THESIS / THÈSE

MASTER IN BUSINESS ENGINEERING PROFESSIONAL FOCUS IN DATA SCIENCE

Graph Mining applied to Portfolio Management

Benchmarking of a Communicability Betweenness CentralityPortfolio Construction model

OHN, Adrien

Award date:
2023

Awarding institution:
University of Namur

Graph Mining applied to Portfolio Management:

Benchmarking of a Communicability Betweenness Centrality Portfolio Construction model

Adrien OHN

Directeur : Prof. Dr. J-Y. GNABO

Mémoire présenté en vue de l'obtention du titre de

Master 120 - Ingénieur de Gestion à
Finalité Spécialisée Data Science

ANNÉE ACADÉMIQUE : 2022-2023

Université de Namur, ASBL

Faculté des Sciences économiques, sociales et de gestion – Département des Sciences de gestion

Rempart de la Vierge 8, B-5000 Namur, Belgique, Tel. +32 [0]81 72 49 58/48 41

0. Abstract / Résumé

The purpose of this study is to verify the relevance and added value of implementing solutions from the field of data science, and more specifically graph mining, in the construction and management of portfolios. The approach adopted in this work is based on the concept of Communicability Betweenness Centrality introduced by Estrada et al. (2009). The research process documents the implementation of this methodology, the aim of which is to minimise impact propagation between entities observed over a given period on the basis of a distance correlation matrix. The performance of the CBC portfolio is then benchmarked against a panel of US large cap equity funds.

L'objectif de ce mémoire est de vérifier la pertinence et la valeur ajoutée de la mise en œuvre de solutions issues de la science des données, et plus particulièrement du graph mining, dans la construction et la gestion de portefeuilles. L'approche adoptée dans ce travail est basée sur le concept de Communicability Betweenness Centrality introduit par Estrada et al. (2009). Le processus de recherche documente la mise en œuvre de cette méthodologie dont l'objectif est de minimiser la propagation de l'impact entre des entités observées sur une période donnée sur la base d'une matrice de corrélation de distance. La performance du portefeuille CBC est ensuite comparée à un panel de fonds d'actions américaines de grande capitalisation.

Contents

1. Introduction

Over the last few decades, the field of portfolio management has undergone major changes, with the approaches used becoming increasingly modern and complex. Traditional approaches and theories of portfolio optimisation are often based on statistical models such as mean variance optimization, part of the modern portfolio theory developed by Markowitz (1952). These models often fail to capture the complexities and interdependencies between assets by assuming that asset returns are normally distributed. Michaud et al. (1989) states that they also tend to be over-reliant on estimates of expected returns. More recent years have seen the emergence of a new movement, involving the use of data science techniques applied to portfolio theory. This study seeks to be part of this trend and to contribute to this contemporary field dealing with the application of techniques from graph and network theory. This study aims to address these limitations by presenting a graph mining approche to portfolio management for weight optimization. Graph mining, a branch of data mining, focuses on analyzing and extracting meaningful patterns and relationships from complex network structures. By representing financial assets as vertices and their relationships as edges in a graph, we can capture the intricate dependencies and interactions among assets. The approach developed in this study is based on Estrada et al. (2009), his introduction of the Communicability Betweenness Centrality measure. This is a network-specific centrality measure that quantifies the extent to which a vertex (or graph) is involved in the circulation of flows within the network. There are, of course, other measures of communicability, centrality and betweenness. However, the latter allows not only the shortest paths between two vertices to be considered, but also alternative paths, which requires greater computational power. This method therefore allows the consideration of interconnections and interdependencies between assets, as well as repeating patterns. This approach has the potential to enhance the risk-return profile of investment portfolios, improve the robustness of the portfolios studied and potentially increase the overall performance of portfolio management strategies. By incorporating graph mining, one can overcome the limitations of traditional statistical models and provide a more comprehensive framework for portfolio weight optimization. The way these techniques can capture the underlying structure and dynamics of the market will be explored, facilitating the identification of influential assets within the network. Then, an optimization framework that integrates the insights gained from graph mining into the portfolio construction process will be developed. By incorporating the previously computed network-based metrics, the approach aims to enhance the diversification and risk management aspects of portfolio optimization. The results of this approach will then be benchmarked against a database containing the performance of a set of US equity large cap funds, which will make it possible to give meaning to and understand the potential dysfunctions of the approach implemented.

The study is organized as follows:

Section 2 provides a literature review tracing back the evolution of the most important studies in classical portfolio theory towards more advanced techniques, and the more recent evolution of the literature in graph theory and its financial applications. Section 3 describes the methodology used from a technical point of view to carry out this study. In this section, the research question is restated, this time in a more technical language. The key concepts of portfolio theory are discussed as well as their peers in graph theory, which are useful for understanding the approach developed. Then, the application to the research area is documented. This section concludes with the conversion of the metrics into optimized portfolio weights. Section 4 discusses the data used to conduct this study. Their selection, the cleaning and filtering process applied, descriptive statistics and an example of a network visualisation of a portfolio studied are presented. Section 5 presents the results of the algorithm used on the data presented in the previous section and discusses the performance of the algorithm used. Then, section 6 possible improvements, further implementations of the methodology and limitations of the study. The conclusion will bring this study to an end, recalling the research process, the results and the areas for development and improvement identified.

2. Litterature review

2.1 Foreword

This literature review will be structured as follows. To begin with, the evolution of the literary context of the subject out of which two main areas of research can be distinguished. The broad stream of literature relating to portfolio theory as well as the more recent field of graph theory and the application of graph mining methods to financial problems, although originating in a development by the mathematician Euler in the 18th century.

For this first component, the literature review will start from the first recognized articles of modern portfolio theory such as Portfolio selection Markowitz (1952). Going through the first articles on weight optimization, the notions of risk adjusted return and the importance of diversification in these portfolios. To reach more recent articles on the use of graph theory in such a context, by considering these portfolios as networks in which each constituent is seen as a linked node and its links to the other constituents, the arrests, can be attributed various phenomena.

2.2 Modern portfolio theory evolution

Modern portfolio theory as we know it today finds its foundations in Markowitz's “Portfolio selection” article. In this article, Markowitz introduces the concept of diversification and emphasizes the importance of considering both returns and associated risks when selecting investment opportunities. His approach to portfolio selection differed from the traditional approaches at that time because it was not solely focused on expected returns. Markowitz underlined the fact that a portfolio should not be considered as a simple aggregation of assets but as a whole. Indeed, Markowitz developed a mathematical model considering the expected return of the assets, their variance as well as their covariances. In doing so, Markowitz was able to demonstrate that diversification could reduce the overall risk of a portfolio without decreasing its returns. Markowitz also introduced the concept of the efficient frontier, representing the portfolios with the highest returns for each given risk level and vice versa. Mr. Markowitz's work has had a profound impact on the financial field as he defined key concepts that are still relevant nowadays.

10 years later, Sharpe (1964), through his paper "Capital asset prices", marked another significant step in portfolio theory by introducing the Capital Asset Pricing Model. A model that predicts the return on an asset as a function of the return on a risk-free asset, the market return and the sensitivity of the asset under analysis to the market. The introduction of the concept of systemic risk will follow from the use of this last variable. William Sharpe will build on this model the graphic concept of the security market line, another pillar of portfolio theory.

Brinson et al. (1986), in their paper "Determinants of portfolio performance", argued that the most important step in the investment process was the definition of the investment policy of a portfolio. They also argued that,

contrary to what was thought at the time, the definition of asset allocation policy contributed to more than 90% of the variation in the return of a portfolio, to the detriment of the prominent stages at the time, security selection and market timing. They then demonstrated that the same pattern also applied to portfolio risk variation. These two breakthroughs led them to the conclusion that investors should focus on strategic asset allocation and the definition of their risk tolerance to achieve their investment objectives. This study led to the development of a variety of portfolio construction methodologies such as target date funds, a fund aiming for optimal performance at a a key date and risk-parity strategies, which allocates capital in a portfolio according to the risk of the different assets.

A few years later, Michaud (1989) discussed Harry Markowtiz's work in his paper entitled "The Markowitz optimization enigma: Is "optimized" optimal?”. This critique is based on a series of simulations and empirical studies showing that the Markowitz approach often produced inefficient and unstable portfolios. According to the author, these shortcomings were due to the model's over-reliance on estimates of expected returns, variances and covariances of assets subject to significant uncertainty and bias. The model therefore tended to overestimate the performance of high-risk assets and underestimate the benefit of diversification. Michaud's work has had a big impact on the field of portfolio optimisation but has also triggered a trend towards more sophisticated and data-driven approaches which have improved the performance and stability of the sector as a whole.

The following years highlighted the beginning of increasingly advanced approaches in portfolio theory, notably marked by the work of Michaud and Ma (2001). They published the book "Efficient asset management: A practical guide to stock portfolio optimization and asset allocation". The later presents a portfolio construction framework based on their own optimisation methodologies, including Resampled Efficiency (RE). This approach uses a bootstrap resampling technique to generate a wide range of asset price movement scenarios based on real data and then calculates the efficient frontier for each of these scenarios. A diversified portfolio more robust to uncertainty and market volatility can then be constructed based on the set of efficient frontiers obtained. Michaud et al. (2010) will join David N. Esch to sign the paper "Portfolio monitoring in theory and practice" in which they introduce "2 new algorithms to overcome the sensitivity of the Michaud rebalancing rule to the likelihood of information overlap in the construction of optimal portfolios compared to current portfolios". These studies, carried out from the 1990s to 2010, marked the beginning of quantitative, automated, data-driven and refined techniques in portfolio management with the appearance of algorithms for simulating and modelling financial assets, automated weight optimisation, etc. The era of computational finance was definitely launched.

Meucci (2005) provides an overview of the latest techniques used in portfolio optimisation, covering areas such as risk management, factor models, mean-variance optimization, downside risk optimization and robust optimization. One of the important developments in this book is the introduction of robust techniques to reduce

the sensitivity of portfolios to input data estimation errors and the use of Bayesian networks. Meucci introduces the use of these Bayesian networks as a tool for modelling the dependence structures of asset returns. Indeed, these networks symbolise events that can impact on the evolution of an asset price, the possible associated outcomes and their probabilities. They allow to model the dependence between the factors impacting the return of a portfolio. This was an important development because it involved the use of theories from outside the traditional financial field.

Fabozzi (2007) described the classical theories of portfolio allocation and robust parameter estimation methods. He also discusses "the black litterman model to overcome the problems of parameter estimation required by classical portfolio theory" and "the Bayesian approach to overcome the assumption that expected returns are random and not fixed as required by the classical approach". He will then present several mathematical developments of robust portfolio optimisation techniques and will outline recent trends and upcoming directions of the field.

2.3 Graph Theory application to Portfolio Management

The article, "Solutio Problematis ad Geometriam Situs Pertinentis" Euler (1736) is a milestone in the history of mathematics and graph theory in particular. In this paper, Euler introduced the concept of graph theory and demonstrated its application to solving a problem related to the geometry of position. Euler's solution to the Königsberg Bridge problem involved the introduction of several key concepts in graph theory, including the definition of a graph as a set of vertices connected by edges, the concept of degree of a vertex, and the notion of path and circuit in a graph.

The academic movement promoting the application of graph theory to portfolio theory originated when Mantegna (1999) published "Hierarchical structure in financial markets", an article in which he investigates the daily time series of the logarithm of stock price and proposes a visualization in the form of a graph connecting the stocks of the analyzed portfolio based on the matrix of the correlation coefficients between each pair of stocks considering the synchronous time evolution of the difference of the logarithm of daily stock price”. This new portfolio modelling method aims to “overcome the main flaw that he attributes to traditional portfolio optimisation methods, their ignorance of the interdependencies between assets and their inability to capture the dynamic changes in correlations between them”.

Tumminello et al. (2010) developped quantitative methods to investigate the properties of correlation matrices. They were able to demonstrate that "hierarchical clustering was able to detect clusters of stocks belonging to the same sector or sub-sector of activity without the need for any supervision during the clustering procedure". These developments laid the foundations of portfolio visualisation as a graph. Clemente et al (2022) have suggested various methods for estimating advanced correlation graphs, such as shrinkage and weigthed depth methods. Kim et al (2019) introduced an approach for categorising relationships between observed stocks on

which to apply various network structure methods, the HATS model. The second major step was to use these new concepts for optimization purposes. Certain notions specific to graphs therefore appeared to be used alongside the notions specific to classical portfolio theory. Vyrost et al (2019) also investigated the application of graph-theoretic measures in portfolio optimisation strategies. They began by descriptively constructing four graphs in order to examine and propose a visualisation of the observed entities, a complete graph, a minimum spnanning tree, a maximal filtered plannair graph and a threshold significance graph. The vertices of the graphs are constructed on the basis of a distance representing the intensity of the correlation between two vertices and not on the basis of the distance correlation as is the case in this study. Then, the authors employ three measures of centrality on which their portfolio optimisation strategies are based: betweenness, which will be one of the components of the method implemented in this study, coupled with the notion of communicability, eigenvector, which is also discussed in the methodology section, and expected force. The authors will also assess the performance of strategies based on combinations of the three measures described. One of the first measures of centrality was the Katz centrality, Katz (1953), which, unlike the shortest path between two nodes, considers all the paths between these two nodes to measure the influence of a node in a network. Mayoral et al. (2022) will develop in their article "Using a hedging network to minimize portfolio risk" the application of centrality to "summarize how an asset behaves in relation to others in a network (hedging relations) and to itself (unhedgeable component)". This study aimed at holding stocks with the lowest centrality will result in "a lower portfolio variance than other traditional strategies such as stock selection by correlation coefficient, minimum-variance portfolio and naive strategies (EWP, etc.)" as well as a "number of stocks allowing to reach a given level of diversification lower than other traditional portfolio strategies" allowing to also reduce the transaction costs for the rebalancing of such a portfolio. Bloch et al. (2021) will produce a glossary of different measures of centrality based on the information they use about vertex positions and the way they weight this information.

These selected articles form the literary basis of the present work. The interaction of these two streams of literature, portfolio theory and graph theory, has a promising future ahead of them, as the interaction between them demonstrates their relevance and performance, whether in recent academic research or in concrete implementations within industry. They are indeed key methods in the context of financial markets, which have become increasingly complex in recent years, and which no longer necessarily correspond to the old macro-economic beliefs.

3. Methodology

3.1 Overview

This section will be the technical part of the study. Its purpose is to propose a quantitative model and a methodology capable of addressing the research problem, the improvement of the risk-adjusted return of a portfolio. The result verification of the approach will be found in the section entitled "empirical results". This section will be structured as follows. First, the research question and the purpose of the study will be technically defined, which will be the starting point of this methodology section. Second, the key concepts and notions of portfolio theory will be defined mathematically and literally in order to provide the reader with a sufficient knowledge base to understand the approach developped. Then, in the same way, key concepts of graph theory will be defined and discussed. Fewer people have been exposed to these concepts as their application to real life problems is more recent than the classical portfolio theory. Then, the proposed approach will be developed, step by step. Finally, the limitations and improvement areas will be discussed.

3.2 Research question and purpose of this study

The overarching purpose of this study is to demonstrate and verify the potential added value of applying methods and concepts specific to graph theory to the financial industry, and more specifically to portfolio management. To do this, a model capable of proposing an optimised allocation was chosen. This model is part of the mathematical branch of Graph Mining. The aim of this approach is to analyse and identify phenomena relating to the relationships between observed entities. The approach developed in this section will be benchmarked in the Results section, using a large panel of US large cap equity funds.

3.3 Key concepts definition of Portfolio Theory

In portfolio theory, two main concepts must be understood. These two fundamental concepts are the risk and return. The concept of return refers to the profit obtained by an investor from holding one or more assets over a given period of time. The concept of risk refers to the uncertainty associated to an asset value and its potential to lead to losses over a given period of time. The relationship between these two concepts is key to portfolio theory. Indeed, the theory states that an investor “may obtain a higher expected rate of return on his holdings only by incurring additional risk” to his portfolio (Sharpe, 1964). This notion is called the risk adjusted return.

One of the most widely used models developed on this tradeoff between risk and return is the CAPM proposed by William Sharpe. This model states that the expected return on an investment is directly proportional to the risk free rate plus a premium that reflects the riskiness of the asset. This premium is calculated as the difference between the expected market return on the asset and the risk free rate, the market risk premium, multiplied by beta, the sensitivity of the asset to market fluctuations.

$$E(R_i) = R_f + \beta_i(E(R_m) - R_f)$$

$$with\ \beta_i = \frac{Cov(R_i, R_m)}{Var(R_m)}$$

The other major development of these two concepts of risk and return, first discussed in the literature review section, are the achievements of a certain Harry Markowitz. Based on the principle that "investors can reduce risk and increase returns by diversifying their investments across a variety of assets", this study proposes statistical methods for constructing optimal portfolios, maximising returns for a given level of risk. The concept of efficient frontier is introduced as the “set of portfolios satisfying the condition that no other portfolio exists with a higher expected return for the same risk level”. Modern portfolio theory assumes that investors are risk averse and rational, thus, "an investor will not invest in a portfolio if another portfolio exists with the same risk-expected return profile".

In this framework, the return of a portfolio is defined as the weighted return of its constituent assets by their weight.

$$E(R_p) = \sum_{i=1}^{n} w_i\ E(R_i)$$

$with,$

$R_p =\ the\ return\ of\ the\ portfolio,$

$R_i\ =\ the\ return\ of\ the\ asset\ i,$

$w_i\ =\ the\ weight\ of\ the\ asset\ i\ in\ the\ portfolio\ \ (\sum_{i=1}^{n} w_i = 1)$

The risk of a portfolio is defined as the standard deviation of the variance of the returns of the constituent assets of the portfolio over a given period of time.

$$\sigma_p^2 = \sum_{i=1}^{n} w_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{m} w_i w_j\ \sigma_i \sigma_j \rho_{ij}$$

$with,$

$\sigma_i =\ the\ standard\ deviation\ of\ the\ periodic\ returns\ on\ asset\ i,$

$\rho_{ij} =\ the\ correlation\ coefficient\ between\ returns\ on\ asset\ i\ and\ j$

Portfolio return volatility can therefore be expressed as,

$$\sigma_p = \sqrt{\sigma_p^2}$$

Finally, having defined these two concepts, the following risk adjusted return measure, the Sharpe ratio, can be defined as,

$$S_a = \frac{E(R_p - R_f)}{\sigma_p}$$

$with,$

$E(R_i - R_f) = the\ expected\ excess\ return\ of\ the\ portfolio\ p\ compared\ to\ the\ risk\ free\ rate$

$\sigma_p = the\ standard\ deviation\ of\ the\ portfolio$

The Efficient Frontier is then found by minimizing the following matrix form expression,

$$W^T \sum w - q\ x\ R^T w$$

$with,$

$w = the\ portfolio\ weights\ vector\ (sum\ of\ which = 1)$

$R^T w = the\ portfolio\ expected\ return$

$W^T \sum w = the\ variance\ of\ the\ portfolio$

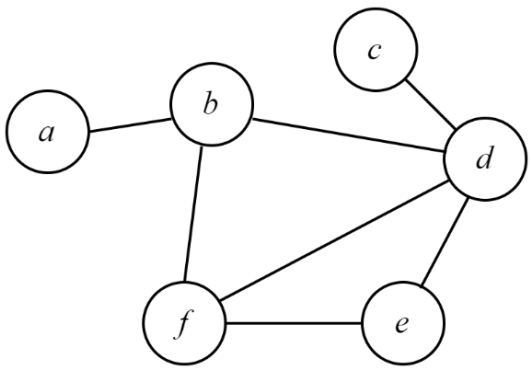
$q = the\ risk\ tolerance\ factor\ (\geq 0)$

3.4 Key concepts definition of Graph Theory

In this section will be discussed the fundamental concepts of graph theory. Graph theory is a branch of mathematics that allows the analysis of data in the form of graphs, consisting of nodes connected or not by edges. This field of mathematics allows the study of the properties of graphs as well as the relationships between these data. It was introduced at the beginning of the 18th century by Leonhard Euler in which he solved the famous Seven Bridges of Konigsberg problem. Nowadays, this theory has much wider applications, whether in the analysis of social networks, communications, transport, but also in physics and biology. A graph, being a visualisation of a set of related data, can be schematised by an adjacency matrix. This matrix indicates whether a pair of nodes are directly connected, adjacent, or not in a graph. For a simple, loop-free, weightless, undirected graph, taking the edges $E = \{e_1, e_2, \dots, e_n\}$, the $n\ x\ n$ adjacency matrix A is such that the element a_{ij} takes the value 1 when vertices i and j are adjacent, connected by an edge, and takes the value 0 otherwise. The following adjacency matrix,

$$A = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Will have the following associated graph,



There are various subfields in graph theory such as shortest path search, connectivity analysis, minimal colouring, spanning trees and determining the planarity of a graph. This analysis will be focused on the concept of vertex centrality. The centrality of a vertex can be measured in several more or less complex ways.

The first and most trivial measure of centrality is the measure of the degrees of the vertices in a graph. In graph theory, the degree of a vertex corresponds to the number of edges incident to that vertex and is denoted, for

vertex *d*, *deg(d)*. A potential approach to measure the centrality of the vertices of a graph is therefore the analysis of their degrees with respect to the minimum degree of a graph, noted $\delta(G)$, as well as the maximum degree of a graph, noted $\Delta(G)$ for the graph G. These degrees are represented by the degree matrix, defined as,

$$D_{i,j} = \begin{cases} deg(v_i) & if\ i = j \\ 0 & otherwise \end{cases}$$

Using the graph specified above, the associated degree matrix is therefore,

$$D = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{bmatrix} \end{array}$$

After having identified a situation by these 2 matrices, of adjacency and of degrees, it is possible to obtain the Laplacian matrix. The latter is obtained by subtracting the adjacency matrix from the degree matrix in the following way,

$$L = D - A$$

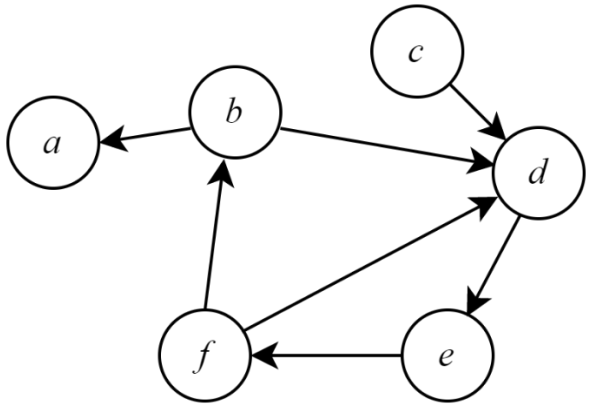
The Laplacian matrix associated to the example graph is therefore the following,

$$L = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 3 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 4 & 1 & 1 \\ 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 1 & 0 & 1 & 1 & 3 \end{bmatrix} \end{array}$$

This matrix, resulting from the subtraction of the two previously described ones, is an important step in the representation of a situation in the form of a graph, as it is the starting point for the analysis of important graph properties. Indeed, this Laplacian matrix allows the search for spanning trees, sparsest cuts as well as eigenvectors and eigenvalues, defined later, via the Fourrier transformation which eigendecomposes the Laplacian matrix.

The Eigenvector centrality measures the importance of a vertex, in a score-weighted vertex network, by edges adjacent to other vertices with a high score. A high Eigenvector score therefore means that a vertex is connected to other high-scoring vertices. A well-known application of this centrality measure is the Pagerank algorithm, founded by Larry Page in the early 2000s and used by the company Google for their search engine to measure the relative importance of web pages. This algorithm is based on the Eigenvector centrality with the difference

that it adds the notion of direction of an edge. A vertex to which many vertices point via their edge is considered influent. A smaller importance is given to the vertices pointing to this influent vertex. Thus, in the above graph, vertex *d* will be considered as influential since many vertices point to it while vertex *c* will only be given a low score,



Another important measure of centrality in a graph is called betweenness centrality. This measure analyses the shortest paths between each pair of vertices to determine which are the most used by these paths. A vertex located on a large number of shortest paths will be considered more important than a vertex located on few shortestpaths. The betweenness centrality of the vertex *b* is as follows,

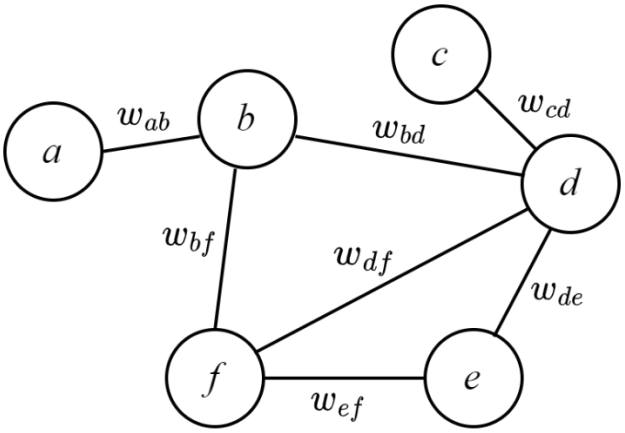
$$g(b) = \sum_{y \neq b \neq z} \frac{\sigma_{yz}(b)}{\sigma_{yz}}$$

$with,$

$y, z \ = any\ other\ vertex\ in\ the\ graph\ than\ b$

$\sigma_{yz}(b) \ = \ total\ number\ of\ shortest\ path\ between\ y\ and\ z\ passing\ throught\ b$

$\sigma_{yz} = the\ total\ number\ of\ shortest\ path\ between\ y\ and\ z$



This measure can be extended to weighted graphs, graphs for which a value is associated with each edge. In the case of weighted graphs, the length of a path will no longer be considered simply as the number of vertices it contains between its two ends but as the sum of the values associated with the edges it contains between its two ends. Thus, a shortest path between two vertices will be the one that minimises the weight of the edges on its total length. The influence of a vertex is therefore measured by the sum of the values associated with the edges that are incident to it,

$$s(x) = \sum_{j=1}^{N} a_{ij}\, w_{ij}$$

After having defined the notion of centrality of a vertex in a network, the notion of communicability will now be addressed. Communicability within a network can be defined as the analysis of the propagation of phemonenon within a network. This concept is therefore linked to the notions described above because intuitively the phenomenon analysed should be more inclined to propagate through the shortest paths and vertices identified as central. Estrada et al. (2008) proposed a method to measure this communicability in a paper entitled "Communicability in complex networks". This method is no longer based solely on the shortest paths, as the assumption that these were the only paths taken during a propagation phenomenon within a network is eroded. They proposed a generalisation based on the shortest path phenomenon, taking into account all the paths that connect a vertex a to a vertex b. Their approach was to give a lower degree of communicability to paths identified as long to get from one vertex to another and a higher degree of communicability to the shortest paths between two vertices. The Communicability as defined by Estrada et al. (2008) between two vertices *p* and *q* can be obtained, according to their approach as,

$$G_{pq} = \frac{1}{s!}\, P_{pq} + \sum_{k>s} \frac{1}{k!}\, W_{pq}^{(k)}$$

$with,$

$P_{pq} = \ the\ number\ of\ shortest\ path\ between\ nodes\ p\ and\ q\ having\ a\ length\ of\ s$

$W_{pq}^{(k)} = \ the\ number\ of\ walks\ connecting\ p\ and\ q\ of\ length\ k\ > \ s$

Using the connection between the powers of the adjacency matrix and the number of walks in the matrix graph, this formula can be rewritten as,

$$G_{pq} = \sum_{j=1}^{n} \varphi_j(p)\ \varphi_j(q)\ e^{\lambda_j}$$

$with,$

$\varphi_j(p) = the p^{th} element of the j^{th} orthonormal\ eigenvector\ of\ the\ adjcency\ matrix,$
$associated\ with\ the\ eigenvalue\ \lambda$

This Communicability measure will serve as the foundation on which Estrada et al. (2009) will introduce the measure on which this research will focus, Communicability Betweenness centrality, defined in the following section.

3.5 Application to research field

In this section, the concepts discussed and defined above will be combined to propose an approach to answer the research question. The aim of this theory application to a concrete case is to verify the potential benefits of using Graph Mining methods to propose an asset allocation in a portfolio.

The first step of the approach is therefore to determine the intensity of the relationships between the assets observed. The method chosen to measure these relationships between stocks contained in a portfolio is the distance correlation. Indeed, the Pearson correlation often preferred has certain limitations. According to Edelmann et al. (2021), distance correlation is not suffering from the disadvantages of Pearson correlation. It can measure both linear and non-linear associations between two random variables, unlike the Pearson correlation which only captures linear ones. The second major advantage of this method of calculating the correlation is that, unlike the Pearson correlation, it is able to take into account time series of different dimensions.

Pearson correlation,

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\ Var(Y)}}$$

$with,$

$Cov(X,Y) = covariance\ of\ X\ and\ Y$

$Var(X) = variance\ of\ X$

$Cor(X,Y) \in [-1,1]$

Distance correlation,

$$dCor(X,Y) = \frac{Cov(X,Y)}{\sqrt{dCov(X,X)\ dCov(Y,Y)}}$$

$with,$

$dCor(X,Y) \in [0,1]$

The correlation distance can thus be calculated between the returns of each asset one by one contained in the observed entities, in order to obtain the distance correlation matrix of the portfolio. In the case of a portfolio of 3 assets, the distance correlation matrix will be as follows,

$$C = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 1 & dCor(A,B) & dCor(A,C) \\ dCor(A,B) & 1 & dCor(B,C) \\ dCor(A,C) & dCor(B,C) & 1 \end{bmatrix} \end{array}$$

$with,$

$dCor(A,B) = distance\ correlation\ coefficient\ between\ asset\ A\ and\ asset\ B$

The second step of the approach is the construction of the distance correlation network between the assets. The vertices of the network will therefore represent the assets contained in the portfolio and the edges incident to these vertices will be weighted according to the cross-correlation between these two asset price movements. Once this network is constructed, the centrality measures defined in the previous section will be applied to it in order to obtain optimized asset weights.

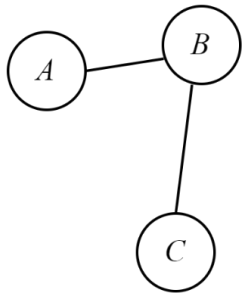
In their research, “A network perspective of the stock market”, published in the Journal of Empirical Finance in 2010, Tse et al. proposes an approach that allows the construction of complex networks to study the correlation of the evolution of the prices of several assets. This study made it possible to assess stocks interdependence, over two periods of about two years. In particular, it showed that "the variation in stock prices is strongly influenced by a relatively small number of stocks". The winner-take-all approach is proposed to establish the edges of the network. It consists of defining a threshold value under which two vertices are considered not connected by an edge. These vertices will be connected if the value associated with their relation is higher than the threshold value.

$$dCor_{ij} = \begin{cases} dCor(X_i, Y_j), & dCor > \rho_c \\ 0, & otherwise \end{cases}$$

$with,$

$$\rho_c = defined\ distance\ correlation\ threshold$$

This method will allow us to visualize, in the form of a network, the identified distance correlation between assets in the following way,



The above distance correlation network would therefore be associated with a three assets portfolio such that,

$$\begin{cases} dCor(A, B) > \rho_c \\ dCor(B, C) > \rho_c \\ dCor(A, C) < \rho_c \end{cases}$$

Then, once the network graph is constructed on the basis of the distance correlation matrix, the purpose of the approach is now to identify an optimized asset allocation according to the detected relationships between the assets of the portfolio. The goal of the approach is to associate a lower weight to the most inter-connected assets and a higher weight to the less inter-connected ones in order to create a portfolio more robust to volatility transmission. At this stage, the intensity of the relationships between assets is quantified but no information is available on the propagation of impact through the portfolio. We will now move on to the concept on which this study is based, the concept of Communicability Betweenness centrality, by Estrada et al. (2009).

To begin with, the adjacency matrix resulting from the distance correlation matrix via the winner-take-all method is as follows,

$$A_{p,q} = \begin{cases} 1 & if\ dCor(P, Q) > \rho_c \\ 0, & otherwise \end{cases}$$

The communicability between vertices p and q is,

$$G_{pq} = (\exp A)_{pq}$$

Therefore, the number of paths passing through vertex r is,

$$G_{prq} = (\exp A)_{pq} - (\exp(A - E(r)))_{pq}$$

The relative risk associated with an asset r is,

$$Relative\ Risk\ of\ Asset\ r = \frac{\omega_r}{\sum_{r'=1}^{N} \omega_{r'}}$$

Where ω_r, the Communicability Betweenness centrality of the vertex r is,

$$\omega_r = \frac{1}{C} \sum_p \sum_q \frac{G_{prq}}{G_{pq}}$$

The meaning attributed to the Communicability Betweenness Centrality measure in this study is therefore the following,

$$CBC = \frac{sum\ of\ all\ weigthed\ paths\ involving\ vertex\ r}{sum\ of\ all\ weigthed\ paths\ involving\ every\ vertex}$$

This measure can indeed predict how an impact, in this case, volatility, may propagate through the asset distance correlation network. The weights can then be optimised to limit the effects of an impact within the portfolio. The strategy developed will therefore allocate significant weight to assets that are poorly correlated with others and that have little capacity to spread in the event of an impact.

3.6 Translate the CBC measure to asset weights

The translation of the Communicability Betweenness Centrality of each vertex into a weight in the portfolio is the final step of the developed approach. This conversion can be done in several ways, correcting the distribution more or less severely depending on the criterion identified previously.

The CBC measures are initially reversed in order to penalise stocks with a high CBC and prioritise those with a lower CBC.

$$w_r = \frac{1}{\omega_r}$$

$with,$

$w_r = the\ final\ weight\ of\ stock\ r\ in\ the\ CBC\ optimized\ portfolio$

Then, in order to homogenise the distribution of asset weights, the recently inverted CBCs are normalised as follows,

$$\omega_r = \frac{\omega_r}{\sum \omega_G}$$

$with,$

$\sum \omega_G = the\ CBC\ sum\ of\ every\ vertex\ in\ graph\ G$

The capital allocation proportion for each stock is thus obtained and only needs to be converted into a number of discrete shares. This is an allocation optimisation problem in which the aim is to allocate a total value of shares for each stock that is as close and as consistent as possible to its weighting as determined by the model. A greedy algorithm is applied for this purpose. This algorithm will first allocate a number of shares that is less than the specified weight of each asset, rounding down. The algorithm will then iterate in several rounds to fill the gaps, sorting them by order of magnitude and prioritising the gaps that, once filled by an additional action, will have the least impact on the weight determined by the model.

4. Data

This part of the study deals with the data used to carry out this research. First, the sources of the data and the architecture of the databases will be discussed. Secondly, the data collection method will be explained. The data processing section will then deal with the processing of the data. This will be followed by a brief descriptive study of the data, drawing on the concepts discussed in the methodology section.

4.1 Data sources & architecture

Two levels of data are to be distinguished in this study. The first level corresponds to funds. The second level corresponds to the underlying assets, the stocks.

- Funds database

The complete funds database contains the performance and asset weights of 5,219 US equity funds. These US equity funds are divided into several categories according to the nature of their underlying assets and the sector they are focused on. First, large cap funds, which invest in large-capitalisation companies. These companies have often been listed on the market for a long time, have acquired a significant reputation and are recognised as leaders in their sectors and/or have grown recently and rapidly, and are seen by the markets as having a promising future. Then there are small cap funds, investing in small-cap companies.These companies with smaller capitalisations are often seen as underdogs in their respective sectors. They tend to have smaller market shares, are less liquid, and often offer a higher potential return as well as a higher risk for an investor. In between are the mid cap funds. Then there are utilities funds, which invest in companies active specifically in certain sectors, generally around energy and raw materials, whether as producers and/or distributors. This work will focus on the first category, US large cap equity funds, primarily for reasons of data availability and validity in order to propose a development that is as universal as possible. A potential area of development for this research would be to verify the performance of the approach analysed in other investment universes. There are 1,652 large cap. funds within the population. From this group of funds, only those covering the entire period of analysis, from the beginning of 2012 to the end of 2018, have been retained in order to guarantee the widest possible research window, common to each selected fund. The final panel of funds thus brings together the performance and weightings of 920 entities over the period 2012 to 2018.

- Stocks database

The second level flows from this first level of data used in the research. Indeed, the first step in the research process aimed at benchmarking a portfolio construction model is to determine a potential investment universe on which to train and then verify the model's performance. This universe must be as representative as possible of the entities being measured in order to guarantee the relevance of the model's benchmarking.

It was therefore decided to scan the underlying assets of the funds selected in the initial database in order to derive an initial population of assets. From these 57,179 instruments, sorted by frequency of occurence in the funds, we deducted all instruments that did not correspond to the desired investment universe for the rest of the research process (cash, treasury bills, derivatives, etc). The 200 most represented stocks were then selected from this list. Of these 200 stocks, the daily close price could be recovered for 174 and of these 174, 14 were dropped, in the identical process as for the funds, since their data was not available for the entire research period (2012-2018). The final population is therefore made up of 160 US large cap. stocks. The data selection process detailed above is shown in Figure 1. The database used during the model training and validation process is therefore based on the daily close prices of the 160 stocks over the research period, between 1st january of 2012 to the 31 december of 2018. The final stock sample is shown in Figure 2.

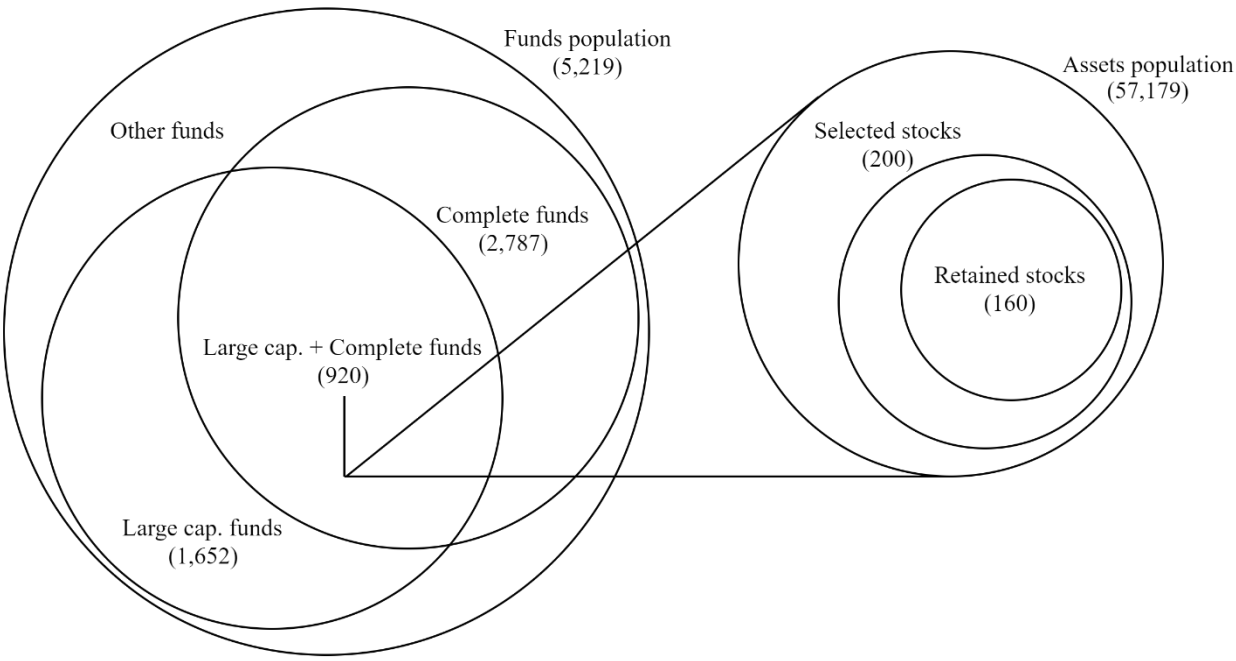


Figure 1: Data architecture involved

A	BAX	COST	EW	ITW	MDT	PEP	TJX
AAPL	BIIB	CRM	F	JNJ	MET	PFE	TMO
ABC	BK	CSCO	FDX	JPM	MMM	PG	TXN
ABT	BKR	CSIQ	FITB	KEY	MO	PM	UAL
ADI	BLK	CSX	GD	KMB	MPC	PNC	UNH
AET	BMY	CTSH	GE	KO	MRK	PPG	UNP
AFL	BRK/A	CVS	GILD	KR	MRO	PRU	UPS
AGN	BWA	CVX	GIS	LLY	MS	PVH	USB

AIG	C	DAL	GLW	LMT	MSFT	PXD	V
ALL	CAH	DE	GM	LOW	MU	QCOM	VFC
AMAT	CAT	DFS	GOOGL	LRCX	NEE	ROST	VLO
AMGN	CB	DG	GS	LUMN	NKE	RTX	VRTX
AMP	CF	DHR	HAL	LUV	NOC	SBUX	VZ
AMT	CI	DIS	HD	LVS	NOV	SCHW	WBA
AMZN	CL	DVN	HES	LYB	NSC	SHW	WDC
ATVI	CMCSA	EA	HON	M	NVDA	SLB	WFC
AXP	CME	EBAY	IBM	MA	ORCL	STT	WMB
AZO	CMI	EMR	ICE	MCD	ORLY	SWK	WMT
BA	COF	EOG	INTC	MCK	OXY	T	XOM
BAC	COP	ETN	INTU	MDLZ	PARA	TEL	YUM

Figure 2: Selected investment universe

4.2 Data processing framework

The initial database provided by the academic team in charge of supervising this research, listing the funds, their weighting by quarter and their monthly performance, was retrieved from the Morningstar platform. The construction of the second level database, concerning the daily close prices of the stocks selected, was built on the basis of data extracted from Bloomberg. These databases were stored in the form of Excel files.

All the operations involved in this research were carried out in Python, using the Visual Studio Code integrated development environment (IDE). The main class and method libraries used in this research were as follows:

- pandas: data extraction, manipulation, dynamic storage in specific structures.
- numpy: data structures, fast-paced quantitative methods.
- networkx: graph construction, transformation, analysis and exploration tools.
- matplotlib / seaborn: data visualisation, representation.
- scipy: scientific / technical computing, optimization algorithms.
- pypfopt: portfolio construction, allocation, optimization methods.

The methods provided by these libraries have occasionally been rewritten to suit the research process. These methods have made it possible to carry out certain stages of the research approach that require a significant amount of computing power (database extraction, calculation of distance correlations, graph construction, model training/validation), which will be a potential limitation in the chapter discussing them specifically.

4.3 Preliminary data visualisation

The starting point of the research process, in accordance with the methodology section, is the quantification of the relationships between the observed entities. The distance correlation was thus calculated one by one between each stock's daily close price to obtain a final matrix displayed in Figure 3.

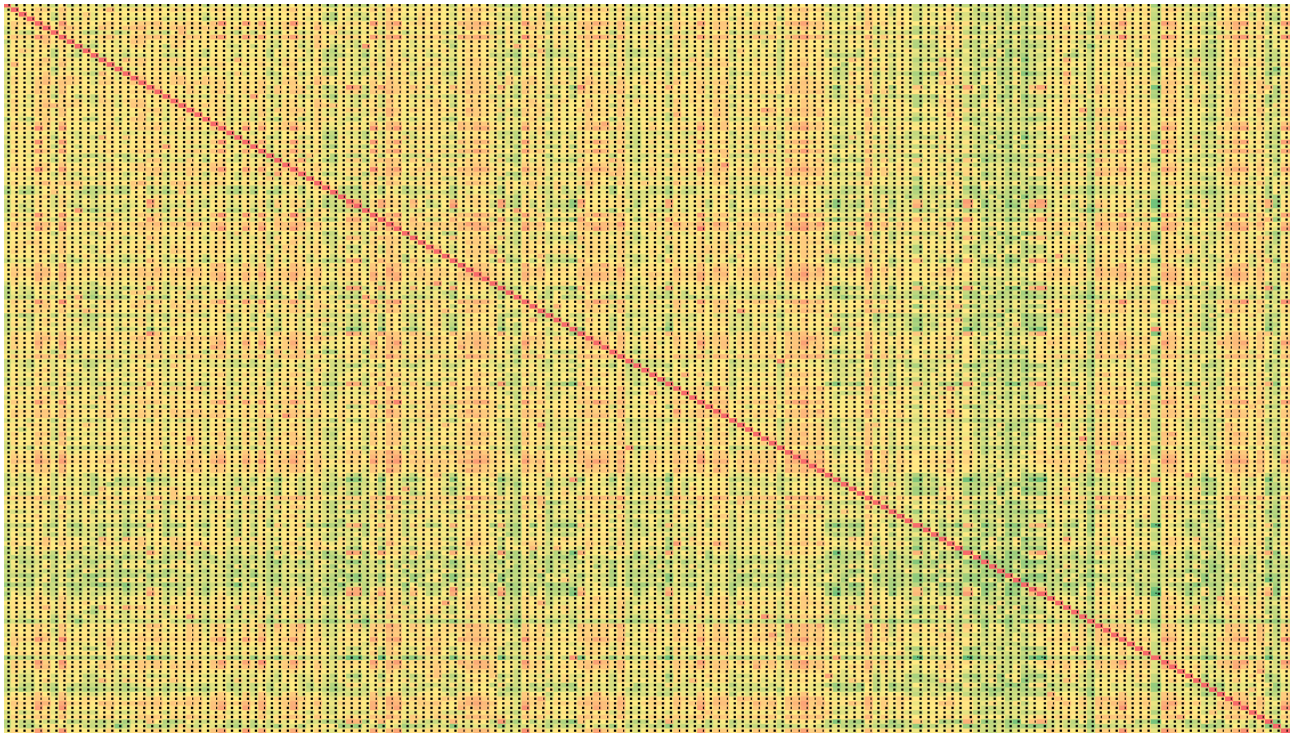


Figure 3: Distance correlation matrix heatmap

A number of observations can be drawn intuitively from this visualisation. The set of entities observed appears globally positively correlated, which corresponds to the nature of the universe of entities selected. As this is a universe of US large cap stocks, it is not surprising that the majority of them follow broadly similar movements and respond to certain events and market parameters in a uniform manner. Secondly, it is possible to observe that the visualisation is not perfectly homogeneous and that it includes a number of singular relationships that are synonyms for non-correlated or even negatively correlated observed entities.

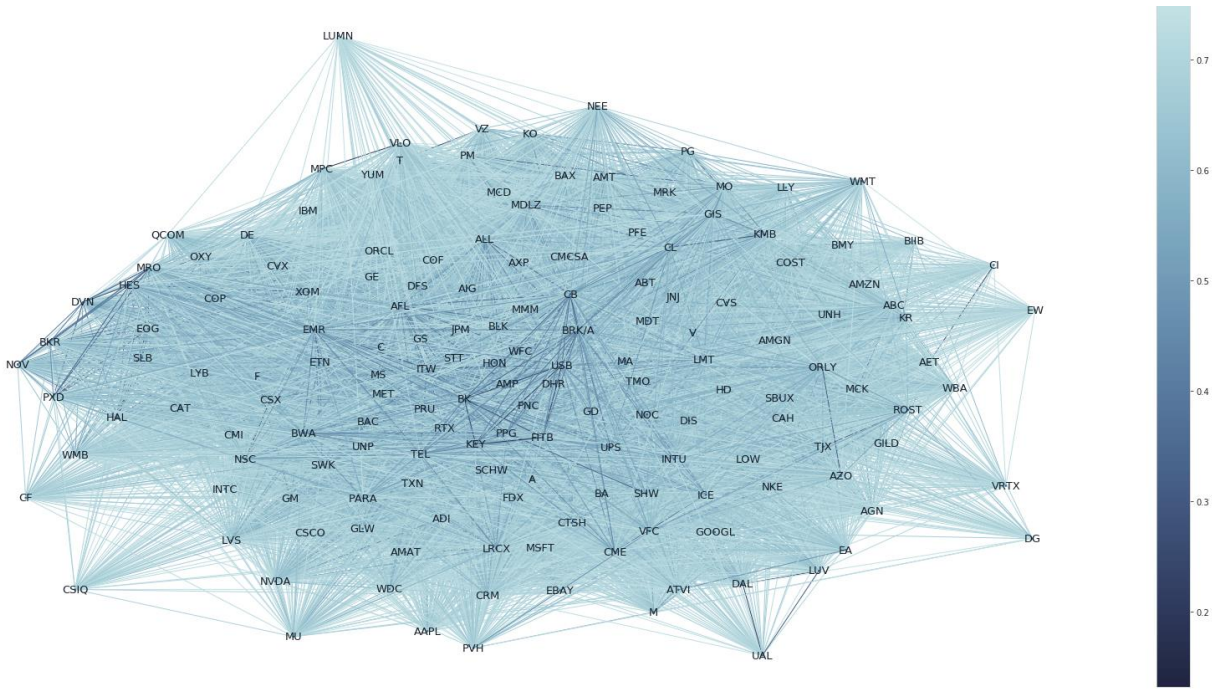


Figure 4: Network Graph visualisation

Based on this distance correlation matrix between observed stocks, it is possible to construct the following network graph as displayed in Figure 4. This network is constructed according to the method explained in the methodology section and therefore considers that two observed entities (vertices) are linked by an edge only if the distance correlation between them is greater than the defined threshold. The decision was made during the research process to define this threshold as being equal to 0.25 for distance correlations between -1 and 1. It is important to take into account the context of the entities observed when setting this parameter, as it has a significant impact on the rest of the research process. Considering the globally and homogeneously positive relationship of the observed entities, a threshold of 0.25 seems acceptable. A threshold that is too tight would potentially pollute the model's performance, while a threshold too excessive would deprive it of important information and blindfold the assessment of inherent risks in the next stage of the research process. As explained in the section on the key concepts of graph theory, the first measure of the centrality of a vertex within a network is its degree, the number of edges connecting it to other vertices. The degree distribution for the previous graph can be seen in Figure 5. This distribution is left skewed, meaning that the median is higher than the mean due to the presence of extreme values on the left of the histogram. These extreme values on the left are therefore the effect of stocks that are less correlated to the others as a whole, whose distance correlations more rarely exceed the threshold set previously. On the other hand, the mode is located to the right of the mean and median, indicating the highest distribution frequency (0.0783). These properties confirm the visual intuitions raised when visualising the distance correlation matrix.

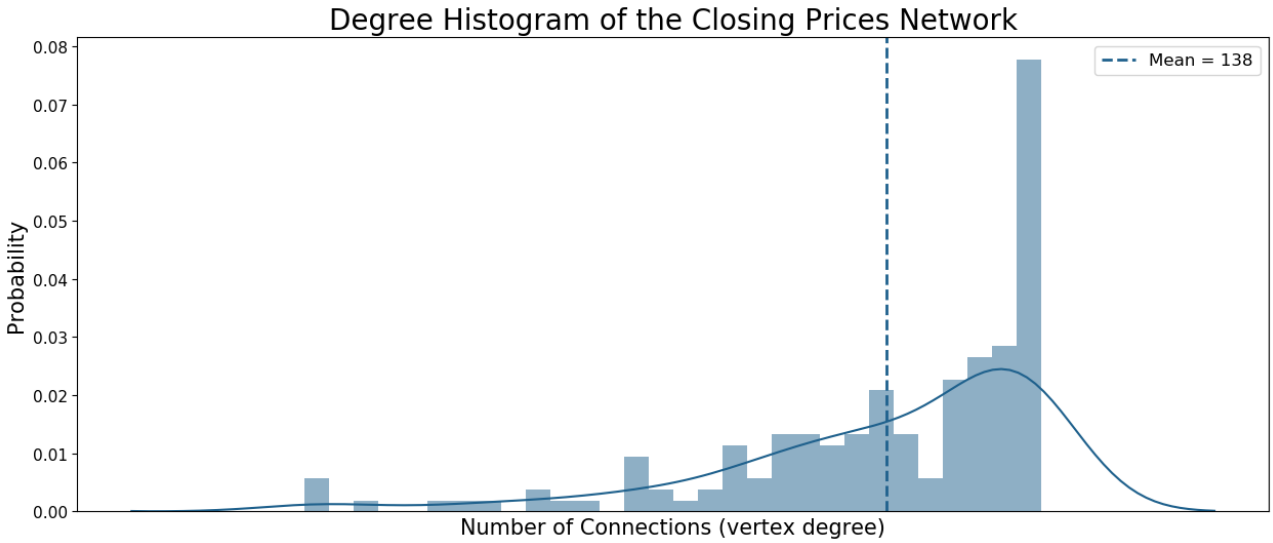


Figure 5: Network Graph degree histogram

5. Results

This section will focus on the outcomes of the research process. As a reminder, the purpose of this study is to implement a graph mining model in the field of portfolio management and to benchmark it against the industry. The results of the model will first be discussed and explained, and then benchmarked against the panel described in the Data section.

5.1 Model outcomes

This section focuses on the outputs of the model on the previously selected investment universe. Starting from the step described previously, the representation of the distance correlation matrix in the form of a network graph provides a visualisation of the interdependency relationships between the various assets observed over the sample period. Nevertheless, the aim of the study is not a static and retroactive analysis of relationships between entities over a given period, but rather to suggest a model for quantifying the risk of interdependence between assets in the first instance, and to recommend an allocation that takes into account this developed measure in a second phase. Using the method of measuring Communicability Betweenness centrality as defined by Estrada et al., the portfolio risk contribution of an asset is defined as its share in the sum of the portfolio risk contributions of all the assets held. These contributions to risk are presented in Figure 7. On this basis, the second step after running the model is to propose a capital allocation set between the different assets belonging to the investment universe. The aim is to create an allocation which is structured to be the most robust possible, by favouring assets whose behaviour is judged to be less susceptible to influence, and on which it would be more difficult for impacts affecting the market to propagate. In this way, capital is inversely allocated to the assets' contribution to portfolio risk. In order to obtain the individual capital allocations of the various assets, their Communicability Betweenness centrality is inverted and then normalised. These

individual capital allocations are shown in Figure 8 of this section. The performance of a portfolio cannot be considered simply as the sum of its weights multiplied by the performance of the individual assets. This is why the ultimate step is to materialise these individual allocations by converting them into shares. A greedy approximation algorithm is used to obtain an integer number of shares for each asset that matches the determined allocations as closely as possible. Once this process has been completed, it is now possible to assess the performance of the portfolio built on the 2nd set of data, the validation period, running from 2016 to 2018.

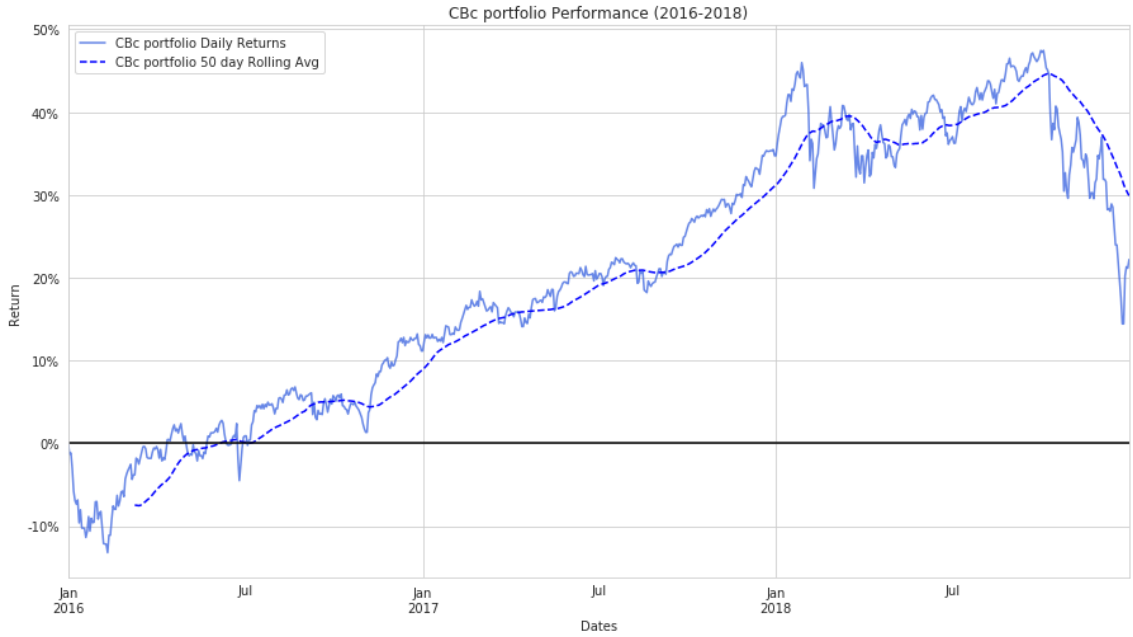


Figure 6: CBC portfolio performance (2016-2018)

Figure 6 shows the performance of the CBC portfolio over 2016, 2017 and 2018. It can be seen that the portfolio fell at the start of 2016 but then performed well until the end of 2017 before having a difficult year in 2018. The following section will provide a better understanding and justification of these phenomena by benchmarking this performance against the panel of US equity large cap funds.

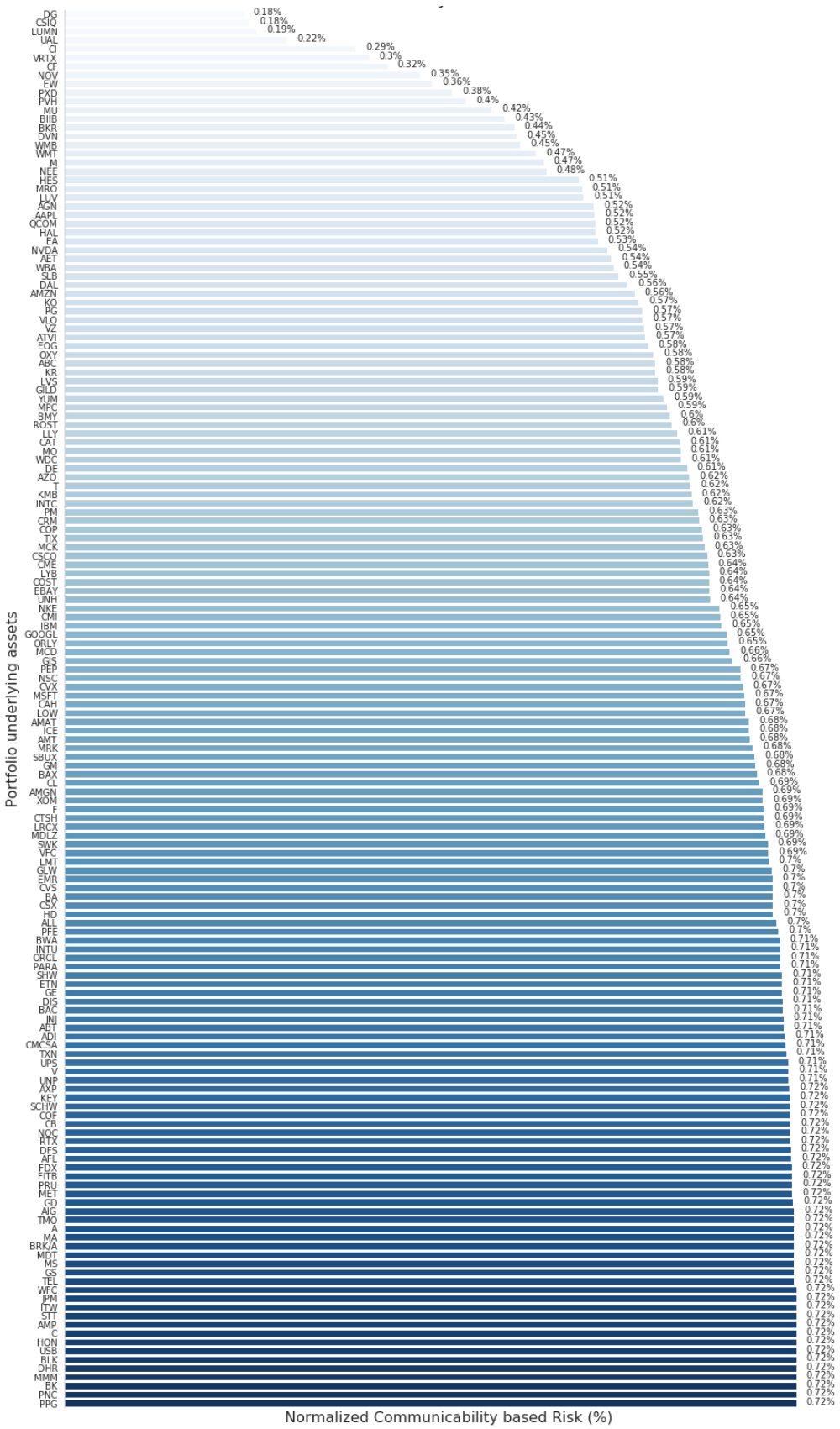


Figure 7: Assets' contribution to CBC portfolio risk (2012-2016)

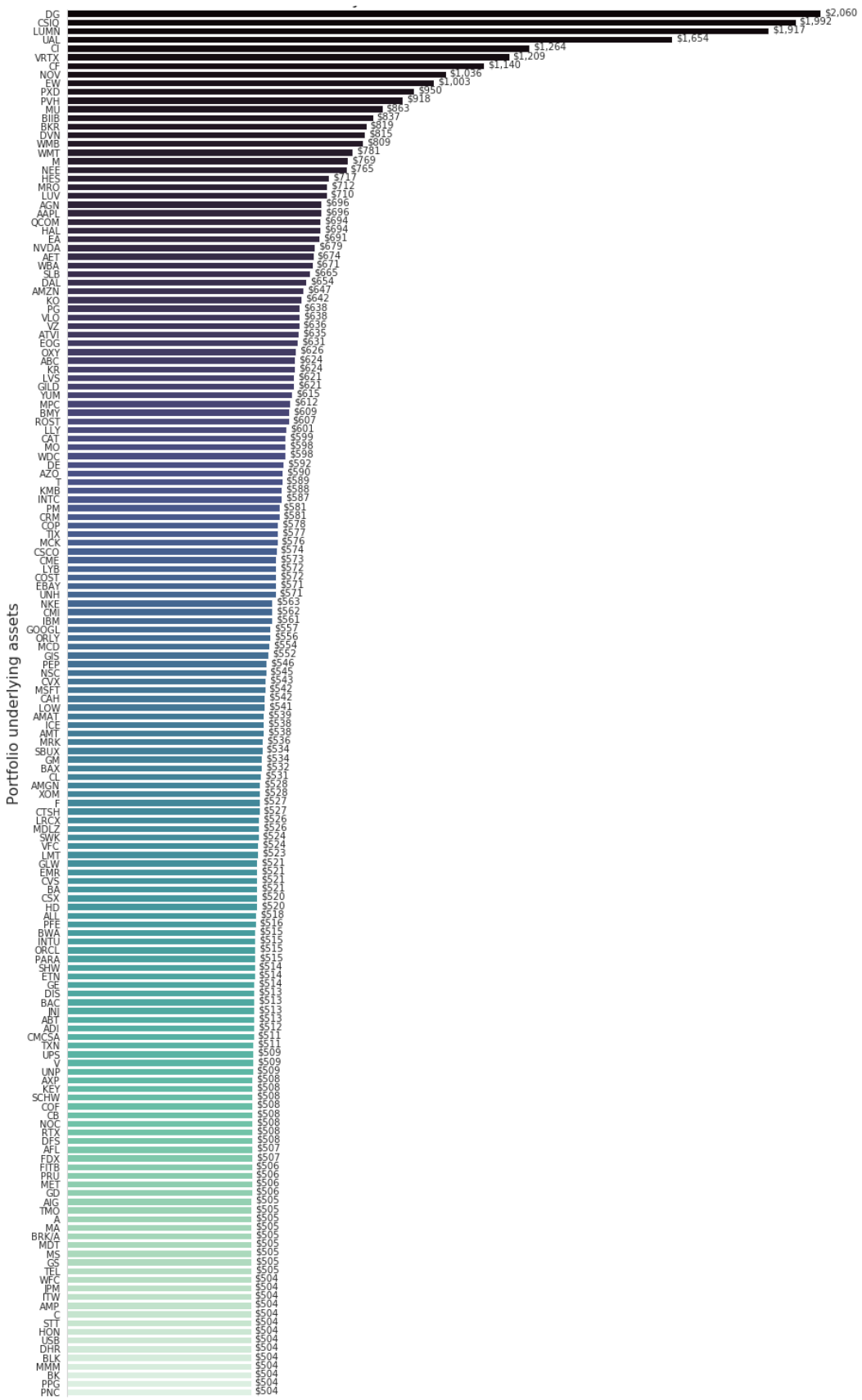


Figure 8: CBC portfolio final capital allocation (USD 100.000)

5.2 Model benchmarking

In this section, we benchmark the performance of the CBC portfolio during the test years, from 2016 to 2018, against the performance of the panel of US equity large cap funds. First of all, it is worth pointing out, in accordance with the section on Data, that the funds covering the entire period only have been retained in order to be able to make a full comparison. This choice in setting up the benchmarking panel could potentially introduce a survivorship bias into the data relating to the funds. Indeed, it can be assumed that funds that have disappeared may have had bad years prior to their disappearance. The performance of entities that have passed the selection filter applied is therefore potentially slightly overestimated in this sense.

	Years	2016	2017	2018	Total
CBC-portfolio	Return	11.34%	23.67%	-11.86%	21.14%
	STD (daily)	12.72%	7.31%	14.29%	11,82%
US Funds	Return Q1	5.87%	17.94%	-8.36%	15.84%
	Return Q2	10.59%	21.28%	-5.68%	26.51%
	Return Q3	14.28%	26.57%	-1.94%	41.83%
	STD (monthly)	3.4%*	1.4%*	4.6%*	3.4%*
SP500	Return	9.54%	19.42%	-6.24%	22.64%

Figure 9: Performance comparison CBC portfolio – US funds panel

The performance comparison between the CBC portfolio and the panel of US equity large cap funds is available in Figure 9. The performance of the CBC portfolio is slightly above the median of the panel of funds for the years 2016 and 2017. However, it has underperformed the panel in 2018. This was a year in which the market as a whole fell, but by half compared with the model. However, a few points need to be made and some of the figures should be taken with a grain of salt. The annual standard deviations of the returns of the panel of US funds are calculated on the basis of monthly returns, i.e. over 12 observations instead of around 250. This is a significant smoothing operation that does not allow us to really take this parameter into account and compare it with that of the CBC portfolio. Returns on monthly US funds are also expressed on a gross basis. However, the cost structure of these funds has an impact on their net returns. According to Morningstar, US large cap funds generally have management fees of around 1.45% charged annually. Although this has the advantage of greatly limiting transaction costs, setting a fixed investment universe remains an unrealistic assumption in the hope of positive benchmarking. In fact, the model used in this research was trivial in the

sense that it had to limit itself to exploring the relationships between 200 stocks, whereas the funds on the panel can sometimes approach a number of underlying assets of 1,000. This is an important constraint for the model. The second is the determination of a static portfolio compared with actively managed funds that are dynamically rebalanced at different frequencies. There is a good chance that the relationships between assets will not be stable over time and that the spread of volatility among them will vary from year to year.

Nevertheless, two different doubts can be raised about the performance of the CBC portfolio under the conditions of this study. Firstly, although it is difficult to compare its performance with that of the fund panel due to the problem mentioned above, the standard deviation of the CBC portfolio's performance remains abnormally high. In fact, as explained above, the very principle of using such a model is to be able to create a portfolio with superior robustness and therefore solid diversification by minimising the propagation of price movements (volatility) between assets. It is therefore surprising to find standard deviations of this magnitude. There is a good chance that this is also due to the selection of the initial investment universe, with the 200 US large cap stocks most represented in the fund panel. Although there are 200 of these stocks, they remain highly correlated overall, as Figure 5 confirms. This point will also be discussed in the Discussion section. The second reason to doubt the performance of the CBC portfolio is its fall in the second half of 2018. As previously stated, the aim of such a model is to limit the impact of an event or trend that affects the market as a whole. The SP500 fell by 13.97% in the last quarter of 2018, and we might have hoped that this event would have been mitigated by the CBC portfolio. However, it fell by around 11.86% over 2018. Once again, this lack of resistance and robustness can be attributed to the model's limited scope. The relationships between assets quantified between 2012 and 2015 potentially no longer reflected what they actually were two and a half years later, at the end of 2018.

6. Discussions & reflections

This section will discuss the problems encountered, their potential reasons and possible future developments for this research. Although the results set out in the results section fell short of expectations in relation to the supposed purpose of the approach implemented, the method is nonetheless extremely relevant, especially at the present time. Indeed, many experts agree that the financial markets have not shown such complexity for a long time. Equipping oneself with tools that analyse the evolution of relationships between assets and/or, at a higher level, between the different indices/funds analysed is therefore essential at a time when even professionals are finding it hard to understand why certain assets respond to particular events and why others seem to avoid any impact, regardless of past theoretical beliefs.

Various possible areas of development could be incorporated into the research process carried out in the course of this work. For example, it would be possible to extend the data framework provided to the model, its investment universe of 200 stocks, either vertically or horizontally. It might indeed be worth thinking about increasing the number of assets observed, while remaining in the large cap segment, in order to provide the model with the possibility of seeking out and prioritising potentially less correlated large cap stocks. This horizontal expansion could improve the model's performance, but to a lesser extent if we take into account the large number of large cap stocks to be observed in order to find some that are less correlated. Vertical expansion could involve considering an investment universe that is still of a reasonable size, but that includes large cap stocks as well as small caps. The limits of the selection of the 160 stocks most present in the panel of funds showed its limits during this work and it would be judicious to consider an extension to the universe of small caps. The diversity and interdependence of these assets is often greater than that of the blue chip stocks studied here. This could therefore be a relevant area for development. Nevertheless, some processes in the implementation are considerably costly in terms of computing power, such as the creation of the correlation distance matrix. This is one of the current limitations encountered, given that the creation of the correlation matrix requires the daily price movements of stocks to be compared one by one. The relationship between the increase in the size of the model's investment universe and the computing power required to carry out this increase is therefore not linear.

Another possible area for development would be to exploit the time dimension of this research. It might be wiser to design a dynamic, rather than static, implementation of this approach. This could be structured as shown in Figure 10, i.e. evolving dynamically from period to period. We could imagine the distance correlations calculated over the different periods being discriminated according to their position in time, giving greater weight to recent correlations and less to more distant correlations. This historisation over time of the relationships between observed assets could also enable greater granularity of analysis, by analysing variations in the intensity of relationships between assets. Another feature that could add value to the way the approach works would be to consider, at each rebalancing period, an additional population that could be included in the

investment universe as well as a part of the actual investment universe that could be excluded on the basis of a contribution to portfolio risk-adjusted return on the portfolio that is attractive in the first case and insufficiently rewarding in the second.

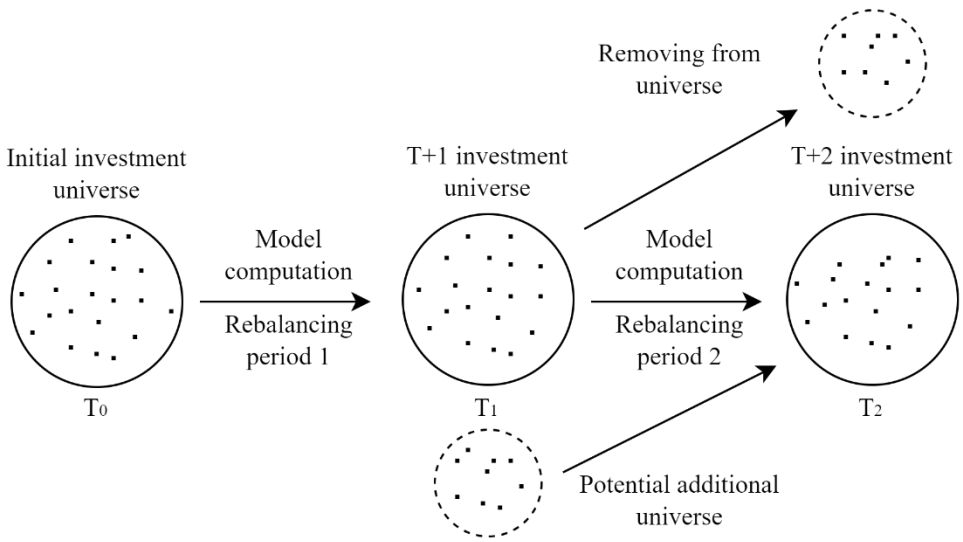


Figure 10: Dynamising the approach over time

We could also see the developped methodology through the prism of an indicator tool, not as a standalone portfolio construction model. Indeed, the concept of diversification, although central to portfolio theory, has often lacked concrete quantification, and it is conceivable that the concept of the contribution of assets to portfolio risk via the Communicability Betweenness centrality measure could prove to be a relevant concept for fund managers. Whether the funds are index-linked or more actively managed, a tool based on this methodology could provide a valuable insight for the person structuring the periodic rebalancings. This methodology could therefore easily be implemented as an external decision-support software package.

7. Conclusion

This study is organised in such a way as to provide an overall introduction to the research subject and to review the evolution of the two streams of literature inherent in the research process. It then outlines the methodology used to carry out this research process, starting with the definition of the key concepts specific to portfolio theory and graph theory and ending with the implementation of this methodology. The next section deals with the nature of the data used to carry out this work, as well as the selection and architecture process. Finally, the fruits of the preceding sections have been presented in the Results section, together with their motivation and/or justification. The Discussions & Reflections section attempted to broaden the framework of the research process in order to identify potential areas of development for this research.

The research question, or rather the problem that this research sought to answer, was to determine to what extent a solution based on concepts specific to graph theory could benefit portfolio construction and management. To this end, a graph mining model based on Communicability Betweenness centrality was used to create a portfolio of US equity large cap stocks. The aim of this process is to promote the allocation of capital to assets that are less interdependent with others in order to limit the propagation of price variations (volatility) between these different assets. This process has shown certain limitations, which are discussed in the section entitled Discussions, but also some potential axes for development. It is therefore possible to scale up the implemented approach towards a more professional and institutionalised approach that can help industry decision-making. The literature on the subject has been growing since the work of Estrada et al. (2008). The versatility of the measure studied in this research should also be emphasised, as it has already proved its worth in a wide variety of fields. The results of this research should be treated with caution and are not the most important element of this work. The appeal of this subject lies in its potential future directions. Whether by industry professionals, academics or individuals wishing to take their understanding of the relationships between the assets in their portfolios a step further.

8. Appendix

8.1 Bibliographie

Barnett, Janet Heine. (2009). « Early Writings on Graph Theory: Euler Circuits and The Königsberg Bridge Problem ». In *Resources for Teaching Discrete Mathematics*, édité par Brian Hopkins, 1re éd., 197-208. Mathematical Association of America. https://doi.org/10.5948/UPO9780883859742.026.

Bloch, Francis, Matthew O. Jackson, et Pietro Tebaldi. (2021). « Centrality Measures in Networks ». arXiv,. http://arxiv.org/abs/1608.05845.

Brinson, Gary P, L Randolph Hood, et Gilbert L Beebower. (1986). « Determinants of Portfolio Performance », s. d.

Clemente, Gian Paolo, Rosanna Grassi, et Asmerilda Hitaj. (2022). « Smart Network Based Portfolios ». *Annals of Operations Research* 316, nº 2 : 1519-41. https://doi.org/10.1007/s10479-022-04675-7.

Edelmann, Dominic, Tamás F. Móri, et Gábor J. Székely. (2021). « On Relationships between the Pearson and the Distance Correlation Coefficients ». *Statistics & Probability Letters* 169: 108960. https://doi.org/10.1016/j.spl.2020.108960.

Estrada, Ernesto, et Naomichi Hatano. (2008). « Communicability in Complex Networks ». *Physical Review E* 77, nº 3: 036111. https://doi.org/10.1103/PhysRevE.77.036111.

Estrada, Ernesto, Desmond J. Higham, et Naomichi Hatano. (2009). « Communicability Betweenness in Complex Networks ». *Physica A: Statistical Mechanics and Its Applications* 388, nº 5: 764-74. https://doi.org/10.1016/j.physa.2008.11.011.

Fabozzi, Frank J. (2007). éd. *Robust Portfolio Optimization and Management*. Frank J. Fabozzi Series. Hoboken, New Jersey: John Wiley.

Katz, Leo. (1953). « A New Status Index Derived from Sociometric Analysis ». *Psychometrika* 18, nº 1: 39-43. https://doi.org/10.1007/BF02289026.

Kim, Raehyun, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, et Jaewoo Kang. (2019). « HATS: A Hierarchical Graph Attention Network for Stock Movement Prediction ». arXiv. http://arxiv.org/abs/1908.07999.

Mantegna, R.N. (1999). « Hierarchical Structure in Financial Markets ». *The European Physical Journal B* 11, nº 1: 193-97. https://doi.org/10.1007/s100510050929.

Markowitz, Harry. (1952). « Portfolio Selection ». *The Journal of Finance*, Mémoire, 7, nº 1: 77-91.

Mayoral, Silvia, David Moreno, et Abalfazl Zareei. (2022). « Using a Hedging Network to Minimize Portfolio Risk ». *Finance Research Letters* 44: 102044. https://doi.org/10.1016/j.frl.2021.102044.

Michaud, Richard O., David N Esch, et Robert Michaud. (2010). « Portfolio Monitoring in Theory and Practice ». *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2658662.

Michaud, Richard O., et Robert O. Michaud. (2008). *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. 2nd ed. New York: Oxford University Press.

Rakib, Mahmudul Islam, Ashadun Nobi, et Jae Woo Lee. (2021). « Structure and Dynamics of Financial Networks by Feature Ranking Method ». *Scientific Reports* 11, nº 1: 17618. https://doi.org/10.1038/s41598-021-97100-1.

Sharpe, William F. (1963). « A Simplified Model for Portfolio Analysis ». *Management Science* 9, nº 2: 277-93.

Sharpe, William F. (1964). « Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk ». *The Journal of Finance* 19, nº 3: 425. https://doi.org/10.2307/2977928.

Silva, Thiago Christiano, Sergio Rubens Stancato De Souza, et Benjamin Miranda Tabak. (2016). « Structure and Dynamics of the Global Financial Network ». *Chaos, Solitons & Fractals* 88: 218-34. https://doi.org/10.1016/j.chaos.2016.01.023.

Surtee, Taariq G.H., et Imhotep Paul Alagidede. (2022). « A Novel Approach to Using Modern Portfolio Theory ». *Borsa Istanbul Review*, S2214845022001284. https://doi.org/10.1016/j.bir.2022.12.005.

Székely, Gábor J, et Maria L Rizzo. (2010). « Brownian Distance Covariance », s. d.

Székely, Gábor J., Maria L. Rizzo, et Nail K. Bakirov. (2007). « Measuring and Testing Dependence by Correlation of Distances ». *The Annals of Statistics* 35, nº 6. https://doi.org/10.1214/009053607000000505.

Tadlaoui, Ghali. (2017). « Intelligent Portfolio Construction: Machine-Learning Enabled Mean-Variance Optimization », s. d.

Titz, Robert. (2020). « Portfolio Optimization with Random Matrix Theory and Artificial Neural Networks », s. d.

Tse, Chi K., Jing Liu, et Francis C.M. Lau. (2010). « A Network Perspective of the Stock Market ». *Journal of Empirical Finance* 17, nº 4: 659-67. https://doi.org/10.1016/j.jempfin.2010.04.008.

Tumminello, Michele, Fabrizio Lillo, et Rosario N. Mantegna. (2010). « Correlation, Hierarchies, and Networks in Financial Markets ». *Journal of Economic Behavior & Organization* 75, nº 1 (juillet 2010): 40-58. https://doi.org/10.1016/j.jebo.2010.01.004.

Výrost, Tomas, Štefan Lyócsa, et Eduard Baumöhl. (2019). « Network-Based Asset Allocation Strategies ». *The North American Journal of Economics and Finance* 47: 516-36. https://doi.org/10.1016/j.najef.2018.06.008.

Yan, Xiangzhen, Hanchao Yang, Zhongyuan Yu, et Shuguang Zhang. (2021). « A Network View of Portfolio Optimization Using Fundamental Information ». *Frontiers in Physics* 9: 721007. https://doi.org/10.3389/fphy.2021.721007.

8.2 Code

This section documents some of the key functions used in the research process.

Distance correlation function

```
def distance_correlation(X, Y):
    X_diff = X[:, None] - X
    Y_diff = Y[:, None] - Y
    A = np.sqrt(np.dot(X_diff, X_diff.T))
    B = np.sqrt(np.dot(Y_diff, Y_diff.T))
    denom = np.mean(A) * np.mean(B)

    if denom == 0:
        return 0

    XY_diff = X[:, None] - Y
    covXY = np.mean(np.dot(X_diff, XY_diff.T))
    dcov2X = np.mean(np.dot(X_diff, X_diff.T))
    dcov2Y = np.mean(np.dot(Y_diff, Y_diff.T))

    dcorr = covXY / np.sqrt(dcov2X * dcov2Y)

    return dcorr
```

Network graph creation function

```
import networkx as nx

def build_corr_nx(df_train):
    cor_matrix = df_train.values.astype('float')
    sim_matrix = 1 - cor_matrix
    G = nx.from_numpy_matrix(sim_matrix)
    stock_names = df_train.index.values
    G = nx.relabel_nodes(G, lambda x: stock_names[x])
    G.edges(data=True)
    H = G.copy()

    for (u, v, wt) in G.edges.data('weight'):
        if wt >= 1 - 0.25:
            H.remove_edge(u, v)
        if u == v:
            H.remove_edge(u, v)
    return H
```

CBC function

```
def communicability_betweenness_centrality(G):
    nodelist = list(G)
    n = len(nodelist)
    A = nx.to_numpy_array(G, nodelist)
    A[np.nonzero(A)] = 1
    expA = sp.linalg.expm(A)
    mapping = dict(zip(nodelist, range(n)))
    cbc = {}
    for v in G:

        i = mapping[v]
        row = A[i, :].copy()
        col = A[:, i].copy()
        A[i, :] = 0
        A[:, i] = 0
        B = (expA - sp.linalg.expm(A)) / expA
        B = np.nan_to_num(B, nan=0)

        B[i, :] = 0
        B[:, i] = 0
        B -= np.diag(np.diag(B))
        cbc[v] = B.sum()

        A[i, :] = row
        A[:, i] = col

    order = len(cbc)
    if order > 2:
        scale = 1.0 / ((order - 1.0) ** 2 - (order - 1.0))
        for v in cbc:
            cbc[v] *= scale
    return cbc
```

Conversion of CBC to portfolio weights function

```
def centrality_to_portfolio_weights(weights):
    for key, value in weights.items():
        weights[key] = 1/value

    norm = 1.0 / sum(weights.values())
    print("sum(weights.values():", sum(weights.values()))
    print("norm:", norm)

    for key in weights:
        weights[key] = round(weights[key] * norm, 3)
```

```
    return weights
```

8.3 Others

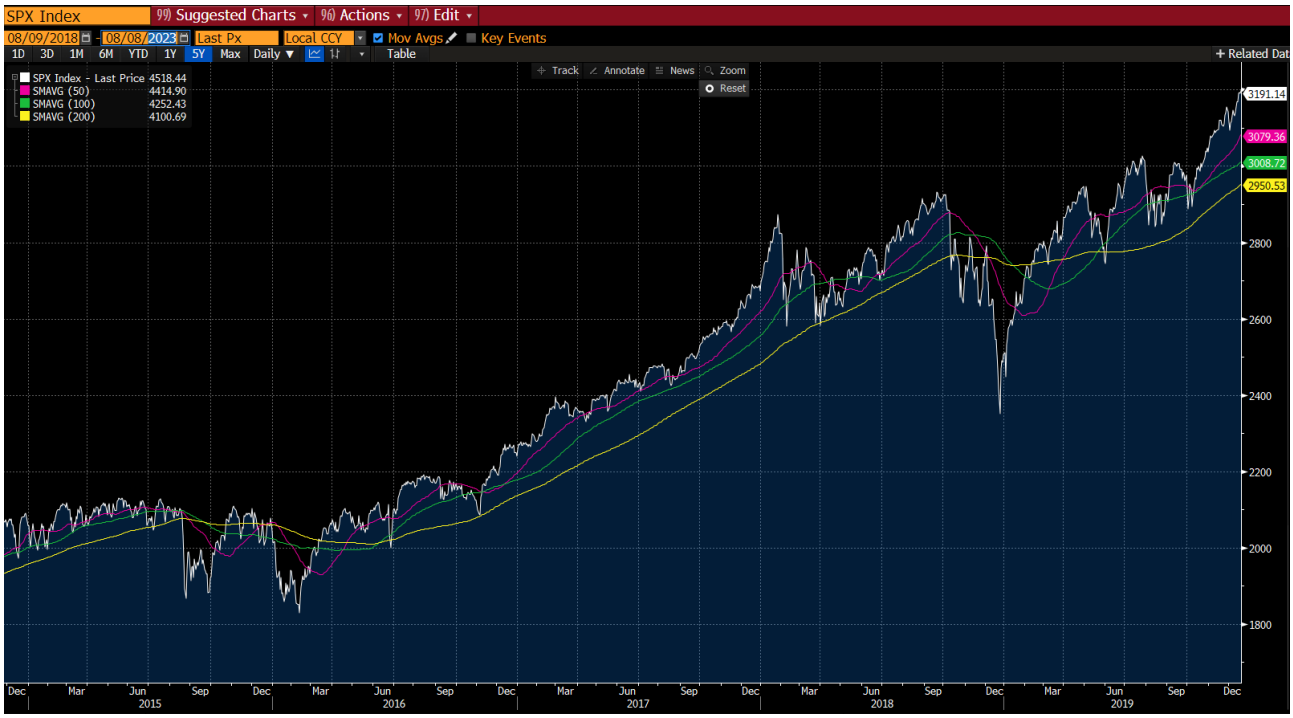


Figure: SP500 performance (2016-2018)