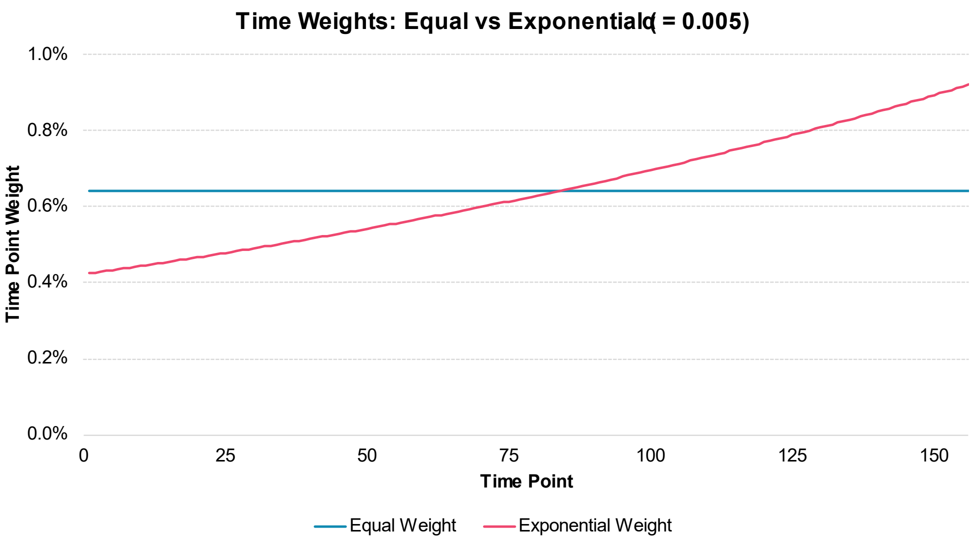


Figure 3.3 Time Weighting Schemes: Equal vs Exponential for a 3-Year Window

Having established the weighting scheme that will be used, an exponentially weighted correlation matrix can be calculated as follows:

Let P_t^i = the price of stock i at time t.

Then $r_t^i \ = \frac{P_t^i}{P_{t-1}^i} - 1$ = the return of stock i at time t.

Let T = the number of return observations.

Let w_t = weight of time point t as described in Equation (6).

Let $\bar{r}^i \ = \ \sum_{i=1}^{T} w_t \times r_t^i$ = the weighted average return of stock i.

$$\text{Then } \rho_{ij} = \frac{\sum_{t=1}^{T} w_t (r_t^i - \bar{r}^i)(r_t^j - \bar{r}^j)}{\sqrt{\sum_{t=1}^{T} w_t (r_t^i - \bar{r}^i)^2 \sum_{t=1}^{T} w_t (r_t^j - \bar{r}^j)^2}} . \qquad \textbf{(7)}$$

Note that the removal of the market mode and the denoising process (which were repeated at each time step) also needed to be adapted to cater for the non-standard time weighting. For the removal of the market mode, the PCA step was conducted using a weighted-PCA approach (in which the weight of each observation was incorporated into the estimation of the PCA scores), and the orthogonalisation/regression step was conducted using a weighted least squares approach. Both steps made use of built-in MATLAB functions. For the denoising process, the MATLAB code of Ledoit (2022) was adapted to cater for the non-standard time weighting scheme.

3.2.10 How to Use Bootstrapping to Assess the Robustness of the Analysis

3.2.10.1 Assessing the Robustness of the Filtered Networks

The bootstrap samples were determined using the MATLAB function “bootstrp”. The statistical bootstrapping process that was used to assess the reliability of the links in the MSTs and PMFGs is described in Figure 3.4. The number of samples needs to be sufficiently large so that the random error due to the bootstrap process is small. For this analysis, the number of bootstrap samples was set to 10 000, as this was deemed to be sufficiently large given that the full period data set consisted of approximately 1 000 weekly observations. This was also confirmed by comparing the results from 1 000 samples to the results using 10 000 samples. We noted that the results were not significantly different between these two scenarios.

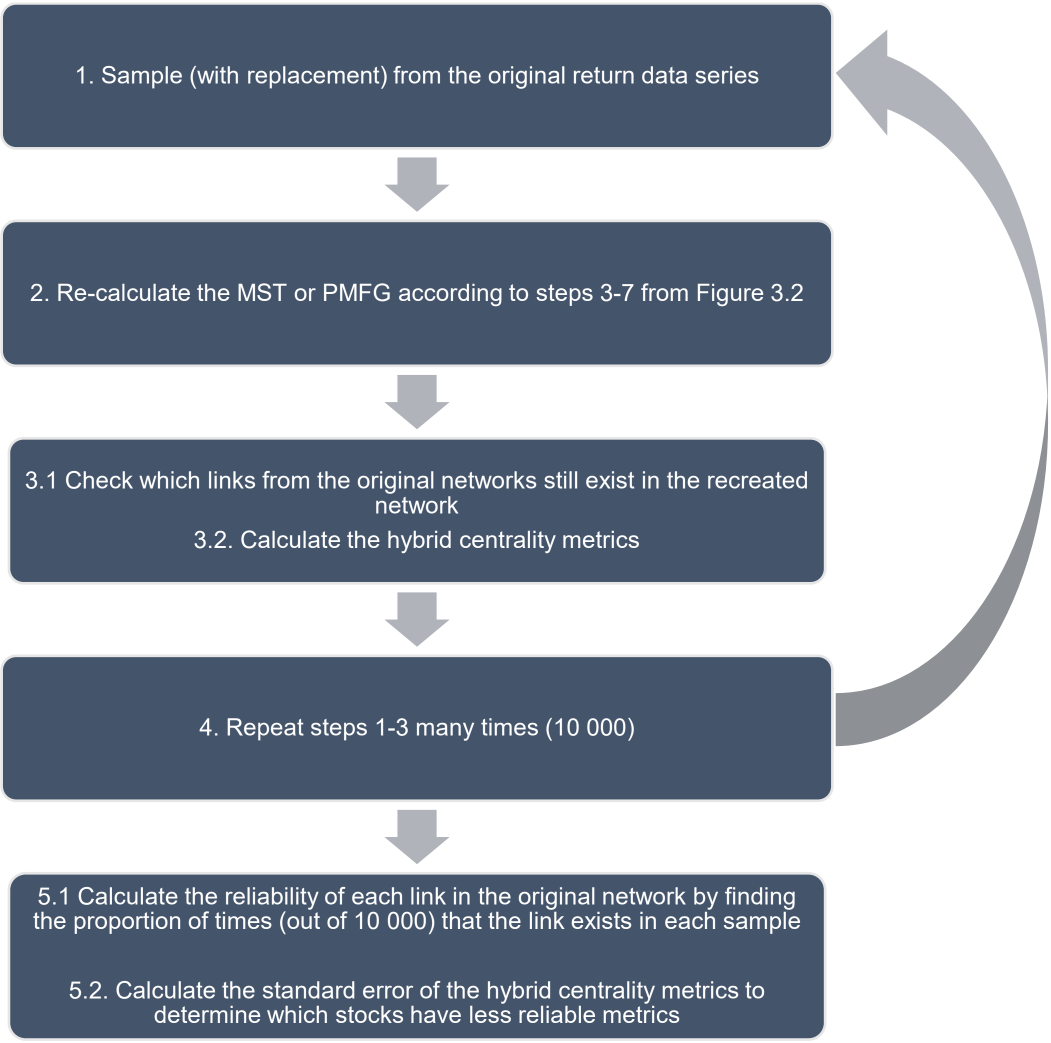

1. Sample (with replacement) from the original return data series
2. Re-calculate the MST or PMFG according to steps 3-7 from Figure 3.2
3.1 Check which links from the original networks still exist in the recreated network
3.2. Calculate the hybrid centrality metrics
4. Repeat steps 1-3 many times (10 000)
5.1 Calculate the reliability of each link in the original network by finding the proportion of times (out of 10 000) that the link exists in each sample
5.2. Calculate the standard error of the hybrid centrality metrics to determine which stocks have less reliable metrics

Figure 3.4 Process Flow: Bootstrap Reliability Analysis for the Filtered Networks

3.2.10.2 Assessing the Robustness of the Number of Clusters

The bootstrapping analysis was also used to determine the robustness of the number of clusters that was extracted from the DBHT. This process is described in Figure 3.5, and similar to the bootstrapping analysis described in the previous section, the number of samples used was 10 000. Although not reported here, it should be noted that the number of clusters seemed to converge to the overall average after a relatively small number of samples (approximately 100 samples).

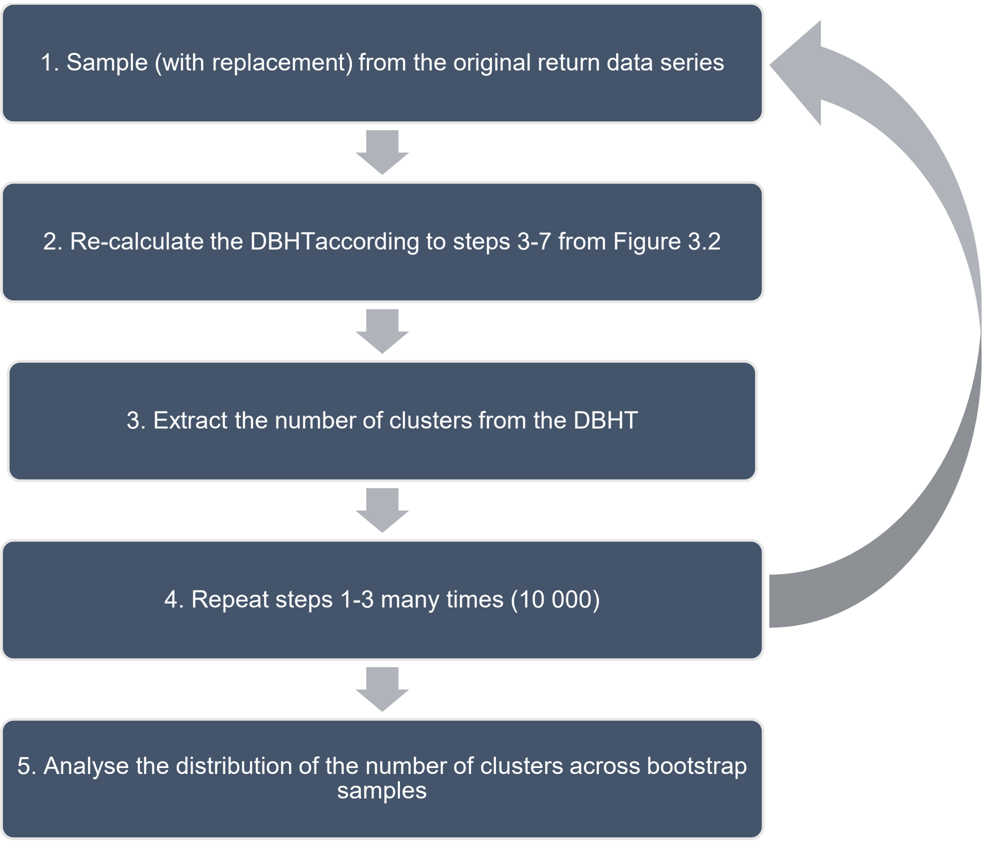

1. Sample (with replacement) from the original return data series
2. Re-calculate the DBHTaccording to steps 3-7 from Figure 3.2
3. Extract the number of clusters from the DBHT
4. Repeat steps 1-3 many times (10 000)
5. Analyse the distribution of the number of clusters across bootstrap samples

Figure 3.5 Process Flow: Bootstrap Reliability Analysis for the DBHT

3.2.11 Network Topology Measures

In this section, we adapt the notation from Oldham et al. (2019) and MATLAB (2022) to precisely define the calculations of the various network topology metrics. A network can be represented as an $N \times N$ adjacency matrix A in which the element $A_{ij} = 1$ if vertices i and j are connected, and $A_{ij} = 0$ otherwise. We denote the adjacency matrix of a weighted network W, where the element W_{ij} contains the weight of the edge between vertices i and j. Following Pozzi, Di Matteo and Aste (2013), we set $W_{ij} = (1 + \rho_{ij})$, where ρ_{ij} is the correlation between stock i and j for the weighted degree, and weighted eigenvector centrality metrics. Using this method, a higher weight is given to stocks that have a higher correlation with other stocks. However, for the weighted closeness, eccentricity, and betweenness centrality metrics we set $W_{ij} = \sqrt{2(1 - \rho_{ij})}$, which is the standard transformation of the correlation measure into a distance metric (as shown in Section 2.8), and is the same distance metric used to filter the networks. Similarly, for the normalised tree length (NTL), we set $W_{ij} = \sqrt{2(1 - \rho_{ij})}$, as we are interested in the total distance of all edges in the network. We also let N_v represent the number vertices in the network, while N_E is the number of edges in the network, and E_k is the weight of the k^{th} edge in the network.

3.2.11.1 Normalised Tree Length (NTL)

$$NTL = \frac{1}{N_E - 1} \sum_{k=1}^{N_E} E_k. \tag{8}$$

3.2.11.2 Degree Centrality (DC)

$$Unweighted\ DC_i = DC_i^u = \sum_{j \neq i} A_{ij}. \tag{9}$$

$$Weighted\ DC_i = DC_i^w = \sum_{j \neq i} W_{ij}. \tag{10}$$

3.2.11.3 Eigenvector Centrality (EVC)

If v^{u} is the eigenvector associated with the largest eigenvalue λ_1^{u} of the adjacency matrix A then

$$Unweighted\ EVC_i = EVC_i^u = \frac{1}{\lambda_1^u} \sum_j A_{ji} v_j^u. \tag{11}$$

If v^{w} is the eigenvector associated with the largest eigenvalue λ_1^{w} of the weighted adjacency matrix W then

$$Weighted\ EVC_i = EVC_i^w = \frac{1}{\lambda_1^w} \sum_j W_{ji} v_j^w. \tag{12}$$

3.2.11.4 Closeness Centrality (CC)

If NR_j is the number of reachable vertices from vertex j (not including j itself) and l_{ij}^u is the shortest unweighted distance between vertices i and j then

$$Unweighted\ CC_i = CC_i^u = \left(\frac{NR_j}{N_v - 1}\right)^2 \frac{1}{\sum_j l_{ij}^u}. \tag{13}$$

If NR_j is the number of reachable vertices from vertex j (not including j itself) and l_{ij}^w is the shortest weighted distance between vertices i and j then

$$Weighted\ CC_i = CC_i^w = \left(\frac{NR_j}{N_v - 1}\right)^2 \frac{1}{\sum_j l_{ij}^w}. \tag{14}$$

3.2.11.5 Eccentricity Centrality (EC)

If l_{ij}^u is the shortest unweighted distance between vertices i and j then

$$Unweighted\ EC_i = EC_i^u = \frac{1}{max(l_{ij}^u)}. \tag{15}$$

If l_{ij}^w is the shortest weighted distance between vertices i and j then

$$Weighted\ EC_i = EC_i^w = \frac{1}{max(l_{ij}^w)}. \tag{16}$$

3.2.11.6 Betweenness Centrality (BC)

If g_{pq} is the number of shortest paths between vertices p and q and $g_{pq}(i)$ is the number of shortest paths between vertices p and q which pass through node i, then the betweenness centrality of node i is

$$Unweighted\ BC_i = BC_i^u = \sum_{p\neq i, p\neq q, q\neq i} \frac{g_{pq}(i)}{g_{pq}}. \tag{17}$$

In a weighted network, the shortest path is calculated as the path with the smallest edge-weighted sum. If g_{pq} is the number of shortest paths between vertices p and q and $g_{pq}(i)$ is the number of shortest paths between vertices p and q which pass through node i, then the betweenness centrality of node i is

$$Weighted\ BC_i = BC_i^w = \sum_{p\neq i, p\neq q, q\neq i} \frac{g_{pq}(i)}{g_{pq}}. \tag{18}$$

3.2.11.7 Hybrid Centrality (X+Y)

If the ten centrality metrics (five unweighted and five weighted) are ranked in ascending order (i.e. the lower the better), then

$$X_i = \frac{Rank(DC_i^u) + Rank(DC_i^w) + Rank(BC_i^u) + Rank(BC_i^w) - 4}{4 \times (N_v - 1)}. \tag{19}$$

$$Y_i = \frac{Rank(EC_i^u) + Rank(EC_i^w) + Rank(CC_i^u) + Rank(CC_i^w)}{6 \times (N_v - 1)} + \frac{Rank(EVC_i^u) + Rank(EVC_i^w) - 6}{6 \times (N_v - 1)}. \tag{20}$$

$$(Hybrid\ Centrality)_i = X_i + Y_i. \tag{21}$$

3.2.12 Data Sources

The constituents of the ALSI, as well as the weekly share price data, volume data, bid-ask spreads, and price/book data for these constituents, was retrieved from Bloomberg. The weekly data for the Rand Dollar exchange rate, was retrieved from IRESS.

Chapter 4 Results

In this section, we present the results of our analysis of the South African stock market. Given that the cross-correlation amongst stocks is the metric that was used to measure the strength of relationships in the network, we begin with a brief empirical analysis of the distribution of values in the correlation matrix. We considered both a *long-term* (full period) analysis, followed by a *dynamic* (rolling) analysis, in which we determined the impact that varying market conditions had on the structure of the correlation matrix. We then proceeded to analyse the MST (both long-term and dynamic), followed by a similar analysis of the PMFG and the DBHT.

Figure 4.1 depicts a flowchart of the five main sections of this chapter, along with a summary of the underlying analysis that was performed in each of them.

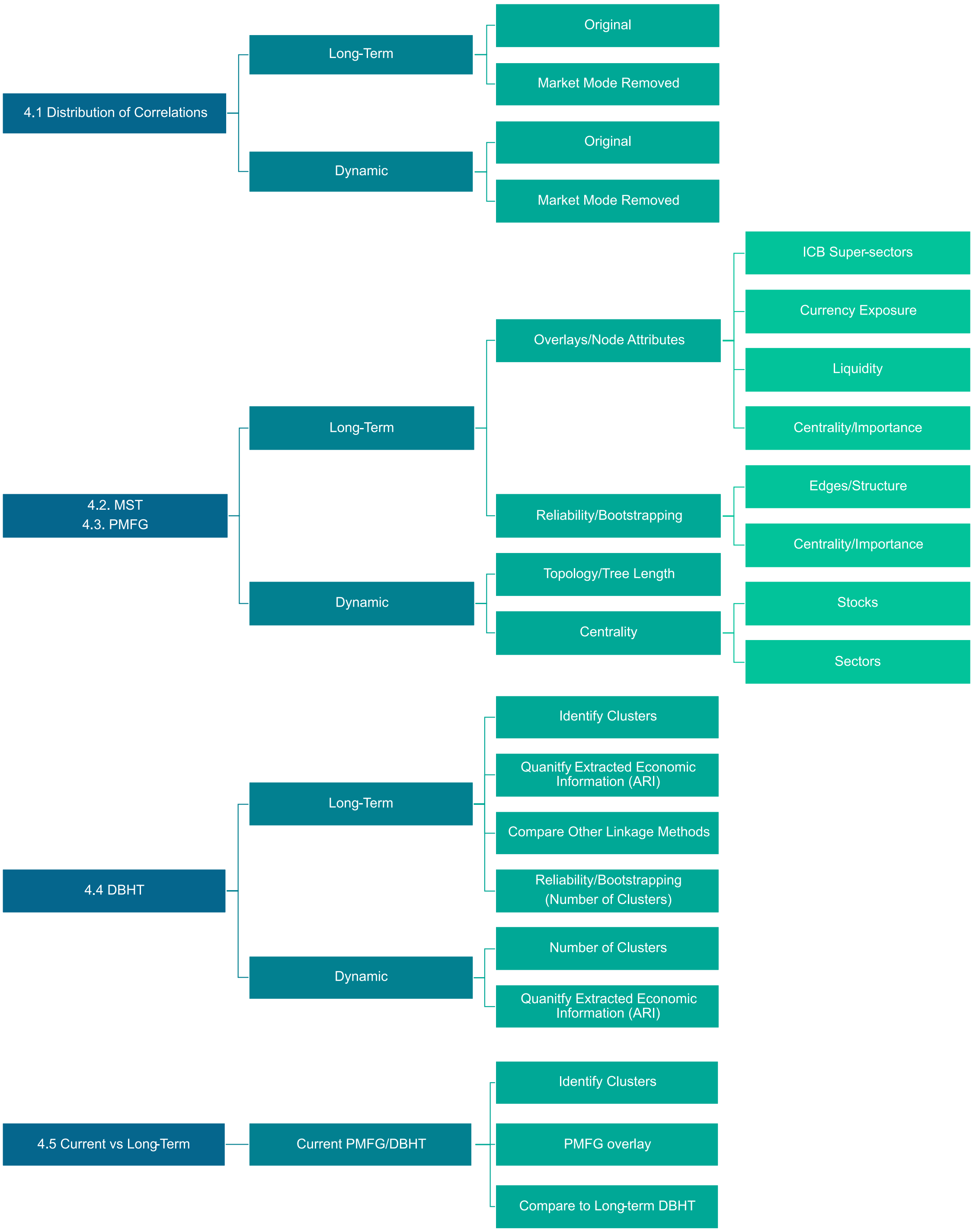


Figure 4.1 Flow Chart: Summary of Analysis

4.1 Empirical Distribution of Financial Correlations

4.1.1 Long-Term Correlations

We begin by depicting the impact of removing the market mode on the distribution of correlations for our universe of stocks. This is shown using an empirical distribution of the correlations in Figure 4.2. One can see that the values of the correlations amongst stocks typically ranged between -0.2 to 0.8, with a peak at a correlation of 0.2. Removing the market mode did lower the level of correlation, with the peak of the distribution moving to a value close to 0. However, the range of correlations was still wide, varying between -0.4 to 0.8. As discussed in 3.1.3, even though the process of removing the market mode lowers the average correlation amongst stocks, it does have other useful benefits from a clustering point of view.

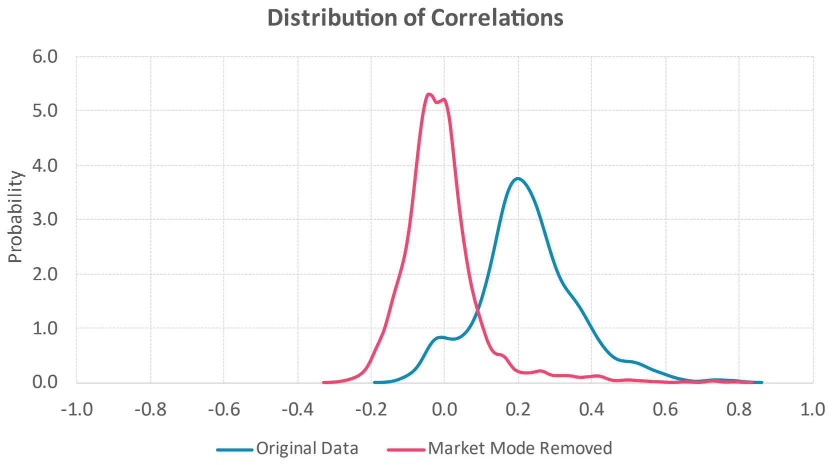


Figure 4.2 Long-Term Correlation Distribution: Normal vs Market Mode Removed

4.1.2 Dynamic Correlations

In Figure 4.3 we show the distribution of the dynamic correlations (calculated using a 3-year exponential-weighing scheme), and the impact of removing the market mode. It was interesting to note that while the range of normal correlations was affected by general market events (such as market crashes), the distribution of correlations with the market mode removed was not affected by these events. Note that in all of the subsequent analyses, the market mode was removed from the data.

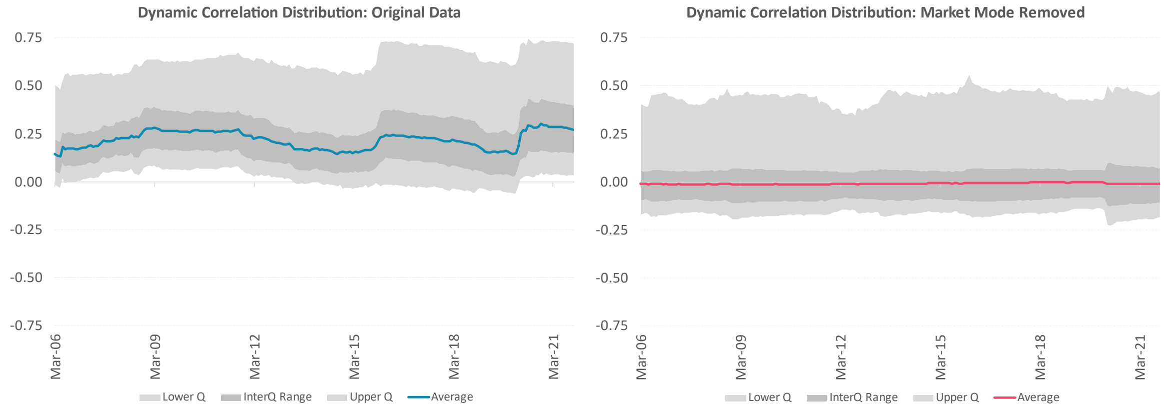


Figure 4.3 Dynamic Correlation Distribution: Normal vs Market Mode Removed

4.2 Analytic Results for the MST

This section contains the results of the application of the MST to the South African stock market, on both a long-term basis, as well as from a dynamic point of view.

4.2.1 Long-Term MST: Sector Overlay

Figure 4.4 depicts the basic, long-term MST for South African stocks based on the correlation matrix. It is important to note that the positions of vertices on the page are not unique and depends on the layout algorithm used. So, the closeness of stocks needs to be considered in the context of their position along the edges of the MST (see Section 2.6 for more information regarding the various layout algorithms).

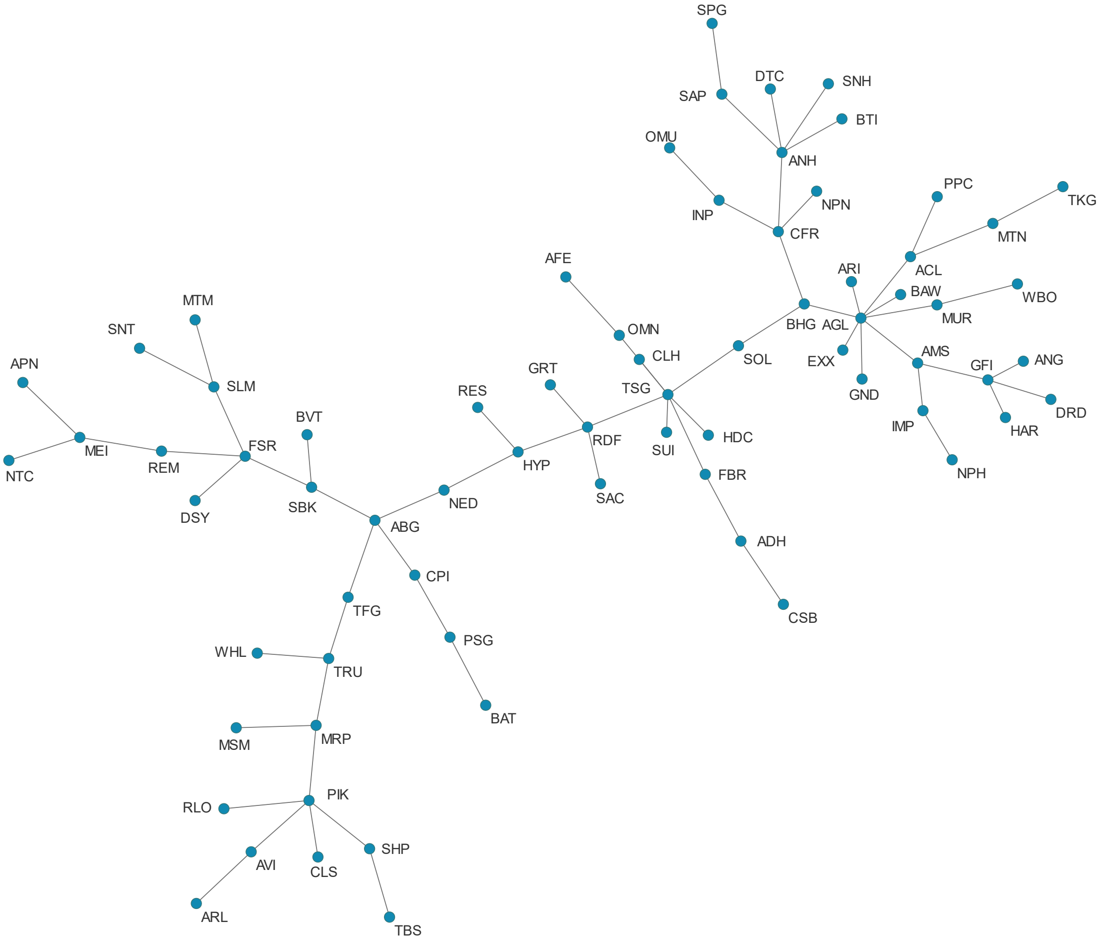


Figure 4.4 Long-Term MST: Standard View

An experienced investment analyst would have been able to glean meaningful insights from here, noticing that shares that were in related ICB sectors were located close to each other on the MST. As discussed in Section 2.4, the visual nature of the MST can be enhanced by overlaying the ICB classification of each share. This would allow the user to determine if the MST recovers the ICB classification in an unsupervised manner (i.e. without any a priori knowledge). This is shown in Figure 4.5, where each vertex was shaded in a colour that reflected the underlying share's ICB super-sector classification.

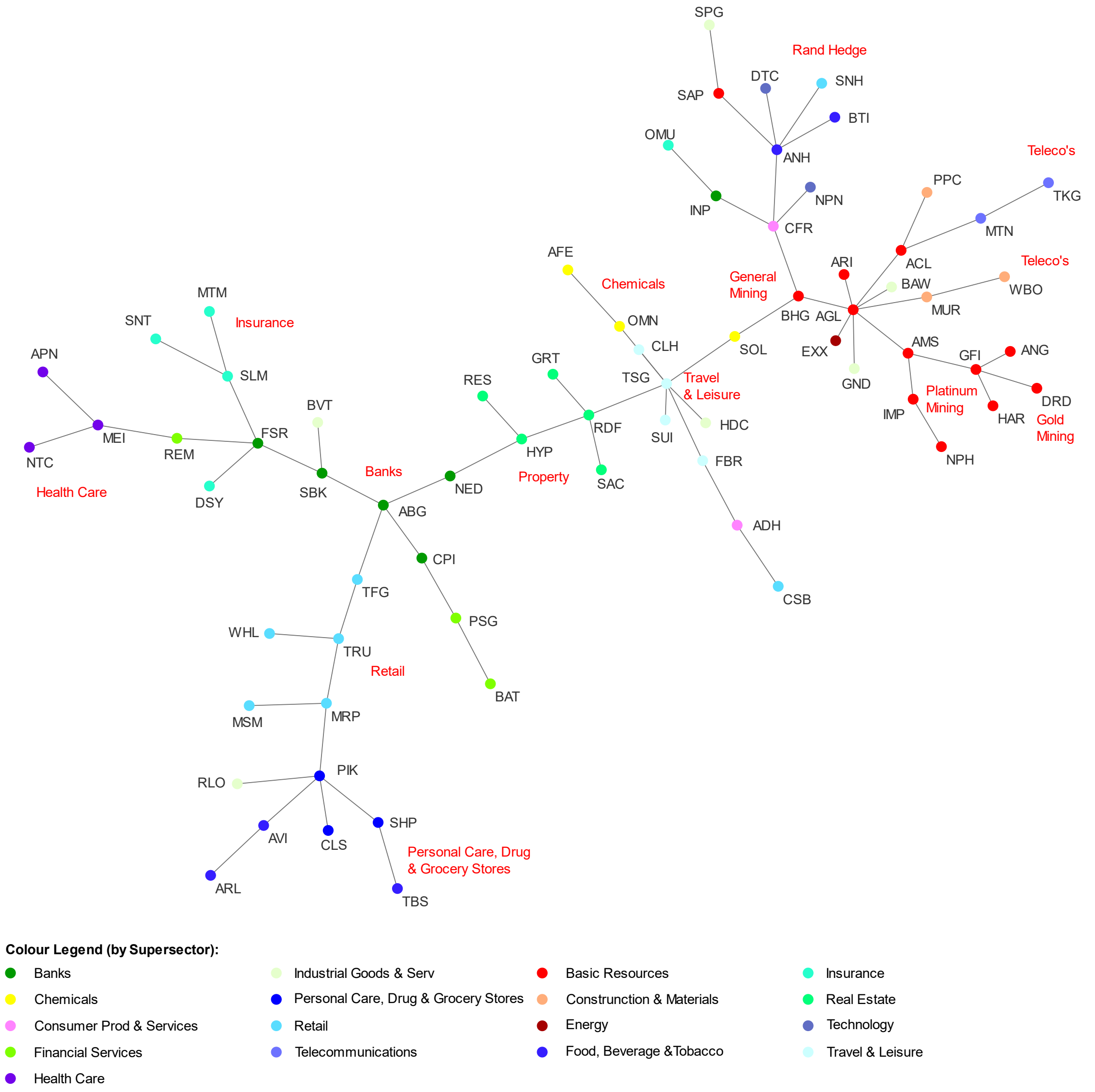


Figure 4.5 Long-Term MST: Super-sector Overlay

Overlaying the ICB super-sector classification onto the MST, highlighted the overlap between each share's positioning on the MST and the sector classification. We also manually labelled various locations on the MST (in red) to highlight the overlap between the position of shares on the MST and the ICB classifications.

It was easy to see that in general, shares that formed part of the same super-sector tended to be located near each other on the MST. If one recalls that the MST was constructed by filtering the return-based correlation matrix, and the fact that the correlation matrix captures the co-movement of shares, the MST was effectively highlighting that the share prices of companies within the same sector tended to move together.

4.2.2 Long-Term MST: Currency Exposure Overlay

Although the positioning of many shares conformed to their ICB classification, certain shares did not follow this classification. For example, shares from the Industrial Goods and Services super-sector tended to be dispersed across the MST. There was also a cluster of shares at the top-right of the MST (e.g. CFR, NPN, INP, BTI, etc.) that emanated from a variety of sectors. This indicated that while the ICB super-sector classification tended to be a good representation of the manner in which shares co-move, the MST captured additional effects over and above this classification. To investigate this further, a variety of alternative information was overlaid onto the MST (by varying the size and colour of each vertex), beginning with the exposure of each share to the Rand Dollar exchange rate in Figure 4.6.

The exposures are relative in nature because the market mode was removed from the data set (as described in Section 3.2.4). The vertices that are shaded in the dark blue colour represent shares that had a positive exposure to weakness in the Rand (i.e. they had a positive beta emanating from a regression of the share’s returns against the Rand Dollar exchange rate). These shares would have been expected to perform relatively well when the Rand weakened. The vertices that are shaded in the teal colour represent shares that had a positive exposure to strength in the Rand (i.e. they had a negative beta emanating from a regression of the share’s returns against the Rand Dollar exchange rate). These shares would have been expected to perform relatively well when the Rand strengthened. The size of each vertex represents the magnitude of these exposures.

One can clearly see that the cluster of shares at the top-right of the MST represents shares that have a large exposure to Rand weakness (with large vertices that are shaded in dark blue), i.e. they would have benefited to a relatively large extent when the Rand weakened. These types of shares have been referred to in the literature (and by practitioners) as *Rand Hedge* shares (e.g. see Barr, Kantor and Holdsworth, 2007). Fundamentally, these companies, although listed on the JSE, have a large proportion of their revenues and costs that are not Rand denominated. Therefore, their economic performance is often independent (i.e. hedged) from the performance of the South African economy. One can therefore posit that *the Rand Hedge nature of these shares overrides that of the ICB classification*, with these shares forming a cluster of their own.

It is also interesting to note that shares which benefited from Rand weakness (with vertices shaded in dark blue) are located on the right of the MST, while shares which benefited from Rand strength (the so-called *Rand Play* shares) are located on the left. This highlights that the Rand Dollar exchange rate was an important overall factor that impacted the co-movement of shares on the JSE in a global manner. It was once again interesting, that this structure emerged from the MST in an unsupervised manner.

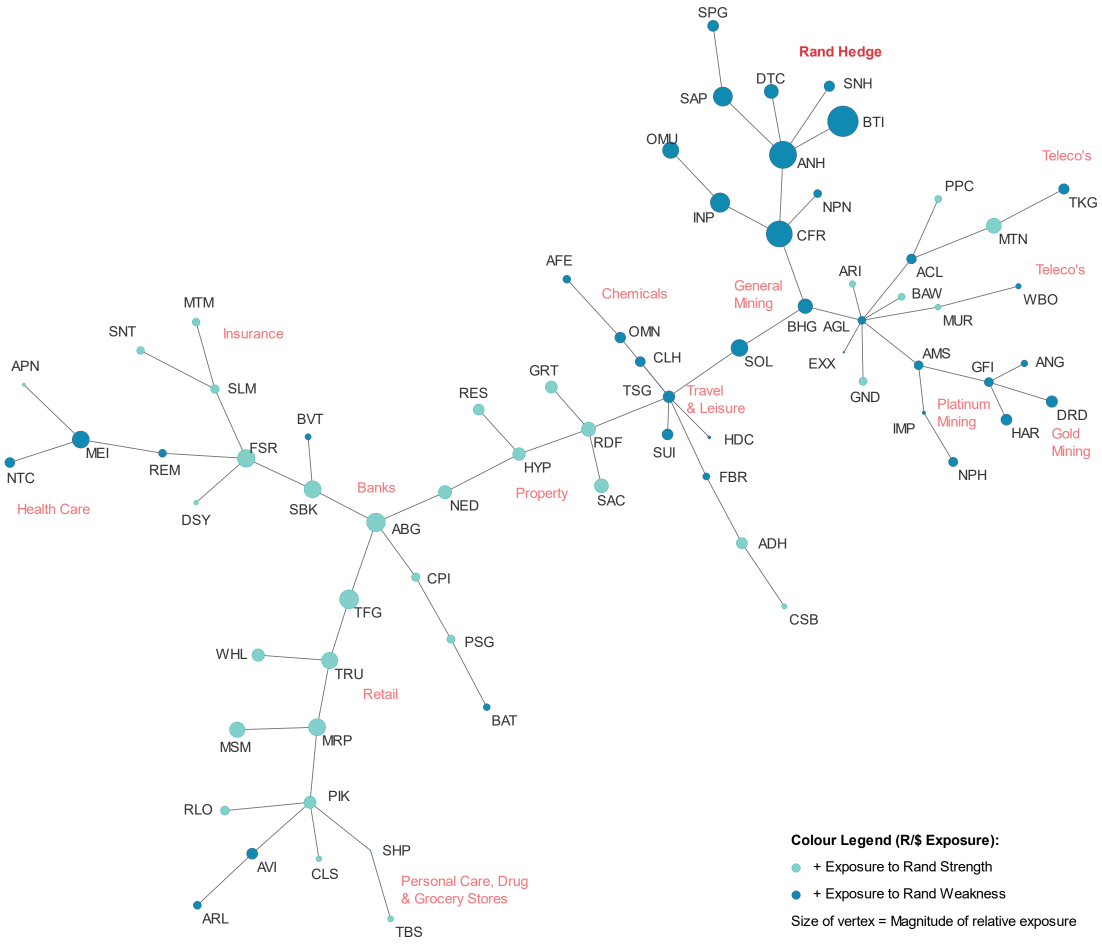


Figure 4.6 Long-Term MST: Currency Exposure Overlay

4.2.3 Long-Term MST: Liquidity Overlay

As discussed in Section 3.1.1, shares that do not trade regularly (i.e. they suffer from lower liquidity, or have wider bid-ask spreads) tend to have biases in metrics that are calculated based on price data (such as market betas, or cross-correlations). This effect was catered for by applying a liquidity filter to the universe of shares that were included in the analysis. Nevertheless, some groups of shares may have traded in significantly lower volumes as compared to larger and more liquid shares. Therefore, in Figure 4.7, we highlight shares that fell into the lower third of our universe for liquidity. The liquidity metrics used here were an average of a ranking based on volume-traded and a ranking based on bid-ask spreads.

From Figure 4.7 one can see that many of the less liquid shares (which are shaded in the dark blue colour) were located on the periphery (or edges) of the MST. However, there was also a cluster of shares in the middle of the MST, centred around TSG, that exhibited lower liquidity. This was informative, as ordinarily, one would not expect sectors such as Travel & Leisure to be located close to the Chemicals sector, as well as other shares such as HDC, ADH, and CSB, which come from a variety of sectors. This highlighted that the *liquidity of shares was also an important factor that impacted the co-movement of shares* on the JSE and was not necessarily captured by the ICB classification. Again, this structure emerged from the MST in an unsupervised manner.

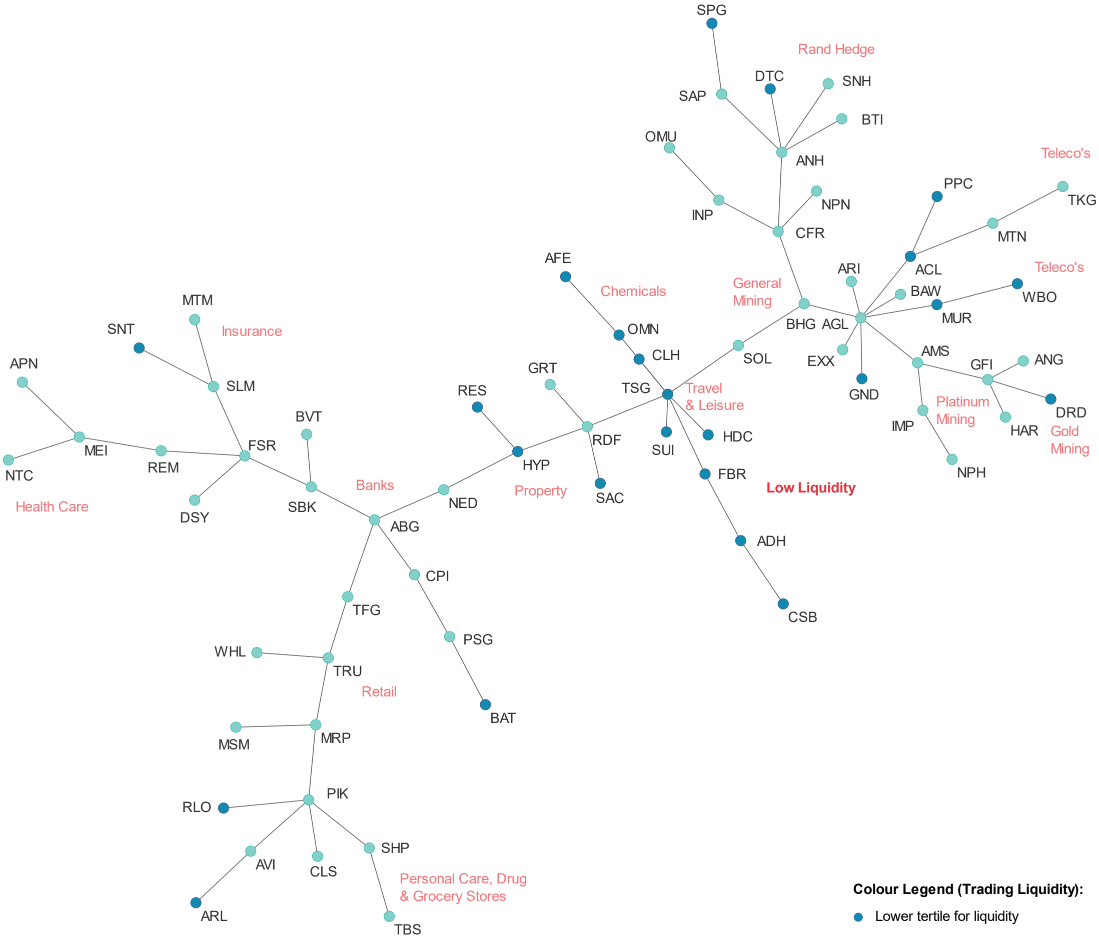


Figure 4.7 Long-Term MST: Low Liquidity Overlay

4.2.4 Long-Term MST: Hybrid Centrality Overlay

As discussed in Section 2.13, there are a variety of measures that can be used to assess the importance of a share in a network, such as the MST. In Figure 4.8 we depict the MST with an overlay of the hybrid centrality metric of Pozzi, Di Matteo and Aste (2013) to obtain a better understanding of the importance of each share on the MST.

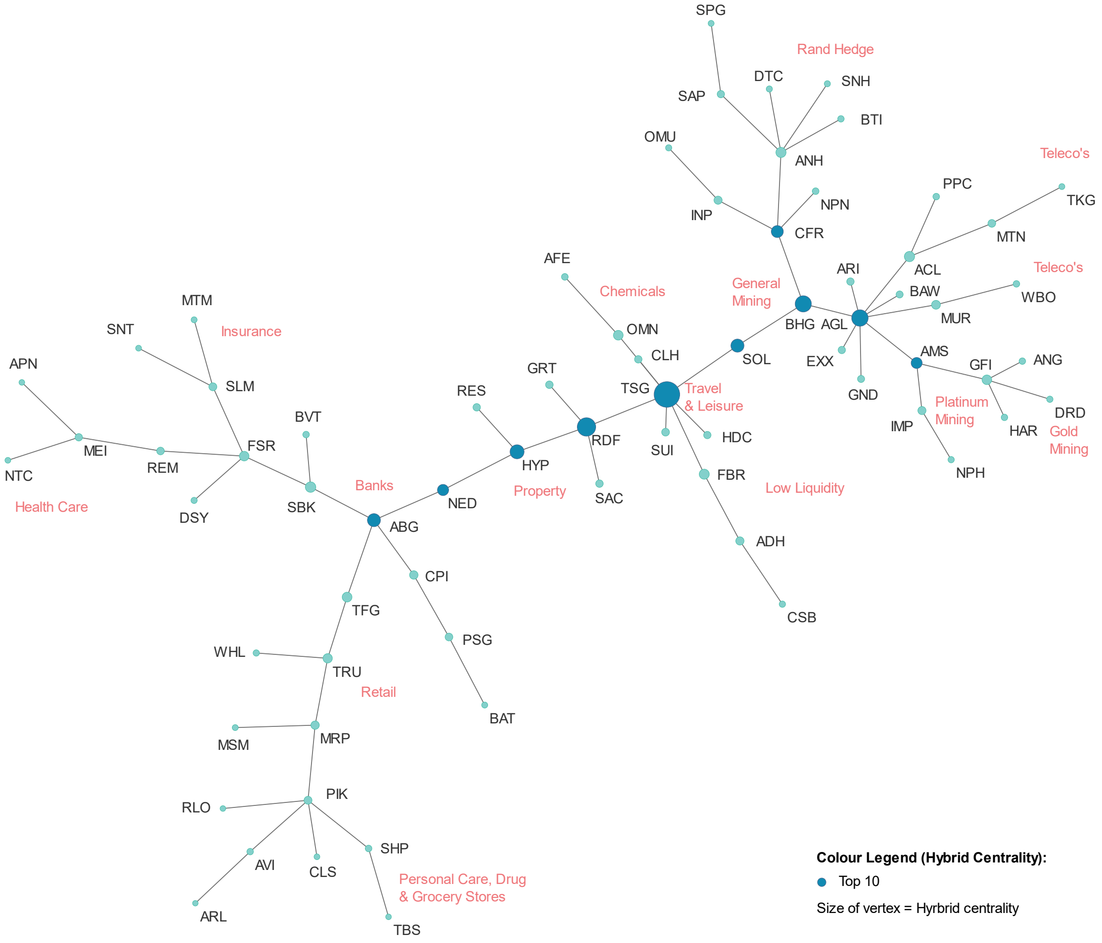


Figure 4.8 Long-Term MST: Hybrid Centrality Overlay

It is clear to see that TSG ranked highly in terms of centrality/importance (with the largest vertex). RDF, AGL, BHG, HYP, SOL, ABG, CFR, NED, AMS also featured among the top ten most central stocks (highlighted in the dark blue colour). One can also see the values of the subcomponents of the hybrid centrality metric in Figure 4.9. The benefit of using the hybrid approach is highlighted here with several shares scoring well on a few metrics, but poorly on others (for example see ABG).

TSG scored the highest in terms of betweenness, and closeness centrality. This implied that it was important in being able to connect vertices and was also close (on average) to other shares. AGL was a top scorer in terms of both degree and eigenvector centrality. It was therefore a well-connected share, but furthermore, it was connected to other important shares.

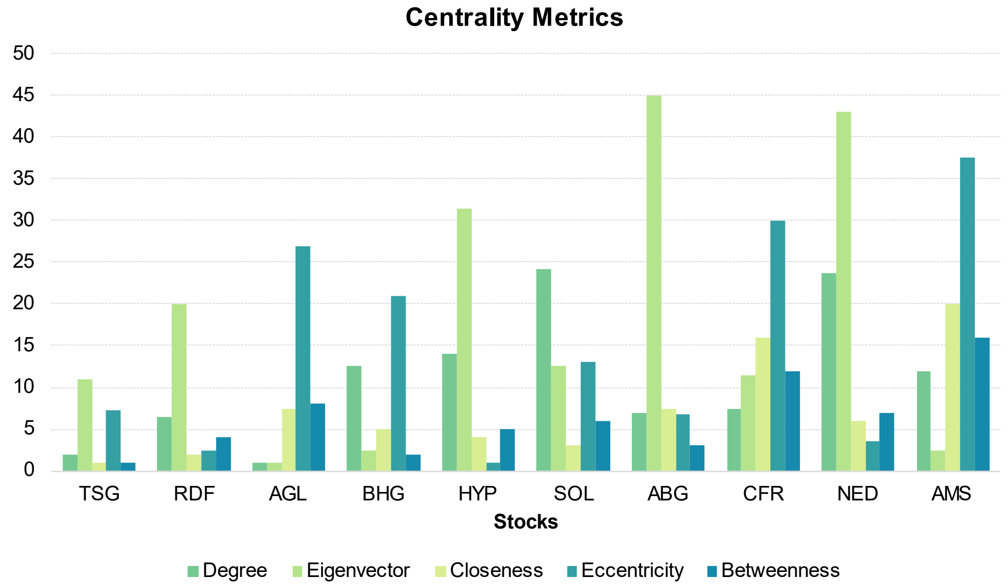


Figure 4.9 Long-Term MST: Centrality Metrics for Best Ranked Stocks by Hybrid Centrality

Curiously, there seemed to be a contradiction between TSG being a highly central vertex (as seen in Figure 4.8), and the fact that it fell into the less liquid group of shares (see Figure 4.7). Even though we had applied the QIS denoising procedure, there may still have been remnants of spurious relationships in the correlation matrix, which may have potentially been driven by the lower liquidity of certain shares (such as TSG). In the following section we report on the application of the bootstrapping methodology, which was used to assess the reliability of the structure of the MST (as discussed in Section 3.1.5).

4.2.5 Long-Term MST: Measuring Reliability

Having used the bootstrapping procedure described in Figure 3.4 to determine the reliability of each edge in the MST, we then calculated the average reliability for each share by averaging across all edges that are connected to that share's vertex. This average reliability factor was then overlaid onto the MST by varying the size of each vertex in accordance with this metric (with small vertices indicating shares that are part of less reliable edges). The vertices of the top third of the most reliable shares were shaded in a light blue colour. This is shown in Figure 4.10.

Examining the number of smaller vertices in the MST, one can see that there are several links in the MST that would have disappeared if the underlying data changed marginally (as is done in a bootstrapping procedure). In particular, TSG showed poorer reliability than one would have expected given its high ranking in terms of the centrality metrics.

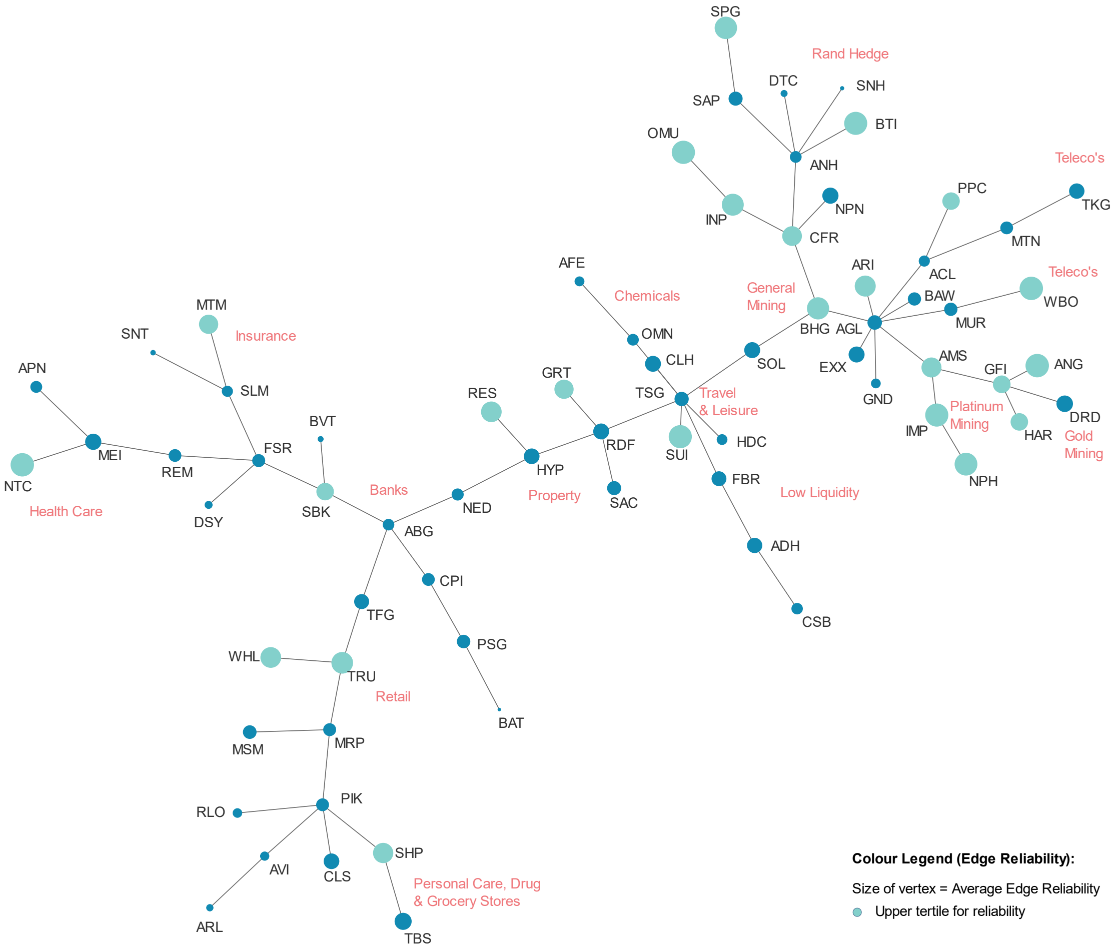


Figure 4.10 Long-Term MST: Average Edge Reliability Overlay

Examining the reliability of each of the seven edges that TSG formed a part of (which is shown in Table 4.1), one can see that many of the links had relatively low reliability (apart from the intra-sector link to SUI), occurring in only 46% to 65% all bootstrap samples. This confirmed our suspicion that TSGs importance in the network may have been overstated if one only considered the full period MST.

Table 4.1 Long-Term MST: Bootstrap Edge Reliability for TSG

Vertex 1	Vertex 2	Reliability
SOL	TSG	54.6%
RDF	TSG	62.4%
CLH	TSG	65.4%
SUI	TSG	98.7%
OMN	TSG	62.2%
FBR	TSG	46.0%
TSG	HDC	48.2%
Average		62.5%

One can also gain interesting insights by examining the relationship between correlation and edge reliability. This was done by plotting, for each edge in the MST, its related correlation (i.e. the correlation between the two shares that are linked by the edge) against the bootstrap reliability value of each edge. This is depicted in Figure 4.11.

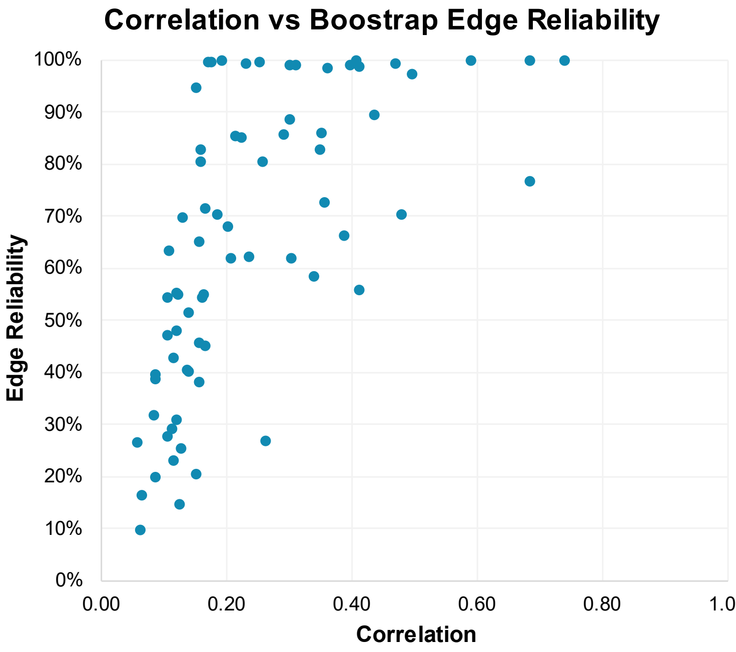


Figure 4.11 Long-Term MST: Bootstrap Edge Reliability vs Correlation

In terms of average reliability across the MST, approximately 24% of the edges in the MST had a reliability above 90%. Interestingly, not all of these edges were formed between highly correlated stocks (with some edges formed between shares that had correlations as low as 0.15). It is also noteworthy that higher correlations did not always imply a higher reliability of an edge in the MST. There were seven pairs of stocks (or 10% of all edges in the MST) that exhibited a correlation above 0.3 but had an edge reliability that was less than 80%.

One can also use the bootstrapping technique to measure the reliability of each share's importance in the MST. For each bootstrap sample the MST was created, and the hybrid centrality metrics were calculated. We then assessed the stability of each stock's importance in the MST by calculating the standard error of their hybrid centrality metrics across all the bootstrap samples. This is shown in Figure 4.10. Note that lower values of the hybrid centrality metric indicate more central and important shares.

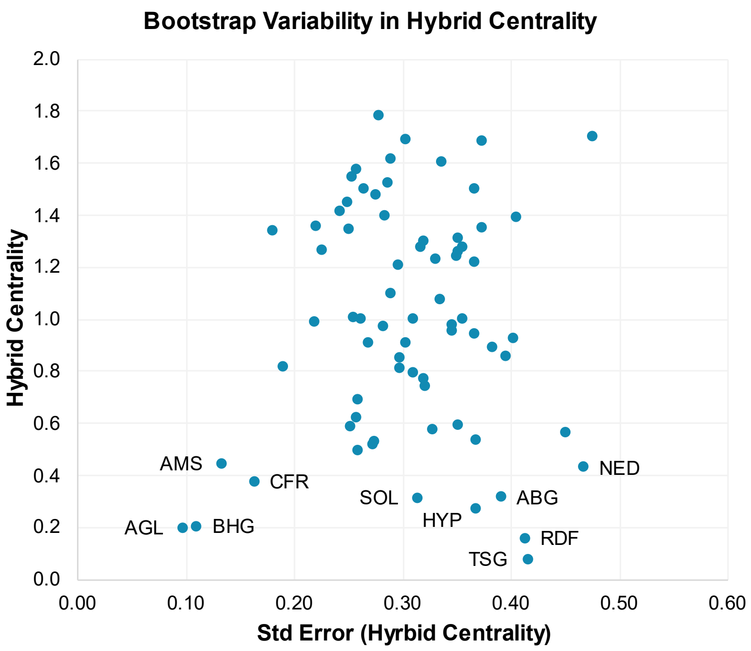


Figure 4.12 Long-Term MST: Centrality vs Bootstrap Variability

Stocks on the bottom-left of the chart (AGL, BHG, AMS, and CFR), are important shares with good centrality metrics, and their centrality metrics are also highly reliable (i.e. they do not fluctuate significantly as the data changes). Shares on the bottom-right of the chart (TSG, RDF, HYP, ABG, and NED), are important shares with good centrality metrics, but their centrality metrics are also less reliable, with the importance of these shares varying significantly across the bootstrap samples. One can again see that TSG fell into the category of highly central, but less reliable shares.

4.2.6 Dynamic MSTs

We now proceed to the analysis in which we examined the topology of the dynamic MSTs through time. In Figure 4.13 we report the Normalised Tree Lengths (NTLs) for the dynamic MSTs that were calculated at each point in time. The most noteworthy point of this figure is that the MST tended to shrink during market crashes (such as the financial crisis in 2008 and the COVID-19 crash in 2020). This implied that the behaviour of shares became more similar over these periods (with the distance between shares shrinking). This was intuitive, as the distance metric that was used for the analysis was an inverse of the correlation coefficient, and it is well established that correlations tend to rise in market crashes. However, it is worthwhile noting that the shrinking of the MST during market crashes was not due to a general market effect, as we have removed the market mode from the data prior to performing any analysis. It was due to an increasing similarity of stocks that was independent of a general market effect.

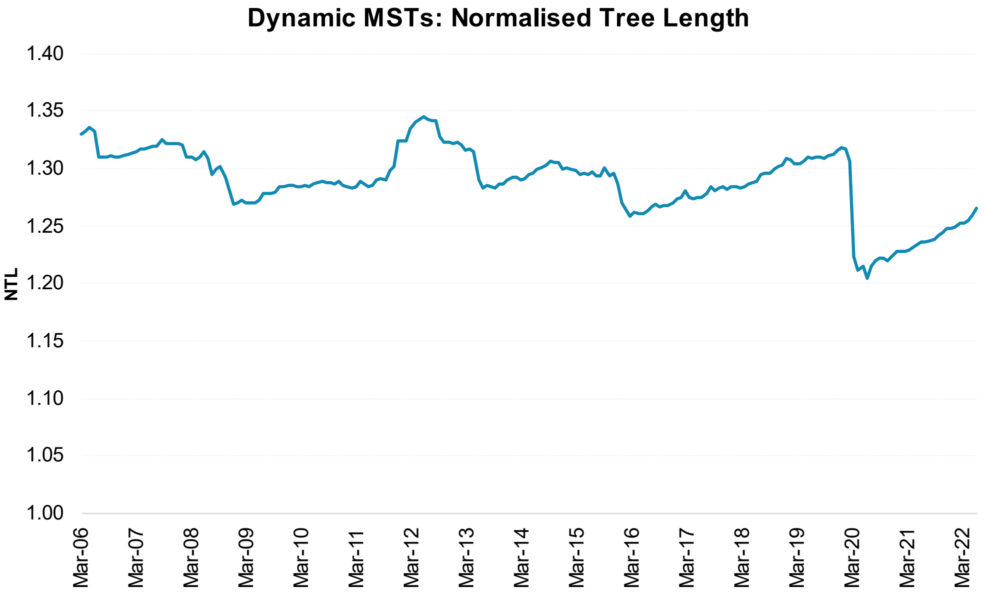


Figure 4.13 Dynamic MSTs: Normalised Tree Length

In the prior section, we calculated the hybrid centrality metrics for the long-term, static, MST, and we saw that TSG, RDF, AGL, BHG, HYP, SOL, ABG, CFR, NED, and AMS featured in the list of the top ten of the most important (or central) stocks in the MST. However, subsequent analysis from a liquidity and reliability point of view did cast some doubt over the role of some stocks, and in particular TSG. A dynamic analysis was used to shed further light on this contradiction by determining which stocks were ranked consistently well in terms of hybrid centrality on the dynamic MSTs. For each of the dynamic MSTs, we calculated the hybrid centrality of each stock and then ranked them from best to worst. We then ranked the stocks according to the amount of time each one of them spent as one of the top five most central stocks. These results are shown in Table 4.2, with TSG included at the bottom of the table for comparative purposes.

Table 4.2 Dynamic MSTs: Consistency in Hybrid Centrality

Share	ICB Super-sector	% Time in Top 5
AGL	Basic Materials	69.5%
BHG	Basic Materials	64.8%
AMS	Basic Materials	23.9%
OMU	Financials	21.1%
TFG	Consumer Discretionary	18.3%
ARI	Basic Materials	17.4%
ANG	Basic Materials	17.4%
BTI	Consumer Staples	16.9%
CFR	Consumer Discretionary	16.9%
IMP	Basic Materials	15.0%
ANH	Consumer Staples	15.0%
TSG	Consumer Discretionary	4.69%

From Table 4.2 one can see that AGL and BHG were consistently ranked highly in terms of their centrality (having respectively spent 69% and 65% of the time as one of the five most central stocks). This was reassuring as AGL and BHG are shares that have a large market capitalisation on the JSE and are often deemed to be important. Shares from the Basic Materials sector dominated the list, as well as Consumer stocks (TFG, BTI, CFR, and ANH). In contrast to these stocks, TSG spent less than 5% of its time amongst the five most central stocks, when considering the dynamic MSTs. This once again highlighted that the influence of TSG on the long-term MST may be overstated.

It is interesting that the ranking of the top three stocks from Table 4.2 (AGL, BHG, and AMS), supported the results from the bootstrap reliability estimates that we saw in Figure 4.12.

Lastly, Figure 4.14 shows which ICB industries were ranked as number one (the chart on the top) and number two (the chart on the bottom) in terms of the average hybrid centrality of the shares within them. Note that for this analysis, Energy was combined with Basic Materials, as there was only one stock (EXX) within the Energy industry. Examining Figure 4.14 one can see the dominance of the Basic Materials sector (in grey) in terms of importance/centrality. However, there were periods when Real Estate (light blue) was the most central industry, as well as Financials (in yellow). Financial stocks were very central after the financial crisis in 2008, being ranked first or second over the bull market from 2009 to 2011 which was driven by monetary easing from central banks.

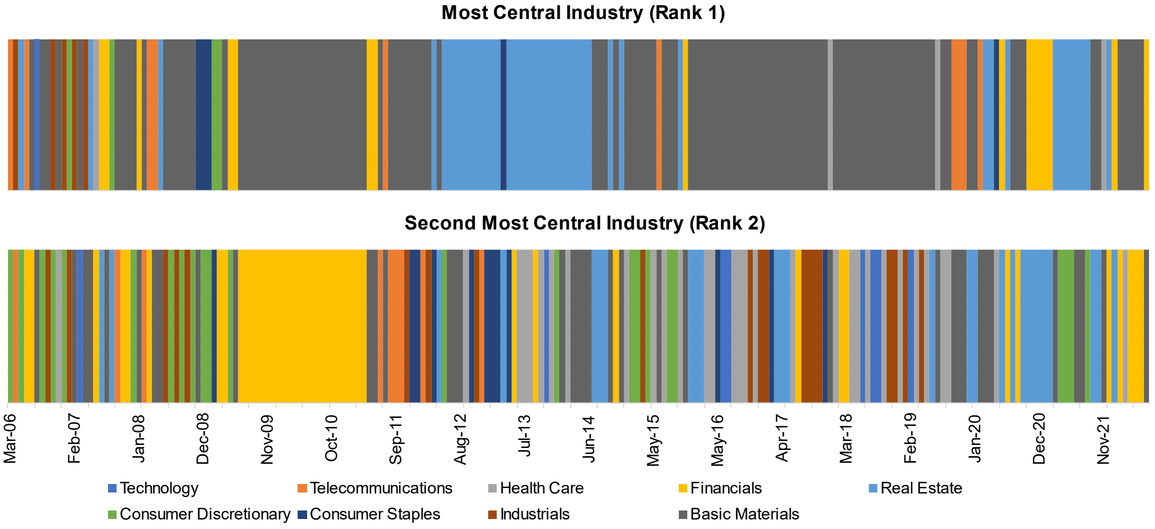


Figure 4.14 Dynamic MSTs: Top Two Most Central Industries

4.3 Analytic Results for the PMFG

This section contains the results of the application of the PMFG to the South African stock market, on both a long-term basis, as well as from a dynamic point of view. As discussed in Section 2.12.4, to our knowledge this is the first time this technique has been applied to the South African stock market.

4.3.1 Long-Term PMFG: Sector Overlay

Figure 4.15 depicts the long-term PMFG for South African stocks based on the correlation matrix. The visual nature of the PMFG was enhanced by overlaying the ICB super-sector classification of each share.

Note that for the sake of brevity, we did not repeat the currency exposure overlay, and the liquidity overlay, that was done for the MST. However, we did use the results from those MSTs (Figure 4.6 and Figure 4.7) to assist in labelling how shares have grouped together on the PMFG.

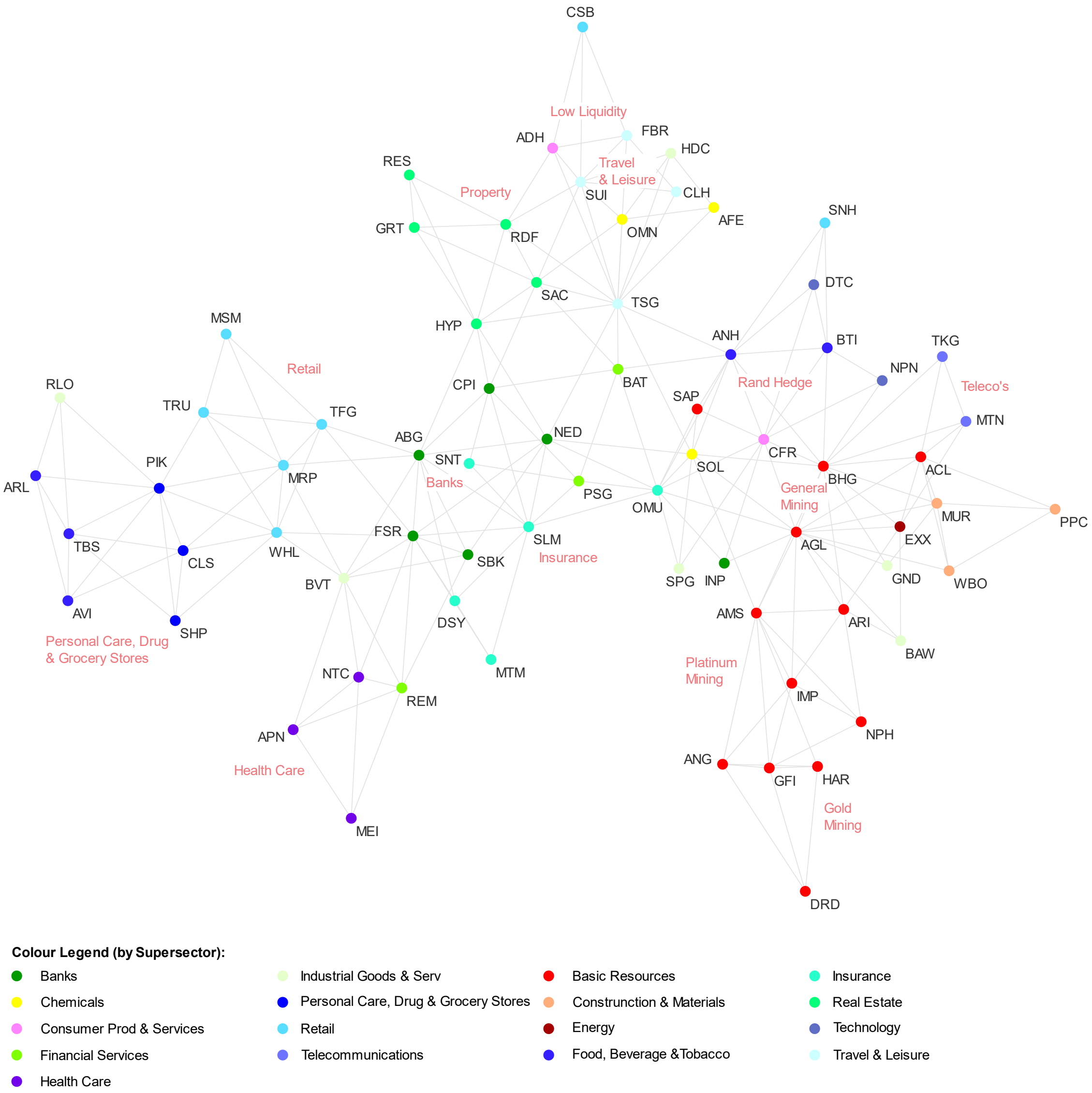


Figure 4.15 Long-Term PMFG: Super-sector Overlay

It was noteworthy to see that the grouping of shares on the PMFG was similar to that of the MST. However, given that the MST is a subgraph of the PMFG, one would not have expected large differences between the two methods. There was still a large amount of grouping by ICB sector on the PMFG, but the Rand Hedge group, as well as the low liquidity group, also featured.

The connectivity of shares as compared to MST was richer (by design), allowing shares such as WHL, FSR, AMS to form part of multiple cliques, as compared to the MST which tended to have a *winner takes all* outcome (i.e. only a few shares featured in multiple connections). The downside of this richer dataset was that the PMFG was visually “noisier” due to the additional links. Fortunately, the PMFG is accompanied by the DBHT, which assists in clustering similar stocks together in a fully unsupervised manner. The long-term DBHT is analysed in greater detail in Section 4.4.

4.3.2 Long-Term PMFG: Hybrid Centrality Overlay

In Figure 4.16 we depict the PMFG with an overlay of the hybrid centrality metric of Pozzi, Di Matteo and Aste (2013). This metric we used to obtain a better understanding of the importance of the shares on the PMFG.

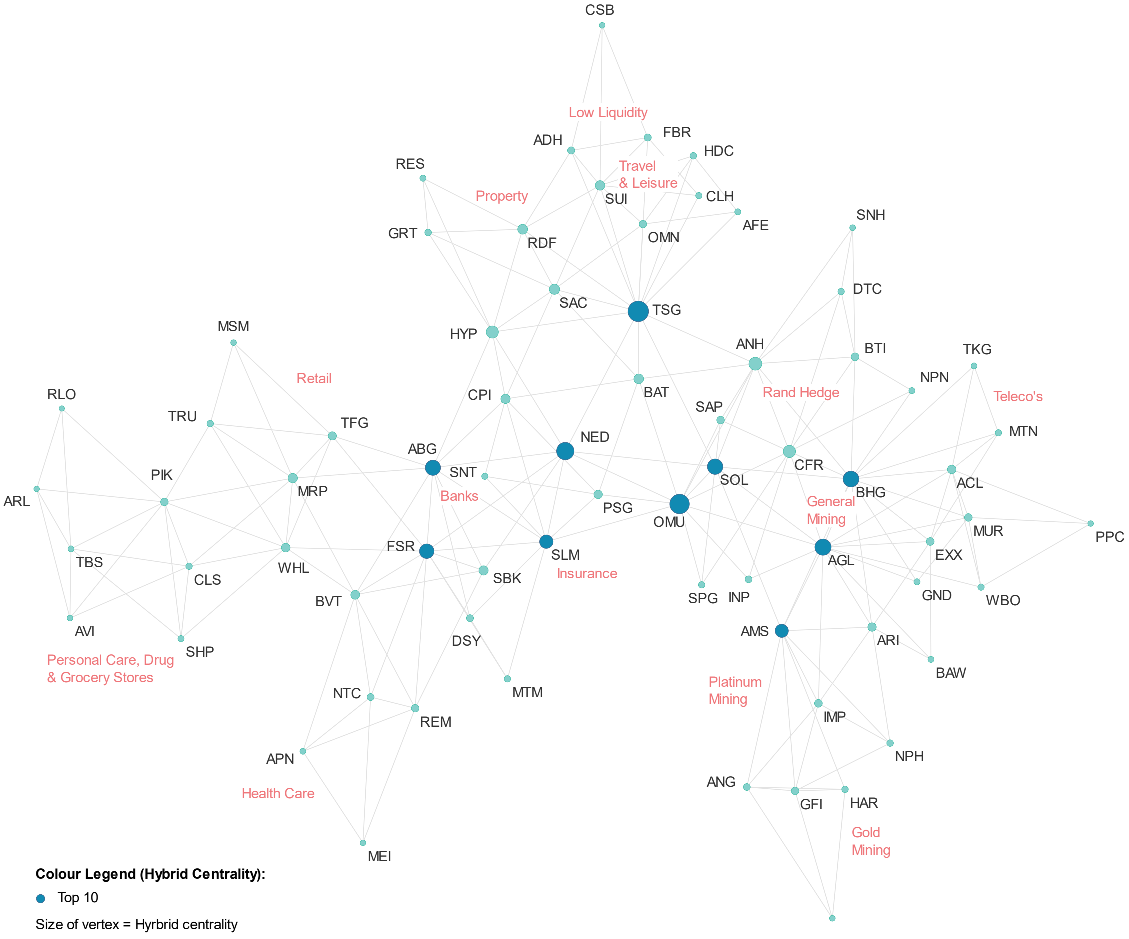


Figure 4.16 Long-Term PMFG: Hybrid Centrality Overlay

TSG again ranked highly in terms of centrality/importance (with the largest vertex). OMU, NED, AGL, BHG, SOL, ABG, FSR, SLM, and AMS also featured in the top ten stocks when ranked by the hybrid centrality metric. One can see the values of the subcomponents of the hybrid centrality metric in Figure 4.17. NED was located more centrally in the PMFG (as compared to the PMFG), with good rankings for betweenness, eccentricity, and closeness centrality, but had fewer connections (with lower rankings for degree centrality, and eigenvector centrality). AGL demonstrated the opposite behaviour, being well-connected with many connections (and therefore had good rankings in terms of degree, and eigenvector centrality), but poorer ranks in terms of betweenness, eccentricity, and closeness centrality.

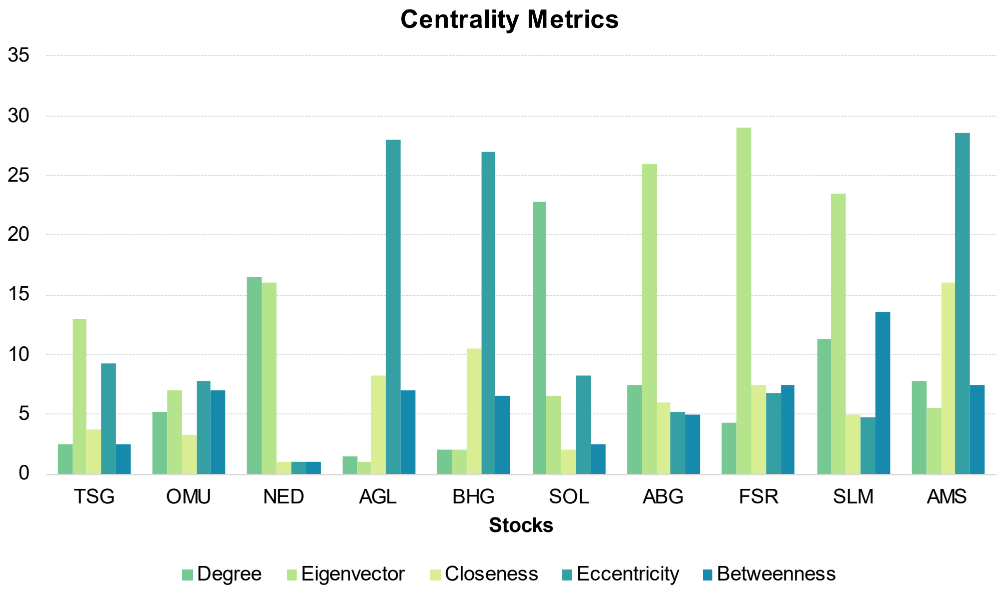


Figure 4.17 Long-Term PMFG: Centrality Metrics for Best Ranked Stocks by Hybrid Centrality

4.3.3 Long-Term PMFG: Measuring Reliability

Using the bootstrapping procedure described in Figure 3.4 we determined the average reliability factor of each vertex in the PMFG (as was done for the MST in Section 4.2.5). This average reliability factor was then overlaid onto the PMFG by varying the size of each vertex in accordance with this metric (with smaller vertices indicating shares that were part of less reliable edges). The vertices of the top third of the most reliable shares were shaded in the light blue colour. These results can be seen in Figure 4.18.

One can see that certain groups of shares (and industries) tended to have better reliability as compared to others (these can be seen as groups of shares with larger vertices). For example, the Mining, Consumer-related, and Real Estate sectors tended to have higher reliability. Similar to the MST, TSG had a lower reliability than its centrality metrics would have indicated. NED also fell into this category with good centrality metrics, but lower reliability scores.

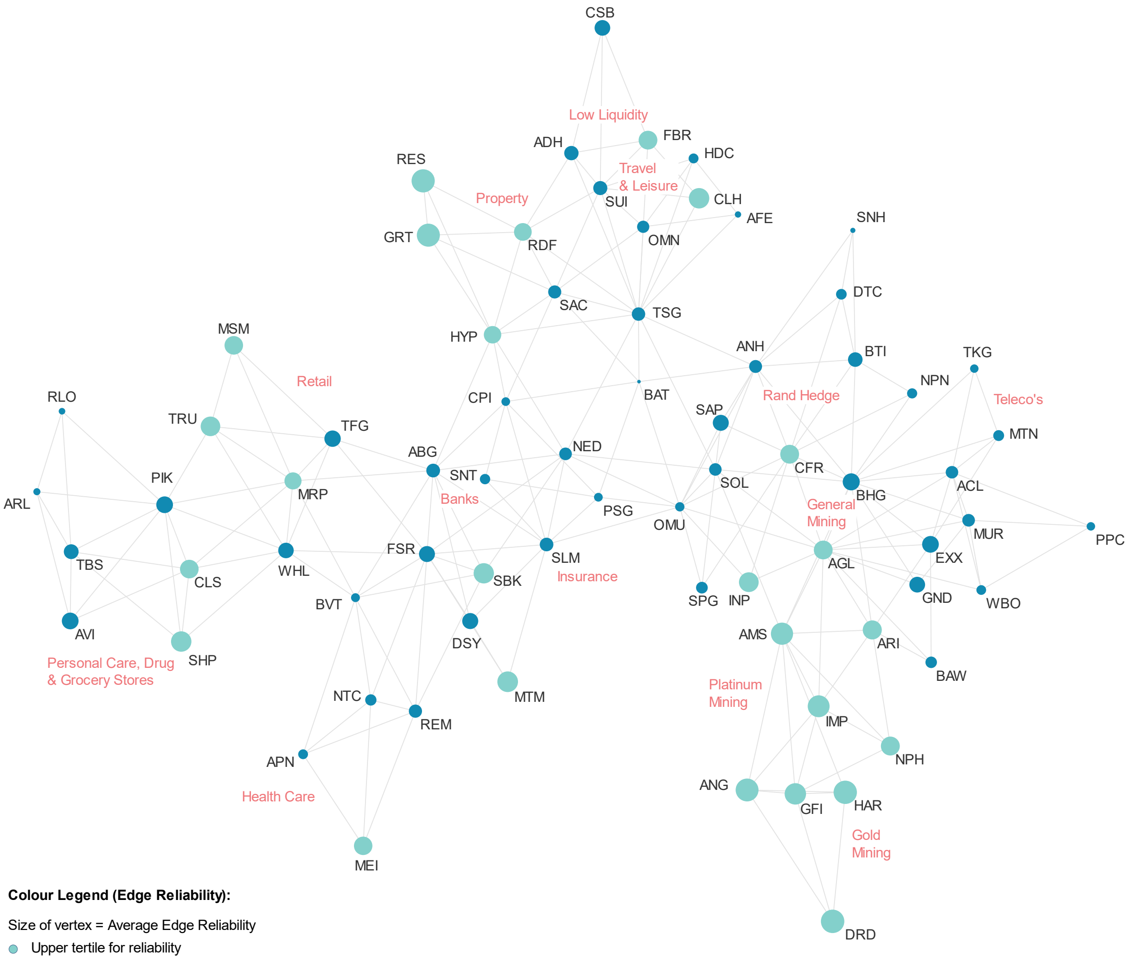


Figure 4.18 Long-Term PMFG: Average Edge Reliability Overlay

Continuing with the reliability analysis, in Figure 4.19 we plotted the correlation associated with each edge in PMFG against its bootstrap reliability.

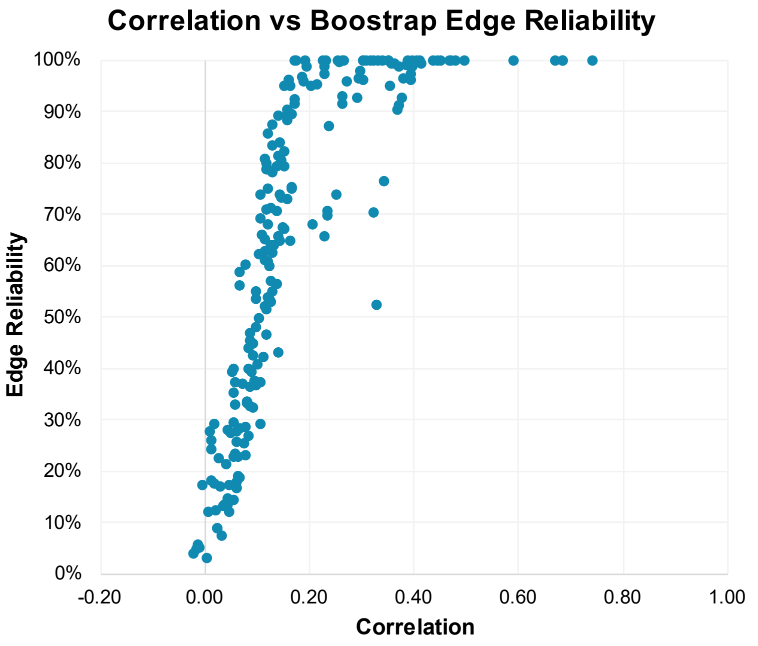


Figure 4.19 Long-Term PMFG: Bootstrap Edge Reliability vs Correlation

In terms of average reliability across the PMFG, approximately 32% of the edges in the PMFG had a reliability score above 90%. This was higher as compared to a value of 24% for the MST. Furthermore, there were only three pairs of stocks (or 1% of all edges in the PMFG) that exhibited a correlation above 0.3, but had an edge reliability that was less than 80% (as compared to the 10% of the edges in the MST). This implied that for the PMFG there was a stronger link between the level of the correlation between two shares and the reliability of the edge between them (as compared to the MST). These results suggested that the *structure of the PMFG is more robust as compared to the MST.*

The results of using the bootstrapping technique to evaluate the reliability of the centrality metrics for the PMFG are shown in Figure 4.20. Similar to the MST, the following shares, AGL, BHG, AMS, and CFR, seemed to be important shares with good centrality metrics, and they stayed reliably central in the PMFG across all of the bootstrap samples. TSG and NED fell into the category of highly central, but less reliable shares (on the bottom-right).

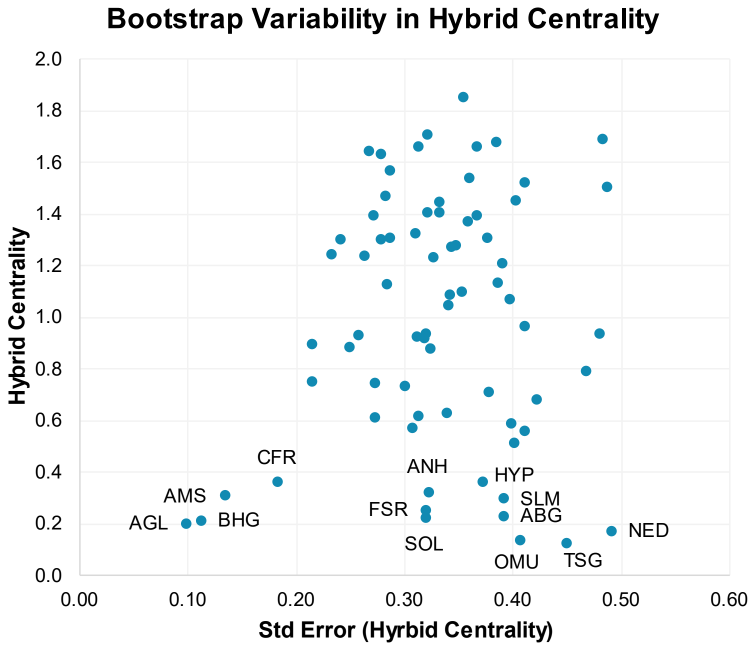


Figure 4.20 Long-Term PMFG: Centrality vs Bootstrap Variability

4.3.4 Dynamic PMFGs

We now report on the results of the dynamic analysis in which we analysed the topology of the PMFG through time. We begin by portraying the NTLs of the dynamic MSTs and PMFGs in Figure 4.21. Note that the NTL for the PMFG will generally be higher than that of the MST. This trait is due to the design of the MST which is constructed to have a globally minimum total distance, while this is not a requirement for the PMFG.

There was a strong correlation between the movement of the NTLs of the dynamic MSTs (which is shown in the dark blue line), and the NTLs of the dynamic PMFGs (shown in light blue). Similar to the MST the PMFG tended to shrink during market crashes such as the financial crisis in 2008 and the COVID-19 crash in 2020. This highlighted a strong correlation in the overall topology of the MST and PMFG.

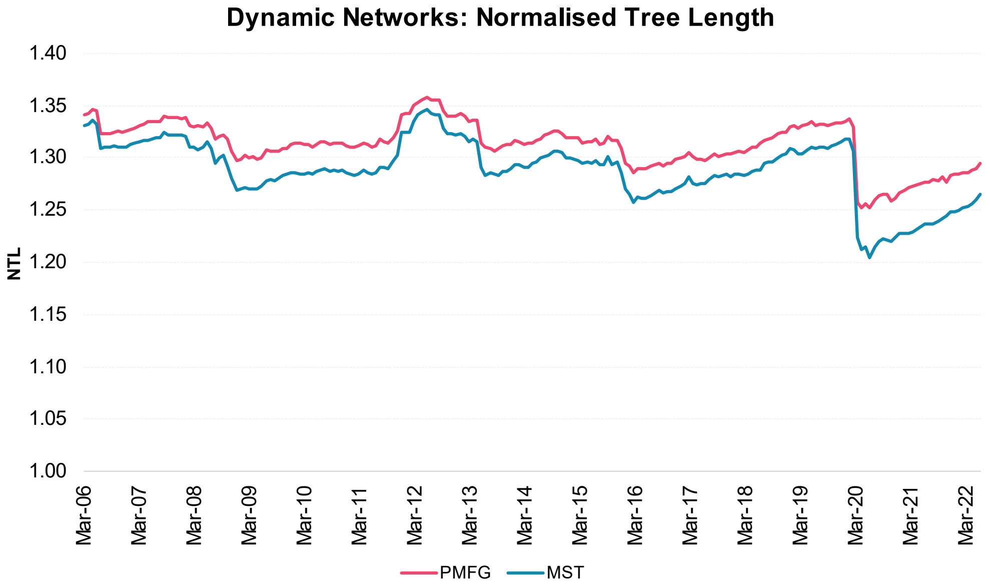


Figure 4.21 Dynamic PMFGs: Normalised Tree Length

In Table 4.3 we show which stocks were the most consistently ranked as the five most central stocks at each point in time, as calculated from the dynamic PMFGs. We have once again included TSG at the bottom of the table for comparative purposes.

Table 4.3 Dynamic PMFGs: Consistency in Hybrid Centrality

Share	ICB Super-sector	% Time in Top 5
AGL	Basic Materials	75.1%
BHG	Basic Materials	60.6%
AMS	Basic Materials	32.9%
CFR	Consumer Discretionary	25.4%
ANH	Consumer Staples	19.2%
IMP	Basic Materials	16.9%
GFI	Basic Materials	16.4%
OMU	Financials	15.5%
SOL	Basic Materials	15.5%
ARI	Basic Materials	14.1%
TSG	Consumer Discretionary	0.9%

The standout shares from Table 4.3 were once again AGL, BHG, and AMS (as seen for the dynamic MSTs in Table 4.2). In contrast to these stocks, TSG spent less than 1% of its time amongst the five most central stocks when considering these dynamic PMFGs. This once again highlighted that the influence of TSG on the long-term PMFG may have been overstated.

Finally, Figure 4.22 shows which ICB industries were ranked in the top two in terms of the average hybrid centrality of the shares within each industry. The Energy industry was again combined with Basic Materials, as there was only one stock in that industry.

The results were similar to those of the MST, and one can again see the dominance of the Basic Materials sector in terms of centrality in the PMFG, with periods when Real Estate or Financial stocks were the most central.

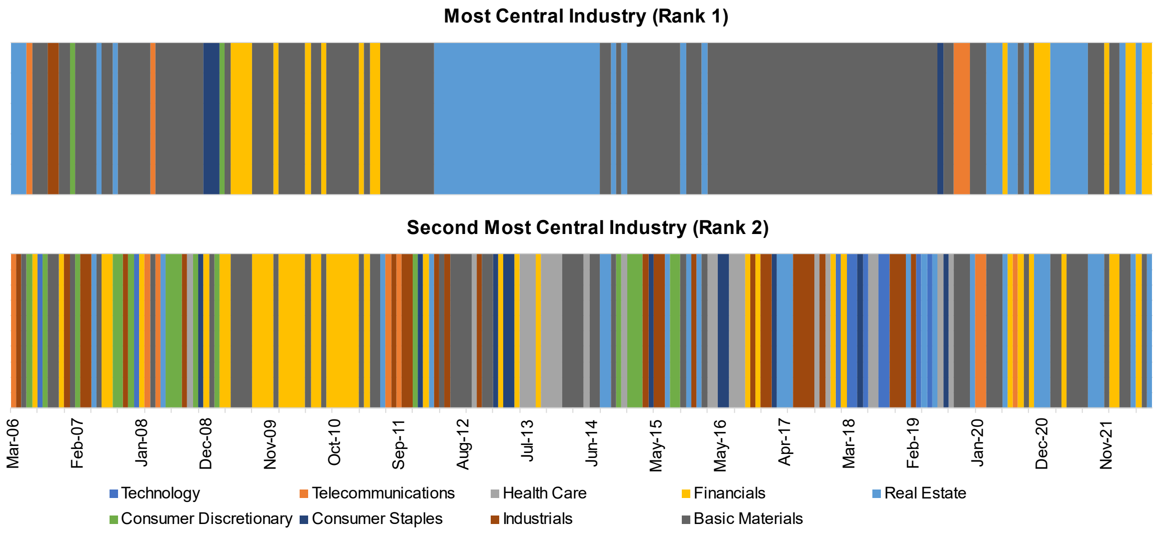


Figure 4.22 Dynamic PMFGs: Top Two Most Central Industries

4.4 Analytic Results for the DBHT

This section contains the results of the application of the DBHT to the South African stock market, on both a long-term basis, as well as from a dynamic point of view. We show the long-term dendrogram of the DBHT, followed by an analysis in which we determined the amount of economic information that was extracted by the DBHT. We also compared the DBHT to other hierarchical clustering methods, and then used a bootstrapping technique to assess the robustness of the estimated number of clusters that emerged from the long-term DBHT.

In the dynamic setting, we determined if a changing market environment had an impact on the estimated number of clusters, as well as on the quantity of economic information that was extracted using the DBHT.

4.4.1 Long-Term DBHT: Sector Overlay

Figure 4.23 depicts the dendrogram of the DBHT for South African stocks that was based on the long-term correlation matrix. Each stock had its associated ICB industry, as well as its cluster number attached to each label. The varying colours of the lines in the dendrogram were used to highlight the various clusters that were automatically extracted.

The DBHT extracted six clusters from the long-term PMFG. Note that the emergence of these clusters was fully unsupervised and as such, it was contingent upon us to interpret them. Given that we have seen that the shares on the MST and PMFG tended to group together in accordance with their economic sectors, we turned to the ICB classifications as a starting point to assist us in labelling the clusters. In Figure 4.24 we show the overlap between stocks in the six clusters and the ICB industry classifications, while Figure 4.25 depicts the overlap between stocks in the six clusters and the ICB super-sector classification.

Figure 4.26 depicts a Sankey chart showing the flow of the allocation of shares from the six clusters (in the middle of the chart) to the ICB industry classification (on the left), and from the six clusters (in the middle) to the ICB super-sector classification (on the right).

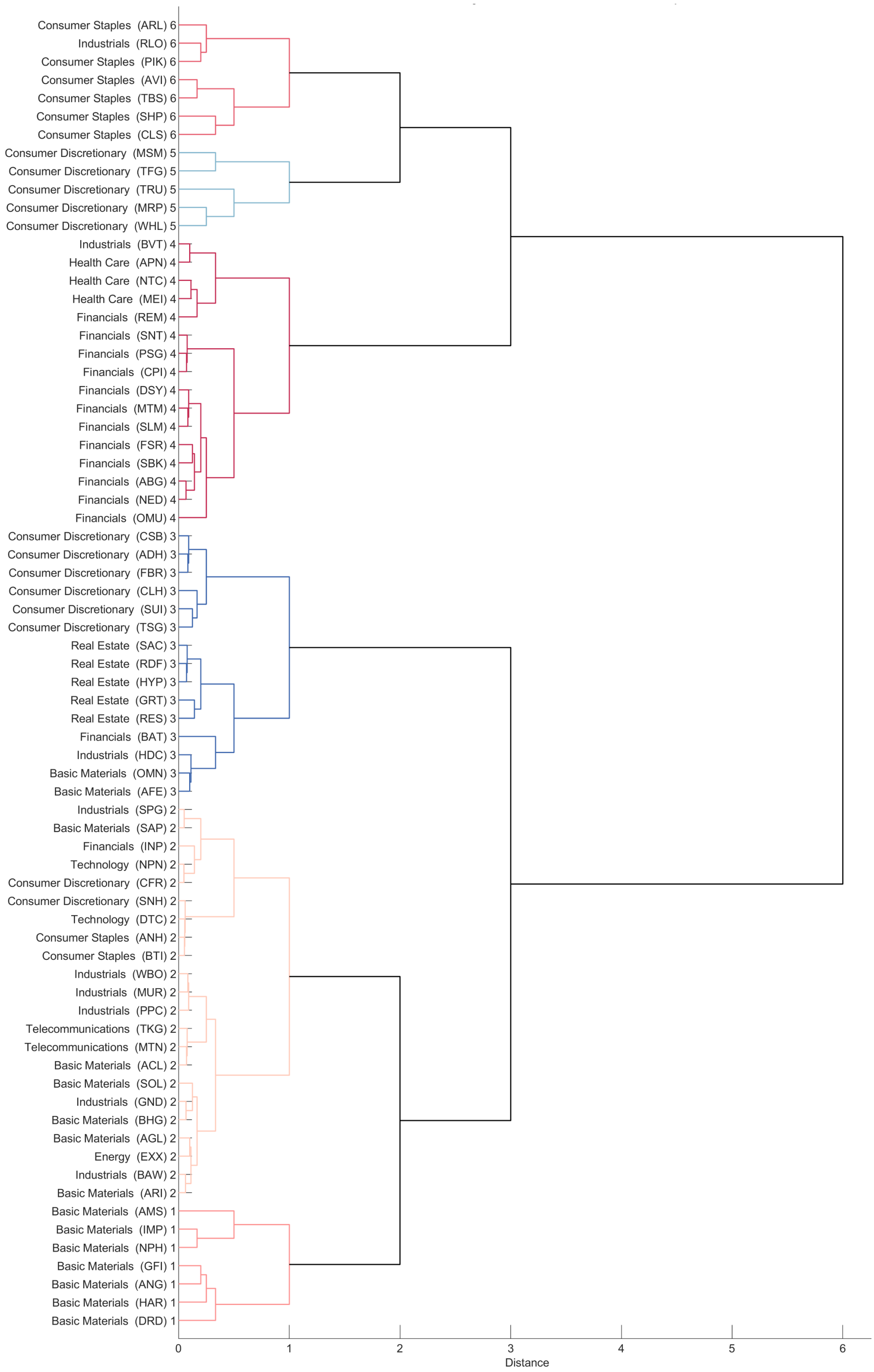


Figure 4.23 Long-Term DBHT: Dendrogram - Six Clusters Highlighted

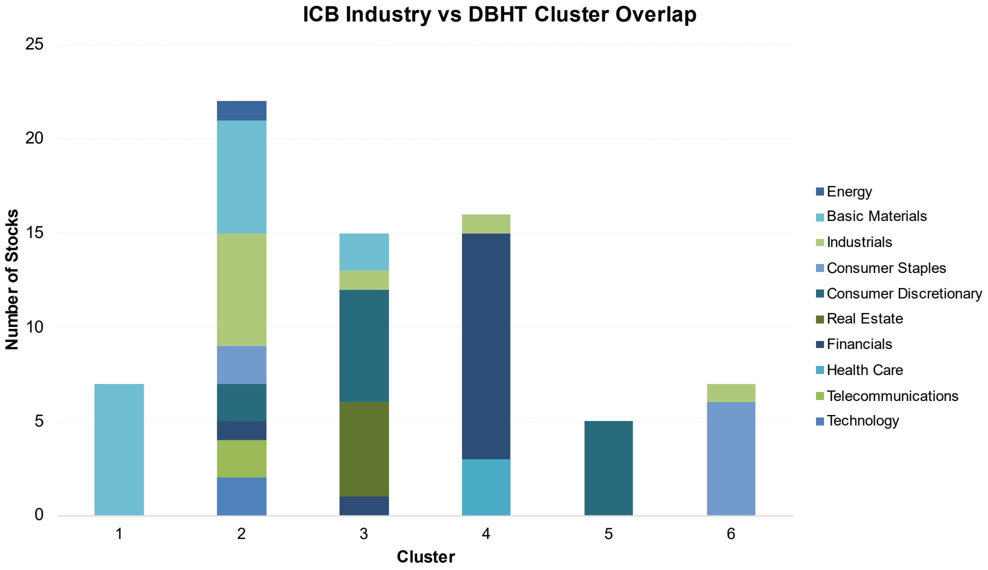


Figure 4.24 Long-Term DBHT: ICB Industry Overlap with Clusters

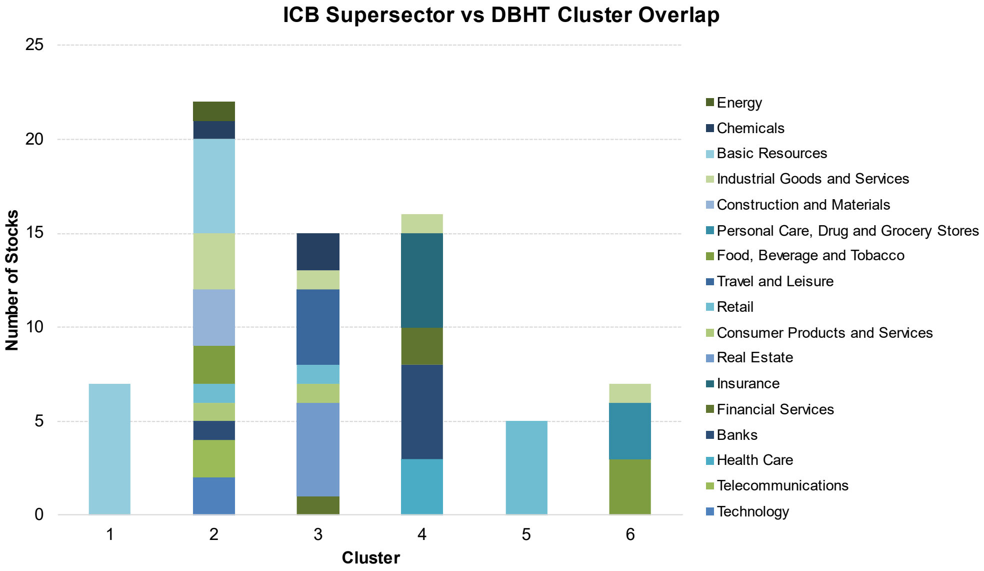


Figure 4.25 Long-Term DBHT: ICB Super-sector Overlap with Clusters

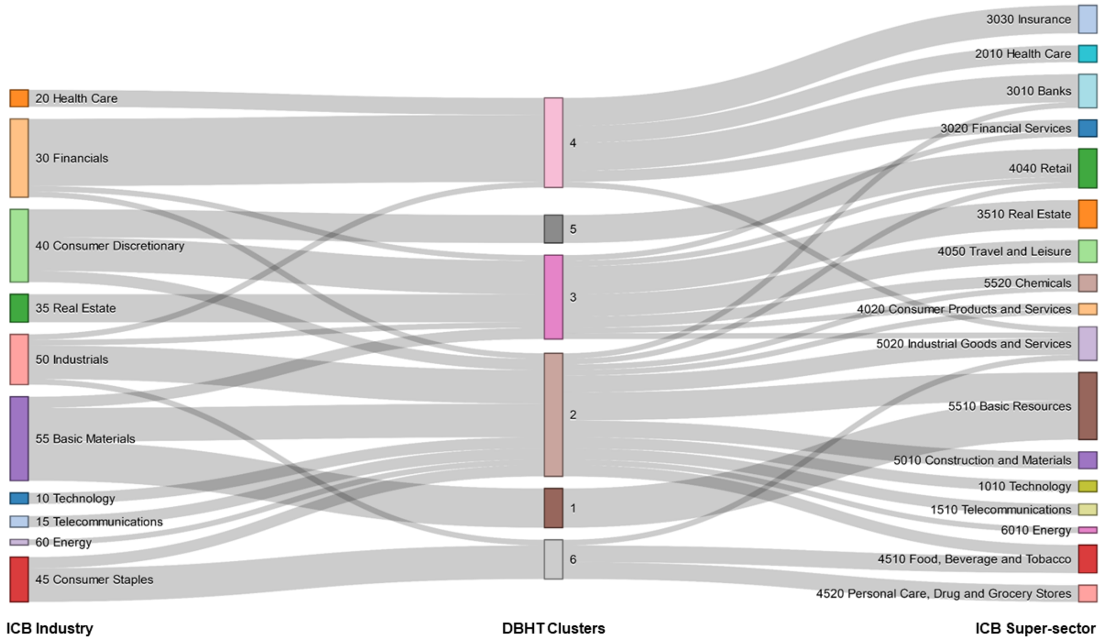


Figure 4.26 Long-Term DBHT: Sankey chart - ICB classification vs Clusters

Having analysed Figure 4.24, Figure 4.25, and Figure 4.26 one can see that **Cluster 1** consisted of stocks from the Basic Materials industry, and in particular stocks from the Gold Mining, and Platinum Mining sectors, while **Cluster 5** contained stocks from the Consumer Discretionary/Retail sector. **Cluster 6** typically contained stocks from the Consumer Staples industry. **Cluster 3** seemed to be a mix of shares from the Consumer Discretionary and Real Estate industries, however, there was also a strong overlap with the low liquidity effect that was outlined in Section 4.2.3. **Cluster 4** certainly had a strong Financial influence with shares from both the Banks and Insurance super-sectors featuring strongly, but it also contained shares from the Health Care sector. And finally, although **Cluster 2** contained shares from a multitude of industries and sectors, there did seem to be some overlap between this cluster and the Rand Hedge effect that was discussed in Section 4.2.2.

Overall, this led to the following broad labelling of the six clusters:

- Cluster 1 – Mining
- Cluster 2 – Rand Hedge
- Cluster 3 – Consumer Discretionary and Real Estate (Low Liquidity)
- Cluster 4 – Financials and Health Care
- Cluster 5 - Consumer Discretionary/Retail
- Cluster 6 - Consumer Staples

In Figure 4.27 we overlayed the clusters from the DBHT onto the long-term PMFG, by varying the colour of each vertex contingent upon which cluster it belonged to. As expected, the clusters consisted of shares that were grouped close together on the PMFG. The DBHT method automatically chooses the points on the PMFG at which to separate the shares.

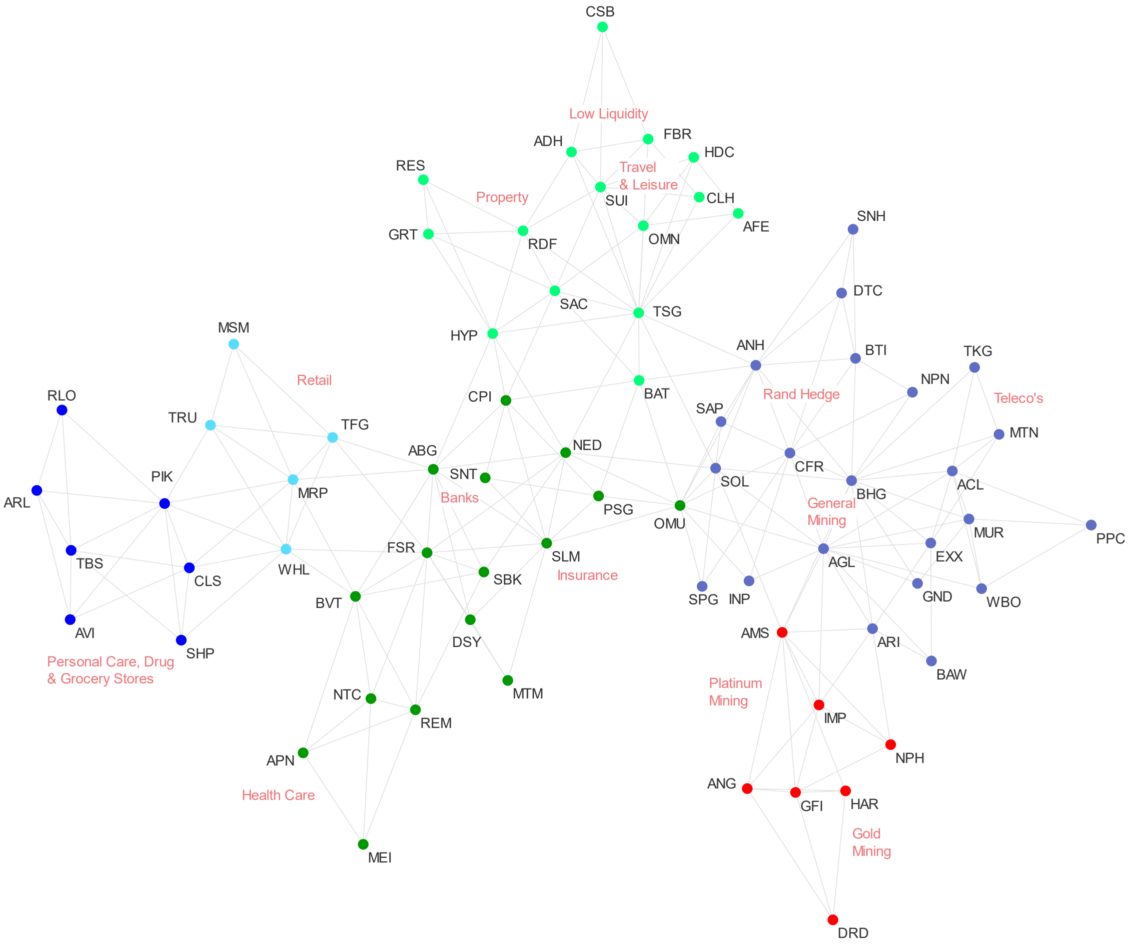


Figure 4.27 Long-Term PMFG: DBHT Cluster Overlay

4.4.2 Long-Term DBHT: Quantifying the Extracted Economic Information

As discussed in Section 3.2.8, one can quantify the overall amount of economic information that is extracted from any clustering technique using the Adjusted Rand Index (ARI). For this dissertation, this was done by assuming that one of the partitions came from the DBHT clusters, while the second partition came from the various ICB classifications. In Figure 4.28 we used the ARI to determine the overlap between each cluster and a specific ICB classification (such as the industry, super-sector, or sector level classification), as we varied the number of clusters that were extracted from the DBHT clustering. By varying the number of clusters (from two to twenty) we were also able to determine if the optimal number of clusters that emerged unsupervised from the DBHT, extracted the maximum amount of economic information. Note that the optimal number of clusters that was extracted automatically from the DBHT (six) is shown by the vertical red line on the chart.

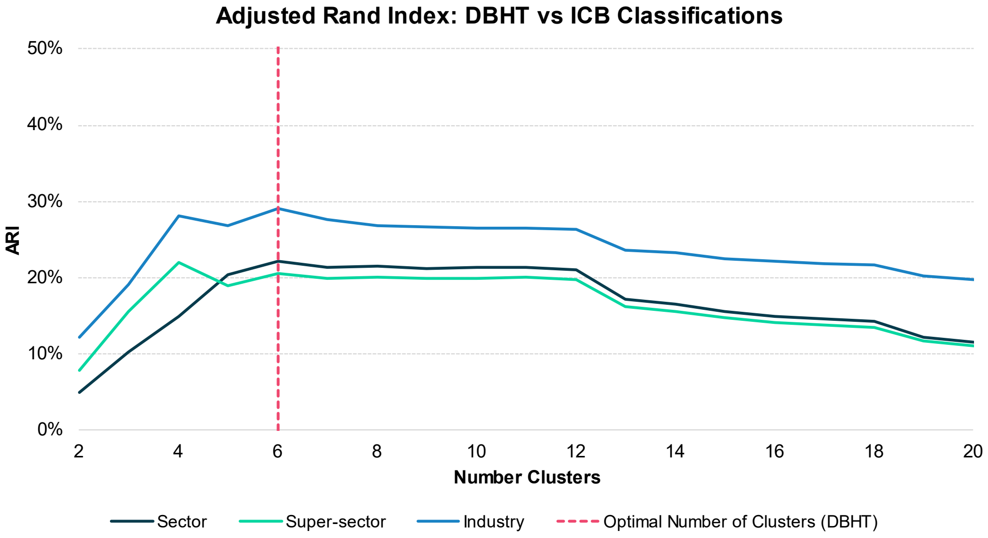


Figure 4.28 Long-Term DBHT: ARI vs ICB classifications

One can see that the DBHT overlapped the ICB classification by 20% to 30% (at the peak of the ARI) depending on the classification that was used. The most overlap came from the ICB industry level classification, which consists of ten categories. It was comforting to see that ARI had a peak (for the industry level ICB classification) near the optimal number of clusters. These results were marginally lower than Musmeci, Aste and Matteo (2015) who found a maximum ARI of 40% for stocks on the NYSE. However, that analysis consisted of a significantly larger universe of approximately 340 stocks.

It should be noted though, that using the ARI, in conjunction with a qualitative classification scheme such as the ICB, as a metric to benchmark the amount of economic information extracted from a specific clustering technique, does have some shortcomings. Firstly, it is reliant on the accuracy of the qualitative classification partitioning, but secondly, a company’s classification can remain the same for a prolonged period of time even if the business model of a company changes. Furthermore, if market participants have a view of a stock that is different to the qualitative classification scheme, they may trade that stock differently as compared to other stocks in the same category. This would lead to a clustering for that stock that is different from the qualitative classification, especially if the clustering was performed using a correlation-based distance metric that was calculated from price changes (given that share price movements reflect the views that market participants have of a share). In the South African stock market, the so-called Rand Hedge stocks (which were discussed in Section 4.2.2), which are shares that react positively to a weakening in the South African Rand, are a group of shares that fall into this category. These shares come from a diverse set of ICB industries and sectors, but a larger proportion of the movements in their share price can be explained by movements in the currency, as compared to their respective industry/sector (although this can vary over time). The group of less liquid shares (that was discussed in Section 4.2.3) also falls into this category, but there was also strong sectoral overlap for that group of shares.

These shortcomings highlight the difficulty in determining the accuracy of a clustering or network filtering technique. Often domain/industry knowledge is required to assess whether or not the outcomes from such techniques are sensible. With these caveats in mind, in the following section, we compared the clustering outcomes of the DBHT technique against the other popular linkage methods that were described in Section 2.11.4, and in particular, in Table 2.11.

4.4.3 Long-Term DBHT: A Comparison to Alternative Linkage Methods

We specifically focused on the SLCA, ALCA, and Ward's linkages, and compared these methods to the DBHT using the ARI. Similar to Figure 4.28, we calculated the ARI as the number of clusters was varied from two up to twenty. The optimal number of clusters that was extracted from the DBHT (six) is again shown by the vertical red line on the chart. Figure 4.29 depicts the ARI calculated against the ICB sector level classification, while Figure 4.30 depicts the ARI calculated against the higher ICB industry level classification.

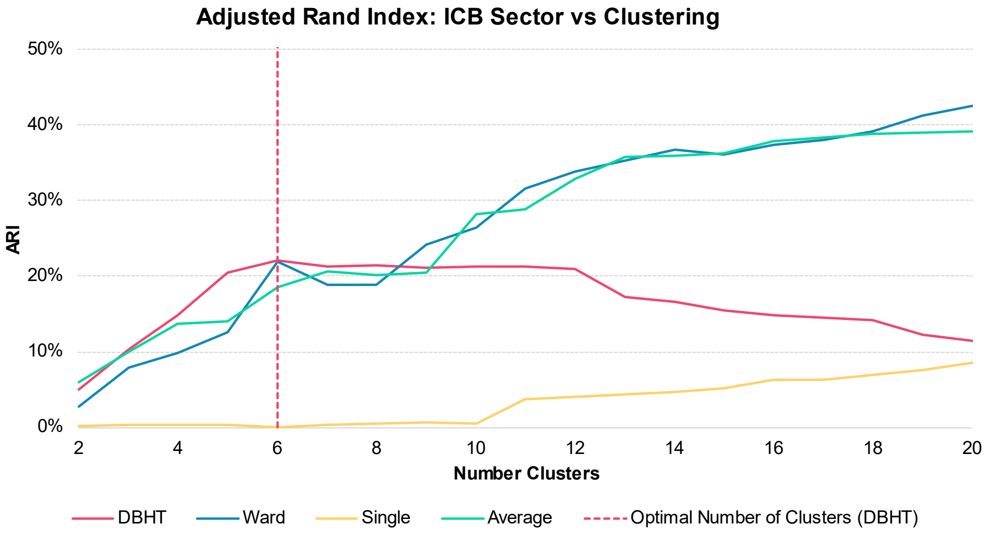


Figure 4.29 Long-Term DBHT vs Alternatives: ARI with ICB sectors

From Figure 4.29, it seemed that the DBHT best represented ICB sectors for a smaller number of clusters (i.e. less than eight). However, as the number of clusters increased, the ARI for Ward's linkage and the ALCA increased significantly. The SLCA (which is the method associated with the MST) led to clusters that had a low overlap with the ICB sectors.

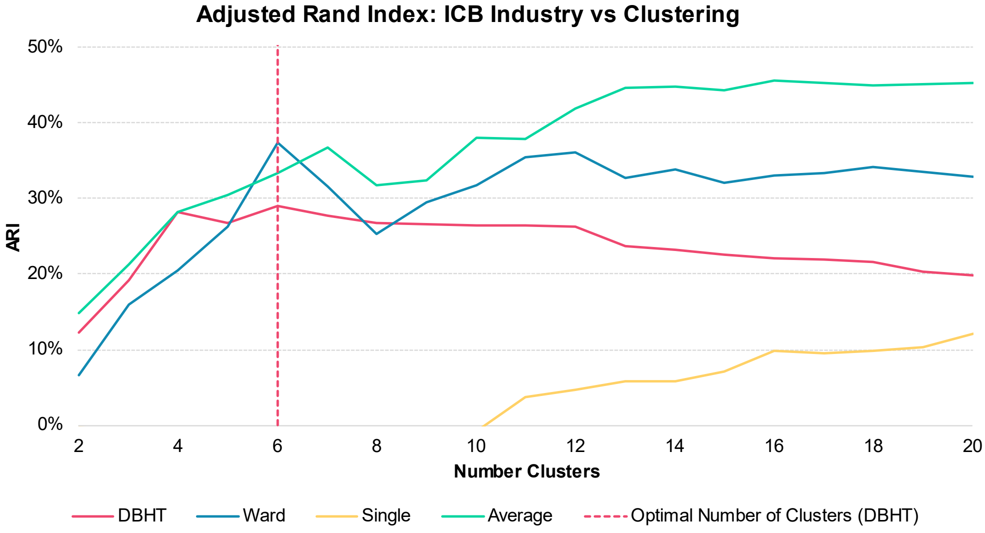


Figure 4.30 Long-Term DBHT vs Alternatives: ARI with ICB Industries

If ones considers the ARI, which was calculated using the ICB industry classification in Figure 4.30, both Ward’s linkage and the ALCA methods outperformed the DBHT. It was interesting that Ward’s method achieved its peak at the same number of clusters that was extracted from the DBHT (six). The SLCA once again performed poorly.

The above results once again re-iterated the stance that it is difficult to compare various clustering techniques using the ARI against a pre-specified industry classification benchmark. Ward’s linkage, ALCA, and the DBHT all seemed to perform well depending on the setting. However, the DBHT does have the advantage of having the network (PMFG) representation and its associated network metrics that can provide valuable information, as well as the fact that the number of clusters is determined in a fully unsupervised manner (whereas one must use a separate cluster validation technique for the other linkage methods).

4.4.4 Long-Term DBHT: How Sensitive is the “Number of Clusters” to Noise?

As we saw in Section 4.2.5 and Section 4.3.3, one can use a bootstrapping technique to assess the reliability of the structure of the filtered networks (the MST and PMFG). One can use a similar methodology to assess the robustness of the number of clusters that emerged from the long-term DBHT. For each bootstrap sample, the DBHT was estimated, and the number of clusters was extracted. Figure 4.31 depicts the frequency distribution of the number of clusters across the bootstrap samples. One can see that 6 clusters was most often extracted from the DBHT (27.2%), followed by 5 clusters (24.2%) and then 7 clusters (19.6%).

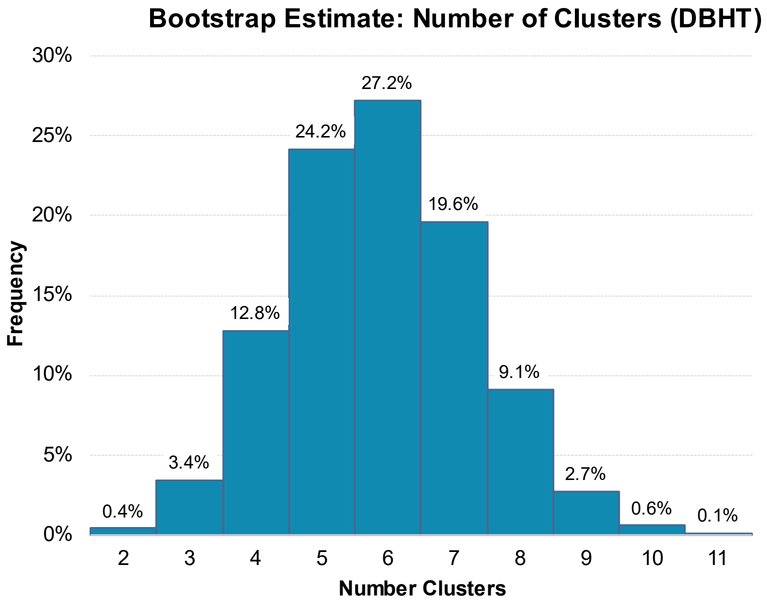


Figure 4.31 Long-Term DBHT: Bootstrap Distribution of the Number of Clusters

Table 4.4 contains sample statistics of the number of clusters, from across the 10 000 bootstrap samples. The average number of clusters was 5.87, while the median was 6. The 5th and 95th percentiles were 4, and 8, respectively.

These statistics indicated that *the six clusters that was obtained using the original data set was relatively robust*. However, there were instances of bootstrap samples that resulted in as little as 2 clusters, and some with as many as 11 clusters. A reasonable range for the number of clusters was between 4 and 8.

Table 4.4 Long-Term DBHT: Bootstrap Statistics of the Number of Clusters

Number of Clusters	
Average	5.87
Median	6
5th Percentile	4
95th Percentile	8

We now turn our attention to the results of the dynamic analysis of the DBHT. We focused particularly on the impact of the changing market environment on the number of clusters that emerged from the DBHT, as well as the ability of the DBHT to extract economic information (as measured by the ARI) during these various environments.

4.4.5 Dynamic DBHTs: Did the Number of Clusters Vary Over Time?

Figure 4.32 shows how the number of clusters has varied over time, while the frequency distribution of the dynamic number of clusters is shown in Figure 4.33. As one can see from Figure 4.33, four and five clusters were most frequently extracted from the dynamic DBHTs. These numbers are highlighted as blue dots in Figure 4.32 to emphasise this point.

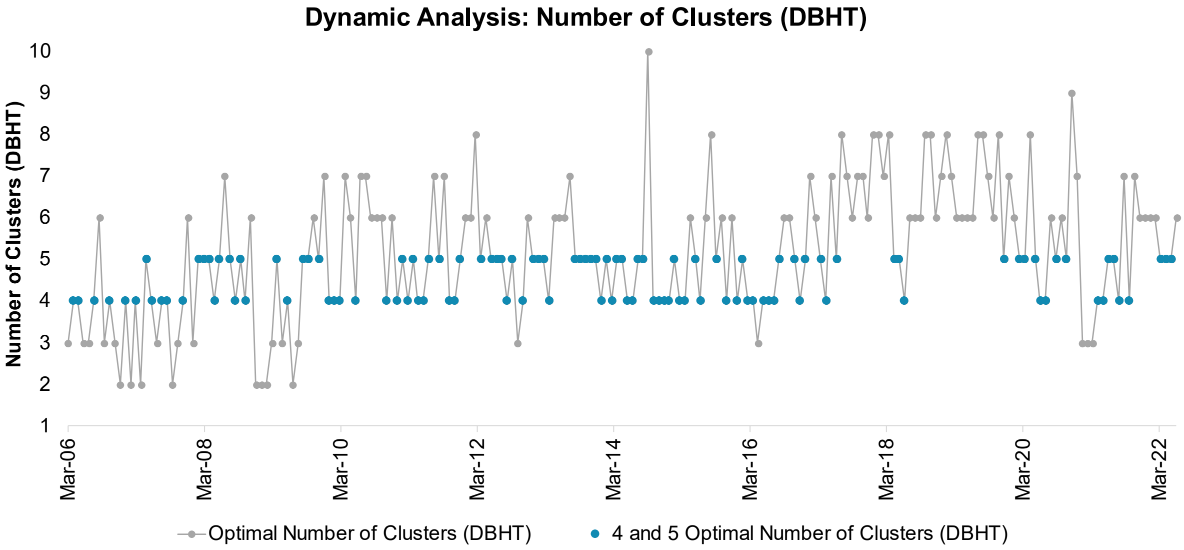


Figure 4.32 Dynamic DBHTs: Number of Clusters

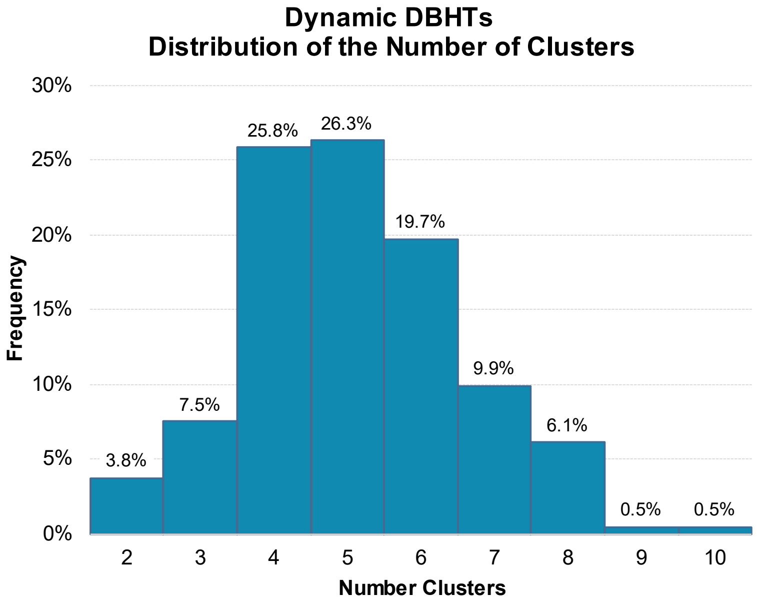


Figure 4.33 Dynamic DBHTs: Distribution of the Number of Clusters

The number of clusters did seem to be dependent upon market moving events. For example, in the bull market phase leading up to the 2008 Global Financial Crisis (GFC), the number of clusters was typically four or less, followed by an increase to a value of five. In the post GFC phase, the number of clusters then dropped to below five as central banks introduced quantitative easing, which resulted in a general recovery (and therefore higher correlations) across market segments. Another example occurred in 2018 when the USA-China trade wars resulted in poor performance across the board leading to the number of clusters dropping to four. Similarly, in 2020, as COVID-19 lockdowns began impacting negatively on economies and stock markets fell, followed by sharp recoveries as governments and central banks rolled out packages to assist ailing economies, the number of clusters dropped to three before rising again.

The question of whether the number of clusters *structurally* changed over time is a difficult one to answer, given the noisy nature of Figure 4.32. To help identify any trends in the number of clusters we applied a 52-week moving average to the data. This can be seen by the dark blue line in Figure 4.34. This chart did seem to suggest that the number of clusters has gradually trended upwards over time. In the period prior to 2010, there were several occasions when the number of clusters was less than four. However, since 2010, two, or three clusters were seldom extracted from the dynamic DBHTs. In the period from 2010 to 2016, the number of clusters seemed to remain near a value of five, but in the 2016 to 2020 period, it had increased to a values between six to eight. As of June 2022, the number of clusters had settled at a value of six. We commented briefly on the current output of the DBHT in Section 4.5.

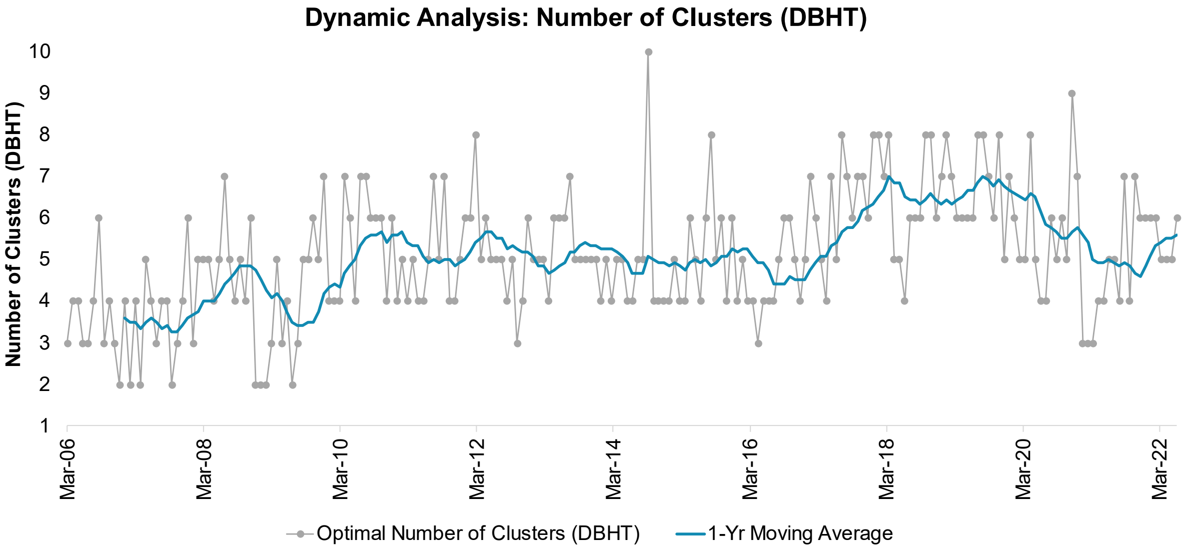


Figure 4.34 Dynamic DBHTs: Number of Clusters (smoothed)

4.4.6 Dynamic DBHTs: Does the Market Environment Impact the Ability of the DBHT to Extract Economic information?

Given that the number of clusters that was extracted from the DBHT varied dynamically over time, one may posit that the ability of the DBHT to extract economic information had also evolved over time. This was measured by dynamically calculating the ARI against the various ICB classification levels for the dynamic DBHTs. A 52-week moving average was also applied to the ARI to remove the impact of noise and highlight any trends in the data. These results are shown in Figure 4.35.

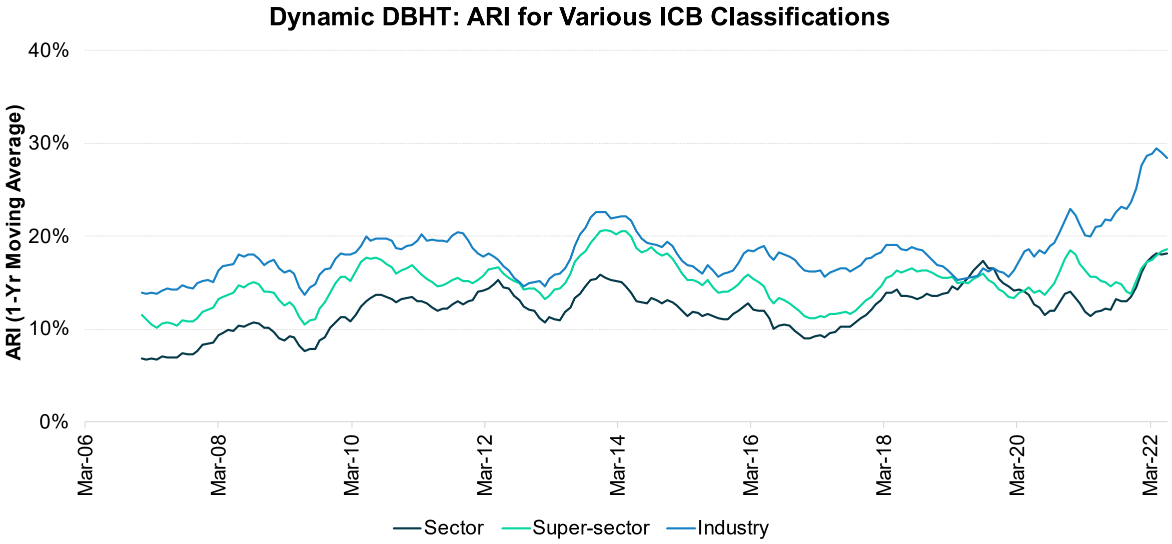


Figure 4.35 Dynamic DBHTs: ARI (smoothed) vs ICB classifications

Up to 2018, the ARI fluctuated between a value of approximately 10% to 20%. It then fell across the board leading up to the COVID-19 crash of 2020 and has increased since then. The ARI for the ICB industry level reached its highest value (close to 30%), as of June 2022.

4.5 Comparing the Current and Long-Term PMFG/DBHT

For the final piece of analysis, we compared the structure of the long-term PMFG and DBHT to the current one (estimated using the latest three years of data, and an exponentially weighted correlation matrix). As shown in Figure 4.32, the current DBHT extracted *six clusters* from the data. Figure 4.36 depicts the current PMFG, with the six clusters overlaid onto the chart by varying the colours of vertices. The size of each vertex represents the hybrid centrality of the corresponding share. Interestingly, some shares which had shown good hybrid centrality over the long-term (such as TSG, OMU, and SOL) were showed less importance in the latest DBHT, while FSR, AGL, ABG, and AMS still exhibited strong centrality characteristics in the latest DBHT. Resource stocks such as AGL, EXX, and ACL, and Financial stocks such as FSR, ABG, SBK, DSY, and SLM all ranked highly in terms of centrality on the current PMFG.

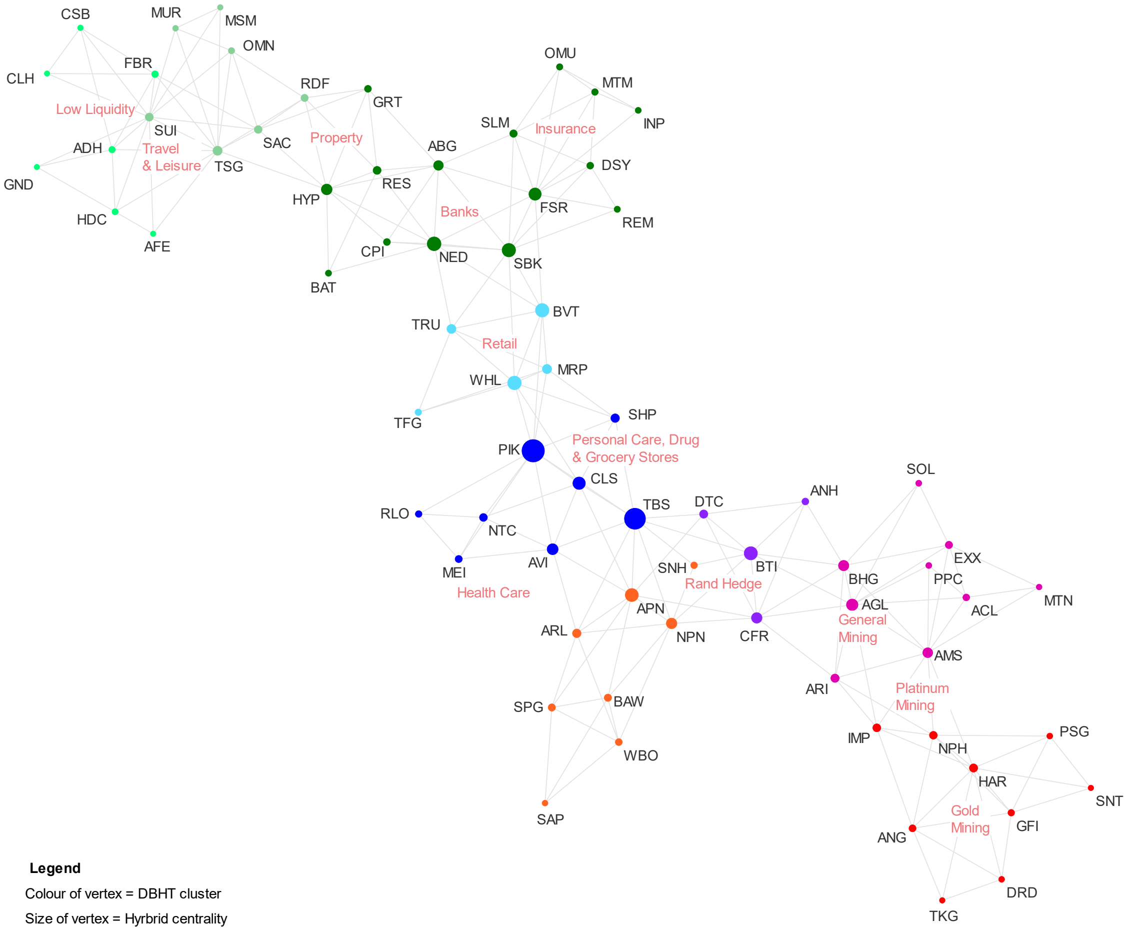


Figure 4.36 Current PMFG (Jun19 – Jun22): DBHT and Hybrid Centrality Overlay

Figure 4.37 uses a Sankey chart to show the differences between the clusters from the long-term DBHT (on the left) as compared to the current DBHT (in the middle). The clusters from the current DBHT were then compared to the ICB industries (on the right).

From Figure 4.37, one of the main differences that was identified between the long-term clustering and the short-term clusters was that Cluster 2, which consisted of the Rand Hedge shares had broken up into other clusters (i.e. into Clusters 2, 3, 4, and 6 in the current DBHT). Similarly, Cluster 3 (Consumer Discretionary, Real Estate, and Low Liquidity) had broken up into Clusters 4 and 6. These effects can occur when the market's perception of shares changes over time and investors begin to trade them in a different manner. In the case of the Rand Hedge shares, perhaps market participants had more recently been trading the shares in line with their economic sector. As can be seen in Figure 4.38, some of the shares that showed a large exposure to a weakening Rand over the long-term (such as INP, SAP, MEI, SOL, OMU, and SPG), had seen this relationship diminish, and even reverse in some cases (i.e. they showed more Rand Play tendencies than Rand Hedge), in the current time period. The theory that the economic sector of the shares was playing a larger role in how shares more recently co-moved was also confirmed by the ARI data that was shown for the dynamic DBHTs in Figure 4.35. In this chart one can see that the ARI had increased in recent times, indicating a greater overlap between the results of the DBHT and the ICB sector classifications.

These results highlighted that the way that market participants viewed shares has changed over time, even though the business models of these companies did not materially change. This was especially true for clusters that formed outside of the economic sectors of the shares. For these shares, there were periods when the market viewed them in one light, and in other market environments, they were viewed differently.

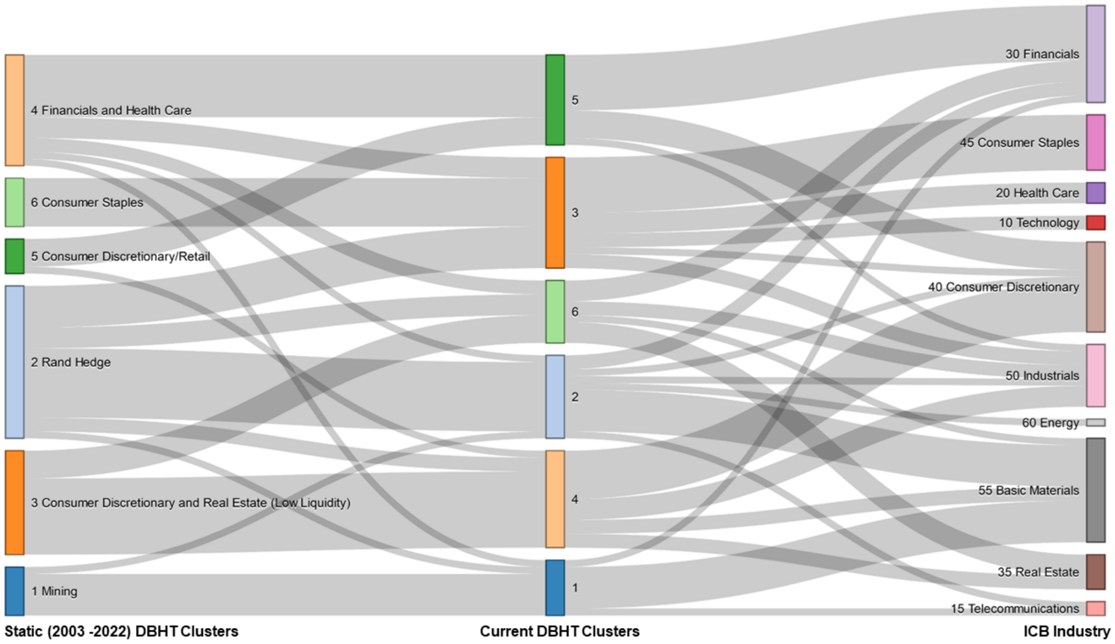


Figure 4.37 Sankey: Long-Term DBHT vs Current DBHT

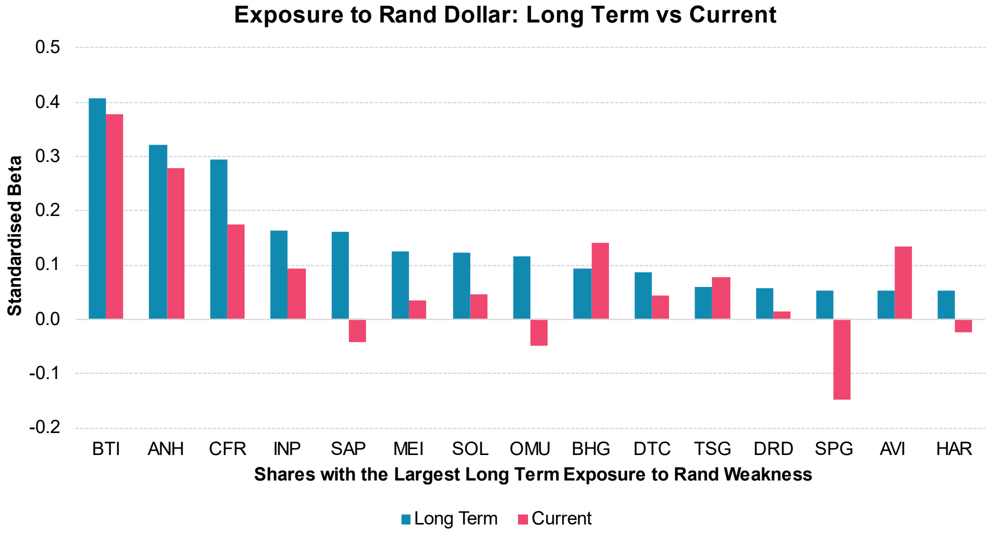


Figure 4.38 Examining the Changes in the Exposure to the Rand Dollar Exchange Rate

Chapter 5 Discussion and Future Research

5.1 Discussion

In this dissertation, we followed the track of engineers, mathematicians, and physicists, who pioneered the field of econophysics by applying methods from physics to model financial markets. These researchers favoured the idea that financial markets are a complex adaptive system, consisting of entities that behave and interact in a diverse manner, leading to non-linear, emergent behaviour of the system. In the last twenty years, there has been an increasing focus on modelling complex adaptive systems using network theory. Correlation-based networks, where stocks are represented as entities in the network, and the relationships amongst the stocks are based on the strength of the co-movements of the stocks, have been widely studied. Researchers have shown how network filtering tools, such as the MST and PMFG, have been useful to prune noise in these networks, thereby allowing important macroscopic and mesoscopic structures to emerge.

We have applied these techniques to analyse the complex interactions amongst stocks in the South African stock market. In particular, we have used MSTs and PMFGs to filter correlation-based networks of share price returns. *In keeping with our research aim, we plotted the long-term MST and saw an emergent structure in which shares from similar ICB sectors tended to cluster together*. These results agreed with other international studies. However, the so-called Rand Hedge shares, and shares which exhibited low liquidity, tended to override the sector effect and clustered together. We also applied calculated various centrality metrics for the MST and saw that AGL and BHG were deemed to be important shares on the MST. Counterintuitively, TSG, which is a share that has a small market capitalisation and trades significantly less as compared to the larger shares, was ranked first in terms of the hybrid centrality metric. However, when we assessed the reliability of the MST we saw that edges emanating from TSG tended to be less robust, confirming our suspicion that TSG's importance was overstated. From a dynamic perspective, the MST seem to shrink during market crashes, while the Basic Materials sector was typically the most central sector over time.

We then focused on the PMFG, which is a network filter that relaxes some of the constraints of the MST, thereby allowing more information that is embedded in the correlation matrix to filter through. While this method did provide a more informative filter, it also resulted in a visually noisier graph. Bootstrapping analysis suggested that the structure of the PMFG was more robust as compared to the MST, while the dynamic analysis showed similar results as compared to the MST.

One of the main benefits of the PMFG is that it is accompanied by a hierarchical clustering algorithm called the DBHT. This method has the benefit of being fully unsupervised in that it does not require the user to decide on the number of clusters that the data should be split into, i.e. the number of clusters is an automatic outcome of the DBHT. Over the long-term, the DBHT divided the stocks on the JSE into six clusters: Cluster 1 – Mining; Cluster 2 – Rand Hedge; Cluster 3 – Consumer Discretionary and Real Estate (Low Liquidity); Cluster 4 – Financials and Health Care; Cluster 5 - Consumer Discretionary/Retail; and Cluster 6 - Consumer Staples. Bootstrapping techniques that were applied to the long-term dataset confirmed that the six clusters obtained from the DBHT was relatively robust.

We also determined the amount of economic information that emerged from the DBHT using the ARI. It was encouraging to see that the ARI for the DBHT was maximised at the number of clusters extracted from the DBHT. When comparing the DBHT to other popular linkage methods, the DBHT best represented the ICB sectors at the optimal number of clusters (six), but other methods had a higher ARI if the data was split into more clusters. All of the linkage methods (except for SLCA) seemed to perform well depending on the setting. These results re-iterated the stance that it is difficult to compare various clustering techniques by using the ARI to compare the clusters against a pre-specified industry classification benchmark. This is because clustering techniques that are based on price data, can produce clusters which may not exist in the ICB classification (the Rand Hedge and Low Liquidity clusters are examples of this). However, the DBHT does have the advantage of having the network (PMFG) representation and its associated network metrics that can provide valuable information, as well as the fact that the number of clusters is determined in a fully unsupervised manner.

The dynamic analysis of the DBHTs provided noteworthy insights into the changing nature of the South African stock market. The number of clusters seemed to change during market moving events, with the average number of clusters trending upward over time. Interestingly, the amount of economic information being extracted by the DBHT had increased in recent years.

The structure of the current PMFG and DBHT (as of June 2022) showed a relatively changed structure as compared to the long-term data. Although six clusters emerged as optimal, which was similar to six clusters for the long-term DBHT, the composition of the clusters had changed. In particular, the long-term Rand Hedge cluster split into clusters that aligned more closely with the economic sectors of the stocks.

5.2 Future Research

It should be noted that the PMFG does have some limitations. The main limitation is that it is computationally costly to construct (as one needs to check that the planarity condition is satisfied at each step of the construction) and it cannot be applied to large data sets. Therefore, for large correlation matrices, or other *big data* exercises, it becomes impractical to use the PMFG. To address this issue Massara, Di Matteo and Aste (2015) introduced the Triangulated Maximally Filtered Graph (TMFG) which is generally faster to construct. The TMFG can also be updated in an *online* manner, i.e. as new data becomes available. This is important for data that is constantly arriving or being updated frequently (such as the prices of stocks that are being updated on a daily or even an intraday basis). The technique can also be *parallelised* to speed up the implementation further. Given the recent introduction of TMFGs in 2015, their application to financial markets has been limited. Table 5.1 contains a summary of academic papers that utilise the TMFG as opposed to the PMFG.

Table 5.1 Studies of TMFGs Applied to Financial Markets

Setting	Reference
Stocks: New York Stock Exchange (NYSE)	Massara, Di Matteo and Aste (2015)
Stocks: NYSE, Italy, Germany, and Israel	Turiel, Barucca and Aste (2020)
Stocks: China	Xu et al. (2022)
Cryptocurrencies	Briola and Aste (2022) Katsiampa, Yarovaya and Zięba (2022)

From a general modelling point of view, even though we made carefully considered decisions based on a mixture of practitioner experience and academic literature, it was clear that these choices had a material impact on the outcome of the analysis. This confirmed the view of Marti et al. (2015). Specifically, the period that was considered for the analysis had a material impact, with some companies switching clusters even though they did not change their business models. This may be due to the market participants viewing them from one perspective during certain regimes and then treating them differently during others. The Rand Hedge shares were a prime example of this, being traded as a group when investors were looking to express a view on the currency and being traded along with other stocks within their respective sectors at other times. It is difficult to model these types of shares using traditional methods of clustering. This is because the methods create *hard partitions*, in which a stock is either in a cluster or out, i.e. it can only form part of one cluster. This works well for shares that are dominated by a single factor (e.g. gold mining shares which are predominantly driven by the movement in the underlying gold price). On the other hand, *soft clustering* techniques such as Fuzzy C-Means (introduced by Dunn, 1973) may be more appropriate for shares that show more of a dual nature (such as the Rand Hedge shares). Gopi (2009) and Gopi (2012a) applied this technique to the South African stock market with interesting results.

In this dissertation, the only *feature* that was used to model the relationships amongst shares was the co-movement of share prices. Given that share price returns are subject to noise, and even though one can apply noise reduction techniques, it is difficult to quantify the amount of noise that will be removed and the amount that will remain. Therefore, using *alternate data sources*, or other *techniques* may prove to be useful. Winton (2018) clustered stocks based on the similarity of the text in the annual reports of companies, while Fodor, Jorgensen and Stowe (2021) clustered stocks based on data from the financial statements of companies. Networks based on these features may provide a complementary viewpoint to the traditional return-based correlation networks. An interesting approach may be to apply to a *multiplex network* to combine networks based on various types of features (such as those mentioned here). These types of networks were introduced in a financial market setting by Musmeci et al. (2016). The authors used a multiplex PMFG to analyse various dependency measures such as linear, non-linear, tail, and partial correlations. These networks were then combined, and the interaction of the various layers of the networks revealed insights that would not have been observed from the analysis of each network in isolation.

Finally, recent developments have seen the application of *machine learning techniques* to networks of financial networks. One such development is the use of algorithms that take the topological features of the networks themselves (such as the centrality metrics, clusters etc.) and uses them to predict the structure of the market in the future (see Castilho *et al.*, 2021). Other developments have seen the use of techniques such as Node2Vec (Grover and Leskovec, 2016) to compress the network into a lower dimensional continuous space, called an embedding. Sarmah et al. (2022), state that these embeddings can be used to model the network in an intuitive manner, as well as for a variety of tasks such as building stock recommender systems, performing analogical inferences, etc. We believe that the use of networks and graphs together with modern machine learning techniques may provide a useful avenue for future research.

Chapter 6 References

Akhanli, S.E. and Hennig, C. (2020) 'Comparing Clusterings and Numbers of Clusters by Aggregation of Calibrated Clustering Validity Indexes', Statistics *and Computing*, 30(5), pp. 1523–1544. Available at: https://doi.org/10.1007/s11222-020-09958-2.

Aslam, F., Mohmand, Y.T., Ferreira, P., Memon, B.A., Khan, Maaz and Khan, Mrestyal (2020) 'Network Analysis of Global Stock Markets at the Beginning of the Coronavirus Disease (Covid-19) Outbreak', *Borsa Istanbul Review*, 20, pp. S49–S61. Available at: https://doi.org/10.1016/j.bir.2020.09.003.

Aste, T. (2012) *Calculate the PMFG from a Matrix of Weights*. Available at: https://www.mathworks.com/matlabcentral/fileexchange/38689-pmfg (Accessed: 10 October 2022).

Aste, T. (2014) *Perform DBHT Clustering*. Available at: https://www.mathworks.com/matlabcentral/fileexchange/46750-dbht (Accessed: 10 October 2022).

Aste, T., Di Matteo, T., Tumminello, M. and Mantegna, R.N. (2005) 'Correlation Filtering in Financial Time Series'. arXiv. Available at: https://doi.org/10.48550/arXiv.physics/0508118.

Aste, T., Shaw, W. and Matteo, T.D. (2010) 'Correlation Structure and Dynamics in Volatile Markets', *New Journal of Physics*, 12(8), p. 085009. Available at: https://doi.org/10.1088/1367-2630/12/8/085009.

Bajramovic, F., Tauber, A., Wozelka, R. and Ferdinand, W. (2011) 'A Taxonomy of Force-Directed Placement Techniques', p. 33.

Barabási, A.-L., Gulbahce, N. and Loscalzo, J. (2011) 'Network Medicine: A Network-based Approach to Human Disease', *Nature reviews. Genetics*, 12(1), pp. 56–68. Available at: https://doi.org/10.1038/nrg2918.

Barbi, A.Q. and Prataviera, G.A. (2019) 'Nonlinear Dependencies on Brazilian Equity Network from Mutual Information Minimum Spanning Trees', *Physica A: Statistical Mechanics and its Applications*, 523, pp. 876–885. Available at: https://doi.org/10.1016/j.physa.2019.04.147.

Barr, G.D.I., Kantor, B.S. and Holdsworth, C.G. (2007) 'The Effect of the Rand Exchange Rate on the JSE Top-40 Stocks: An Analysis for the Practitioner', *South African Journal of Business Management*, 38(1), pp. 45–58. Available at: https://doi.org/10.4102/sajbm.v38i1.577.

Bastian, M., Heymann, S. and Jacomy, M. (2009) 'Gephi: An Open Source Software for Exploring and Manipulating Networks', *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1), pp. 361–362.

Batagelj, V. and Mrvar, A. (2004) 'Pajek — Analysis and Visualization of Large Networks', in M. Jünger and P. Mutzel (eds) *Graph Drawing Software*. Berlin, Heidelberg: Springer (Mathematics and Visualization), pp. 77–103. Available at: https://doi.org/10.1007/978-3-642-18638-7_4.

Bennett, S., Cucuringu, M. and Reinert, G. (2022) 'Lead-Lag Detection and Network Clustering for Multivariate Time Series with an Application to the Us Equity Market'. arXiv. Available at: https://doi.org/10.48550/arXiv.2201.08283.

Beveridge, A. and Shan, J. (2016) 'Network of Thrones', *Math Horizons*, 23(4), pp. 18–22. Available at: https://doi.org/10.4169/mathhorizons.23.4.18.

Billio, M., Getmansky, M., Lo, A.W. and Pelizzon, L. (2012) 'Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors', *Journal of Financial Economics*, 104(3), pp. 535–559. Available at: https://doi.org/10.1016/j.jfineco.2011.12.010.

Birch, J. (2016) *Modelling Financial Markets Using Methods from Network Theory.* phd. University of Liverpool. Available at: https://livrepository.liverpool.ac.uk/2028739 (Accessed: 28 September 2022).

Birch, J., Pantelous, A.A. and Soramäki, K. (2016) 'Analysis of Correlation Based Networks Representing DAX 30 Stock Price Returns', *Computational Economics*, 47(4), pp. 501–525. Available at: https://doi.org/10.1007/s10614-015-9481-z.

Boginski, V., Butenko, S. and Pardalos, P.M. (2005) 'Statistical Analysis of Financial Networks', *Computational Statistics & Data Analysis*, 48(2), pp. 431–443. Available at: https://doi.org/10.1016/j.csda.2004.02.004.

Bonanno, G., Caldarelli, G., Lillo, F. and Mantegna, and R.N. (2003) 'Topology of Correlation Based Minimal Spanning Trees in Real and Model Markets', *Physical Review E*, 68(4), p. 046130. Available at: https://doi.org/10.1103/PhysRevE.68.046130.

Bonanno, G., Vandewalle, N. and Mantegna, R.N. (2000) 'Taxonomy of Stock Market Indices', *Physical Review E*, 62(6), pp. R7615–R7618. Available at: https://doi.org/10.1103/PhysRevE.62.R7615.

Borghesi, C., Marsili, M. and Miccichè, S. (2007) 'Emergence of Time-Horizon Invariant Correlation Structure in Financial Returns by Subtraction of the Market Mode', *Physical Review E*, 76(2), p. 026104. Available at: https://doi.org/10.1103/PhysRevE.76.026104.

Borůvka, O. (1926) 'O Jistém Problému Minimálním (About a Certain Minimal Problem)', *Praca Moravske Prirodovedecke Spolecnosti*, 3(1), pp. 36–58.

Bradfield, D. (2003) 'Investment Basics XLVI. On Estimating the Beta Coefficient', *Investment Analysts Journal*, 32(57), pp. 47–53. Available at: https://doi.org/10.1080/10293523.2003.11082448.

Brandes, U. and Pich, C. (2007) 'Eigensolver Methods for Progressive Multidimensional Scaling of Large Data', in M. Kaufmann and D. Wagner (eds) *Graph Drawing*. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 42–53. Available at: https://doi.org/10.1007/978-3-540-70904-6_6.

Brida, J.G. and Risso, W.A. (2008) 'Multidimensional Minimal Spanning Tree: The Dow Jones Case', *Physica A: Statistical Mechanics and its Applications*, 387(21), pp. 5205–5210. Available at: https://doi.org/10.1016/j.physa.2008.05.009.

Brin, S. and Page, L. (1998) 'The Anatomy of a Large-Scale Hypertextual Web Search Engine', *Computer Networks and ISDN Systems*, 30(1), pp. 107–117. Available at: https://doi.org/10.1016/S0169-7552(98)00110-X.

Briola, A. and Aste, T. (2022) 'Dependency Structures in Cryptocurrency Market from High to Low Frequency'. arXiv. Available at: https://doi.org/10.48550/arXiv.2206.03386.

Brockmann, D. and Helbing, D. (2013) 'The Hidden Geometry of Complex, Network-Driven Contagion Phenomena', *Science*, 342(6164), pp. 1337–1342. Available at: https://doi.org/10.1126/science.1245200.

Buldú, J.M., Busquets, J., Echegoyen, I. and Seirul. lo, F. (2019) 'Defining a Historic Football Team: Using Network Science to Analyze Guardiola's F.C. Barcelona', *Scientific Reports*, 9(1), p. 13602. Available at: https://doi.org/10.1038/s41598-019-49969-2.

Bun, J., Bouchaud, J.-P. and Potters, M. (2017) ‘Cleaning Large Correlation Matrices: Tools from Random Matrix Theory’, *Physics Reports*, 666, pp. 1–109. Available at: https://doi.org/10.1016/j.physrep.2016.10.005.

Castilho, D., Souza, T.T.P., Kang, S.M., Gama, J. and de Carvalho, A.C.P.L.F. (2021) ‘Forecasting Financial Market Structure from Network Features using Machine Learning’. arXiv. Available at: https://doi.org/10.48550/arXiv.2110.11751.

Clayton, M.A. (2018) *Correlation Estimates from Asynchronously Observed Series*. SSRN Scholarly Paper 3270305. Rochester, NY: Social Science Research Network. Available at: https://doi.org/10.2139/ssrn.3270305.

Coelho, R., Hutzler, S., Repetowicz, P. and Richmond, P. (2007) ‘Sector Analysis for a FTSE Portfolio of Stocks’, *Physica A: Statistical Mechanics and its Applications*, 373, pp. 615–626. Available at: https://doi.org/10.1016/j.physa.2006.02.050.

Coletti, P. (2016) ‘Comparing Minimum Spanning Trees of the Italian Stock Market Using Returns and Volumes’, *Physica A: Statistical Mechanics and its Applications*, 463, pp. 246–261. Available at: https://doi.org/10.1016/j.physa.2016.07.029.

Cook, S., Soramäki, K. and Laubsch, A. (2016) ‘A Network-Based Method for Visual Identification of Systemic Risks’, *Journal of Network Theory in Finance* [Preprint]. Available at: https://www.risk.net/node/2454528 (Accessed: 24 November 2022).

Coronnello, C., Tumminello, M., Lillo, F., Micciche`, S. and Mantegna, R.N. (2007) ‘Economic Sector Identification in a Set of Stocks Traded at the New York Stock Exchange: A Comparative Analysis’, *arXiv:physics/0609036*, p. 66010T. Available at: https://doi.org/10.1117/12.729619.

Coronnello, C., Tumminello, M., Lillo, F., Miccichè, S. and Mantegna, R.N. (2005) ‘Sector Identification in a Set of Stock Return Time Series Traded at the London Stock Exchange’. arXiv. Available at: https://doi.org/10.48550/arXiv.cond-mat/0508122.

Denkowska, A. and Wanat, S. (2020) ‘A Tail Dependence-Based MST and Their Topological Indicators in Modelling Systemic Risk in the European Insurance Sector’, *arXiv:2001.06567 [econ, q-fin]* [Preprint]. Available at: http://arxiv.org/abs/2001.06567 (Accessed: 16 February 2021).

Di Matteo, T., Pozzi, F. and Aste, T. (2010) ‘The Use of Dynamical Networks to Detect the Hierarchical Organization of Financial Market Sectors’, *The European Physical Journal B*, 73(1), pp. 3–11. Available at: https://doi.org/10.1140/epjb/e2009-00286-0.

Dias, J. (2012) ‘Sovereign Debt Crisis in the European Union: A Minimum Spanning Tree Approach’, *Physica A: Statistical Mechanics and its Applications*, 391(5), pp. 2046–2055. Available at: https://doi.org/10.1016/j.physa.2011.11.004.

Dunn, J.C. (1973) ‘A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters’, *Journal of Cybernetics*, 3(3), pp. 32–57. Available at: https://doi.org/10.1080/01969727308546046.

Efron, B. (1979) ‘Bootstrap Methods: Another Look at the Jackknife’, *The Annals of Statistics*, 7(1), pp. 1–26. Available at: https://doi.org/10.1214/aos/1176344552.

Eryiğit, M. and Eryiğit, R. (2009) ‘Network Structure of Cross-Correlations Among the World Market Indices’, *Physica A: Statistical Mechanics and its Applications*, 388(17), pp. 3551–3562. Available at: https://doi.org/10.1016/j.physa.2009.04.028.

Fodor, A., Jorgensen, R.D. and Stowe, J.D. (2021) ‘Financial Clusters, Industry Groups, and Stock Return Correlations’, *Journal of Financial Research*, 44(1), pp. 121–144. Available at: https://doi.org/10.1111/jfir.12236.

Fornito, A., Zalesky, A. and Bullmore, E. (2016) *Fundamentals of Brain Network Analysis*. 1st edition. Amsterdam ; Boston: Academic Press.

Fruchterman, T.M.J. and Reingold, E.M. (1991) ‘Graph Drawing by Force-Directed Placement’, *Software: Practice and Experience*, 21(11), pp. 1129–1164. Available at: https://doi.org/10.1002/spe.4380211102.

Garas, A. and Argyrakis, P. (2007) ‘Correlation Study of the Athens Stock Exchange’, *Physica A: Statistical Mechanics and its Applications*, 380, pp. 399–410. Available at: https://doi.org/10.1016/j.physa.2007.02.097.

Giada, L. and Marsili, M. (2002) ‘Algorithms of maximum likelihood data clustering with applications’, *Physica A: Statistical Mechanics and its Applications*, 315(3–4), pp. 650–664. Available at: https://doi.org/10.1016/S0378-4371(02)00974-3.

Gibson, H., Faith, J. and Vickers, P. (2013) ‘A Survey of Two-Dimensional Graph Layout Techniques for Information Visualisation’, *Information Visualization*, 12(3–4), pp. 324–357. Available at: https://doi.org/10.1177/1473871612455749.

Giudici, P. and Polinesi, G. (2021) ‘Crypto Price Discovery Through Correlation Networks’, *Annals of Operations Research*, 299(1), pp. 443–457. Available at: https://doi.org/10.1007/s10479-019-03282-3.

Goh, Y.K., Hasim, H.M. and Antonopoulos, C.G. (2018) ‘Inference of Financial Networks Using the Normalised Mutual Information Rate’, *PLOS ONE*, 13(2), p. e0192160. Available at: https://doi.org/10.1371/journal.pone.0192160.

Gopi, Y. (2008) *The JSE is a Tree!* Cadiz Quantitative Research, p. 34.

Gopi, Y. (2009) *Finding Clarity by Blurring the Lines. Fuzzy Cluster Analysis on the JSE*. Cadiz Quantitative Research.

Gopi, Y. (2010) *What’s Under the JSE Christmas Tree? Depicting Value, Growth and Performance on a Minimal Spanning Tree*. Cadiz Quantitative Research, p. 9.

Gopi, Y. (2012a) *Fuzzy Clustering on the JSE - 2012 Update. Analysis, Explanations and Opportunities*. Cadiz Quantitative Research.

Gopi, Y. (2012b) *The JSE is a Tree! 2012 Update Depicting Value, Growth and Momentum on a Minimal Spanning Tree*. Cadiz Quantitative Research, p. 10.

Gopi, Y. (2014) *Quantum Leaps: February 2014 (Barking up the JSE tree: a 2014 update)*. BNP Paribas Cadiz Quantitative Research, p. 23.

Gower, J.C. and Ross, G.J.S. (1969) ‘Minimum Spanning Trees and Single Linkage Cluster Analysis’, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1), pp. 54–64. Available at: https://doi.org/10.2307/2346439.

Graham, R.L. and Hell, P. (1985) ‘On the History of the Minimum Spanning Tree Problem’, *Annals of the History of Computing*, 7(1), pp. 43–57. Available at: https://doi.org/10.1109/MAHC.1985.10011.

Grandjean, M. (2016) ‘A Social Network Analysis of Twitter: Mapping the Digital Humanities Community’, *Cogent Arts & Humanities*. Edited by A. Mauro, 3(1), p. 1171458. Available at: https://doi.org/10.1080/23311983.2016.1171458.

‘Graph Theory – The Network Pages’ (no date). Available at: https://www.networkpages.nl/graph-theory/ (Accessed: 11 September 2022).

Grover, A. and Leskovec, J. (2016) ‘node2vec: Scalable Feature Learning for Networks’. arXiv. Available at: https://doi.org/10.48550/arXiv.1607.00653.

Grunspan, D.Z., Wiggins, B.L. and Goodreau, S.M. (2014) ‘Understanding Classrooms through Social Network Analysis: A Primer for Social Network Analysis in Education Research’, *CBE Life Sciences Education*, 13(2), pp. 167–178. Available at: https://doi.org/10.1187/cbe.13-08-0162.

Guo, H., Yu, H., An, Q. and Zhang, X. (2022) ‘Risk Analysis of China’s Stock Markets Based on Topological Data Structures’, *Procedia Computer Science*, 202, pp. 203–216. Available at: https://doi.org/10.1016/j.procs.2022.04.028.

Guo, X., Zhang, H. and Tian, T. (2018) ‘Development of Stock Correlation Networks Using Mutual Information and Financial Big Data’, *PLOS ONE*, 13(4), p. e0195941. Available at: https://doi.org/10.1371/journal.pone.0195941.

Hagberg, A., Swart, P. and S Chult, D. (2008) *Exploring network structure, dynamics, and function using NetworkX*. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Han, C., He, Z. and Toh, A.J.W. (2022) ‘Pairs Trading Via Unsupervised Learning’, *European Journal of Operational Research* [Preprint]. Available at: https://doi.org/10.1016/j.ejor.2022.09.041.

Harel, D. and Koren, Y. (2004) ‘Graph Drawing by High-Dimensional Embedding’, *Journal of Graph Algorithms and Applications*, 8(2), pp. 195–214. Available at: https://doi.org/10.7155/jgaa.00089.

Hayashi, T. and Yoshida, N. (2005) ‘On Covariance Estimation of Non-Synchronously Observed Diffusion Processes’, *Bernoulli*, 11(2), pp. 359–379. Available at: https://doi.org/10.3150/bj/1116340299.

Hendricks, D., Wilcox, D. and Gebbie, T. (2016) ‘High-Speed Detection of Emergent Market Clustering Via an Unsupervised Parallel Genetic Algorithm’, *South African Journal of Science*, Volume 112(Number 1/2). Available at: https://doi.org/10.17159/sajs.2016/20140340.

Henningsen, E.B. (2019) *On the Accuracy of the SARD Approach Across Country Borders*. Master of Science in Finance and Accounting. Copenhagen Business School. Available at: https://research.cbs.dk/en/studentProjects/on-the-accuracy-of-the-sard-approach-across-country-borders (Accessed: 25 September 2022).

Heywood, G.C., Marsland, J.R. and Morrison, G.M. (2003) ‘Practical Risk Management for Equity Portfolio Managers’, *British Actuarial Journal*, 9(5), pp. 1061–1123. Available at: https://doi.org/10.1017/S1357321700004463.

Hong, M.Y. and Yoon, J.W. (2022) ‘The Impact of COVID-19 on Cryptocurrency Markets: A Network Analysis Based on Mutual Information’, *PLOS ONE*, 17(2), p. e0259869. Available at: https://doi.org/10.1371/journal.pone.0259869.

Hu, Y. (2005) ‘Efficient and High Quality Force-Directed Graph Drawing’, *The Mathematica Journal*, 10, pp. 37–71.

Hu, Y. (2011) ‘Algorithms for Visualizing Large Networks’, *Combinatorial Scientific Computing*, 5. Available at: https://doi.org/10.1201/b11644-20.

Huang, W., Wang, H. and Wei, Y. (2020) ‘Mapping the Illegal International Ivory Trading Network to Identify Key Hubs and Smuggling Routes’, *EcoHealth*, 17(4), pp. 523–539. Available at: https://doi.org/10.1007/s10393-020-01511-x.

Hubert, L. and Arabie, P. (1985) ‘Comparing partitions’, *Journal of Classification*, 2(1), pp. 193–218. Available at: https://doi.org/10.1007/BF01908075.

Inkscape Project (2022) ‘Inkscape’. Available at: https://inkscape.org.

Jacomy, M., Venturini, T., Heymann, S. and Bastian, M. (2014) ‘ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software’, *PLOS ONE*, 9(6), p. e98679. Available at: https://doi.org/10.1371/journal.pone.0098679.

James, W. and Stein, C. (1961) ‘Estimation with Quadratic Loss’, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–380.

Jang, W., Lee, J. and Chang, W. (2011) ‘Currency crises and the evolution of foreign exchange market: Evidence from minimum spanning tree’, *Physica A: Statistical Mechanics and its Applications*, 390(4), pp. 707–718. Available at: https://doi.org/10.1016/j.physa.2010.10.028.

Ji, Q. and Fan, Y. (2016) ‘Evolution of the World Crude Oil Market Integration: A Graph Theory Analysis’, *Energy Economics*, 53, pp. 90–100. Available at: https://doi.org/10.1016/j.eneco.2014.12.003.

Jung, W.-S., Chae, S., Yang, J.-S. and Moon, H.-T. (2006) ‘Characteristics of the Korean stock market correlations’, *Physica A: Statistical Mechanics and its Applications*, 361(1), pp. 263–271. Available at: https://doi.org/10.1016/j.physa.2005.06.081.

Kamada, T. and Kawai, S. (1989) ‘An Algorithm for Drawing General Undirected Graphs’, *Information Processing Letters*, 31(1), pp. 7–15. Available at: https://doi.org/10.1016/0020-0190(89)90102-6.

Katsiampa, P., Yarovaya, L. and Zięba, D. (2022) ‘High-Frequency Connectedness Between Bitcoin and Other Top-Traded Crypto Assets During the COVID-19 Crisis’, *Journal of International Financial Markets, Institutions and Money*, 79. Available at: https://doi.org/10.1016/j.intfin.2022.101578.

Kenett, D.Y., Tumminello, M., Madi, A., Gur-Gershgoren, G., Mantegna, R.N. and Ben-Jacob, E. (2010) ‘Dominating Clasp of the Financial Sector Revealed by Partial Correlation Analysis of the Stock Market’, *PLOS ONE*, 5(12), p. e15032. Available at: https://doi.org/10.1371/journal.pone.0015032.

Keskin, M., Deviren, B. and Kocakaplan, Y. (2011) ‘Topology of the Correlation Networks Among Major Currencies Using Hierarchical Structure Methods’, *Physica A: Statistical Mechanics and its Applications*, 390(4), pp. 719–730. Available at: https://doi.org/10.1016/j.physa.2010.10.041.

King, B. (1966) ‘Market and Industry Factors in Stock Price Behavior’. Available at: https://doi.org/10.1086/294847.

Knudsen, J.O., Kold, S. and Plenborg, T. (2017) ‘Stick to the Fundamentals and Discover Your Peers’, *Financial Analysts Journal*, 73(3), pp. 85–105. Available at: https://doi.org/10.2469/faj.v73.n3.5.

Kruskal, J.B. (1956) ‘On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem’, *Proceedings of the American Mathematical Society*, 7(1), pp. 48–50. Available at: https://doi.org/10.2307/2033241.

Kukreti, V., Pharasi, H.K., Gupta, P. and Kumar, S. (2020) ‘A Perspective on Correlation-Based Financial Networks and Entropy Measures’, *Frontiers in Physics*, 8. Available at: https://www.frontiersin.org/articles/10.3389/fphy.2020.00323 (Accessed: 8 July 2022).

Kwapien, J., Gworek, S., Drozdz, S. and Gorski, A. (2009) ‘Analysis of a Network Structure of the Foreign Currency Exchange Market’. arXiv. Available at: https://doi.org/10.48550/arXiv.0906.0480.

Laloux, L., Cizeau, P., Bouchaud, J.-P. and Potters, M. (1999) ‘Noise Dressing of Financial Correlation Matrices’, *Physical Review Letters*, 83(7), pp. 1467–1470. Available at: https://doi.org/10.1103/PhysRevLett.83.1467.

Laloux, L., Cizeau, P., Potters, M. and Bouchaud, J.-P. (2000) ‘Random matrix theory and financial correlations’, *International Journal of Theoretical and Applied Finance*, 03(03), pp. 391–397. Available at: https://doi.org/10.1142/S0219024900000255.

Ledoit, O. (2022) ‘covShrinkage - A Package for Shrinkage Estimation of Covariance Matrices’. Available at: https://github.com/oledoit/covShrinkage (Accessed: 20 November 2022).

Ledoit, O. and Wolf, M. (2003) ‘Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection’, *Journal of Empirical Finance*, 10(5), pp. 603–621. Available at: https://doi.org/10.1016/S0927-5398(03)00007-0.

Ledoit, O. and Wolf, M. (2004a) ‘A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices’, *Journal of Multivariate Analysis*, 88(2), pp. 365–411. Available at: https://doi.org/10.1016/S0047-259X(03)00096-4.

Ledoit, O. and Wolf, M. (2004b) ‘Honey, I Shrunk the Sample Covariance Matrix’, *The Journal of Portfolio Management*, 30(4), pp. 110–119. Available at: https://doi.org/10.3905/jpm.2004.110.

Ledoit, O. and Wolf, M. (2012) ‘Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices’, *The Annals of Statistics*, 40(2), pp. 1024–1060. Available at: https://doi.org/10.1214/12-AOS989.

Ledoit, O. and Wolf, M. (2017) ‘Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks’, *The Review of Financial Studies*, 30(12), pp. 4349–4388. Available at: https://doi.org/10.1093/rfs/hhx052.

Ledoit, O. and Wolf, M. (2020) ‘Quadratic Shrinkage for Large Covariance Matrices’. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.3486378.

Ledoit, O. and Wolf, M. (2022) ‘The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation’, *Journal of Financial Econometrics*, 20(1), pp. 187–218. Available at: https://doi.org/10.1093/jjfinec/nbaa007.

Lohre, H., Rother, C. and Schäfer, K.A. (2020) *Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi-Asset Multi-Factor Allocations*. SSRN Scholarly Paper ID 3513399. Rochester, NY: Social Science Research Network. Available at: https://doi.org/10.2139/ssrn.3513399.

López de Prado, M. (2016) ‘Building Diversified Portfolios that Outperform Out of Sample’, *The Journal of Portfolio Management*, 42(4), pp. 59–69. Available at: https://doi.org/10.3905/jpm.2016.42.4.059.

Lucey, B.M. (2010) ‘Comovements in Government Bond Markets: A Minimum Spanning Tree Analysis’. Available at: https://doi.org/10.1016/j.physa.2010.06.057.

MacMahon, M. and Garlaschelli, D. (2015) ‘Community Detection for Correlation Matrices’, *Physical Review X*, 5(2), p. 021006. Available at: https://doi.org/10.1103/PhysRevX.5.021006.

Majapa, M. and Gossel, S.J. (2016) ‘Topology of the South African Stock Market Network Across the 2008 Financial Crisis’, *Physica A: Statistical Mechanics and its Applications*, 445, pp. 35–47. Available at: https://doi.org/10.1016/j.physa.2015.10.108.

Mantegna, R.N. (1999) 'Hierarchical Structure in Financial Markets', *The European Physical Journal B*, 11(1), pp. 193–197. Available at: https://doi.org/10.1007/s100510050929.

Marčenko, V.A. and Pastur, L.A. (1967) 'Distribution of Eigenvalues for Some Sets of Random Matrices', *Mathematics of the USSR-Sbornik*, 1(4), p. 457. Available at: https://doi.org/10.1070/SM1967v001n04ABEH001994.

Marti, G., Andler, S., Nielsen, F. and Donnat, P. (2016) 'Clustering Financial Time Series: How Long is Enough?', *arXiv:1603.04017 [q-fin, stat]* [Preprint]. Available at: http://arxiv.org/abs/1603.04017 (Accessed: 16 February 2021).

Marti, G., Nielsen, F., Bińkowski, M. and Donnat, P. (2021) 'A Review of Two Decades of Correlations, Hierarchies, Networks and Clustering in Financial Markets', *arXiv:1703.00485 [q-fin]* [Preprint]. Available at: https://doi.org/10.1007/978-3-030-65459-7.

Marti, G., Very, P., Donnat, P. and Nielsen, F. (2015) 'A Proposal of a Methodological Framework with Experimental Guidelines to Investigate Clustering Stability on Financial Time Series', *arXiv:1509.05475 [cs, q-fin]* [Preprint]. Available at: http://arxiv.org/abs/1509.05475 (Accessed: 16 February 2021).

Martin, G.R.R. (2002) *A Storm of Swords: A Song of Ice and Fire: Book Three*. New York: Bantam.

Martin, S., Brown, W.M., Klavans, R. and Boyack, K.W. (2011) 'OpenOrd: an open-source toolbox for large graph layout', in. *IS&T/SPIE Electronic Imaging*, San Francisco Airport, California, USA, p. 786806. Available at: https://doi.org/10.1117/12.871402.

Massara, G.P., Di Matteo, T. and Aste, T. (2015) 'Network Filtering for Big Data: Triangulated Maximally Filtered Graph', *arXiv:1505.02445 [cond-mat]* [Preprint]. Available at: http://arxiv.org/abs/1505.02445 (Accessed: 18 March 2019).

MATLAB (2022) 'version 9.12.0.1884302 (R2022a)'. Natick, Massachusetts: The MathWorks Inc.

Mauboussin, M.J. (2012) *Think Twice: Harnessing the Power of Counterintuition*. Boston, Mass: Harvard Business Review Press.

Mayer, R.E. (1997) 'Multimedia Learning: Are We Asking the Right Questions?', *Educational Psychologist*, 32(1), pp. 1–19. Available at: https://doi.org/10.1207/s15326985ep3201_1.

Mbambiso, B. (2008) *Dissecting the South African Financial Markets into Sectors and States*. University of Cape Town.

Mbatha, V.M. and Alovokpinhou, S.A. (2022) 'The Structure of the South African Stock Market Network During COVID-19 Hard Lockdown', *Physica A: Statistical Mechanics and its Applications*, 590, p. 126770. Available at: https://doi.org/10.1016/j.physa.2021.126770.

McDonald, M., Suleman, O., Williams, S., Howison, S. and Johnson, N.F. (2005) 'Detecting a Currency's Dominance or Dependence Using Foreign Exchange Network Trees', *Physical Review E*, 72(4). Available at: https://doi.org/10.1103/PhysRevE.72.046106.

Miccichè, S., Bonanno, G., Lillo, F. and N. Mantegna, R. (2003) 'Degree Stability of a Minimum Spanning Tree of Price Return and Volatility', *Physica A: Statistical Mechanics and its Applications*, 324(1–2), pp. 66–73. Available at: https://doi.org/10.1016/S0378-4371(03)00002-5.

Miceli, M.A. and Susinno, G. (2004) 'Ultrametricity in Fund of Funds Diversification', *Physica A: Statistical Mechanics and its Applications*, 344(1), pp. 95–99. Available at: https://doi.org/10.1016/j.physa.2004.06.094.

Millington, T. and Niranjan, M. (2020) 'Partial Correlation Financial Networks', *Applied Network Science*, 5(1), p. 11. Available at: https://doi.org/10.1007/s41109-020-0251-z.

Millington, T. and Niranjan, M. (2021) 'Construction of Minimum Spanning Trees from Financial Returns using Rank Correlation', *Physica A: Statistical Mechanics and its Applications*, 566, p. 125605. Available at: https://doi.org/10.1016/j.physa.2020.125605.

Münnix, M.C., Schäfer, R. and Guhr, T. (2010) 'Compensating Asynchrony Effects in the Calculation of Financial Correlations', *Physica A: Statistical Mechanics and its Applications*, 389(4), pp. 767–779. Available at: https://doi.org/10.1016/j.physa.2009.10.033.

Musciotto, F., Marotta, L., Miccichè, S. and Mantegna, R.N. (2018) 'Bootstrap Validation of Links of a Minimum Spanning Tree', *Physica A: Statistical Mechanics and its Applications*, 512, pp. 1032–1043. Available at: https://doi.org/10.1016/j.physa.2018.08.020.

Musmeci, N., Aste, T. and Matteo, T.D. (2015) 'Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Methods', *PLOS ONE*, 10(3), p. e0116201. Available at: https://doi.org/10.1371/journal.pone.0116201.

Musmeci, N., Nicosia, V., Aste, T., Di Matteo, T. and Latora, V. (2016) 'The Multiplex Dependency Structure of Financial Markets', *arXiv:1606.04872 [physics, q-fin]* [Preprint]. Available at: http://arxiv.org/abs/1606.04872 (Accessed: 11 May 2020).

Nešetřil, J. (1997) 'A Few Remarks on the History of MST-Problem', *Archivum Mathematicum*, 033, pp. 15–22.

Newman, M. (2010) *Networks: An Introduction*. 1st edition. Oxford ; New York: Oxford University Press.

Newman, M., Barabási, A.-L. and Watts, D.J. (2006) *The Structure and Dynamics of Networks*. 1st edition. Princeton: Princeton University Press.

Nguyen, A.P.N., Mai, T.T., Bezbradica, M. and Crane, M. (2022) 'The Cryptocurrency Market in Transition before and after COVID-19: An Opportunity for Investors?', *Entropy*, 24(9), p. 1317. Available at: https://doi.org/10.3390/e24091317.

Nguyen, Q., Nguyen, N.K.K. and Nguyen, L.H.N. (2019) 'Dynamic Topology and Allometric Scaling Behavior on the Vietnamese Stock Market', *Physica A: Statistical Mechanics and its Applications*, 514, pp. 235–243. Available at: https://doi.org/10.1016/j.physa.2018.09.061.

Oldham, S., Fulcher, B., Parkes, L., Arnatkevičiūtė, A., Suo, C. and Fornito, A. (2019) 'Consistency and Differences Between Centrality Measures Across Distinct Classes of Networks', *PLOS ONE*, 14(7), p. e0220061. Available at: https://doi.org/10.1371/journal.pone.0220061.

Omran, M.G.H., Salman, A. and Engelbrecht, A.P. (2005) 'Dynamic Clustering Using Particle Swarm Optimization with Application in Image Segmentation', *Pattern Analysis and Applications*, 8(4), p. 332. Available at: https://doi.org/10.1007/s10044-005-0015-5.

Onnela, J.-P. (2002) *Taxonomy of Financial Assets*. Available at: http://www.jponnela.com/web_documents/t1.pdf (Accessed: 5 December 2019).

Onnela, J.-P., Chakraborti, A., Kaski, K. and Kertész, J. (2003) 'Dynamic Asset Trees and Black Monday', *Physica A: Statistical Mechanics and its Applications*, 324(1), pp. 247–252. Available at: https://doi.org/10.1016/S0378-4371(02)01882-4.

Onnela, J.-P., Chakraborti, A., Kaski, K., Kertesz, J. and Kanto, A. (2003a) 'Asset Trees and Asset Graphs in Financial Markets', *Physica Scripta*, T106(1), p. 48. Available at: https://doi.org/10.1238/Physica.Topical.106a00048.

Onnela, J.-P., Chakraborti, A., Kaski, K., Kertesz, J. and Kanto, A. (2003b) 'Dynamics of Market Correlations: Taxonomy and Portfolio Analysis', *Physical Review E*, 68(5), p. 056110. Available at: https://doi.org/10.1103/PhysRevE.68.056110.

Pavel N. Krivitsky, M.S.H., Hunter, D.R., Butts, C.T., Klumb, C., Goodreau, S.M. and Morris, M. (2003) 'Statnet: Tools for the Statistical Modeling of Network Data'. Statnet Development Team. Available at: https://statnet.org.

Peralta, G. and Zareei, A. (2016) 'A Network Approach to Portfolio Selection', *Journal of Empirical Finance*, 38, pp. 157–180. Available at: https://doi.org/10.1016/j.jempfin.2016.06.003.

Pike, W. (2015) *Do You Actually Know Who the Influencers Are or Is Your Influencer Marketing a Joke?*, *Bizcommunity*. Available at: https://www.bizcommunity.com/Article/196/669/123688.html (Accessed: 18 October 2022).

Pozzi, F., Di Matteo, T. and Aste, T. (2012) 'Exponential Smoothing Weighted Correlations', *The European Physical Journal B*, 85(6), p. 175. Available at: https://doi.org/10.1140/epjb/e2012-20697-x.

Pozzi, F., Di Matteo, T. and Aste, T. (2013) 'Spread of Risk Across Financial Markets: Better to Invest in the Peripheries', *Scientific Reports*, 3(1), p. 1665. Available at: https://doi.org/10.1038/srep01665.

Prim, R.C. (1957) 'Shortest Connection Networks and Some Generalizations', *The Bell System Technical Journal*, 36(6), pp. 1389–1401. Available at: https://doi.org/10.1002/j.1538-7305.1957.tb01515.x.

Qu, S., Mesomeris, S., Davies, C., Natividade, C., Ward, P., Capra, J., Osiol, J. and Anand, V. (2016) *Mean Reversion II: Pairs Trading Strategies*. Deutche Bank, p. 35.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Raffinot, T. (2017) 'Hierarchical Clustering-Based Asset Allocation', *The Journal of Portfolio Management*, 44(2), pp. 89–99. Available at: https://doi.org/10.3905/jpm.2018.44.2.089.

Rhys, H.I. (2020) *Machine Learning with R, the tidyverse, and mlr*. 1st edition. Shelter Island, NY: Manning.

Ross, S.A. (1976) 'The Arbitrage Theory of Capital Asset Pricing', *Journal of Economic Theory*, 13(3), pp. 341–360. Available at: https://doi.org/10.1016/0022-0531(76)90046-6.

Roy, R.B. and Sarkar, U.K. (2011) 'Identifying Influential Stock Indices from Global Stock Markets: A Social Network Analysis Approach', *Procedia Computer Science*, 5, pp. 442–449. Available at: https://doi.org/10.1016/j.procs.2011.07.057.

Samal, A., Kumar, S., Yadav, Y. and Chakraborti, A. (2021) 'Network-Centric Indicators for Fragility in Global Financial Indices', *Frontiers in Physics*, 8, p. 624373. Available at: https://doi.org/10.3389/fphy.2020.624373.

Sarmah, B., Nair, N., Mehta, D. and Pasquali, S. (2022) 'Learning Embedded Representation of the Stock Correlation Matrix using Graph Machine Learning'. arXiv. Available at: https://doi.org/10.48550/arXiv.2207.07183.

Sarmento, S.M. and Horta, N. (2020) 'Enhancing a Pairs Trading strategy with the application of Machine Learning', *Expert Systems with Applications*, 158, p. 113490. Available at: https://doi.org/10.1016/j.eswa.2020.113490.

Seabrook, I., Caccioli, F. and Aste, T. (2021) 'An Information Filtering approach to stress testing: an application to FTSE markets'. arXiv. Available at: https://doi.org/10.48550/arXiv.2106.08778.

Sharpe, W.F. (1963) ‘A Simplified Model for Portfolio Analysis’, *Management Science*, 9(2), pp. 277–293. Available at: https://doi.org/10.1287/mnsc.9.2.277.

Sieczka, P. and Hołyst, J.A. (2009) ‘Correlations in Commodity Markets’, *Physica A: Statistical Mechanics and its Applications*, 388(8), pp. 1621–1630. Available at: https://doi.org/10.1016/j.physa.2009.01.004.

Song, D.-M., Tumminello, M., Zhou, W.-X. and Mantegna, R.N. (2011) ‘Evolution of Worldwide Stock Markets, Correlation Structure and Correlation Based Graphs’, *Physical Review E*, 84(2), p. 026108. Available at: https://doi.org/10.1103/PhysRevE.84.026108.

Song, J.Y., Chang, W. and Song, J.W. (2019) ‘Cluster Analysis on the Structure of the Cryptocurrency Market Via Bitcoin–Ethereum Filtering’, *Physica A: Statistical Mechanics and its Applications*, 527, p. 121339. Available at: https://doi.org/10.1016/j.physa.2019.121339.

Song, W.-M., Di Matteo, T. and Aste, T. (2011) ‘Nested Hierarchies in Planar Graphs’, *Discrete Applied Mathematics*, 159(17), pp. 2135–2146. Available at: https://doi.org/10.1016/j.dam.2011.07.018.

Song, W.-M., Di Matteo, T. and Aste, T. (2012) ‘Hierarchical information clustering by means of topologically embedded graphs’, *PLoS ONE*, 7(3), p. e31929. Available at: https://doi.org/10.1371/journal.pone.0031929.

Stein, C. (1956) ‘Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution’, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics*. University of California Press, p. 197.

Stein, C. (1975) ‘Estimation of a Covariance Matrix’, in *39th Annual Meeting IMS, Atlanta, GA, 1975*.

Tabak, B.M., Serra, T.R. and Cajueiro, D.O. (2010a) ‘Topological Properties of Commodities Networks’, *The European Physical Journal B*, 74(2), pp. 243–249. Available at: https://doi.org/10.1140/epjb/e2010-00079-4.

Tabak, B.M., Serra, T.R. and Cajueiro, D.O. (2010b) ‘Topological Properties of Stock Market Networks: The Case of Brazil’, *Physica A: Statistical Mechanics and its Applications*, 389(16), pp. 3240–3249. Available at: https://doi.org/10.1016/j.physa.2010.04.002.

Tufte, E.R. (2001) *The Visual Display of Quantitative Information, 2nd Ed.* 2nd edition. Cheshire, Conn: Graphics Press.

Tumminello, M., Aste, T., Di Matteo, T. and Mantegna, R.N. (2005) ‘A Tool for Filtering Information in Complex Systems’, *Proceedings of the National Academy of Sciences*, 102(30), pp. 10421–10426. Available at: https://doi.org/10.1073/pnas.0500298102.

Tumminello, M., Coronnello, C., Lillo, F., Micciche’, S. and Mantegna, R.N. (2007) ‘Spanning Trees and Bootstrap Reliability Estimation in Correlation Based Networks’, *International Journal of Bifurcation and Chaos*, 17(07), pp. 2319–2329. Available at: https://doi.org/10.1142/S0218127407018415.

Turiel, J.D., Barucca, P. and Aste, T. (2020) ‘Simplicial Persistence of Financial Markets: Filtering, Generative Processes and Portfolio Risk’. arXiv. Available at: https://doi.org/10.48550/arXiv.2009.08794.

Ullmann, T., Hennig, C. and Boulesteix, A.-L. (2022) ‘Validation of Cluster Analysis Results on Validation Data: A Systematic Framework’, *WIREs Data Mining and Knowledge Discovery*, 12(3), p. e1444. Available at: https://doi.org/10.1002/widm.1444.

Van Rossum, G. and Drake, F.L. (2009) *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vass, C. (2019) *Industry Classification Benchmark (ICB) Reclassification*. FTSE Russell. Available at: https://content.ftserussell.com/sites/default/files/research/Industry%20Classification%20Benchmark%20reclassification-expanded%20and%20improved%20FINAL.pdf (Accessed: 5 October 2022).

Výrost, T., Lyócsa, Š. and Baumöhl, E. (2019) ‘Network-Based Asset Allocation Strategies’, *The North American Journal of Economics and Finance*, 47, pp. 516–536. Available at: https://doi.org/10.1016/j.najef.2018.06.008.

Wang, G.-J. and Xie, C. (2015) ‘Correlation Structure and Dynamics of International Real Estate Securities Markets: A Network Perspective’, *Physica A: Statistical Mechanics and its Applications*, 424, pp. 176–193. Available at: https://doi.org/10.1016/j.physa.2015.01.025.

Wang, G.-J. and Xie, C. (2016) ‘Tail Dependence Structure of the Foreign Exchange Market: A Network View’, *Expert Systems with Applications*, 46, pp. 164–179. Available at: https://doi.org/10.1016/j.eswa.2015.10.037.

Wang, G.-J., Xie, C., Chen, Y.-J. and Chen, S. (2013) ‘Statistical Properties of the Foreign Exchange Network at Different Time Scales: Evidence from Detrended Cross-Correlation Coefficient and Minimum Spanning Tree’, *Entropy*, 15(5), pp. 1643–1662. Available at: https://doi.org/10.3390/e15051643.

Wang, G.-J., Xie, C., Han, F. and Sun, B. (2012) ‘Similarity Measure and Topology Evolution of Foreign Exchange Markets Using Dynamic Time Warping Method: Evidence from Minimal Spanning Tree’, *Physica A: Statistical Mechanics and its Applications*, 391(16), pp. 4136–4146. Available at: https://doi.org/10.1016/j.physa.2012.03.036.

Wang, G.-J., Xie, C., Zhang, P., Han, F. and Chen, S. (2014) ‘Dynamics of Foreign Exchange Networks: A Time-Varying Copula Approach’, *Discrete Dynamics in Nature and Society*, 2014, p. e170921. Available at: https://doi.org/10.1155/2014/170921.

Wen, F., Yang, X. and Zhou, W.-X. (2019) ‘Tail Dependence Networks of Global Stock Markets’, *International Journal of Finance & Economics*, 24(1), pp. 558–567. Available at: https://doi.org/10.1002/ijfe.1679.

Wigner, E.P. (1955) ‘Characteristic Vectors of Bordered Matrices With Infinite Dimensions’, *Annals of Mathematics*, 62(3), pp. 548–564. Available at: https://doi.org/10.2307/1970079.

Winton (2018) *Systematic Methods for Classifying Equities*. Available at: https://assets.winton.com/cms/Images/Research/Systematic-Methods-for-Classifying-Equities/2018-08_Winton-Research_Systematic-Methods-for-Classifying-Equities.pdf (Accessed: 20 June 2019).

Xu, Q., Wang, L., Jiang, C., Jia, F. and Chen, L. (2022) ‘Tail Dependence Network of New Energy Vehicle Industry in Mainland China’, *Annals of Operations Research*, 315(1), pp. 565–590. Available at: https://doi.org/10.1007/s10479-022-04729-w.

Yan, X.-G., Xie, C. and Wang, G.-J. (2015) ‘Stock Market Network’s Topological Stability: Evidence from Planar Maximally Filtered Graph and Minimal Spanning Tree’, *International Journal of Modern Physics B*, 29(22), p. 1550161. Available at: https://doi.org/10.1142/S0217979215501611.

Yan, Y., Wu, B., Tian, T. and Zhang, H. (2020) ‘Development of Stock Networks Using Part Mutual Information and Australian Stock Market Data’, *Entropy*, 22(7), p. 773. Available at: https://doi.org/10.3390/e22070773.

Zhao, L., Li, W. and Cai, X. (2016) ‘Structure and Dynamics of Stock Market in Times of Crisis’, *Physics Letters A*, 380(5), pp. 654–666. Available at: https://doi.org/10.1016/j.physleta.2015.11.015.

Appendix A Universe of Stocks

A1 Full List of Shares and ICB Classifications

Table A-1 contains the full list of the 136 stocks that we considered for the analysis, along with the JSE share code, company names, and the various ICB classifications.

Table A-1 Full List of Universe of Stocks in ALSI (Jun22) and ICB Classifications

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
ABG	Absa Group Ltd	Financials	Banks	Banks	Banks
ACL	ArcelorMittal South Africa Ltd	Basic Materials	Basic Resources	Industrial Metals and Mining	Iron and Steel
ADH	Advtech Ltd	Consumer Discretionary	Consumer Products and Services	Consumer Services	Education Services
AEL	Altron Ltd	Technology	Technology	Software and Computer Services	Computer Services
AFE	AECI Ltd	Basic Materials	Chemicals	Chemicals	Chemicals: Diversified
AFH	Alexander Forbes Group Holding	Financials	Financial Services	Investment Banking and Brokerage Services	Asset Managers and Custodians
AFT	Afrimat Ltd	Industrials	Construction and Materials	Construction and Materials	Building Materials: Other
AGL	Anglo American PLC	Basic Materials	Basic Resources	Industrial Metals and Mining	General Mining

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
AIL	African Rainbow Capital Invest	Financials	Financial Services	Closed End Investments	Closed End Investments
AIP	Adcock Ingram Holdings Ltd	Health Care	Health Care	Pharmaceuticals and Biotechnology	Pharmaceuticals
AMS	Anglo American Platinum Ltd	Basic Materials	Basic Resources	Precious Metals and Mining	Platinum and Precious Metals
ANG	AngloGold Ashanti Ltd	Basic Materials	Basic Resources	Precious Metals and Mining	Gold Mining
ANH	Anheuser-Busch InBev SA/NV	Consumer Staples	Food, Beverage and Tobacco	Beverages	Brewers
APN	Aspen Pharmacare Holdings Ltd	Health Care	Health Care	Pharmaceuticals and Biotechnology	Pharmaceuticals
ARI	African Rainbow Minerals Ltd	Basic Materials	Basic Resources	Industrial Metals and Mining	General Mining
ARL	Astral Foods Ltd	Consumer Staples	Food, Beverage and Tobacco	Food Producers	Farming, Fishing, Ranching and Plantations
ATT	Attacq Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
AVI	AVI Ltd	Consumer Staples	Food, Beverage and Tobacco	Food Producers	Food Products
BAT	Brait PLC	Financials	Financial Services	Investment Banking and Brokerage Services	Diversified Financial Services
BAW	Barloworld Ltd	Industrials	Industrial Goods and Services	General Industrials	Diversified Industrials
BHG	BHP Group Ltd	Basic Materials	Basic Resources	Industrial Metals and Mining	General Mining
BID	Bid Corp Ltd	Consumer Staples	Personal Care, Drug and Grocery Stores	Personal Care, Drug and Grocery Stores	Food Retailers and Wholesalers
BLU	Blue Label Telecoms Ltd	Telecoms	Telecoms	Telecoms Service Providers	Telecoms Services

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
BTI	British American Tobacco PLC	Consumer Staples	Food, Beverage and Tobacco	Tobacco	Tobacco
BVT	Bidvest Group Ltd/The	Industrials	Industrial Goods and Services	General Industrials	Diversified Industrials
BYI	Bytes Technology Group PLC	Technology	Technology	Software and Computer Services	Software
CCO	Capital & Counties Properties	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
CFR	Cie Financiere Richemont SA	Consumer Discretionary	Consumer Products and Services	Personal Goods	Luxury Items
CLH	City Lodge Hotels Ltd	Consumer Discretionary	Travel and Leisure	Travel and Leisure	Hotels and Motels
CLS	Clicks Group Ltd	Consumer Staples	Personal Care, Drug and Grocery Stores	Personal Care, Drug and Grocery Stores	Drug Retailers
CML	Coronation Fund Managers Ltd	Financials	Financial Services	Investment Banking and Brokerage Services	Asset Managers and Custodians
COH	Curro Holdings Ltd	Consumer Discretionary	Consumer Products and Services	Consumer Services	Education Services
CPI	Capitec Bank Holdings Ltd	Financials	Banks	Banks	Banks
CSB	Cashbuild Ltd	Consumer Discretionary	Retail	Retailers	Home Improvement Retailers
DCP	Dis-Chem Pharmacies Ltd	Consumer Staples	Personal Care, Drug and Grocery Stores	Personal Care, Drug and Grocery Stores	Drug Retailers
DGH	Distell Group Holdings Ltd	Consumer Staples	Food, Beverage and Tobacco	Beverages	Distillers and Vintners
DRD	DRDGOLD Ltd	Basic Materials	Basic Resources	Precious Metals and Mining	Gold Mining

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
DSY	Discovery Ltd	Financials	Insurance	Life Insurance	Life Insurance
DTC	DataTec Ltd	Technology	Technology	Software and Computer Services	Computer Services
EMI	Emira Property Fund Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
EQU	Equites Property Fund Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Industrial REITs
EXX	Exxaro Resources Ltd	Energy	Energy	Oil Gas and Coal	Coal
FBR	Famous Brands Ltd	Consumer Discretionary	Travel and Leisure	Travel and Leisure	Restaurants and Bars
FFA	Fortress REIT Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
FFB	Fortress REIT Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
FSR	FirstRand Ltd	Financials	Banks	Banks	Banks
FTB	Fairvest Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
GFI	Gold Fields Ltd	Basic Materials	Basic Resources	Precious Metals and Mining	Gold Mining
GLN	Glencore PLC	Basic Materials	Basic Resources	Industrial Metals and Mining	General Mining
GND	Grindrod Ltd	Industrials	Industrial Goods and Services	Industrial Transportation	Transportation Services
GRT	Growthpoint Properties Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
HAR	Harmony Gold Mining Co Ltd	Basic Materials	Basic Resources	Precious Metals and Mining	Gold Mining
HCI	Hosken Consolidated Investment	Financials	Financial Services	Investment Banking and Brokerage Services	Diversified Financial Services

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
HDC	Hudaco Industries Ltd	Industrials	Industrial Goods and Services	Industrial Support Services	Industrial Suppliers
HMN	Hammerson PLC	Real Estate	Real Estate	Real Estate Investment Trusts	Retail REITs
HYP	Hyprop Investments Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Retail REITs
IAP	Irongate Property Fund I	Real Estate	Real Estate	Real Estate Investment Trusts	Office REITs
IMP	Impala Platinum Holdings Ltd	Basic Materials	Basic Resources	Precious Metals and Mining	Platinum and Precious Metals
INP	Investec PLC	Financials	Banks	Banks	Banks
IPF	Investec Property Fund Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
ITE	Italtile Ltd	Consumer Discretionary	Retail	Retailers	Home Improvement Retailers
JSE	JSE Ltd	Financials	Financial Services	Investment Banking and Brokerage Services	Investment Services
KAP	KAP Industrial Holdings Ltd	Industrials	Industrial Goods and Services	General Industrials	Diversified Industrials
KIO	Kumba Iron Ore Ltd	Basic Materials	Basic Resources	Industrial Metals and Mining	Iron and Steel
KRO	Karooooo Ltd	Technology	Technology	Software and Computer Services	Software
KST	PSG Konsult Ltd	Financials	Financial Services	Investment Banking and Brokerage Services	Diversified Financial Services
L2D	Liberty Two Degrees Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Retail REITs
L4L	Long4Life Ltd	Consumer Discretionary	Consumer Products and Services	Leisure Goods	Recreational Products

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
LBR	Libstar Holdings Ltd	Consumer Staples	Food, Beverage and Tobacco	Food Producers	Food Products
LHC	Life Healthcare Group Holdings	Health Care	Health Care	Health Care Providers	Health Care Facilities
LTE	Lighthouse Properties plc	Real Estate	Real Estate	Real Estate Investment and Services	Real Estate Holding and Development
MCG	MultiChoice Group	Telecoms	Telecoms	Telecoms Service Providers	Cable Television Services
MEI	Mediclinic International PLC	Health Care	Health Care	Health Care Providers	Health Care Facilities
MKR	Montauk Renewables Inc	Energy	Energy	Alternative Energy	Alternative Fuels
MLI	Industrials REIT Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Industrial REITs
MNP	Mondi PLC	Industrials	Industrial Goods and Services	General Industrials	Containers and Packaging
MRP	Mr Price Group Ltd	Consumer Discretionary	Retail	Retailers	Apparel Retailers
MSM	Massmart Holdings Ltd	Consumer Discretionary	Retail	Retailers	Diversified Retailers
MSP	MAS P.L.C.	Real Estate	Real Estate	Real Estate Investment and Services	Real Estate Holding and Development
MTA	Metair Investments Ltd	Consumer Discretionary	Automobiles and Parts	Automobiles and Parts	Auto Parts
MTH	Motus Holdings Ltd	Consumer Discretionary	Retail	Retailers	Specialty Retailers
MTM	Momentum Metropolitan Holdings	Financials	Insurance	Life Insurance	Life Insurance
MTN	MTN Group Ltd	Telecoms	Telecoms	Telecoms Service Providers	Telecoms Services

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
MUR	Murray & Roberts Holdings Ltd	Industrials	Construction and Materials	Construction and Materials	Engineering and Contracting Services
N91	Ninety One PLC	Financials	Financial Services	Investment Banking and Brokerage Services	Asset Managers and Custodians
NED	Nedbank Group Ltd	Financials	Banks	Banks	Banks
NPH	Northam Platinum Holdings Ltd	Basic Materials	Basic Resources	Precious Metals and Mining	Platinum and Precious Metals
NPN	Naspers Ltd	Technology	Technology	Software and Computer Services	Consumer Digital Services
NRP	NEPI Rockcastle NV	Real Estate	Real Estate	Real Estate Investment and Services	Real Estate Holding and Development
NTC	Netcare Ltd	Health Care	Health Care	Health Care Providers	Health Care Facilities
OCE	Oceana Group Ltd	Consumer Staples	Food, Beverage and Tobacco	Food Producers	Farming, Fishing, Ranching and Plantations
OMN	Omnia Holdings Ltd	Basic Materials	Chemicals	Chemicals	Chemicals: Diversified
OMU	Old Mutual Ltd	Financials	Insurance	Life Insurance	Life Insurance
PAN	Pan African Resources PLC	Basic Materials	Basic Resources	Precious Metals and Mining	Gold Mining
PIK	Pick n Pay Stores Ltd	Consumer Staples	Personal Care, Drug and Grocery Stores	Personal Care, Drug and Grocery Stores	Food Retailers and Wholesalers
PPC	PPC Ltd	Industrials	Construction and Materials	Construction and Materials	Cement
PPH	Pepkor Holdings Ltd	Consumer Discretionary	Retail	Retailers	Diversified Retailers
PSG	PSG Group Ltd	Financials	Financial Services	Investment Banking and Brokerage Services	Diversified Financial Services
QLT	Quilter PLC	Financials	Financial Services	Investment Banking and Brokerage Services	Asset Managers and Custodians

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
RBP	Royal Bafokeng Platinum Ltd	Basic Materials	Basic Resources	Precious Metals and Mining	Platinum and Precious Metals
RBX	Raubex Group Ltd	Industrials	Construction and Materials	Construction and Materials	Construction
RDF	Redefine Properties Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
REM	Remgro Ltd	Financials	Financial Services	Investment Banking and Brokerage Services	Diversified Financial Services
RES	Resilient REIT Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Retail REITs
RFG	RFG Holdings Ltd	Consumer Staples	Food, Beverage and Tobacco	Food Producers	Food Products
RLO	Reunert Ltd	Industrials	Industrial Goods and Services	Electronic and Electrical Equipment	Electrical Components
RMI	Rand Merchant Investment Holdi	Financials	Financial Services	Investment Banking and Brokerage Services	Diversified Financial Services
SAC	SA Corporate Real Estate Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Diversified REITs
SAP	Sappi Ltd	Basic Materials	Basic Resources	Industrial Materials	Paper
SBK	Standard Bank Group Ltd	Financials	Banks	Banks	Banks
SHP	Shoprite Holdings Ltd	Consumer Staples	Personal Care, Drug and Grocery Stores	Personal Care, Drug and Grocery Stores	Food Retailers and Wholesalers
SLM	Sanlam Ltd	Financials	Insurance	Life Insurance	Life Insurance
SNH	Steinhoff International Holdin	Consumer Discretionary	Retail	Retailers	Diversified Retailers
SNT	Santam Ltd	Financials	Insurance	Non-life Insurance	Property and Casualty Insurance
SOL	Sasol Ltd	Basic Materials	Chemicals	Chemicals	Chemicals: Diversified

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
SPG	Super Group Ltd/South Africa	Industrials	Industrial Goods and Services	Industrial Transportation	Transportation Services
SPP	SPAR Group Ltd/The	Consumer Staples	Personal Care, Drug and Grocery Stores	Personal Care, Drug and Grocery Stores	Food Retailers and Wholesalers
SRE	Sirius Real Estate Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Office REITs
SSS	Stor-Age Property REIT Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Storage REITs
SSW	Sibanye Stillwater Ltd	Basic Materials	Basic Resources	Precious Metals and Mining	Platinum and Precious Metals
SUI	Sun International Ltd/South Af	Consumer Discretionary	Travel and Leisure	Travel and Leisure	Casinos and Gambling
TBS	Tiger Brands Ltd	Consumer Staples	Food, Beverage and Tobacco	Food Producers	Food Products
TCP	Transaction Capital Ltd	Financials	Financial Services	Finance and Credit Services	Consumer Lending
TFG	Foschini Group Ltd/The	Consumer Discretionary	Retail	Retailers	Apparel Retailers
TGA	Thungela Resources Ltd	Energy	Energy	Oil Gas and Coal	Coal
TGO	Southern Sun Ltd	Consumer Discretionary	Travel and Leisure	Travel and Leisure	Hotels and Motels
THA	Tharisa PLC	Basic Materials	Basic Resources	Industrial Metals and Mining	General Mining
TKG	Telkom SA SOC Ltd	Telecoms	Telecoms	Telecoms Service Providers	Telecoms Services
TRU	Truworths International Ltd	Consumer Discretionary	Retail	Retailers	Apparel Retailers
TSG	Tsogo Sun Gaming Ltd	Consumer Discretionary	Travel and Leisure	Travel and Leisure	Casinos and Gambling

JSE Share Code	Company Name	ICB Industry	ICB Super-sector	ICB Sector	ICB Subsector
TXT	Textainer Group Holdings Ltd	Industrials	Industrial Goods and Services	Industrial Transportation	Transportation Services
VKE	Vukile Property Fund Ltd	Real Estate	Real Estate	Real Estate Investment Trusts	Retail REITs
VOD	Vodacom Group Ltd	Telecoms	Telecoms	Telecoms Service Providers	Telecoms Services
WBO	Wilson Bayly Holmes-Ovcon Ltd	Industrials	Construction and Materials	Construction and Materials	Engineering and Contracting Services
WHL	Woolworths Holdings Ltd/South	Consumer Discretionary	Retail	Retailers	Diversified Retailers
ZED	Zeder Investments Ltd	Financials	Financial Services	Investment Banking and Brokerage Services	Asset Managers and Custodians

A2 Exclusion Process and Final List of Shares

Table A-2 shows the stocks that were excluded due to insufficient history and those that were excluded due to liquidity concerns, leaving us with the final list of 72 stocks in the final column of the table.

Table A-2 Screening of Stocks to Final List

ALSI Stocks	Company Name	Sufficient History	Trading Liquidity	Final Universe of Stocks
ABG	Absa Group Ltd	✓	✓	✓
ACL	ArcelorMittal South Africa Ltd	✓	✓	✓
ADH	Advtech Ltd	✓	✓	✓
AEL	Altron Ltd	✓	✗	
AFE	AECI Ltd	✓	✓	✓
AFH	Alexander Forbes Group Holding	✗		
AFT	Afrimat Ltd	✗		

ALSI Stocks	Company Name	Sufficient History	Trading Liquidity	Final Universe of Stocks
AGL	Anglo American PLC	✓	✓	✓
AIL	African Rainbow Capital Invest	✗		
AIP	Adcock Ingram Holdings Ltd	✗		
AMS	Anglo American Platinum Ltd	✓	✓	✓
ANG	AngloGold Ashanti Ltd	✓	✓	✓
ANH	Anheuser-Busch InBev SA/NV	✓	✓	✓
APN	Aspen Pharmacare Holdings Ltd	✓	✓	✓
ARI	African Rainbow Minerals Ltd	✓	✓	✓
ARL	Astral Foods Ltd	✓	✓	✓
ATT	Attacq Ltd	✗		
AVI	AVI Ltd	✓	✓	✓
BAT	Brait PLC	✓	✓	✓
BAW	Barloworld Ltd	✓	✓	✓
BHG	BHP Group Ltd	✓	✓	✓
BID	Bid Corp Ltd	✗		
BLU	Blue Label Telecoms Ltd	✗		
BTI	British American Tobacco PLC	✓	✓	✓
BVT	Bidvest Group Ltd/The	✓	✓	✓
BYI	Bytes Technology Group PLC	✗		
CCO	Capital & Counties Properties	✗		
CFR	Cie Financiere Richemont SA	✓	✓	✓
CLH	City Lodge Hotels Ltd	✓	✓	✓
CLS	Clicks Group Ltd	✓	✓	✓
CML	Coronation Fund Managers Ltd	✗		
COH	Curro Holdings Ltd	✗		
CPI	Capitec Bank Holdings Ltd	✓	✓	✓
CSB	Cashbuild Ltd	✓	✓	✓

ALSI Stocks	Company Name	Sufficient History	Trading Liquidity	Final Universe of Stocks
DCP	Dis-Chem Pharmacies Ltd	✗		
DGH	Distell Group Holdings Ltd	✓	✗	
DRD	DRDGOLD Ltd	✓	✓	✓
DSY	Discovery Ltd	✓	✓	✓
DTC	DataTec Ltd	✓	✓	✓
EMI	Emira Property Fund Ltd	✗		
EQU	Equites Property Fund Ltd	✗		
EXX	Exxaro Resources Ltd	✓	✓	✓
FBR	Famous Brands Ltd	✓	✓	✓
FFA	Fortress REIT Ltd	✗		
FFB	Fortress REIT Ltd	✗		
FSR	FirstRand Ltd	✓	✓	✓
FTB	Fairvest Ltd	✗		
GFI	Gold Fields Ltd	✓	✓	✓
GLN	Glencore PLC	✗		
GND	Grindrod Ltd	✓	✓	✓
GRT	Growthpoint Properties Ltd	✓	✓	✓
HAR	Harmony Gold Mining Co Ltd	✓	✓	✓
HCI	Hosken Consolidated Investment	✗		
HDC	Hudaco Industries Ltd	✓	✓	✓
HMN	Hammerson PLC	✗		
HYP	Hyprop Investments Ltd	✓	✓	✓
IAP	Irongate Property Fund I	✗		
IMP	Impala Platinum Holdings Ltd	✓	✓	✓
INP	Investec PLC	✓	✓	✓
IPF	Investec Property Fund Ltd	✗		
ITE	Italtile Ltd	✓	✗	

ALSI Stocks	Company Name	Sufficient History	Trading Liquidity	Final Universe of Stocks
JSE	JSE Ltd	✗		
KAP	KAP Industrial Holdings Ltd	✓	✗	
KIO	Kumba Iron Ore Ltd	✗		
KRO	Karooooo Ltd	✗		
KST	PSG Konsult Ltd	✗		
L2D	Liberty Two Degrees Ltd	✗		
L4L	Long4Life Ltd	✗		
LBR	Libstar Holdings Ltd	✗		
LHC	Life Healthcare Group Holdings	✗		
LTE	Lighthouse Properties plc	✗		
MCG	MultiChoice Group	✗		
MEI	Mediclinic International PLC	✓	✓	✓
MKR	Montauk Renewables Inc	✗		
MLI	Industrials REIT Ltd	✗		
MNP	Mondi PLC	✗		
MRP	Mr Price Group Ltd	✓	✓	✓
MSM	Massmart Holdings Ltd	✓	✓	✓
MSP	MAS P.L.C.	✗		
MTA	Metair Investments Ltd	✓	✗	
MTH	Motus Holdings Ltd	✗		
MTM	Momentum Metropolitan Holdings	✓	✓	✓
MTN	MTN Group Ltd	✓	✓	✓
MUR	Murray & Roberts Holdings Ltd	✓	✓	✓
N91	Ninety One PLC	✗		
NED	Nedbank Group Ltd	✓	✓	✓
NPH	Northam Platinum Holdings Ltd	✓	✓	✓
NPN	Naspers Ltd	✓	✓	✓

ALSI Stocks	Company Name	Sufficient History	Trading Liquidity	Final Universe of Stocks
NRP	NEPI Rockcastle NV	✗		
NTC	Netcare Ltd	✓	✓	✓
OCE	Oceana Group Ltd	✓	✗	
OMN	Omnia Holdings Ltd	✓	✓	✓
OMU	Old Mutual Ltd	✓	✓	✓
PAN	Pan African Resources PLC	✗		
PIK	Pick n Pay Stores Ltd	✓	✓	✓
PPC	PPC Ltd	✓	✓	✓
PPH	Pepkor Holdings Ltd	✗		
PSG	PSG Group Ltd	✓	✓	✓
QLT	Quilter PLC	✗		
RBP	Royal Bafokeng Platinum Ltd	✗		
RBX	Raubex Group Ltd	✗		
RDF	Redefine Properties Ltd	✓	✓	✓
REM	Remgro Ltd	✓	✓	✓
RES	Resilient REIT Ltd	✓	✓	✓
RFG	RFG Holdings Ltd	✗		
RLO	Reunert Ltd	✓	✓	✓
RMI	Rand Merchant Investment Holdi	✗		
SAC	SA Corporate Real Estate Ltd	✓	✓	✓
SAP	Sappi Ltd	✓	✓	✓
SBK	Standard Bank Group Ltd	✓	✓	✓
SHP	Shoprite Holdings Ltd	✓	✓	✓
SLM	Sanlam Ltd	✓	✓	✓
SNH	Steinhoff International Holdin	✓	✓	✓
SNT	Santam Ltd	✓	✓	✓
SOL	Sasol Ltd	✓	✓	✓

ALSI Stocks	Company Name	Sufficient History	Trading Liquidity	Final Universe of Stocks
SPG	Super Group Ltd/South Africa	✓	✓	✓
SPP	SPAR Group Ltd/The	✗		
SRE	Sirius Real Estate Ltd	✗		
SSS	Stor-Age Property REIT Ltd	✗		
SSW	Sibanye Stillwater Ltd	✗		
SUI	Sun International Ltd/South Af	✓	✓	✓
TBS	Tiger Brands Ltd	✓	✓	✓
TCP	Transaction Capital Ltd	✗		
TFG	Foschini Group Ltd/The	✓	✓	✓
TGA	Thungela Resources Ltd	✗		
TGO	Southern Sun Ltd	✗		
THA	Tharisa PLC	✗		
TKG	Telkom SA SOC Ltd	✓	✓	✓
TRU	Truworths International Ltd	✓	✓	✓
TSG	Tsogo Sun Gaming Ltd	✓	✓	✓
TXT	Textainer Group Holdings Ltd	✗		
VKE	Vukile Property Fund Ltd	✗		
VOD	Vodacom Group Ltd	✗		
WBO	Wilson Bayly Holmes-Ovcon Ltd	✓	✓	✓
WHL	Woolworths Holdings Ltd/South	✓	✓	✓
ZED	Zeder Investments Ltd	✗		

Appendix B Source Code

This section contains tables with short descriptions of the MATLAB functions and scripts that were used to conduct the analysis. The scripts were used to set up the input variables, and to call the relevant functions for each piece of analysis.

Samples of the actual code can be found at https://github.com/YashinG/Network_Filtering_Matlab.

B1 MATLAB Functions

Table B-1 List of General MATLAB Functions

Function	Description
preProcessRets.m	Performs the pre-processing tasks of standardising the returns and removing the market mode.
EWMA_Wts.m	Calculates the exponential weights for a given number of observations and a given exponential decay factor.
covMatWtd.m	Calculates a weighted covariance matrix, i.e. each observation can have a different weight.
QIS_Wtd.m	Calculates a covariance matrix using the QIS method of Ledoit and Wolf to reduce the impact of noise. This function has been adapted to cater for weighted observations.
correlToDistMetric.m	Converts a correlation matrix to a distance matrix using Equation (3).
networkMetrics.m	Calculates the various metrics that describe the topology of a network, e.g. centrality metrics, normalised tree length etc.

Table B-2 List of MATLAB Functions to Analyse Correlation Matrices

Function	Description
correlationAnalysis.m	Determines the distribution of cross correlations in the correlation matrix over a static period.
correlationAnalysis_Dynamic.m	Determines the distribution of cross correlations in the covariance matrix dynamically through time.

Table B-3 List of MATLAB Functions to Run the Network Filters

Function	Description
getFilteredNetwork.m	Filters the correlation matrix using either the MST or the PMFG methods over a static period.
getFilteredNetwork_Dynamic.m	A function to run the network filters dynamically through time.

Table B-4 List of MATLAB Functions to Run the Cluster Analysis

Function	Description
DBHTs.m	Performs DBHT clustering. This function was adapted to allow the creation of a DBHT from a PMFG or TMFG.
getClusters.m	Get the hierarchical clustering for various linkages, including the DBHT method over a static period.
getClusters_Dynamic.m	Get the hierarchical clustering for various linkages, including the DBHT method, dynamically through time.

Table B-5 List of MATLAB Functions to Run the Bootstrap Reliability Analysis

Function	Description
bootstrapDBHT.m	Use bootstrap resampling to assess the reliability of the number of clusters that results from the DBHT.
bootstrapNetwork.m	Use bootstrap resampling to assess the reliability of edges in a network and network metrics.

B2 MATLAB Scripts

Table B-6 List of MATLAB Scripts That Were Used to Run the Analysis

Script	Description
scr_RunBootstrapDBHT.m	A script to run the bootstrap reliability analysis of the number of clusters that results from the DBHT.
scr_RunBootstrapNetwork.m	A script to assess the reliability of the edges in a network and the reliability of the network metrics using bootstrap resampling.
scr_RunClustering.m	A script to determine the hierarchical clusters over a static period.
scr_RunClustering_Dynamic.m	A script to determine the hierarchical clusters dynamically through time.
scr_RunCorrelAnalysis.m	A script to determine the distribution of cross correlations in the correlation matrix over a static period, as well as dynamically.
scr_RunCurrent.m	A script to run the latest network filter and DBHT.
scr_RunNetworkFilter.m	A script to run a network filter over a static period.
scr_RunNetworkFilter_Dynamic.m	A script to run a network filter dynamically through time.

Below is an example of the script that was used to create the static PMFG (i.e. the scr_RunNetworkFilter.m).

```
% Script to run MST or PMFG and plot in MATLAB

% Add paths to relevant folders
addpath('../Functions');
addpath('../External Functions/PMFG')
addpath('../External Functions/matlab_bgl-4.0.1/matlab_bgl')

% Load data
load('../Data/Data.mat')

% Data inputs
stInputs.data.inReturns = stStockRets.rets;
stInputs.data.inNames   = stStockRets.aNames;
stInputs.data.inDates   = stStockRets.mlDates;
stInputs.data.pajekLabels = stStockRets.pajekStrData;

% Pre-processing inputs
stInputs.preProp.blnStdize     = 1;
stInputs.preProp.blnRemMktMode = 1;
stInputs.preProp.EWMA_Alpha    = 0; % {0, 0.005}

% Network inputs
stInputs.network.distanceMethod = 'QIS_correlDistMetric';  % {'QIS_correlDistMetric','correlDistMetric'}
stInputs.network.filter          = 'PMFG';  % {'MST','PMFG'}
stInputs.network.blnPlot         = 1;

% Get EWMA weights
stInputs.data.timeWts = EWMA_Wts(size(stInputs.data.inReturns,1), stInputs.preProp.EWMA_Alpha);

% Run filter
% stOut = getFilteredNetwork(inReturns, inTimeWts, inNames, inNetworkFilter, ...
%    inBlnStdize, inBlnRemMktMode, inDistMethod, inPlotID)
clear stOutNetworkFilter

stOutNetworkFilter = getFilteredNetwork(stInputs.data.inReturns, stInputs.data.timeWts,...
    stInputs.data.inNames, stInputs.network.filter, stInputs.preProp.blnStdize,...
    stInputs.preProp.blnRemMktMode, stInputs.network.distanceMethod, stInputs.network.blnPlot);

stOutNetworkFilter.dateStamp = datestr(clock,'_yyyymmdd_HHMMSS');

% Store results and generate pajek file
stOutNetworkFilter.edgeData = table2array(graph(stOutNetworkFilter.filteredNetwork_D).Edges);

stOutNetworkFilter.filename = [stOutNetworkFilter.dateStamp,' ', stInputs.network.filter ,' Method_', ...
    stInputs.network.distanceMethod,' ModesRem_' ,num2str(stInputs.preProp.blnRemMktMode),' EWMA_' ,...
    num2str(stInputs.preProp.EWMA_Alpha)];

% Create pajek .net file
adj2pajek2(stOutNetworkFilter.edgeData, stInputs.data.pajekLabels, stOutNetworkFilter.filename, 0);

% Store results in labelled variable
eval(['stOut_',stInputs.network.filter,'_', stInputs.network.distanceMethod,'= stOutNetworkFilter;']);
```