Visualization of a Stock Market Correlation Matrix

Alethea Rea[1], William Rea[2]
1. Data Analysis Australia, Perth, Australia
2. Department of Economics and Finance, University of Canterbury, New Zealand

November 13, 2012

Abstract

This paper presents a novel application of software developed for constructing a phylogenetic network to the correlation matrix for 48 stocks listed on the New Zealand Stock Exchange. We show that by visualizing the correlation matrix using a Neighbor-Net network some of the problems associated with understanding the large number of correlations between the individual stocks can be overcome. This yields greater insight into how closely individual stocks are related to each other in terms of their correlations and suggests new avenues of research into how to construct small diversified stock portfolios.

Keywords: Visualization, Neighbour-Nets, Correlation Matix, Diversification

1 Introduction

For a single investment opportunity, potential investors are interested in the expected return (the mean or average return) and the potential spread of the return (that is the variance). When considering two or more investment opportunities there is also an interest in the correlation, the extent to which investments move together. As Epstein and Tanny (1980) note, the more risk averse the individual the greater likelihood that the investor is also "correlation averse". Therefore all investors need to quantify and consider correlations between investments.

However, the problem is that the number of correlations between stocks rises in proportion to the square of the number of stocks (the correlation matrix is a square, symmetric matrix with ones on the diagonal, giving $n(n-1)/2$ unique correlations for n stocks). When Markowtiz (1952) proposed his mean-variance optimization method of portfolio selection this was computationally infeasible for reasonable values of n, but even as memory and CPU power problems were overcome, the problem remained that the vast numbers of correlations were beyond the human ability to comprehend them.

There are two solutions to this problem reported in the literature. The first reduced the number of correlations by estimating the correlations of an individual stock against a relevant index, hence the growth in correlations was reduced to linear. This was the approach of Sharpe (1964) and Lintner (1965).

The second used optimisation software, which is essentially a "black-box" approach. For example, the mean-variance analysis of Markowtiz (1952), which requires forecasts of expected returns, correlations and variances, is commonly solved using optimisation software. The problem with trying to use forecasts is that it is notoriously difficult to generate forecasts which yield a consistent correlation matrix without the use of additional specialized software to ensure the forecasts are consistent. For example, the full correlation matrix for the S&P500 requires 124,750 unique, but not independent, forecasts. However, the method needs to avoid using historical data because this often leads to exceptionally poor diversification for (as Bernstein (2001, p69) warns the small investor who is thinking of using a mean-variance optimizer with historical data for inputs to do asset allocation), "it is overly fond of assets with a recent history of high returns."

The basic problem of vast amounts of numerical data overwhelming human capacity to understand it is well known to many fields. In particular, with the development of supercomputers the need to understand the often enormous output of data from simulations required a new approach. The human ability to comprehend large amounts of data when presented in graphical form has long been known and is encapsulated in the old adage "a picture is worth a thousand words" though that proverb may well underestimate the value of a good picture. There is a burgeoning field of data visualization, see for example Few (2009), Steel and Iliinsky (2010), Lima (2011) and Yau (2011) among many others. The purpose of our paper is to show that, with appropriate visualization methods, in our case using Neighbor-Nets (Bryant and Moulton, 2004), it is possible to use the full correlation matrix for a given set of stocks to gain deeper insight into their behaviour. This should aid the stock market analyst or other investor in the process of selecting a limited number of stocks to form as well diversified a portfolio as possible given the financial limitations they face.

A method of visualising the correlations has implications for creating well diversi-

fied stock portfolios, or alternatively it can act as a check for other methods. The recommended number of stocks required to form a well diversified portfolio has risen dramatically over time. Evans and Archer (1968) concluded that holding a portfolio of eight stocks was sufficient to obtain a well diversified portfolio and doubted that there was economic justification of increasing portfolio sizes beyond 10. In the introductory section of Statman (1987) he provided a range of quotes on the numbers of stocks required to form a diversified portfolio from then standard finance textbooks. The highest recommendation in Statman's survey was to hold 12 to 18 stocks given by Reilly (1985). In his study Statman (1987) went on to conclude that a minimum of 30 stocks for the borrowing investor and 40 stocks for the lending investor was required for good diversification. Recently, Domian et al. (2007) concluded that 100 stocks was not enough. Therefore modern investors are interested in correlations between hundreds of potential investments and in comparing subsets of the stocks. To this end a visualisation method is necessary.

The remainder of the paper is structured as follows. Section (2) provides an overview of the Neighbor-Net clustering algorithm. Section (3) applies Neighbor-Nets to data from the New Zealand Stock Exchange. Section (4) contains the discussion and our conclusions.

2 Neighor-Net

Neighbor-Net was developed to represent the relationships between DNA strands, once each pair to genetic sequences was converted to a distance measure. Neighbor-Net takes a matrix of pairwise distances and produces a network based on "splits". A split is a partition of the set of nodes or objects (in our case companies on the stock exchange) into two disjoint, non-empty groups.

The construction of neighbor-Net networks has four key components: the agglomerative process, selection formulae, distance reduction and estimation of the split weights. The agglomerative process describes how the hierarchy of nodes is determined, selection formulae describe the system used in determining the hierarchy and distance reduction describes how the distances are adjusted as the hierarchy is built. The result of the these three steps is a circular collection of splits. Formally a set of circular splits is one which satisfies that condition that there is an ordering of the nodes $x_1, x_2, \ldots, x_n$ such that every split is of the form $\{x_i, x_{i+1}, \ldots, x_j\}|X - \{x_i, \ldots, x_j\}$ for some i and j satisfying $1 \leq i \leq j < n$. The advantage of this is that the splits can be represented on a plane.

We describe the algorithm following Bryant and Moulton (2004). All the nodes

start out as singletons and the selection formulae finds the two closest nodes. These nodes are not grouped immediately but remain as singletons until a node has two neighbors. At this stage the three nodes, the node and its two neighbors, are merged into two nodes. Here we present the selection formula for grouping nodes. Let neighboring relations group the n nodes into m clusters. Let d_{xy} be the distance between nodes x and y. Let $C_1, C_2, \ldots, C_m, m \leq n$ be the m clusters. The distance $d(C_i, C_j)$ between two clusters is

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d_{xy}, \tag{1}$$

that is, an average of the distances between elements in each cluster.

The closest pair of clusters is given by finding the i and j that minimise

$$Q(C_i, C_j) = (m-2)d(C_i, C_j) - \sum_{k=i, k \neq i} md(C_i, C_k) - \sum_{k=i, k \neq i} md(C_j, C_k), \tag{2}$$

and denote them C_{i*} and C_{j*}

To choose particular nodes within clusters we select the node from each cluster that minimises

$$\hat{Q}(x_i, x_j) = (\hat{m}-2)d(x_i, x_j) - \sum_{k=i, k \neq i} \hat{m}d(x_i, C_k) - \sum_{k=i, k \neq i} \hat{m}d(x_j, C_k). \tag{3}$$

where $x_i \in C_{i*}$ and $x_j \in C_{j*}$ and $\hat{m} = m + |C_{i*}| + |C_{j*}| - 2$.

The distance reduction updates the distance matrix with the distance from the two new clusters to all the other clusters. The distance reduction formulae calculate the distances between the existing nodes and the new combined nodes. If y has two neighbors, x and z, then the three nodes will be combined and replaced by two nodes which we can denote as u and v. The neighbor-Net algorithm uses

$$\begin{aligned} d(u, a) &= \alpha d(x, a) + \beta d(y, a) && (4) \\ d(v, a) &= \beta d(y, a) + \gamma d(z, a) && (5) \\ d(u, v) &= \alpha d(x, y) + \beta d(x, z) + \gamma d(y, z) && (6) \end{aligned}$$

where α, β and γ are non-negative real numbers with $\alpha + \beta + \gamma = 1$.

The process stops when all the nodes are in a single cluster.

The neighbor-Net method of Bryant and Moulton (2004) used non-negative least squares to estimate the split weights given the distance vector and a set splits known as the circular splits.. Suppose that the splits in the network are numbered $1, 2, \ldots, m$ and that the nodes are numbered $1, 2, \ldots, n$. Let $\mathbf{X}$ be the be the *splits*

matrix with the dimensions $n(n-1)/2 \times m$ matrix with rows indexed by pairs of nodes, columns indexed by splits, and entry $\mathbf{X}_{ij,k}$ given by

$$\mathbf{X}_{ij,k} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ are on opposite sides of the split} \\ 0 \text{ if } i \text{ and } j \text{ are on the same side of the split.} \end{cases} \quad (7)$$

Similar nodes will be clustered together in the network. This is a direct result of each pair of neighboring nodes in the ordering being close together in terms of distance, and separated from node where the distance measure reveals dissimilarity.

3 Example

The overall aim of this section is to demonstrate that visualisation of the correlation matrix can be informative when considering subsets of stocks (or other investment opportunities). This example focuses on data from the New Zealand Stock Exchange and comparing and interpreting the correlations observed in several industry categories.

3.1 Data

We downloaded closing prices for companies listed on the New Zealand Stock exchange from Datastream for the 10 year period between 5 October 1999 and 5 October 2009 giving 2610 trading days. A number of stocks were removed from the sample which had either not been listed for the whole period or were so thinly traded that typically there was no price movement over a trading day. This left us with 48 stocks for which there was complete data and sufficient trading activity that correlations could be estimated.

The price time series was used to calculate one, five, and twenty day return series from which the correlations were estimated using the function `cor` in base `R` (R Development Core Team, 2009). The decision to use five trading day returns in our analysis was a subjective one based on several factors. Firstly, the stocks on the New Zealand exchange are generally low in price, often well under $10, and minimum price movements were typically one cent increments. This gives rise to noticeably discrete levels of returns which in turn makes estimating correlations difficult. Secondly, there was a lot of noise in the one day return series which we presume depended on whether the closing price of the day was based on an at market buy or sell. Thirdly, at longer time periods, such as using 20 trading

day returns the number of observations per year were small, at most 13 in the 20 trading day case. For stocks which pay dividends quarterly nearly one in three returns would be affected by the dividend payment. So the selection of five day returns, while not ideal, seemed the best choice.

Because input to the Neighbor-Net software requires distances we simply subtracted the estimated correlation from one to yield an estimated distance between stocks. A zero distance corresponds to a perfect correlation, hence no distance between them. Similarly the maximum distance was two, corresponding to a perfect negative correlation. The full "distance" matrix obtained from the correlations was formatted and augmented with appropriate taxa label data for input into Neighbor-Net and a network generated.

3.2 Neighor-Net Networks for New Zealand Stocks

In Figure (1) we have a plot of the network for the 48 New Zealand stocks using correlations obtained using five trading day return data. The network is unlabelled to give the best view of the network structure. In the top left corner of the diagram is a distance scale of 0.1 correlation units. The labelled network is presented in Figure (2). The labels are the three letter stock market codes assigned to each security by the New Zealand Stock Exchange. A list of stock codes, company name and industry groupings are listed in Table (1) in Appendix (A). Clearly there are several "arms" on the diagram interspersed with gaps. The "arms" represent stocks which are close to each other in terms of the full correlation matrix when converted to "distances".

When reading the network it is important to note that the internal network structure has no significance when used with stock market data. With biological data the places where the network bifurcates represents speciation events and the places where the network rejoins are recombination events where previously isolated populations have exchanged genetic material. In a biological context the internal network represents the best estimate of the evolutionary history given the current genetic distances of the species.

In our case the "species" are individual stocks and the key piece of information is the final location of the stock on the periphery of the evolutionary network.

The "arms" in Figure (1) are groups of stocks which not only have high correlations with, or low "distances" from, each other but which have similar correlation "distances" with the stocks in the remainder of the market. For example, the "arm" between 10 and 11 o'clock is a cluster of five property trusts, (Kiwi Income Property Trust (KIP), AMP NZ Office Trust (APT) , Goodman Property Trust

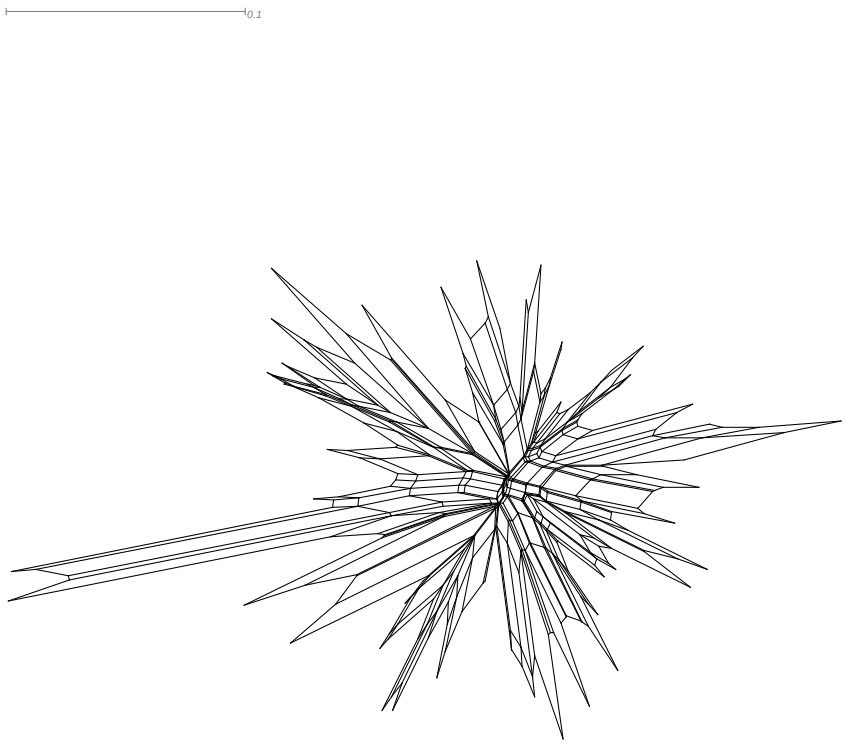

Figure 1: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using weekly returns to estimate correlations and hence distances.

(GMT), Property for Industry (PFI) and ING Medical Properties Trust (IMP)) and Ryman Health Care (RYM), a company whose business involves substantial property holdings. This can be seen from the labelled network in Figure (2). (Note that some of these have undergone name changes after the close of the study period.) It is perhaps unsurprising that these stocks are highly correlated but it also gives us confidence that the groupings reflect useful information.

From this network diagram one can see at a glance which groups of stocks should be considered to obtain maximum diversification for a typical portfolio size held by private investors. The distance scale in the top left corner indicates that the greatest correlation "distance" between any two stocks is not much more than 0.3 units out of a maximum possible distance of 2.0. The reader should note that Figure (2) was generated using the "magnifier" tool within SplitsTree which increases the size of the network as well as slightly altering its shape in order to

better see its structure but that this renders the distance scale meaningless. In the remaining network diagrams in Figures (1) and (6) through (8) the distance scale is meaningful.

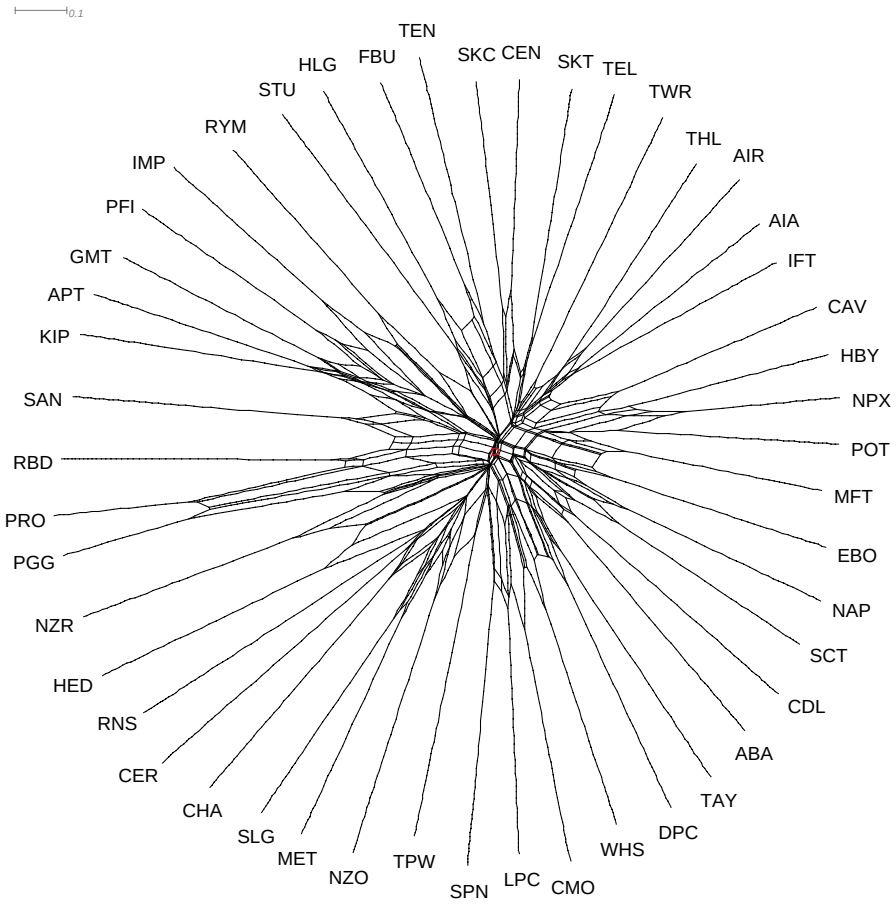

Figure 2: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using weekly returns to estimate correlations and hence distances.

From the network diagram we chose two pairs of stocks to illustrate their similarity or difference depending on their location in the network. From within the above named group of property stocks it is clear that AMP NZ Office Trust (APT) and Kiwi Income Property Trust (KIP) are very close to each other. Their price series are presented in Figure (3). The similarity of these two series are obvious even to the untrained eye.

Secondly, we present in Figure (4) the price series of Fletcher Building (FBU) and the Warehouse Group (WHS). This graph illustrates a second problem namely, that one cannot blindly choose stocks based on their correlations, expected returns are also critically important. Much of the lower correlation between these

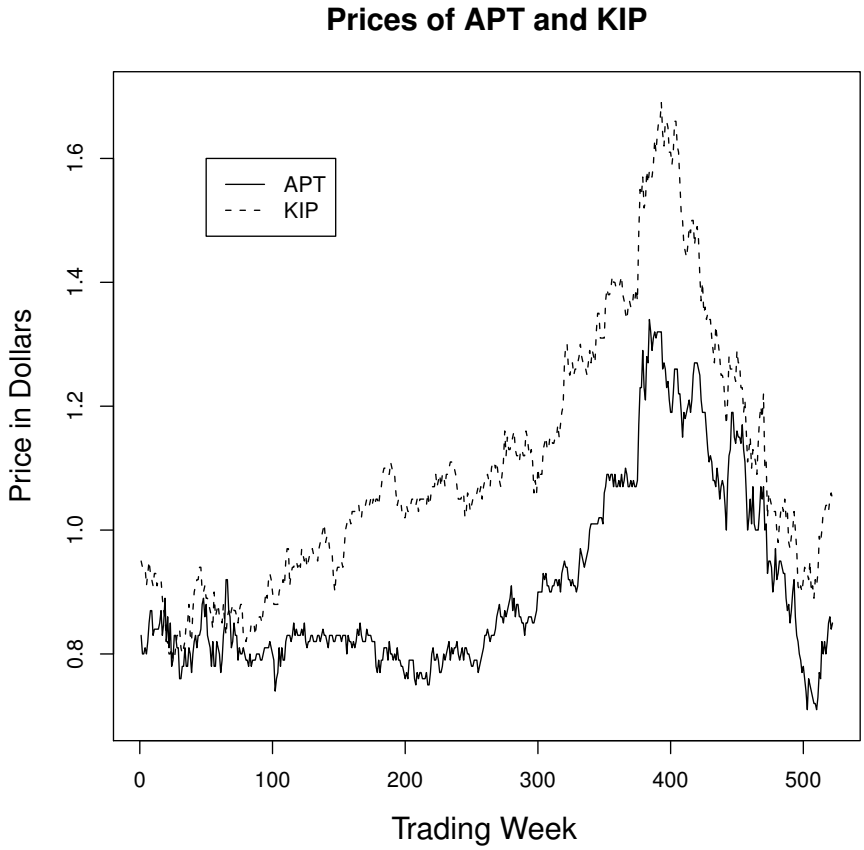

Figure 3: The price series for AMP NZ Office Trust (APT) and Kiwi Income Property Trust (KIP) over the ten year study period.

two stocks is due to the poor performance of WHS relative to FBU. WHS started the study period at a price of $3.75 and finished at $4.25 while FBU started at $2.39 and finished at $8.33.

In previous studies of diversification a common methodology was to compare portfolios of randomly selected stocks with identically sized portfolios diversified by selecting stocks based on their industry groupings, see Domian et al. (2007) for an example. Given that the effectiveness of diversification depends on the specific correlations between stocks, the Neighbor-Net networks allow us to see whether, in fact, stocks in the same industries are "close" to each other. For the stocks in our study we present four examples of industry groupings. (We have made diagrams for all NZX industry groupings, these are available on request from the authors.)

Figure (5) presents locations of the property stocks in the network. As indicated

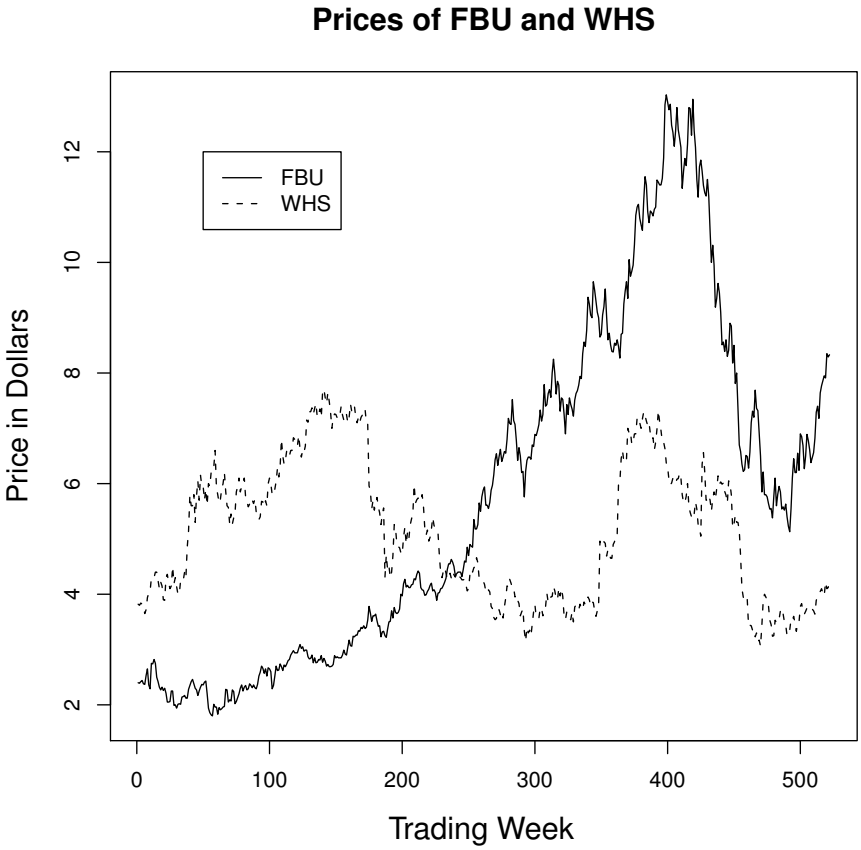

Figure 4: The price series for Fletcher Building (FBU) and the Warehouse Group (WHS) over the ten year study period.

above, five of the stocks are located in one cluster but the two remaining stocks (NAP and CDL) lie on the other side of the network. Although CDL is classified by the exchange as a property stock it is a hotel chain. While CDL has substantial property holdings in order to conduct its business, the rentals are largely on a day-to-day basis rather than the multi-year leases entered into by the other companies and trusts classified as property stocks.

Figure (6) shows the locations of companies in the energy sector, while Figure (7) shows the locations of companies in the consumer services sector. Here the traditional wisdom that companies in the same industry group should behave similarly is not supported.

Figure (8) presents the location of the companies in the media sector. In this case, SKT and TEL lie next to each other confirming traditional wisdom, but this observation must be tempered by the fact that there are only two such companies

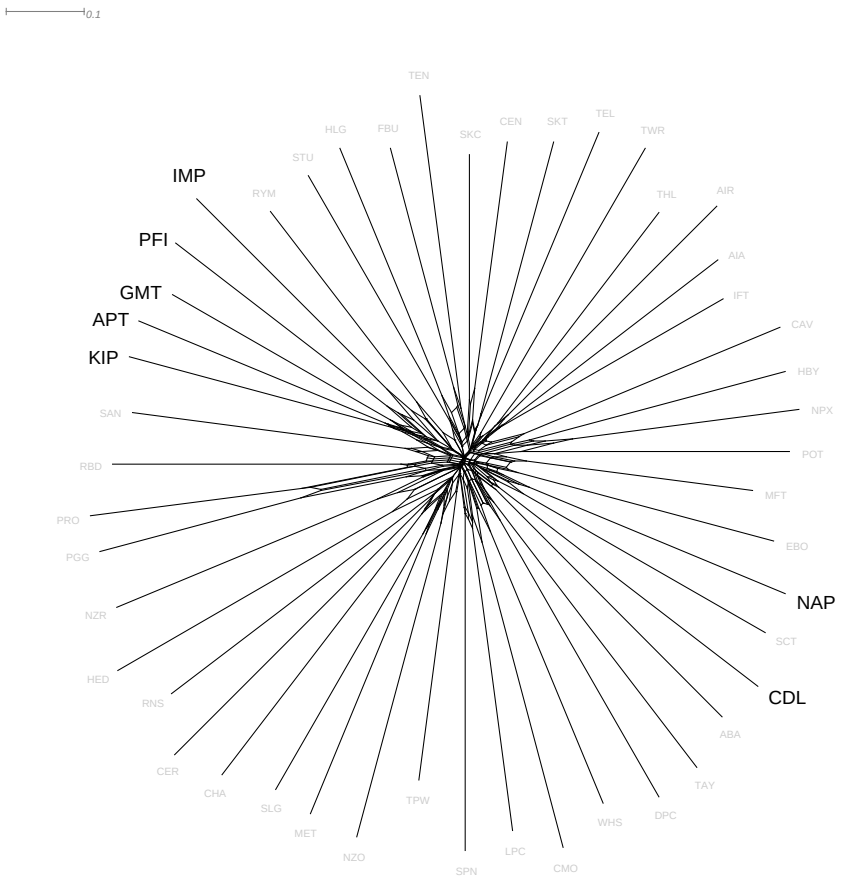

Figure 5: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Property in bold.

in our sample.

4 Discussion and Conclusions

The problem of understanding the correlation matrix for a set of investment opportunites (not limited to stocks) has posed a significant barrier to applications.

The method and results presented here shows significant promise in being able to add value to an analysis of the correlation matrix. We used historical data primarily because by its very nature it yields a consistent matrix. While it is well known that correlation forecasts rather than historical correlations should be

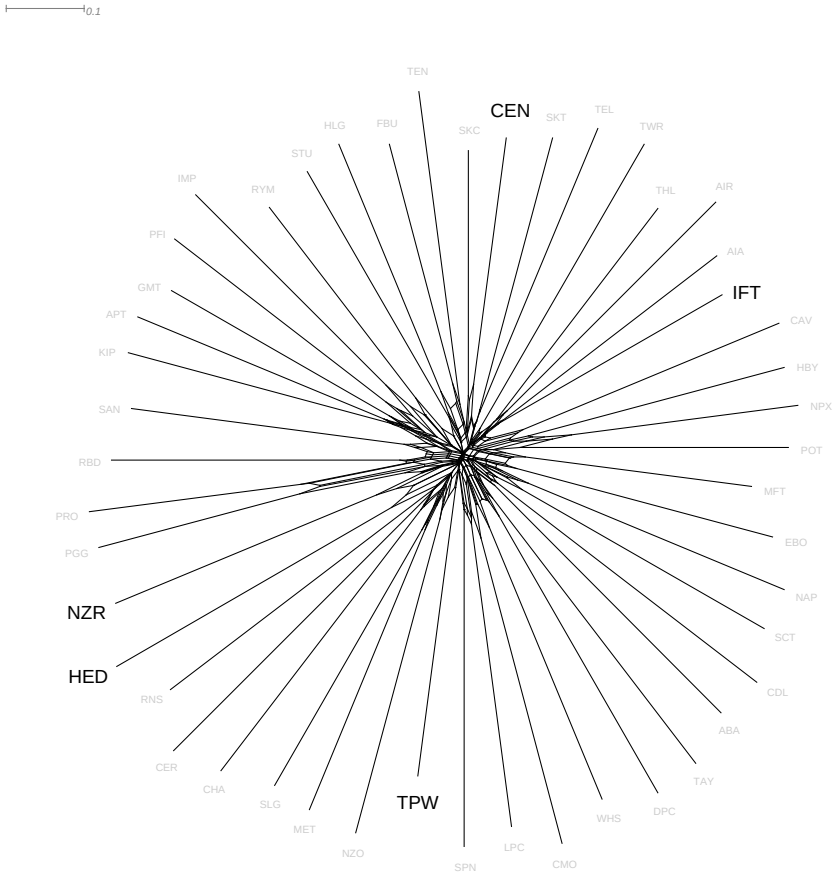

Figure 6: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Energy Processing companies in bold.

used, should a set of consistent forecasts be available, no change in the methods presented here would be required. Further, any other security analysis which yeilds a matrix of "distances" between stocks (or other investment opportunities) can be visualized using Neighbor-Net networks.

Much more needs to be done to prove the worth of this particular visualization method and further visualization and clustering methods need to be investigated. The New Zealand stock market is small so a larger stock market, particulary one with higher priced stocks to aid in the estimation of the correlations, needs to be studied using these methods. While past studies of diversification have typically focussed on random selection from the full stock universe versus selection from within industry groupings, the above results strongly suggest that these studies need to be repeated and add a third method of diversification by selecting from

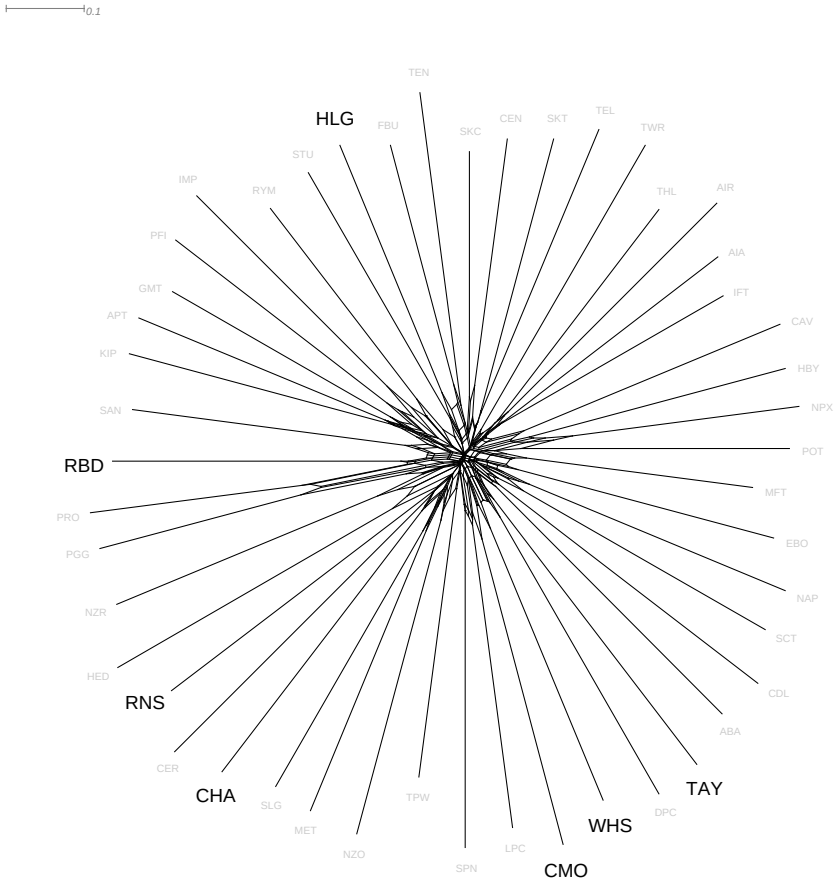

Figure 7: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Consumer Services in bold.

groups of stocks which lie close to each other on the network.

References

Bernstein, W. (2001). *The Intelligent Asset Allocator*. McGraw-Hill.

Bryant, D. and V. Moulton (2004). Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol. 21*(2), 255–265.

Domian, D. L., D. A. Louton, and M. D. Racine (2007). Diversification in Portfolios of Individual Stocks: 100 Stocks Are Not Enough. *The Financial Review 42*, 557–570.

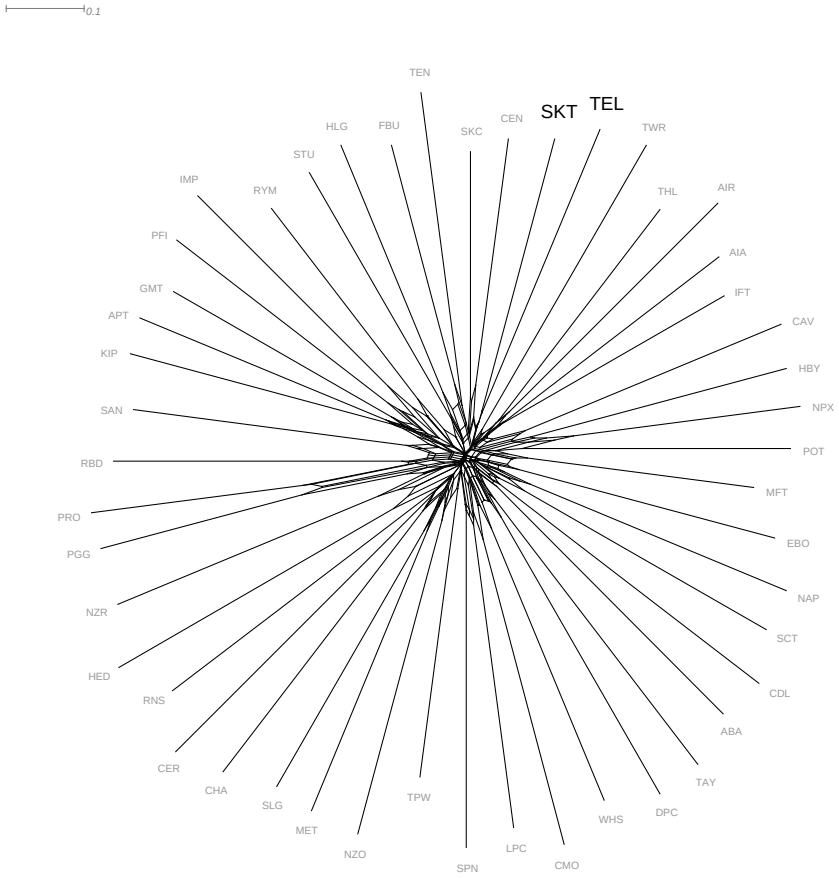

Figure 8: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Media companies in bold.

Epstein, L. G. and S. M. Tanny (1980). Increasing Generalized Correlation: A Definition and Some Economic Consequences. *The Canadian Journal of Economics 13*(1), 16–34.

Evans, J. L. and S. H. Archer (1968). Diversification and the Reduction of Dispersion: An Empirical Analysis. *The Journal of Finance 23*(5), 761–767.

Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press.

Lima, M. (2011). *Visual Complexity: Mapping Patterns of Information*. Princeton Architectural Press.

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky

Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics 47*(1), 13–37.

Markowtiz, H. (1952). Portfolio Selection. *The Journal of Finance 7*(1), 77–91.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reilly, F. K. (1985). *Investment Analysis and Portfolio Management 2nd Edition*. Dryden Press.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *The Journal of Finance 19*(3), 13–37.

Statman, M. (1987). How Many Stocks Make a Diversified Portfolio. *The Journal of Financial and Quantitative Analysis 22*(3), 353–363.

Steel, J. and N. Iliinsky (2010). *Beautiful Visualization: Looking at Data through the Eyes of Experts (Theory in Practice)*. O'Reilly Media.

Yau, N. (2011). *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*. Wiley.

A Stock Codes and Industry Segments

Table 1: Stock market codes, company names and industrial sector of stocks in the study.

Code	Name	Industry Segment
ABA	Abano Healthcase	Services/Finance and Other Services
AIA	Auckland International Airport	Services/Ports
AIR	Air New Zealand	Services/Transport
APT	AMP NZ Office Trust	Property
CAV	Cavalier	Goods/Textiles and Apparel
CDL	CDL Investments NZ.	Property
CEN	Contact Energy	Energy/Energy Processing
CER	CER Group	Investments
CHA	Charlie's Group	Services/Consumer
CMO	Colonial Motors	Services/Consumer
DPC	Dorchester Pacific	Services/Finance and Other Services
EBO	EBOS Group	Goods/Intermediate and Durables

Table 1: Stock market codes, company names and industrial sector of stocks in the study.

Code	Name	Industry Segment
FBU	Fletcher Building	Primary/Building
GMT	Goodman Property Trust	Property
HBY	Hellaby Holdings	Investment
HED	Horizon Energy Distributors	Energy/Energy Processing
HLG	Hallenstein Glasson Holdings	Services/Consumer
IFT	Infratil	Energy/Energy Processing
IMP	ING Medical Properties Trust	Property
KIP	Kiwi Income Property Trust	Property
LPC	Lyttleton Port	Services/Ports
MET	Metlifecare	Services/Finance and Other Services
MFT	Mainfreight	Services/Transport
NAP	National Property Trust	Property
NPX	Nuplex Industries	Primary/Building
NZO	New Zealand Oil and Gas	Primary/Mining
NZR	New Zealand Refining	Energy/Energy Processing
PFI	Property for Industry	Property
PGG	PGG Wrightson	Primary/Agriculture and Fishing
POT	Port of Tauranga	Services/Ports
PRO	Provenco Cadmus	Services/Finance and Other Services
RBD	Restaurant Brands NZ	Services/Consumer
RNS	Renaissance	Services/Consumer
RYM	Ryman Healthcare	Services/Finance and Other Services
SAN	Sanford	Primary/Agriculture and Fishing
SCT	Scott Technology	Goods/Intermediate and Durables
SKC	Sky City Entertainment Group	Services/Leisure and Tourism
SKT	Sky Network Television	Services/Media and Communications
SLG	Sealegs	Investment
SPN	South Port New Zealand	Services/Ports
STU	Steel and Tube Holdings	Primary/Building
TAY	Taylors Group	Services/Consumer
TEL	Telecom Corporation of NZ	Services/Media and Communications
TEN	Tenon	Primary/Forestry
THL	Tourism Holdings	Services/Leisure and Tourism
TPW	Trustpower	Energy/Energy Processing
TWR	Tower	Services/Finance and Other Services
WHS	Warehouse Group	Services/Consumer

B Other Networks

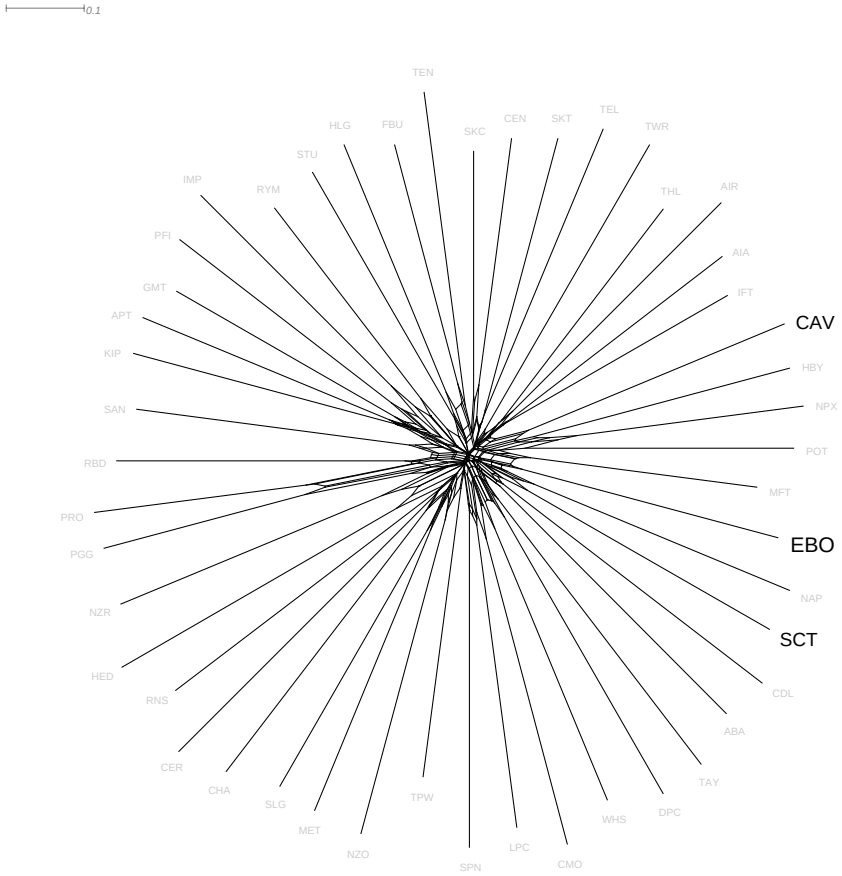

Figure 9: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Goods companies in bold. Scott Technology (SCT) and EBOS Group (EBO) are intermediate and durable goods companies while Cavalier (CAV) is a textiles company.

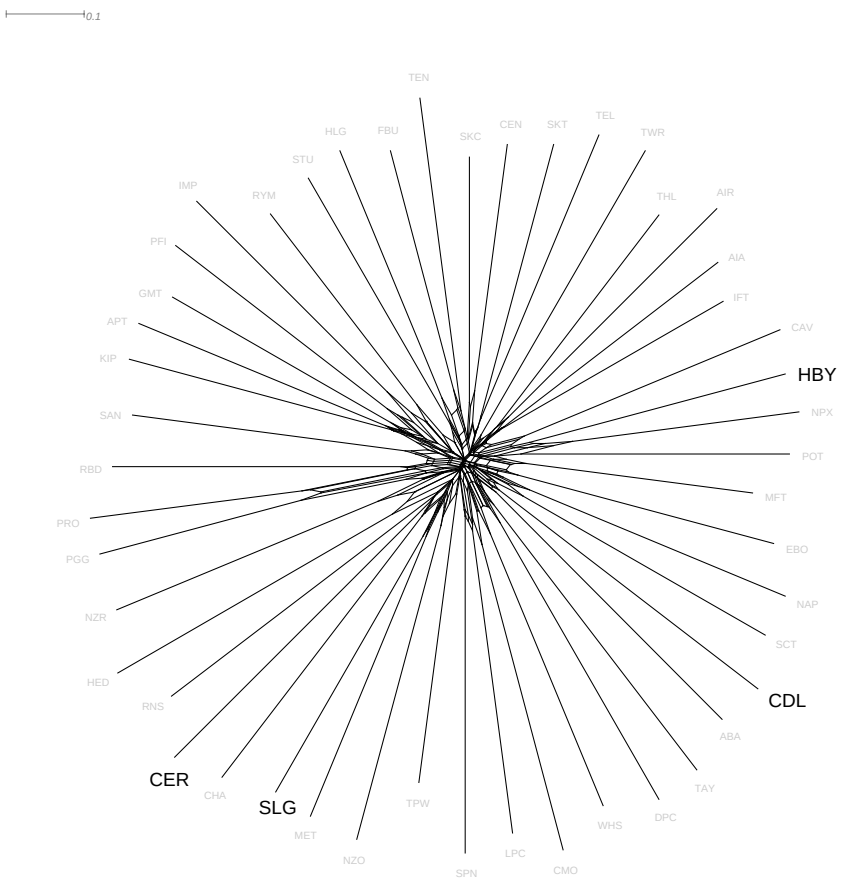

Figure 10: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Investments in bold.

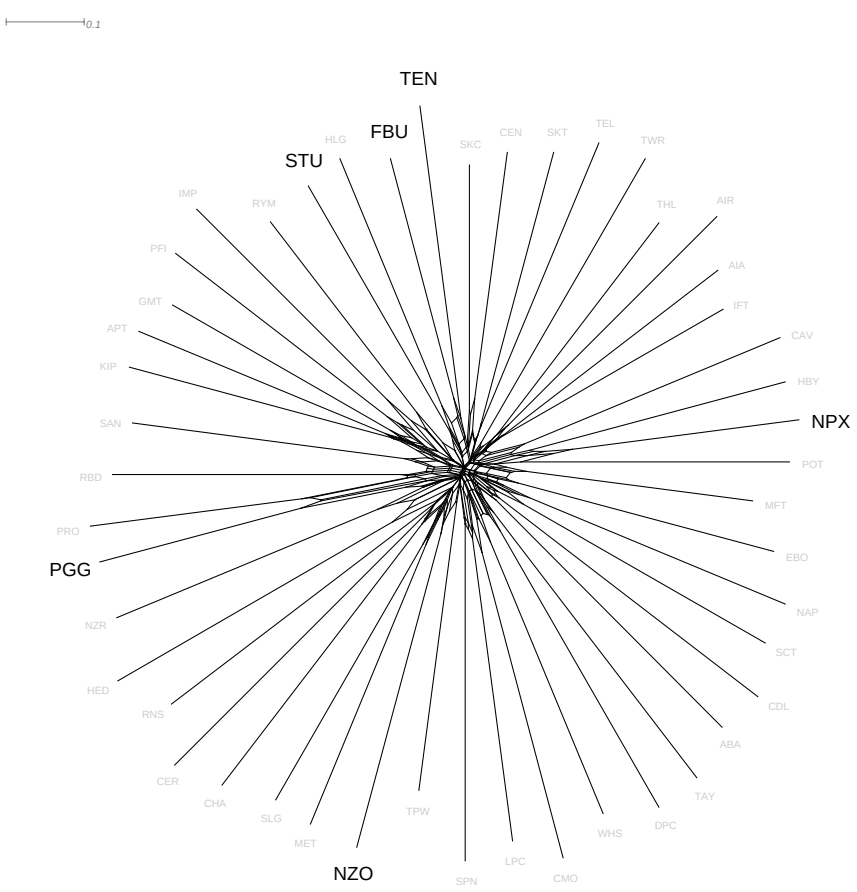

Figure 11: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing all companies from the Primary Industries in bold.

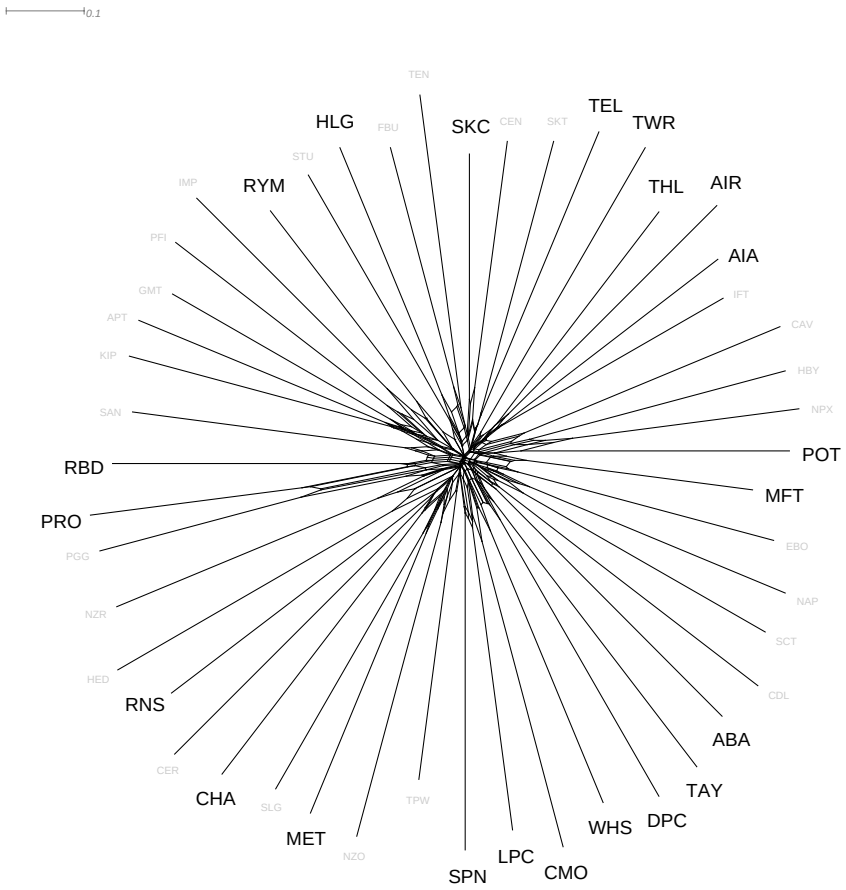

Figure 12: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing all Service companies in bold.

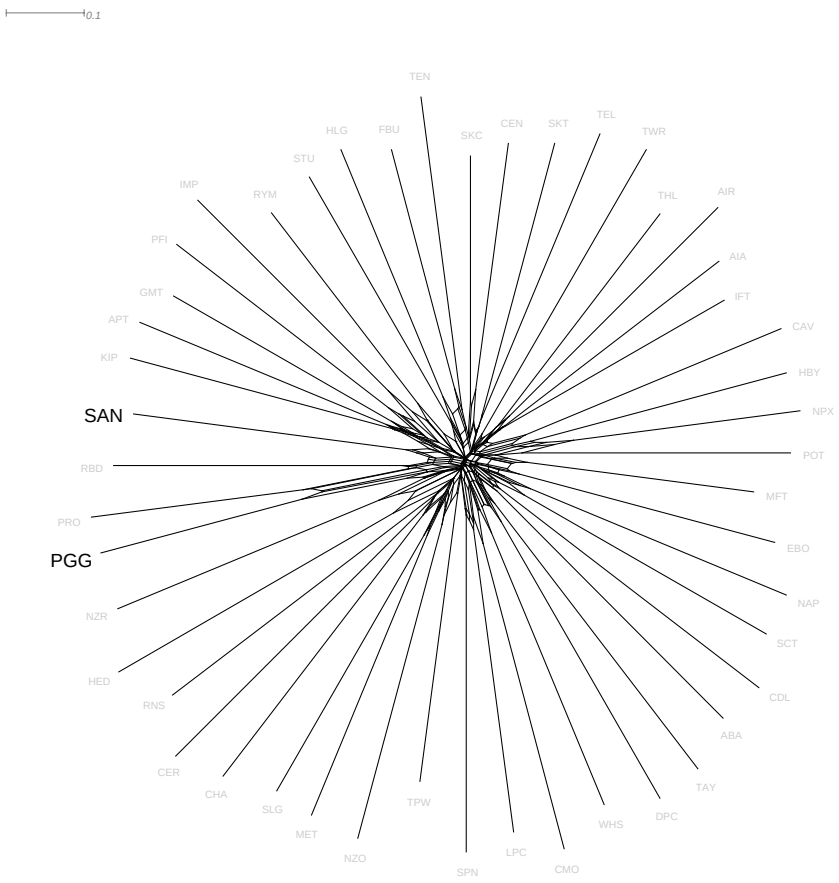

Figure 13: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Agriculture and Fishing companies in bold.

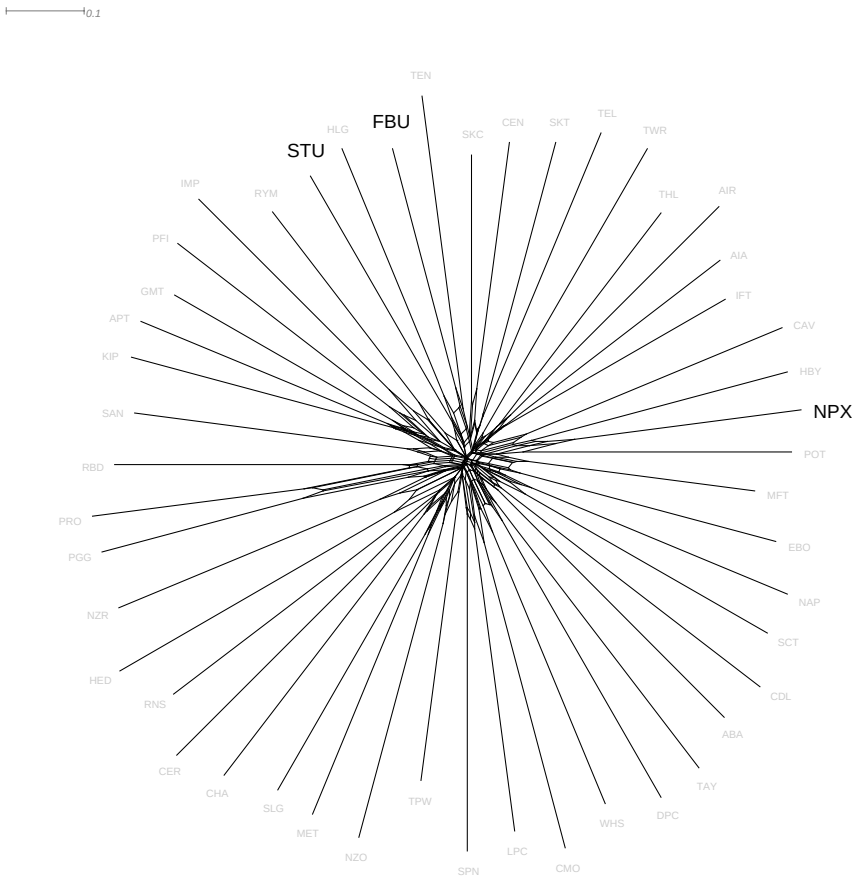

Figure 14: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Building companies in bold.

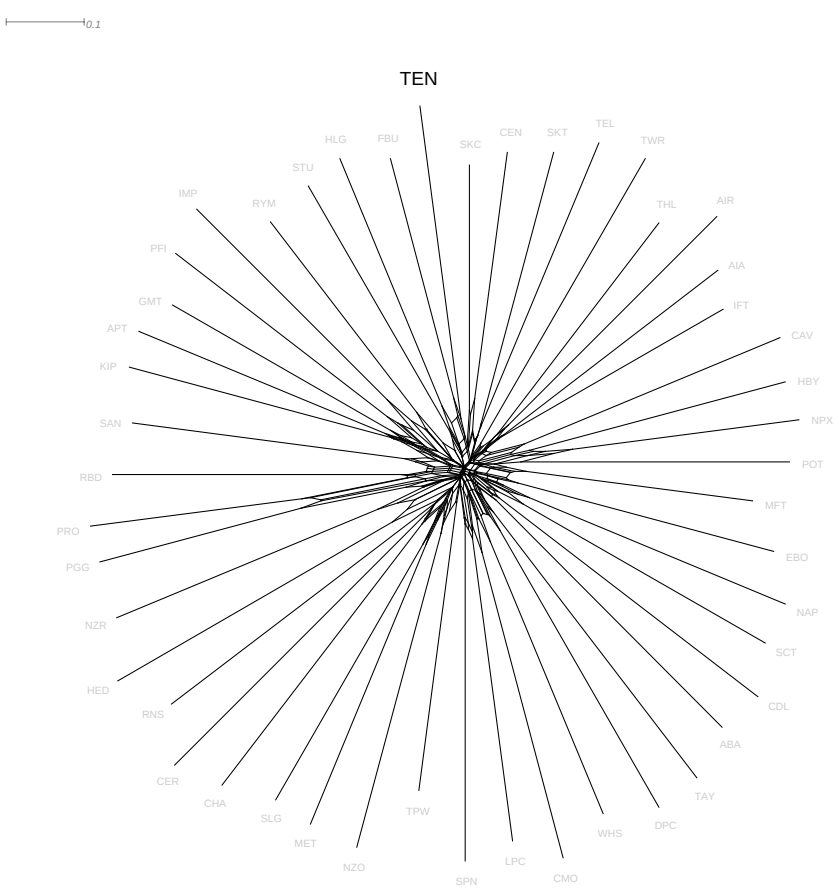

Figure 15: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Forestry companies in bold.

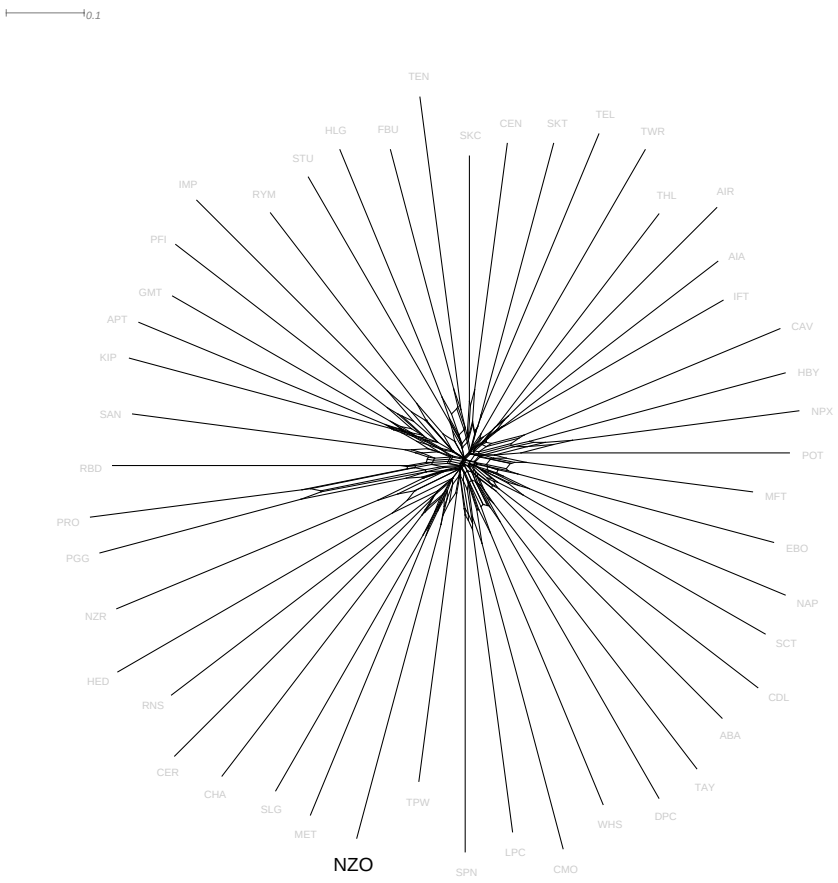

Figure 16: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Mining companies in bold.

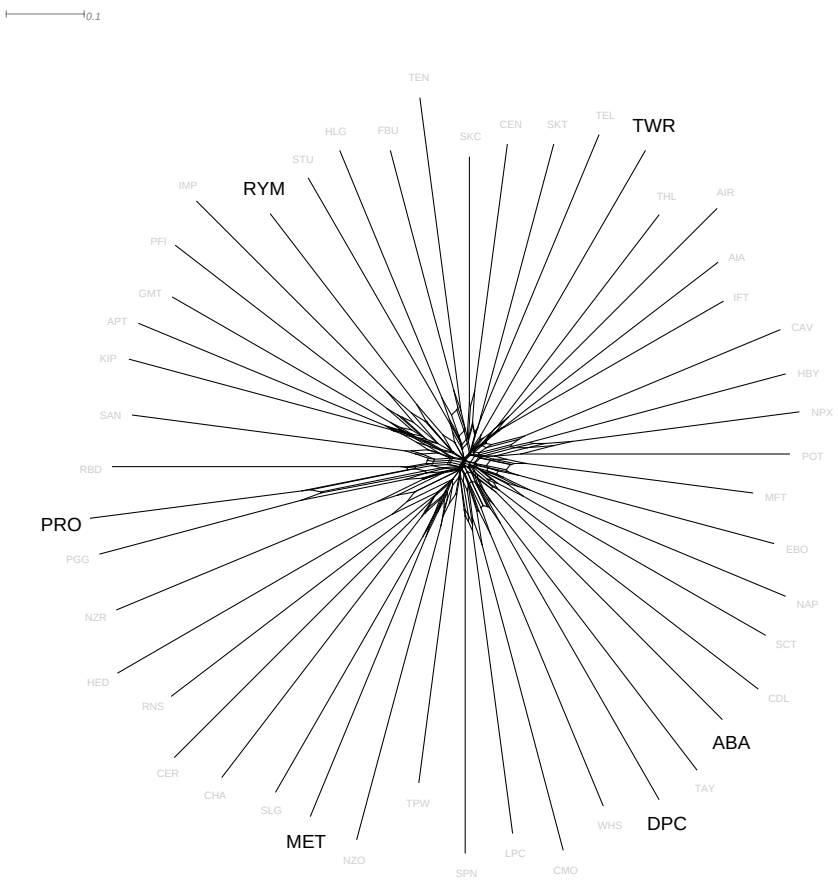

Figure 17: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Finance and Other Services in bold.

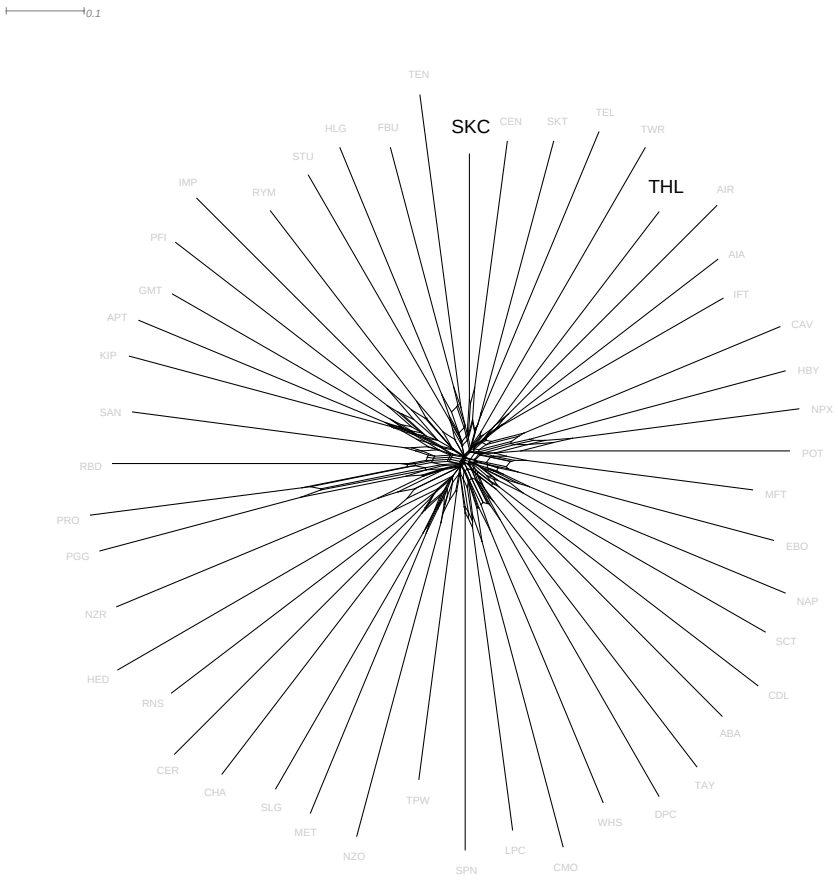

Figure 18: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Leisure companies in bold.

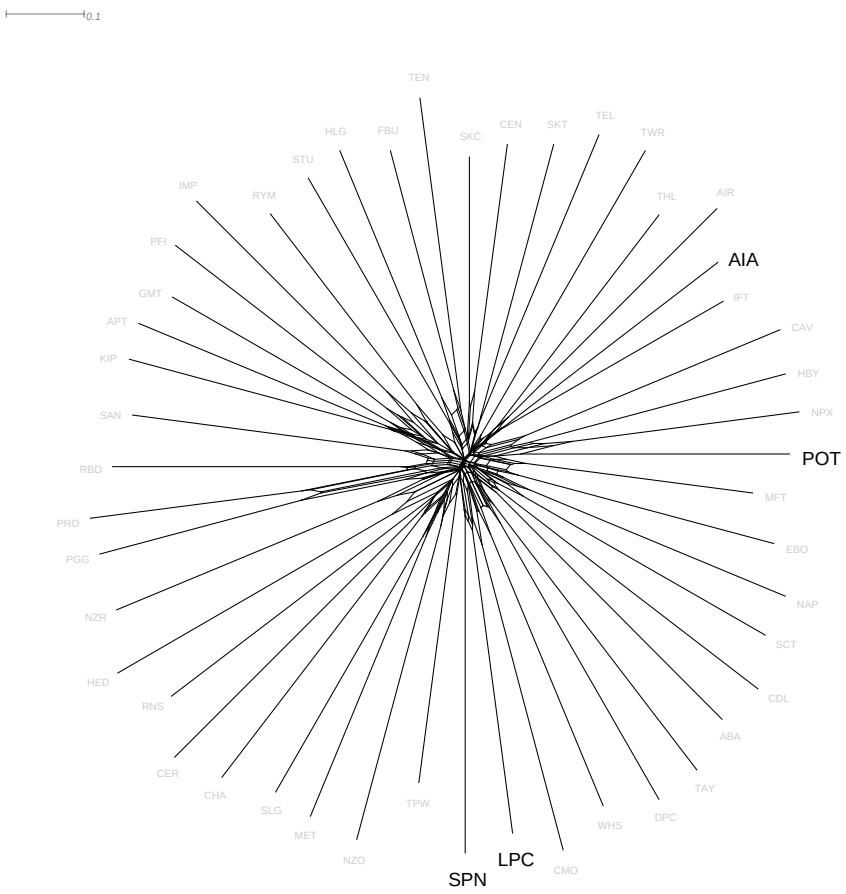

Figure 19: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Ports in bold.

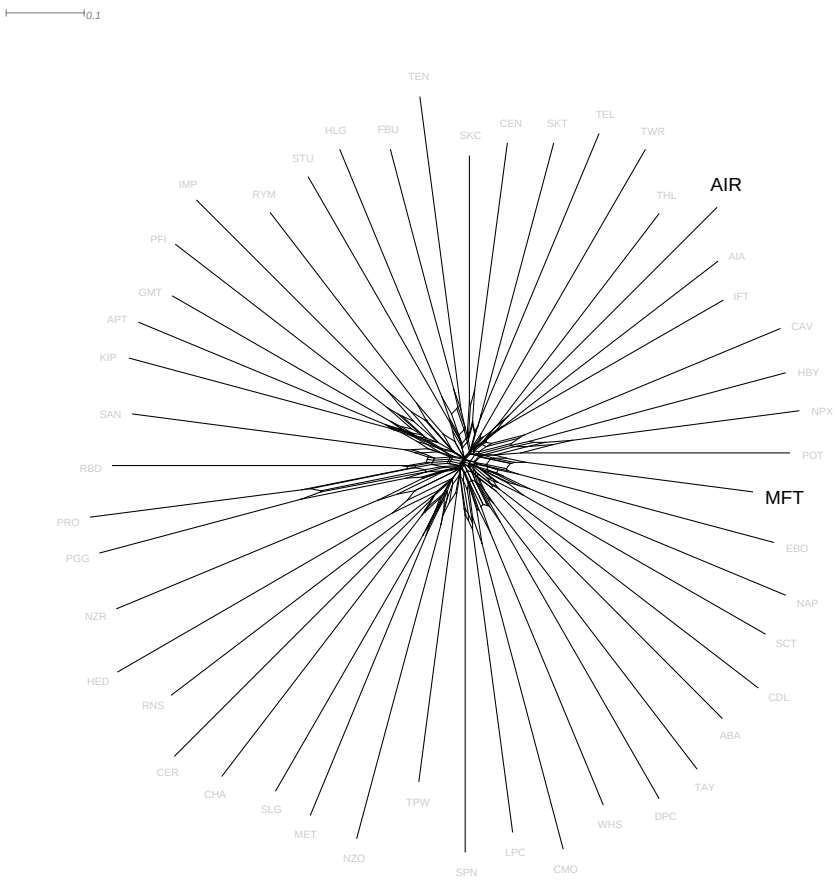

Figure 20: SplitsTree network for 48 stocks from the New Zealand Stock Exchange using daily returns to estimate correlations and hence distances showing Transport Services in bold.