Zhan, Cheng Juan; Rea, William; Rea, Alethea

Article

Stock selection as a problem in phylogenetics: Evidence from the ASX

International Journal of Financial Studies

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Zhan, Cheng Juan; Rea, William; Rea, Alethea (2016) : Stock selection as a problem in phylogenetics: Evidence from the ASX, International Journal of Financial Studies, ISSN 2227-7072, MDPI, Basel, Vol. 4, Iss. 4, pp. 1-19, https://doi.org/10.3390/ijfs4040018

This Version is available at:
https://hdl.handle.net/10419/167819

International Journal of
Financial Studies

MDPI

Article

Stock Selection as a Problem in Phylogenetics—Evidence from the ASX

Cheng Juan Zhan [1], William Rea [1,*] and Alethea Rea [2]

[1] Department of Economics and Finance, University of Canterbury, Christchurch 8140, New Zealand; cheng.zhan@pg.canterbury.ac.nz
[2] Data Analysis Australia, Perth 6009, Australia; alethea.rea@gmail.com
* Correspondence: bill.rea@canterbury.ac.nz; Tel.: +64-3-364-3474

Academic Editor: Nicholas Apergis
Received: 17 March 2016; Accepted: 20 September 2016; Published: 29 September 2016

Abstract: We report the results of fifteen sets of portfolio selection simulations using stocks in the ASX200 index for the period May 2000 to December 2013. We investigated five portfolio selection methods, random selection, selection within industrial groups, and three based on neighbor-Net phylogenetic networks. We report that using random, industrial groups, or neighbor-Net phylogenetic networks alone rarely produced statistically significant reduction in risk, though in four out of the five cases in which it did so, the portfolios selected using the phylogenetic networks had the lowest risk. However, we report that when using the neighbor-Net phylogenetic networks in combination with industry group selection that substantial reductions in portfolio return spread were achieved.

Keywords: stock selection; ASX200; neighbor-Net networks; portfolio risk

JEL Classification: G11

1. Introduction

Portfolio diversification is critical for risk management because it aims to reduce the variance of returns compared with a portfolio of a single stock or similarly undiversified portfolio. The academic literature on diversification is vast, stretching back at least as far as [1]. The modern science of diversification is usually traced to [2] which was expanded upon in great detail in [3].

In one sense, the approach of [2] is optimal and cannot be improved if either the correlations and expected returns of the assets are not time-varying (thus can be accurately estimated from historical data) or they can be forecasted accurately. Unfortunately, neither of these conditions hold in real markets leaving the door open to other approaches.

The literature covers a wide range of approaches to portfolio diversification, such as; the number of stocks required to form a well diversified portfolio, which has increased from eight stocks in the late 1960's [4] to over 100 stocks in the late 2000's [5], what types of risks should be considered, [6–8], factors intrinsic to each stock [9,10], the age of the investor, [11], and whether international diversification is beneficial, [12,13], among others.

In recent years a significant number of papers have appeared in the econophysics literature which apply graph theoretical methods to the study of a stock or other financial markets, see for example, [14–22] among others. Much of this literature uses correlations, partial correlations, or both in their analysis. While these papers are insightful and a valuable addition to the literature, they do not address the question of how to apply their insights to portfolio formation and management.

The mean returns and variances of the individual contributing stocks are insufficient for making an informed decision on selecting a suite of stocks because selecting a portfolio requires an understanding

of the correlations between each of the stocks available for consideration for inclusion in the portfolio. The number of correlations between stocks rises in proportion to the square of the number of stocks, meaning that for all but the smallest of stock markets the very large number of correlations is beyond the human ability to comprehend them. Rea et al. [22] presented a method to visualise the correlation matrix using neighbor-Net networks [23], yielding insights into the relationships between the stocks.

Another key aspect of stock correlations is the potential change in the correlations with a significant change in market conditions (say comparing times of general market increase with recession and post-recession periods).

In order to compare some of the approaches to portfolio diversification [24] lists 15 different methods for forming portfolios and reported results from their study which evaluated 14 portfolio selection methods against a naive $1/N$ strategy. They reported that no single portfolio formation method consistently delivered a higher Sharpe ratio than that of the $1/N$ portfolio, which also had the added benefit of a lower turnover, hence lower transactions costs. Absent from these 15 methods were any which utilized the above-mentioned graph theory approaches. This leaves as an open question whether these graph theory approaches can usefully be applied to the problem of portfolio selection.

The goal of this paper is to compare three network methods with two simple portfolio selection methods for small private-investor sized portfolios.

There are two motivations for looking at very small portfolios sizes. The first is that, despite the recommendation of authorities like [5], [25] reported that in a large sample of American private investors the average portfolio size of individual stocks was only 4.3 individual stocks though, obviously, each investor must hold an integer number of stocks. While comparable data does not appear to be available for private Australian investors, it seems unlikely that they hold substantially larger portfolios. Thus there is a practical need to find a way of maximising the diversification benefits for these investors. The second is that testing the methods on small portfolios gives us a chance to evaluate the potential benefits of the network methods because the larger the portfolio size, the more closely the portfolio resembles the whole market and the less likely any potential benefit is to be discernible.

In this paper we explore investment opportunities on the Australian Stock Exchange using data from the stocks in the ASX200 index.

This paper contributes to the literature by evaluating three portfolio selection methods based on a neighbor-Net network analysis of the correlation matrix and comparing those with two simple, but widely used methods. The five methods are to pick stocks for a portfolio

1. at random;
2. from different industry groups;
3. from different, maximally distant, correlation clusters identified using neighbor-Nets networks;
4. from the dominant industry group within the identified correlation clusters;
5. from non-dominant industry groups within the identified correlation clusters.

These selection methods are described in detail in Section 2.2 together with the motivation for using maximally distant correlation clusters. The identification of maximally distant correlation clusters is described in Section 3.

Our results show that knowledge of correlation clusters, as identified using the neighbor-Net networks, together with the industry groups within these clusters can reduce the portfolio risk.

The outline of this paper is as follows: Section 2 discusses the data and methods used in this paper, Section 3 discusses identifying the correlation clusters, Section 4 presents the results of the simulations of the portfolio selection methods and Section 5 contains the discussion and our conclusions.

2. Data and Methods

We used the weekly price data for the stocks in the ASX200 as our data set. Weekly prices along with the dividend rate and payment date for the period 3 May 2000 to 4 December 2013 were obtained

from DataStream. We appended one or two letters to each ticker symbol in order to identify the industry group for each stock.

The weekly returns were calculated from the price and dividend data and used for both the portfolio formation simulations and for estimating the correlations. The correlations were estimated using the function `cor` in base R [26]. We also calculated period returns for each stock in each of periods two to six for use in the simulations.

We divided the whole period into six shorter periods shown in Figure 1 and used out-of-sample testing to test the effectiveness of each the five methods of diversifying portfolios on reducing risk.

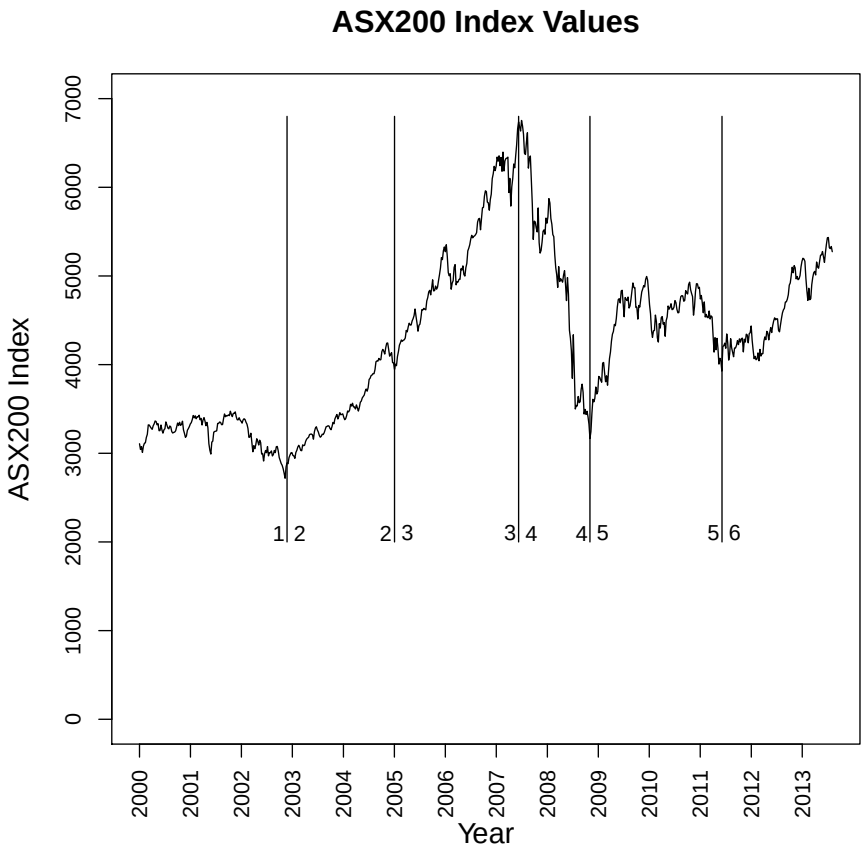

Figure 1. A plot of the ASX200 Index with the boundaries of the study periods marked.

2.1. Neighbor-Net Splits Graphs

A neighbor-Net network is a distance-based clustering method which visualises the relationships between a set of objects based on a set of pairwise distances. The objects are presented around a network according to the (circular) ordering determined by the neighbor-Net algorithm.

The neighbor-Net network algorithm was developed to display the relationships between DNA strands, once each pair to genetic sequences had been 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. Despite its origins within the field of phylogenetics, the neighbor-Net algorithm is, in fact, quite general because it works with distances rather than DNA data. This means neighbor-Net can be applied to any data set which can be expressed as a set of distances between entities (here stocks). Appendix A gives greater detail on the neighbor-Net algorithm.

To be able to use neighbor-Net or any of the other clustering algorithms we need to create from the data a numerical estimate of the "distance" between all pairs of stocks in the sample. In the

literature a distance measure is created by converting the numerical values in the correlation matrix, which range from -1 to 1, to the range 0 to 2 so that a 0 "distance" represents two stocks which are perfectly correlated. The most common way to do the conversion is by using the so-called ultra-metric given by,

$$d_{ij} = \sqrt{2(1-\rho_{ij})} \tag{1}$$

where d_{ij} is the distance corresponding to the the estimated correlation, ρ_{ij}, between stocks i and j, see [14] or [21] for details.

A typical stock market correlation matrix for n stocks is of full rank which means that after converting to a distance matrix according to Equation (1), the location of the points, here stocks, can only be fully represented in $(n-1)$-dimensional space. In visualization, the high dimensional data space is collapsed to a sufficiently low dimensional space that the data can be represented on 2-dimensional surface, such as a page or computer screen, for viewing. Information loss is often unavoidable in the reduction of the dimension of the data space. One of the goals of visualization is to minimize the information loss while making the structures within the data visible to the human eye.

Using the conversion in Equation (1) we formatted the converted correlation matrix and augmented it with the appropriate stock codes for reading into the neighbor-Net software, SplitsTree, available from http://www.splitstree.org [27]. Using the SplitsTree software we generated the neighbor-Nets splits graphs. Because the splits graphs are intended to be used for visualization we defer the discussion of the identification of correlation clusters and their uses to Section 3 below.

2.2. Simulated Portfolios

Recently [28] discussed so-called risk-based asset allocation (sometimes called risk budgeting). In contrast to strategies which require both expected risk and expected returns for each investment opportunity as inputs to the portfolio selection process, risk-based allocation considers only expected risk. The five methods of portfolio selection we present below can be considered to be risk-based allocation methods. This probably reflects private investor behaviour in that often they have nothing more than broker buy, hold, or sell recommendations to assess likely returns.

The portfolio formation methods were compared using simulations of 1000 iterations. There were two sets of simulations. For the first set of simulations a portfolio was sampled based on the rules governing the portfolio type using the period return data. We recorded the mean and standard deviation of the returns for the 1000 portfolios. The second set of simulations was carried out in exactly the same manner except weekly return data was used in order to obtain an estimate of the weekly volatility of the portfolios. Each set covered five portfolio formation strategies.

The five portfolio formation strategies are each described in turn while merging the description of methods four and five listed in the introduction.

Random Selection: The stocks were selected at random using a uniform distribution without replacement. In other words each stock was given equal chance of being selected according but with no stock being selected twice within a single portfolio.

By Industry Groups: There were 11 industry groups represented among the stocks. Some of the groups were very small. For example, the telecommunications group only had two representatives in the early periods but this increased over time as additional stocks classified as being in the telecommunications industry were either listed or grew to sufficient size that they were included in the index. Thus, when the groups were small, it was necessary to merge some of them into larger groups for the purposes of the simulations. This need lessened as the number of stocks grew. We had eight such groups in periods two, three and four and nine groups in periods five and six.

Because the maximum portfolio size was eight stocks the industries were chosen at random using a uniform distribution without replacement. Within each industry group, stocks were selected using a uniform distribution.

By Correlation Clusters: A detailed description, with examples, of identifying the correlation clusters is given in Section 3 below. Briefly, the correlation clusters were determined by examining the neighbor-Net network for the relevant periods (periods one through five). Each stock was assigned to exactly one cluster and each cluster can be defined by a single split (or bipartition) of the circular ordering of the neighbor-Net for the relevant period. The clusters determined in periods one through five were used to generate the stock groups for out-of-sample testing in periods two through six respectively. Because the goal of portfolio building is to reduce risk, each cluster was paired with another cluster which was considered most distant from it. If the correlation clusters represent useful financial groupings of stocks we would expect that choosing a pair of stocks from distant correlation clusters would be likely to give a greater reduction in risk than two stocks selected randomly. This method is discussed in detail below.

If there were fewer clusters than the desired portfolio size, cluster pairs were selected at random and a stock selected from within each correlation cluster pair. The simulation code was written so that if the desired portfolio size was larger than the number of correlation clusters then each cluster group pair had at least s stocks selected, where s is the quotient of the portfolio size divided by the number of clusters. Some (the remainder of the portfolio size divided by number of clusters) correlation groups will have $s + 1$ stocks selected and the cluster pairs this applied to were chosen using a uniform distribution without replacement. However, in all cases reported in this paper, the number of clusters equalled or exceeded the number of stocks in the simulated portfolios.

By Dominant or Non-Dominant Industry Group within Clusters: The final two methods relate to selecting stocks from industry groups within correlation clusters. Each stock within each cluster has an associated industry group. Therefore each correlation cluster could be subdivided into up to eleven sub-clusters based on industry. However, it was clear that in a number of clusters one industry was dominant, sometimes containing more than half of the stocks within that cluster. This lead us to assign each stock to either the dominant industry group or the non-dominant industry grouping creating two groups of stocks within each cluster. This created two disjoint sets of clusters with each stock in exactly one group.

From these two distinct sets of stocks, simulations were run in the same manner as that described in "By Correlation Clusters" above. These simulations were not comparable with the three above because the sample sizes were different and, obviously, each deals with a subset of the data. However, care was taken to ensure that the sample sizes of both subsets was as close to equal in size as was practical.

A problem arose, particularly with the non-dominant industry group stocks when the number of stocks in the cluster was considered too small. In these cases we took advantage of the circular ordering produced by neighbor-Nets and combined the small cluster with a neighbouring cluster.

Unchanged from above, each cluster was paired with the one most distant from it. Once a cluster was selected for inclusion, so was the paired cluster, and a selection was made using a uniform distribution and, if necessary, without replacement.

All simulations were coded and run in R [26].

We used the stocks' weekly return data in period one to determine the clusters, then observed period two's return distributions of the simulated portfolios (1000 replications) which are picked from the different correlation clusters. Because out-of-sample testing was used in our analysis, the simulation then was continued for period three, four, five and six based on the splits graphs produced from the weekly returns in period two, three, four and five respectively.

3. Neighbor-Net Splits Graphs

For three of the five types of portfolio simulation methods discussed above we need to identify correlation clusters and for this paper we define the correlation clusters using the neighbor-Net

splits graphs. In this section we explain how to identify the clusters, then proceed to present the neighbor-Net splits graphs and the clusters identified for each of the first five periods.

At its simplest a neighbor-Net splits graph is a type of map. The ability to identify correlation clusters depends on the user's skill in reading it. As an analogy, all readers of a topographic map read the map in the same way. The information the reader extracts depends on their needs. One person may read a map to extract information about mountain ranges, another for information on river catchments, and still another on the distribution of human settlements. But in all cases the map readers agree which features are mountains, which are rivers and which are towns and cities, no confusion arises because the map is read visually. In the same way all readers of neighbor-Net splits graphs agree on which features are splits, which are recombinations and which are the terminal nodes.

Because this is a visual approach, the information extracted from reading a neighbor-Net splits graph depends on the researcher or financial analyst balancing whatever competing requirements they may have. In this application we know that the sizes of the portfolios we will generate will be two, four, or eight stocks. Consequently, we do not need large numbers of clusters and we would like each cluster to have a sufficient number of stocks such that when selecting stocks at random from within the cluster there are a large number of combinations available to make the simulations meaningful. These requirements guide us when identifying clusters in the neighbor-Net splits graphs. The numbers of clusters and cluster membership is determined visually and it is important not to confuse visual with subjective.

Figures 2–6 present the neighbor-Nets splits graphs for periods one through five. These were used to determine the stock groups for out-of-sample testing using data from periods two through six respectively. Each of the splits graphs has the stocks laid out in a circle (according to the circular ordering) and presented with the split network. The clusters generally were chosen based on the longer distances between two neighbors as seen by the fact that the splits between them are longer and/or deeper.

For each graph we identify the clusters using colour, note that because the ordering of the stocks around the outside represents the circular ordering, each group is a contiguous block of stocks. We then consider each cluster noting which industry is the dominant industry based on the industry with the largest number of stocks. These two features are then used in the portfolio selection methods.

In Figure 2 eight clusters were identified and are colour coded in order to distinguish them. The `SplitsTree` software generates a large amount of statistical information about the network. Within each cluster the stocks are split into stocks which belong to the dominant industry group (larger typeface) and stocks which are in other groups (smaller typeface). For example, in the aqua coloured group between eight and nine o'clock, the dominant industry is financials and all but three stocks belong to this sector.

Table 1. Below are listed the one or two letter industry designations that were appended to a stock's three letter ticker symbol to indicate the main industry within which it operates. The industry groups were assigned by the exchange and are used in Figures 2–6.

Industry	Appended Code	Industry	Appended Code
Consumer Goods	CG	Mining and Materials	M
Consumer Services	CS	Oil and Gas	O
Financials	F	Telecoms	TC
Health	H	Technology	TN
Industrials	I	Utilities	U

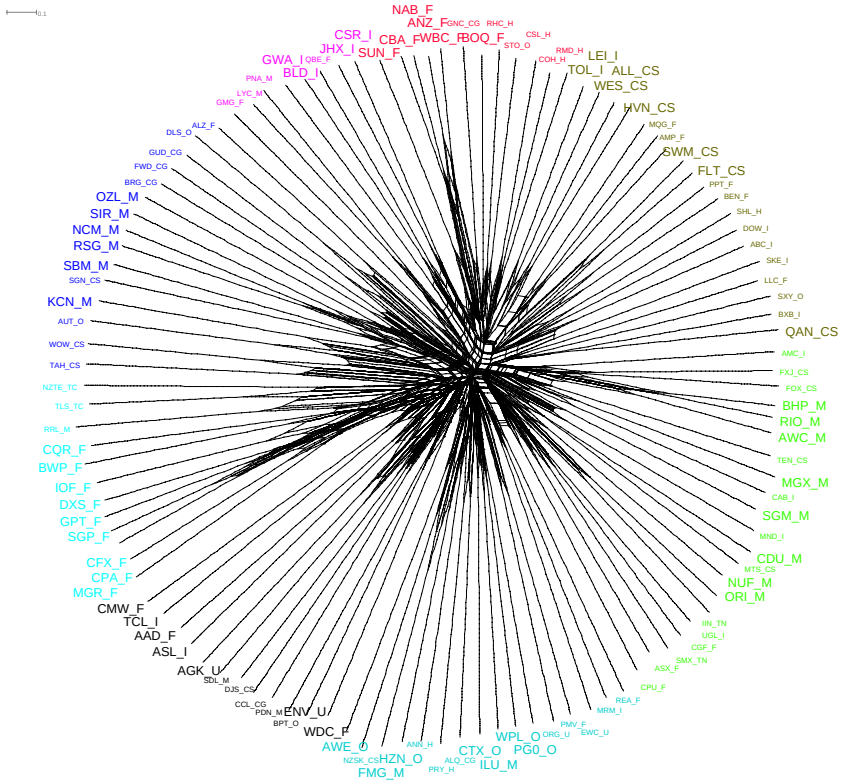

Figure 2. Period 1 neighbor-Nets splits graph with the clusters split into dominant and non-dominant industries. A one or two letter code has been appended to each stock's three letter ticker symbol to indicate what industry group it belongs to using the codes assigned in Table 1.

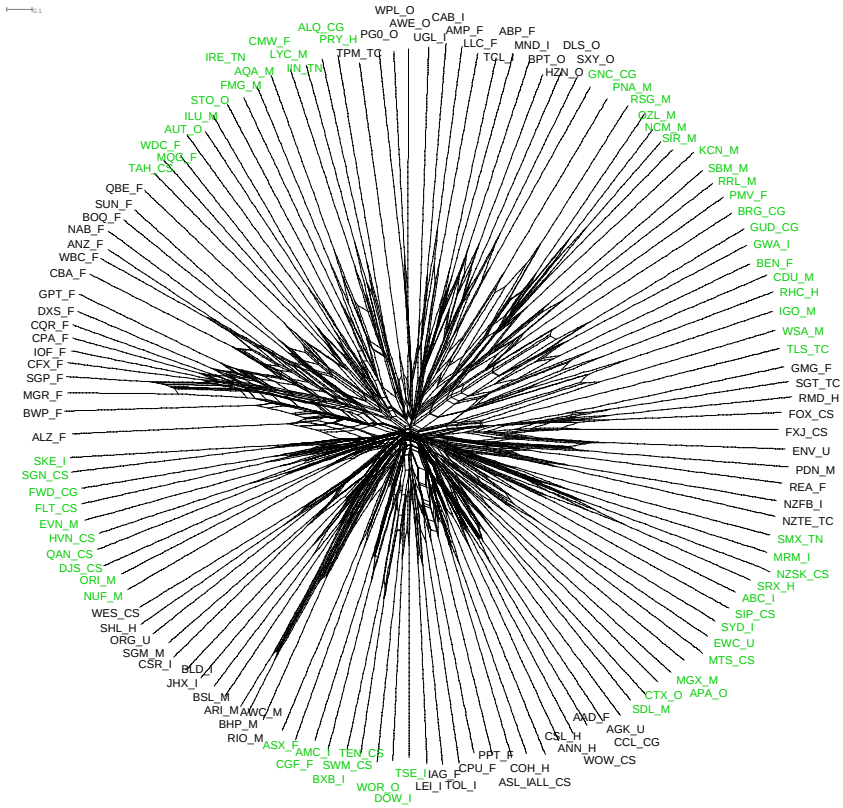

Figure 3. Period 2 neighbor-Nets splits graph with 10 clusters. A one or two letter code has been appended to each stock's three letter ticker symbol to indicate what industry group it belongs to using the codes assigned in Table 1.

Figure 3 presents the neighbor-Nets splits graph for period 2. While the clusters in this figure have not been visually separated into dominant and non-dominate industry groups, the cluster between nine and ten o'clock is dominated by financial stocks. It is also straight-forward to see how the clusters were paired. The cluster just mentioned would be paired with the black coloured cluster between about 2:30 and 3:30 on the opposite side of the network. These stocks are the most distant in terms of the circular ordering. As indicated above, if the correlation clusters represent useful financial groupings of stocks we would expect that choosing a pair of stocks from these two clusters would be likely to give a greater reduction in risk than two stocks selected randomly because of the distance between them.

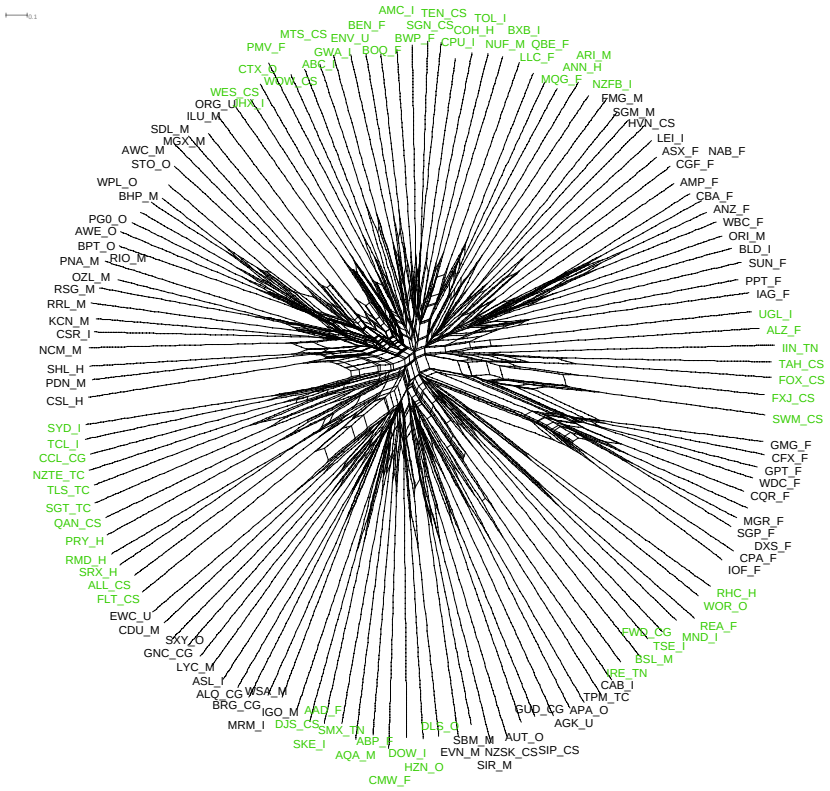

Figure 4. Period 3 neighbor-Nets splits graph with 10 clusters. A one or two letter code has been appended to each stock's three letter ticker symbol to indicate what industry group it belongs to using the codes assigned in Table 1.

Figure 4 gives a good example of where the cluster pairing may not be one-to-one. Consider the green coloured cluster at the top of the splits graph. It is relatively large and we shall call it cluster one. At the bottom of the graph are three smaller clusters, two black and one green coloured which we shall call clusters six, seven and eight, reading the circular ordering in a clockwise direction. The most distant cluster from cluster one (the cluster at 12 o'clock) would be cluster seven (the green coloured cluster at six o'clock). However, we should pair both clusters six and seven (the black coloured cluster at five o'clock and the green coloured cluster at six o'clock respectively) with cluster one. This illustrates the fact that while all clusters have a cluster they are paired with, not all clusters are a reciprocal pair.

To summarize, a stock market analyst or portfolio manager looks for breaks in the structure of the neighbor-Net network when dividing the stocks into correlation clusters. The `SplitsTree` software has considerable flexibility to magnify sections of the network to aid in decision making which cannot be easily captured in the static image outputs included in this paper. The circular ordering can be very useful when splitting a correlation cluster into its component industry groups because if one or more of the resulting groups are too small to be useful they can be joined with groups next to them in the circular order.

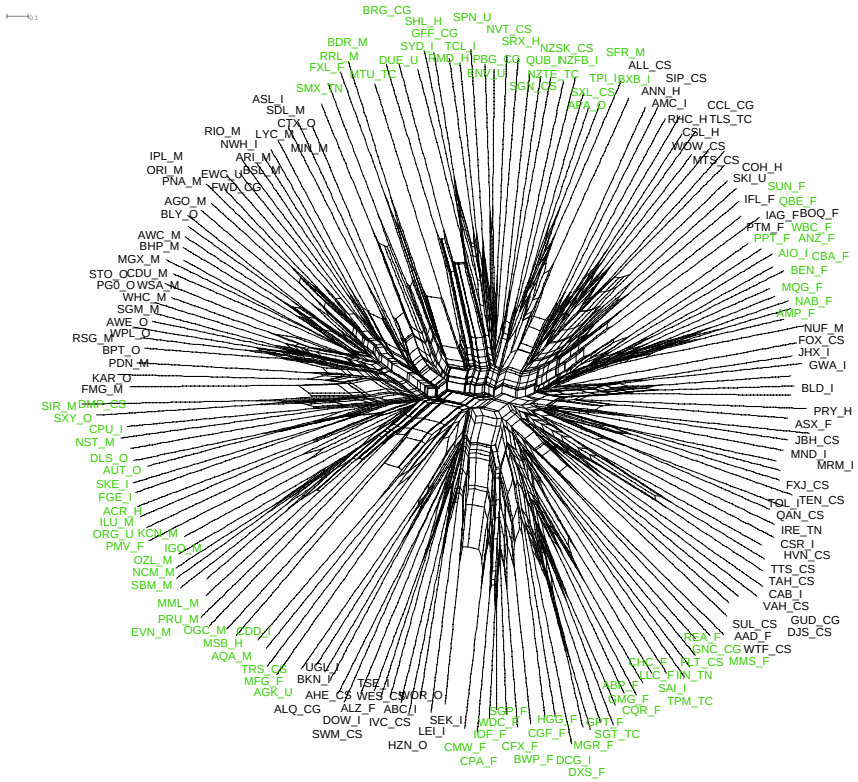

Figure 5. Period 4 neighbor-Nets splits graph with eight clusters. A one or two letter code has been appended to each stock's three letter ticker symbol to indicate what industry group it belongs to using the codes assigned in Table 1.

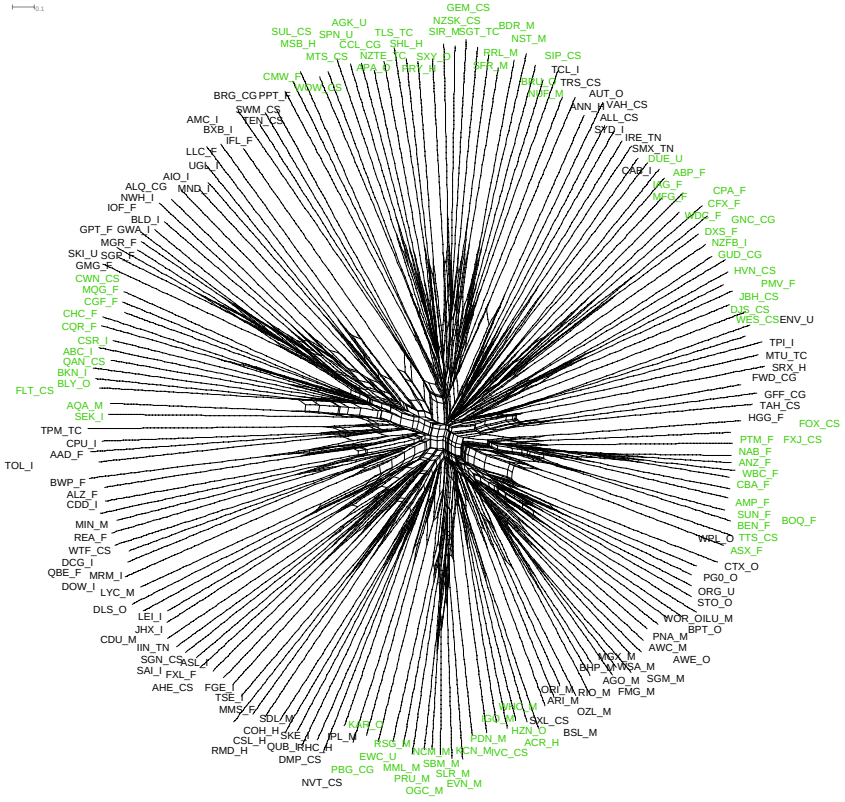

Figure 6. Period 5 neighbor-Nets splits graph with 10 clusters. A one or two letter code has been appended to each stock's three letter ticker symbol to indicate what industry group it belongs to using the codes assigned in Table 1.

4. Results

In the simulated portfolios section we outlined five methods for selecting portfolios. Three of methods depend on correlations clusters (among other things such as industry). In the previous section we presented the neighbor-Net splits graphs for periods 1 to 5 and described how each of these graphs were amenable to defining correlation clusters. In this section we present the results of the five methods for selecting portfolios as applied to the ASX200. Tables 2–6 together with Figures 7–11 present the results of the portfolio selection simulations.

Table 2. Period 2 portfolio performances with standard deviation for three different portfolio sizes. The selection method with the lowest standard deviation is marked with an dagger (†). The neighbor-Nets method is out-of-sample testing using the correlation clusters determined from Period 1 data. The first Levene test is the *p*-value whether the standard deviation of the three different portfolio selection methods are equal. The second result is for the neighbor-Net clusters split into dominant and non-dominant industry groupings.

Period 2 Simulation Results	Random	Neighbor-Net's Correlation Cluster	Industry Group	Correlation Cluster with Industry Group	Correlation Cluster without Industry Group
Mean return					
(2-stock)	106.49	94.93	98.52	101.80	97.24
(4-stock)	101.51	98.45	96.93	97.55	97.12
(8-stock)	104.66	96.90	100.82	98.892	96.14
Std. Dev.					
(2-stock)	79.34	† 70.51	73.87	81.90	† 62.47
(4-stock)	† 49.39	51.85	48.96	53.14	† 43.62
(8-stock)	† 33.61	35.64	36.56	42.33	† 29.81
Sharpe Ratio					
(2-stock)	1.34	1.35	1.33	1.24	1.56
(4-stock)	2.06	1.90	1.98	1.84	2.23
(8-stock)	3.11	2.72	2.69	2.70	3.23
Levene Tests					
(2-stock)			0.001		1.0×10^{-7}
(4-stock)			0.30		1.7×10^{-6}
(8-stock)			0.13		1.3×10^{-12}

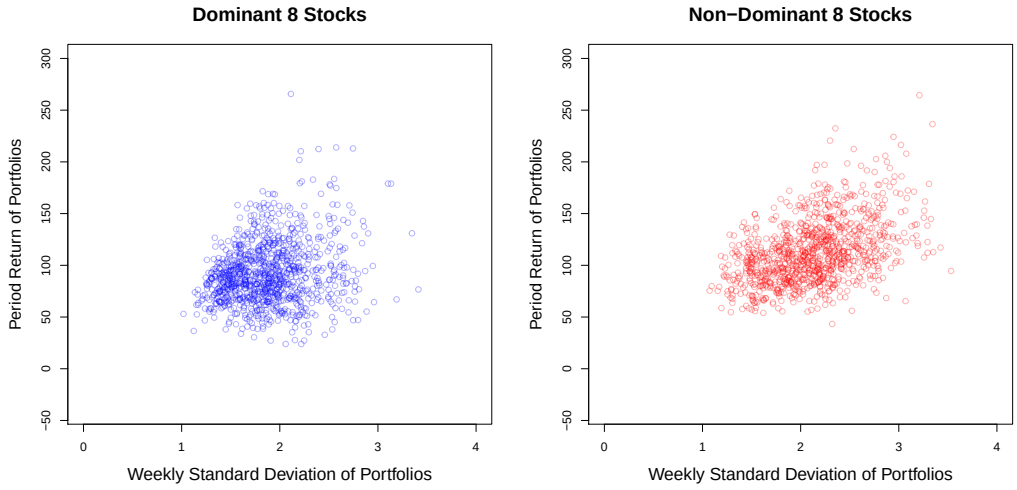

Figure 7. Scatter plots of Period 2 returns against the weekly volatility for the eight-stock portfolios selected on the basis of the clusters identified in neighbor-Nets splits graph from Period 1 with the clusters split into dominant and non-dominant industry groups.

In the third part of each table we have labelled the results presented there the "Sharpe Ratio" though this is clearly not the ratio of [29] (the Sharpe ratio is properly applied to single assets or single portfolios and estimates the reward to risk ratio). We have divided the mean portfolio return by the standard deviation of the returns estimated from the replications. A higher ratio indicates either a higher mean return or a lower spread of returns, or some combination of both, generated by that selection method. We believe the results are worth reporting but do not discuss them further in this paper.

Table 2 presents the results for the Period 2 simulations which was a period of strongly rising equity prices. The model building period (Period 1) was a period when the market largely tracked sideways with a small decline over the period. Thus the out-of-sample test represents a strong test of stock selection methods between the two periods which do not resemble each other.

We are primarily concerned with reducing the risk of the portfolios. The results in the table give the mean return and standard deviation of those returns of the 1000 replications of the portfolio selection method. A lower standard deviation indicates that the returns of the portfolios were more concentrated about the mean portfolio return. The Levene test is a statistical test of whether the standard deviation of two or more groups are equal. In each of the tables two Levene test p-values are reported. The vertical line between columns four and five in each table separates the two Levene tests grouping the results with the methods to which they apply. The first reported p-value tests whether the standard deviation of portfolios formed using the random, industry group, and neighbor-Net correlation cluster selection methods had equal standard deviations. The second reported p-value tests whether the standard deviation of portfolios formed using the dominant and non-dominant industry groups within the correlation clusters selection methods had equal standard deviations.

The first set of Levene tests, in column four, show that, of the three portfolio sizes simulated, only the portfolios of two stocks had a significantly different spread of returns. The difference between the random selection method and neighbor-Nets selection was almost nine percentage points different.

The final two columns of Table 2 present the results for the simulations in which the correlation clusters were divided into dominant and non-dominant industry groups. In this case, according to the Levene test p-values in column six, the correlation clusters of non-dominant industries showed statistically significantly lower levels of spread of the portfolio returns compared to the correlation clusters of dominant industries for all three portfolio sizes simulated.

Figure 7 plots the weekly standard deviation of the portfolio returns, a measure of volatility, against the period portfolio return for the eight-stock portfolios dividing the stocks in to dominant and non-dominant industry groups. The differences are not pronounced but the spread of returns is smaller for the correlation clusters with non-dominant industry groups though the weekly volatility appears comparable.

Table 3 presents the results for the Period 3 simulations which was a period of strongly rising equity prices. The model building period (Period 2) was also a period of market increases. Thus the out-of-sample test and model building periods closely resemble each other.

The Levene tests show that only the two stock portfolios had a significantly different spread of returns, though the results for the eight-stock portfolios almost reached statistical significance. The difference between the neighbor-Nets selection and industry group selection was almost 10 percentage points different.

The final two columns of Table 3 present the results for the simulations in which the correlation clusters were divided into dominant and non-dominant industry groups. In this case the correlation clusters of the dominant industries showed statistically significantly lower levels of spread of the portfolio returns for all three portfolios sizes.

Figure 8 plots the weekly standard deviation of the portfolio returns, a measure of volatility, against the period portfolio return for the eight-stock portfolios dividing the stocks in to dominant and non-dominant industry groups. It is noticeable that the spread of returns is smaller for the correlation clusters with dominant industry groups though the weekly volatility is higher.

Table 3. Period 3 portfolio performances with standard deviation for three different portfolio sizes. The selection method with the lowest standard deviation is marked with an dagger (†). The neighbor-Nets method is out-of-sample testing using the correlation clusters determined from Period 2 data. The first Levene test is the *p*-value whether the standard deviation of the three different portfolio selection methods are equal. The second result is for the neighbor-Net clusters split into dominant and non-dominant industry groupings.

Period 3 Simulation Results	Random	Neighbor-Net's Correlation Cluster	Industry Group	Correlation Cluster with Industry Group	Correlation Cluster without Industry Group
Mean return					
(2-stock)	156.32	149.30	135.11	137.09	150.97
(4-stock)	152.99	151.27	134.02	132.85	155.49
(8-stock)	154.30	149.05	137.79	134.82	152.94
Std. Dev.					
(2-stock)	108.11	109.64	† 99.57	† 66.02	116.40
(4-stock)	75.93	75.47	† 71.01	† 45.62	83.45
(8-stock)	51.34	50.15	† 48.26	† 32.49	57.63
Sharpe Ratio					
(2-stock)	1.45	1.36	1.36	2.07	1.27
(4-stock)	2.01	2.00	1.89	2.91	1.86
(8-stock)	3.01	2.97	2.86	4.15	2.65
Levene Tests					
(2-stock)			0.05		$<10^{-16}$
(4-stock)			0.15		$<10^{-16}$
(8-stock)			0.06		$<10^{-16}$

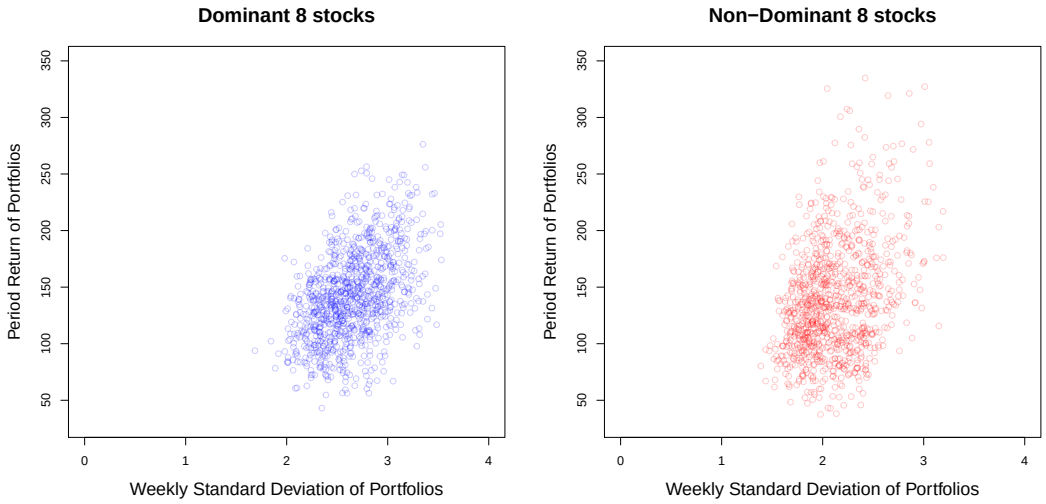

Figure 8. Scatter plots of Period 3 returns against the weekly volatility for the eight-stock portfolios selected on the basis of the clusters identified in neighbor-Nets splits graph from Period 2 with the clusters split into dominant and non-dominant industry groups.

Table 4 presents the result for the Period 4 simulations. Period 4 was a period of strongly falling equity prices. The model building period (Period 3) was a period of strong market increases. Thus the out-of-sample test and model building periods are effectively opposites of each other.

The Levene tests show that no stock portfolios had a significantly different spread of returns.

Table 4. Period 4 portfolio performances with standard deviation for three different portfolio sizes. The selection method with the lowest standard deviation is marked with an dagger (†). The neighbor-Nets method is out-of-sample testing using the correlation clusters determined from Period 3 data. The first Levene test is the *p*-value whether the standard deviation of the three different portfolio selection methods are equal. The second result is for the neighbor-Net clusters split into dominant and non-dominant industry groupings.

Period 4 Simulation Results	Random	Neighbor-Net's Correlation Cluster	Industry Group	Correlation Cluster with Industry Group	Correlation Cluster without Industry Group
Mean return					
(2-stock)	−46.58	−49.18	−45.56	−54.61	−43.26
(4-stock)	−47.52	−49.46	−43.93	−54.81	−42.19
(8-stock)	−47.48	−48.52	−44.09	−54.70	−42.18
Std. Dev.					
(2-stock)	22.80	† 21.30	21.70	† 17.16	23.49
(4-stock)	† 15.16	15.41	15.68	† 11.29	16.80
(8-stock)	10.65	10.49	† 10.42	† 8.03	11.11
Sharpe Ratio					
(2-stock)	−2.04	−2.31	−2.10	−3.18	−1.85
(4-stock)	−3.13	−3.21	−2.80	−4.85	−2.51
(8-stock)	−4.46	−4.62	−4.25	−6.81	−3.80
Levene Tests					
(2-stock)			0.78		1.6×10^{-9}
(4-stock)			0.32		$<10^{-16}$
(8-stock)			0.73		$<10^{-16}$

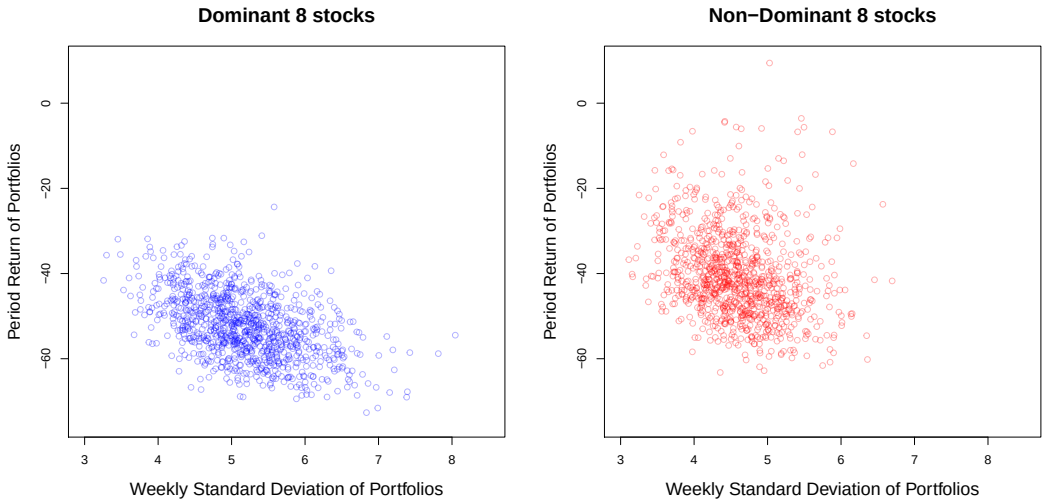

Figure 9. Scatter plots of Period 4 returns against the weekly volatility for the eight-stock portfolios selected on the basis of the clusters identified in neighbor-Nets splits graph from Period 3 with the clusters split into dominant and non-dominant industry groups.

The final two columns of Table 4 present the results for the simulations in which the correlation clusters were divided into dominant and non-dominant industry groups. In this case the correlation clusters of the dominant industries showed statistically significantly lower levels of spread of the portfolio returns. The Levene tests were highly significant for all portfolio sizes.

Figure 9 plots the weekly standard deviation of the portfolio returns against the period portfolio return for the eight-stock portfolios dividing the stocks in to dominant and non-dominant industry groups. It is noticeable that the spread of returns is substantially smaller for the correlation clusters with dominant industry groups though, again, the weekly volatility is higher.

Table 5. Period 5 portfolio performances with standard deviation for three different portfolio sizes. The selection method with the lowest standard deviation is marked with an dagger (†). The neighbor-Nets method is out-of-sample testing using the correlation clusters determined from Period 4 data. The first Levene test is the *p*-value whether the standard deviation of the three different portfolio selection methods are equal. The second result is for the neighbor-Net clusters split into dominant and non-dominant industry groupings.

Period 5 Simulation Results	Random	Neighbor-Net's Correlation Cluster	Industry Group	Correlation Cluster with Industry Group	Correlation Cluster without Industry Group
Mean return					
(2-stock)	164.49	159.20	156.22	142.77	160.07
(4-stock)	162.29	150.17	160.37	149.74	164.06
(8-stock)	162.43	154.90	162.25	151.25	163.34
Std. Dev.					
(2-stock)	144.70	149.58	† 141.50	† 124.24	155.57
(4-stock)	105.14	† 94.92	99.47	† 83.90	106.43
(8-stock)	70.68	† 68.04	70.61	† 60.30	73.00
Sharpe Ratio					
(2-stock)	1.14	1.06	1.10	1.15	1.03
(4-stock)	1.54	1.58	1.61	1.78	1.54
(8-stock)	2.30	2.28	2.38	2.51	2.24
Levene Tests					
(2-stock)			0.46		3.9×10^{-5}
(4-stock)			0.003		3.3×10^{-8}
(8-stock)			0.35		1.4×10^{-8}

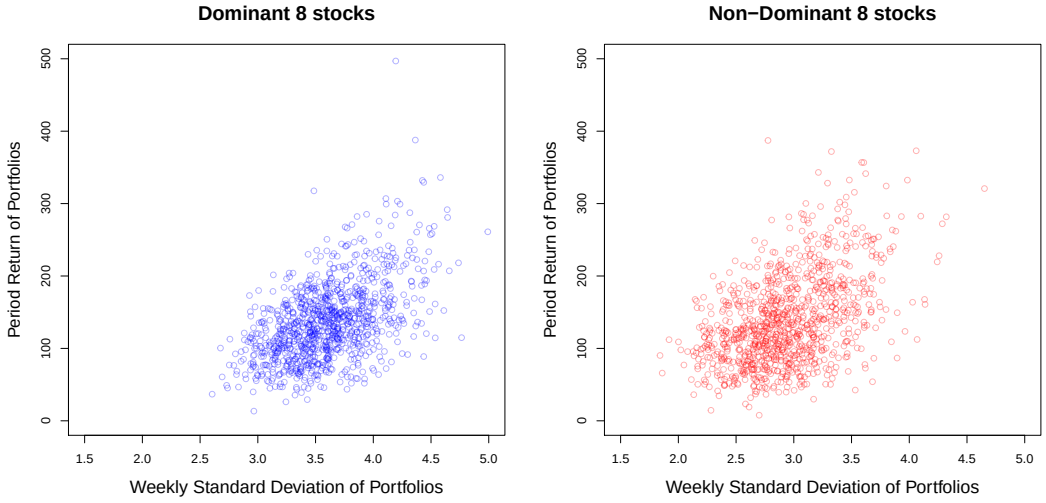

Figure 10. Scatter plots of Period 5 returns against the weekly volatility for the eight-stock portfolios selected on the basis of the clusters identified in neighbor-Nets splits graph from Period 4 with the clusters split into dominant and non-dominant industry groups.

Table 5 presents the results for the Period 5 simulations which was a period of equity prices initially rebounding then tracking sideways. The model building period (Period 4) was a period of strong market decreases. Thus the out-of-sample test and model building periods are substantially different.

The Levene tests show that only the four stock portfolios had a significantly different spread of returns.

The final two columns of Table 5 present the results for the simulations in which the correlation clusters were divided into dominant and non-dominant industry groups. In this case the correlation clusters of the dominant industries showed statistically significantly lower levels of spread of the portfolio returns. The Levene tests were strongly significant for all portfolios sizes.

Figure 10 plots the weekly standard deviation of the portfolio returns against the period portfolio return for the eight-stock portfolios dividing the stocks in to dominant and non-dominant industry groups. It is noticeable that the spread of returns is smaller for the correlation clusters with dominant industry groups though, again, the weekly volatility is higher.

Table 6 presents the result for the Period 6 simulations which was a period of rising equity prices with significant volatility. The model building period (Period 5) was a period of rebound followed by a time of stacking sideways. Thus the out-of-sample test and model building periods have some similarities.

Table 6. Period 6 portfolio performances with standard deviation for three different portfolio sizes. The selection method with the lowest standard deviation is marked with an dagger (†). The neighbor-Nets method is out-of-sample testing using the correlation clusters determined from Period 5 data. The first Levene test is the *p*-value whether the standard deviation of the three different portfolio selection methods are equal. The second result is for the neighbor-Net clusters split into dominant and non-dominant industry groupings.

Period 6 Simulation Results	Random	Neighbor-Net's Correlation Cluster	Industry Group	Correlation Cluster with Industry Group	Correlation Cluster without Industry Group
Mean return					
(2-stock)	45.96	46.60	54.36	34.77	65.16
(4-stock)	48.39	51.10	55.19	35.85	61.87
(8-stock)	46.66	50.32	55.62	35.36	62.09
Std. Dev.					
(2-stock)	52.03	† 49.87	51.62	† 44.92	51.32
(4-stock)	37.28	† 33.18	34.11	† 28.49	35.65
(8-stock)	25.56	† 21.63	22.94	† 15.04	24.04
Sharpe Ratio					
(2-stock)	0.88	0.93	1.05	0.77	1.27
(4-stock)	1.30	1.54	1.62	1.26	1.74
(8-stock)	1.76	2.33	2.42	2.35	2.58
Levene Tests					
(2-stock)			0.64		0.004
(4-stock)			0.007		4.0×10^{-9}
(8-stock)			1.1×10^{-9}		$<10^{-16}$

The Levene tests show that the four and eight stock portfolios had a significantly different spread of returns. In both cases the neighbor-Nets portfolio selection method had the lowest spread of returns.

The final two columns of Table 6 present the results for the simulations in which the correlation clusters were divided into dominant and non-dominant industry groups. In this case the correlation clusters of the dominant industries showed statistically significantly lower levels of spread of the portfolio returns. The Levene tests were highly significant for all portfolios sizes.

Figure 11 plots the weekly standard deviation of the portfolio returns against the period portfolio return for the eight-stock portfolios dividing the stocks in to dominant and non-dominant industry groups. It is noticeable that the spread of returns is smaller for the correlation clusters with dominant industry groups though, as with some previous periods, the weekly volatility is higher.

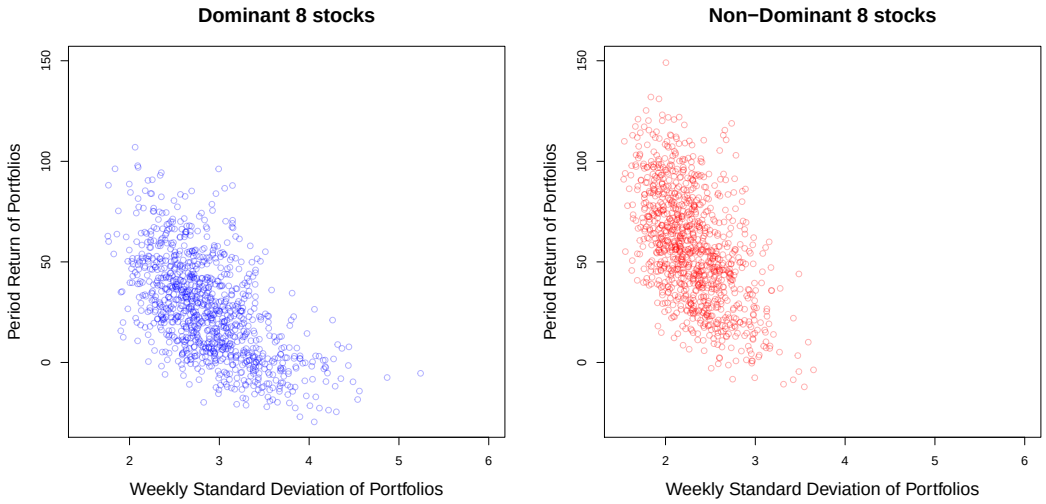

Figure 11. Scatter plots of Period 6 returns against the weekly volatility for the eight-stock portfolios selected on the basis of the clusters identified in neighbor-Nets splits graph from Period 5 with the clusters split into dominant and non-dominant industry groups.

5. Discussion and Conclusions

The simulation tests performed here represent a particularly challenging test of portfolio diversification because of the long out-of-sample test periods coupled with the fact that the market conditions in the model building phase were often very different from the market conditions in the test phase.

In the 15 sets of simulations comparing, random, industry group, and neighbor-Net correlation cluster selection methods, statistically significant differences in the portfolios standard deviation were obtained only five times. In four cases the neighbor-Net correlation cluster produced the lowest standard deviation and in one case the industry group selection method was the lowest.

Considering the neighbor-Net correlation clusters split into dominant and non-dominant industries, all 15 cases had statistically significant differences in the portfolios' standard deviation. Of these 12 were correlation clusters with dominant industry groups in Periods 3–6 and three were correlation clusters with non-dominant industry groups in Period 2.

Figures 7–11 present this last observation in graphical form and it is clear that the distribution both in terms of portfolio returns and weekly volatility are different for all periods. Graphs of the two and four stock portfolios show similar results but are not reported here.

These results suggest that within each correlation cluster there are two distinct sub-populations of stocks. Intuitively, the dominant industry group are simply stocks within the same industry with similar risk characteristics. We would expect such stocks to be strongly correlated, hence fall into the same correlation cluster. We would also expect that they would continue to be strongly correlated into the future. On the other hand, the stocks in the non-dominant industry group within a correlation cluster would seem far less likely to remain strongly correlated in the future.

At this stage we can only recommend that this be investigated further. The differences between the two groups of stocks appear significant both in a statistical and economic sense but it is not yet clear how to exploit this difference for financial gain.

This paper is primarily concerned about reducing risk in small portfolios. It would be of interest to study a larger set of stocks and investment options and consider the impact the methods would have on diversification of larger portfolios, the next logical sizes would be 16 and 32 stocks. It would also be of interest to see if the method yield better results if the out-of-sample testing phase was substantially reduced. Both these consideration wait for further research.

Investors are also concerned about returns and in three of the five periods (two, three and five) the randomly selected portfolios had the highest returns among the three methods which included all stocks. Truly negative correlations among stocks are uncommon and in a strongly rising market a negative correlation between a pair of stocks often indicates that one stock is suffering from some form of financial distress and hence a falling share price. The correlation cluster method excludes many stock combinations available to the random selection method but also increases the probability of a stock being paired with one in financial distress hence depressing overall portfolio performance. If this explanation is correct then the application of financial analysis aimed at removing financially distressed stocks may well enhance the performance of all portfolio selection methods. Again, this must wait for further research.

Initially we asked the question whether graph theory portfolio selection methods could yield useful results for the portfolio manager. The results given here suggest that when used alone the answer is likely to be no, but when coupled with other financial information, here the industry group the stock is in, the answer is a provisional yes.

Acknowledgments: The authors would like to thank David Bryant for his helpful comments.

Author Contributions: This paper is drawn from Hannah Zhan's Master of Commerce thesis entitled "An Alternative Approach to Visualizing Stock Market Correlation Matrices—An Empirical study of forming portfolios that contain only small numbers of stocks using both existing and newly discovered visualization methods" . William Rea and Alethea Rea were, respectively, the senior and associate supervisors of the thesis. They jointly wrote the paper based on Hannah's work.

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

Abbreviations

The following abbreviations are used in this manuscript:

ASX200	Australian Stock Exchange 200 Index

Appendix A. Neighbor-Net

This appendix gives a more technical description of the Neighbor-Net algorithm.

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 $1 \leq i \leq j < n$. As highlighted above, the advantage of this set of splits is that they can be represented on a plane.

We describe the algorithm following [23]. 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

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d_{xy}, \tag{A1}$$

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

$$Q(C_i, C_j) = (m-2)d(C_i, C_j) - \sum_{k=i, k \neq i} md(C_i, C_k) - \sum_{k=i, k \neq i} md(C_j, C_k), \tag{A2}$$

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

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

$$\hat{Q}(x_i, x_j) = (\hat{m}-2)d(x_i, x_j) - \sum_{k=i, k \neq i} \hat{m}d(x_i, C_k) - \sum_{k=i, k \neq i} \hat{m}d(x_j, C_k) \tag{A3}$$

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

$$d(u, a) = (\alpha + \beta)d(x, a) + \gamma d(y, a) \tag{A4}$$

$$d(v, a) = \alpha d(y, a) + (\beta + \gamma)d(z, a) \tag{A5}$$

$$d(u, v) = \alpha d(x, y) + \beta d(x, z) + \gamma d(y, z) \tag{A6}$$

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 [23] 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

$$\mathbf{X}_{ij,k} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ are on opposite sides of the split} \\ 0 \text{ if } i \text{ and } j \text{ are on the same side of the split.} \end{cases} \tag{A7}$$

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.

The network, or splits graph, generated by Neighbor-Nets has three biologically meaningful components. The places where a line splits represents a speciation event, where a single population becomes two genetically isolated populations. The places where two lines join to become one represents a recombination event, where two genetically isolated populations exchange genetic material. The lengths of the individual lines represent the time the population evolves without either a speciation or recombination event. The interpretation of these three components in a financial context is an active area of research for the authors.

References

1. Lowenfeld, H. *Investment, an Exact Science*; Financial Review of Reviews: London, UK, 1909.
2. Markowtiz, H. Portfolio Selection. *J. Finance* **1952**, *7*, 77–91.

3. Markowitz, H.M. *Portfolio Selection: Efficient Diversification of Investments*, 2nd ed.; Blackwell Pubishing: Malden, MA, USA, 1991.
4. Evans, J.L.; Archer, S.H. Diversification and the Reduction of Dispersion: An Empirical Analysis. *J. Finance* **1968**, *23*, 761–767.
5. Domian, D.L.; Louton, D.A.; Racine, M.D. Diversification in Portfolios of Individual Stocks: 100 Stocks Are Not Enough. *Financ. Rev.* **2007**, *42*, 557–570.
6. Cont, R. Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance* **2001**, *1*, 223–236.
7. Goyal, A.; Santa-Clara, P. Idiosyncratic Risk Matters! *J. Finance* **2003**, *58*, 975–1007.
8. Bali, T.G.; Cakici, N.; Yan, X.; Zhang, Z. Does Idiosyncratic Risk Really Matter? *J. Finance* **2005**, *60*, 905–929.
9. Fama, E.F.; French, K.R. The Cross-Section of Expected Stock Returns. *J. Finance* **1992**, *47*, 427–465.
10. French, K.R.; Fama, E.F. Common Risk Factors in the Returns on Stocks and Bonds. *J. Financ. Econ.* **1993**, *33*, 3–56.
11. Benzoni, L.; Collin-Dufresne, P.; Goldstein, R.S. Portfolio Choice over the Life-Cycle when the Stock and Labor Markets are Cointegrated. *J. Finance* **2007**, *62*, 2123–2167.
12. Jorion, P. International Portfolio Diversification with Estimation Risk. *J. Bus.* **1985**, *58*, 259–278.
13. Bai, Y.; Green, C.J. International Diversification Strategies: Revisited from the Risk Perspective. *J. Bank. Finance* **2010**, *34*, 236–245.
14. Mantegna, R.N. Hierarchical structure in financial markets. *Eur. Phys. J. B* **1999**, *11*, 193–197.
15. Onnela, J.P.; Chakraborti, A.; Kaski, K.; Kertèsz, J.; Kanto, A. Asset Trees and Asset Graphs in Financial Markets. *Phys. Scr.* **2003**, *T106*, 48–54.
16. Onnela, J.P.; Chakraborti, A.; Kaski, K.; Kertèsz, J.; Kanto, A. Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E* **2003**, *68*, 0561101.
17. Bonanno, G.; Calderelli, G.; Lillo, F.; Micciché, S.; Vandewalle, N.; Mantegna, R.N. Networks of equities in financial markets. *Eur. Phys. J. B* **2004**, *38*, 363–371. .
18. Micciché, S.; Bonannon, G.; Lillo, F.; Mantegna, R.N. Degree stability of a minimum spanning tree of price return and volatility. *Physica A* **2006**, *324*, 66–73.
19. Naylor, M.J.; Rose, L.C.; Moyle, B.J. Topology of foreign exchange markets using hierarchical structure methods. *Physica A* **2007**, *382*, 199–208.
20. Kenett, D.Y.; Tumminello, M.; Madi, A.; Gur-Gershgoren, G.; Mantegna, R.N.; Ben-Jacob, E. Systematic analysis of group identification in stock markets. *PLoS ONE* **2010**, *5*, e15032.
21. Djauhari, M.A. A Robust Filter in Stock Networks Analysis. *Physica A* **2012**, *391*, 5049–5057.
22. Rea, A.; Rea, W. Visualization of a stock market correlation matrix. *Physica A* **2014**, *400*, 109–123.
23. Bryant, D.; Moulton, V. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Mol. Biol. Evol.* **2004**, *21*, 255–265.
24. DiMiguel, V.; Garlappi, L.; Uppal, R. Optimal versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *Rev. Financ. Stud.* **2009**, *22*, 1915–1953.
25. Barber, B.M.; Odean, T. All That Glitters: The Effect of Attention and News on the Buying Behaviour of Individual and Institutional Investors. *Rev. Financ. Stud.* **2008**, *21*, 785–818.
26. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.
27. Huson, D.H.; Bryant, D. Application of Phylogenetic Networks in Evolutionary Studies. *Mol. Biol. Evol.* **2006**, *23*, 255–265.
28. Lee, W. Risk-Based Asset Allocation: A New Answer to an Old Question? *J. Portf. Manag.* **2011**, *37*, 11–28.
29. Sharpe, W.F. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *J. Finance* **1964**, *19*, 13–37.