2018

Study of three Network Portfolio Selection Methods

SHUBHAM PAWAR, MANSI SHETH AND ANURAG DUTT

INTRODUCTION

Portfolio diversification is a strategy to combine a variety of assets to reduce the overall risk of an investment portfolio. It is essential for risk management as it reduces the variance of returns compared with a portfolio of a single stock. An investor can reduce portfolio risk simply by holding combinations of instruments that are not perfectly positively correlated. In other words, investors can reduce their exposure to individual asset risk by holding a diversified portfolio of assets. Diversification may allow for the same portfolio expected return with reduced risk. Modern Portfolio Theory assumes that investors are risk averse, meaning that given two portfolios that offer the same expected return, investors will prefer the less risky one. Thus, an investor will take on increased risk only if compensated by higher expected returns. Conversely, an investor who wants higher expected returns must accept more risk. The exact trade-off will be the same for all investors, but different investors will evaluate the trade-off differently based on individual risk aversion characteristics.

Markowitz assumed the following behavioral traits before postulating his mean-variance model:

1. Risk of a portfolio is based on the variability of returns from the said portfolio.
2. An investor is risk averse.
3. An investor prefers to increase consumption.
4. The investor's utility function is concave and increasing, due to his risk aversion and consumption preference
5. Analysis is based on single period model of investment.
6. An investor either maximizes his portfolio return for a given level of risk or maximizes his return for the minimum risk.
7. An investor is rational in nature.

The mean variance portfolio theory has certain limitations though:

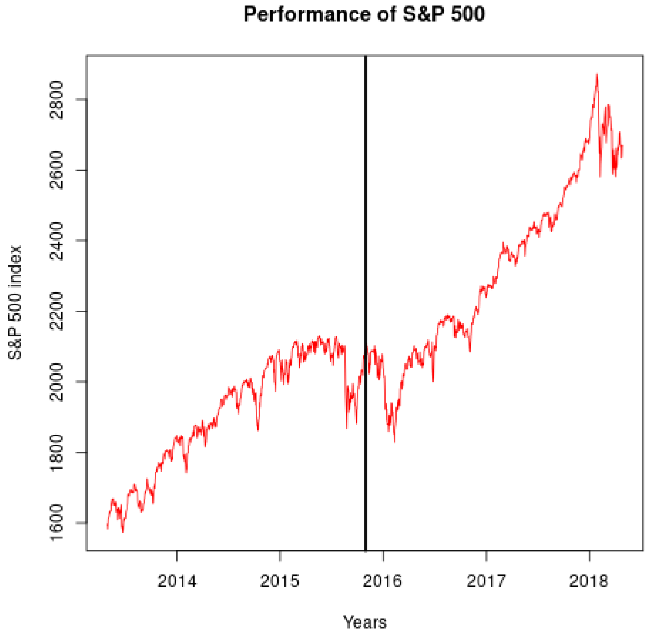
1. The efficient is calculated for single period of investment and thus needs frequent re-balancing as new data is received. This is not necessarily a significant issue but it has been observed that even small changes in correlation values between the stocks leads to major re-balancing signals in the portfolio which is not viable due high transaction costs to maintain such a portfolio.
2. The model assumes that correlations and expected returns of the assets are not time-varying thus can be accurately estimated from historical data or, alternatively, they can be forecast accurately, which leads to estimation error maximization.
3. The amount of information (the covariance matrix, specifically, or a complete joint probability distribution among assets in the market portfolio) needed to compute a mean-variance optimal portfolio is often intractable and certainly has no room for subjective measurements.

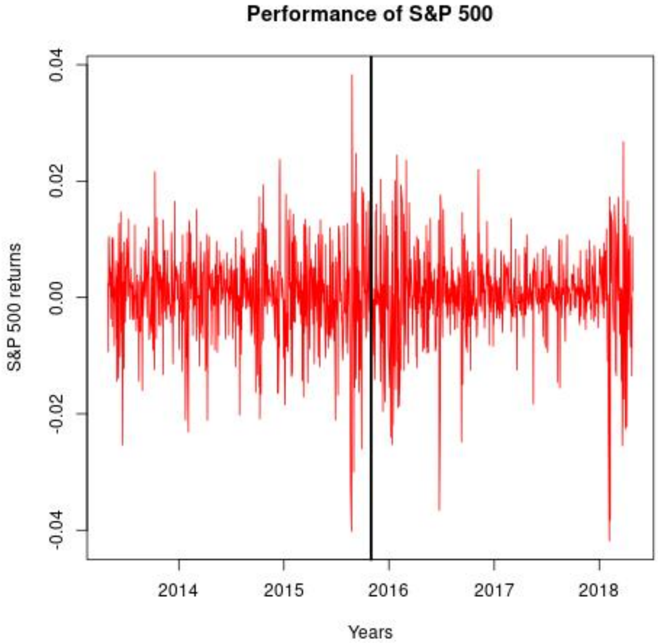
This leaves us to explore the application of network methods for the purpose of portfolio diversification. We apply these methods to the problem of stock selection confining our investible universe to the 30 stocks of the Standard and Poor's S&P-500 Index. The stocks are chosen on three primary categories:

1) We replicate the stocks that are present in the investible universe of the paper "A Comparison of Three Network Portfolio Selection Methods – Evidence from the Dow Jones" by Zhuan et. al[1]. So that we have a viable source to compare our results (All stocks belong to top-50 listed companies ordered by market capitalization.)
2) We have an uninterrupted data source available for downloading the closing prices of the stocks and provide complete time-series information about the stocks including stock splits and stock mergers between 04/28/2013 and 04/28/2018, which is also our period of analysis.

DATA

We downloaded the data for the daily closing prices of 30 S&P 500 stocks for the period April 28, 2013 to April 28, 2018 from Yahoo finance. During our period of analysis, the markets were relatively stable.





This is where we differ from the paper. The period of analysis for the paper is between 2001 and 2013, where the US. stock markets experienced two major upheavals namely the dot-com bubble and the 2008 financial crisis. Our period of analysis is between 2013 and 2018, where the markets were relatively stable as can be seen from the

S&P-500 spot graph. For simplicity, we assumed that the closing prices of the stocks incorporated complete information regarding growth and performance of a stock. We also assumed the dividends paid were reinvested into the stock which issued them, at the closing price on the day the dividend payment was made. The returns are calculated as the log-difference of the prices between two consecutive days so as counter the stationary nature of the time-series of stock-prices. We used the R package "Quantmod" to incorporate the changes in stock-mergers and stock-splits and adjust the returns accordingly.

In our study, we defined 2 periods which are as follows : -

1. April 28, 2013 to October 31, 2015 (633 observations)
2. November 1, 2015 to April 28, 2018 (626 observations)
3.

Period 1 was used for model building and in-sample testing and period 2 was used for out-of-sample testing. The stocks in the sample and their ticker symbols are given below.

Company Name	**Symbol**	**Company Name**	**Symbol**
Goldman Sachs	GS	Home Depot	HD
3M	MMM	Intel	INTC
Alcoa Corporation	AA	IBM	IBM
American Express	AXP	Johnson & Johnson	JNJ
AT&T	T	JPMorgan Chase	JPM
Bank of America	BAC	McDonald's	MCD
Boeing	BA	Merck	MRK
Caterpillar	CAT	Microsoft	MSFT
Chevron	CVX	Pfizer	PFE
Cisco	CSCO	Travelers Companies	TRV
Coca-Cola	CCE	United Technologies	UTX
DuPont Inc	DWDP	United Health	UNH
Exxon Mobil	XOM	Verizon	VZ
General Electric	GE	WalMart	WMT
HP	HPQ	Disney	DIS

CLUSTERING DIAGRAMS

The correlation matrices were generated using R, for each of the two testing periods. Since the correlation values range from -1 to 1, we converted each correlation value to a completely positive metric using the measure:

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

The reasoning behind using such a metric is two-fold:

1. It provides a positive weight for the edges, for us to construct graphs.
2. It provides a higher weight to uncorrelated or negatively correlated stock-pairs, which should be the case when we desire to have effective portfolio diversification.

After clustering the stocks were selected at random from each of the four clusters using a uniform distribution without replacement. In other words, each stock was given an equal chance of being selected, but no stock was selected twice within a portfolio.

Portfolio Simulation from Clusters

The we simulated 1000 portfolios from each of the clusters obtained. For selecting a portfolio, we randomly sampled one stock from each cluster. Since we formed 4 clusters by all of our clustering methods, we present analysis of a portfolio of 4 stocks.

HIERARCHICAL CLUSTERING

Hierarchical Clustering is an agglomerative approach based clustering wherein we find the closest two items, put them together and find the next closest. The components at each iterative step are always a subset of other structures. Hence, the subsets can be represented using a tree diagram or dendrogram.

The leaves at the bottom of the dendrogram represent individual observations or nodes (in our case, stocks). The leaves are combined to form branches until we have grouped all the items. This combination process is reversed to give a rooted Tree structure.

Clustering Hierarchical trees is simple. We maintain a list of splits while agglomerating and move down this list until k splits are seen, k being the number of clusters. For example, to implement 4 clusters, we move down the tree until the split gives 4 clusters.

Algorithm

Given: A set of N items to be clustered and an N x N distance matrix.

1. Begin with N observations and treat each observation as its own cluster.
2. For i = N, N - 1, …, 2
   i) Find the closest pair of clusters based on the linkage between them and merge them into a single cluster.
   ii) Compute distances between the new cluster and the remaining i - 1 clusters.

Linkages

Linkage represents the measure of dissimilarity used by the clustering algorithm while evaluating which nodes to agglomerate. The linkage method used by us during our trial was ‘Complete’ Linkage, where maximum distance is the dissimilarity between two clusters.

HCT Clustering

The cluster assignment as obtained by using HCT clustering on the distance matrix given by ultra-metric in the period one log returns were –

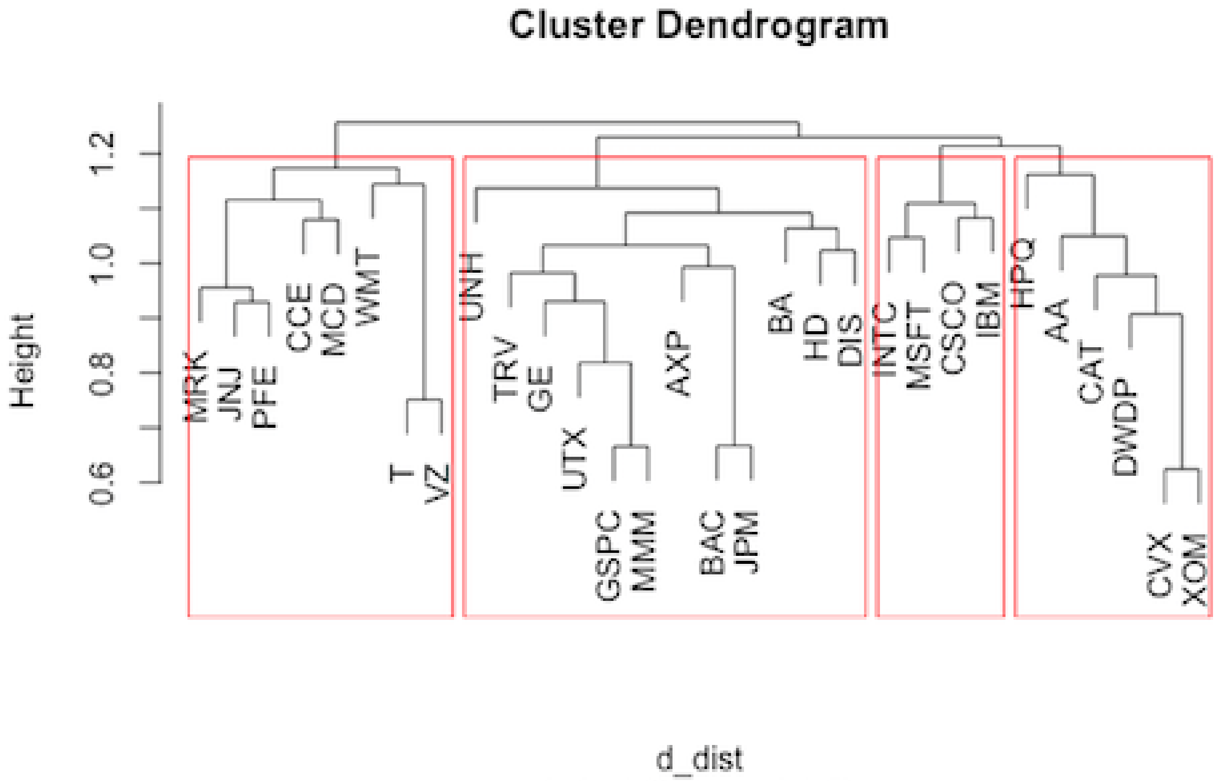
Cluster 1: MRK, JNJ, PFE, CCE, MCD, WMT, T, VZ

Cluster 2: UNH, TRV, GE, UTX, GSPC, MMM, AXP, BAC, JPM, BA, HD, DIS

Cluster 3: INTC, MSFT, CSCO, IBM

Cluster 4: HPQ, AA, CAT, DWDP, CVX, XOM

As can be seen by the cluster assignments, HCT gives highly unbalanced clusters, which can be an issue as this might cause the portfolio to have a high variance and hence might lead to a riskier portfolio.



MINIMUM SPANNING TREE

As already seen in class, the Kruskal's algorithm for finding the minimum spanning tree can be used for clustering the nodes by stopping the algorithm k-1 before, where k is the number of clusters desired. In the paper[1], the authors discuss a more visual clustering based on MST which is suitable of such a small space of stocks. The authors propose finding clear breaks in the tree structure to cluster the stocks together. In our approach, we tried both the methods of clustering – as suggested in class, and as suggested in the paper. We find that, with the clustering algorithms seen in class applied to this problem, the clusters obtained were highly unbalanced (two clusters of a single stock, one cluster of two stocks and all the other stocks in one clusters). Therefore, the results are only for the clusters obtained with the method discussed by the authors in their paper.

MST Clustering

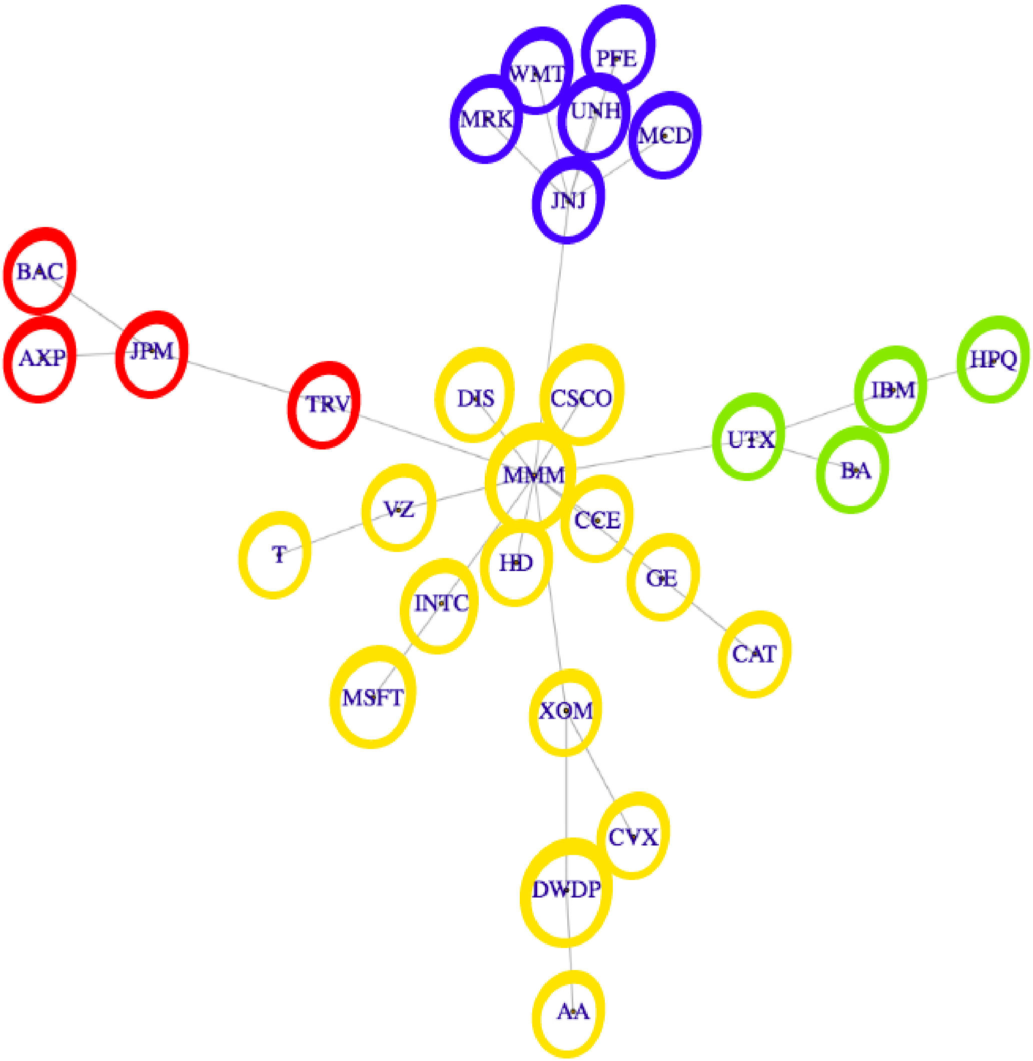
The cluster assignment as obtained by using MST clustering on the distance matrix given by ultra-metric in the period one log returns were –

Cluster 1: MMM, DIS, CSCO, VZ, T, CCE, HD, GE, CAT, INTC, MSFT, XOM, CVX, DWDP, AA

Cluster 2: BAC, AXP, JPM, TRV

Cluster 3: MRK, WMT, PFE, UNH, JNJ, MCD

Cluster 4: UTX, IBM, HPQ, BA



NEIGHBOR-NETS

The next graph theory based method that we used for clustering the stocks together was Neighbor-Net splits graph. As the paper noted, Neighbor-Net Splits graphs give the stocks a circular ordering which gives it an advantage over other techniques considered for clustering. The circular ordering ensures that cluster assignment errors are minimized.

Neighbor-Net is a distance based method for constructing phylogenetic networks which is based on the Neighbor-Joining (NJ) algorithm. The NJ algorithm uses an agglomerative process which converts an arbitrary distance matrix to a fully resolved phylogenetic tree. Phylogenetic networks generalize phylogenetic trees because they permit incompatible splits which may represent alternative phylogenetic histories. A split is a partition of the set of nodes into two disjoint non-empty sets. A collection splits is compatible if it is contained within the set of some phylogenetic tree. Neighbor-Net constructed splits are, generally, incompatible but this affords the ability to represent the phylogeny in presence of conflicting signals when the underlying evolutionary history is not treelike.

Method:

Agglomerative Process

All tree linkage algorithms follow the same general scheme. Start with one node for each taxa (stocks in our case), and at each iteration a pair of nodes is selected and replaced by a new composite node. At the conclusion of the agglomeration, the process is reversed to obtain a tree-like representation of the taxa. Neighbor-Net follows a similar agglomeration procedure but, when a pair of nodes is selected they are not combined and replaced immediately. The combination procedure is deferred until a node is paired twice, at which point the three linked nodes are replaced by two composite nodes. When this process is reversed, the representation of the distance matrix is not of the form of a tree but of a network. The agglomeration process is represented below, the figure is taken from original Neighbor-Net paper[2].

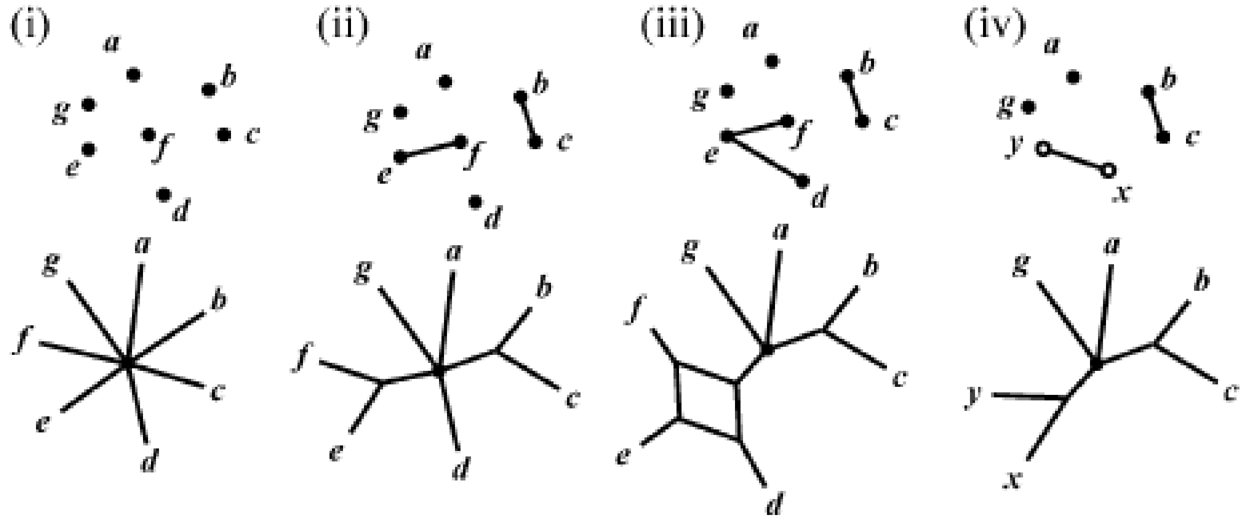


Fig: The Agglomerative Process for Neighbor-Net. (i) We begin with each node representing a single taxon. (ii) using the Selection Criteria, we identify b and c as neighbors also we identify e and f as neighbors. Unlike NJ, algorithm, we don't amalgamate immediately. (iii) We identify e as a neighbor of d. (iv) As e has two neighbors, we perform a reduction replacing d, e, f by x, y.

Selection Formulae

The neighboring relations group the n nodes into clusters C_1, C_2, ..., C_m, $m \leq n$, where each cluster contains a single node or a pair of nodes. The distance between two clusters is the average distance between elements in two clusters –

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

The selection of neighboring nodes proceeds in two steps. First, we find a pair of clusters that minimizes the standard NJ formula given below. Second, we choose particular nodes belonging to the selected clusters which minimizes the formula as the new neighboring nodes. The NJ formula to minimize is given by –

$$Q(C_i, C_j) = (m-2)d(C_i, C_j) - \sum_{k=1, k \neq i}^{m} d(C_i, C_k) - \sum_{k=1, k \neq j}^{m} d(C_j, C_k)$$

Distance-Reduction Formulae

If, after selection step a node is neighboring two different nodes, the three nodes are selected and reduced two nodes. Suppose node *y* has two neighbors, *x* and *z*, the three nodes *x*, *y*, *z* are replaced by two composite nodes *u* and *v*. The distances from these two nodes with another node *a* is then given by the reduction formulae –

$$d(u, a) = \alpha d(x, a) + \beta d(y, a)$$
$$d(v, a) = \beta d(y, a) + \gamma d(z, a)$$
$$d(u, v) = \alpha d(x, y) + \beta d(x, z) + \gamma d(y, z)$$

where,

$$\alpha + \beta + \gamma = 1$$

By, default –

$$\alpha = \beta = \gamma = 1/3$$

Neighbor-Nets Clustering:

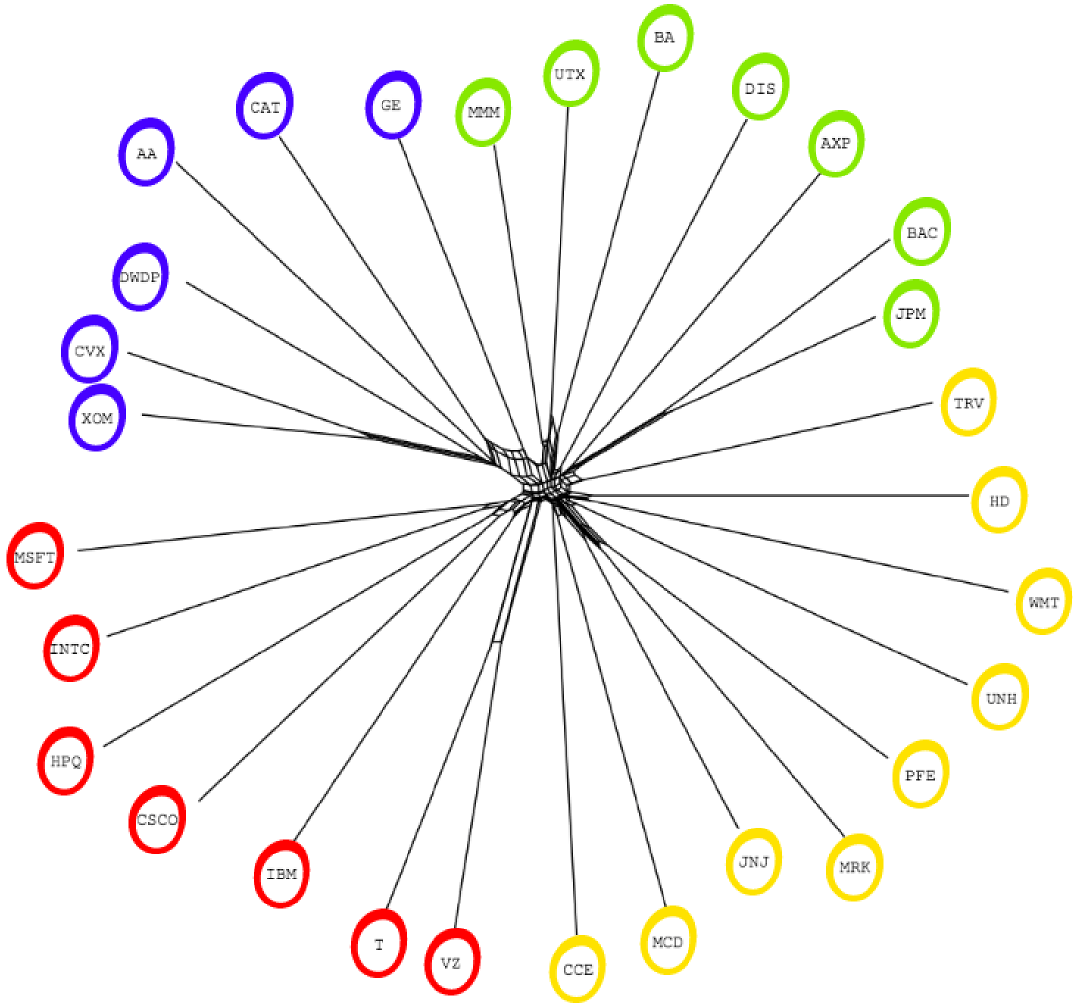
After using Neighbor-Nets Splits Graph on the distance matrix given by the ultra-metric on stock returns from period 1 and visually inspecting the network obtained (given in figure below), the clusters were given as –

Cluster 1: XOM, CVX, DWDP, AA, CAT, GE

Cluster 2: BA, UTX, MMM, DIS, AXP, BAC, JPM

Cluster 3: MSFT, INTC, HPQ, CSCO, IBM, T, VZ

Cluster 4: CCE, MCD, JNJ, MRK, PFE, UNH, WMT, HD, TRV



The circular ordering obtained by Neighbor-Nets ensures that cluster assignment error is minimized. For example, in the clusters obtained by our graph MMM is assigned to cluster ‘Green’ though it could have been easily assigned to cluster ‘Blue’ as the two closest neighbors to MMM are GE and UTX. Though, this would not result in a big change in the cluster assignments as only the cluster nodes that are at the edges of cluster are susceptible to change in cluster assignment.

Note: We constructed the neighbor nets using the method given in ‘phangorn’ Library of R but the splits were exported in nexus format to the SplitsTree4 package which we used to plot the network obtained as SplitsTree figures were a lot more aesthetically pleasing than the native plots obtained in R.

RESULTS

We compare the three clustering algorithms by simulating 1000 different portfolios from each of them. The comparison is done to show the robustness of the algorithms to the time varying nature of the return dynamics in stock returns. The results are presented below –

	Hierarchical Clustering Trees	**Minimum Spanning Trees**	**Neighbor-Nets Splits**
Mean	1.42	1.46	1.40
Std	0.096	0.095	0.096
Sharpe Ratio	0.05	0.056	0.046

CONCLUSION

We used three network based algorithms to cluster a group of stocks in the Dow Jones Industrial Average (DJIA) and used the clusters formed to construct a small private investor sized portfolio. The clusters formed were shown to be robust to time-varying characteristics of returns which leads to estimation errors in the Markowitz Mean Variance Portfolio Optimization framework.

The simulation results show that portfolio constructed on the clusters formed by minimum spanning trees (MST) outperformed those constructed on the clusters of HCT and Neighbor-Nets. This was mostly due to the high returns observed on the clustered stocks in the second period. This discrepancy can be attributed to the unbalanced cluster sizes in MST and results remain largely unexplained. Given a longer time period dataset and a larger universe of the stocks, these empirical findings can be extended to gather some patterns. Nevertheless, the study successfully shows that network based algorithms can be employed in passive portfolio management.

References:

Hannah Cheng Juan Zhan, William Rea, Alethea Rea; A Comparison of Three Network Portfolio Selection Methods -- Evidence from the Dow Jones, arXiv:1512.01905

David Bryant, Vincent Moulton; Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks, *Molecular Biology and Evolution*, Volume 21, Issue 2, 1 February 2004, Pages 255–265, https://doi.org/10.1093/molbev/msh018