Research Article Open Access

Amelie Hüttner*, Jan-Frederik Mai, and Stefano Mineo

Portfolio selection based on graphs: Does it align with Markowitz-optimal portfolios?

https://doi.org/10.1515/demo-2018-0004
Received December 11, 2017; accepted March 23, 2018

Abstract: Some empirical studies suggest that the computation of certain graph structures from a (large) historical correlation matrix can be helpful in portfolio selection. In particular, a repeated finding is that information about the portfolio weights in the minimum variance portfolio (MVP) from classical Markowitz theory can be inferred from measurements of centrality in such graph structures. The present article compares the two concepts from a purely algebraic perspective. It is demonstrated that this heuristic relationship between graph centrality and the MVP does not originate from a structural similarity between the two portfolio selection mechanisms, but instead is due to specific features of observed correlation matrices. This means that empirically found relations between both concepts depend critically on the underlying historical data. Repeated empirical evidence for a strong relationship is hence shown to constitute a stylized fact of financial return time series.

Keywords: Portfolio selection, correlation matrix, minimum spanning tree, network centrality, Markowitz

MSC: 91G10, 62H20

1 Introduction

The problem of optimal investment in a universe of d assets is central in mathematical finance. It was first formalized in the seminal work of [22, 23], where optimal investment is considered in terms of a covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and, if desired, an expected return estimate $\mu \in \mathbb{R}^{d}$. A Markowitz-optimal portfolio is one that minimizes variance for a given expected target return (resp. maximizes return for a given variance), the optimal solution of this quadratic optimization problem under linear side constraint being known in closed form using matrix notation. Among all these optimal portfolios, the so-called *minimum variance portfolio (MVP)* is the one with smallest variance, and it depends solely on Σ (independent of μ). This approach has been extended in different directions, for example to the optimization of alternative risk or return measures as in [34], or to the inclusion of nonnegativity or cardinality constraints, or discrete-type constraints related to trading restrictions, which are highly relevant for practitioners as in, e.g., [12, 19]. The problem of robust covariance matrix estimation is a challenging topic of its own, relevant in different applications, and has also received considerable attention, see, e.g., [16, 27].

Recently, a more descriptive approach to portfolio selection has emerged: Pioneered by some remarkable works by Mantegna, e.g. [21], graph-based methods have found their way into finance literature, and recent studies, for example [13, 29, 30, 33], explore their usefulness for optimal investment purposes. In this context, portfolio selection is essentially also based on Σ (or some other $d \times d$-matrix composed of pairwise dependence measures), but relies on a more descriptive approach compared to the Markowitz paradigm. The matrix

***Corresponding Author: Amelie Hüttner:** Technical University of Munich, Chair of Mathematical Finance, Parkring 11, D-85748 Garching, Germany, E-mail: amelie.huettner@tum.de
Jan-Frederik Mai: XAIA Investment GmbH, Sonnenstraße 19, D-80331 München, Germany, E-mail: jan-frederik.mai@xaia.com
Stefano Mineo: XAIA Investment GmbH, Sonnenstraße 19, D-80331 München, Germany, E-mail: stefano.mineo@xaia.com

Σ is reduced to essential information in the form of a planar graph derived from it, such as, e.g., a *minimum spanning tree (MST)*. The resulting graph structure is used as an easy-to-grasp visualization of the essential aspects of interconnectedness between the assets. Investment decisions are then based on the idea of choosing 'central' or 'non-central' assets from the graph, according to certain centrality measures, as, intuitively, these should be related to risk propagation. Indeed, [29] find empirically that the non-central assets in an MST computed from the historical stock return correlation matrix are prominently represented in the associated MVP. Similarly, [33] detect that portfolio performance is improved if the constituent assets are selected among the non-central ones in an MST (or in a maximally filtered planar graph) derived from the correlation matrix. Using the same idea but a slightly differing technique, [13] bases his analysis on a matrix containing pairwise mutual information of the assets in order to make the dependence measurement more robust. He finds that more central assets yield higher returns, and concludes that portfolio selection should favor central names with low volatility, which is slightly opposite to the aforementioned references. [30] study the relation between Markowitz-optimal portfolios and graph-centrality not only empirically, but also provide a heuristic algebraic connection between both concepts. Like [29, 33], they find evidence for Markowitz-optimal portfolios favoring non-central assets. However, they also find that during certain time periods, in which the correlation between individual and systemic performance is high, more central assets gain more weight in Markowitz-optimal portfolios.

The present article analyzes whether there is a significant, inner-mathematical relationship between Markowitz variance minimization and graph-based portfolio selection based on the covariance (resp. correlation) matrix. Both approaches essentially base their investment decision solely on Σ, and in such a situation the Markowitz solution is optimal in a well-defined sense, namely variance minimization. When believing in the Markowitz setting, graph-based portfolio selection in general can lead to suboptimal results in this sense. Investment strategies based on graph structures seem only reasonable either if (i) data additional to Σ, for example higher-dimensional dependence measures (cf. Section 4), are incorporated into the graph somehow, if (ii) there is good reason to believe that variance minimization is not necessarily the target goal leading to optimal portfolios, if (iii) there is a hidden structural similarity between the selection mechanisms, or if (iv) it is known a priori that the underlying historical data structure guarantees a strong relationship with Markowitz portfolios. We demonstrate that condition (iii) is not algebraically given. Existing evidence that graph-centrality relates to portfolio performance is purely empirical. Since our analysis demonstrates that the inner-mathematical link between both concepts is rather weak, we conclude that existing findings highlighting a strong relation rely on the specific underlying data used for the estimation of Σ. In Section 3 we carry out our own study conducted on historical market data of credit default swaps (CDS), in order to extend previous investigations that have all considered equity return data to returns of assets dominated by credit risk. For small to moderate portfolio sizes, we find examples of MVPs overweighting central assets. With increasing portfolio size, however, we find that the percentage of MVPs favoring non-central assets grows. This is in line with the previous findings of [29, 30, 33], which all refer to large portfolios ($d > 200$ assets). The persistence of such empirical results for large portfolios in studies on different asset classes and time horizons indicates that correlation matrices calculated from large market data sets indeed tend to exhibit a special, non-random structure, which was also already observed by [36]. We further identify stylized facts of financial correlation matrices and investigate which of these might be responsible for the repeated empirical findings in favor of a relation between the two approaches. We find that a realistic eigenvalue structure alone does not result in similar outcomes of both portfolio selection methods.

The remainder of the article is organized as follows: Section 2 introduces the required concepts, Section 3 investigates the mathematical common grounds between Markowitz-optimization and graph-based portfolio selection, Section 4 points out general issues to be aware of when using graph-based portfolio selection methods, and Section 5 concludes.

2 Concepts

We consider an investment universe of $d \in \mathbb{N}$ assets. Each asset $k = 1, \dots, d$ is associated with its annualized log-return R_k, and all investigated methods of portfolio selection are based on an algorithm which depends on the probability distribution of the vector $\mathbf{R} := (R_1, \dots, R_d)$.

2.1 Minimum variance portfolio

In classical Markowitz portfolio theory, cf. [22, 23], the distribution of $\mathbf{R}$ is described in terms of its mean vector μ and its covariance matrix Σ. A portfolio comprised of the d assets is given by a vector $\mathbf{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ satisfying $\mathbf{1}^T \mathbf{x} = x_1 + \ldots + x_d = 1$, with $\mathbf{1}$ denoting a d-dimensional column vector with all entries being equal to one. Component x_i gives the portfolio weight of asset i, with negative value corresponding to shortselling the asset. The side condition $\mathbf{1}^T \mathbf{x} = 1$ demands that the portfolio is fully invested, shortselling being allowed. An optimal portfolio is one that minimizes portfolio variance $\mathbf{x}^T \Sigma \mathbf{x}$ for a given expected return $\mu^T \mathbf{x} = c$, with c an input constant. If one omits the constraint $\mu^T \mathbf{x} = c$ in the optimization problem, one obtains the portfolio $\bar{\mathbf{x}}$ with the smallest variance, the so-called *minimum variance portfolio (MVP)*. The latter is independent of μ and is given by

$$\bar{\mathbf{x}} = \bar{\mathbf{x}}(\Sigma) = \frac{\Sigma^{-1} \mathbf{1}}{\mathbf{1}^T \Sigma^{-1} \mathbf{1}}.$$

We will also occasionally use the abbreviation MVP(Σ) (resp. MVP(Ω)) for the minimum variance portfolio associated with a covariance matrix Σ (correlation matrix Ω).

2.2 Graphs associated with the covariance matrix

We investigate several graph-based portfolio selection algorithms, which all depend on the distribution of $\mathbf{R}$ solely through the covariance matrix. This choice is made to ensure comparability with the classical Markowitz approach. Graph-based methods can more generally also be based on any matrix $\Sigma \in \mathbb{R}^{d \times d}$ containing pairwise dependence measurements, with diagonal element Σ_{ii}, $i = 1, \dots, d$, interpreted as a measure of risk associated with asset i. For the sake of later reference we denote the sets of covariance and correlation matrices by

$$\begin{aligned} V_d &:= \left\{\Sigma \in \mathbb{R}^{d \times d} \;:\; \Sigma \text{ symmetric and non-negative definite}\right\}, \\ C_d &:= \left\{\Omega \in V_d \cap [-1, 1]^{d \times d} \;:\; \text{all diagonal entries are equal to } 1\right\}. \end{aligned}$$

Notice that $C_d \subset V_d$, i.e. every correlation matrix is a covariance matrix. The set C_d may be considered a compact subset of $\mathbb{R}^{d\,(d-1)/2}$, while V_d may be considered an unbounded subset of $\mathbb{R}^{d\,(d+1)/2}$. Regarding notation, for every $\Sigma \in V_d$ there is a unique $\Omega \in C_d$ such that

$$\Sigma = \operatorname{diag}(\sqrt{\Sigma_{ii}})\, \Omega \, \operatorname{diag}(\sqrt{\Sigma_{ii}}),$$

$$\operatorname{diag}(\sqrt{\Sigma_{ii}}) = \begin{bmatrix} \sqrt{\Sigma_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{\Sigma_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\Sigma_{dd}} \end{bmatrix}.$$

With each $\Sigma \in V_d$ we associate the weighted, undirected graph $G_w(\Sigma)$ with vertex set $\{1, \dots, d\}$, edge set[1] $\{(i, j) \;:\; 1 \le i < j \le d\}$, and associated edge weights $w(\Sigma_{ij})$, for a given strictly monotone function $w : \mathbb{R} \to \mathbb{R}$.

[1] We exclude diagonal entries, as these imply self-loops in the graph, which are not meaningful in the considered applications.

A popular weight function is the so-called correlation distance, $w(x) = \sqrt{2\,(1-x)}$, initially proposed by [21]. Sometimes covariances/correlations are also directly used as weights, $w(x) = x$. While for increasing w the interpretation of the edge weights is a measure of connectedness, for decreasing w the interpretation is a measure of distance, manifesting a "sign change" in interpretation. For a weighted graph G, a connected subgraph with the same vertex set and without cycles is called a *spanning tree* of G. Any connected G has a spanning tree, and among all spanning trees the one with minimal sum over all its edge weights is called *minimum spanning tree (MST)* of G. If all off-diagonal entries of Σ are mutually different, then there is exactly one minimum spanning tree of $G_w(\Sigma)$, see Matoušek and Nešetřil [25, Ch. 5.4, Ex. 4], which we denote by $\mathrm{MST}(\Sigma)$.

2.3 Centrality measures

We investigate the relation between 'central' resp. 'peripheral' assets in the graph associated with Σ and their weights in the corresponding MVP. An intuitive way of identifying non-central assets in a tree is to consider its **leaves**, i.e. vertices with only a single neighbor. More sophisticated definitions of centrality, which will be used in the remainder of this article, are the following:

- **Eigenvector centrality of a graph**: The adjacency matrix $\mathbf{A}$ of a finite connected graph has entries in $\{0, 1\}$ with $A_{ij} = 0$ (resp. $A_{ij} = 1$) meaning that there is no (resp. an) edge between vertex i and vertex j. By the Perron-Frobenius Theorem, the largest eigenvalue of $\mathbf{A}$ is positive and the associated eigenvector $\mathbf{v}_1$ has non-negative components. Consequently, by normalizing $\mathbf{v}_1$ in such a way that $\mathbf{v}_1^T\,\mathbf{1} = 1$, the dominant eigenvector $\mathbf{v}_1$ gives a probability distribution on the vertices. These probabilities can be interpreted as measurements of centrality in the graph, since $\mathbf{v}_1$ is the limit of $\mathbf{A}^n\,\mathbf{1}/\mathbf{1}^T\,\mathbf{A}^n\,\mathbf{1}$, i.e. the normalized version of $\mathbf{A}^n\,\mathbf{1}$, as $n \to \infty$. The i-th entry of $\mathbf{A}^n\,\mathbf{1}$ gives precisely the number of all paths in the graph of length n starting at i (including stopovers, i.e. all paths of length $\leq n$ without stopovers). Consequently, the largest entry of $\mathbf{v}_1$ corresponds to the vertex from which most different paths are possible, i.e. which is most connected to other vertices.

  While this centrality notion is originally based on unweighted (and interesting only for incomplete) graphs, [30] heuristically extend it to the weighted graph $G_w(\Sigma)$ replacing (A_{ij}) by $(w(\Sigma_{ij}))$ for increasing w, see Section 3.1 for details and comments.
- **Mean occupation layer of a tree**[2] This notion, cf. [29], is also called *closeness centrality* in [28]. The central vertex of a tree T according to this criterion is defined as the vertex $r(T)$ minimizing the so-called *mean occupation layer*

  $$\ell(T) := \frac{1}{d}\sum_{v=1}^{d}\mathcal{L}\big(r(T), v, T\big),$$

  where

  $$\mathcal{L}(r, v, T) := \text{length of (unique) tree-path from } r \text{ to } v.$$

  Intuitively, $\ell(T)$ gives the average length of a path in T from its root $r(T)$ to a vertex, and the root $r(T)$ is chosen such that this average length is minimal. Later on, we will apply this notion to a minimum spanning tree $\mathrm{MST}(\Omega)$ derived from a correlation matrix Ω. In this case, we will abbreviate $r(\Omega) = r(\mathrm{MST}(\Omega))$ and $\ell(\Omega) = \ell(\mathrm{MST}(\Omega))$.

For other centrality concepts the interested reader is referred to [13, 28]. Furthermore, we like to point out that the central vertex of a tree computed via the notion of eigenvector centrality can differ from the one computed via the notion of mean occupation layer, as Figure 1 shows.

[2] A tree is a connected graph without cycles.

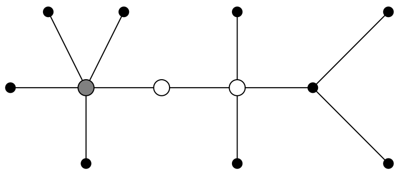


Figure 1: Central nodes implied by eigenvector centrality (gray) and mean occupation layer (white) may differ.

3 Graph-based portfolio selection

How do the weights in the variance-minimizing portfolio potentially relate to measurements of centrality in some associated graph structures?

Here is the heuristic idea: Imagine a graph whose vertices represent the assets and whose weighted edges are associated with measurements of dependence between the respective assets, such as $G_w(\Sigma)$ or a MST derived from it. It is then intuitive to assume that a variance-minimizing portfolio consists of rather non-central vertices in this graph because, heuristically, these should form a well-diversified portfolio, i.e. there should be a significant relationship between centrality measurements in graphs and the MVP. This is the fundamental idea on which the discussed graph methods are based.

- **Empirically:** Based on historical stock return data, some studies provide empirical findings that in financial asset return data one is likely to detect a strong relationship between centrality measurements in graphs and the MVP. In Section 3.3 we conduct our own study on historical CDS data, which provides further empirical evidence for this hypothesis.
- **Algebraically:** The MVP is known in closed form as a function of the historical covariance matrix of the asset returns, and portfolios resulting from graph-based portfolio selection methods are ultimately also functions of a dependence matrix of these assets. Consequently, one might wonder whether the mathematical functions transforming the given input matrices into the portfolio outputs share a great level of similarity. This would imply that the aforementioned empirical findings do not really detect special structure in the data, but instead are simply due to the fact that one unknowingly looks at the given data in two quite similar ways.

In the present article, we aim to show that there is no fundamental relation between the centrality measurements on graphs associated with the correlation matrix introduced in Section 2.3 and the weights in the corresponding MVP for more than 3 assets, and indicate that some of the proposed graph-based portfolio selection methods can lead to completely different portfolios than the MVP, hence be suboptimal in the sense of variance minimization. In other words, both concepts are truly different from a purely algebraic viewpoint. This makes clear that the persistent evidence for such relations in market data depends critically on the special structure of the observed data. We further attempt to identify which special structure of the correlation matrix might cause a relation between centrality and MVP weights.

3.1 MVP and eigenvector centrality

We first consider a possible relation between eigenvector centrality and MVP weights, which can be approached from the viewpoint of matrix algebra, cf. [30]. By means of an eigenvalue decomposition of the correlation matrix Ω, the MVP associated with the covariance matrix $\Sigma = \operatorname{diag}(\sqrt{\Sigma_{ii}})\,\Omega\,\operatorname{diag}(\sqrt{\Sigma_{ii}})$ can be rewritten as follows:

$$\bar{\mathbf{x}}(\Sigma) = \frac{\operatorname{diag}(1/\sqrt{\Sigma_{ii}})}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\left(\sum_{k=1}^{d}\frac{1}{\lambda_k}v_k v_k^T\right)\operatorname{diag}(1/\sqrt{\Sigma_{ii}})\mathbf{1}, \tag{1}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ denote the eigenvalues of Ω with associated orthonormal basis of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$.

In the definition of eigenvector centrality, the entries of the dominant eigenvector associated with the adjacency matrix $\mathbf{A}$ of an unweighted connected graph are non-negative and allow to be interpreted as measurements of centrality. However, [30] consider the weighted graph $G_w(\Omega)$ derived from a correlation matrix and relax the notion of eigenvector centrality in an intuitive, but algebraically questionable way. They consider the entries of the dominant eigenvector $\mathbf{v}_1$ of Ω (instead of $\mathbf{A}$) as measurements of centrality in $G_w(\Omega)$, although these entries need not be non-negative. In fact, [4] show that, considering purely random correlation matrices, they typically are not all non-negative. This renders the interpretation of the entries of $\mathbf{v}_1(\Omega)$ as measurements of centrality less intuitive. However, a major percentage of empirical correlation matrices exhibits a dominant eigenvector with non-negative entries, and according to [4], this percentage has been constantly growing. We are able to confirm this finding in our data set described in Section 3.3 consisting of 395 CDS time series: When considering the correlation matrices of randomly chosen portfolios of 20 assets, over 99.9% exhibited dominant eigenvectors with only non-negative entries.

[30] represent the minimum variance portfolio (1) as the sum of three parts:

$$\bar{\mathbf{x}}(\Sigma) = \frac{\operatorname{diag}(1/\sqrt{\Sigma_{ii}})}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\Big(I + N + R\Big),$$

$$I = \begin{bmatrix} \frac{1}{\sqrt{\Sigma_{11}}} \\ \vdots \\ \frac{1}{\sqrt{\Sigma_{dd}}} \end{bmatrix}, \quad N = \Big(\frac{1}{\lambda_1} - 1\Big)(\mathbf{v}_1^T I)\,\mathbf{v}_1, \quad R = \sum_{k=2}^{d}\Big(\frac{1}{\lambda_k} - 1\Big)(\mathbf{v}_k^T I)\,\mathbf{v}_k.$$

The term I is interpreted as **i**ndividual performance part, because its i-th entry is decreasing in the volatility $\sqrt{\Sigma_{ii}}$ of asset i, while the term N is interpreted as containing information about the location of asset i in the **n**etwork, and R is a **r**emainder part. In their Corollary 1, from this representation [30] draw the conclusion that under the conditions

$$\lambda_1 > 1 \quad \text{and} \quad \mathbf{v}_1^T I > 0, \tag{2}$$

non-central assets in $G_w(\Omega)$ receive large weights in the minimum variance portfolio.

The given conditions (2) are introduced purely for technical reasons, namely to ensure that N has negative entries and the centrality measurements in $\mathbf{v}_1$ enter the MVP with negative sign. It is important to note, however, since the eigenvalues of a correlation matrix sum up to its dimension, that $\lambda_1 > 1$ holds almost surely. The only possible case of $\lambda_1 \leq 1$ is $\lambda_1 = \ldots = \lambda_d = 1$, which corresponds to having the identity as correlation matrix, and almost surely never happens. The condition $\mathbf{v}_1^T I > 0$ is also fulfilled for almost all correlation matrices: If $\mathbf{v}_1^T I < 0$, it suffices to take $-\mathbf{v}_1$ instead of $\mathbf{v}_1$. This, too, is an eigenvector corresponding to the largest eigenvalue, and orthogonal to all others. Further, the mere fact that this condition holds true does not ensure a connection between low eigenvector centrality and high MVP weights, as can be shown by means of a Monte Carlo study as described in the sequel: The set C_d of all correlation matrices is compact, when considered as a subset of $\mathbb{R}^{d(d-1)/2}$. [10, 18] present efficient methods to simulate from the uniform distribution on C_d, denoted $\mathcal{U}(C_d)$ in the sequel. Intuitively, a realization of a correlation matrix $\Omega \sim \mathcal{U}(C_d)$ is completely random in the sense that no knowledge about the structure of Ω is taken into account and each element of C_d is equally likely to be a realization. We consider MVPs constructed from $\Sigma = \Omega$, where $\Omega \sim \mathcal{U}(C_d)$, which fulfill the above conditions, and are interested in the quantity

$$e_d(\Omega) := \frac{\text{portfolio weight of the 20\% least central assets}}{0.2}.$$

The numerator of e_d is the sum of the MVP weights assigned to the 20% least central assets according to [30]'s version of eigenvector centrality. If the centrality measurements did not play a role in the construction of the MVP, we would expect that these assets get assigned a total weight of about 20% (since MVP weights sum up to 1), so the denominator is chosen in order to normalize e_d. A value of $e_d > 1$ thus indicates an overrepresentation of the 20% least central assets in the MVP. Figure 2 visualizes the density of $e_d(\Omega)$ as a histogram of its law based on $n = 1,000,000$ independent simulations. We find indeed that there is a large probability $P(e_d > 1)$ for overrepresentation of the 20% least central assets, confirming [30]'s result where they regress MVP weights on centrality measurements and find a significant negative relation. However, in all considered dimensions d there exists a nonempty set of correlation matrices that fulfill [30]'s technical conditions, and yet exhibit an underrepresentation of the 20% least central assets in the MVP. Moreover, the probability of underweighting these assets in the MVP increases with the dimension of the correlation matrix, cf. Figure 2 and Table 5.

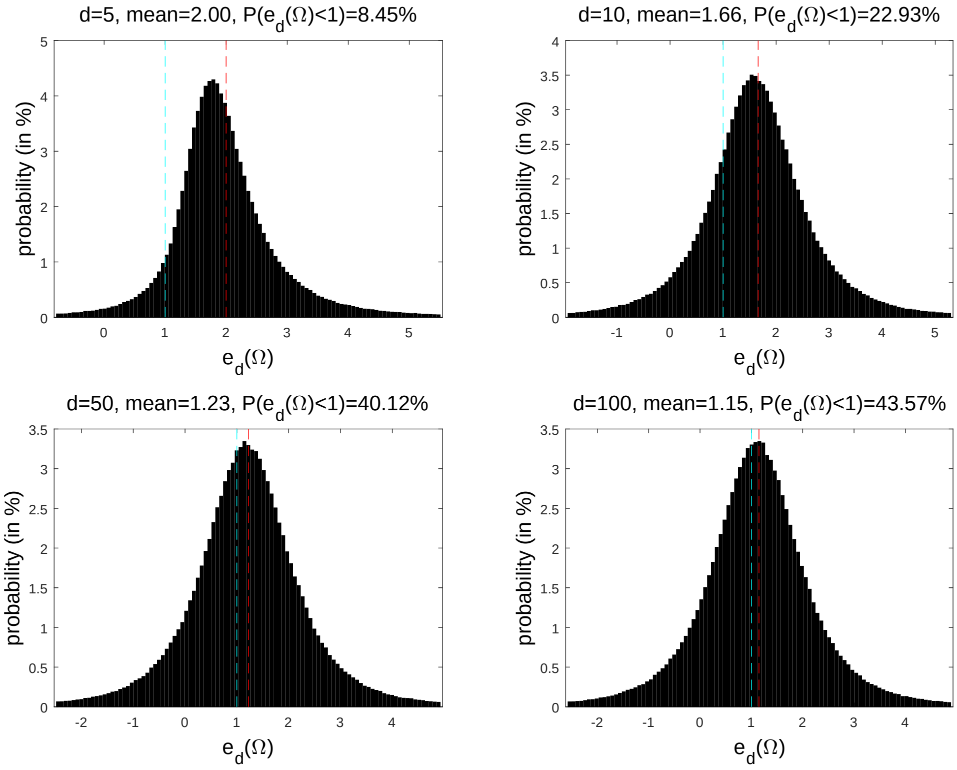


Figure 2: Histogram of the probability distribution of $e_d(\Omega)$ with $\Omega \sim \mathcal{U}(C_d)$ based on $n = 1,000,000$ simulations, for $d \in \{5, 10, 50, 100\}$. The vertical, red line gives the mean, and the blue line represents the border 1 between over- and underrepresentation of the non-central assets.

[30] do not discuss the influence of the remainder term R in their decomposition of the MVP, which can be quite large and indeed offset the influence of the network centrality related part N, as we illustrate in Example 1 below. Indeed, according to [15], 'the composition of the least risky portfolio has a large weight on the eigenvectors with the smallest eigenvalues', as can be adumbrated also from the formulas for R and N, which contain the eigenvalues in the denominator.

Example 1 (An example in $d = 5$). *Consider the 5-dimensional correlation matrix*

$$\Sigma = \Omega = \begin{bmatrix} 1 & 0.2 & 0.4 & 0.1 & -0.3 \\ 0.2 & 1 & 0.4 & 0.1 & -0.1 \\ 0.4 & 0.4 & 1 & -0.7 & -0.2 \\ 0.1 & 0.1 & -0.7 & 1 & 0 \\ -0.3 & -0.1 & -0.2 & 0 & 1 \end{bmatrix}.$$

It can easily be checked numerically that Ω is positive definite and has leading eigenvalue $\lambda_1 = 1.9646$. The corresponding eigenvector is

$$\mathbf{v}_1 = (0.4035,\ 0.3517,\ 0.6737,\ -0.4106,\ -0.3016)^T,$$

and the condition $\mathbf{v}_1^T I = \mathbf{v}_1^T \mathbf{1} > 0$ is fulfilled. The MVP $\bar{\mathbf{x}} = \bar{\mathbf{x}}(\Sigma)$ is given and decomposed as

$$\bar{\mathbf{x}} = \begin{bmatrix} -0.1466 \\ -0.1828 \\ 0.6299 \\ 0.5548 \\ 0.1446 \end{bmatrix} = 0.0809 \left(\underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}_{I} + \underbrace{\begin{bmatrix} -0.1420 \\ -0.1238 \\ -0.2371 \\ 0.1445 \\ 0.1062 \end{bmatrix}}_{N} + \underbrace{\begin{bmatrix} -2.6701 \\ -3.1353 \\ 7.0234 \\ 5.7131 \\ 0.6816 \end{bmatrix}}_{R} \right).$$

It is observed that the largest weight in the MVP is assigned to asset 3, the most central asset according to the entries of the first eigenvector, as the remainder part R offsets the negative influence of the centrality measurements in part N.

Remark 1 (Further related work)**.** Many authors argue that the (normalized) dominant eigenvector, i.e. the eigenvector corresponding to the largest eigenvalue, of the correlation matrix Ω of stock returns provides a reasonable proxy for the so-called *market portfolio*; see the references in [4]. The i-th component of the latter by definition equals the market share of asset i (among the d assets considered). The market portfolio plays an important role in Markowitz theory and the capital asset pricing model (CAPM). According to the mean-variance tautology in Roll's critique, cf. [35], the market portfolio lies on the efficient frontier (i.e. it is mean-variance efficient) if and only if the CAPM holds. This means that under the assumption of the CAPM framework, the market portfolio is mean-variance efficient in the sense of Markowitz. Apparently the market portfolio has non-negative components, while the dominant eigenvector of an arbitrary correlation matrix can have negative components, see [4] for examples and a thorough investigation of this issue. This shows that the dominant eigenvector in general is not equal to the market portfolio, and the aforementioned findings are merely approximations that work well empirically.

3.2 MVP and MST

The arguments presented in the previous section already raise first doubts regarding a fundamental relation between centrality on a graph associated with the covariance matrix and the corresponding MVP weights. Whereas we have just dealt with a 'weighted' centrality measure on the complete graph $G_w(\Sigma)$, in the following we will focus on leaves and closeness centrality on the associated MST. Empirical findings are in favor of an existing relation between MST and MVP. Based on historical data, [29, 33] find that non-central assets in MST(Ω) dominate Markowitz-optimal portfolios. For instance, it is claimed that 'the companies of the minimum risk Markowitz portfolio [MVP] are always located on the outer leaves of the [minimum spanning] tree', cf. [29, p. 1].

The following lemma shows that at least in the simplest case $d = 3$ there is a fundamental relation between MVP weights and the MST, if variances are ignored and only a correlation matrix is considered. Notice that the statement remains valid also for covariance matrices as long as all their diagonal entries are identical.

Lemma 1 (MVP and MST for $d = 3$)**.** *Consider a 3×3 correlation matrix $\Omega \in C_3$.*
(a) The MVP associated with the matrix $\Sigma = \Omega$ is $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \bar{x}_3)^T$, where

$$\bar{x}_i := \frac{(1-\Omega_{jk})(1+\Omega_{jk}-\Omega_{ij}-\Omega_{ik})}{D},$$

with (i, j, k) some permutation of $(1, 2, 3)$,

$$D := 4\,(1-\Omega_{13})\,(1-\Omega_{23}) - (1+\Omega_{12}-\Omega_{23}-\Omega_{13})^2.$$

(b) *Let T be an MST associated with Ω, computed from $G_w(\Omega)$ with a decreasing weight function w. The unique*[3] *central vertex of T corresponds to the minimum weight in the MVP. More formally, letting $\{1, 2, 3\} = \{i, j, k\}$, if $\Omega_{ij} = \min\{\Omega_{12}, \Omega_{13}, \Omega_{23}\}$, then $\bar{x}_k = \min\{\bar{x}_1, \bar{x}_2, \bar{x}_3\}$.*

Proof.

(a) Tedious but straightforward computation.

(b) By symmetry, it suffices to verify the statement for $k = 3$, i.e. we may assume w.l.o.g. that Ω_{12} is the smallest entry of Ω. We also assume w.l.o.g. that $\Omega_{23} \geq \Omega_{13}$ (the opposite case is treated symmetrically). We have to show (i) $\bar{x}_3 \leq \bar{x}_2$ and (ii) $\bar{x}_3 \leq \bar{x}_1$. Using part (a) and some basic algebra, the inequality (i) is seen to be equivalent to

$$\Omega_{13}\left(\Omega_{13} - (1 + \Omega_{23})\right) \leq \Omega_{12}\left(\Omega_{12} - (1 + \Omega_{23})\right). \tag{3}$$

The function $f_{23}(u) := u\,(u - (1 + \Omega_{23}))$ is a parabola with global minimum at $u_{23} := (1 + \Omega_{23})/2$. Since $\Omega_{12} \leq \Omega_{13} \leq u_{23}$ by assumption, it follows that $f_{23}(\Omega_{12}) \geq f_{23}(\Omega_{13})$, which is equivalent to (3), hence to (i). Using part (a) and some basic algebra, the inequality (ii) is seen to be equivalent to

$$\Omega_{23}\left(\Omega_{23} - (1 + \Omega_{13})\right) \leq \Omega_{12}\left(\Omega_{12} - (1 + \Omega_{13})\right). \tag{4}$$

The function $f_{13}(u) := u\,(u - (1 + \Omega_{13}))$ is a parabola with global minimum at $u_{13} := (1 + \Omega_{13})/2$. In order to verify (ii), it suffices to verify (4), which is equivalent to showing $f_{13}(\Omega_{12}) \geq f_{13}(\Omega_{23})$. If $\Omega_{23} \leq u_{13}$, the assertion follows precisely as in the previous case (i). If not, then we have $\Omega_{12} \leq u_{13} < \Omega_{23}$. Since f_{13} is a parabola, the assertion holds true if and only if $\Omega_{23} - u_{13} \leq u_{13} - \Omega_{12}$. The last inequality is equivalent to $\Omega_{23} - 1 \leq \Omega_{13} - \Omega_{12}$, which is true since the left-hand side is non-positive and the right-hand side non-negative by assumption.

□

A statement similar to the one of Lemma 1(b), algebraically hard-coding a relation between centrality in MST(Ω) and weights in MVP(Ω), becomes more difficult to obtain in larger dimensions, as the following Example 2 shows in $d = 5$.

Example 2 (Non-centrality in MST ≠ large weight in MVP). *Consider the 5-dimensional correlation matrix of Example 1, whose MVP is given by*

$$\bar{\mathbf{x}} = (-0.1466, -0.1828, 0.6299, 0.5548, 0.1446)^T.$$

In particular, the assets 3 and 4 have by far the largest weights in the MVP. However, it is readily checked that none of these two assets is a leaf in any MST associated with Ω. There is an MST with leaves 1 and 5, and an MST with leaves 2 and 5.

While Example 2 shows that there exist correlation matrices for which the dominating constituents of the MVP form a subset of assets disjoint from the subset of leaves in an associated MST, it is natural to ask how many correlation matrices of this type do exist, i.e. how pathological Example 2 is. For a given dimension $d \geq 2$ this question can be answered by means of a Monte Carlo study similarly as in Section 3.1. Recall that all entries of $\Omega \sim \mathcal{U}(C_d)$ are almost surely mutually distinct, so there is a unique minimum spanning tree MST(Ω) of $G_w(\Omega)$, with w a decreasing weight function. We denote by $|B|$ the cardinality of a finite set B and by L(Ω) the set of all leaves of MST(Ω), and are interested in the random variable

$$f_d(\Omega) := \frac{\text{portfolio weight of L}(\Omega)\text{ in MVP}(\Omega)}{|\text{L}(\Omega)|/d}, \quad \Omega \sim \mathcal{U}(C_d).$$

[3] In the case $d = 3$ the two leaves i and j of an MST are obviously such that Ω_{ij} is the minimal entry of Ω. The MST is unique if this minimal entry is unique.

The numerator of $f_d(\Omega)$ gives the MVP-weight of the leaves in MST(Ω), while the denominator gives the share of leaves of MST(Ω) in all d assets. Intuitively, $f_d(\Omega)$ is $>$ 1 ($<$ 1) if and only if the leaves are over- (under-) represented in the MVP. Figure 3 visualizes the density of $f_d(\Omega)$ in terms of a histogram for the law of $f_d(\Omega)$ based on $n = 1,000,000$ independent simulations. It is observed that the mean of $f_d(\Omega)$ is indeed greater than 1, indicating that there are more correlation matrices overweighting the leaves in MVP(Ω) than underweighting them. However, with increasing dimension d the mean $\mathbb{E}[f_d(\Omega)]$ decreases and the probability of underweight $\mathbb{P}(f_d(\Omega) < 1)$ increases. This suggests that there is not really a strong relation between MVP(Ω) and MST(Ω) for large d, unless one knows something about the structure of the correlation matrix which rules out certain subsets of C_d. Indeed, there exist many correlation matrices that even underweight the leaves of MST(Ω) in MVP(Ω).

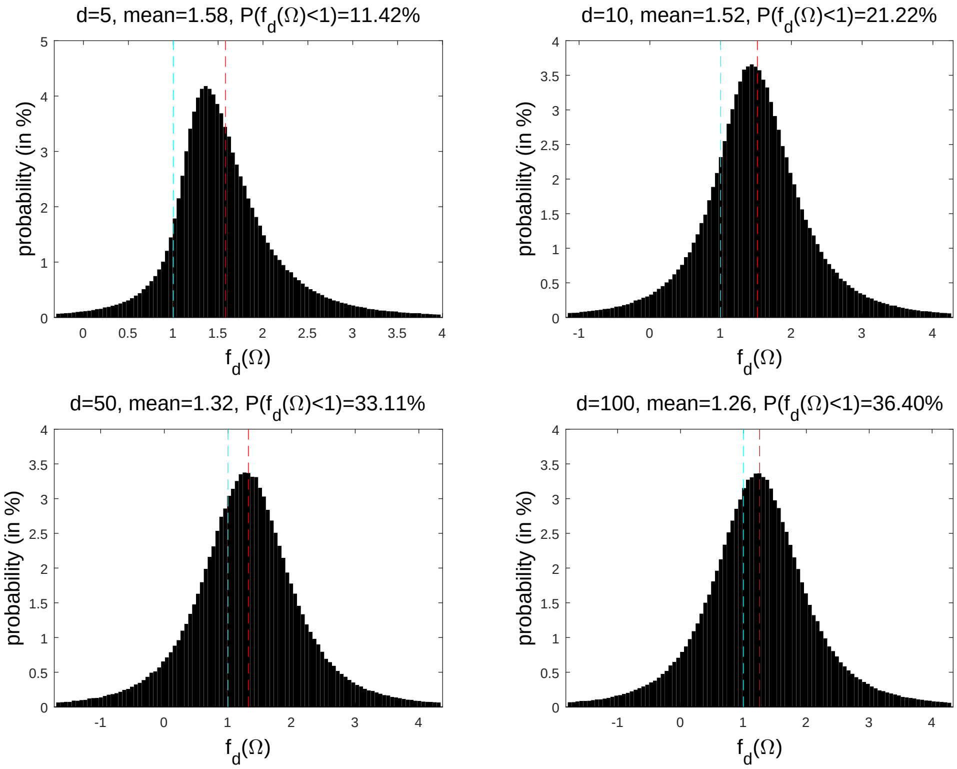


Figure 3: Histogram of the probability distribution of $f_d(\Omega)$ with $\Omega \sim \mathcal{U}(C_d)$ based on $n = 1,000,000$ simulations, for $d \in \{5, 10, 50, 100\}$. The vertical, red line gives the mean, and the blue line represents the border 1 between over- and underrepresentation of the leaves.

We have seen that there is a huge number of correlation matrices, for which the associated MVP is dominated by non-leaf assets. A related, but clearly much weaker question is, whether there exists at least one leaf which is overweighted in the MVP. To this end, instead of $f_d(\Omega)$, we repeat the analysis above with the random variable

$$g_d(\Omega) := \max_{B \subset L(\Omega)} \left\{ \frac{\text{portfolio weight of } B \text{ in MVP}(\Omega)}{|B|/d} \right\}, \quad \Omega \sim \mathcal{U}(C_d).$$

Table 1 shows that the answer to this question is by far more affirmative, i.e. for almost every correlation matrix there is at least one leaf prominently represented in the MVP, and for $d \geq 50$ this statement becomes practically certain. Recall that this statement is universal, i.e. follows from the structure of C_d and has nothing to do with empirical data.

Table 1: Mean of the probability distribution of $g_d(\Omega)$ and probability that no subset of leaves is overweighted in the MVP, with $\Omega \sim \mathcal{U}(C_d)$ based on $n = 1,000,000$ simulations, for $d \in \{5, 10, 50, 100\}$.

	$d = 5$	$d = 10$	$d = 50$	$d = 100$
$\mathbb{E}[g_d(\Omega)]$	2.39	3.59	11.07	17.47
$\mathbb{P}(g_d(\Omega) < 1)$	2.29%	1.54%	0.00%	0.00%

Instead of just focusing on leaves, [29] use the more sophisticated concept of *mean occupation layer* introduced in Section 2 to relate centrality in MST(Ω) and the associated MVP weights: Choosing the central vertex $r(\Omega)$ such that the mean occupation layer $\ell(\Omega)$ is minimized, and associating each vertex v with a layer $\mathcal{L}(r(\Omega), v, \Omega)$ corresponding to the length of the unique MST(Ω)-path connecting this vertex with the central one, they find that the *MVP-weighted portfolio layer*

$$\sum_{v=1}^{d} \bar{x}_v(\Sigma)\, \mathcal{L}\big(r(\Omega), v, \Omega\big)$$

is larger than the mean occupation layer $\ell(\Omega)$. In words, this means that the MVP assigns more weight to non-central assets than an equally-weighted basket does (i.e. more weight on non-central assets than on central ones). To analyze this finding, we also consider the following two random variables, with $\mathrm{O}(\Omega)$ denoting the subset of $\mathrm{L}(\Omega)$ consisting only of the leaves $v \in \mathrm{L}(\Omega)$ with maximal length $\mathcal{L}\big(r(\Omega), v, \Omega\big)$ (we call them *outer leaves*):

$$\begin{aligned} h_d(\Omega) &:= \frac{\sum_{v=1}^{d} \bar{x}_v(\Omega)\, \mathcal{L}\big(r(\Omega), v, \Omega\big)}{\ell(\Omega)}, \\ i_d(\Omega) &:= \frac{\text{portfolio weight of } \mathrm{O}(\Omega) \text{ in } \mathrm{MVP}(\Omega)}{|\mathrm{O}(\Omega)|/d}, \qquad \Omega \sim \mathcal{U}(C_d). \end{aligned}$$

Again, a fundamental relation can not be detected, as shown in Figure 4 and Tables 2 and 5: While for a lower number of assets the probability of overweighting non-central assets, resp. outer leaves, in the MVP is substantial, this finding is not persistent for larger dimensions. For portfolios consisting of 100 assets, only in about half of the cases non-central assets, resp. outer leaves, are dominating the MVP.

3.3 Empirical results from CDS portfolios

In the light of our previous analyses, we conclude that the strong relations between non-centrality in a graph and an MVP, both associated with the correlation matrix Ω, as observed by [13, 29, 30, 33] are not inner-mathematical, but purely data-dependent. Looking at historical data of credit default swaps (CDS), we are able to confirm this suspicion: We find that large portfolios tend to exhibit a strong overweighting of non-central assets. Portfolios underweighting non-central assets are only found for small to moderate portfolio

Table 2: Mean of the probability distribution of $i_d(\Omega)$ and probability of underweighting outer leaves, with $\Omega \sim \mathcal{U}(C_d)$ based on $n = 1,000,000$ simulations, for $d \in \{5, 10, 50, 100\}$.

	$d = 5$	$d = 10$	$d = 50$	$d = 100$
$\mathbb{E}[i_d(\Omega)]$	1.54	1.51	1.34	1.30
$\mathbb{P}(i_d(\Omega) < 1)$	20.19%	33.89%	45.74%	47.55%

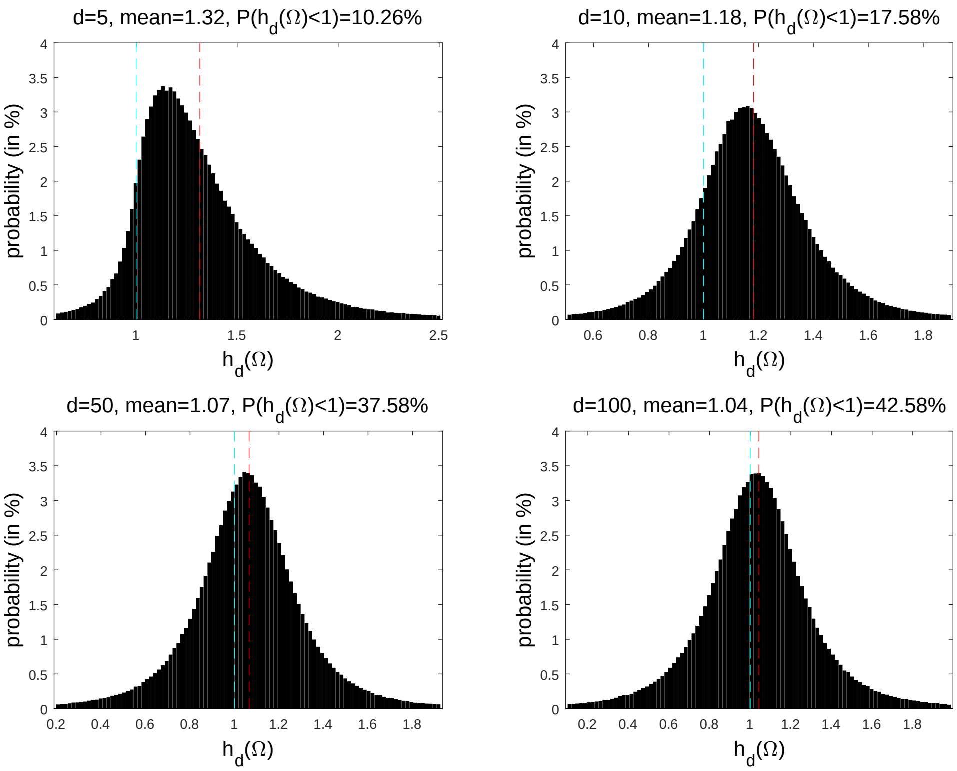


Figure 4: Histogram of the probability distribution of $h_d(\Omega)$ with $\Omega \sim \mathcal{U}(C_d)$ based on $n = 1,000,000$ simulations, for $d \in \{5, 10, 50, 100\}$. The vertical, red line gives the mean, and the blue line represents the border 1 between over- and underrepresentation of the non-central assets.

sizes. Our data set[4] consists of 5Y-CDS mid upfront time series of the constituents of the four major credit indices, namely ITRX EUR, ITRX XO, CDX IG, and CDX HY, observed daily from July 30, 2015 to May 2, 2017. For each asset we consider the trading strategy of selling 5Y CDS protection. Notice that CDS maturities are standardized to be always on 20 June or 20 December of a year. Furthermore, the observed market price (=upfront) of a 5Y CDS switches from the CDS with maturity in June (December) to the one with maturity in December of the same year (June of the next year) on 20 September (20 March). On these CDS roll dates 20 March and 20 September the trading strategy closes out the old CDS and rolls into the new one, in order to be in accordance with the observed market prices and to keep the duration of the CDS as constant as possible over time. If u_t denotes the upfront of a CDS on day t, we define the log-return at the next day $t + 1$ by $\log\left(\frac{1-u_{t+1}}{1-u_t}\right)$. This is because the value $1 - u_t$, sometimes called the *bond-equivalent value* of the CDS, can be considered the value of the investment at time t. Clearly, $-u_t$ is the value of the CDS, but the amount 1 needs to be held in cash because it is at stake in case of a potential credit event at t, followed by a CDS auction yielding zero recovery

[4] Source: ICE Data Services.

rate.[5] After deleting series with missing data, we are left with 395 assets. We study the quantities

$$\tilde{e}_d(\Sigma) := \frac{\text{portfolio weight of 20\% least } \Omega\text{-eigenvector-central assets in MVP}(\Sigma)}{0.2}$$

$$\tilde{f}_d(\Sigma) := \frac{\text{portfolio weight of L}(\Omega) \text{ in MVP}(\Sigma)}{|\mathrm{L}(\Omega)|/d},$$

$$\tilde{h}_d(\Sigma) := \frac{\sum_{v=1}^{d} \bar{x}_v(\Sigma)\, \mathcal{L}\big(r(\Omega), v, \Omega\big)}{\ell(\Omega)},$$

for each of the four indices, where d is the number of assets of the respective index included in our data set. Here, Σ refers to the covariance matrix of the considered CDS investment log-return time series, Ω is the correlation matrix associated with Σ, and $\bar{\mathbf{x}}$ are the MVP weights calculated from Σ. Table 3 shows that $\tilde{e}_d(\Sigma)$, $\tilde{f}_d(\Sigma)$ and $\tilde{h}_d(\Sigma)$ are larger than 1 for all four indices, thus indicating an overweighting of leaves, respectively non-central assets according to both the mean occupation layer criterion and [30]'s eigenvector centrality. It is worth noting that all indices do not only fulfill the plausible conditions of [30], but also the way more restrictive condition that the first eigenvector has only non-negative components.

Table 3: In all major credit indices we detect a systematic overweighting of leaves resp. non-central assets in the MVP. The number of constituents of these indices (after deleting series with missing data) is given in the rightmost column.

	$\tilde{e}_d(\Sigma)$	$\tilde{f}_d(\Sigma)$	$\tilde{h}_d(\Sigma)$	constituents
ITRX EUR	3.11	2.56	1.35	123
ITRX XO	1.57	1.58	1.07	64
CDX IG	3.09	1.70	1.23	123
CDX HY	1.89	1.86	1.20	85

To get a more profound impression, we further calculate $\tilde{e}_d(\Sigma)$, $\tilde{f}_d(\Sigma)$ and $\tilde{h}_d(\Sigma)$ for $n = 1{,}000{,}000$ random drawings of $d = 20$ assets out of our pool comprising 395 firms. There are $\binom{395}{d} \approx 2.1547 \cdot 10^{33}$ possibilities, so enough that the probability of choosing the same set twice is negligible. Unlike the aforementioned references and Table 3, we cannot confirm a systematic overweighting of non-central assets for arbitrary baskets of CDS, cf. Figure 5. A significant number of the randomly chosen portfolios exhibits an underweighting of leaves, respectively peripheral assets. An example is given in Table 4.

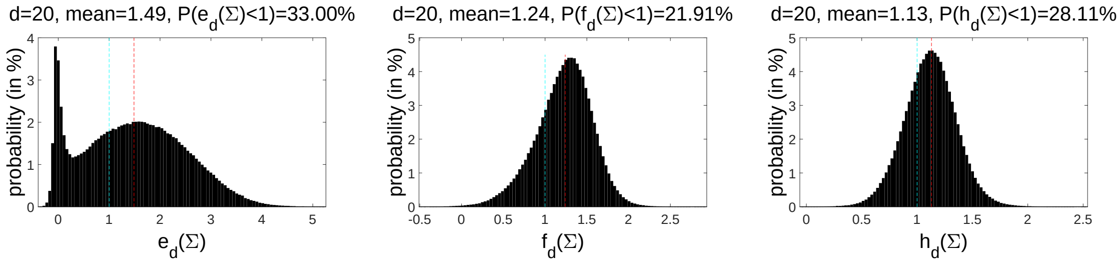


Figure 5: Histograms of the probability distributions of $\tilde{e}_{20}(\Sigma)$ (left), $\tilde{f}_{20}(\Sigma)$ (middle) and $\tilde{h}_{20}(\Sigma)$ (right) with Σ being the covariance matrix of 20 randomly chosen CDS upfront time series out of the 395 considered entities. The vertical, red lines give the respective mean, and the blue lines represent again the border 1 between over- and underrepresentation of non-central assets.

[5] If this value was not held in cash, the investment must be considered levered, which we do not.

Table 4: The following portfolio exhibits a systematic underweighting of leaves, respectively peripheral assets, in the considered time period: $\tilde{e}_d(\Sigma) = 0.2392, \tilde{f}_d(\Sigma) = 0.8847, \tilde{h}_d(\Sigma) = 0.8777$.

firm	MVP-weight	MST-layer	$\mathbf{v}_1$
Centrica	0.0796	1	0.2754
Halliburton	0.0151	3	0.1994
DISH DBS	-0.0336	2	0.2361
Koninklijke Ahold Delhaize	0.2249	1	0.2568
Whirlpool	0.0453	3	0.2113
Meritor	-0.0191	2	0.2069
Quest Diagnostics	0.3829	3	0.2065
BMW	-0.0090	2	0.2308
AirFrance-KLM	-0.0129	2	0.2044
Gas Natural	-0.0971	0	0.2795
Danone	0.2307	1	0.2263
Orange	0.2481	1	0.2623
Ziggo Bond Finance	-0.0389	1	0.2548
Best Buy	-0.0266	4	0.2074
Lincoln National	0.0078	1	0.1809
Deutsche Lufthansa	-0.0034	2	0.2314
Newmont Mining	0.0170	4	0.0387
Stena	-0.0162	2	0.2620
Astaldi	-0.0026	3	0.2118
Nordstrom	0.0079	5	0.1739

However, for increasing dimension, the probability of finding an overweighting of non-central assets increases, cf. Figure 7 and Table 5. This aligns with the results of [29, 30, 33], who study data sets of 200, 300, and 477 stocks, respectively, and with our previous observations for the four major credit indices.

3.4 Influence of structures of observed correlation matrices

The persistence of the empirical finding that there is indeed a strong relation between non-centrality in the MST and comparatively large MVP weights indicates that large financial return data sets have a special covariance/correlation structure. An interesting line of research is to identify the features of covariance resp. correlation matrices that cause this relation. There are mainly two different strands of research concerned with the special structure of market correlation matrices: One strand approaches the problem by investigating corresponding graph structures, e.g. [2, 36], the other utilizes results from random matrix theory to assess the degree of randomness in a given empirical correlation matrix, e.g. [3, 5, 15, 31, 32].

In the following, we briefly summarize stylized facts observed in market correlation matrices:

1. **Large first eigenvalue**
   Considering the spectrum, random matrix theory predicts a certain range $[\lambda^-, \lambda^+]$ and density f_λ for the eigenvalues of a random correlation matrix constructed from data matrices with iid entries[6], which is usually violated by the eigenvalues of market correlation matrices. The largest empirical eigenvalues lie well above the theoretical upper bound λ^+, cf. [3, 5, 15, 31, 32].
   In our data set, we find that the first eigenvalue explains about 40% of the variance: 47% in $d = 5$,

[6] These bounds and the density are dependent on the ratio of the number of simulated time series to their length.

declining in d to 37% in $d = 50$. Correlation matrices simulated from $\mathcal{U}(C_d)$ however fail to produce such large first eigenvalues in our simulations. Regardless of the dimension d of the simulated correlation matrix, its eigenvalues almost all lie in the interval [0,4], with eigenvalues larger than 6 never observed in $d = 5, 10, 50, 100$ in $n = 1,000,000$ simulations.

2. **Perron-Frobenius property**
   As already mentioned in Section 3.1, [4] observe in a data set of S&P1500 stocks that the major percentage of market correlation matrices exhibits a dominant eigenvector with only positive entries, and this percentage has been increasing up to 100% in their considered time period from 1994 to 2013. Indeed, also in our data set, more than 99.9% of correlation matrices ($d = 20$, $n = 1,000,000$ simulations) have this property.
   When simulating from $\mathcal{U}(C_d)$ however, correlation matrices with the Perron-Frobenius property are realized with a very small probability, which declines to zero quickly with growing dimension, cf. [4]: In $d = 5$, about 6% of the correlation matrices simulated according to $\mathcal{U}(C_5)$ have the Perron-Frobenius property, in contrast to about 0.8% in $d = 8$.
3. **Distribution of pairwise correlations is significantly shifted to the positive**
   [14, 31] find that the distribution of pairwise correlations, i.e. the off-diagonal entries of market correlation matrices, displays a positive mean, as opposed to the Beta($d/2$, $d/2$)-distribution on $[-1, 1]$ of Ω_{ij} when $\Omega \sim \mathcal{U}(C_d)$, cf. [18], which has mean 0. We are able to confirm these observations in our data set: Drawing $n = 1,000,000$ portfolios in different dimensions ranging from $d = 5$ to $d = 50$, the mean of pairwise correlations is always positive, and on average equal to 0.33. Figure 6 contrasts the histogram of pairwise correlations in our market of 395 assets to that of a random correlation matrix of the same size drawn from $\mathcal{U}(C_d)$.
4. **Scale-free property of the corresponding MST**
   [2, 36] investigate MSTs constructed on financial correlation matrices, and find that these MSTs exhibit the special structure of a so-called scale-free graph, which corresponds by definition to a power-law type degree distribution. This also results in a much higher probability of observing nodes with a high degree, i.e. a large number of neighbors, in contrast to random trees where this probability decays with increasing number of vertices, cf. [36]. Correspondingly, in scale-free trees one also observes more leaves, as the sum over the degrees of all vertices is fixed and the nodes with high degree already 'use up' a significant portion of this capacity.[7]
   This finding can already be confirmed in comparatively small portfolios of 20 assets: We compare the number of leaves encountered in $n = 1,000,000$ simulations of MSTs from random correlation matrices $\Omega \sim \mathcal{U}(C_{20})$ with the number of leaves encountered in the same number of MSTs from correlation matrices of 20 randomly drawn assets from our data pool, and find that MSTs based on market data exhibit in general a higher number of leaves, cf. Figure 6.

Considering the influence of these stylized facts on our quantities e_d, f_d, h_d, it is hard to anticipate which of these features triggers the completely different behavior of e_d in market and random correlation matrices. For f_d, the scale-free graph structure of empirical correlation matrices implies that the denominator is larger than in the simulated case. However, as we typically observe $f_d > 1$ for empirical correlation matrices, there must be an even stronger influence on the MVP weights that counters the influence of the higher portion of leaves. To analyze possible effects of the stylized facts on h_d, we first observe that the more leaves a tree structure on d nodes has, the smaller we expect its mean occupation layer to be. This expectation is extrapolated from observations of low-dimensional tree structures. Clearly a 'chain-like' graph has the highest mean occupation layer (with values $d(d + 1)/(2d + 1)$ for an odd number d of vertices, resp. $d/2$ for an even number of vertices), and a 'star-like' graph has the lowest possible mean occupation layer with value $(d - 1)/d$. Therefore, the denominator is smaller in the empirical case, leading to a higher value of h_d.

[7] [1] argue that this behaviour originates in a growing network with preferential attachment, i.e. new nodes are more likely to attach to highly connected nodes in the existing network, an interpretation that seems intuitive in a financial context. (Note that there exist other generating mechanisms that may result in a scale-free network.)

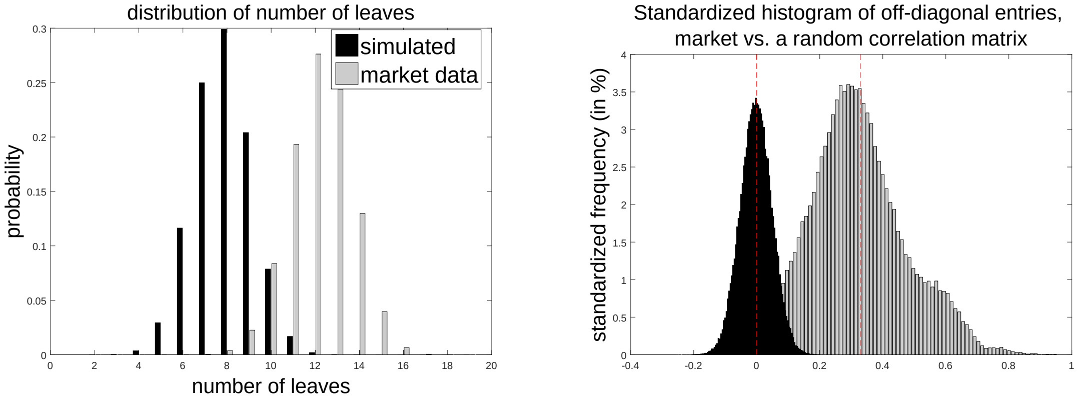


Figure 6: Left: Number of leaves encountered in uniformly simulated (black) vs. CDS data-based (gray) MSTs. Right: Standardized frequencies of the off-diagonal elements of a uniformly simulated (black) and our market correlation matrix (gray). The latter is significantly shifted to the right.

As in the quantity f_d, the numerator is affected both by MVP weights and graph structure. As we observe a higher value of h_d in the empirical case, higher MVP weights on the outskirts of the network compensate for the overall shortening of paths from the central node to the other nodes.

To analyze the effect of the observed features 1.-4. on the quantities e_d, f_d, h_d in more detail, the previous Monte Carlo studies should be rerun using simulation algorithms that are able to produce correlation matrices that display only a subset of these stylized facts, as opposed to completely random correlation matrices (which display none of them) and market correlation matrices (which exhibit them all). However, the choice of available algorithms for this task is limited:

- [7] present an algorithm that generates random correlation matrices with specified eigenvalues. We rerun our simulation study using this algorithm, combined with a realistic distribution for the eigenvalues, cf. Paragraph 3.4.1.
- Whereas there are some references that generate factor models (typically displaying stylized facts 1.-3., but not 4., at least in the one-factor case, cf. [2]), e.g. [6, 9], these papers rely strongly on specific characteristics of the market the authors are considering, and thus do not qualify as completely random realizations of correlation matrices exhibiting properties 1.-3. Nevertheless, to gain some insight in whether such correlation matrices may yield a relation between centrality measurements and MVP weights, we rerun our simulation study with one- and 3-factor model correlation matrices, cf. Paragraph 3.4.2.
- To the best of our knowledge, there is no algorithm available for the generation of reasonably random correlation matrices with the Perron-Frobenius property. [11] present an algorithm for the generation of covariance matrices whose diagonal, resp. off-diagonal elements follow a distribution with specified moments. However, this algorithm is not readily adaptable to the generation of correlation matrices with off-diagonal entries with specified moments. Concerning the generation of correlation matrices whose MSTs exhibit the scale-free property, to the best of our knowledge there is no algorithm available, and due to the generating mechanism of the MST we expect the task of finding such correlation matrices to be highly complex.

3.4.1 Correlation matrices with a realistic eigenvalue structure

Using the algorithm of [7], which is implemented in Matlab as `gallery('randcorr',...)`, we are able to analyze the influence of the presence of a large first eigenvalue. For dimensions $d = 5, 10, 50, 100$, we generate $n = 1,000,000$ simulations of eigenvalues $(\lambda_1, \dots, \lambda_d)$, where we fix the first eigenvalue $\lambda_1 = 0.4 \cdot d$, according to the typical size of the first eigenvalue in random portfolios drawn from our data set, and simulate

$\lambda_2, \dots, \lambda_d$ according to the density

$$f_\Omega(\lambda) = \frac{2}{(\lambda + 1)^3}, \tag{5}$$

and rescale in order to cover the remaining 60% of total variance. This is a special form of the power-law type density given in [3, 5], which is found to capture the distribution of the bulk of eigenvalues of market correlation matrices fairly well.[8] An artificial spectrum simulated in this way is very similar to the observed spectrum of an arbitrary correlation matrix from our data set. In the next step, we generate for each $(\lambda_1, \dots, \lambda_d)$ a random correlation matrix having this particular spectrum according to the algorithm by [7], and calculate e_d, f_d, h_d. This procedure is able to reproduce stylized fact 1., but not the others: Similar to $\mathcal{U}(C_d)$ the percentage of simulated correlation matrices with the Perron-Frobenius property is small and decreases fast with increasing dimension d. Pairwise correlation entries have a bimodal distribution, symmetric about 0, with mean close to 0. The histogram of leaves for MSTs of correlation matrices with these realistically simulated eigenvalues looks very similar to that obtained from the uniform distribution, so on average graphs derived from correlation matrices simulated from this algorithm exhibit a lot fewer leaves than those derived from market correlation matrices, which hints at stylized fact 4. also not being present. The results show that just the fact of displaying a realistically large first eigenvalue with a realistic distribution of the spectrum is not enough to explain the empirically observed relation between graph centrality and MVP weights. As for uniformly random simulated correlation matrices, the percentage of correlation matrices simulated according to the algorithm by [7] that exhibit a significant overweighting of central assets grows with dimension d, contrary to market correlation matrices where this percentage declines with d, cf. Figure 7 and Table 5. Similar results were obtained when simulating for each dimension a fixed spectrum according to (5), and generating n random correlation matrices with this fixed spectrum.

3.4.2 Factor model correlation matrices

Following a methodology similar to Fan et al. [9, Section 4], we simulate correlation matrices corresponding to a one-factor model. The distributional characteristics of the parameters are obtained from a fit of a one-factor model to our data set described in Section 3.3:

$$X_i(t) = b_i M(t) + \epsilon_i(t),$$

where M denotes the market factor (as a proxy we choose an equally weighted portfolio of all assets), $X_i, i = 1, \dots, 395$, is the i-th time series in our data set, b_i its factor loading, and ϵ_i the associated time series of errors. We find that the factor loadings b_i and the standard deviations σ_i of the error terms ϵ_i are both approximately gamma distributed, with parameters $\alpha_b = 0.4819$, $\beta_b = 1.6533$, and $\alpha_\sigma = 0.5400$, $\beta_\sigma = 0.0052$, respectively. In the simulation, the market factor is taken to be normally distributed[9], with mean and standard deviation matching the observed values, $\mu_M = 4.7 \cdot 10^{-5}$, and $\sigma_M = 0.0018$, factor loadings and error standard deviations are simulated from the above gamma distributions, and the error time series are simulated independently from normal distributions with zero mean and the respective simulated standard deviations. In $n = 1,000,000$ simulations, we obtain $d = 5, 10, 50, 100$ time series from a factor model with the above characteristics, and calculate the corresponding (sample) correlation matrices.

The largest eigenvalue of the simulated matrices explains on average about 40% of total variance for $d = 10, 50, 100$, and about 45% for $d = 5$. Contrary to our expectations, correlation matrices simulated from this one-factor model do not regularly exhibit the stylized fact 2.: The proportion of simulated correlation

[8] [3, 5]'s formula relies on the lower bound λ^- for the eigenvalues of a random correlation matrix in the sense of random matrix theory. We set $\lambda^- = 0$, which conforms to the random matrix theory limit when the data matrix is $N \times N$ and $N \to \infty$.

[9] This assumption would be questionable if one intends to describe our data set accurately. However, since we intend to construct time series from an artificial factor model, fitting the exact distribution of the factor returns is not crucial, and we stick with normality for the sake of simplicity.

matrices with Perron-Frobenius property steadily declines, from 62.50% in dimension 5 to 0.02% in dimension 100. The mean of pairwise correlations in our simulations is 0.23 on average, so stylized fact 3. is typically present. The MSTs associated with the one-factor correlation matrices exhibit on average more leaves than those obtained from uniformly random correlation matrices, but fewer leaves than those associated with empirically observed correlation matrices.

Concerning eigenvector centrality, the MVPs related to the simulated factor correlation matrices almost certainly overweight the 20% least central assets, regardless of the number of assets considered, cf. e_d 1-factor in Table 5. Also leaves seem to be consistently overweighted: f_d is smaller than 1 for only a low percentage of the simulated correlation matrices, with only a slight growth in dimension. In terms of mean occupation layer, the probability of underweighting peripheral assets, $\mathbb{P}(h_d < 1)$, grows with dimension, from 4.11% in dimension 5 to 27.71% in dimension 100. Thus, concerning h_d, correlation matrices simulated from this 1-factor model unexpectedly behave similar to correlation matrices simulated uniformly or from the randcorr algorithm described in Paragraph 3.4.1.

To shed more light on the behavior of the quantities e_d, f_d, h_d for factor models, we repeat our analysis with correlation matrices simulated according to the characteristics described in Fan et al. [9, Section 4]: There, a Fama-French 3-factor model was fit to daily data of 30 stock portfolios obtained from French's website (`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`), and factor loadings were found to approximately follow a trivariate normal distribution, and error standard deviations were found to approximately follow a gamma distribution.

In correlation matrices simulated from this model, we find that stylized facts 1.-3. are present: Regardless of dimension, over 99% of the simulated correlation matrices exhibit the Perron-Frobenius property, the first eigenvector on average explains more than 60% of total variance, and the distribution of the pairwise correlation entries is shifted to the positive with a mean of 0.61. We further find that the MSTs associated with 3-factor correlation matrices typically exhibit more leaves than those associated with empirically observed correlation matrices, thus hinting at a denser graph structure than typically exhibited by scale-free trees. This is in line with [2]'s finding for one-factor correlation matrices[10].

Concerning the quantities e_d, f_d, h_d, we find that regardless of dimension or centrality measure, peripheral assets are almost certainly overweighted in the MVPs associated with the 3-factor correlation matrices, cf. Table 5 and Figure 7.

4 Issues of graph-based asset allocation

Having demonstrated that graph-based portfolio selection mechanisms lack a fundamental connection to the traditional Markowitz approach, we further want to draw the reader's attention to potential problems that may arise in the context of graph-based portfolio selection. On the one hand, the chosen dependence and centrality measures may heavily influence the graph structure and the graph-based portfolio selection; on the other hand, by just taking into account correlations (or pairwise dependence measures), one loses the information captured by the marginal distribution of the assets or by higher-order dependence structures.
As a side remark, it is worth noting that certain graphs derived from the correlation matrix correspond to clustering techniques, e.g. the MST corresponds to single linkage clustering. Issues of clustering-based portfolio selection have been documented e.g. in [17].

[10] [2] simulate correlation matrices from a 1-factor model previously fitted to a large stock data set with the S&P500 index as market factor.

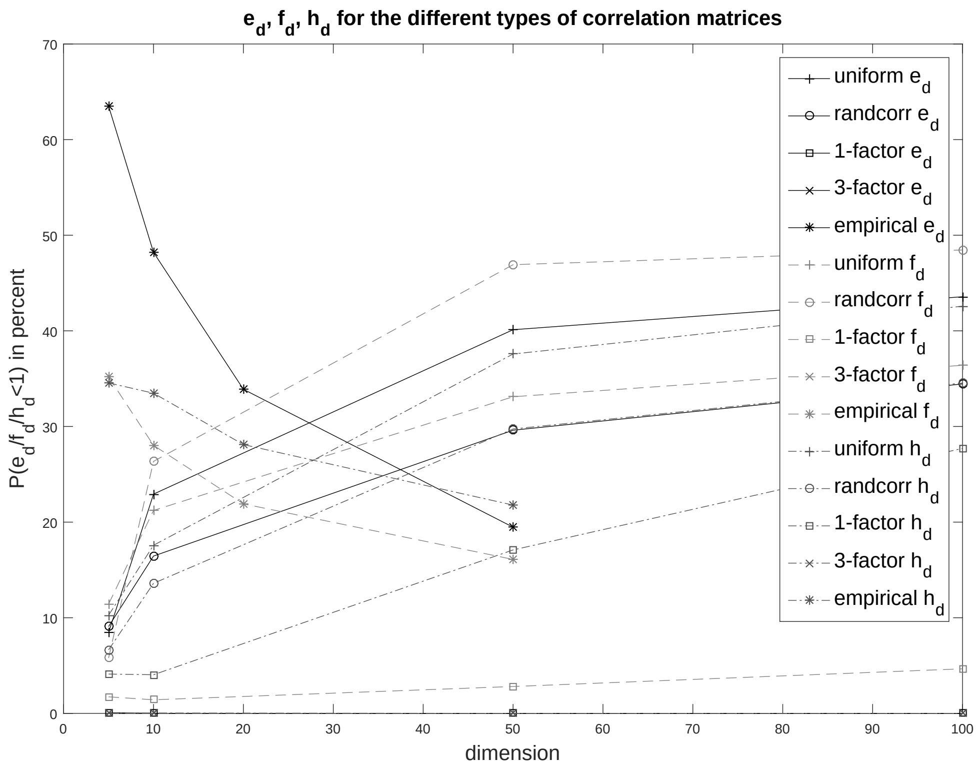


Figure 7: Probability of underweight for different types of correlation matrices: uniform, gallery:randcorr, one-and 3-factor, and empirical.

4.1 Graph structure depends on chosen dependence measure

For graph-based portfolio selection methods, any risk measure can be used for the construction of Σ. The resulting dependence matrix will be symmetric, and, unlike in the Markowitz setting, positive definiteness is not required[11] in the selection algorithms. However, one has to keep in mind that different dependence measures may yield different MSTs, as can be seen from the following toy example:[12] Consider $\mathbf{R} = (R_1, R_2, R_3)$, where R_i is lognormally distributed with parameters $\mu_i = 0$ and $\sigma_i > 0$ for $i = 1, 2, 3$, with $\sigma_1 = 0.5$, $\sigma_2 = 0.5$, and $\sigma_3 = 3$. The dependence structure of $\mathbf{R}$ is characterized by a Gaussian copula parameterized by the matrix

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix},$$

with $\rho_{12} = 0.1$, $\rho_{13} = 0.4$, and $\rho_{23} = 0.8$. Spearman's ρ and Kendall's τ are given as

$$\rho_S(X_i, X_j) = \frac{6}{\pi} \arcsin\left(\frac{\rho_{ij}}{2}\right), \qquad \tau(X_i, X_j) = \frac{2}{\pi} \arcsin(\rho_{ij}).$$

[11] Although not required, positive definiteness is a nice-to-have, as in this case the often used correlation distance provides a pseudometric on the set of considered assets.

[12] For an overview of the shortcomings of correlation-based MST and alternative dependence measures proposed in the literature, see e.g. the review [24].

Table 5: Probability of underweighting non-central assets in terms of $e_d/f_d/h_d$, $\mathbb{P}(e_d/f_d/h_d(\Omega) < 1)$, for different correlation matrices. Whereas the uniform and randcorr algorithms produce correlation matrices whose probability of underweighting non-central assets grows with dimension, factor models on the other hand almost certainly overweigh non-central assets.

$\mathbb{P}(\cdot_d < 1)$	$d = 5$	$d = 10$	$d = 20$	$d = 50$	$d = 100$
e_d uniform	8.45%	22.93%	-	40.12%	43.57%
e_d randcorr	9.08%	16.45%	-	29.61%	34.45%
e_d 1-factor	0.07%	0.01%	-	0%	0%
e_d 3-factor	0%	0%	-	0%	0%
e_d empirical	63.51%	48.21%	33.89%	19.47%	-
f_d uniform	11.42%	21.22%	-	33.11%	36.40%
f_d randcorr	5.79%	26.35%	-	46.92%	48.46%
f_d 1-factor	1.73%	1.42%	-	2.81%	4.66%
f_d 3-factor	0%	0%	-	0%	0%
f_d empirical	35.18%	28.03%	21.91%	16.10%	-
h_d uniform	10.26%	17.58%	-	37.58%	42.58%
h_d randcorr	6.64%	13.60%	-	29.74%	34.55%
h_d 1-factor	4.11%	4.04%	-	17.09%	27.71%
h_d 3-factor	0.03%	0.04%	-	0%	0%
h_d empirical	34.56%	33.48%	28.11%	21.78%	-

Both are strictly increasing transformations of the ρ_{ij}, so from $\rho_{12} < \rho_{13} < \rho_{23}$ it follows $\rho_S(X_1, X_2) < \rho_S(X_1, X_3) < \rho_S(X_2, X_3)$ and $\tau(X_1, X_2) < \tau(X_1, X_3) < \tau(X_2, X_3)$. On the other hand,

$$\begin{aligned}\operatorname{Cor}(X_1, X_3) = \frac{(e^{\rho_{13}\sigma_1\sigma_3} - 1)}{\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_3^2} - 1)}} \approx 0.017 < \operatorname{Cor}(X_2, X_3) &= \frac{(e^{\rho_{23}\sigma_2\sigma_3} - 1)}{\sqrt{(e^{\sigma_2^2} - 1)(e^{\sigma_3^2} - 1)}} \approx 0.048 \\ < \operatorname{Cor}(X_1, X_2) &= \frac{(e^{\rho_{12}\sigma_1\sigma_2} - 1)}{\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)}} \approx 0.089,\end{aligned}$$

so constructing the MST from (Pearson) correlations results in a different MST than construction from Spearman's ρ or Kendall's τ, as the central nodes differ, cf. Figure 8.

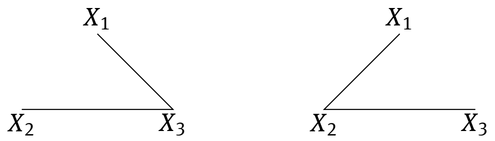


Figure 8: MSTs constructed from the different dependence measures using correlation distance as weight function. Left: ρ_S, τ. Right: Cor.

In an example with just three nodes, this may seem a minor issue at first glance. In practice, however, different tree structures for different dependence measures are often encountered, and the differences can be dramatic, as illustrated in Figure 9: The two MSTs constructed on return data of the SMI index constituents[13] using correlation resp. Spearman's ρ matrices in combination with a decreasing weight function are fundamentally different. Striking differences are e.g. the position of ZURN, which is rather central in the Spearman's

[13] Data from May 2015 to May 2017; Source: Bloomberg.

ρ MST, but is a leaf in the correlation MST, or the branch descending from BAER (UBSG, CSGN, ADEN), which is located in the center of the correlation MST, but rather peripheral in the Spearman's ρ MST. Table 6 presents the eigenvalues of the correlation resp. Spearman's ρ matrices, which differ only marginally, thus indicating that the two matrices are quite similar. The different tree structure is exclusively inferred by the marginal distributions of the return time series, which enter the calculation of the correlation coefficient, but not the calculation of Spearman's ρ. Table 6 further shows the annualized volatilities of the time series, indicating that the marginal distributions are diverse.

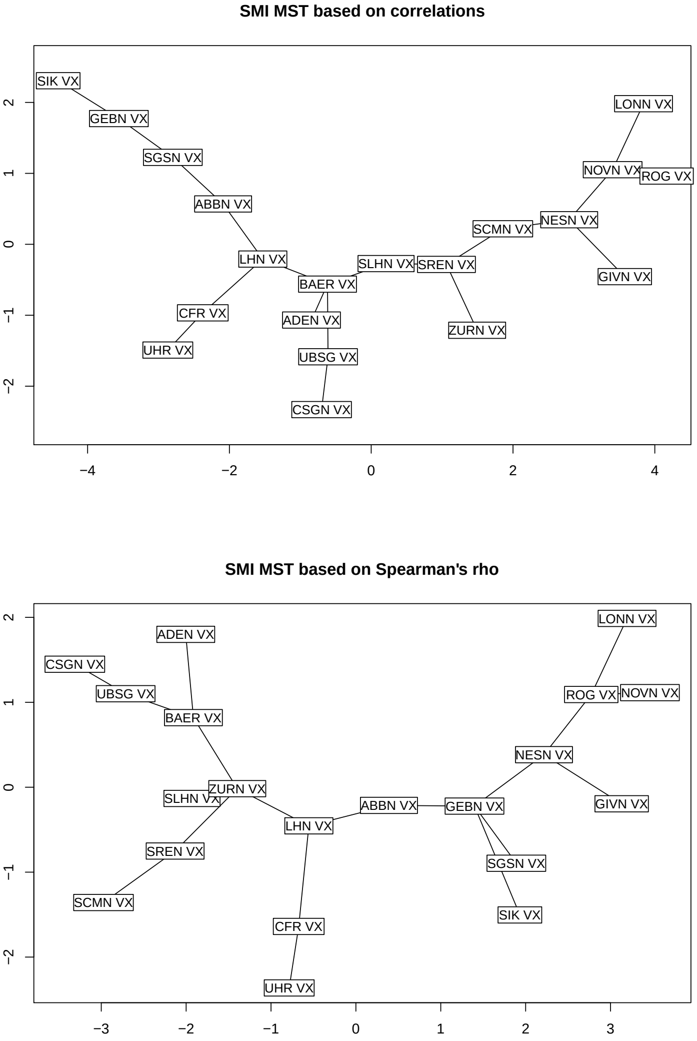


Figure 9: MSTs constructed from return data of the SMI Index constituents. Striking differences are for example the respective positions of ZURN and BAER in the networks.

4.2 Variances matter! And maybe even more matters!

Generally speaking, the 'performance' of a portfolio return $\mathbf{x}^T\mathbf{R}$ should be a measurement depending on the full distribution of $\mathbf{R}$. Only taking into account a partial aspect of the latter distribution bears the risk to

Table 6: Left: The eigenvalues of the correlation (EV corr.) and Spearman's ρ (EV rho) matrices indicate that the two matrices are quite similar. Right: Annualized volatilities of the SMI return time series.

EV corr.	EV rho
0.1583	0.1461
0.1759	0.1703
0.2139	0.1900
0.238	0.2292
0.2582	0.2662
0.2756	0.2871
0.2848	0.2976
0.3105	0.3281
0.3654	0.3568
0.3763	0.3824
0.3814	0.4001
0.4133	0.4218
0.4343	0.4368
0.5092	0.4600
0.5668	0.5451
0.6174	0.5836
0.8395	0.8020
1.0862	1.0741
1.4723	1.3879
11.0227	11.2347

firm	vol.
ABB ('ABBN VX')	0.2046
Adecco Group ('ADEN VX')	0.2696
Julius Baer Gruppe ('BAER VX')	0.2777
Cie. Fin. Richemont ('CFR VX')	0.2737
Credit Suisse Group ('CSGN VX')	0.3719
Geberit ('GEBN VX')	0.1874
Givaudan ('GIVN VX')	0.1874
Lafargeholcim ('LHN VX')	0.3203
Lonza Group ('LONN VX')	0.2332
Nestle ('NESN VX')	0.1597
Novartis ('NOVN VX')	0.1985
Roche Holding ('ROG VX')	0.1987
Swisscom ('SCMN VX')	0.1741
SGS ('SGSN VX')	0.1727
Swiss Life Holding ('SLHN VX')	0.2073
Swiss Re ('SREN VX')	0.1908
Sika ('SYNN VX')	0.2825
UBS ('UBSG VX')	0.3101
Swatch Group ('UHR VX')	0.2769
Zurich ('ZURN VX')	0.2381

overlook better performing portfolios.[14] Furthermore, if it is necessary to rely only on partial aspects, it is important to respect the hierarchy of the effects of the respective aspects on the result. What does that mean concretely? The approaches of [29, 33] base portfolio selection only on the correlation matrix Ω of $\mathbf{R}$, whereas Markowitz portfolio selection relies on the covariance matrix Σ. The information in the covariance matrix comprises the information in the correlation matrix, and in addition uses the information about the variances of the margins. The latter information, which is fully discarded in the selection processes proposed by [29, 33], has a massive effect on diversification when measured in terms of portfolio variance. In particular, if some components of $\mathbf{R}$ have a variance that is significantly larger than that of others, they are underweighted in the MVP irrespectively of the correlation matrix, which only has a secondary effect. For example, if the covariance matrix is

$$\Sigma = \begin{bmatrix} 100 & -0.1 & 0 \\ -0.1 & 100 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

then it is easily seen that the associated MVP assigns most weight to asset 3, simply because it has by far the smallest variance. However, it is easily checked, cf. Lemma 1(b), that asset 3 is not a leaf in any MST computed

[14] In the clustering literature, [8] make an effort to overcome this risk by designing a distance measure that incorporates both information from the margins and the dependence structure of the assets.

from the associated correlation matrix

$$\Omega = \begin{bmatrix} 1 & -0.001 & 0 \\ -0.001 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In this artificial example, the effect of the correlation structure is clearly dominated by the effect of the one-dimensional margins (variances). The MST-based selection process simply overlooks the fact that assets 1 and 2 have a high variance.

Furthermore, all graph-based methods (but also Markowitz's approach) essentially rely on dependence information between bivariate pairs only. Consequently, they may be prone to overlook important characteristics of the distribution of **R** resulting from higher-level dependence structures beyond those observed through bivariate pair measurements (such as included in Σ). Typically, these effects are of secondary importance in practice, but there are cases in which they do matter, as the following example emphasizes. Consider the following two stochastic models for **R**, denoted $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$, which both have exactly the same covariance matrix.

(1) Each $R_i^{(1)}$ is normally distributed with mean $\mu = 0.08$ and standard deviation $\sigma = 0.3$, and the survival copula of $\mathbf{R}^{(1)}$ is given by

$$C(u_1, \ldots, u_d) = u_{[1]} \prod_{k=2}^{d} u_{[k]}^{2^{1-k}},$$

where $u_{[1]} \leq \ldots \leq u_{[d]}$ denotes the ordered list of $u_1, \ldots, u_d$, i.e.

$$\mathbb{P}(R_1^{(1)} > x_1, \ldots, R_d^{(1)} > x_d) = C\left(1 - \Phi\left(\frac{x_1 - \mu}{\sigma}\right), \ldots, 1 - \Phi\left(\frac{x_d - \mu}{\sigma}\right)\right),$$

where Φ denotes the distribution function of a standard normally distributed random variable.

(2) Each $R_i^{(2)}$ is normally distributed with mean $\mu = 0.08$ and standard deviation $\sigma = 0.3$, and the survival copula of $\mathbf{R}^{(2)}$ is given by

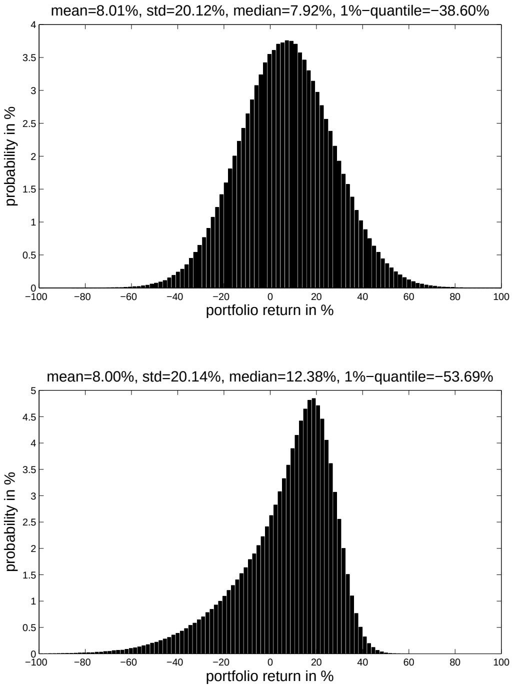
$$C(u_1, \ldots, u_d) = u_{[1]} \prod_{k=2}^{d} u_{[k]}^{\frac{1}{2}}.$$

Both copulas are within the family of Lévy-frailty copulas; see [20] for background on the latter, and the standard textbook [26] for background on copulas in general. Obviously, both models are such that $(R_i^{(1)}, R_j^{(1)})$ has the same distribution as $(R_i^{(2)}, R_j^{(2)})$, thus both models share the same covariance matrix Σ. All off-diagonal elements of Σ are equal, as are all its diagonal entries. Consequently, the MVP $\bar{\mathbf{x}}$ is an equally weighted portfolio in both cases. Regarding the portfolio derived from an MST, there is complete freedom. One finds an MST with $k \in \{2, \ldots, d-1\}$ arbitrary leaves.[15] While this example shows that the MST-based portfolio selection clearly needs further criteria, how different are the distributions of $\bar{\mathbf{x}}^T \mathbf{R}^{(1)}$ and $\bar{\mathbf{x}}^T \mathbf{R}^{(2)}$? Figure 10 illustrates that the variances and means of both portfolio returns are identical, but the shapes of their distributions differ dramatically. In particular, the second model is negatively skewed and has a significantly larger downside risk than the first.

5 Conclusion

It was demonstrated that there is no significant evidence for an inner-mathematical relation between Markowitz-optimality and centrality in a graph derived from a random correlation matrix. The intuitive

[15] The considered portfolios exhibit constant correlation matrices, i.e. all pairwise correlations are equal. The corresponding complete graph has the same weight on all edges, thus any of its spanning trees is minimal.


mean=8.01%, std=20.12%, median=7.92%, 1%−quantile=−38.60%

mean=8.00%, std=20.14%, median=12.38%, 1%−quantile=−53.69%

Figure 10: Visualization of the probability distribution of the MVP for $d = 20$. Top: model $\mathbf{R}^{(1)}$. Bottom: model $\mathbf{R}^{(2)}$.

heuristic argument stating that a group of non-central assets in a graph form a well-diversified portfolio in a Markowitz setting cannot be backed by mathematical arguments for correlation matrices with no special structure. Consequently, empirical findings in this direction must be considered highly data-dependent. Nevertheless, for large data sets of financial asset returns, this finding is persistent, thus should originate in specific features of the underlying correlation/covariance matrix. It was demonstrated that imposing a realistic eigenvalue structure (but none of the other features of observed financial correlation matrices) on simulated correlation matrices could not produce a relation between graph-centrality and Markowitz portfolio weights.

Acknowledgement: This work was supported by the German Research Foundation (DFG) and the Technical University of Munich within the funding programme Open Access Publishing. We also thank the anonymous referees for valuable comments and suggestions.

References

[1] Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *Science 286*, 509–512.

[2] Bonanno, G., G. Caldarelli, F. Lillo, and R. N. Mantegna (2003). Topology of correlation-based minimal spanning trees in real and model markets. *Phys. Rev. E 68*(4), 046130.

[3] Bouchaud, J.-P. and M. Potters (2011). Financial applications of random matrix theory: a short review. In G. Akemann, J. Baik, and P. Di Francesco (Eds.), *The Oxford Handbook of Random Matrix Theory*, pp. 824–850. Oxford University Press.

[4] Boyle, P. P., S. Feng, D. Melkuev, and J. Zhang (2014). Correlation matrices with the Perron-Frobenius property. Available at https://ssrn.com/abstract=2493844.

[5] Bun, J., J.-P. Bouchaud, and M. Potters (2017). Cleaning large correlation matrices: tools from random matrix theory. *Phys. Rep. 666*, 1–109.

[6] Cizeau, P., M. Potters, and J.-P. Bouchaud (2001). Correlation structure of extreme stock returns. *Quant. Finance 1*(2), 217–222.

[7] Davies, P. I. and N. J. Higham (2000). Numerically stable generation of correlation matrices and their factors. *BIT 40*(4), 640–651.

[8] Donnat, P., G. Marti, and P. Very (2016). Toward a generic representation of random variables for machine learning. *Pattern Recogn. Lett. 70*, 24–31.

[9] Fan, J., Y. Fan, and J. Lv (2008). High-dimensional covariance matrix estimation using a factor model. *J. Econometrics 147*(1), 186–197.

[10] Ghosh, S. and S. Henderson (2003). Behavior of the NORTA method for correlated random vector generation as the dimension increases. *ACM Trans. Model. Comput. Simul. 13*(3), 276–294.

[11] Hirschberger, M., Y. Qi, and R. E. Steuer (2007). Randomly generating portfolio-selection covariance matrices with specified distributional characteristics. *Eur. J. Oper. Res. 177*(3), 1610–1625.

[12] Jagannathan, R. and T. Ma (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *J. Finance 58*(4), 1651–1683.

[13] Kaya, H. (2015). Eccentricity in asset management. *J. Netw. Theory in Finance 1*(1), 1–32.

[14] Kazakov, M. and V. Kalyagin (2016). Spectral properties of financial correlation matrices. In V. Kalyagin, P. Koldanov, and P. Pardalos (Eds.), *Models, Algorithms and Technologies for Network Analysis*, pp. 135–156. Springer, New York.

[15] Laloux, L., P. Cizeau, J.-P. Bouchaud, and M. Potters (1999). Noise dressing of financial correlation matrices. *Phys. Rev. Lett. 83*(7), 1467–1470.

[16] Ledoit, O. and M. Wolf (2004). Honey, I shrunk the sample covariance matrix. *J. Portfol. Manage. Summer 30*(4), 110–119.

[17] Lemieux, V., P. S. Rahmdel, R. Walker, B. L. W. Wong, and M. Flood (2014). Clustering techniques and their effect on portfolio formation and risk analysis. *Proceedings of the International Workshop on Data Science for Macro-Modeling*, pp. 1–6. ACM.

[18] Lewandowski, D., D. Kurowicka, and H. Joe (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal. 100*(9), 1989–2001.

[19] Li, D., X. Sun, and J. Wang (2006). Optimal lot solution to cardinality constrained mean-variance formulation for portfolio selection. *Math. Finance 16*(1), 83–101.

[20] Mai, J.-F. and M. Scherer (2009). Lévy-frailty copulas. *J. Multivariate Anal. 100*(7), 1567–1585.

[21] Mantegna, R. N. (1999). Hierarchical structure in financial markets. *Eur. Phys. J. B 11*(1), 193–197.

[22] Markowitz, H. (1952). Portfolio selection. *J. Finance 7*(1), 77–91.

[23] Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*. Wiley, New York.

[24] Marti, G., F. Nielsen, M. Bińkowski, and P. Donnat (2017). A review of two decades of correlations, hierarchies, networks and clustering in financial markets. Available at https://arxiv.org/abs/1703.00485.

[25] Matoušek, J. and J. Nešetřil (2007). *Diskrete Mathematik: Eine Entdeckungsreise*. Second edition. Springer, Berlin.

[26] Nelsen, R. B. (2006). *An Introduction to Copulas*. Second edition. Springer, New York.

[27] Neuberg, R. and P. Glasserman (2017). Estimating a covariance matrix for market risk management and the case of credit default swaps. Columbia Business School Research Paper No. 16-39, available at SSRN: https://ssrn.com/abstract=2782107.

[28] Newman, M. E. (2008). Mathematics of networks. In S. Durlauf and L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics*, pp. 4059–4064. Second edition. Palgrave Macmillan, London.

[29] Onnela, J.-P., A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto (2003). Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E 68*(5), 056110.

[30] Peralta, G. and A. Zareei (2016). A network approach to portfolio selection. *J. Empirical Finance 38*(A), 157–180.

[31] Plerou, V., P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley (2002). Random matrix approach to cross correlations in financial data. *Phys. Rev. E 65*(6), 066126.

[32] Plerou, V., P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley (1999). Universal and nonuniversal properties of cross correlations in financial time series. *Phys. Rev. Lett. 83*(7), 1471–1474.

[33] Pozzi, F., T. Di Matteo, and T. Aste (2013). Spread of risk across financial markets: Better to invest in the peripheries. *Sci. Rep. 3*, Article ID 1665, 7 pages.

[34] Rockafellar, R. T. and S. Uryasev (2000). Optimization of conditional value-at-risk. *J. Risk 2*(3), 21–41.

[35] Roll, R. (1977). A critique of the asset pricing theory's tests part I: On past and potential testability of the theory. *J. Finan. Econ. 4*(2), 129–176.

[36] Vandewalle, N., F. Brisbois, and X. Tordoir (2001). Non-random topology of stock markets. *Quant. Finance 1*(3), 372–374.