Application of Clustering Methods to Trading Strategies in the US Equity Market

by

Yilang Lu (CID: 01407813)

Department of Mathematics
Imperial College London
London SW7 2AZ
United Kingdom

Thesis submitted as part of the requirements for the award of the MSc in Mathematics and Finance, Imperial College London, 2017-2018

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Signature and date:

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Pietro Siorpaes, Mr. Sherif Khalifa and Mr. Anthony Yong, for their engagement and contribution throughout the process of this project, as well as insightful suggestions. I'm grateful to Mr. Tang Chao for spending his time to proofread my thesis.

I dedicate this thesis to my parents and my girlfriend, for providing me with unfailing support and continuous encouragement throughout my year of study and the completion of researching and writing this thesis. This accomplishment would not have been possible without them.

Contents

1 Introduction

There are thousands of stocks in the worldwide equity market including dozens of industries varying from Banking to Software. In fact, Bloomberg designed a Bloomberg Industry Classification System (BICS) [1] to categorise stocks into different industries based on their general business activities. BICS contains 10 sectors which is the first level of the classification and each sector is further partitioned into industries while industries can be further divided into sub-industries. Therefore, BICS is a three-level hierarchical system of stocks: sector, industry and sub-industry where names are detailed in Table 12 attached in Appendix. Every stock is given a code, it is named as BICS code to identify its position. For example, Microsoft Corporation's code is '181211'. The first two digits '18' represents Technology sector which Microsoft belongs to, the third and fourth digits '12' represents Semiconductor industry where Microsoft belongs to, and '11' represents the sub-industry.

BICS codes are widely used when investing in stocks since plenty of valuable information is contained in BICS code. For instance, one can expect stocks in the same industries are highly correlated and thus it is appropriate to assume their stock prices will have similar movement [2], because firms in the same sector or industry have similar business activities or products so the fluctuation of their stock prices will be influencing each other. If stocks in a sector move together as one supposed, one can say that this sector has a *good clustering property*. However, an interesting phenomenon occurred during the research of BICS code. That is even though MDU[1] is in the Construction Materials industry, this stock has a close relationship with three stocks: NJR[2], IDA[3] and SR[4], which are in Utilities sector. Figure 1 presents annualised cumulative sums of daily returns of the four stocks during the timespan ranging from

One can see from Figure 1 the four stocks somehow have similar momentum, which means their stock prices move up or down very similarly. There are two reasons why MDU has a similar price fluctuation as the other three. The first one is influence of expected market return [3]. The second one is that all of their business is related to natural gas. Therefore, if someone assumes stocks in Construction Materials industry are moving together and trade them based on this assumption, he/she will lose profit since MDU is not consistent with other in Construction Materials industry in term of the stock price fluctuation. This example suggests that some stocks are divided into "wrong" sectors, and this may be because the way of the classification standard has been defined.

For a trading purpose, the stock market is supposed to be split into groups where stocks have a close relationship within the same group while stocks in different groups are more independent. To analyse this and reconstruct the stock model, one applies clustering algorithms. As a machine

[1] MDU is the ticker of MDU Resources Group Incorporation
[2] NJR is the ticker of New Jersey Resources Corporation
[3] IDA is the ticker of IDACORP Incorporation
[4] SR is the ticker of Spire Incorporation

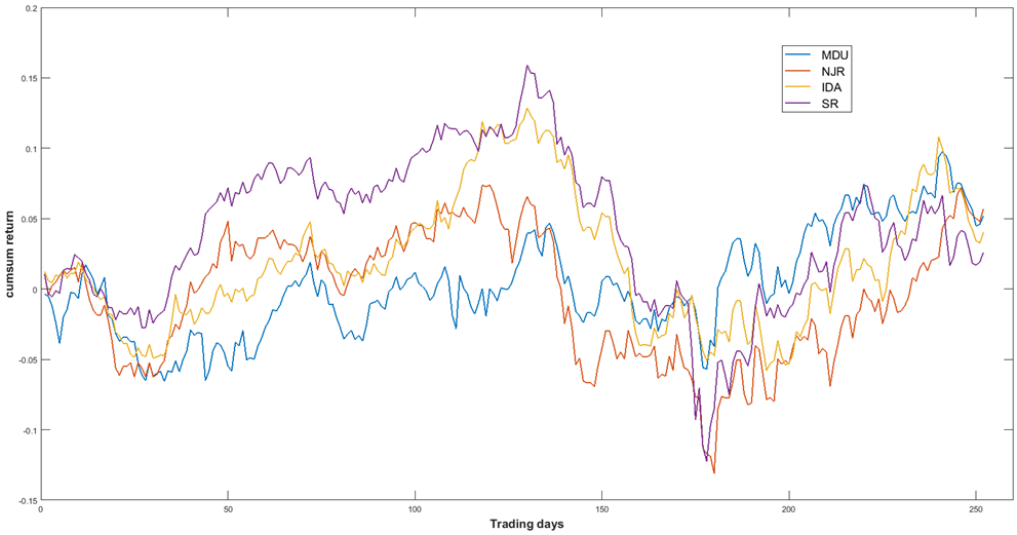


Figure 1: First discovery of BICS code with 4 time series of stock prices.

learning technique, the clustering algorithm can classify data points into groups under some constraints. Data points which are divided into the same groups should have some similar properties while data points in different groups have very different features [4].

Among stock markets in the worldwide, the US stock market has the greatest capitalisation and largest number of stocks. It is the most influential one too, hence this project will focus on the US stock market where many investors are very keen to understand its structure. In the following, distance functions including Pearson's Correlation, Spearman's Rank Correlation and Cosine measure will be stated in Section 2.1, then three clustering methods containing K-means clustering, Hierarchical Clustering and Planar Maximally Filtered Graph are introduced in the remaining of Section 2. After these, applications of three clustering methods to US stock market will be presented in Section 3.

2 Preliminaries

In this section, the concept of distances will be discussed and it will become the foundation of the model building. Then it will present 3 different models for modelling the stock universe and classify stocks into clusters.

2.1 Distance Introduction

Assume $X = \{x_1, x_2, ..., x_n\}$ is a non-empty set, and let distance $d : X \times X \to \mathbb{R}$ be a function where for all $x, y, z \in X$, the following four principles are: satisfied

(p1)	$d(x, y) \geq 0$	non-negativity,
(p2)	$d(x, y) = d(y, x)$	symmetry,
(p3)	$d(x, y) = 0$ iff $x = y$	identity,
(p4)	$d(x, y) \leq d(x, z) + d(y, z)$	triangle inequality.

The function d can be also referred as metric on X, and we denote the pair (X, d) as a metric space [5].

2.1.1 Pearson Correlation Distance

Correlation, a very well-known measure in Statistics, was developed by Karl Pearson in [6] from a related idea introduced by Francis Galton [7]. Here, the correlation is been used as a distance function to describe the similarity of two elements x_i and x_j in the set X. The correlation between x_i and x_j is defined as:

$$\rho_{ij} = \frac{\mathbb{E}[(x_i - \mu_{x_i})(x_j - \mu_{x_j})]}{\sigma_{x_i}\sigma_{x_j}}, \tag{2.1}$$

where μ_{x_i} is the mean of x_i and σ_{x_i} is the standard deviation of x_i, similarly with x_j. Especially, the estimating formula can be expressed as:

$$\hat{\rho}_{ij} = \frac{(x_i - \bar{x}_i)^T(x_j - \bar{x}_j)}{\|x_i - \bar{x}_i\| \cdot \|x_j - \bar{x}_j\|}. \tag{2.2}$$

The value of correlation is in the interval $[-1, 1]$, and it is closer to 1 if x_i and x_j are highly correlated, which suggests that the distance between x_i and x_j is small. By the meaning of distance, the further apart the two vectors are, the more different they are, but correlation is a measure of how similar two quantities are which negative value is undesirable.

To satisfy four principles of metric definition, the correlation distance on set X, is given as:

$$d_{ij} = \sqrt{2 * (1 - \hat{\rho}_{ij})}. \tag{2.3}$$

It is easy to check that value of correlation distance is in the interval $[0, 2] \in \mathbb{R}^+$. $d_{ij} = 0$ when x_i and x_j are positively and linearly correlated, $d_{ij} = 2$ when x_i and x_j are negatively and linearly correlated. This distance function is well defined as a distance by the definition of distance in 2.1. Clearly the formula (2.3) satisfies non-negativity, symmetry and identity. Let's prove the triangle inequality as it is not as straightforward as the previous 3 conditions.

Proof. Assume X, Y and Z are three variables, ρ_{XY}, ρ_{XZ} and ρ_{YZ} are pairwise correlations between them. What we need to prove is

$$\sqrt{2 * (1 - \rho_{XY})} + \sqrt{2 * (1 - \rho_{XZ})} > \sqrt{2 * (1 - \rho_{YZ})}.$$

Take square on both sides, we get:

$$\sqrt{2 * (1 - \rho_{XY})} * \sqrt{2 * (1 - \rho_{XZ})} > 2 * (1 - \rho_{YZ}) - 2 * (1 - \rho_{XY}) - 2 * (1 - \rho_{XZ}).$$

Take square again on both sides, and cancel the same terms on both sides:

$$4(1 - \rho_{XY})(1 - \rho_{YZ}) \geq (\rho_{XY} + \rho_{XZ} - \rho_{YZ} - 1)^2.$$

Open the brackets, and cancel the same terms, we get:

$$(1 - \rho_{XY})^2 + (1 - \rho_{XZ})^2 + (1 - \rho_{YZ})^2 \geq (\rho_{YZ} - \rho_{XY})^2 + (\rho_{XY} - \rho_{XZ})^2 + (\rho_{XZ} - \rho_{YZ})^2.$$

Looking at the first term on both sides, clearly that $1 - \rho_{XY}$ is greater than $\rho_{YZ} - \rho_{XY}$ since correlation is between -1 and 1 by definition, hence former square is still greater than the latter, the same as the remaining two terms. Therefore, the left hand side is greater than or equal to the right hand side, which implies that triangle inequality holds for this Person's Correlation Distance function. □

Indeed, this proof also holds for the following two distance functions, where the value of Spearman's Rank Correlation and the value of Cosine are also in the interval [-1,1].

2.1.2 Spearman's rank correlation

Spearman's rank correlation coefficient, named by Charles Spearman, is a non-parametric rank statistic, that is either distribution free or has a distribution with unspecified parameters, and it is proposed as a measure of the strength of the association between two variables [26]. As a measure of a monotonic relationship, Spearman's correlation is used when the distribution of data makes Pearson's correlation coefficient undesirable or misleading, since Pearson's correlation is parametric. The advantage of Spearman's correlation is that when we analyse a complex system, which has a unknown distribution or the variables are not linearly correlated, Spearman's correlation gives a good description of the relationships through a monotonic function.

Assume X and Y are two random variables, $x_1, x_2, ...x_n$ and $y_1, y_2, ...y_n$ are their respective samples of size n. Denote $rank(X) = a_1, a_2, ..., a_n$ and $rank(Y) = b_1, b_2, ..., b_n$ as the rank of samples, where a_i and b_i are corresponding x_i and y_i's rank, hence $rank(X)$ and $rank(Y)$ are two particular sequences of integer 1 to n. The covariance of $rank(X)$ and $rank(Y)$ is defined as:

$$\text{cov}(rank(X), rank(Y)) = \mathbb{E}[(rank(X) - \mathbb{E}[rank(X)])(rank(Y) - \mathbb{E}[rank(Y)])].$$

Specifically, the estimating formula of covariance is computed by samples:

$$\hat{\text{cov}}(rank(X), rank(Y)) = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b}),$$

where $\bar{a}$ and $\bar{b}$ are means of $rank(X)$ and $rank(Y)$ respectively. According to the proof included in Appendix A, the estimated covariance can be also expressed as:

$$\hat{\text{cov}}(rank(X), rank(Y)) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_j)(b_i - b_j), \tag{2.4}$$

The definition of Spearman's rank correlation is given by:

$$\rho_s = \frac{\text{cov}(rank(X), rank(Y))}{\sigma_{rank(X)} \cdot \sigma_{rank(Y)}}, \tag{2.5}$$

where $\sigma_{rank(X)}$ and $\sigma_{rank(X)}$ are standard deviations of $rank(X)$ and $rank(Y)$ respectively [27]. Since both $rank(X)$ and $rank(Y)$ are 1 to n integer sequence, whatever the order is, the standard deviation is fixed, that is $\frac{n^2-1}{12}$. Indeed, we can use the formula introduced above considering variance as a special case of covariance.

$$\begin{aligned}
\hat{\sigma}^2_{rank(X)} &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_j)^2 \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i^2 + a_j^2 - 2a_i a_j) \\
&= \frac{1}{2n^2}\left(2n\sum_{i=1}^{n}a_i^2\right) - \frac{1}{2n^2}\left(2\sum_{i=1}^{n}\sum_{j=1}^{n}a_i a_j\right)
\end{aligned}$$

Since for the sequence $(1, ..., n)$, the sum of square of elements and sum of elements can be determined, we have:

$$\begin{aligned}
\hat{\sigma}^2_{rank(X)} &= \frac{1}{2n^2}\left(2n \cdot \frac{n(n+1)(2n+1)}{6}\right) - \frac{1}{2n^2}\left(2\left(\sum_{i=1}^{n}a_i\right)^2\right) \\
&= \frac{(n+1)(2n+1)}{6} - \frac{1}{2n^2}\left(2 \cdot \frac{n^2(n+1)^2}{4}\right) \\
&= \frac{n^2-1}{12}.
\end{aligned}$$

This result is also the denominator of Spearman's Rank correlation. Then, to compute correlation, let's begin by considering $\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_j)(b_i - b_j)$:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_j)(b_i - b_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n}a_i b_i + \sum_{i=1}^{n}\sum_{j=1}^{n}a_j b_j - \sum_{i=1}^{n}\sum_{j=1}^{n}a_i b_j - \sum_{i=1}^{n}\sum_{j=1}^{n}a_j b_i \\ &= 2n\sum_{i=1}^{n}a_i b_i - 2\sum_{i=1}^{n}a_i \sum_{j=1}^{n}b_j \\ &= 2n\sum_{i=1}^{n}a_i b_i - \frac{1}{2}n^2(n+1)^2.\end{aligned}$$

Denote $d_i = a_i - b_i$, and consider $\sum_{i=1}^{n} d_i^2 = 2\sum_{i=1}^{n} a_i^2 - 2\sum_{i=1}^{n} a_i b_i$, substitute this into above equation, so that we have:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_j)(b_i - b_j) &= 2n\sum_{i=1}^{n}a_i^2 - n\sum_{i=1}^{n}d_i^2 - \frac{1}{2}n^2(n+1)^2 \\ &= \frac{1}{6}n^2(n^2-1) - n\sum_{i=1}^{n}d_i^2\end{aligned}$$

Hence, if we substitute the results above into Spearman's correlation definition, we obtain a formula which can be seen more frequently in the literature as Spearman rank correlation:

$$\hat{\rho}_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}. \tag{2.6}$$

From this definition, it is easy to see that the value of this correlation is between $[-1, 1]$, and it equals 1 when two sample sequences have the same order, that is $a_i = b_i$ for i from 1 to n, while it equals -1 when two sample sequences have the opposite order, that is $a_i = b_{n+1-i}$ for i from 1 to n. The following is an example of Spearman's rank correlation:

Table 1: An example of Spearman's rank correlation

Math	Rank 1	Physics	Rank 2	Difference	d_i^2
100	1	83	4	-3	9
96	2	79	5	-3	9
92	3	97	1	2	4
85	4	85	3	1	1
79	5	53	10	-5	25
77	6	75	6	0	0
73	7	71	7	0	0
67	8	91	2	6	36
61	9	68	8	1	1
53	10	61	9	1	1

The Table 1 shows 10 students' math and physics grades. Rank 1 is the rank of math while rank 2 is physics. Difference, represented as d_i, is the difference of everyone's rank difference. Hence, the Spearman' rank correlation is 0.4788, while Pearson's correlation is 0.4892.

2.1.3 Cosine function

Cosine function $cos(\theta)$ is a measure of similarity for two vector variables in an inner product space, it is a function of the angle θ of two variables. Intuitively speaking, for two n dimensional vectors, the smaller the angle is, the higher similarity they have. In the extreme case, the angle is 0 if they are exactly the same, and the angle is π (180°) if they have the same magnitude with opposite signs in every dimension. For the corresponding cosine value, cosine equals 1 when the angle is 0, cosine equals -1 when the angle is π and cosine equals 0 when the angle is $\pi/2$ (90°) which implies that two vectors are orthogonal. However, we would like to use distance to picture out the relationship between variables.

The definition of cosine of two vectors $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\} \in \mathbb{R}^{n\times 1}$ is:

$$\cos(\hat{\theta}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}}, \tag{2.7}$$

where θ is the angle of X and Y and the corresponding distance function is given as:

$$d_{XY} = \sqrt{2 * (1 - cos(\hat{\theta}))}. \tag{2.8}$$

It can be easily shown that the cosine distance function also satisfies the four principles of distance (since its value is between -1 and 1, triangle inequality also holds), hence it is a well defined distance.

2.2 K-means Clustering

K-means Clustering is a type of unsupervised learning of which the goal is to discover the underlying structure of input data. While supervised learning is looking for the mapping function between input data and the corresponding output data which is known, unsupervised learning is used to draw out the hidden structure of input data. Hence, the result is unknown until the learning is applied, and how to assess the result is an essential topic this report will be focusing on.

The name 'K-means' was firstly proposed by James MacQueen in 1967 [8], though the algorithm was put forward by Stuart Lloyd in 1957 [9]. As the name indicates, K-means clustering is to partition $N \in \mathbb{N}$ unassigned variables into $k \in \mathbb{N}$ assigned clusters according to a predefined distance function. Based on this function, the algorithm will iterate until every unassigned variable is assigned to a cluster.

Let $\mathbf{S} = \{s_1, s_2, ...s_n\}$ be a set of variables as the input of K-means clustering, then we denote d_{ij} as the distance between variable s_i and variable s_j. Let $c_1, c_2, ...c_k$ be centroids of k clusters,

which are initialised by randomly choosing from set S and then they are computed depend on the type of variables. The mapping function f is defined as:

$$f = \min \sum_{i=1}^{k} \sum_{j=1}^{n_i} d_{ji}, \tag{2.9}$$

where k is the number of categories, n_i is the number of variables belonging to i^{th} cluster, and total number of variables is $n = \sum_{i=1}^{k} n_i$, d_{ji} is the distance measure between variable s_j and its corresponding centroid c_i, let's call it distance between s_j and c_i.

This mapping function will minimise the sum of distance summation between centroids and variables within clusters. Based on this mapping function, the standard algorithm of K-means Clustering can be described in four steps:

Step 1: Randomly choose k variables from set S as initial centroids of clusters.

Step 2: Compute pairwise distances between variables and centroids and label variables to its nearest clusters.

Step 3: For every cluster, recompute its centroid and repeat step 2.

Step 4: Algorithm stops when centroids are not changed anymore.

Input unlabelled data set S and choose the type of distance function, the algorithm will output labels of data points which represent their clusters. This algorithm keeps iterating until the value of mapping function f increases by moving one variable arbitrarily from its original cluster to another cluster [10]. It is worth mentioning that the way to compute centroids of clusters depends on the type of variables. For instance, if variables are points in $\mathbb{R} \times \mathbb{R}$ surface, coordinates of a centroid is defined as the mean of coordinates of points in this cluster. If the variables are stocks, every variable is a series of daily return of stocks, a centroid is the mean of stock daily returns in its cluster.

Indeed, the algorithm can be seen as two steps. Firstly, treat c_i as fixed and minimise the value of f with respect to s_j. Secondly, hold s_j fixed and minimise the value of f with respect to c_i, and keep iterating until the value of f converges to a constant and this algorithm always converges. It always give a solution no matter what the input is, as every data point is assigned into a cluster and if we move arbitrary one data point into another cluster, the value of f will increase.

Note that the result of K-means Clustering has non-convex property, to prove this, let's rewrite formula (2.9) as follow:

$$f(W, C) = \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} d(s_j, c_i), \tag{2.10}$$

subject to:

$$\sum_{k}^{i=1} w_{ij} = 1, \text{ for } j = 1, 2, \ldots, m$$

$$w_{ij} = 0 \text{ or } 1 \text{ for } j = 1, 2, \ldots, m \text{ and } i = 1, 2, \ldots, k,$$

where $W = [w_{ij}]$ is a $k \times m$ matrix of variables assignments. $w_{ij} = 1$ implies that variable s_j is assigned to cluster i while $w_{ij} = 0$ means not, hence the sum of every column of W is 1. $C = (c_1, c_2, \ldots, c_k)$ is a column vector of centroids. Then let's consider the reduced function of f, F, with fixed centroids C:

$$F(W) = \min\{f(W, C) : C \ fixed\}. \tag{2.11}$$

$F(W)$ is a non-convex function, since if we assume W_1 and W_2 are arbitrary two points satisfied constraints and α is a parameter in [0,1], then for a certain C we have:

$$\begin{aligned} F(\alpha W_1 + (1-\alpha)W_2) &= \min\{f(\alpha W_1 + (1-\alpha)W_2, C)\} \\ &= \min\{\alpha f(W_1, C) + (1-\alpha) f(W_2, C)\} \\ &\geq \alpha \min\{f(W_1, C)\} + (1-\alpha)\min\{f(W_2, C)\} \\ &= \alpha F(W_1) + (1-\alpha)F(W_2). \end{aligned}$$

This inequality holds because $f(W, C)$ is not only positive but also linear with respect to W. In fact, it can be proved that minimising $f(W, C)$ is equivalent to minimising $F(W)$ [28] which illustrates that minimising f is a non-convex problem. Figure 2 points out that convex only has one minimum point while non-convex can have several minimum points where algorithm will stop.

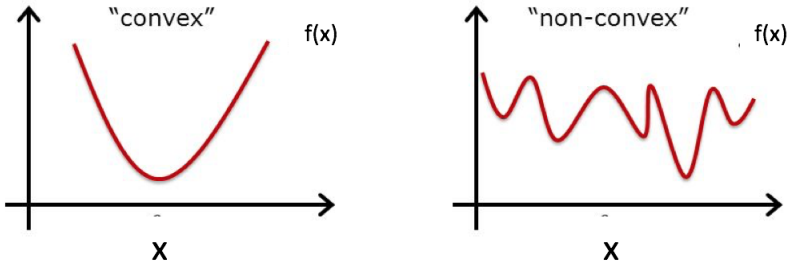


Figure 2: Convex and non-convex

The last thing worth to pay attention with K-means Clustering is the stability of results. Result of this optimisation problem is not only a local optimal solution but also not determined. It is fair to say that every time algorithm were ran with the same data set, output is different. Non-deterministic of results is due to the initialisation of centroids as they are randomly chosen from data set. Since optimising function f is a non-convex problem, the problem can have many local

optimal solutions. Algorithm will stop when a local optimal solution is found and different initial value of centroid vector C may leads the algorithm stops at different local optimal solution.

2.3 Hierarchical Agglomerative Clustering

Hierarchical Clustering is a widely-used method when dealing with complex systems, for example, when elements of a complex system can be divided into N clusters and these clusters can be further divided into M sub-clusters ($M > N$), until a level when the number of clusters is required [18]. The data which contains 724 US stocks, it can be initially partitioned into 10 sectors, and then can be partitioned into 48 industries, 166 sub-industrials and 724 stocks. Hence, the sample data is a complex system with 4 levels. Based on this hierarchical property, Hierarchical Agglomerative Clustering (HAC) will be applied. It is a method that merges the clusters into bigger cluster and in the next section it will be following Zura Kakushadze's idea to introduce Hierarchical Agglomerative clustering [2].

Assume X is a data set of a complex system, which contains n elements $\{x_1, x_2, \ldots, x_n\}$, and the target is to cluster elements into K clusters. The clustering algorithm is:

Step 1: Define distance between clusters, distance between elements and distance between element and cluster.

Step 2: Compute distance between elements pair-wisely, find the pair which have the smallest distance and merge them into a new cluster.

Step 3: Compute distance between clusters and distance between elements and clusters, then merge the pair with the smallest distance. Keep merging until the number of clusters reduces to K.

Note that there are three definitions of distance mentioned above, distance between clusters, distance between elements and distance between element and cluster. If we consider elements as singleton clusters, all the distance can be seen as distance between clusters. Theoretically, the result of agglomerative clustering could not be determined if some different element or cluster pairs have exactly same distance. However, in practice, this would not happen. Once the pairwise distances are determined, the result of hierarchical clustering is fixed, compare to K-means clustering where it depends on the randomly chosen initial points. An example will be given in in Figure 3 later to illustrate this.

2.3.1 Measure Introduction

Defining distance is the most important process in Hierarchical Clustering, different definitions will lead to different results. In practice, one need to define the distance in a proper way. The definition needs to be meaningful and well-presented for the characteristic of the data.

The next sections will introduce some frequently-used distances or metrics such as: Single, Complete and Hausdorff. They are defined as distance between clusters which corresponding to Single linkage, Complete linkage and Hausdorff linkage.

2.3.2 Single Metric

For two non-empty subsets A and B which belong to a complex or large set $\subseteq \mathcal{S}$, δ is a predefined distance of elements, the Single distance is defined as:

$$d_S(A, B) = \inf_{a \in A, b \in B} \delta(a, b). \tag{2.12}$$

Single distance is the minimum of the distances between all pairs of elements randomly chosen from A and B, it is a sub-dominant ultra-metric distance [24], which means that even though the function d_S does not satisfy the triangle inequality it can be treated as a pseudo-distance function hence can be applied. However, let's call it Single distance for convenience.

Single Linkage Clustering, which is a method came up by Sneath and Sokal in 1973 [23], it is based on Single distance associated with a metric between elements. The procedure of Single Linkage is been used in many clustering methods, it can extract a Hierarchical Model which is Minimum Spanning Tree (MST) for the metric matrix given by the predefined metric.

MST from Graph Theory, is a subgraph of a weighted graph that connects all the nodes together, it does not contain any cycles and minimises the total distance within nodes [25, Chapter 5, page 66]. The algorithm of Single Linkage is:

> Step 1: Consider N elements as N singular clusters. Construct a list of pairwise distance between clusters and sort the list in ascending order.
>
> Step 2: Start from the top of the list, merge the corresponding two clusters, then recompute the list. Repeat until all the elements are connected, that is the number of left cluster is 1, or the number of edges of the graph is $N - 1$ equivalently.
>
> Step 3: Display result through a dendrogram(section 3.3).

We can decide how many clusters the set will be partitioned into at the beginning, since the number of clusters reduces by one every time when two clusters merge together. Hence we can give a threshold (the final number of clusters) to the algorithm first and stop the algorithm when this threshold is reached. In particular, if we merge all elements together into one cluster, the result of Single Linkage is a MST.

Theoretically, the MST we get from Single Linkage could not be deterministic. Figure 3 gives an example, according to our Single Linkage algorithm, node D and E will be linked firstly, then B and C is the next pair. The distance between cluster $\{B, C\}$ and $\{D, E\}$ is 0.3, so this two cluster will be merged at third iteration. Finally, the distance between A and cluster $\{B, C, D, E\}$ is 0.4.

	A	B	C	D	E
A	0	0.6	0.4	0.4	0.5
B	0.6	0	0.2	0.7	0.8
C	0.4	0.2	0	0.3	0.7
D	0.4	0.7	0.3	0	0.1
E	0.5	0.8	0.7	0.1	0

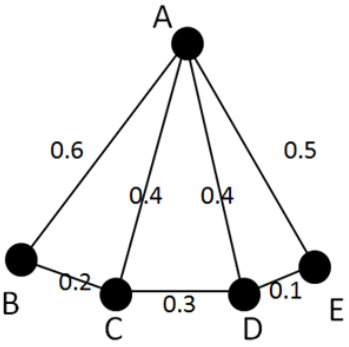


Figure 3: An example of non-unique MST

However there are two cases, that is linking A and C or A and D. Both cases give a MST, whereas this two MST are different. In other word, the result is not unique. Fortunately, this situation almost never happen in practice as distances between different pairs are never identical.

2.3.3 Complete Metric

Complete Linkage is very similar to Single Linkage. It is just another way to define the distance between cluster, named Complete distance. For two non-empty subsets A and $B \subseteq \mathcal{S}$, a complex or large set, and δ is a predefined distance of elements, the complete distance d_S is defined as:

$$d_S(A, B) = \sup_{a \in A, b \in B} \delta(a, b). \tag{2.13}$$

Complete distance is the maximum of the distances between all pairs of elements random chosen from A and B, similar to Single distance, Complete distance is also only a sub-dominant ultra-metric distance [24] because it does not satisfy the identity principle. However, same as Single distance, we call it Complete distance for convenience.

2.3.4 Hausdorff Metric

Hausdorff distance is named after F. Hausdorff [21] and developed by Nicolas Basalto [19], it is introduced in Topology as a derivative definition of Hausdorff Space [20]. Given a metric space $(\mathcal{F}, \delta)$, where δ is the predefined metric, the distance between an element $a \in \mathcal{F}$ and a non-empty subset $B \subseteq \mathcal{F}$ is described as:

$$d(a, B) = \inf_{b \in B} \delta(a, b). \tag{2.14}$$

Based on this, let's define a function between a non-empty subset $A \subseteq \mathcal{F}$ and a non-empty subset $B \subseteq \mathcal{F}$:

$$d(A, B) = \sup_{a \in A} d(a, B) = \sup_{a \in A} \inf_{b \in B} \delta(a, b). \tag{2.15}$$

This function is not symmetric, that is $d(A, B) \neq d(B, A)$, hence it cannot be the distance between A and B. To satisfy the four constraints 2.1, the Hausdorff Distance is defined as the maximum

of two functions $d(A, B$ and $d(B, A$:

$$d_H(A,B) = \max\{d(A,B), d(B,A)\} = \max\left\{ \sup_{a\in A} \inf_{b\in B} \delta(a,b), \sup_{b\in B} \inf_{a\in A} \delta(a,b) \right\}, \tag{2.16}$$

which is a well defined distance.

By comparing the three distances above, it can be easily found that for arbitrary non-empty subsets A and $B \subseteq \mathcal{F}$, we have:

$$d_S(A,B) \leq d_H(A,B) \leq d_C(A,B). \tag{2.17}$$

Indeed, d_S underestimates the distance between two given sets while d_C overestimates it and this implies important consequences when we cluster a complex system [22]. Both Single and Complete distances are extreme situations of distance, and they are not well defined as distance.

2.4 Planar Maximally Filtered Graph

The method Planar Maximally Filtered Graph (PMFG) is introduced by Tumminello et al. [13]. Since PMFG is related to graphs, let's state some basic definitions from Graph Theory before introducing PMFG.

In the context of Graph Theory, a graph is a mathematical structure which is used to model relationships between subjects. As showed in Figure 4, graph G consists of *vertices* (points) which are linked by *edges* (lines). A *subgraph* of graph G is a graph formed from a subset of vertices and edges of G, such as the red component of graph G.

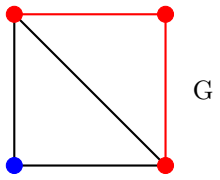


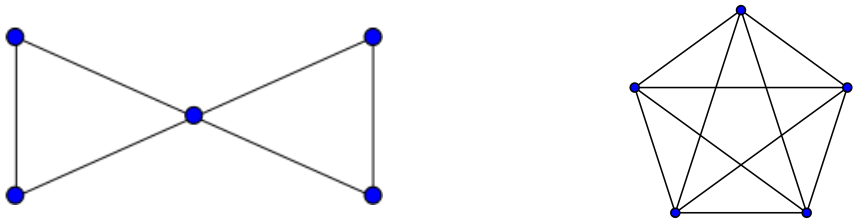
Figure 4: Example of graph G, red part is the subgraph of G.

Furthermore, a *planar graph* is a graph that can be embedded into a plane [17, Chapter 1, page 5], so there is way to draw this graph in the plane such that edges will only intersect at vertices. For instance, for the following two graphs in Figure 5, the graph on the left is a planar graph, while another is not since edges intersect with others in the middle of the figure. Besides, in the set of planar graphs, edges are needed as many as possible since the more edges the graph includes, the more information the graph has. There is a kind of special planar graph named as *Maximal Planar Graph*, which is a planar graph that has as many edges as possible. That is, if we add one more edge to the graph, the planar property will be no longer satisfied.

One important property of Maximal Planar Graph is that the quantity of edges is fixed. In fact, the number of edges of Maximal Planar Graph is $3(n-2)$, where n is the number of vertices.

This property comes from the corollary based on Euler's Formula, which indicates that the number of edges of a planar graph is small or equal to $3(n-2)$ [17, Corollary 13.3, Chapter 5, page 67].

A graph is called *complete* if every vertex of this graph links with all other vertices of this graph, that is equal to say the graph has $\frac{n(n-1)}{2}$ edges. Then if a subgraph is complete, we call it *clique*. For instance, 3-clique means a subgraph with 3 pairwise linked vertices which indeed is a triangle, and 5-Clique is the figure 5 showed below on the right hand side.



Planar graph

Non-planar graph

Figure 5: Example of planar and non-planar graph.

Based on the above definitions and properties, let's now explain the definition of PMFG and its idea. PMFG is a Maximal Planar Graph with n vertices linked by $3(n-2)$ edges. However, since a graph with n vertices can have at most $\frac{n(n-1)}{2}$ edges, not every pair of vertices will be linked when $n > 5$ and thus we have to remove some edges. What has to be done is to assess how strongly the vertices relate to each other by a predefined distance function and give a priority to the pairs having a strong correlation. Hence, for a complex system $S = \{s_1, s_2, \ldots, s_n\}$, let the vertices represent variables s_i which will be linked by edges if they have strong correlations and describe the construction of this system by a PMFG [16].

According to the definitions and the properties above, the constructing algorithm is straightforward.

- Step 1: Given a distance function and compute all pairwise distance between variables.
- Step 2: Sort the distance (in descending order), thus has an ordered sequence $\left\{c_1, c_2, \ldots, c_{\frac{n(n-1)}{2}}\right\}$, where n is the number of variables and $c_i > c_j$ if $i > j$.
- Step 3: Start from c_1 to end, connect the corresponding two vertices if the planar graph property is not violated. Keeping adding the edges until graph has $3(n-2)$ edges.
- Step 4: Partition PMFG into 3-cliques and 4-cliques.

Now, let's state some benefits of using PMFG for analysing a complex system. Firstly, as mentioned above, the number of edges of PMFG is $3(n-2)$, where n is the number of vertices. In a PMFG, Edges are used to convey underlying information of system to us, thus the more edges

the graph contains, the better understanding we have to the system. As it has been mentioned in the last subsection 2.3, the number of edges of MST is $n-1$, PMFG definitely contains more information than MST because it has roughly twice more edges than MST.

Secondly, cliques and loops are allowed in PMFG. Compared with MST, which does not allow cliques or loops due to its definition, PMFG is more flexible and has more edges.

Note that PMFG consists of 3-cliques and 4-cliques. Before explaining this property, let me introduce Kuratowski's theorem [17, Theorem 12.2, Chapter 5, page 62].

Theorem 2.1 (Kuratowski's theorem)**.** *A graph is a planar graph if and only if it does not contain K_5 or $K_{3,3}$ as subgraphs, where K_5 is complete graph with 5 vertices and $K_{3,3}$ is complete bipartite.*

A *bipartite* graph is the graph in which vertices can be divided into two disjoint sets such that two vertices within the same set are not allowed to be linked.

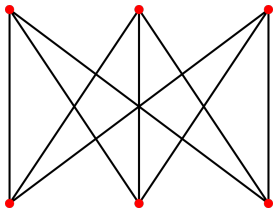


Figure 6: $K_{3,3}$ - A complete bipartite graph

This theorem is sufficient to show that it is impossible to contain a n-clique where $n >= 5$ in a PMFG, since PMFG is a planar graph and 4-cliques are the maximum sized cliques which are also planar. Hence, in theory, only 3-cliques and 4-cliques can exist in a PMFG. According to Tummomello's [13] description, PMFG only consists of 3-cliques and 4-cliques. Every edge of PMFG must be an edge of a triangle and there is no singular vertex or segment occurring in PMFG [13]. This property suggests that we can divide out complex system into smaller groups which only contain 3 or 4 members, and in such groups, variables are strongly correlated.

The last important property of PMFG is that it includes the basic hierarchical construction of MST [15]. From this aspect of Graph Theory, MST is a subgraph of PMFG. As both MST and PMFG are based on correlation, and as they have similar construction rules, their structures are similar. But PMFG has more relaxed constraints than MST, so it will have a more complicated structure. Indeed, every edge of MST is a *bridge*, that is, if we remove this edge, the graph will partition into two disconnected subgraph. Tummomello used mathematical induction to prove that PMFG includes all the bridges in its graph so that it must contain MST [13].

Indeed, PMFG is only the first extension of MST. Aste [16] argued that with the respect of MST, a graph storing more information can not only be constructed by linking the most strongly connected vertices iteratively but also be embedded on a surface of a given genus $g = k$. Genus is the largest number of nonisotopic simple closed curves that can be drawn on the surface without

separating it [14]. Such a surface can at most include $3(n-2+2k)$ edges, where n is the number of vertices. However, the improvement of information stored in the graph mostly depends on n when $n \gg k$. Thus, choosing the number of genus to be equal to 0 can reduce the complexity of the graph with slight information loss and in this case, the embedded graph is a planar graph in the plane.

3 Applying Clustering Methods to US Stock Market

3.1 Data Introduction

This section will introduce the data that will be used with the methods introduced in the previous sections. The data set consists of 724 stocks from US stock market, including 10 sectors and 48 industries, they are represented by the first and second levels of BICS code. Details are showed in Table 2. Besides, the timespan of the data set ranges from 2^{nd} Jun 2017 to 21^{st} May 2018.

Table 2: BICS Classification of the dataset

Sector Name	short BICS code	Number of Stocks	Number of Industries
Communications	10	23	2
Consumer Discretionary	11	107	11
Consumer Staples	12	39	3
Energy	13	45	2
Financials	14	165	6
Health Care	15	78	3
Industrials	16	101	9
Materials	17	44	6
Technology	18	92	5
Utilities	19	30	1

1. Short BICS code is the first two numbers of BICS codes which represents the sector which stock belongs to.
2. Number of stocks is the number of stocks contained in the sector.
3. Number of idustries is the number of sub-sectors in the sector.

Here is the formula for the *daily return*:

$$r_{i+1} = \frac{p_{i+1}}{p_i} - 1, \tag{3.1}$$

where r_{i+1} is $(i+1)^{th}$ daily return, and p_i is i^{th} close price of the stock. Daily return is an important and useful piece of information in stock analysis, it can be used to compute the correlation between two stocks. Note that, the weekly return is simply the sum of 5 consecutive daily returns in a week, that is the difference between the close price of Monday and the close price of next Monday .

3.2 Result of K-means Clustering

Pearson's correlation of two stocks is straight forward to define and compute. Assume that $\{r_{A1}, r_{A2}, ..., r_{An}\}$ and $\{r_{B1}, r_{B2}, ..., r_{Bn}\}$ are n-day returns of stock A and stock B in a certain

period. The correlation of A and B can be easily computed by using (2.2), and thus the distance follows equation (2.3). Hence, the K-means clustering can be applied to stock market by using this correlation based distance. Stocks are divided into 10 clusters, each of them consists of stocks from 10 BICS code sectors. The reason for choosing the 10 clusters for K-means Clustering is because the BICS code have 10 sectors.

Let's call the result of K-means Clustering as *clusters* to distinguish with sectors which is the name of first level BICS code. Table 3 shows the result of K-means Clustering, where the n^{th} row shows components of the n^{th} cluster.

Table 3: Result of K-means Clustering

Num	Com	C-D	C-S	Ene	Fin	Hea	Ind	Mat	Tec	Uti
111	0.9%	14.4%	4.5%	0.9%	11.7%	6.3%	39.6%	15.3%	6.3%	0%
107	0.9%	22.4%	2.8%	0%	3.7%	9.3%	37.4%	10.3%	13.1%	0%
52	1.9%	1.9%	1.9%	0%	0%	84.6%	1.9%	0%	7.7%	0%
52	0%	0%	7.7%	0%	90.4%	0%	0%	0%	0%	1.9%
91	4.4%	8.8%	3.3%	1.1%	2.2%	15.4%	4.4%	1.1%	59.3%	0%
77	15.6%	54.5%	26.0%	0%	2.6%	0%	1.3%	0%	0%	0%
89	0%	0%	1.1%	0%	97.8%	0%	1.1%	0%	0%	0%
49	2.0%	0%	0%	87.8%	0%	0%	6.1%	0%	0%	4.1%
33	3.0%	0%	0%	0%	9.1%	0%	0%	3.0%	3.0%	81.8%
63	3.2%	25.4%	3.2%	0%	11.1%	4.8%	11.1%	22.2%	19.0%	0%

1. The top line contains ten abbreviations of BICS first level sectors' names. Check Table 2 for their full names.
2. First column on the left is the quantity of stocks in each clusters.
3. Every row has 10 Percentage values, which represent percentages of stocks from BICS codes.

Every cluster consists of stocks from at least 3 sectors, and some of them have a significant component (>80%) from one sector, such as cluster number 3, 4, 7, 8, and 9. Red numbers in Table 3 are those significant components in high percentage, which suggests that not only their represented sectors are the main components of these clusters, but also implies that these sectors are very strongly correlated. Specifically, Health Care, Financials, Energy and Utilities have good clustering property compared with other sectors.

For instance, the cluster 4 has 52 stocks, in which 47 (90.4%) stocks come from Financials sectors, the rest consists of 4 stocks from Consumer Staples and 1 stock from Utilities. This cluster can be seen as a modified Financials sector or a inherited cluster of Financials sectors.

However, one interesting discovery is that similar to the cluster 4, the main component of cluster

7 comes from Financials sectors. It includes 87 (97.8%) stocks, 1 (1.1%) Consumer Staples stock and 1 (1.1%) Industrials stock. By checking the second level BICS codes of two clusters, one would see that Financials stocks in clusters 4 all come from REIT[1], which is a sub-sector of Financials sector (details in BICS code Table 12), on the other hand Financials stocks in cluster 7 consists of stocks from Asset Management, Banking, Institutional Financial Services and Insurance, which are another four industries of Financials sector. This means that the K-means algorithm partitioned Financials sector into two components. Details of these two clusters are presented in Table 4:

Table 4: Details of cluter 4 and 7

Cluster	Industry Name (BICS code)	Total number of stocks	Number of stocks in cluster
4	REIT (1415)	55	47
	Consumer Products (1210)	25	4
	Utilities (1910)	30	1
7	Consumer Products (1210)	25	1
	Asset Management (1410)	11	4
	Banking (1411)	56	55
	Institutional Financial Services (1412)	14	5
	Insurance (1413)	8	6
	Real Estate Oper&Services (1414)	21	17
	Waste&Envrnmt Srvc Equip&Fac (1618)	8	1

To assess the result of K-means Clustering, there are two indices that needs to be introduced first which is related to correlations and dependence. The ideal situation for trading strategy is that stocks in the same sector are strongly correlated (correlation is close to 1) while stocks in different sectors are weakly correlated or even independent (correlation close to 0). If so, taking or removing positions in sector 1 will not influence the price of sector 2. Though two independent sectors could never be found, it is worth to reduce the influence between two sectors when applying the method. To assess this ideal situation, *inside correlation* and *dependence* are defined to evaluate the strength of correlation within sectors and dependence between sectors.

Assume $A = \{a_1, a_2, ..., a_n\}$ and $B = \{b_1, b_2, ..., b_m\}$ are two sectors containing n stocks and m stocks respectively, each a_i and b_j represent a series of daily returns. The *inside correlation* is the mean of pairwise Pearson's correlations of stocks within the sector. For instance, the inside correlation of sector A is:

$$R_A = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \rho_{ij}, \tag{3.2}$$

[1]Real-estate Invesment Trust

where ρ_{ij} is the Pearson's correlation between stock a_i and a_j.

Dependence between two sectors are defined as the mean of pairwise Pearson's correlations of two stocks from different sectors. For instance, dependence between sector A and sector B is:

$$D_{AB} = \frac{1}{nm} \sum_{1 \le i \le n} \sum_{1 \le j \le m} \rho_{ij}, \tag{3.3}$$

where ρ_{ij} is the Pearson's correlation between stock a_i and b_j.
Inside correlation and dependence can be used to assess the result of K-means Clustering. The most desirable result is that the inside correlation for K-means is higher than the inside correlation of BICS code. By contrast, the value of dependence of the K-means Clustering is as small as possible. Table 5 listed ten inside correlations of K-means result and BICS code.

Table 5: Inside Correlation of K-means and BICS

Clusters	1	2	3	4	5	6	7	8	9	10	Mean
K-means	0.36	0.25	0.17	0.45	0.32	0.24	0.52	0.38	0.44	0.2	0.3317
Sectors	Com	C-D	C-S	Ene	Fin	Hea	Ind	Mat	Tec	Uti	Mean
BICS	0.19	0.21	0.23	0.41	0.33	0.19	0.31	0.28	0.28	0.45	0.2839

The red numbers in Table 5 are the weighted mean of previous ten inside correlations, that is giving the weights to clusters according to quantities of stocks of clusters. This suggests that K-means Clustering efficiently improves the inside relationship of a cluster. Specifically, we can see clusters 4, 7, 8 and 9 have the top 4 inside correlations of K-means result, which corresponds to big percentage values in Table 3. However, clusters 3 and 10 contain stocks from many sectors, they have lower inside correlation than BICS code. These two discoveries implied that K-means clustering method has strengthen the inside correlation of clusters which has a good clustering property. But unfortunately, it weakened the inside correlation of the sector which doesn't have good clustering property. For example, in Financials sector, the inside correlation is 0.33, it is inherited by cluster 4 and cluster 7 where the inside correlations are 0.45 and 0.52. On the other hand, stocks of Consumer Staples sectors are partitioned into many smaller groups which belong to different clusters.

Dependence matrices of K-means and BICS code are attached in the Appendix A. Since dependence matrix is symmetric, let's consider the upper triangular part only, and the mean of these pairwise dependence of BICS code in Table 14 is 0.1504 while the mean of pairwise dependence of K-means in Table 13 is 0.1476. Dependence of these two tables indicated that K-means Clustering method does not reduce the dependence of this stock classification system. If we list these pairwise dependence and compute the variance of them, one can see that the variance of K-means is 0.0623, which is much higher than the variance of BICS code, which is 0.0479.

Similar dependence values with high difference in variance illustrate the previous point. K-means Clustering method optimises the sector which has strong inside correlation, removes stocks which have weak correlation with other stocks in the same sector and adds stocks which are more correlated with this sector than the original sector they belong to.

The algorithm works well only with sectors which have good clustering property. In fact, sectors are unions of industries, and industries may be independent with each other even they belong to the same sector. For example, REIT industry in Financials sector, is less correlated with Banking and Asset Management.

Considering the stocks which have more solid correlation with other sectors than the sector they belong to. In the algorithm, stocks will be classified into a cluster which is the closest one to them. However, some stocks has no close relationship with any other stocks. For instance, In Figure 7, the blue and green stars can be seen as two sectors, and black points are the centroids of the blue and green stars respectively. The green star in the middle of Figure will be classified to the blue cluster according to K-means Clustering algorithm, since the distance between the green star and blue stars' centroid is shorter than the distance between the green star and green stars' centroid. Although the green star in the middle of figure is assigned into blue cluster, it is far away from any other star. It is assigned only because the algorithm will give every element a label.

Figure 7: Exampe of isolated stock

The same example can also be found in the stock market, some stocks are not only uncorrelated to the stocks within their sector but also have a weaker relation with stocks in other sectors. Let's call them as *isolated stocks*. There are two reasons why one stock is isolated. First reason is that this stock is uncorrelated with all others, for example, stock CMLSQ[1], which bankrupted in November 2017, still exists in the stock market until 4^{th} June 2018. Since the data set ends at 21^{st} May 2018, this stock is included. Second reason is that the data set is a subset of the whole US stock universe, hence stocks which are correlated with the isolated stock may not be included in this data set.

The clusters 4 and 7 are containing over 90% of stocks from Financials sector, hence it is more

[1] $CMLSQ$ is the ticker of Cumulus Media Incorporation

suitable to increase the number of cluster from 10 to 11 because the Financials sector may take over the space for other sector to be classified into a cluster. On the other hand, some industries should be considered together rather than separately, for example, Banking and Asset Management. How to choose a optimal value for K? The number of clusters is an essential question one needs to answer when applying K-means Clustering method.

K-means results suggest that K is bigger than 10 (number of sectors), but smaller than 48 (number of industries). In general, there is no robust method to determine the optimal value of K, however, Andrea Trevino introduced an accurate estimate method for K-means Clustering [12]. The mean distance between stocks and their belonging centroids is a common metric used to evaluate clustering results with a range of K value. The reason is when we increase the number of centroids, the mean distance will reduce. In an extreme case when the mean distance is zero when the number of centroids equals the number of stocks. From Andrea's methodology, with K increasing in the range [1,100] for the first step, the mean correlation distance for every K value can be computed. Figure 8 shows the result.

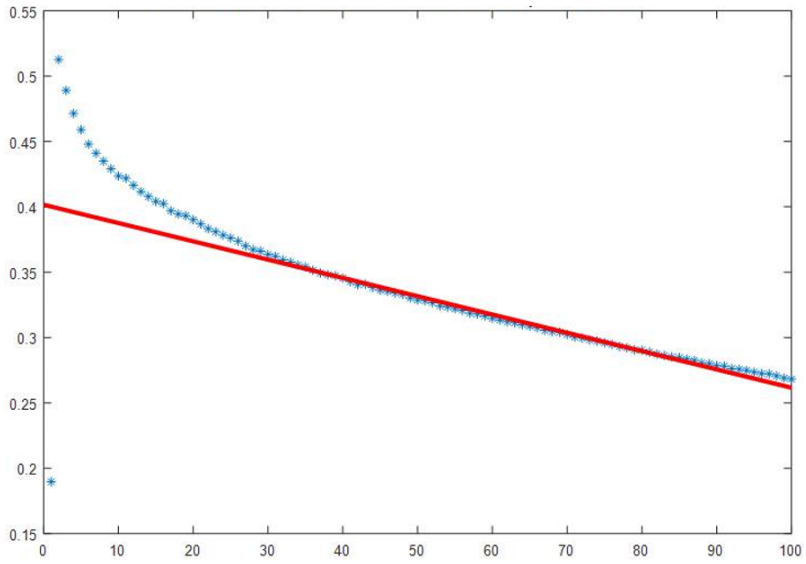


Figure 8: Relation of number of clusters and mean distance

Blue stars are the mean distances corresponding to different K values and the red line indicates the convergence of the mean distances. Straightforwardly, mean distance keeps decreasing when number of clusters increases. However, we can see that the rate of decrease is getting lower, this suggests that the number of clusters K affects the mean distance significantly at small number of clusters, and the influence reduces with a constant rate when K is big enough. Moreover, when K is between 35 and 80, the curve and the red line roughly coincide, this indicates that the influence

of K to mean distance is approximately constant. Hence, a number between 30 and 35 can be an appropriate number of clusters, this range has narrowed down even further with the initial guess $(10,\ 48) \in \mathbb{N}$.

3.3 Result of Hierarchical Clustering

In this section, the distance functions considered are Pearson's correlation, Spearman's Rank Correlation and Cosine. The distances between clusters, Single, Complete and Hausdorff, are referred to as *metrics*.

Let's consider Pearson's correlation as distance funcion between stocks and Complete distance as metric between two stock sets. Assume $r_i = \{r_{i1}, r_{i2}, \ldots, r_{im}\}$ is the daily return sequence of stock i, and r_{ik} is the return of k^{th} day, where $k = 1, 2, \ldots, m$. The correlation between two arbitrary stocks i and j is computed by equation (2.2) and therefore, distance between stock i and j is derivatived according to equation (2.3). The Complete distance between stock sets A and B is defined as the biggest distance among stock pairs in which one comes from set A and another one comes from set B. Now, the Pearson-Complete result as an example to introduce dendrogram, a tree diagram which perfectly illustrates the result of Hierarchical Clustering.

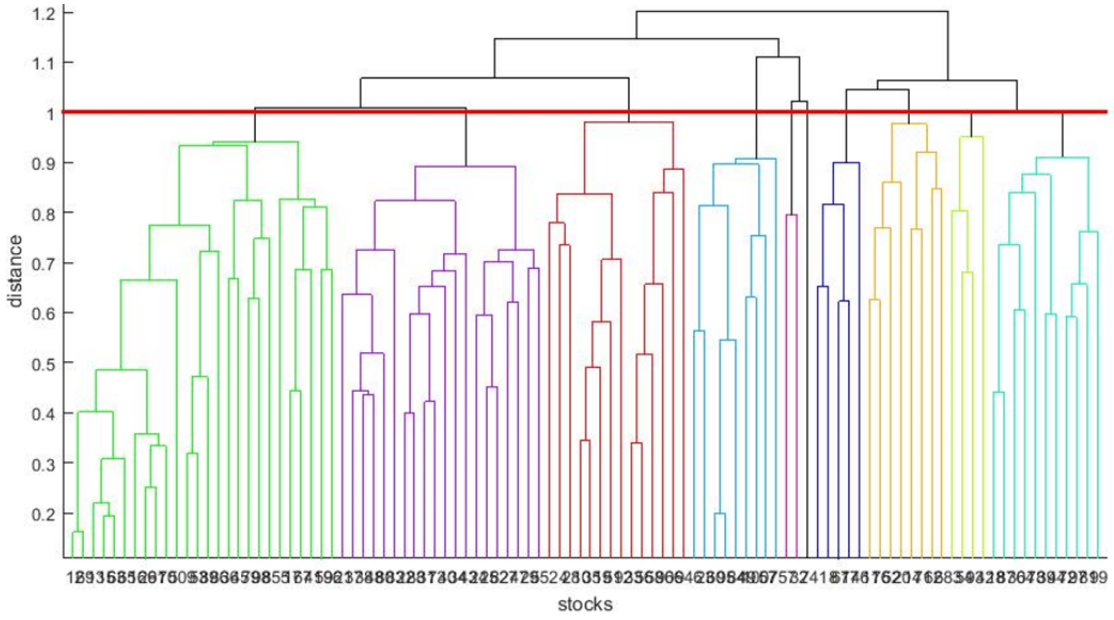


Figure 9: Dendrogram with Pearson's correlation and Complete distance

The Figure 9 is a dendrogram which clearly describes the hierarchical property of stock system. The algorithm of Hierarchical Clustering method keeps merging clusters until there is only one cluster left and this process perfectly expressed in the dendrogram. In Figure 9, every vertical line begins from the bottom X-axis represents a stock and top ends of two vertical lines will be linked if algorithm decides to put them into a cluster. Similarly, every horizontal line represents a stock

cluster and a new vertical line beginning from the middle of a horizontal line will be linked with another vertical line if algorithm decides to merge two clusters or a cluster and a stock. What's more, each colour represents one cluster and Y coordinates of horizontal lines indicate distances between two clusters. For instance, the red cluster consists of two sub-clusters and their distance is about 0.98 since the horizontal line is a little below than 1.

When a dendrogram is drawn, a threshold is required to stop the merging at a certain level. For example, in the dendrogram of Figure 9, the threshold has been set as 1 which is presented as the horizontal red line. The black lines means the stock is isolated and unsassigned on the condition of threshold. if the threshold keep increasing, these 9 clusters will continue to merge into a one cluster. Especially, the longest black vertical line started from X axis lies between pink cluster and dark blue cluster in the right side of the figure, it represent an unassigned stock. This means that under the constraint of the given threshold, this stock is not assigned to a cluster and isolated. However, if threshold increases, this stock will firstly be merged into the pink cluster and then the light blue cluster.

Now, let's use a variable control method to compare results based on different distance function and metrics.

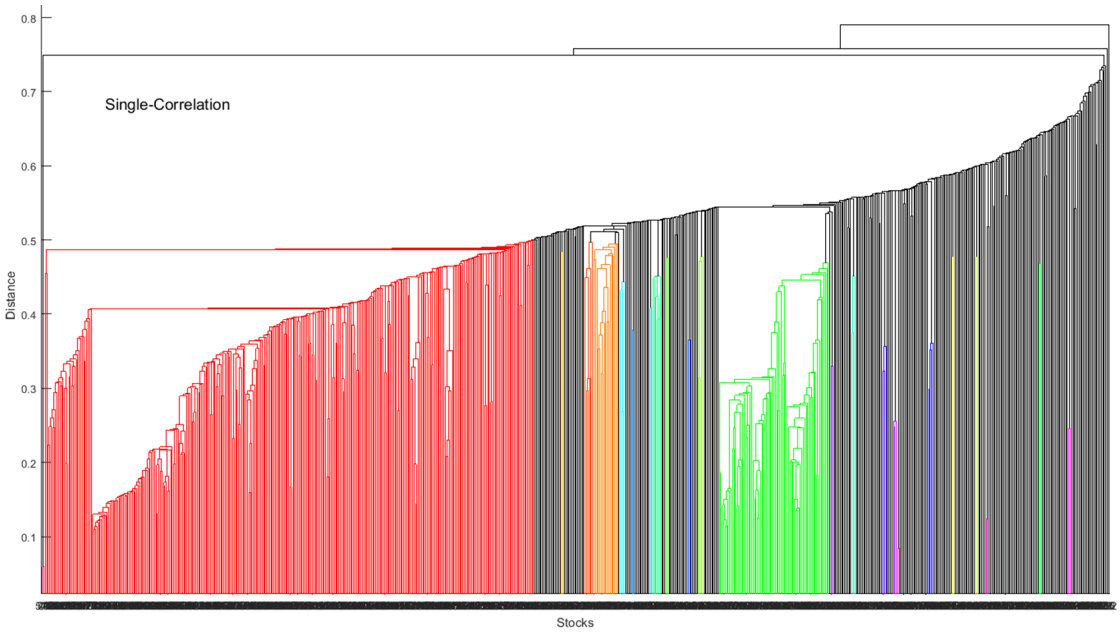


Figure 10: Dendrogram with Single and Pearson's correlation result

The single distance and Complete distance were used to draw the Figures 10 and Figure 11 respectively, the thresholds will be different also they need to be able to achieve a reasonable number of clusters. The result of Single distance was not very good, almost half of the stocks are in the red cluster with rest either keep isolated (black) or in a small cluster which includes less than 5 stocks. Moreover, if a threshold over 0.7 is given, all the stocks will be in the same cluster.

What's more, for red cluster, one can see from left to right, the height of cluster (Y coordinate) increases, which implies that every iteration of algorithm merges a unassigned stock with an existed cluster and this cluster becomes bigger when we keep iterating. The reason of this result is that no stock can be an isolated one, it must correlate with other stocks. Hence when we choose Single distance, what happens is a cluster with many stocks is more attractive to a unassigned stock than a isolated one, such that stocks will be merged into a large cluster rather than several smaller clusters.

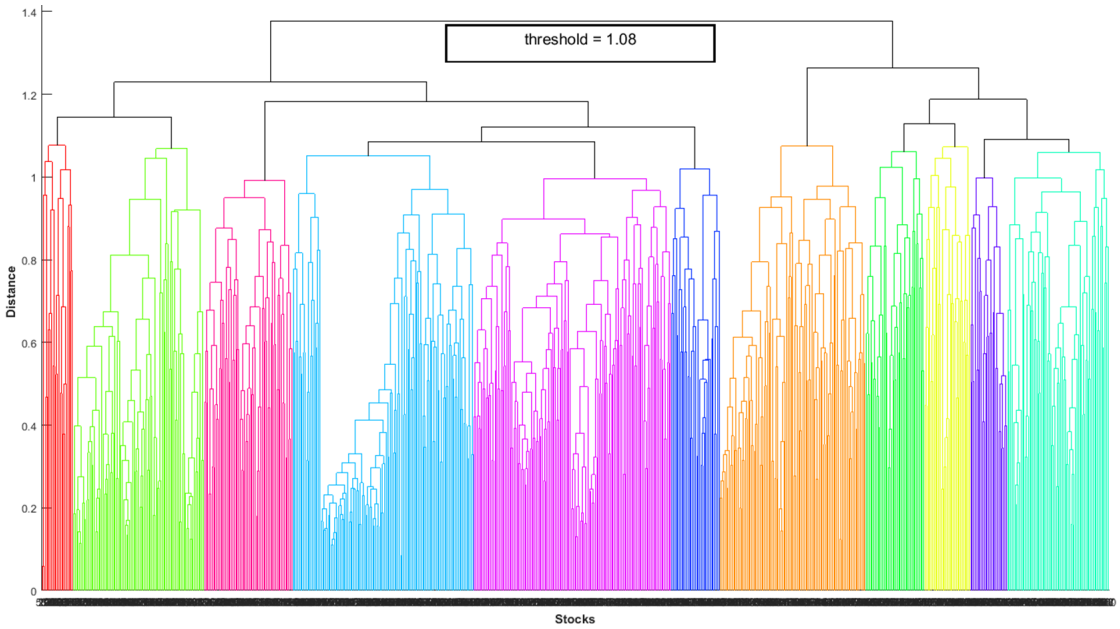


Figure 11: Dendrogram with Complete and Pearson's correlation result

Compared with the result of single distance, the Figure 11 of Complete distance shows a much more clear hierarchical structure than Single distance. The stock universe is divided into 11 clusters with a given threshold 1.08. For convenience and computation time, this function defines $1-\rho$ as measure between two stocks rather than $\sqrt{2*(1-\rho)}$, and it is only a metric rather than a distance due to its definition. The range of $1-\rho$ is $[0,2]$, so that if the value is higher than 1 it implies two stocks are negatively correlated.

The result of Pearson's correlation and Hausdorff distance presented in Figure 12 is better than Single distance but worse than Complete distance. The number of unassigned stocks has been reduced compared with Single distance while the massive red cluster in 10 disappeared. However, the hierarchical construction of Hausdorff distance result is not as clear as Complete distance result shows, small clusters containing less than 5 stocks still exist and a significant number of stocks are still unassigned.

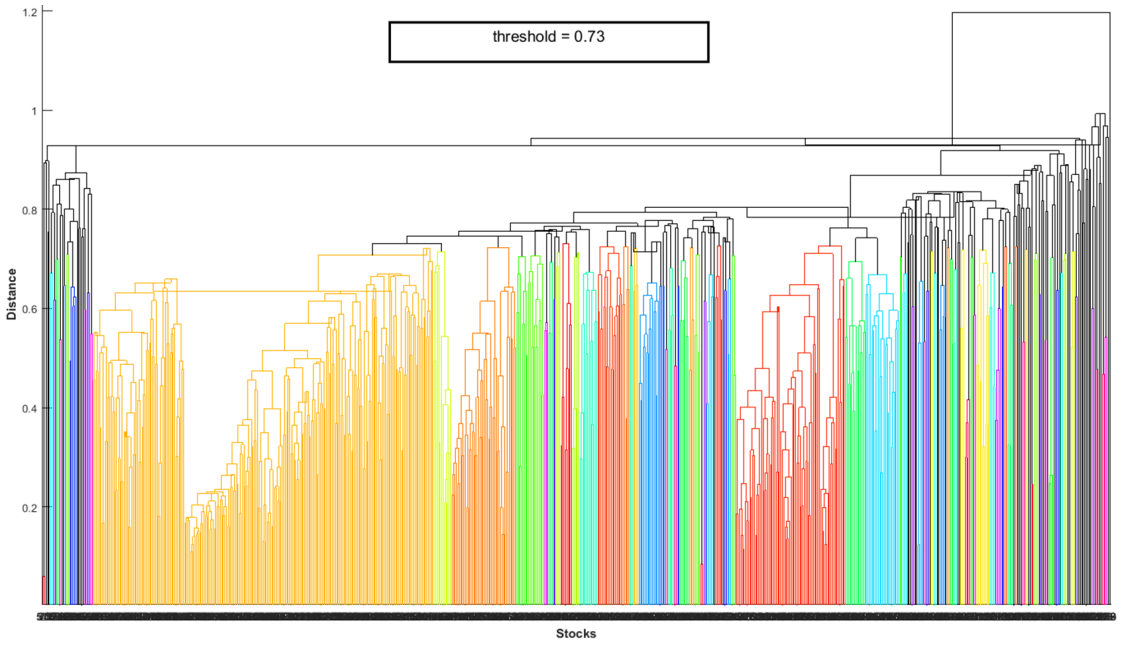


Figure 12: Dendrogram with Hausdorff and Pearson's correlation result

Furthermore, in Figure 13, vertical lines in the red circles start from the middle of horizontal lines and drop down when two clusters merge. This implies that distance between them is smaller than distances between their inside subsets. In general, distances between two clusters will increase when clusters become bigger, just like the result of Complete distance, in which vertical lines always go up. Nicolas et al. named this phenomenon as 'backstep' and gave a mathematical explanation of it [29] which the details are included in Appendix A. However, the practical meaning of backstep in our stock system is that though two groups of stocks belong to different clusters, they are more correlated with each other than stock groups in the same cluster.

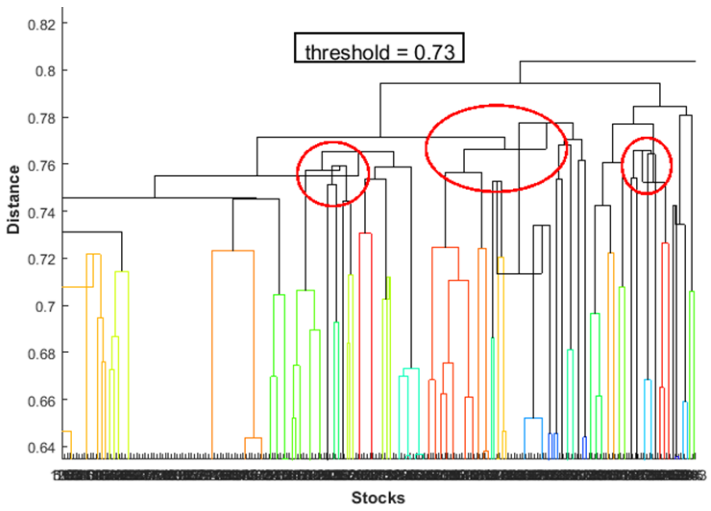


Figure 13: Hausdorff discovery

Since Complete distance result has the best results, we use it to compare results of different distance functions under the constraint of Complete distance. Figure 22, Figure 23 and Figure 24 in Appendix are the results of Pearson's correlation, Spearman's correlation and Cosine using weekly returns respectively. The reason why daily returns are not suitable is that weekly returns have less volatility and are more stable.

The above three results are equally good since all of them have a clear hierarchical construction without unassigned stocks and small clusters. Therefore, let's investigate the inside correlation and dependence function which was introduced during the analysis of K-means clustering. Table 6 has recorded the inside correlation of every cluster in results of different distance function and tables of components of three results are presented in Appendix A.

Table 6: Inside Correlation of K-means and BICS

Cluster	1	2	3	4	5	6	7	8	9	10	mean
Pearson	0.16	0.14	0.24	0.20	0.25	0.36	0.28	0.27	0.22	0.15	0.23
Spearman	0.31	0.33	0.31	0.44	0.14	0.21	0.18	0.29	0.26	0.13	0.26
Cosine	0.23	0.19	0.10	0.22	0.36	0.25	0.40	0.22	0.15	0.16	0.23
BICS	0.19	0.21	0.23	0.41	0.33	0.19	0.31	0.28	0.28	0.45	0.28

From the last column of Table 6 one can see that the means of inside correlations of all three results are 0.23 ,0.26 and 0.23 respectively and their means of pairwise dependence function values between clusters are 0.1370, 0.1357 and 0.1397 respectively.

By comparing with BICS code, the inside correlation and dependence matrix, the inside correlation means of all three results are lower than inside correlation mean of BICS code. The means of dependence (0.1370, 0.1357 and 0.1397) have a better performance than BICS code (0.1476).

Among three results, the Spearman correlation is best since not only it has the highest inside correlation average but also has the lowest dependence average. Moreover, in Table 16, we can see that the result of Pearson's correlation and Complete distance using weekly stock returns includes a small cluster which only contains 3 stocks. Besides, the other cluster consists of stocks from at least 6 BICS code sector. Both details implies that the result is worse than result of K-means or BICS code since small cluster exists and some clusters have too many components.

Table 15 and 17 are the results of Spearman-Complete and Cosine-Complete, both of them have a large cluster which includes over 200 stocks, it indicates that stock universe is not well partitioned and this is similar with what happened when Single distance is applied. In conclusion, Complete distance has a better performance than Single distance and Hausdorff distance while Spearman's correlation is better than Pearson's correlation and Cosine measure. However, using Complete distance and Spearman's correlation can only partly solve the problem of the existence

of overlarge clusters.

As we know from previous introduction (Section 2.3) that the result of Hierarchical clustering is a MST when the Single metric is applied. The stock can only link with one other stock if it is a leaf and total number of edges is $n-1$ where n is number of stocks. However, practically speaking any given stock is not only correlated with one other stock. Indeed, one stock will usually have strong correlations with several stocks. Hence, more edges are needed if we want to draw out the graph of stock system which is too complicated for a MST cannot adequately explain the structure.

Therefore, when different distance functions and metrics are used for Hierarchical clustering, the only change can make is the way to link stocks but not increase the number of edges. This is the reason why Complete distance and Spearman's correlation are only a partial solution. Fortunately, this fact suggests that even though as one result of Hierarchical clustering does not work, one can give more edges based on MST so that enough information are contained in the graph to analyse the stock system.

The final point worth mentioning about Hierarchical clustering is its uniqueness. That is given a data set, the result is deterministic. Usually one would care about this since every time the K-means is used, results will converge to a arbitrary local optimal solution if initial points are randomly chosen and thus many iterations are needed to approximate the global optimal solution.

Fortunately, the result of Hierarchical clustering is unique. Consider that data set, distance function and metrics are given, so that $n \times n$ distance matrix which includes all distance between pairwise stocks is determined. Hence for every iteration, one would only need to find the minimum of this distance matrix and merge its corresponding two stocks. The result would not change since all distances are determined at the start.

3.4 Result of PMFG

Based on the consequence of Hierarchical clustering, we want to find an extension of MST which contains more information. Fortunately, PMFG could be a good tool to analyse stock system since it includes MST as a subgraph and have more edges than MST. Let's refer the stocks as vertices and apply Pearson's correlation to compute distance between stocks.

The result of algorithm is a $n \times n$ sparse matrix, where n is the number of stocks in the universe, in this case the size of the matrix is 724×724. The matrix is symmetric and every row represents a stock, thus row i and column i is equivalent to each other. In this sparse matrix, most elements are 0 which indicates that its two corresponding stocks are not linked while each nonzero element represents an edge between two corresponding stocks. For example, $a_{ij} = 0.5$, as an element of PMFG sparse matrix, implies an edge between stock i and stock j with Pearson's correlation value at 0.5.

In the PMFG matrix, some stocks are very popular since they are connected with more than

10 stocks while others only connect with 3 or 4 stocks. For example, JPM[1] is linked with 16 stocks while $IDCC$[2] is only linked with 3 stocks. The reason of this phenomenon is that a major corporation such as JPM has many different association with many companies so that it is a popular company while $IDCC$ only focuses on mobile and wireless technology which simply correlates with relevant companies.

To quantify the *popularity* of a sector, let's define it as the mean number of stocks which link with stocks in the sector that is for sector i, we have:

$$POP_i := \frac{1}{n_i} \sum_{j=1}^{n_i} p_j, \tag{3.4}$$

where n_i is the number of stocks sector i and p_j is the number of stocks linked with stock j based on PMFG or the number of nonzero elements of row i of PMFG matrix. Furthermore, for sector i, one can define the *inside popularity* and *outside popularity* as the mean number of stocks from sector i which link with stocks in sector i and the mean number of stocks out of sector i which link with stocks in sector i respectively. That is:

$$POP\_in_i := \frac{1}{n_i} \sum_{j=1}^{n_i} a_j \quad \text{and} \quad POP\_out_i := \frac{1}{n_i} \sum_{j=1}^{n_i} b_j, \tag{3.5}$$

where a_j is the number of stocks from sector i which link with stocks in sector i and b_j is the number of stocks out of sector i which link with stocks in sector i, thus we have:

$$p_j = a_j + b_j \quad \text{and} \quad POP_i = POP\_in_i + POP\_out_i.$$

Table 7 shows details of popularity of every sectors. It is straightforward that stocks in the Industrial sector and Financials sector are the most popular (popularity is 7.1 and 6.9 respectively) while stocks of Communication sector are the least popular. POP_in is higher than POP_out for most sectors except Communications and Materials. Among them, one can see POP_in of Energy, Financials and Utilities are much higher than their POP_out which indicates that stocks in these sectors have a stronger correlation with stocks in the same sector than stocks in other sectors. In other word, these sectors have a good clustering property which means they have a solid inside correlation. One thing worth noting is that one of the consequences of the result with K-means clustering (Section 3.2), it is consistent with the result of popularity.

What's more, as mentioned in Section 2.4, PMFG consists of 3-cliques and 4-cliques in which are strongly correlated stocks. Specifically, for the data set made up of daily returns of 724 US stocks in a timespan ranging from 2^{nd} Jun 2017 to 21^{st} May 2018, it contains 2 3-cliques and 718 4-cliques. Let's start by analyse the structure of cliques, since there are 724 stocks in universe, the number of edges is 2166 according to properties of PMFG. However, 720 cliques contains 2878

[1] JPM is the ticker of JPMORGAN CHASE & CO
[2] $IDCC$ is the ticker of INTERDIGITAL INC

Table 7: popularity of sectors

	Com	ConDis	ConSta	Energy	Fin	Heal	Ind	Mat	Tech	Uti
POP	3.9	5.1	4.8	6.3	6.9	4.8	7.1	5.8	6.0	6.0
POP_in	1.6	2.8	2.8	5.1	5.6	2.5	3.6	2.3	3.5	5.1
POP_out	2.3	2.3	2.0	1.2	1.3	2.3	3.5	3.5	2.5	0.9

vertices and 4314 edges which is much larger than 724 vertices and 2166 edges. The reason maybe that some vertices and edges must be included in several cliques. For example, two 3-cliques have the form presented in Figure 14. Stock TKR, stock AOS and edge between them are in public for both two 3-cliques.

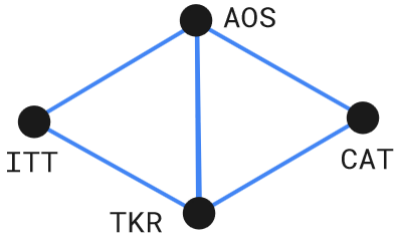


Figure 14: Example of 3-cliques

As shown in Figure 15, 4-cliques can only share vertices but have distinct edges. What's more, the structure can be even more complicated, as the red vertex in Figure 15 which is shared by three cliques in different ways. These three figures are just some examples of components of the whole clique structure, since some stocks are so popular, they are included in over 10 cliques, like JPM which is contained in 14 cliques.

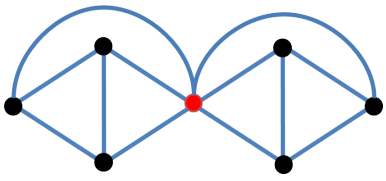


Figure 15: Example 1 of 4-cliques

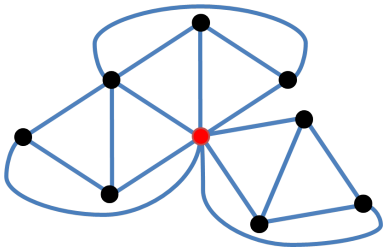


Figure 16: Example 2 of 4-cliques

Among 720 cliques, some of them have a special structure. As the example been given in the Introduction Section 1, some stocks have a higher correlation with stocks in other sectors due to their underlying products and business aspects.

Now, let's use the Communications sector as an example to study this phenomenon. Figure 27 attached in Appendix A contains the stocks in the Communications sector and edges between

stocks belonging to this sector, the edges between stocks in Communications sector and stocks in other sectors are removed. Red rectangles with tickers are vertices representing stocks while double arrow segments are edges given by algorithm. We can see that there is a group of 6 stocks which consists of 3 4-cliques and a smaller group of 4 stocks which is just a 4-clique. What's more, 10 isolated vertices lay around the figure. The reason why these stocks are separated is that they are only linked with stocks from other sectors, so that they become isolated when I remove these edges.

Since edges are given according to pairwise Pearson's correlation between stocks, we can conclude that isolated stocks showed in Figure 27 have a stronger relation with other sectors than its belonging sector. Here are three possible explanations of isolated stock formation:

Firstly, several companies might rely heavily on an underlying product and hence will strongly correlate. This effect can be stronger than the effect of a company's own sector.

Secondly, the data set is incomplete. Since 724 stocks in our data set are randomly chosen from whole US equity market, stocks which are relevant with isolated stocks may not be chosen into our data set.

Thirdly, this may be just a coincidence, hence we need to do experiments to check it. Moreover, one can see that there is a pair of stocks which only link with each other. However, one should know that the PMFG only consists of cliques, hence this pair must be a component of a clique in which other vertices belong to other sectors. This pair indicates that some stocks are not only close with stocks in the same sector but also stocks in different sectors. A convincing reason of this is that some major corporations have many aspects of business with companies both in the same sector or other sectors so that stock price of major corporations is influenced by many companies.

Even some isolated stocks are fairly strongly correlated with the stocks in the same sector, but the PMFG graphs shows most of them are clustered in a different sector. This simply means the correlation between the isolated ones and other sector stocks are slightly stronger, that's why PMFG classifies them in to same cluster.

Finally, among the BICS code 10 sectors, the amount of isolated stocks of Energy, Financials and Utilities is less than 3, it's significantly fewer than other sectors. This illustrates previous remark that these 3 sectors have a better clustering property than others. Besides, Financials sector is clearly divided into two groups: a small group corresponds to REIT industrial and a big group which corresponds to the union of other industries in Financials sector. This is another way to demonstrate the consequence we got from K-means clustering mentioned in Section 3.2.

Since the stock market has been divided into cliques, the cliques can be treated as elements to construct clusters so that these clusters will have better *inside correlations* and *dependence* than BICS code sectors. The descirbed method can be explained through the following algorithm:

- Step 1: Compute the average daily return sequence of stocks in the same clique and regard

them as pseudo stocks.

- Step 2: Draw a PMFG of pseudo stocks and divide it into cliques.
- Step 3: Keep iterating until the amount of cliques reduces to a preset threshold k.
- Step 4: Each clique represents a cluster, go back to find stocks belong to each clique.

However, some cliques need to be broken up before the start the algorithm. The reason being many stocks are included in multiple cliques so that if the cliques are directly merged into clusters, some stocks will belong to multiple clusters.

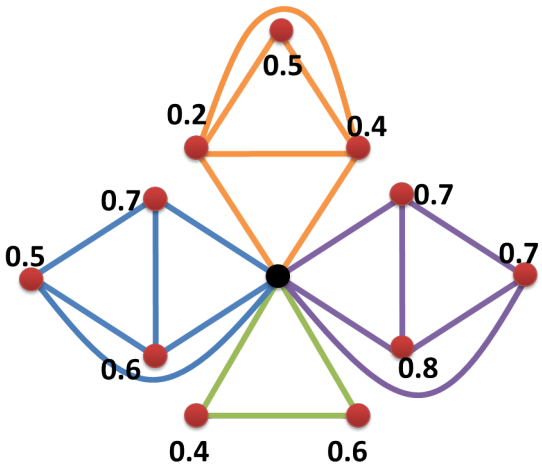


Figure 17: Clique Breaking up

For a stock, one would like to assign the stock in to a clique which is most correlated to this stock. For instance, in Figure 17, black point belongs to one 3-clique and three 4-cliques. The numbers present distances between their corresponding red points and black point, it is straightforward that the purple clique on the right has the highest average correlation (0.73) so that the black point belongs to the purple clique while blue clique and orange clique will be broken up into 3 clique and green clique will be a segment.

The same strategy is applied to the PMFG cliques, they will be broken into either 3-cliques, segments or isolated stocks which can be regarded as elements of next iteration. For every element, the average daily return sequence is used to construct a dash stock such that it's ready to draw a new PMFG graph. The result of this algorithm details in Table 8.

Table 8: Result of PMFG Clustering

Num	Com	C-D	C-S	Ene	Fin	Heal	Ind	Mat	Tec	Uti
48	0%	0%	0%	89.6%	0%	0%	6.3%	0%	0%	4.2%
103	0%	13.6%	1.0%	1.0%	12.6%	2.9%	46.6%	17.5%	4.9%	0%
59	13.6%	66.1%	15.3%	0%	3.4%	0%	1.7%	0%	0%	0%
95	0%	25.3%	2.1%	0%	4.2%	4.2%	37.9%	10.5%	15.8%	0%
35	5.7%	2.9%	62.9%	0%	2.9%	8.6%	0%	2.9%	14.3%	0%
52	0%	0%	0%	0%	0%	92.3%	0%	0%	7.7%	0%
89	4.5%	10.1%	2.2%	1.1%	1.1%	19.1%	3.4%	1.1%	57.3%	0%
88	0%	0%	0%	0%	98.9%	0%	1.1%	0%	0%	0%
76	10.5%	26.3%	3.9%	0%	10.5%	3.9%	11.8%	17.1%	15.8%	0%
32	3.1%	0%	0%	0%	9.4%	0%	0%	3.1%	0%	84.4%
47	0%	0%	0%	0%	97.9%	0%	0%	0%	0%	2.1%

The threshold has been set as 10, the algorithm will stop at point when it has 11 clusters. One can see Table 8 is quite similar with the result from K-means clustering Table 3. Specifically, 89.6% of cluster 1 is mainly the Energy sector, similarly cluster 9 are mainly the Utilities sector. Moreover, Financials sector is mainly distributed in cluster 8 and cluster 11.

In fact, all Financials stocks in cluster 11 came from REIT, and the same result can be seen in result of K-means clustering. Furthermore, other BICS code sectors are located in many clusters without significant distributions except Health Care sector. Indeed, Health Care sector is a special one since it has a good clustering property with a low inside correlation. Finally, to compare result of PMFG and result of K-means clustering, need the *inside correlation* and *dependence* in Table 9.

Table 9: Inside Correlation of PMFG and BICS

PMFG	0.41	0.38	0.26	0.28	0.27	0.19	0.35	0.54	0.22	0.48	0.49	0.3518
K-means	0.36	0.25	0.17	0.45	0.32	0.24	0.52	0.38	0.44	0.20		0.3317
BICS	0.19	0.21	0.23	0.41	0.33	0.19	0.31	0.28	0.28	0.45		0.2839

From Table 9, one can see the average *inside correlation* of PMFG is 6% higher than K-means and 24% higher than BICS. Indeed, the PMFG can increase the strength of correlations inside a cluster and has a better performance than K-means. The blue numbers corresponding *inside correlation* of Health Care sector or clusters whose main component is Health Care sector, which are very small comparing to the others.

According to results of PMFG and K-means, Health Care sector has a good clustering property.

However, its *inside correlation* is below 0.2, which is even smaller than sector which does not have a good clustering property.

In general, sector with good clustering property normally have a high inside correlation since algorithms are based on distance or metric so that if stocks in the same sector do not have strong pairwise correlations, they will be distributed into many clusters. The only reason these blue numbers can be explained is the stocks in Health Care sector have even worse relations with stocks outside of Health Care sector than those which are within the same sector.

As Figure 8 is shown, with the K-means result, when the number of clusters is increasing, the average distance between stocks and centroids will gradually decrease. The decrease in average distance is equivalent to the increase of inside correlation, the outperformance of PMFG is due to either the increase of cluster number or the method since the result of PMFG has one more cluster than K-means and BICS code.

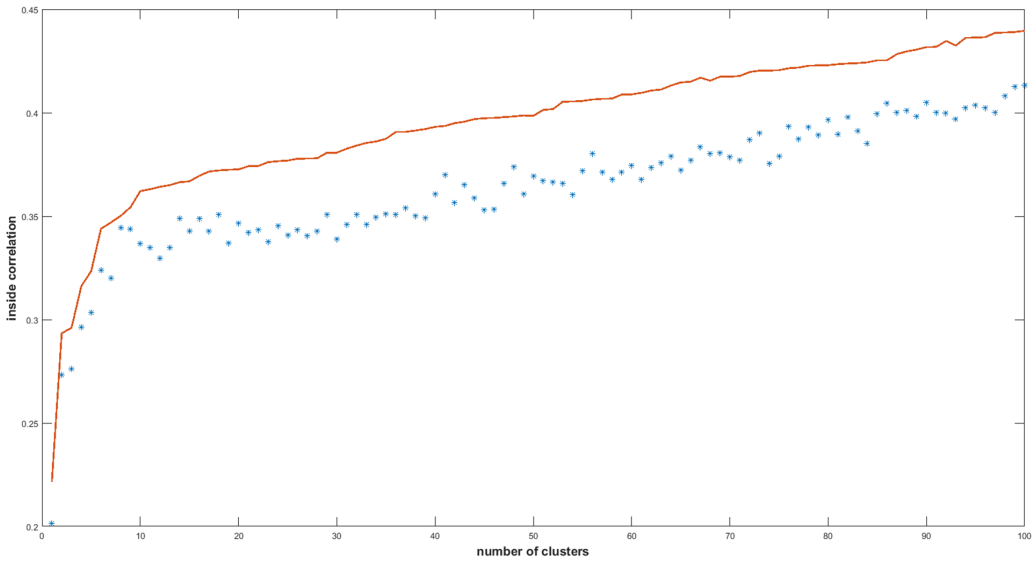


Figure 18: relation between inside correlation and number of clusters, red line is the reult of PMFG, the blue star is the result of K-means.

Figure 18 compares inside correlations of PMFG and K-means when the number of cluster increases. Whatever the number of cluster is, the inside correlation of PMFG is higher than K-means. Besides, the red line is more stable than blue stars which is due to the unstability of K-means result. What's more, both inside correlations of PMFG and K-means increase sharply when the number of cluster is small than 10, then the rate of lines reduces to a certain level after the number of clsuter is over 10. From the aspect of inside correlation, 10 or 11 is a suitable number of cluster.

4 Application in Trading

In this section, two existing trading strategies will be introduced, Momentum Strategy and Betting Against Beta (BAB) Strategy. Then the BICS code and the results of K-means clustering and PMFG will be compared by the Profit and Lose (PnL) and statistics.

4.1 Sector-Momentum

The Sector-Momentum strategy seeks to invest sectors which have the best performance over a specific timespan. James O'Shaunessey [30] stated in his book that Momentum investing is one of the best performing strategies in over the last fifty years. He argued that the stronger sectors tend to get stronger while the weaker sectors tend to get weaker since 'Wall-Street loves winners and hates losers'. The following steps are based on Mebane Faber's work [31].

The strategy does sector rotation trading where monthly returns data is applied. Specifically, for BICS code: Firstly, at the beginning of every month, three best performing sectors are chosen from 10 BICS code sectors according to their previous month performance.

Secondly, buying all the stocks in these three sectors and holding the positions for the next one month. Then at the next beginning of month, we rebalance the portfolio.

The strategy uses the 12-month moving average as a threshold to decide to get into the market or not. That is at the beginning of every month, one will buy the chosen stocks if the S&P500 Index is above the 12-month moving average and one will not buy stocks and/or remove positions as soon as S&P500 Index is below the 12-month moving average.

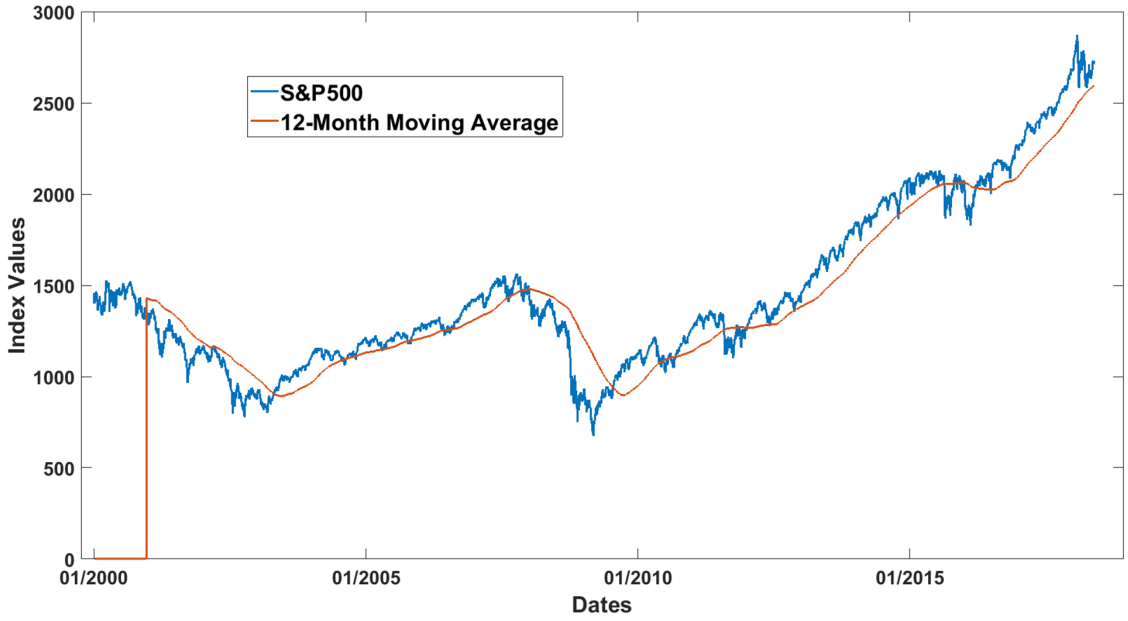


Figure 19: S&P500 and its 12-Month Moving Average.

The price of the S&P500 and its moving average are presented in Figure 19. Most of the time, the S&P500 is above its moving average, which indicates that the market return is positive. However, during the Financials Crisis in 2008, the moving average is above the S&P500, which indicates that one should sell his holdings. Therefore, the moving average ensures that investors are out of the market during extended down-trends and in the market during extended up-trends.

The strategy also works based on the clusters from the result of K-means and PMFG. Thus, following the steps introduced above and applying the 724 stock universe with timespan ranging from 30^{th} Aug 2003 to 21^{st} May 2018, Figure 20 presents the PnL lines of this strategy based on different classifications and Table 10 presents values of some statistics in terms of different methods.

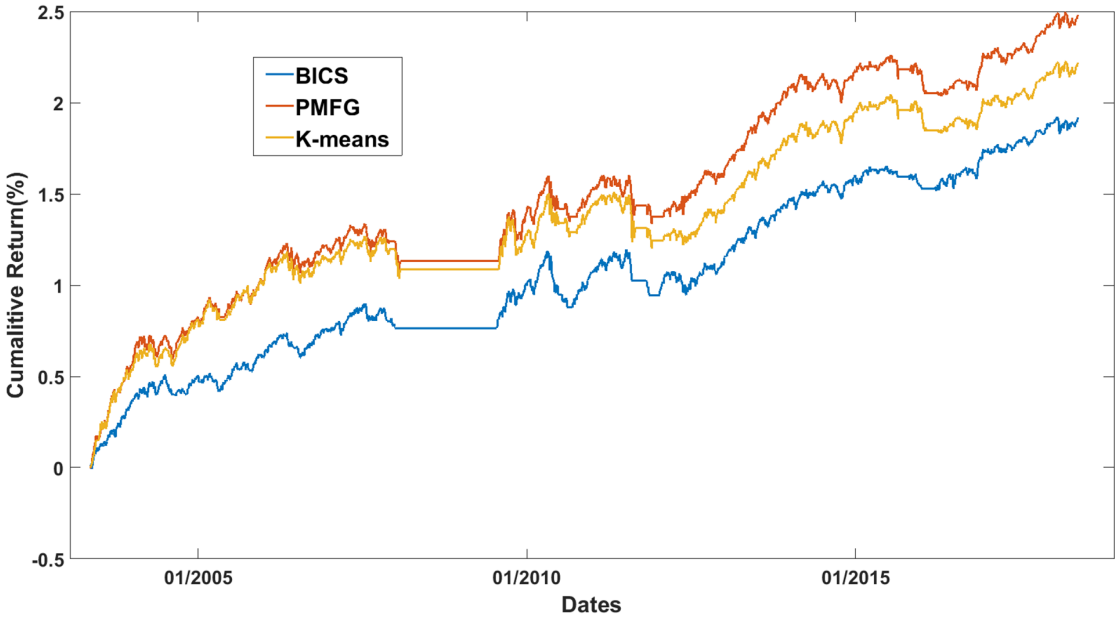


Figure 20: PnL of Momentum Strategy based on BICS code, K-means and PMFG.

Table 10: Statistical data of PnL Lines

	Annualised Return	Annualised Volatility	Sharpe Ratio	t-Statistic
K-means	0.1203	0.1463	0.8391	3.3132
PMFG	0.1335	0.1434	0.9126	3.6036
BICS	0.1039	0.1189	0.8739	3.4506

In Figure 20, difference between PnL lines is not remarkable. In fact, PMFG beats others with the respect of Sharpe Ratio and Annualised Return. All t-Statistic values for three methods are not small, which implies that results are reliable. In conclusion, for momentum strategy, clusters given by PMFG perform slightly better than BICS code sectors and clusters from K-means, which

implies that PMFG adds value when desiging a trading system.

4.2 Betting Against Beta

The Betting Against Beta (BAB) Strategy is detailed by Cliff Asness [32]. The idea of this strategy is that safer (low-risk) stocks deliver higher risk-adjusted returns than riskier stocks. Hence, the strategy buys low-beta stocks and sells high-beta stocks. In the following, steps of constructing the strategy are stated.

For a stock sector $S = \{s_1, s_2, ...s_n\}$:

Firstly, Beta β_i of a stock s_i is estimated by product of the rolling one-year daily standard deviation std_i which is normalised by the rolling one-year daily standard deviation of S&P500 and the rolling five-year three-day correlations ρ_i between the stock returns and S&P500 index:

$$\beta_i = \frac{std_i}{std_{S\&P500}} * \rho_i.$$

Correlation of three-day returns is applied here since it is more stable than correlation of one-day returns.

Secondly, all stocks in the sector S is ranked in ascending order on the basis of their estimated Beta. Denote $z_i = rank(\beta_i)$ as the rank of stock s_i and $\bar{z}$ as the average of $\{z_1, z_2, ...z_n\}$.

Thirdly, give weights to stocks. Dividing stocks into two portfolio: low-risk portfolio and high-risk portfolio. Stock s_i belongs to low-risk protfolio if $z_i - \bar{z} < 0$ otherwise belongs to high-risk portfolio. For stock s_i in low-risk portfolio, weight is given by:

$$w_L^i = \frac{(z_i - \bar{z})^-}{\beta_L' * z_L},$$

where β_L is a column vector of Betas for all stocks in low-risk portfolio and z_L is a column vector of the values of $(z_i - \bar{z})^-$ for all stocks in low-risk portfolio.

Similarly, for stock s_j in high-risk portfolio, weight is given by:

$$w_H^j = \frac{(z_j - \bar{z})^+}{\beta_H' * z_H},$$

where β_H is a column vector of Betas for all stocks in high-risk portfolio and z_H is a column vector of the values of $(z_j - \bar{z})^+$ for all stocks in high-risk portfolio.

Fourthly, the return of low-risk portfolio r_t^L and the return of high-risk portfolio r_t^H at day t is constructed as:

$$r_t^L = r_L * w_L,$$
$$r_t^H = r_H * w_H,$$

where $r_L = (r_t^1, r_t^2, ..., r_t^n)$ is a row vector of stock returns in low-risk portfolio at day t, $w_L = (w_L^1, w_L^2, ..., w_L^n)'$ is a column vectors of stock weights in low-risk portfolio, $r_H = (r_t^1, r_t^2, ..., r_t^m)$ is a

row vector of stock returns in high-risk portfolio at day t and $w_H = (w_H^1, w_H^2, ..., w_H^n)'$ is a column vectors of stock weights in high-risk portfolio.

Finally, the return of strategy at time $t+1$ is computed by:

$$r_{t+1} = r_{t+1}^L - r_{t+1}^H. \tag{4.1}$$

Since the strategy is defined on a sector S, BICS code sectors, result of K-means and result of PMFG are applied respectively to the data set of this strategy. For the BICS code, the strategy will work based on 10 sectors respectively. For K-means and PMFG, stocks will be divided into 10 clusters and then clusters will be applied to the strategy. Thus, following the steps introduced above and applying the 724 stock universe with timespan ranging from 31^{st} Aug 2009 to 21^{st} May 2018, Figure 21 presents the PnL lines of this strategy based on different classifications and Table 11 details values of diffrent statistics with respect to different methods.

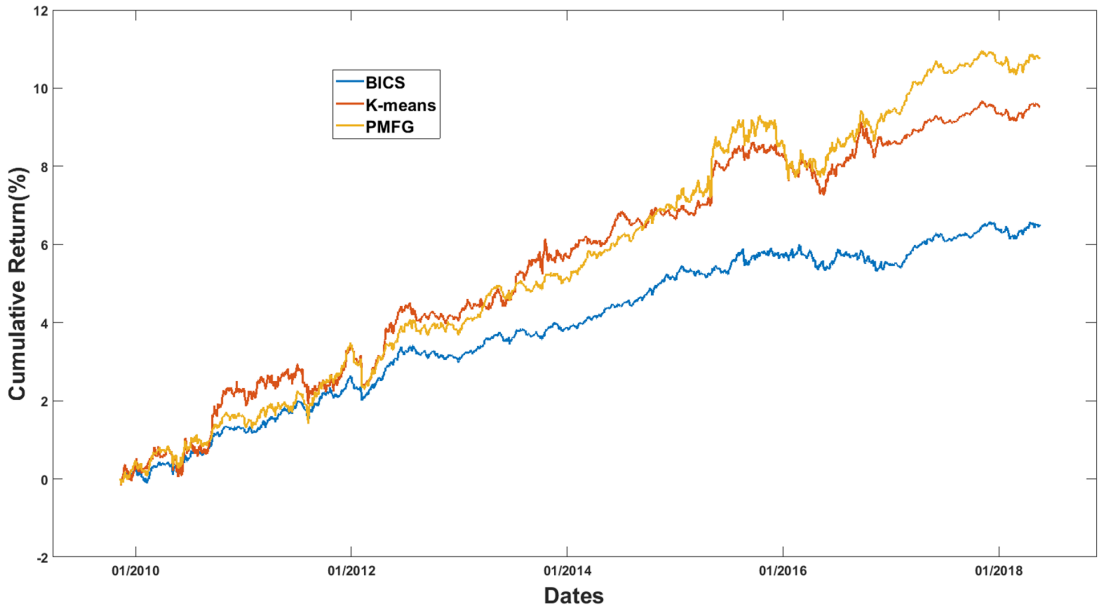


Figure 21: PnL of BAB Strategy based on BICS code, K-means and PMFG.

Table 11: Statistical data of PnL Lines

	Annualised Return	Annualised Volatility	Sharpe Ratio	t-Statistic
K-means	0.5012	0.7433	0.6743	2.0264
PMFG	0.5666	0.6908	0.8202	2.4651
BICS	0.3420	0.4130	0.8282	2.4890

In Figure 21, both PMFG and K-means have a higher PnL than BICS, while PMFG is significantly higher than K-means. However, the Sharpe Ratio of K-means is 0.6743, which is much more

lower than PMFG and BICS. Hence PMFG and BIC beat K-means in terms of Sharpe Ratio. The lower Sharpe Ratio of K-means implies that its volatility is higher than others. In other words, the highest Sharpe Ratio of BICS suggests that this is risk-lowest one. Hence, the performance of K-means are not as good as one supposed.

The reason why this happened is because the result of K-means and PMFG are both only based on distance functions, that is to say only one parameter is used. In fact, more parameters are required to classify stocks since stock market is too complicated to be estimated by one parameter. For instance, stock volatility and trading volumes.

5 Conclusion

We have introduced three clustering methods and applied them to analyse a stock universe which includes 724 stocks from US stock market. Based on the daily return sequences of stocks, the distance functions were used to assess correlation between the stocks. The strong correlations are picked out by clustering algorithms and thus structure of stock market has been described by clustering results.

Applying Pearson's correlation in K-means clustering method, it pointed out the extent of clustering property of every BICS code sector. Among them, Energy, Financials and Utilities sectors have a good clustering property while others are partitioned into several subsets and randomly distributed into many clusters. However, even though the K-means clustering method offers a remarkable result which improved the *inside correlations* of clusters, its algorithm can only find local optimal solutions.

Hierarchical clustering cares not only about the distance between stocks but also distance between stock groups. Three different but related distance functions were defined to evaluate relations between groups and applied to construct a hierarchical model for stock system. Indeed, it came out that the result of this method is a minimum spanning tree (MST), which contains too little information to draw out a super complex system like stock universe. However, this method became the foundation for constructing a even more complicated graph which is more informative.

PMFG is regarded as the extension of MST which have much more edges than MST and thus can convey more information. As components of PMFG, the cliques illustrated that the stocks are widely correlated with many others so that MST is not an adequate model. Besides, discovery of isolated stocks indicate that the existence of stocks which only have strong correlations with other sectors reduce the inside correlations. However, although the result of PMFG is better than K-means, stocks which have relations with many sectors remain to be difficult to split. The reason why this happens is that the distance which is based on daily return sequence cannot represent all the features of stock system. For example, technical data such as volatility and trading volume and fundamental data such as price earning ratio.

Both Momentum Strategy and BAB Strategy have a better performance when clusters based on PMFG are applied in terms of annualised return. Though K-means has a higher annualised return than BICS, Sharpe Ratio indicates that K-means is not as good as BICS. Indeed, the PnL based on PMFG is not as good as supposed. The Sharpe Ratio of PMFG is just slightly higher than or equal to BICS code. Hence, PMFG only makes a limited contribution to correction of BICS code.

In conclusion, clustering methods are useful tools to analyse structure of stock universe. However, an algorithm which has multivariate parameters is needed to draw out the structure of data in many different aspects.

A Appendix

Proof of derivative covariance formula

Assume X and Y are random variables, $x_1, x_2, ...x_n$ and $y_1, y_2, ...y_n$ are their n samples respectively. The covariance of X and Y can be expressed as:

$$\text{cov}(X,Y) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(y_i - y_j). \tag{A.1}$$

Proof. Denote $\vec{x} = (x_1, x_2, x_3, ...x_n)^T$, $\vec{y} = (y_1, y_2, y_3, ...y_n)^T$ and $\vec{E} = (1, 1, 1, ...1)^T$. Then construct two matrices:

$$A = \begin{pmatrix} x_1y_1 & & & & \\ & x_2y_2 & & & \\ & & x_3y_3 & & \\ & & & \ddots & \\ & & & & x_ny_n \end{pmatrix} \qquad B = \begin{pmatrix} x_1y_1 & x_1y_2 & x_1y_3 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & x_2y_3 & \cdots & x_2y_n \\ x_3y_1 & x_3y_2 & x_3y_3 & \cdots & x_3y_n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_ny_1 & x_ny_2 & x_ny_3 & \cdots & x_ny_n \end{pmatrix}$$

Hence we have:

$$\begin{aligned} \text{cov}(X,Y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_iy_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_iy_i) - \bar{x}\cdot\bar{y} = \frac{1}{n^2}[n\sum_{i=1}^{n}(x_iy_i) - (\sum_{i=1}^{n}x_i)(\sum_{j=1}^{n}y_i)] \\ &= \frac{1}{n^2}(nE^TAE - E^TBE) \end{aligned}$$

It is easy to compute that:

$$A - B = \begin{pmatrix} (n-1)x_1y_1 & -x_1y_2 & -x_1y_3 & \cdots & -x_1y_n \\ -x_2y_1 & (n-1)x_2y_2 & -x_2y_3 & \cdots & -x_2y_n \\ -x_3y_1 & -x_3y_2 & (n-1)x_3y_3 & \cdots & -x_3y_n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -x_ny_1 & -x_ny_2 & -x_ny_3 & \cdots & (n-1)x_ny_n \end{pmatrix},$$

so if we time n^2 on both sides, the equation will be:

$$\begin{aligned}
n^2\text{cov} &= ((n-1)x_1 - x_2 - x_3 - \cdots - x_n)y_1 + ((n-1)x_2 - x_1 - x_3 - \cdots - x_n)y_2 \\
&\quad + \cdots + ((n-1)x_n - x_1 - x_2 - \cdots - x_{n-1})y_n \\
&= (nx_1 - \sum_{i=1}^{n} x_i)y_1 + (nx_2 - \sum_{i=1}^{n} x_i)y_1 + \cdots + (nx_n - \sum_{i=1}^{n} x_i)y_n \\
&= \sum_{i=1}^{n}(x_1 - x_i)y_1 + \sum_{i=1}^{n}(x_2 - x_i)y_1 + \cdots + \sum_{i=1}^{n}(x_n - x_i)y_n \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n}(x_j - x_i)y_j.
\end{aligned}$$

Then let us do a subscript exchange:

$$\begin{aligned}
n^2\text{cov} &= \frac{1}{2}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}(x_j - x_i)y_j + \sum_{j=1}^{n}\sum_{i=1}^{n}(x_j - x_i)y_j\right) \\
&= \frac{1}{2}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}(x_j - x_i)y_j + \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)y_i\right) \\
&= \frac{1}{2}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}(x_j - x_i)y_j + \sum_{j=1}^{n}\sum_{i=1}^{n}(x_i - x_j)y_i\right) \\
&= \frac{1}{2}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}(x_j - x_i)y_j - \sum_{i=1}^{n}\sum_{j=1}^{n}(x_j - x_i)y_i\right) \\
&= \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}(x_j - x_i)(y_j - y_i).
\end{aligned}$$

Hence, the covariance is as expressed on (A.1). □

BICS Code Table

Table 12: First two levels of the Bloomberg BICS stocks system

Level 1		**Level 2**	
Code	**Macro Sector**	**Code**	**First level Microsector**
10	Communications	1010	Media Content
		1011	Telecom
11	Consumer Discretionary	1110	Apparel & Textile Products
		1111	Automotive
		1112	Consumer Discretionary Srvcs
		1113	Distributors
		1114	Home & Office Products
		1115	Leisure Products
		1116	Recreation Facilities & Srvcs
		1117	Retail Discretionary
		1118	Travel, Lodging & Dining
		1119	Distributors
		1120	Retail
12	Consumer Staples	1210	Consumer Products
		1211	Dist/Whsl-Consumer Staples
		1212	Retail Staples
13	Energy	1310	Oil, Gas & Coal
		1311	Renewable Energy
14	Financials	1410	Asset Management
		1411	Banking
		1412	Institutional Financials Srvcs
		1413	Insurance
		1414	Real Estate Oper & Srvcs
		1415	REIT
15	Health Care	1510	Biotech & Pharma
		1511	Health Care Facilities/Srvcs
		1512	Medical equipment/Devices

Code	Macro Sector	Code	First level Microsector
16	Industrials	1610	Aerospace & Defense
		1611	Electrical Equipment
		1612	Engineering & Const Srvcs
		1613	Industrial Distribution
		1614	Machinery
		1615	Manufactured Goods
		1616	Transportation & Logistics
		1617	Transportation Equipment
		1618	Waste & Envrnmt Srvc Equip & Fac
17	Materials	1710	Chemicals
		1711	Construction Materials
		1712	Containers & Packaging
		1713	Forest & Paper products
		1714	Iron & Steel
		1715	Metals & Distribution
18	Materials	1810	Design, Mfg & Distribution
		1811	Hardware
		1812	Semiconductors
		1813	Software
		1814	Iron & Technology Services
19	Utilities	1910	Utilities

Dependence Matrices of K-means and BICS code

Table 13: dependencekmeans

Clus	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8	No.9	No.10
No.1	0	0.26	0.18	0.19	0.28	0.20	0.32	0.20	0.14	0.21
No.2		0	0.14	0.14	0.21	0.17	0.29	0.15	0.10	0.18
No.3			0	0.11	0.16	0.13	0.16	0.11	0.06	0.14
No.4				0	0.15	0.15	0.15	0.11	0.31	0.14
No.5					0	0.13	0.21	0.14	0.12	0.18
No.6						0	0.23	0.15	0.08	0.15
No.7							0	0.18	0.10	0.21
No.8								0	0.02	0.17
No.9									0	0.08
No.10										0

Table 14: dependenceBICS

	Com	ConDis	ConSta	Energy	Fin	Health	Ind	Mat	Tech	Uti
Com	0	0.17	0.15	0.15	0.18	0.13	0.18	0.18	0.16	0.09
ConDis		0	0.18	0.17	0.22	0.15	0.22	0.21	0.19	0.09
ConSta			0	0.12	0.19	0.14	0.18	0.18	0.16	0.15
Energy				0	0.17	0.12	0.20	0.21	0.16	0.03
Fin					0	0.17	0.25	0.23	0.20	0.17
Health						0	0.18	0.18	0.19	0.09
Ind							0	0.27	0.24	0.10
Mat								0	0.22	0.11
Tech									0	0.10
Uti										0

Scoring function of three similarity measures

Table 15: Result of Spearman's correlation and Complete distance

Num	Com	C-D	CS-	Ene	Fin	Heal	Ind	Mat	Tec	Uti
29	0%	82.8%	13.8%	0%	3.4%	0%	0%	0%	0%	0%
46	0%	2.2%	0%	69.6%	10.9%	2.2%	4.3%	4.3%	6.5%	0%
34	2.9%	0%	8.8%	0%	85.3%	0%	0%	0%	0%	2.9%
43	0%	0%	2.3%	0%	44.2%	0%	0%	2.3%	0%	51.2%
30	13.3%	40%	6.7%	0%	6.7%	10%	0%	6.7%	16.7%	0%
43	0%	32.6%	23.3%	2.3%	27.9%	2.3%	2.3%	0%	0%	9.3%
38	2.6%	13.2%	18.4%	0%	10.5%	31.6%	5.3%	2.6%	7.9%	7.9%
286	2.8%	13.3%	2.4%	0.7%	28.7%	13.6%	22.4%	9.8%	6.3%	0%
141	4.3%	9.2%	3.5%	1.4%	3.5%	8.5%	19.9%	6.4%	43.3%	0%
34	8.8%	0%	0%	23.5%	17.6%	29.4%	11.8%	2.9%	5.9%	0%

Table 16: Result of Pearson's correlation and Complete distance

Num	Com	C-D	C-S	Ene	Fin	Heal	Ind	Mat	Tec	Uti
3	0%	0%	33.3%	0%	0%	66.7%	0%	0%	0%	0%
20	10%	10%	5%	0%	15%	30%	5%	0%	20%	5%
43	2.3%	23.3%	0%	0%	2.3%	16.3%	32.6%	11.6%	11.6%	0%
46	8.7%	26.1%	2.2%	4.3%	13.0%	21.7%	0%	6.5%	17.4%	0%
122	0.8%	4.9%	0%	27.9%	1.6%	11.5%	11.5%	4.9%	33.6%	0%
115	7.0%	7.8%	0.9%	5.2%	61.7%	3.5%	5.2%	3.5%	3.5%	1.7%
93	0%	2.2%	10.7%	0%	45.2%	6.5%	1.1%	3.2%	2.2%	29.0%
159	1.3%	18.9%	3.1%	0.6%	12.6%	11.3%	29.6%	10.7%	11.9%	0%
94	4.3%	36.2%	21.3%	2.1%	17.0%	1.1%	10.6%	6.4%	1.1%	0%
29	3.4%	6.9%	0%	0%	13.8%	34.5%	13.8%	0%	27.6%	0%

Table 17: Result of Cosine measure and Complete distance

Num	Com	C-D	C-S	Ene	Fin	Heal	Ind	Mat	Tec	Uti
48	0%	37.5%	4.2%	6.3%	14.6%	6.3%	12.5%	10.4%	8.3%	0%
47	4.3%	53.2%	6.4%	4.3%	6.4%	12.8%	8.5%	2.1%	2.1%	0%
11	9.1%	18.2%	0%	9.1%	27.3%	36.4%	0%	0%	0%	0%
59	3.4%	8.5%	18.6%	1.7%	20.3%	8.5%	3.4%	15.3%	20.3%	0%
132	4.5%	9.1%	5.3%	0%	58.3%	5.3%	7.6%	6.1%	3.8%	0%
248	3.6%	9.7%	2.0%	14.9%	6.5%	11.3%	25.4%	6.5%	19.4%	0.8%
67	0%	3.0%	6.0%	0%	49.3%	0%	0%	1.5%	0%	40.3%
50	6%	10%	4%	0%	12%	24%	26%	6%	10%	2%
27	0%	33.3%	18.5%	3.7%	11.1%	25.9%	7.4%	0%	0%	0%
35	0%	14.3%	0%	0%	14.3%	17.1%	2.9%	2.9%	48.6%	0%

Comparison of distance function under Complete metric

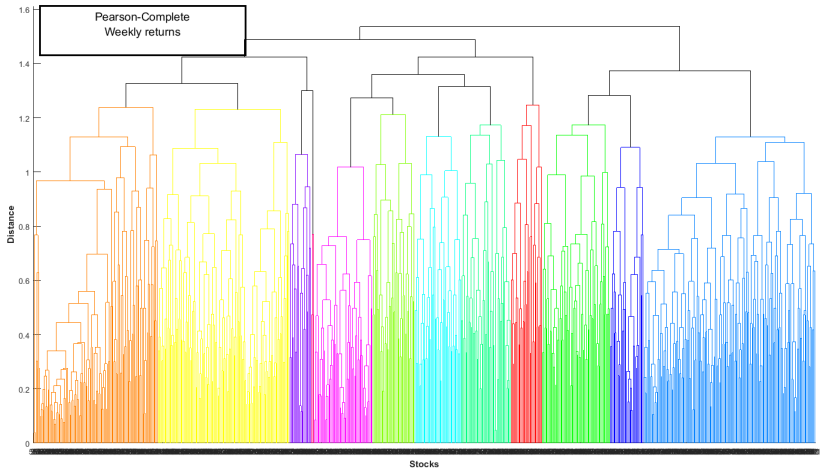


Figure 22: Complete and Pearson's correlation of the weekly returns result

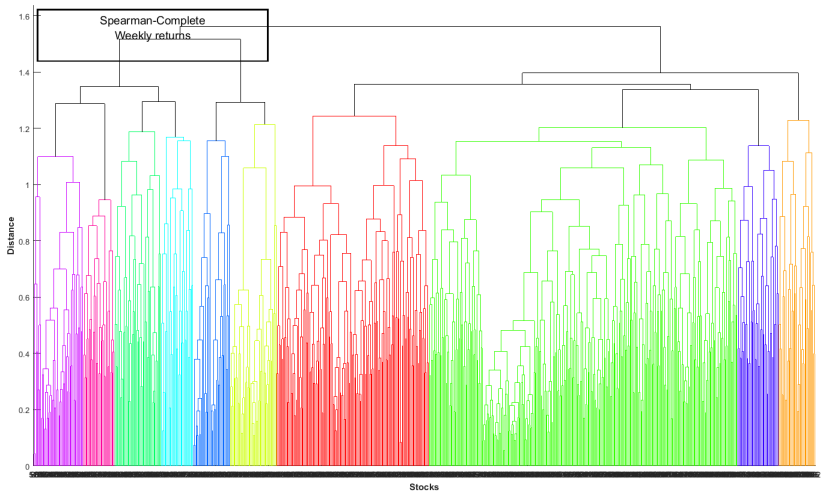


Figure 23: Complete and Spearman's correlation on weekly returns result

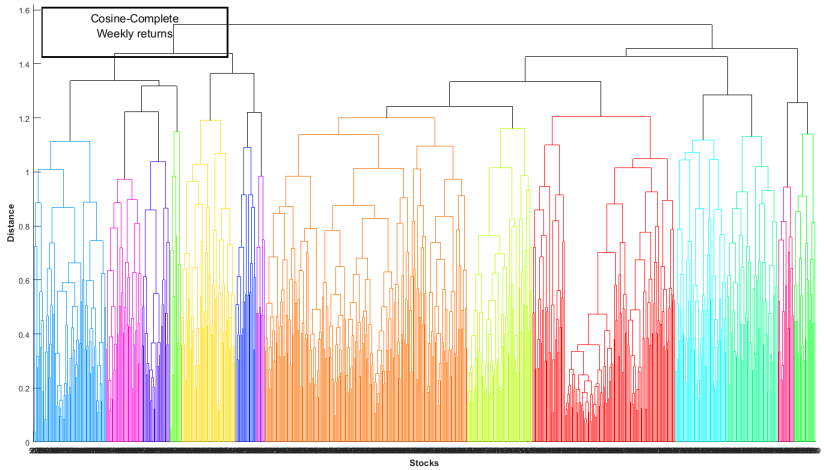


Figure 24: Complete and Cosine's correlation of weekly returns result

Mathematical explanation of backstep

Figure 25: Single and Pearson's correlation result

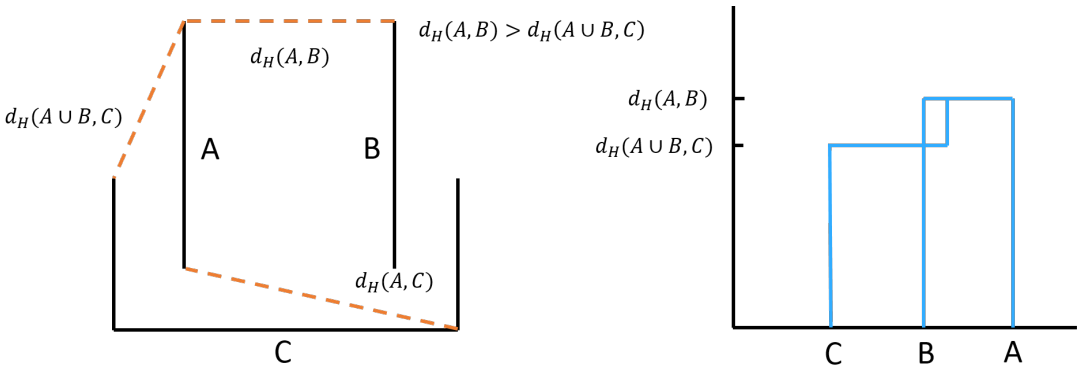


Figure 26: Single and Pearson's correlation result

Biograph of Communications sector

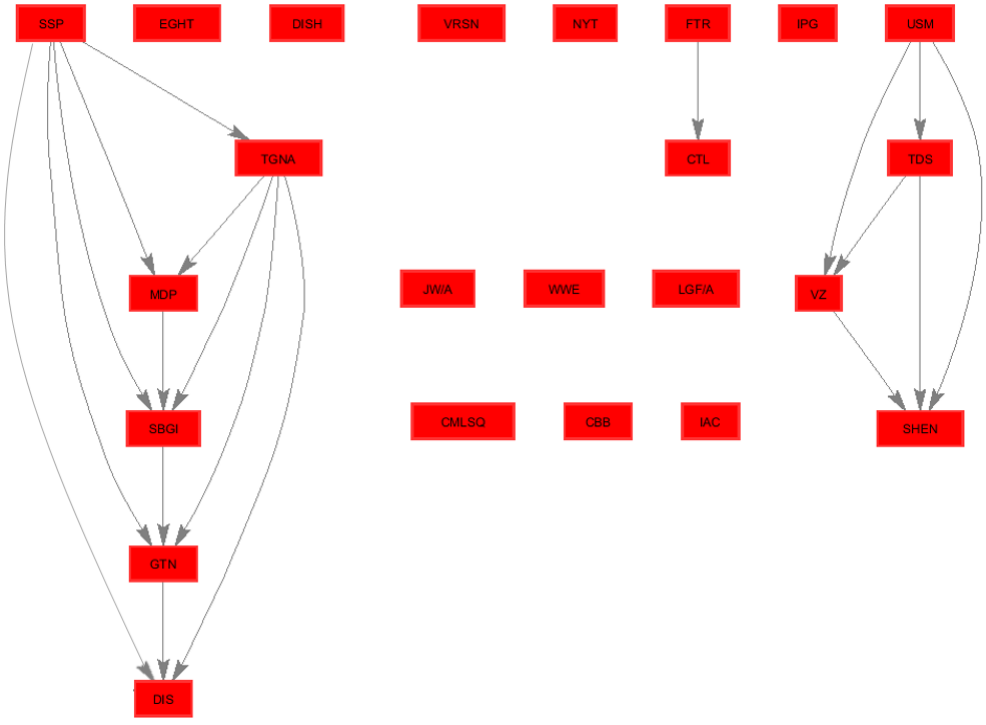


Figure 27: Biograph of Communications sector

References

[1] Riccardo Di Clemente and Guido L. Chiarotti. Supplementary Information andd Diversification versus specialization in complex ecosystems. *London Institute for Mathematical Sciences, 35a South St, Mayfair London United Kingdom.*

[2] Kakushadze Z, and Yu W. Statistical industry classification.

[3] Banz RW. The relationship between return and market value of common stocks. *Journal of Financials economics*, 1981 Mar 1;9(1):3-18.

[4] George Seif. The 5 Clustering Algorithms Data Scientists Need to Know.

[5] Dyer RH and Edmunds DE. From Real to Complex Analysis. *Springer*, 2014 May 14.

[6] Pearson K. LIII. On lines and planes of closest fit to systems of points in space. textitThe London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901 Nov 1;2(11):559-72.

[7] Galton, F. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 1886, 15:246–263.

[8] MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* , `Vol. 1, No. 14, pp. 281-297`.

[9] Lloyd S. Least squares quantization in PCM. *IEEE transactions on information theory*, 1982 Mar;28(2):129-37.

[10] Kakushadze, Z. and Yu, W. Statistical industry classification. `2016.`

[11] Shirali, S. and Vasudeva, Harkrishan L. Metric Spaces. `2006.`

[12] Andrea Trevino. Introduction to K-means Clustering. https://www.datascience.com/blog/K-means-clustering, `2016`

[13] Tumminello, M., Aste, T., Di Matteo, T., and Mantegna, R. N. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 2005 Jul 26;102(30):10421-6.

[14] West, D. B. An Introduction to Graph Theory. *Prentice–Hall, Englewood Cliffs, NJ*, 2001 pp. 247–281.

[15] Mantegna, R. N. and Stanley, H. E. An Introduction to Econophysics: Correlations and Complexity in Finance *Cambridge Univ. Press, Cambridge, U.K.*, pp. 120–128, 2000.

[16] Aste T, Di Matteo T, and Hyde ST. Complex networks on hyperbolic surfaces. *Physica A: Statistical Mechanics and its Applications.* 2005 Feb 1;346(1-2):20-6.

[17] Wilson, R. Introduction to graph theory (4rd ed.). *London: Longman*, 1985.

[18] Tumminello M, Lillo F, and Mantegna RN. Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior and Organization*, 2010 Jul 1;75(1):40-58.

[19] Basalto N, Bellotti R, De Carlo F, Facchi P, Pantaleo E and Pascazio S. Hausdorff clustering of financial time series. *Physica A: Statistical Mechanics and its Applications*, 2007 Jun 15;379(2):635-44.

[20] Munkres JR. Topology. *Prentice Hall*, 2000.

[21] Hausdorff, F. Grundzüge der Mengenlehre *von Veit, Leipzig, 1914*. [Republished as Set Theory, 5th ed. (Chelsea, New York, 2001).]

[22] Basalto N, Bellotti R, De Carlo F, Facchi P, Pantaleo E and Pascazio S. Hausdorff clustering. *Physical Review E*, 2008 Oct 28;78(4):046112.

[23] Sneath P and Sokal R. Numerical Taxonomy. *Freeman, London.*

[24] Rammal R, Toulouse G and Virasoro MA. Ultrametricity for physicists. *Reviews of Modern Physics*. 1986 Jul 1;58(3):765.

[25] Ruohonen K. Graph theory. 2013.

[26] Hauke J and Kossowski T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 2011 Jun 1;30(2):87-93.

[27] Fieller EC, Hartley HO and Pearson ES. Tests for rank correlation coefficients. *I. Biometrika*, 1957 Dec 1;44(3/4):470-81.

[28] Selim SZ and Ismail MA. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, 1984 Jan(1):81-7.

[29] Basalto N, Bellotti R, De Carlo F, Facchi P, Pantaleo E and Pascazio S. Hausdorff clustering. *Physical Review E*, 2008 Oct 28;78(4):046112.

[30] O'Shaughnessy JP. What works on Wall Street: A guide to the best-performing investment strategies of all time. *McGraw-Hill*, 1998 May.

[31] Faber M. Relative strength strategies for investing.

[32] Asness CS, Frazzini A, and Pedersen LH. Low-risk investing without industry bets. *Financial Analysts Journal*, 2014 Jul;70(4):24-41.