

UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE

Ph.D Programme in

ECONOMICS AND POLICIES OF MARKETS AND FIRMS

Curriculum: Statistical Methods

Cycle XXXVIII

Doctoral Thesis in

ADVANCES IN NETWORK METHODS FOR FINANCIAL TIME SERIES ANALYSIS

Supervisor

Prof. Pietro CORETTO

PhD Candidate

Alfonso PELUSO

A.Y. 2025/2026

Contents

1 Literature Review **5**

1.1 Introduction 5

1.2 Networks: Theoretical foundations 6

1.2.1 Centrality measures 7

1.2.2 Community detection on networks 9

1.3 Networks from stock returns time series 13

1.4 Stock correlation networks 16

1.4.1 Threshold filters 17

1.4.2 Topological filters 18

1.5 Community detection in stock correlation networks 22

1.6 Applications 28

1.7 Other Financial Networks 31

1.8 Concluding remarks 32

2 Robust Multi-Source Financial Network Analysis **34**

2.1 Introduction 34

2.2 Standard analysis framework 36

2.3 The case for robustness 39

2.4 Information from different data sources 49

2.4.1 Data integration and Similarity Network Fusion 52

2.4.2 Cross-variable SNF 54
2.4.3 Cross-frequency SNF 56
2.5 Detecting communities via spectral clustering 59
2.6 Empirical Analysis 62
2.6.1 Robustness Consistently Pays Off 66
2.6.2 Uncovering Enhanced Structures via Horizontal Data Integration 69
2.6.3 Vertical Data Integration Reveals the Most Coherent Group Structures 73
2.6.4 Meta-Clusters and Co-Occurrence Networks: A Higher-Level Interpretation and Validation Framework 79
2.7 Concluding remarks 84

3 Stochastic Block Models for Large-Scale Financial Network Analysis 86

3.1 Introduction 86
3.2 Joint reconstruction-community detection 88
3.3 Graph Clustering and Network Reconstruction Evaluation 95
3.4 Empirical analysis 100
3.4.1 Static 102
3.4.2 Dynamic 106
3.5 Concluding remarks 111
Conclusions 116

Introduction

The thesis is composed of three interconnected chapters, each examining community detection in financial networks from a distinct perspective. The first chapter provides a broad overview of the most relevant literature on network studies in finance, with a particular focus on networks reconstructed from financial time series. It highlights their community structures and discusses how these structures have been used to implement risk diversification and asset allocation strategies, whether based on the overall network topology or on the organization of communities. The second chapter constitutes the main contribution of this thesis and seeks to advance the state of the art by improving standard methodologies along three dimensions: robustness, the joint consideration of multiple data types, and the integration of different frequencies. The chapter demonstrates how these improvements lead to more reliable and informative analyses, showing clear advantages over existing approaches across various aspects of community detection. The third chapter then investigates the application of generative community detection methods to financial time series, offering an alternative perspective to the correlation-based approaches employed in the previous chapter and prevailing in most studies of this kind. This exploration aims to extend the analytical framework by considering how generative models can capture structural regularities that might not be directly observable from empirical correlations alone.

Chapter 1

Literature Review

1.1 Introduction

Network science is an interdisciplinary field whose origins trace back to the classical Königsberg Bridges problem, the solution of which laid the foundations of graph theory. Its interdisciplinary nature stems not only from the diversity of tools it draws upon, including mathematics, statistics, physics, computer science, and sociology, but also from the wide range of scientific domains in which it has been successfully applied. Over the years, network science has contributed to major advances in fields such as biology Barabasi and Oltvai (2004), epidemiology Pastor-Satorras et al. (2015), medicine Barabási et al. (2011), neuroscience Bassett and Sporns (2017), sociology Borgatti et al. (2009), computer science Zhou et al. (2020), the study of physical and digital infrastructures Boccaletti et al. (2006), as well as various branches of economics Jackson et al. (2008); Hidalgo and Hausmann (2009) and finance Caccioli et al. (2018).

Network science has been extensively applied in finance since the early 2000s, and its relevance grew substantially after the 2008 global financial crisis, which highlighted the inability of traditional models to account for phenomena such as cascad-

ing failures, effects that arise naturally in network-based representations of system interconnectedness. Because network concepts have been deployed across a wide range of financial contexts, the term "financial network" has become too broad to be associated with a single type of structure or methodology. As a result, financial networks now encompass diverse constructions that differ in their underlying data, the analytical techniques applied, and the specific objectives pursued in each line of research.

The following section introduces the fundamental theoretical concepts underpinning network science. It begins with the formal definition of a graph in its various representations, proceeds with an examination of centrality measures, and concludes with an introduction to community structures, a central component in the study of network organization.

Subsequently, the discussion turns to networks constructed from stock return time series, with a particular focus on stock correlation networks. Both methodological developments and their applications to asset allocation and risk diversification are reviewed. A broader overview of additional successful applications of network science in finance is then provided.

1.2 Networks: Theoretical foundations

Network science builds around the concept of graph, a mathematical object conceived to describe relations among objects. A graph can be defined as a collection of vertices V and edges E:

$$G = (V, E)$$

in which vertices V represent the objects and edges E represent some kind of pairwise relations between pairs of vertices.

A graph is usually defined by a square matrix, the adjacency matrix A whose entries

indicate the existence and, as it will be explained later, the entity of relations between any two vertices. A first classification of graphs can be made based upon the type of pairwise relations they represent. The simplest form is the unweighted undirected graph $G = (V, E)$ where relations between vertices are represented by 0/1 entries in the adjacency matrix and the edges have no directions ($v_1 \to v_2 = v_2 \to v_1$). Unweighted undirected graphs therefore provide the most basic information about the relations between entities concerning the existence / absence of a relation between the vertices.

A more informative type of graph is the weighted graph $G(V, E, W)$ in which edges have an associated weight W which represents the magnitude of the relation between the entities, providing a distinction and a notion of importance for the edges.

A different type of information is provided by directed graphs, in which the relations have an associated direction indicating that the order in the edge set matter and, generally, $v_1 \to v_2 \neq v_2 \to v_1$. A further refinement providing the most complete information is represented by the weighted directed graph in which edges are both directional and have an associated weight.

As shown below the adjacency matrices associated with the different type of graphs: unweighted (A_u), weighted (A_w) and directed (A_d) are significantly different and have different properties.

$$A_u = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \qquad A_w = \begin{pmatrix} 0 & 0.4 & 1.2 \\ 0.4 & 0 & 0 \\ 1.2 & 0 & 0 \end{pmatrix} \qquad A_d = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

1.2.1 Centrality measures

The notion of connectedness naturally associated to the graph gives the space to the definition of some measures of importance of a vertex or an edge based on the

connection patterns it is part of. The research on centrality metrics has a long history and there are several important centrality metrics introduced over the years, each one conveying a different definition of centrality with it. Here we will briefly introduce the most relevant measures whose basic understanding is fundamental to fully appreciate the remainder of the thesis. The first and most straightforward centrality measure is the vertex degree D defined as the number of connections a vertex has, or, differently, the number of vertices a vertex is connected to. For and unweighted undirected graph the degree of a node i is defined as:

$$D(i) = \sum_j A_{ij},$$

where the $A_{i,j}$s are the entries of the adjacency matrix $\mathbf{A}$ corresponding to node i. This measure can be generalized to weighted networks where it is defined as strength and corresponds to the total weights of the edges connected to node i. For what concerns directed networks we can distinguish between the in-degree given by the sum of incoming edges (from other vertices to vertex i) and out-degree (from vertex i to other vertices).

A complementary perspective on node centrality is given by the betweenness centrality, which basically measures the extent to which a vertex stands between other vertices. Betweenness centrality is based on measuring how often a vertex lies on shortest paths [1] between other vertices. For a node i betwenness centrality is defined as:

$$B(i) = \sum_{j \neq i \neq k} \frac{SP_{jk}(i)}{SP_{jk}},$$

where SP_{jk} is the total number of shortest paths between vertices j and k, and $SP_{jk}(i)$ is the number of such shortest paths that pass through vertex i.

[1] a shortest path between two vertices is a sequence of vertices and edges connecting them such that no other route uses fewer steps

Shortest paths form the basis also for the closeness centrality which basically measures how close a vertex is to all other vertices in the graph. Closeness centrality is given by the reciprocal of the sum of the shortest path (connecting a vertex to all other vertices) lengths. For a vertex i

$$C(i) = \frac{1}{\sum_j \ell_{SP}(i,j)},$$

where ℓ_{SP} is the length of the shortest path between i and j. The last two measures, betweenness and closeness centrality can be generalized to both weighted and directed networks by switching to, respectively, weighted shortest paths[2] and directional shortest paths[3]. Here we presented a concise, basic overview of a few important centrality measures as a comprehensive treatment of all existing metrics is neither feasible nor required within the scope of this work.

1.2.2 Community detection on networks

Besides measuring the centrality of vertices another important task in network science is understanding the mesoscale structure of a system by grouping the vertices into communities. This is one of the most widely studied topics since the early days of network science and an extended exposition of the topic would be too much for the scope of this thesis; therefore, we will briefly introduce the intuition behind the most relevant approaches in the literature and, in the next sections, focus on the specifics of some of these methodologies applied in finance.

The main goal of community detection is to partition a graph $G = (V, E)$ into groups whose internal connections are denser than the external ones. This can be achieved through a variety of methods building upon different theoretical frameworks.

[2]Instead of counting steps, the weighted shortest path is the one whose total weight is minimal

[3]In this case the route of the shortest path must follow the direction of the edges, so only paths consistent with the edge directions can be considered.

A first method stems directly from the goal defined above and consists of dividing the network by "cutting" edges between groups. Here the community partition is achieved by minimizing the edges to "cut" (Wu and Leahy (2002)), hence the name min-cut. For subset of vertices $S \subset V$ from a weighted graph $G(V, E, W)$ the cut is defined as:

$$\text{cut}(S, \bar{S}) = \sum_{i \in S, j \in \bar{S}} W_{ij},$$

where $\bar{S}$ denotes the set of vertices V that are not in S and W_{ij} is the weight of the edge between nodes i and j. The min-cut algorithm seeks to minimize the cut. It is important to note that min-cut produces a bipartition of the network, dividing it into exactly two groups. To obtain finer community structures with more than two communities, the method must be applied recursively to each resulting partition. Furthermore, minimizing $\text{cut}(S, \bar{S})$ alone tends to select unbalanced partitions. To avoid that the normalized cut was introduced by Shi and Malik (2000):

$$\text{Ncut}(S, \bar{S}) \ = \ \frac{\text{cut}(S, \bar{S})}{\text{vol}(S)} + \frac{\text{cut}(S, \bar{S})}{\text{vol}(\bar{S})},$$

where $\text{vol}(S) = \sum_{i \in S} \sum_j W_{ij}$ is the total volume defined by the sum of the strengths (weighted degrees) of all the vertices in subset S. The normalized cut has a strong connection with spectral partitioning methods. Meilă and Shi (2001) proved that minimizing Ncut is approximately equivalent to finding the eigenvector of the second smallest eigenvalue of the normalized graph Laplacian, showing that both edge-cutting and spectral approaches solve the same underlying optimization from a random-walk [4] perspective.

Another class of methods is based on the maximization of a quantity called modularity (Girvan and Newman (2002), Newman and Girvan (2004)). The basic idea

[4] here "random walk" denotes a graph-based random walk, a Markov chain on the vertex set with transition probabilities determined by the network structure, in contrast to the classical random-walk processes extensively used in finance.

is that a good partitioning should have an internal edge density greater than the one expected in a random graph. Formally, for a partition of the vertices V into communities c, the modularity is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \Big(W_{ij} - \frac{k_i k_j}{2m} \Big) \delta(c_i, c_j),$$

where $k_i = \sum_j W_{ij}$ is the strength of node i, $m = \frac{1}{2} \sum_{i,j} W_{ij}$ represents the total weight of all edges, and $\delta(c_i, c_j) = 1$ if nodes i and j belong to the same community and 0 otherwise. The quantity $\frac{k_i k_j}{2m}$ represents the expected weight of the edge between nodes i and j in the configuration model (Newman et al. (2001)), a null model that preserves the degree sequence of the original graph while randomizing connections. Modularity based algorithms seek to find the community partition which maximizes Q. Despite its popularity, modularity maximization has important conceptual and practical limitations. As pointed out by Peixoto (2023) modularity maximization can lead to overfitting, ambiguous interpretations, and resolution limits.

A further different perspective is provided by Infomap Rosvall and Bergstrom (2008), which leverages the dynamics of information flow to reveal community structure. Infomap interprets the communities as regions of the graph where random walks [5] tend to be trapped for longer periods. Formally, considering a random walker on a graph $G = (V, E, W)$, with transition probabilities

$$P_{ij} = \frac{W_{ij}}{k_i}, \quad k_i = \sum_j W_{ij},$$

[5] also here the term "random walk" denotes a graph-based random walk, a Markov chain on the vertex set with transition probabilities determined by the network structure, in contrast to the classical random-walk processes extensively used in finance.

and stationary distribution π_j satisfying $\pi_j = \sum_i \pi_i P_{ij}$, Infomap seeks a partition $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ of the nodes into communities that minimizes the expected description length of a trajectory of the random walker under a two-level coding scheme. The associated *map equation* is

$$L(\mathcal{C}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{k} p_{i\circlearrowright} H(\mathcal{P}^i),$$

where $q_{\curvearrowright}$ is the probability of exiting any module, $H(\mathcal{Q})$ is the entropy of the module codebook, $p_{i\circlearrowright}$ is the probability of visiting nodes inside module i and $H(\mathcal{P}^i)$ is the entropy of the codebook for module i. The minimization of $L(\mathcal{C})$ identifies a partition where most of the random walk remains within modules. This flow-based perspective allows Infomap to detect smaller or nested communities that might be missed by purely density-based measures, and it naturally extends to weighted and directed networks.

Lastly, another class of methods that has gained prominence in recent years is represented by community detection frameworks based on generative modeling, particularly the Stochastic Block Model (SBM). These methods have been shown to perform very well in recovering meaningful community structure in diverse networks. Since SBMs are the central subject of Chapter 3, a full description is deferred to that chapter.

Besides the approaches introduced here, several other methods exist in the literature (Fortunato (2010)). Some of these, based on traditional clustering techniques, remain the standard in certain types of financial networks and will be described in the following sections.

1.3 Networks from stock returns time series

The methodologies introduced above operate on an explicitly defined graph and therefore assume that the network structure is already given. In many empirical settings this is not the case, and the graph itself must be reconstructed from raw data before any centrality metric or community detection method can be meaningfully applied. Finance is no exception since when dealing with collections of asset price or return time series the underlying network is not directly observed. Therefore, the network must be inferred from the statistical relationships among the asset returns time series. In this context the construction of a meaningful and robust financial network becomes a crucial preliminary step, as it determines the structure on which the techniques discussed earlier will be applied.

There exist numerous contributions from both econometrics and econophysics that propose different ways of reconstructing a network from financial time series, often tailored to distinct modeling assumptions and empirical objectives. Depending on the purpose, be it risk propagation, market structure identification, portfolio diversification or systemic stability analysis, different reconstruction strategies may be more appropriate. In what follows, we outline the most widely applied approaches for building networks from stock returns, which form the foundations for much of the subsequent contributions in the literature.

Spillover Networks. One of the first contributions from the Econometrics literature is the work of Diebold and Yilmaz (Diebold and Yilmaz (2009), Diebold and Yilmaz (2012), Diebold and Yılmaz (2014)) which can be considered the first and probably most successful methodological framework combining econometric and network science tools. The fundamental idea of this approach is that of using forecast error variance decomposition within a VAR framework to measure the interconnectedness of financial institutions. In detail they implement a two step approach in

which the first step consists of estimating a $VAR(p)$ model for an N-dimensional set of time series:

$$\mathbf{y}_t = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \Sigma).$$

In the second step, using the moving average (MA) representation of the VAR, the H-step-ahead generalized forecast error variance decomposition ($GFEVD$) is computed.

This quantifies the contribution of shocks in variable j to the forecast error variance of variable i:

$$\theta_{ij}^{(H)} = \frac{\sigma_{jj}^{-1} \sum_{h=0}^{H-1} (e_i' \Phi_h \Sigma e_j)^2}{\sum_{h=0}^{H-1} (e_i' \Phi_h \Sigma \Phi_h' e_i)},$$

where Φ_h is the MA coefficient matrix at horizon h, e_i is a selection vector, and Σ is the residual covariance matrix. The resulting matrix $\Theta = [\theta_{ij}^{(H)}]$ is interpreted as a weighted, directed adjacency matrix representing a network of shock transmission and is used to compute both node-level and global spillover metrics.

Network Estimation for Time Series (NETS). Barigozzi and Brownlees (2019) introduced NETS, a methodology for recovering sparse causal networks from high-dimensional financial time series by estimating a penalized VAR model. Formally, the approach considers a N-dimensional time series $\mathbf{Y}_t$ following a $VAR(p)$ process:

$$\mathbf{y}_t = \sum_{j=1}^{N} \mathbf{A}_j \mathbf{y}_{t-j} + \boldsymbol{\varepsilon}_t,$$

where both the autoregressive coefficient matrices $\mathbf{A}_j$ and the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ of innovations are assumed sparse. Two distinct network structures are extracted:

- a contemporaneous network from the sparsity pattern of $\boldsymbol{\Omega}$ representing instantaneous relationships and shock transmission;

- a dynamic Granger network from the $\mathbf{A}_j$ matrices capturing predictive relationships and lead-lag dependencies.

The NETS estimation procedure employs a two-step LASSO-based approach: first estimating VAR coefficients:

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \left\{ \frac{1}{2T} \sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{A}\mathbf{x}_{t-1}\|^2 + \lambda_A \|\mathbf{A}\|_1 \right\},$$

and then estimating the precision matrix using residuals through:

$$\hat{\mathbf{\Omega}} = \arg\min_{\mathbf{\Omega}} \left\{ -\log\det(\mathbf{\Omega}) + \mathrm{tr}(\hat{\mathbf{S}}\mathbf{\Omega}) + \lambda_\Omega \|\mathbf{\Omega}\|_1 \right\},$$

with regularization parameters selected via information criteria or cross-validation. They apply the framework to a panel of volatilities of US blue chips stocks between 2004-2015 and using general centrality metrics find a general high degree of interconnectedness particularly for financial firms.

Granger causality networks. Another widely used methodology from the Econometrics literature is the one introduced in Billio et al. (2012) in which the key step is to build a directed network from pairwise Granger causality tests between the time series associated to the nodes in the network.

This work is considered foundational in the field and introduced several measures of connectedness that are widely used both within and beyond academia. Its primary innovation, however, lies in the application of Granger causality tests Granger (1969) to construct directed networks for the purpose of understanding how shocks propagate through the financial system.

The network is built by testing whether institution j's lagged returns contain information useful for predicting institution i's future returns beyond what i's own history provides, the approach is able to capture lead-lag relationships that could emerge

from portfolio rebalancing, funding constraints, information asymmetries, and contagion effects that are central to understanding how financial distress spreads across institutions and sectors. The network construction methodology employs VAR models estimated over rolling windows, where each institution's returns are regressed on lagged values of all other institutions' returns according to:

$$r_{i,t} = \alpha_i + \sum_{k=1}^{p} \beta_{i,i}^{(k)} r_{i,t-k} + \sum_{j \neq i} \sum_{k=1}^{p} \beta_{i,j}^{(k)} r_{j,t-k} + \varepsilon_{i,t},$$

with the null hypothesis $H_0 : \beta_{i,j}^{(k)} = 0$ for all lags k tested using F-statistics to determine whether j Granger-causes i.

The authors use this methodology to analyze a set of monthly returns time series for hedge funds, broker/dealers, banks, and insurers and find that the Granger-causality networks exhibit dramatic structural changes over time, indicating that predictive relationships proliferate when the financial system comes under stress. They also show that network-based measures provide superior early warning capabilities for identifying periods of heightened systemic risk, with network density and centrality metrics exhibiting significant predictive power for subsequent market downturns.

1.4 Stock correlation networks

Among the various methodologies for reconstructing networks from stock return time series, the approach introduced by Mantegna (1999) is the most widely applied, particularly in the context of community detection. This methodology constructs the network directly from the stock return correlation matrix by transforming it into

a dissimilarity measure using the following transformation:

$$d_{i,j} = \sqrt{2(1 - \rho_{i,j})}. \tag{1.1}$$

Since the correlation matrix, and consequently the dissimilarity matrix, contains no inherently non-zero entries it is necessary to apply a filtering procedure to extract a meaningful network structure. Over the years various methods have been proposed to obtain a sparse representation of the dissimilarity matrix. These methods can be broadly classified into two categories: i) threshold-based filters, which retain only the elements not exceeding a specified threshold, and ii) topological filters, which enforce a specific topological structure on the resulting network.

1.4.1 Threshold filters

Threshold filters Onnela et al. (2003a), as the name implies, are based on the application of a threshold θ to the dissimilarity matrix. Each entry of the dissimilarity matrix for which $d > \theta$ is set to zero, resulting in the following filtered dissimilarity matrix:

$$d_{i,j} = \begin{cases} d_{i,j} & \text{if } d_{i,j} < \theta \\ 0 & \text{otherwise} \end{cases}$$

This approach has been widely employed in numerous studies, however, it presents certain limitations. In particular, weak but potentially important correlations are discarded when set to zero. Another critical issue is the selection of an appropriate threshold, as a threshold that is too stringent may produce a disconnected network, whereas a more permissive threshold can introduce spurious edges, thereby compromising the analysis. These challenges are especially pronounced when analyzing time-evolving networks.

As a result, topological filters have emerged as a common solution, and they are

described in detail in the next section.

1.4.2 Topological filters

Topological filters constitute another class of filtering methodologies for the dissimilarity matrix and are among the most widely used in practice. Indeed, the seminal work of Mantegna (1999) employed a topological filter to extract the backbone of the network.

The fundamental idea is to impose a specific topological structure on the resulting graph and to filter the dissimilarity matrix conditional on this imposed topology. Three principal topological filtering methodologies are commonly employed: the minimum spanning tree (MST), the planar maximally filtered graph (PMFG), and the triangulated maximally filtered graph (TMFG).

Minimum Spanning Tree. The Minimum Spanning Tree (MST) was the first technique employed to extract a network from the stock return correlation matrix, as introduced in Mantegna (1999). The MST imposes a tree structure on the network, which is an acyclic graph that connects all nodes with $N - 1$ edges while minimizing the total sum of dissimilarities. It is commonly constructed using Kruskal's algorithm:

1. Compute the dissimilarity matrix $\mathbf{D}$ from the correlation matrix;
2. Create a list of all possible edges (i, j) sorted in ascending order by d_{ij};
3. Initialize each node as a separate tree;
4. Iterate through the sorted edge list:
    - For each edge (i, j), add it to the MST if it connects two different trees (i.e., no cycle is formed);

5. Continue until $N-1$ edges have been added;

The MST therefore preserves the most significant relationships while ensuring that the graph remains connected and free of loops or cliques. However, as the simplest possible network structure, the tree can be considered overly restrictive, since it discards a substantial amount of information.

Planar Maximally Filtered Graph. The Planar Maximally Filtered Graph (PMFG) Tumminello et al. (2005) generalizes the MST by permitting additional edges, up to $3(N-2)$, while preserving planarity. The key distinction from the MST is that the resulting graph is not restricted to a tree but is a general planar graph. This allows more edges to be retained and therefore captures additional information, albeit at the cost of increased computational complexity. Specifically, the PMFG preserves the most significant links under the constraint that the graph can be embedded in a plane without edge crossings, and it is constructed using a greedy algorithm:

1. Compute the dissimilarity matrix $\mathbf{D}$;
2. Sort all pairwise dissimilarities d_{ij} in ascending order;
3. Initialize an empty graph;
4. Iterate through the sorted edge list:
    - Add edge (i, j) if it does not violate planarity (checked using planarity test algorithms);
5. Stop when the graph reaches $3(N-2)$ edges;

The PMFG construction shares many similarities with the MST algorithm, but it includes an additional planarity check that significantly increases computational

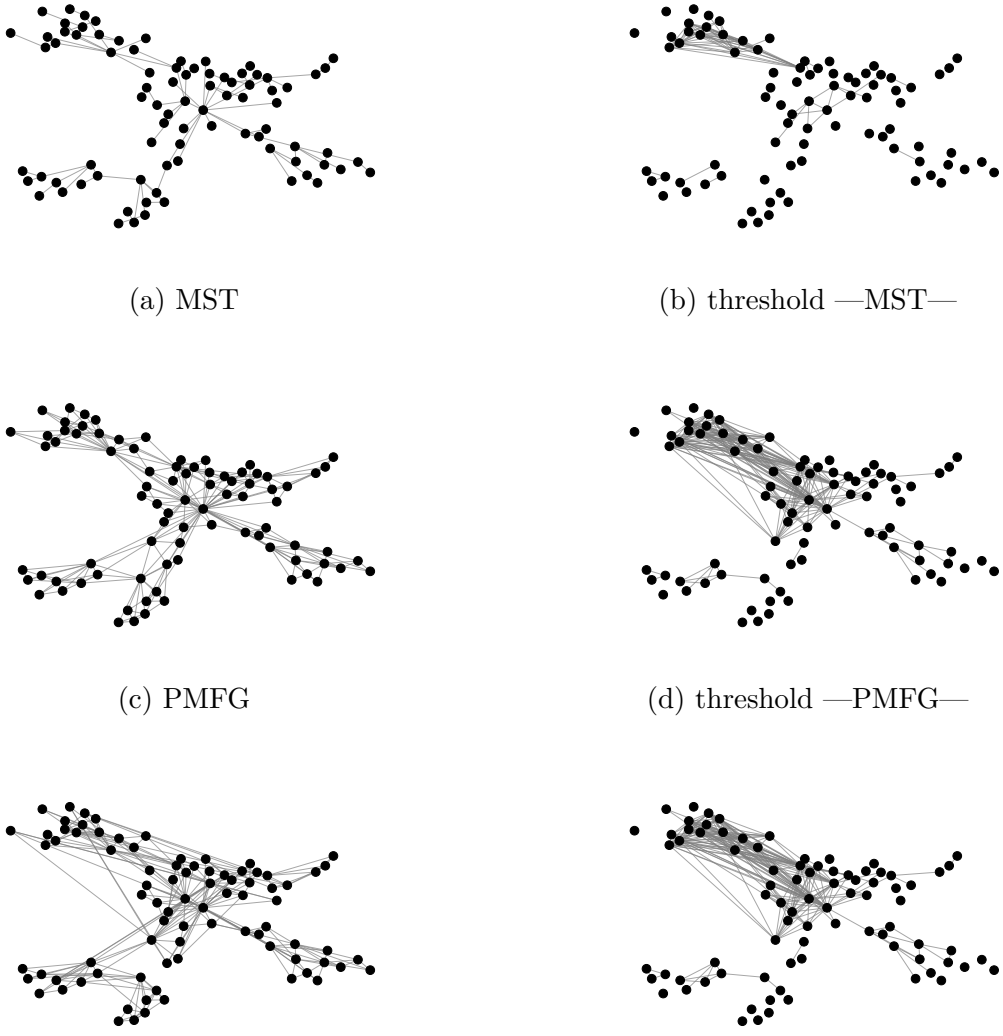
complexity. At each iteration, the procedure must verify whether adding a candidate edge preserves the planarity of the resulting graph. Since the number of edges in a complete graph grows proportionally to the square of the number of nodes, the overall computational complexity of the algorithm can be shown to scale as $\mathcal{O}(N^3)$.

Triangulated Maximally Filtered Graph. The Triangulated Maximally Filtered Graph (TMFG) Massara et al. (2016) is a planar graph constructed through local operations that triangulate the network while maximizing the information retained. Because it relies on local moves rather than global planarity checks, it is significantly faster to compute than the PMFG. Its original formulation is based on a similarity matrix, but, as shown in Briola and Aste (2022), can also be adapted to dissimilarity matrices. The construction follows the procedure outlined below:

1. Create an ordered list of all node pairs (i, j) sorted by increasing dissimilarity $d_{i,j}$;
2. Identify the 4 nodes with the lowest total dissimilarity ;
3. Connect these 4 nodes into a tetrahedron (a 4-clique with 4 triangular faces);
4. While the number of edges is less than $3N - 6$:
    - For each node not yet in the graph, consider all existing triangular faces;
    - Compute the gain (or minimized cost) of inserting that node into each triangle, forming three new triangular faces;
    - Select the insertion (node and triangle) that optimizes this criterion;
    - Insert the node, connect it to the three nodes of the selected triangle, and update the graph;

The TMFG offers two primary advantages compared to the PMFG: i) it is more computationally efficient due to its reliance on local moves, and ii) it is a chordal

graph, which allows its use in graph-based statistical inference models. For these reasons, it is increasingly adopted as the standard topological filter in such studies. Figure 1.1 illustrates the MST, PMFG, and TMFG on the left, alongside the corresponding threshold networks on the right, obtained by retaining the same number of edges as in the respective filtered graphs. These networks are constructed from the stocks composing the S&P 100 index, and the resulting structures differ markedly, potentially leading to substantially different outcomes in analytical applications.



(a) MST

(b) threshold —MST—

(c) PMFG

(d) threshold —PMFG—

(e) TMFG

(f) threshold —TMFG—

Figure 1.1: Different filtering methodologies.

1.5 Community detection in stock correlation networks

As discussed above, understanding the mesoscale organization of networks through community detection is a key area of research in complex network theory. Identifying communities not only provides deeper insight into the structure and dynamics of a system but also enables dimensionality reduction by representing each community as a single node. In the context of financial networks, the study of community structures has been a central topic since the field's early development, beginning with the pioneering work of Mantegna (1999).

Traditional Hierarchical Clustering. Hierarchical clustering is the most commonly applied methodology in such studies, serving both exploratory analyses and asset allocation purposes. This approach generates a hierarchical organization of partitions, providing a theoretically grounded interpretation of the dissimilarities between individual elements as well as between clusters. Hierarchical clustering methods employ a linkage criterion to measure dissimilarities between clusters. The most widely used linkage criteria are single linkage, complete linkage, and average linkage. Single linkage computes the distance between two clusters as the minimum distance between any pair of points from the respective clusters, often resulting in elongated, chain-like cluster formations. In contrast, complete linkage defines the distance as the maximum distance between any pair of points, which promotes the formation of compact, spherical clusters. Average linkage calculates the distance as the average of all pairwise distances between points in the two clusters, yielding results that fall between the two extremes. For further details on these and other linkage criteria approaches, see Hastie et al. (2009).

These methods use as input the dissimilarity matrix derived from the cross-

correlations of stock return time series and are closely related to spanning tree extraction. Specifically, the MST corresponds to the hierarchy obtained through single linkage clustering, whereas an alternative spanning tree, the Average Linkage Minimum Spanning Tree Tumminello et al. (2007a), corresponds to the hierarchical structure produced by average linkage clustering.

Directed Bubble Hierarchical Tree. While hierarchical clustering is commonly linked to filtered graphs like the MST, the Directed Bubble Hierarchical Tree (DBHT) Song et al. (2012) is directly informed by the topology of the Planar Maximally Filtered Graph (PMFG), integrating its structural information into the resulting hierarchy.

The method starts from the PMFG built from the dissimilarity matrix and exploits the structural role of the 3-cliques[6] in the graph to extract clusters.

Specifically, it makes a distinction between separating 3-cliques, cliques whose removal disconnects the graph, and non-separating 3-cliques. The separating cliques identify boundaries between distinct subgraphs, called bubbles. The bubbles are organized into a bubble tree, a directed structure where edges are oriented based on the weight distribution of links across the separating clique. Converging bubbles, nodes with only incoming edges, are taken as cluster centers and all other nodes are assigned to these centers based on topological proximity and strength of connection, without needing to pre-specify the number of clusters.

This assignment is then used to build a three-tier hierarchy:

- **Intra-bubble hierarchy** representing clustering within individual bubbles;
- **Intra-cluster hierarchy** among bubbles assigned to the same cluster;
- **Inter-cluster hierarchy** representing the hierarchical organization among

[6] A 3-clique of is a fully connected triplet of vertices, that is, a set $\{V_i, V_j, V_l\} \subset V$ such that all three edges (V_i, V_j), (V_j, V_l), and (V_i, V_l) belong to E.

the resulting clusters;

This method has been applied for the first time to stock return correlation networks in Musmeci et al. (2015) which shown that DBHT clustering provides better results overall with respect to other methodologies.

Partition Decoupling Method. Another notable approach is the Partition Decoupling Method (PDM) introduced by Leibon et al. (2008), which provides a multi-scale framework for detecting community structures. PDM uncovers multiple interacting partitions across hierarchical layers and is capable of identifying topological features such as cycles within the community graph.

The method combines two algorithms: the partition scrubbing method and hierarchical spectral clustering. The spectral clustering algorithm is specifically adapted for correlation matrices. Let $\boldsymbol{\rho}$ denote a symmetric correlation matrix computed from a set of time series. The first step involves transforming the correlations into dissimilarities using a different transformation compared to the previously discussed methods. In particular, the dissimilarity matrix $d = (d_{ij})$ is defined as:

$$d_{ij} = \sin\left(\frac{1}{2}\arccos(\rho_{ij})\right),$$

representing half the chordal dissimilarity between two points on the unit sphere, which respects the geometry of correlations.
The dissimilarities are then used to build a complete graph from the affinity matrix:

,

and then the graph Laplacian L is computed as:

$$L = I - D^{-\frac{1}{2}} \times W \times D^{-\frac{1}{2}},$$

where D be the diagonal degree matrix with entries $D_{ii} = \sum_j W_{ij}$. Then the next step is to compute the eigenspectrum $\{\lambda_i, \mathbf{v}_i\}$ of L, and select the k most relevant eigenvectors.

The number of clusters k is determined by comparing the spectrum of the Laplacian to that of a suitable null model, such as the Laplacian derived from random correlation matrices. The rows of the matrix corresponding to the selected eigenvalues are then normalized to unit length, and k-means clustering is applied to these normalized rows to partition the nodes into k communities.

After performing spectral clustering, the partition scrubbing method is applied iteratively in combination with spectral clustering to construct a hierarchy of partitions. Specifically, for each cluster obtained from spectral clustering, a prototype is generated by computing the mean time series:

$$\bar{X}^{(\alpha)}(t) = \frac{1}{|c_\alpha|} \sum_{i \in c_\alpha} X_i(t),$$

so that each prototype represents the average dynamics of cluster c_α.

Then each original time series $X_i(t)$ is orthogonally projected onto the subspace spanned by the set of cluster prototypes $\{\bar{X}^{(\alpha)}(t)\}$. This is basically a linear regression step:

$$X_i(t) = \sum_{\alpha=1}^{C} \beta_{i\alpha} \bar{X}^{(\alpha)}(t) + \epsilon_i^{(k)}(t),$$

where $\epsilon_i^{(k)}(t)$ is the exogenous component of $X_i(t)$, orthogonal to the span of the cluster prototypes and is used to build another set of time series:

$$X_i^{(k+1)}(t) = \epsilon_i^{(k)}(t),$$

which are then used to construct the next-layer correlation matrix:

$$\rho_{ij}^{(k+1)} = \text{Corr}(X_i^{(k+1)}, X_j^{(k+1)}),$$

capturing the correlations not explained by the first layer partitions. The two steps are then iteratively repeated to uncover deeper layers of modular organization:

$$C^{(0)} \to C^{(1)} \to \cdots \to C^{(K)}.$$

The authors apply the method to a set of 4 years time series for 2547 stocks and find that the US equity market exhibits rich, multiscale community structure and nontrivial topology, including loops that reflect sector rotation.

Random Matrix Theory and Modularity Maximization. The methods discussed so far are applied to the dissimilarity matrix derived from correlations and are largely associated with traditional clustering techniques for community detection. In contrast, MacMahon and Garlaschelli (2015) adopt a different approach by operating directly on the correlation matrix and adapting community detection methods based on modularity to account for the specific properties of empirical correlation matrices. Their method is closely linked to random matrix theory, which provides a key set of tools for filtering noise from correlation matrices. In particular, the authors redefine the modularity function using null models that are consistent with the spectral characteristics of empirical correlation matrices. Formally, a correlation matrix $\boldsymbol{\rho}$ is decomposed as:

$$\boldsymbol{\rho} = \boldsymbol{\rho}^{(r)} + \boldsymbol{\rho}^{(g)} + \boldsymbol{\rho}^{(m)},$$

- $\boldsymbol{\rho}^{(r)}$ represents a noise component corresponding to the eigenvalues within the

Marčenko–Pastur bounds;

- $\boldsymbol{\rho}^{(m)}$ represents the global mode corresponding to the largest eigenvalue also known as the market mode;
- $\boldsymbol{\rho}^{(g)}$ represents the signal which contains information about the group structure

By using this decomposition of the correlation matrix the authors derive three re-definitions of modularity for different levels of complexity of the underlying time series:

1. **Infinite time series, no global mode:**

$$Q_1(C) = \frac{1}{\boldsymbol{\rho}_{\text{norm}}} \sum_{i,j} (\rho_{ij} - \delta_{ij}) \delta(c_i, c_j);$$

2. **Finite time series, noise-filtered:**

$$Q_2(C) = \frac{1}{\boldsymbol{\rho}_{\text{norm}}} \sum_{i,j} (\rho_{ij} - \rho_{ij}^{(r)}) \delta(c_i, c_j);$$

3. **Finite time series, noise and global mode filtered:**

$$Q_3(C) = \frac{1}{\boldsymbol{\rho}_{\text{norm}}} \sum_{i,j} (\rho_{ij} - \rho_{ij}^{(r)} - \rho_{ij}^{(m)}) \delta(c_i, c_j),$$

where $C = \{c_1, c_2, \ldots, c_N\}$ denotes a partition of the N assets into communities, with c_i indicating the community to which asset i belongs, $\delta(c_i, c_j) = 1$ if $c_i = c_j$ and 0 otherwise, and $\boldsymbol{\rho}_{\text{norm}} = \sum_{i,j} \rho_{ij}$ is a normalization factor ensuring that the modularity is properly scaled.

The authors then adapt three widely used algorithms to maximize the redefined modularity. They further extend the framework by introducing a multiresolution approach, which recursively applies community detection to submatrices of $\boldsymbol{\rho}$. Each

submatrix corresponds to a detected community and is re-filtered by removing community-specific modes. This methodology is consistent with the intrinsic properties of correlation matrices and allows for a principled analysis of hierarchically nested community structures in high-dimensional systems.

1.6 Applications

With regard to practical applications, two main strands of literature link financial networks to portfolio construction. Although both build on the methodologies described above, they differ in whether portfolios are constructed using (i) network centrality metrics or (ii) the hierarchical community structure. Given the extensive literature, we focus on the most influential approaches and indicate when contributions are primarily exploratory and when they include extensive backtesting against standard portfolio optimization benchmarks.

Network centrality-based methods. The intersection of network topology and portfolio design has been explored through several approaches, all converging on a central insight: an asset's position within the correlation network carries information about diversification potential and systemic exposure.

The first explicit link between filtered networks and portfolio construction is found in Onnela et al. (2003b), where 477 NYSE stocks (1980–1999) are analyzed using MST- and hierarchical-based representations. Besides introducing concepts such as the central vertex, market "backbone," and mean occupation layer (MOL), the authors show that assets appearing at the periphery of the MST systematically dominate optimal Markowitz portfolios. The work demonstrates how tree-filtered networks provide interpretable representations that preserve essential economic structure while reducing noise in the correlation matrix.

A more systematic exploitation of network position is proposed by Pozzi et al. (2013),

who construct hybrid centrality indicators combining several metrics into composite indices X and Y. High $X + Y$ values identify peripheral assets, while $X - Y$ captures differences in link quality. Portfolios consisting of peripheral stocks consistently achieve superior signal-to-noise ratios, and even portfolios containing only about 20 peripheral assets match or outperform the full 300-stock benchmark. The approach provides a transparent rule-based selection procedure and shows how network layouts visually encode diversification opportunities.

In Peralta and Zareei (2016), the authors establish a theoretical link between Markowitz optimization and network centrality, proving that the optimal Markowitz weight is negatively related to node centrality. Building on this insight, they introduce the ρ-dependent portfolio rule, which selects assets within a specified centrality range. Extensive in-sample and out-of-sample tests across markets and time periods show that the resulting portfolios achieve lower realized variance and higher Sharpe ratios than standard benchmarks.

Overall, despite methodological differences, these works lead to the same conclusion: assets located on the network periphery tend to provide better risk-adjusted performance, suggesting that investors should "invest in the peripheries(Pozzi et al. (2013))".

Clustering-based methods. A second line of research focuses on the hierarchical community structure derived from the dissimilarity matrix and its use in portfolio construction. A seminal reference is Tola et al. (2008), which studies how hierarchical clustering can reduce estimation noise in correlation matrices. Assuming fixed expected returns and volatilities, the authors treat uncertainty as arising only from correlation estimation. They transform correlations into dissimilarities and apply hierarchical clustering to obtain an ultrametric structure; the resulting filtered corre-

lations $\rho_{ij}^{<}$ are positive-definite and retain essential structure while attenuating noise. Using data on over 1,000 NYSE stocks (1988–1998), they show that average-linkage filtered matrices produce more reliable risk forecasts, reducing the error between ex-ante and realized portfolio variance. The study provides early evidence that clustering-based filtering is effective for improving Markowitz optimization.

A different path is taken by Lopez de Prado (2016), which abandon mean-variance optimization entirely and introduce the Hierarchical Risk Parity (HRP) methodology. HRP begins by computing dissimilarities and then a dissimilarity-of-dissimilarities matrix, followed by hierarchical clustering to reveal the nested asset structure. After reordering the covariance matrix into quasi-diagonal form, a recursive bisection allocates capital inversely to cluster-level risk. This prevents concentration, ensures diversification across and within clusters, and avoids matrix inversion entirely. Through extensive Monte Carlo simulations, the authors show that HRP reduces out-of-sample variance by more than 70% compared to the Critical Line Algorithm (CLA) and by 38% relative to inverse-variance portfolios.

Building on HRP, Raffinot (2018) proposes Hierarchical Clustering-Based Asset Allocation (HCAA), a general framework that extends the dissimilarity-of-dissimilarities approach of Lopez de Prado (2016). Multiple clustering algorithms are tested, and a statistically motivated stopping rule based on the gap statistic Tibshirani et al. (2001) is introduced to determine the optimal depth of the dendrogram, in contrast to HRP's full recursive bisection. Once cluster-level weights are assigned, capital is distributed equally within each cluster. Extensive backtesting shows that HCAA provides more robust out-of-sample performance than traditional risk-based strategies. While no single clustering method dominates, DBHT tends to yield slightly better risk-adjusted returns, whereas average linkage offers greater robustness.

The same author further refines the hierarchical framework with the Hierarchical Equal Risk Contribution (HERC) portfolio, which integrates the stopping rule of

HCAA with the recursive ERC-based allocation scheme of HRP. Large-scale empirical tests show that HERC improves risk-adjusted performance and further stabilizes portfolio weights relative to its predecessors.

1.7 Other Financial Networks

Beyond networks constructed from stock return data, a wide variety of financial systems have been analyzed using network science, providing insights into the structure and dynamics of interconnections of those systems.

A first major strand concerns *interbank networks*, where nodes are banks and edges capture bilateral exposures. Typically directed and weighted, these networks are used to study shock propagation and systemic fragility. Theoretical work relates topology to contagion amplification (e.g., Acemoglu et al. (2015)), while metrics such as DebtRank Battiston et al. (2012) assess institutional importance. Empirical studies reveal robust structural patterns and early-warning signals Squartini et al. (2013), and multilayer models account for heterogeneity across instruments Aldasoro and Alves (2018).

A second major category is that of portfolio-overlap networks, where bipartite institution–asset structures capture indirect contagion through common holdings. Their projections link portfolio similarity to potential fire-sale spillovers, and contagion models quantify amplification channels (e.g., Greenwood et al. (2015); Cont and Schaanning (2017)). Empirical evidence shows the increasing relevance of overlaps before stress episodes Gualdi et al. (2016), and multilayer frameworks combining direct and indirect exposures reveal that direct links alone underestimate systemic risk Poledna et al. (2021).

Payment networks constitute another important class, reflecting settlement systems

such as Fedwire or TARGET2. These networks exhibit scale-free and small-world characteristics Soramäki et al. (2007), efficient liquidity distribution Becher et al. (2008), and community structures tied to institutional roles Galbiati and Stanciu-Vizeteu (2015). Simulation studies examine alternative settlement mechanisms (Leinonen and Soramäki (1999)), and measures such as SinkRank Soramäki and Cook (2013) evaluate systemic relevance.

Finally, network approaches extend to areas including counterparty exposures under central clearing Duffie and Zhu (2011), overnight lending relationships Iori et al. (2008), and global corporate ownership Vitali et al. (2011), underscoring their broad applicability in modern financial systems.

1.8 Concluding remarks

This chapter has provided a broad overview of the main theoretical and methodological foundations underlying the study of networks in financial markets, with a particular focus on stock correlation networks. By presenting the essential concepts, tools, and established approaches used in the literature, the chapter lays the foundation for understanding how network-based methods can be applied to financial time series and why they have become a central component in the analysis of market structure and dynamics. The chapter highlights how networks offer a flexible and intuitive framework for representing relationships between financial assets and identifying patterns not immediately visible through traditional analyses. It also shows that the study of correlation networks intersects with a wide range of applications, demonstrating the breadth of contexts in which these methods have been employed. This chapter does not introduce methodological advances per se. Rather, its role is to provide the reader with the background and context necessary to interpret and fully appreciate the contributions developed in the subsequent chapters. The

subsequent chapters of the thesis build on this foundation to explore more specific questions and applications within the broader field of financial networks.

Chapter 2

Robust Multi-Source Financial Network Analysis

2.1 Introduction

The previous chapter provided a general overview of methodological advances in applying network science concepts to financial time series, defining a wide range of financial networks. Among these, stock correlation networks Mantegna and Stanley (1999); Mantegna (1999) have emerged as a prominent research avenue with numerous studies covering both theoretical innovations and practical applications in asset allocation and risk diversification frameworks.

However, despite the substantial and continuously expanding literature on stock correlation networks, several fundamental aspects remain insufficiently explored. This chapter addresses these gaps through two complementary methodological advances, each supported by extensive empirical evidence through the most comprehensive empirical study of its kind in the literature.

Our first contribution addresses a systematic yet largely neglected challenge in stock correlation network analysis. Extreme observations in financial returns intro-

duce substantial distortions in community detection algorithms, with these effects becoming particularly pronounced at higher sampling frequencies Bhachech et al. (2022). By implementing robust statistical methods for correlation computation, we demonstrate performance improvements exceeding 50% in sector recovery at intra-day frequencies compared to conventional approaches. These findings fundamentally challenge the widespread practice of applying sample correlation estimators to inherently noisy financial data.

Our second contribution investigates the joint consideration of multiple information sources in community detection, a dimension of analysis almost absent from the stock correlation networks literature despite its established importance in related domains. We examine this along two complementary perspectives. The first concerns cross-variable integration, which incorporates returns, volume, and volatility dynamics rather than relying exclusively on returns. The second explores cross-frequency integration, synthesizing information across multiple temporal resolutions. Both approaches yield substantial improvements over traditional single-source methodologies: cross-variable integration achieves nearly twofold enhancement in sector overlap relative to returns-only analysis, while cross-frequency integration attains unprecedented alignment with economic fundamentals, substantially surpassing all existing benchmarks.

While each contribution represents a substantial advancement on its own, their joint implementation through robust cross-frequency integration could play an important role in shaping subsequent developments in the field. The resulting community structures achieve the highest sector alignment ever documented in the literature while simultaneously exhibiting superior economic efficiency. Portfolios constructed from these data-driven communities generate positive risk-adjusted returns over the market average, with markedly reduced volatility compared to sector-based groupings. When these empirically derived communities diverge from traditional sector

classifications, they capture economically meaningful regime shifts that remain invisible to conventional methods.

We validate these findings using a novel analytical framework that enables a direct economic interpretation of stock groupings over time. The validation relies on comprehensive empirical analysis (Section 2.6) leveraging twenty years of financial data observed across multiple sampling frequencies, representing the most extensive multi-frequency investigation in the stock correlation networks literature with respect to both data variety and temporal resolution. Collectively, these results demonstrate that financial markets possess an intrinsic capacity for efficient self-organization, a property that becomes empirically observable only when multiple information dimensions are jointly considered within robust estimation frameworks.

The remainder of this chapter is organized as follows. Section 2.2 provides essential context by reviewing the historical development of correlation-based network methods and their evolution toward the current standard framework. Sections 2.3, 2.4, and 2.5 establish the theoretical foundations underlying our methodological contributions, while Section 2.6 presents the comprehensive empirical analysis supporting our findings and, lastly, Section 2.7 concludes.

2.2 Standard analysis framework

The current standard framework originated from Mantegna (1999) which introduced the idea of building a sparse graph from stock return time series cross-correlations. The method consisted of computing the Pearson cross-correlation between the stock returns time series and then transform the correlations to dissimilarities using the transformation 1.1. The dissimilarity matrix was then used to extract a sparse graph by computing the Minimum Spanning Tree (MST) and to obtain a hierarchical tree and the associated community structure by using the single linkage clustering. The

original work analyzed both the MST and the hierarchical tree extracted from the log-returns time series of the stocks composing the DJIA index and the S&P 500 index and found a correspondence between clusters and clear economic groupings such as industry sectors, suggesting the exposure to common industry specific risk factors.

This approach has been extended and improved in many ways over the last 25 years with both methodological contributions and practical applications that use either the network topology or the community structure as the backbone of asset allocation and diversification strategies. A key refinement to the standard procedure originates from the empirical findings of Borghesi et al. (2007), who conducted an extensive analysis across different sampling frequencies and methodological settings. Their results emphasize the importance of accounting for the common market component, the systematic factor driving broad co-movements in asset returns, before constructing the network and identifying communities. The rationale is that this pervasive component dominates the correlation structure, inducing high average correlations that mask the genuine cross-sectional dependencies among individual stocks. Removing or controlling for this market-wide effect allows the residual correlations to capture more accurately the idiosyncratic relations among assets.

The idiosyncratic component of stock returns is typically extracted using a standard single-factor model with respect to the market return Sharpe (1964), which can be represented either by a broad market index or by the cross-sectional average of all stock returns. Although the financial econometrics literature has progressed substantially since the foundational contributions to asset pricing, most notably with the introduction of the Fama–French three-factor model Fama and French (1993) and its subsequent extension to the five-factor specification Fama and French (2015), our analysis is primarily empirical and relies on real financial data observed at multiple frequencies. Unfortunately, the additional firm-level variables required to estimate

such multi factor models are not available. For this reason we adopt the traditional single-factor specification which remains a parsimonious and widely accepted approach for isolating the idiosyncratic component of returns. Let $r_i(t)$ be the return of asset i at time t and $r_M(t)$ the market return at time t, typically represented by a market index, the single factor model is defined as:

$$r_i(t) = \alpha_i + \beta_i \, r_M(t) + \epsilon_i(t) \tag{2.1}$$

where α_i is the intercept, β_i is the slope coefficient capturing the sensitivity of asset i's return to the market return and $\epsilon_i(t) = r_i(t) - E[r_i(t)|r_M(t)]$ is a random variable representing the portion of asset i's return not explained by the market factor, under the assumption of onthogonality of ϵ with respect to r_M ($E[\epsilon|r_M] = 0$). Accordingly, the correlations are computed on the idiosyncratic component of returns, $\epsilon_i(t)$, as opposed to the raw returns $r_i(t)$. Both the extraction of the idiosyncratic return component through the single-factor model and the subsequent computation of correlations constitute the two fundamental steps of the analysis outlined in figure 2.1. In principle, one could assess the statistical adequacy of both the factor model estimation and the resulting Pearson correlations. However, given the high dimensionality of the datasets used in these analyses, usually in the order of hundreds of stocks, it becomes practically infeasible to monitor the performance of each individual regression and correlation estimate. Consequently, the overall procedure operates in a de facto unsupervised manner. This setting underscores the necessity for estimation techniques that are intrinsically robust, as standard methods such as OLS and the classical Pearson correlation are known to be sensitive to deviations from model assumptions and the presence of outliers. To address these limitations, we introduce robust counterparts for both steps, designed to enhance the stability and reliability of the analysis. The following section introduces these methods and

illustrates how robustness can fundamentally reshape the interpretation of dependence structures in financial markets. A vast amount of literature has also been addressed to the exploration of alternative clustering methods to apply to the dissimilarity matrix. In this regard Coronnello et al. (2007) applies the Average Linkage Hierarchical clustering algorithm and finds that it provides more structured information with respect to the Single Linkage. The most complete work in this regard is represented by Musmeci et al. (2015) which analyzes the results from five different clustering algorithms, namely: i) Single Linkage Hierarchical Clustering Anderberg (2014), ii) Average Linkage Hierarchical Clustering, iii) Complete Linkage Hierarchical Clustering, iv) K-medoids Kaufmann and Rousseeuw (1987), v) Directed Bubble Hierarchical Tree (DBHT) Song et al. (2012). This study is particularly relevant because, beyond comparing the performance of a broad set of clustering algorithms, it also examines how each algorithm performs when applied to correlation matrices computed from raw returns and from returns purged of the systematic component. The findings show that, when correlations are estimated from idiosyncratic returns, the resulting clusters display a substantially stronger alignment with the underlying economic sector classification, consistently across all algorithms considered. The analytical framework developed in Musmeci et al. (2015) serves as the reference standard for this type of analysis; its overall structure is illustrated in Figure 2.1.

2.3 The case for robustness

As already anticipated in the previous section, a key limitation of the standard approach lies in its sensitivity to extreme observations in the data. Estimating the single-factor model for each asset via OLS can lead to distorted coefficient estimates when outliers are present, producing outcomes that misrepresent the true idiosyncratic component. Since these idiosyncratic returns form the basis for the

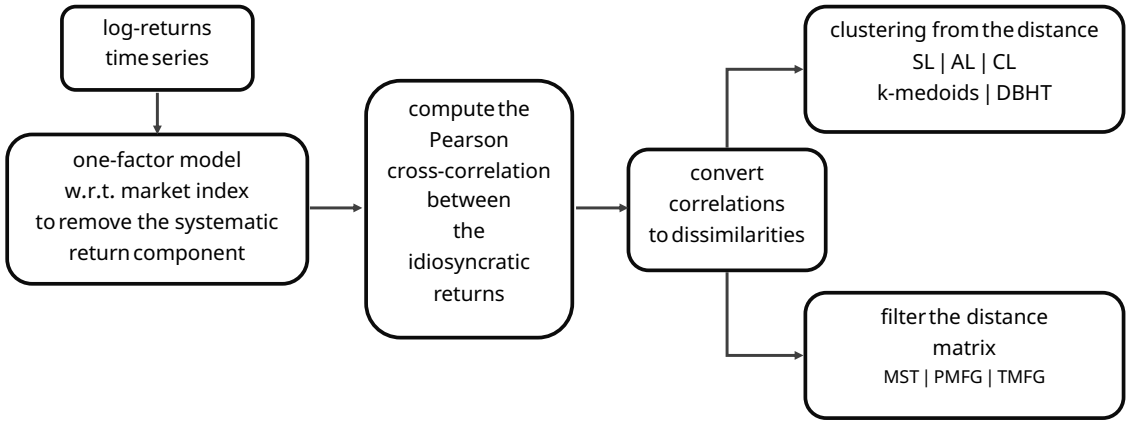


Figure 2.1: Schematic representation of the standard procedure

subsequent correlation calculations, even a few extreme values can propagate distortions through the entire analysis. This issue is particularly relevant in financial time series, where large deviations occur more frequently than in most other contexts. To address these challenges and improve the stability and reliability of the results, we adopt robust regression techniques Maronna et al. (2006); Ronchetti and Huber (2009), which limit the influence of outliers.

Several robust regression methods have been proposed in the literature, with the most widely used being M-estimators Huber (1964), MM-estimators Yohai (1987), and Least Trimmed Squares (LTS) Rousseeuw (1984). All of these approaches share the common principle of reducing the influence of extreme observations, thereby achieving a compromise between bias and efficiency. For the present work we employ the LTS estimator due to its simplicity, interpretability and favorable performance in small samples, as well as the straightforward interpretation of its single hyperparameter.

The basic idea of LTS is to reduce the influence of extreme values by minimizing

not the sum of all squared residuals, like the OLS, but the sum of the h smallest squared residuals. Sorting the squared residuals in ascending order:

$$\epsilon^2_{(1)}(\alpha,\beta) \leq \epsilon^2_{(2)}(\alpha,\beta) \leq \cdots \leq \epsilon^2_{(T)}(\alpha,\beta),$$

the LTS estimator solves:

$$\min_{\alpha,\beta} \sum_{t=1}^{h} \epsilon^2_{(t)}(\alpha,\beta).$$

where the parameter h determines the proportion of observations contributing to the fit. In our empirical analyses, we set $h = 0.8\,T$, meaning that only the 80% of observations with the smallest residuals are used to estimate the regression coefficients. This choice corresponds to a breakdown point of 20%, indicating that up to one fifth of the data can be arbitrarily contaminated without compromising the validity of the estimates. By trimming the largest residuals, LTS concentrates the regression fit on the central bulk of the data, producing coefficient estimates that more accurately reflect the underlying systematic relation and residuals that better capture the idiosyncratic component of returns. From a computational perspective, solving the LTS problem exactly is NP-hard, as it requires evaluating all possible subsets of size h. In practice, efficient approximation algorithms are employed to obtain the estimates. As an example of the impact of robust regression methods we show, in figure 2.12, the scatter plot of the monthly estimated beta coefficients over time for two stocks in our sample, representing Colgate and Procter & Gamble. The coordinates of the points in the scatter represent the $\hat{\beta}$ of the one factor models estimated through OLS on the x axis and LTS on the y axis. Some of the points deviate significantly from the diagonal demonstrating a huge difference between the two beta estimates, with some points showing remarkably different results. The difference in the parameter estimates directly affects the resulting idiosyncratic return component, as shown in figure 2.3 which shows the time series of the differences be-

tween the idiosyncratic return components from the OLS models and the residuals from the LTS models, with the green dots highlighting the biggest ones in magnitude (1%). These differences are directly reflected in the correlations between the time series, as shown by figure 2.4 which shows the Pearson cross-correlation between the idiosyncratic return component extracted from, respectively, the OLS and LTS one factor models of the two stocks analyzed above on the left and on two other stocks in the Real Estate sector on the right. We decided to show the correlations between two stocks belonging to Consumer Staples, a traditionally defensive sector weakly correlated to the market conditions, and two Real Estate stocks, a sector which has been the epicenter of the biggest financial crisis of the 21st century, to show how different the two correlations dynamics are. In particular the two Consumer Staples stocks show a steadily increasing correlation over time while the two Real Estate ones have some steep variations right after the 2008 subprime crisis and the COVID-19 and, interestingly, a significant and persistent difference between correlation computed on OLS and LTS residuals after the COVID-19 pandemic.

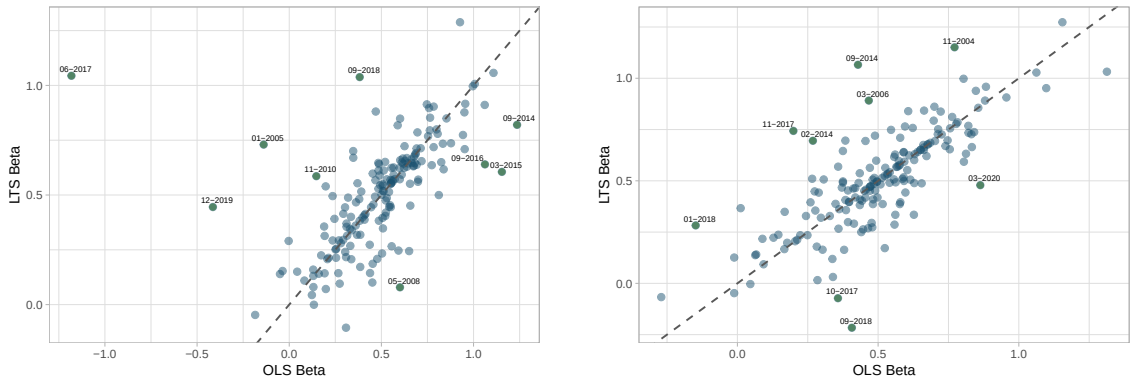


Figure 2.2: Comparison of monthly beta estimates (30-day windows) obtained with OLS and LTS methods for Colgate (left) and Procter & Gamble (right).

The differences in residuals resulting from robust regression, as illustrated in the previous figures, naturally propagate to the computation of cross-correlations between assets. Even moderate deviations in the estimated idiosyncratic components

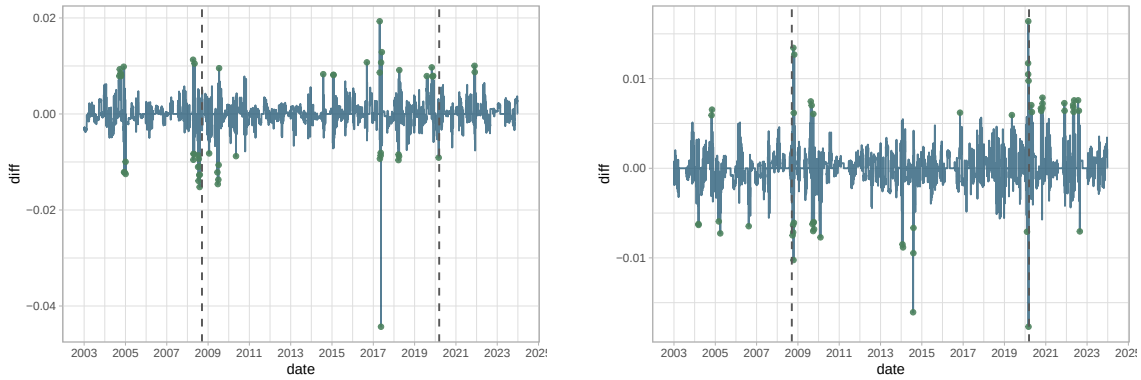


Figure 2.3: Time series of deviations between idiosyncratic return components extracted from one-factor models estimated via OLS and LTS methods for Colgate (left) and Procter & Gamble (right).

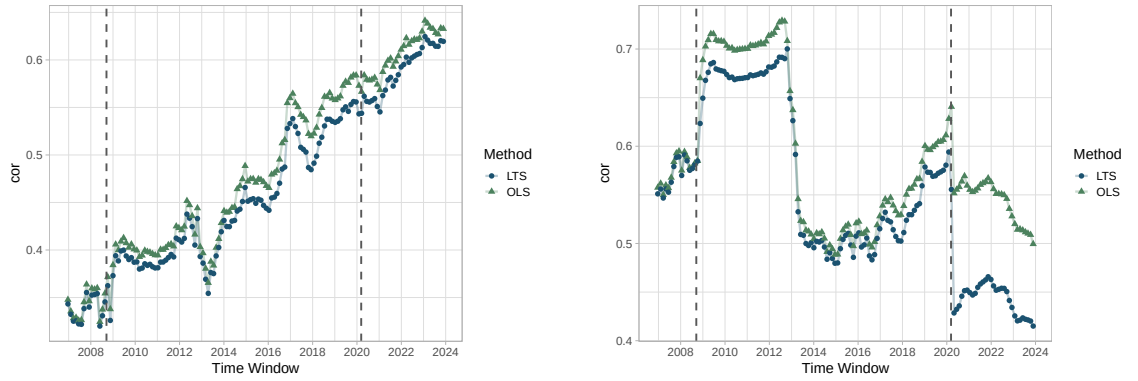


Figure 2.4: Comparison of rolling window Pearson correlations computed from residuals estimated using OLS and LTS methods for Colgate and Procter & Gamble (left) and for Equity Residential and Realty Income Corporation (right). The correlations are computed on 1000 days rolling windows with a 30 days shift between windows.

can lead to substantial changes in the measured pairwise dependence, particularly in high-dimensional financial datasets. While the sample correlation is asymptotically consistent under relatively mild regularity conditions, it lacks robustness in finite samples. Specifically, the estimator's efficiency deteriorates substantially and it exhibits high sensitivity to outlying observations when the underlying distribution is heavy-tailed. This is particularly problematic in the context of financial time series, where leptokurtosis, asymmetry, and volatility clustering constitute well-documented stylized facts, especially at higher sampling frequencies Cont (2001). Consequently, the finite-sample properties of the sample estimator may be severely compromised in practical applications, potentially producing misleading conclusions about the underlying dependence structure. These considerations highlight the need for a robust counterpart to the sample correlation that is able to capture the central dependence patterns while mitigating the effects of outliers.

A wide range of robust covariance estimators have been proposed, often based on M-estimation principles or adaptive weighting schemes that down-weight the influence of extreme observations Maronna et al. (2006). In this work, we employ the Tyler's scatter estimator Tyler (1987) for its distribution-free nature and its scale invariance properties. The key idea of the Tyler estimator is to work with a matrix that is proportional, but not identical, to the covariance matrix: the scatter matrix, obtained by normalizing the observations as follows:

$$\boldsymbol{s}_i = \frac{\boldsymbol{x}_t}{||\boldsymbol{x}_t||_2}$$

and then applying MLE on the normalized points s. While normalization removes scale information from the observations, it preserves their relative differences and directional properties, thereby providing numerical stability. The probability density

function of the normalized points is thus represented by their angular distribution:

$$f(s) \propto \frac{1}{\sqrt{|\mathbf{\Sigma}|}} \left(\boldsymbol{s}^T \mathbf{\Sigma}^{-1} \boldsymbol{s}\right)^{-N/2}$$

with $\mathbf{\Sigma}$ being the parameter to estimate.

Observing that $\left(\boldsymbol{s}_t^T \mathbf{\Sigma}^{-1} \boldsymbol{s}_t\right)^{-N/2} \propto \left(\boldsymbol{x}_t^T \mathbf{\Sigma}^{-1} \boldsymbol{x}_t\right)^{-N/2}$ we can write the MLE for T observations as:

$$\min_{\Sigma} \log\det(\mathbf{\Sigma}) + \frac{N}{T} \sum_{t=1}^{T} \log\left(\boldsymbol{x}_t^T \mathbf{\Sigma}^{-1} \boldsymbol{x}_t\right)$$

if we express the gradient with respect to $\mathbf{\Sigma}^{-1}$ we get:

$$\Sigma = \frac{1}{T} \sum_{t=1}^{T} w_t(\mathbf{\Sigma}) \times \boldsymbol{x}_t \boldsymbol{x}_t^T$$

and the weights w_t given by:

$$w_t(\Sigma) = \frac{N}{\boldsymbol{x}_t^T \mathbf{\Sigma}^{-1} \boldsymbol{x}_t}$$

this fixed-point equation can be solved through the following procedure, for $k = 0, 1, 2, ...$, at $k = 0$ set $\mathbf{\Sigma}^{(0)}$ equal to the sample covariance matrix and then iteratively update it:

$$\tilde{\mathbf{\Sigma}}^{(k+1)} = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{N}{\mathbf{x}_t^\top (\mathbf{\Sigma}^{(k)})^{-1} \mathbf{x}_t} \right) \mathbf{x}_t \mathbf{x}_t^\top$$

After each iteration normalize the scatter by the trace:

$$\mathbf{\Sigma}^{(k+1)} = \frac{\tilde{\mathbf{\Sigma}}^{(k+1)}}{\mathrm{tr}\left(\tilde{\mathbf{\Sigma}}^{(k+1)}\right)}$$

The procedure then stops when reaching convergence:

$$\|\mathbf{\Sigma}^{(k+1)} - \mathbf{\Sigma}^{(k)}\|_F < \epsilon$$

However, when applied to financial settings with a large number of assets N relative to the available time span T, the standard unregularized estimator may become unstable or not defined, as it requires $T >> N$. This high-dimensional settings are pretty common in empirical financial studies, where the trade-off between temporal resolution and dimensionality is unavoidable. Longer estimation windows increase the reliability of covariance estimates but risk mixing heterogeneous market regimes, whereas shorter windows preserve local stationarity at the cost of increased noise in the estimates. In such contexts, a regularized version of the Tyler's scatter estimator Sun et al. (2014) provides a principled compromise. Regularization stabilizes the estimation procedure by augmenting the log-likelihood with a penalization term, typically promoting shrinkage toward a structured target such as the identity matrix.

The regularized Tyler estimator addresses the finite-sample instability that arises when the sample-to-dimension ratio N/T is large, even when $N < T$. Indeed, the degeneracy and noise amplification in multivariate estimators manifest well before the singular case where $N > T$. The regularization penalty shrinks the estimate toward a target matrix, thereby numerically stabilizing the estimator and reducing the contribution of weak spurious correlations arising from sampling noise. Therefore, in this work, we adopt the regularized Tyler scatter estimator to obtain robust and stable correlation structures that remain reliable under both heavy-tailed distributions and challenging dimensionality regimes. As already anticipated, the regularized Tyler's scatter is obtained by adding a penalization term to the likelihood:

$$\min_{\Sigma} \frac{T}{2} \log\det(\boldsymbol{\Sigma}) + \frac{N}{2} \sum_{t=1}^{T} \log\left(\boldsymbol{x}_t^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t\right) + \alpha \left(N \log\left(Tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\right)\right) + \log\det(\boldsymbol{\Sigma})\right),$$

where $\mathbf{\Omega}$ is the target matrix and the main step of the iterative procedure becomes:

$$\mathbf{\Sigma}^{(k+1)} = (1-\rho)\frac{N}{T}\sum_{t=1}^{T}\frac{\bar{x}_t\bar{x}_t^T}{\bar{x}_t\mathbf{\Sigma}_k^{-1}\bar{x}_t^T} + \rho\frac{N}{Tr\left(\mathbf{\Sigma}_k^{-1}\mathbf{\Omega}\right)}\mathbf{\Omega}$$

where $\rho = \frac{\alpha}{T/2+\alpha}$. The correlations are then extracted by normalizing the scatter matrix by its diagonal entries.

To illustrate the practical relevance of adopting robust methodologies in the correlation computations, figure 2.5 compares pairwise correlations estimated using the conventional Pearson estimator and the regularized Tyler's estimator. The left panel reports the correlations between Procter & Gamble (PG) and Colgate (CL), while the right panel shows those between Equity Residential (EQR) and Realty Income (O). All correlations are computed over 1,000-day rolling windows using OLS residuals.

For the two Consumer Staples firms, CL and PG, the differences between estimators are moderate and the temporal evolution of the correlations remains largely consistent across methods, in line with the patterns previously observed in Figure 2.3. In contrast, for the two Real Estate stocks, EQR and O, the discrepancies are substantial, particularly during and after major market events. Following the 2008 subprime crisis, the Pearson correlations rise markedly above the Tyler estimates, while the opposite occurs in the aftermath of the COVID-19 shock. These divergences highlight how the regularized Tyler estimator attenuates the disproportionate impact of extreme co-movements and provides a smoother and arguably more reliable depiction of the underlying dependence structure.

The benefits of robustness become even more apparent when both stages of the analysis, the removal of the market mode and the estimation of the correlation structure, are performed using robust methods. Figure 2.6 compares the correlations obtained

using the Pearson estimator on OLS residuals with those obtained using the Tyler estimator on residuals from the LTS regressions. The combination of LTS regression and the regularized Tyler estimator yields correlations that are simultaneously robust to extreme observations in the factor modeling stage and in the correlation estimation stage. Once again, the Consumer Staples pair shows stable correlations across specifications, whereas the Real Estate pair exhibits pronounced differences, consistent with the greater volatility and tail risk characteristic of that sector.

Overall, these results demonstrate that robust methods can substantially alter the empirical assessment of financial dependencies, especially in sectors prone to heavy-tailed behavior and regime shifts. By mitigating the influence of extreme events, the resulting correlation structures better reflect the typical co-movements of assets rather than transient shocks. In section 2.6 we will show how robustness impacts community detection performances, demonstrating its contribution to clearer and more economically coherent groupings. The same section then reports results based on the integration of heterogeneous information sources, a dimension often neglected in prior studies. Therefore the next paragraphs introduce this further dimension in detail, outlining the underlying rationale and methodological approach.

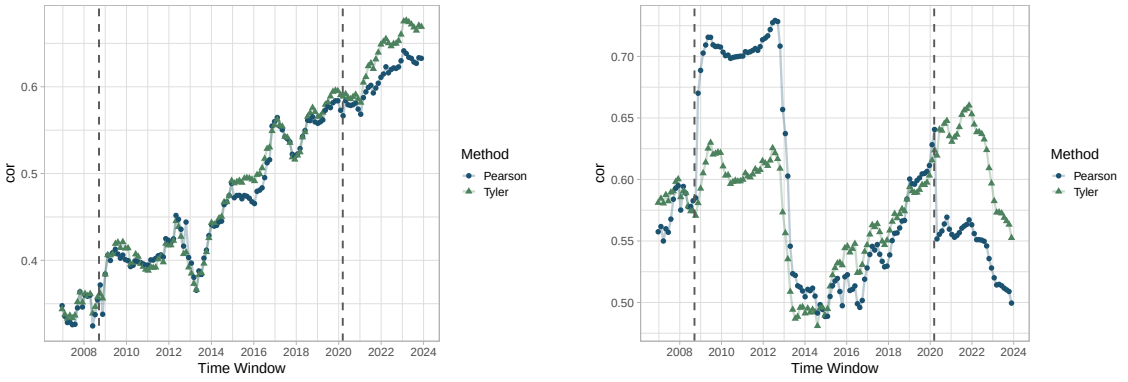


Figure 2.5: Comparison of rolling window correlations on OLS residuals computed from the Pearson and regularized Tyler estimators for Colgate and Procter & Gamble (left) and for Equity Residential and Realty Income Corporation (right). The correlations are computed on 1000 days rolling windows with a 30 days shift between windows.

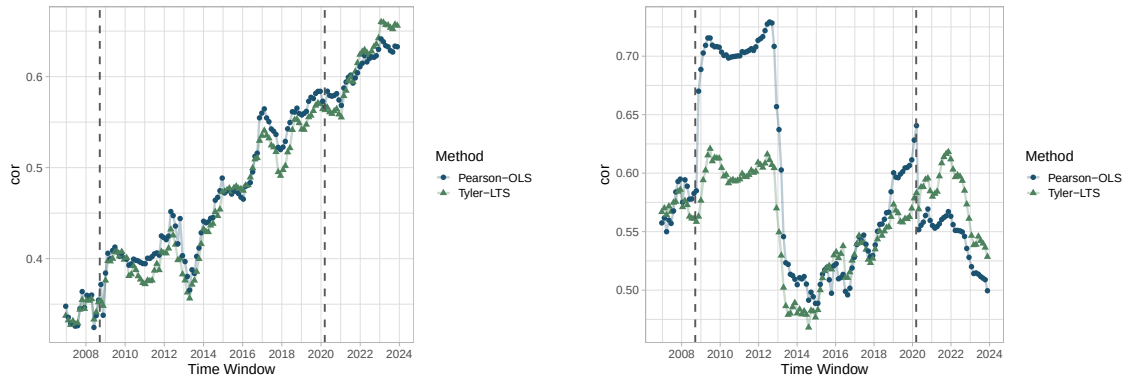


Figure 2.6: Comparison of rolling window correlations computed with the Pearson estimator on OLS residuals and the regularized Tyler estimator on LTS residuals for Colgate and Procter & Gamble (left) and for Equity Residential and Realty Income Corporation (right). The correlations are computed on 1000 days rolling windows with a 30 days shift between windows.

2.4 Information from different data sources

Aa already anticipated, another aspect that has received relatively little attention in the literature concerns the integration of the informative content carried by different data types. This idea can be conceived along two complementary directions. Horizontally, the focus is on the joint behavior of different measures such as returns, trading volumes, and volatilities observed at a common frequency, capturing their mutual relationships and co-movements. Vertically, attention shifts to the same variable examined across different temporal resolutions, revealing how its statistical properties change from high to low frequency representations. In both cases, most works in the literature are based on comparisons of reconstructed networks and community structures obtained from different data, addressing cross-variable or cross-frequency relationships in an indirect manner rather than explicitly integrating information, leaving the joint dynamics across variables and frequencies largely unexplored.

Within the cross-variable dimension, a number of studies briefly introduced below have explored how different market variables interact when observed at a common

frequency. These works often focus on the co-movement of returns, trading volumes, and volatilities, typically by comparing networks and community structures derived from each data type. While such analyses shed light on the patterns of interdependence across variables, they generally examine each data type separately, without providing a unified view that integrates multiple sources of market information, with a few notable exceptions that attempt to combine different measures to capture more comprehensive relationships.

Among them Micciché et al. (2003) built Minimum Spanning Trees from both returns and volatility and analyzed their structure and degree stability over time and found that return MSTs are more stable in time with respect to volatility MSTs, while Musmeci et al. (2017) used a method based on multiplex networks and PMFGs built from four different dependence measures between the same set of stocks and found that different dependency measures capture complementary aspects of the structure and dynamics of financial markets. Among the attempts to integrate different data types, some works have focused on embedding trading volume alongside returns in network constructions. Brida et al. (2016) combined log-returns and volumes in a single time series by taking the element-wise product between return and volume series, applying time series symbolization, computing a dissimilarity from the symbolized series, and then building the MSTs. Similarly, Coletti (2016) created a new time series as the element-wise product of closing prices and volumes, directly embedding the symbolized series while giving more weight to returns than to volumes.

With respect to the cross-frequency dimension, the literature has been primarily concerned with investigating and comparing correlations, reconstructed networks and communities between returns measured at different sampling frequencies. The most important work in this regard is Epps (1979) which documented what is now known as Epps effect: the fact that empirical correlations between stock returns

decrease as the sampling frequency of the data increases. Epps also found that this phenomenon appears to be caused by the nonstationarity of price changes and by correlations between transactions.

In the specific context at hand this difference in the correlations would mean that both the filtered networks and the detected communities would be different for each frequency, being based on fundamentally different dissimilarities. This has been the object of study of many works since the early days, Bonanno et al. (2001) analyzed the MST built from returns at different frequencies and found that the hierarchical organization strongly depends on the sampling frequency of the returns used to compute the correlations, while Tumminello et al. (2007b) analyzes the topological properties of the PMFGs between 300 stocks built from returns sampled at different time horizons and find that they exhibit markedly distinct topological structures, indicating that the correlation architecture of the market depends on the sampling frequency. In this sense, PMFGs derived from different frequencies encode complementary aspects of market organization. More recently Wang et al. (2017) analyzed MSTs and PMFGs constructed from US stock returns decomposed via wavelets at multiple time scales and found that network topologies, sectoral clustering, and link persistence vary significantly across scales, indicating that data at different frequencies provide complementary information about the market structure.

To the best of our knowledge, despite the growing literature on financial data at different frequencies, no study has yet attempted to jointly exploit the information provided across multiple frequencies. To address this gap, both in the cross-frequency and the cross-variable dimension, an existing methodology originally developed in genomics has been adapted. This approach allows for the joint exploitation of information across different views of the same set of observational units, and will be introduced in detail in the following subsection.

2.4.1 Data integration and Similarity Network Fusion

Data integration refers to the combination of heterogeneous data sources, representing different views of the same sampling units, into a unified representation that captures both shared and view-specific information.

Despite the historically large and heterogeneous amount of financial data available, finance has not traditionally been the field in which data-integration methods have been developed or most widely applied.

Indeed, many of the methodological advances in this domain stem from genomics and bioinformatics, where researchers routinely integrate transcriptomic, proteinomic, clinical and other types of data. Comprehensive reviews of data integration methods, such as Reel et al. (2021) and Subramanian et al. (2020), outline several major families of approaches. These include subspace alignment methods that project different datasets into a common latent space, multi-view representation learning approaches that jointly embed heterogeneous data, network-based integration frameworks that operate on pairwise similarity structures, and deep multimodal learning methods capable of capturing nonlinear relationships across modalities.

Among these approaches, Similarity Network Fusion Wang et al. (2014) stands out as particularly suitable method for the problem at hand. SNF constructs a similarity graph for each data modality and iteratively fuses them through a diffusion process that emphasizes both shared and complementary information, resulting in an integrated similarity network that preserves the consensus structure across data types. More in detail, the process starts by computing dissimilarities separately from q different data sources, leading to q different dissimilarity matrices D between the same observational units $D^{(1)}, D^{(2)}, ..., D^{(q)}$. The dissimilarity matrices are then turned to similarity matrices W through a scaled exponential kernel, for a single

element of a single dissimilarity matrix:

$$W(i,j) = \exp -\frac{d^2(i,j)}{\mu\delta_{i,j}} \tag{2.2}$$

where $d(i,j)$ is the i^{th}, j^{th} element of the dissimilarity matrix, μ is an hyperparameter of the method and $\delta_{i,j}$ controls the scaling and is defined as:

$$\delta_{i,j} = \frac{\text{avg}\left\{d\left(i,\mathcal{N}_i\right)\right\} + \text{avg}\left\{d\left(j,\mathcal{N}_j\right)\right\} + d(i,j)}{3}$$

where $d\left(i,\mathcal{N}_i\right)$ is the dissimilarity between element i and each one of his k nearest-neighbors, k is therefore another hyperparameter of the method. This leads to q similarity matrices $W^{(1)}, W^{(2)}, ..., W^{(q)}$ derived directly from the dissimilarity matrices, these similarity matrices are then used to build two different normalized similarity matrices for each single original similarity matrix:

$$P^{(q)}(i,j) = \begin{cases} \frac{W^{(q)}(i,j)}{2\sum_{k\neq i} W^{(q)}(i,k)}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases} \qquad H^{(q)}(i,j) = \begin{cases} \frac{W^{(q)}(i,j)}{\sum_{k\in N_i} W^{(q)}(i,k)}, & j \in N_i \\ 0, & \text{otherwise} \end{cases}$$

where the P matrices embed the full information about the original similarity matrices W, while the H matrices represent a sparse representation of the original W matrices retaining information only about the k nearest-neighboring points. The core step of the methodology consists of cross-diffusing the sparse matrices H on the full ones P through a procedure that iteratively updates the full similarity matrices P at each iteration t, in the case of $q = 2$:

$$P^{(1)}_{t+1} = H^{(1)} P^{(2)}_t \left(H^{(1)}\right)^T$$
$$P^{(2)}_{t+1} = H^{(2)} P^{(1)}_t \left(H^{(2)}\right)^T$$

where $P_{t+1}^{(1)}$ and $P_{t+1}^{(2)}$ are, respectively, the similarity matrices of the data type 1 and 2. The number of iterations τ is another hyperparameter of the methodology, at the last iteration τ the final fused similarity matrix is obtained as:

$$P_{\tau}^{*} = \frac{1}{2}\left(P_{\tau}^{(1)} + P_{\tau}^{(2)}\right)$$

The final fused similarity P^{*} thus embeds information from all the q different data sources and can be used for tasks like clustering and classification.

Although originally developed for genomic data, SNF offers a natural framework for financial applications. Furthermore, as already explained before, in financial time series this multiview framework can be further refined to capture distinct forms of heterogeneity, both across variables and across time scales. On one hand, we can consider horizontal views of a set of stocks, corresponding to distinct data types capturing complementary aspects of market dynamics. On the other hand, we can define vertical views, obtained by observing the same data type at different temporal resolutions. For example, returns computed from high-frequency intraday data provide a short-term view of market fluctuations, while daily or monthly aggregates reflect longer-term dependencies.

We adapted the SNF methodology to the context of financial time series by declining it into two complementary frameworks covering, respectively, the horizontal and the vertical multi-view perspectives. These two frameworks will be described in the following subsections.

2.4.2 Cross-variable SNF

To embed information horizontally across different data types, the first step is to transform the raw OHLCV time series into informative measures. While for price

data there exists a standard transformation, namely the logarithmic return

$$r_t = \log(P_t) - \log(P_{t-1}),$$

the choice of suitable transformations for other variables such as trading volume and volatility proxies is less straightforward. Regarding volume, we employ the differences between the logarithms of traded volumes,

$$v_t = \log(V_t) - \log(V_{t-1}),$$

which have been shown to share many statistical properties with returns, such as heavy tails and volatility clustering Podobnik et al. (2009). This makes them a natural choice for ensuring consistency across data types.

For volatility proxies, we need a measure computable at any sampling frequency and capable of capturing intra-period price variations. Among the various alternatives available in the literature, we use the Parkinson Parkinson (1980) range-based estimator, which takes into account both the highest and lowest prices within a given period:

$$w_t = \frac{1}{4\log(2)} \left(\log \left(\frac{H_t}{L_t} \right) \right)^2.$$

This estimator is non-parametric, efficiently exploits high–low information, and performs well across different time frequencies, justifying its choice among other volatility measures. Table 2.1 summarizes the three transformations applied to the data to obtain information on price variations, traded volumes, and volatility.

After computing these three measures, the next step consists in calculating the correlations over time series of T observations for each variable, thus obtaining three correlation matrices $\{\boldsymbol{\rho}^{(r)}, \boldsymbol{\rho}^{(v)}, \boldsymbol{\rho}^{(w)}\}$. These matrices are converted into dissimilarities through the transformation 1.1, resulting in $\{D^{(r)}, D^{(v)}, D^{(w)}\}$, which serve as

input dissimilarity matrices for the SNF algorithm. The SNF then returns a single fused similarity matrix that integrates information from returns, volumes, and volatilities at a given frequency. The procedure is summarized in figure 2.7.

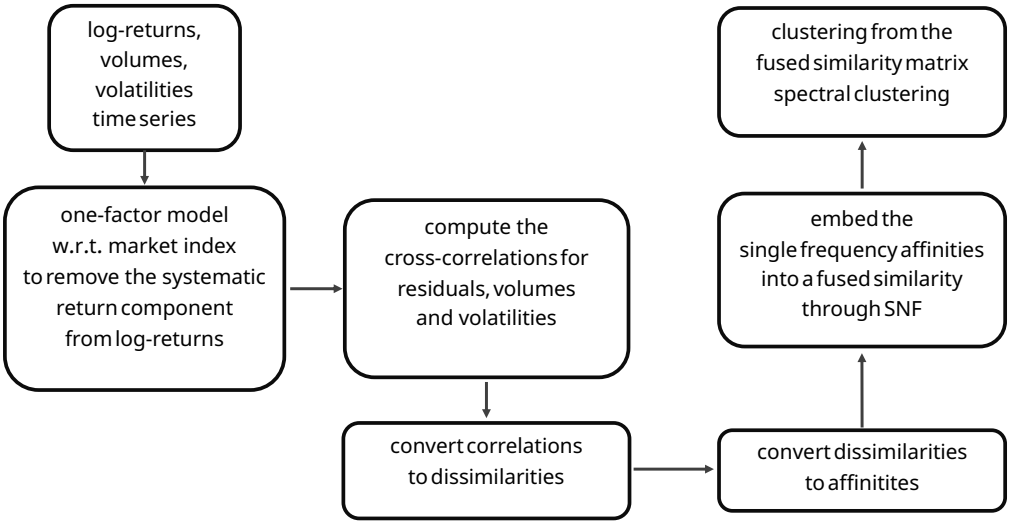


Figure 2.7: Schematic representation of the procedure used to embed different data types

2.4.3 Cross-frequency SNF

The framework for embedding information vertically across frequencies is illustrated in figure 2.8, though some details go beyond what the schematic representation alone can convey. The procedure starts from the daily time series, selecting a time window of T days. For each window, the idiosyncratic return component is extracted using the one-factor model (equation 2.1), from which the cross-correlations are computed to obtain the daily correlation matrix $\boldsymbol{\rho}^{(day)}$. The analysis then moves to a higher frequency while keeping the same time window, selecting the last T observations and repeating the same steps, idiosyncratic returns extraction and correlation computation, to obtain another correlation matrix $\boldsymbol{\rho}^{(f)}$ for the same set of stocks but at

Table 2.1: Summary of all the data used in the analysis

Measure	**Raw data**	**Pre-processing**	**Definition**
returns	adjusted closing prices (split/dividend adjusted) at 1-minute to daily frequency	missing values replaced with last observed value (LOCF)	log-returns: $r_t = \log(P_t) - \log(P_{t-1})$
volumes	trading volume at 1-minute to daily frequency	missing values replaced with last observed value (LOCF)	log-volume changes Podobnik et al. (2009): $v_t = \log(V_t) - \log(V_{t-1})$
volatilities	High (H) and Low (L) prices at 1-minute to daily frequency	missing values replaced with last observed value (LOCF)	Parkinson range-based estimator Parkinson (1980): $w_t = \frac{1}{4\log(2)} \left(\log\left(\frac{H_t}{L_t}\right)\right)^2$
detrended returns	log-returns of each stock i vs. SPY at 1-minute to daily frequency	—	one-factor model: $r_{i,t} = \alpha + \beta r_{M,t} + \epsilon_{i,t}$ idiosyncratic returns $\epsilon_{i,t}$ used in analysis
sectors	GICS-11 classification (11 sectors)	—	"Ground truth" labels used to evaluate community detection performance

a different frequency. This process continues up to the highest frequency, resulting in a set of correlation matrices $\{\boldsymbol{\rho}^{(day)}, \ldots, \boldsymbol{\rho}^{(1min)}\}$. These are converted into dissimilarities using equation 1.1, giving $\{D^{(day)}, \ldots, D^{(1min)}\}$, which are used as input dissimilarity matrices for the SNF algorithm. The SNF then outputs a single fused similarity matrix embedding information from returns at all considered frequencies.

Given that SNF here operates across multiple frequencies, we adopt a fixed observation window for correlation estimation at each frequency, thereby maintaining consistent sample-to-dimension ratios. Alternative specifications based on keeping a constant calendar time window were explored. The results for these alternative specifications are not shown here, but they generally produced inconsistent estimates, particularly at higher frequencies.

Furthermore, a clarification is needed on the role that correlation estimates play in our analysis framework. A central issue in the econometrics literature is the development of sophisticated estimators that account for microstructure noise, asynchronous trading, and other high-frequency effects, in order to obtain precise correlation estimates. Here, however, the goal is different since we are not interested in precisely estimating correlations, but in capturing the overall dependence structure among stocks. In this context, microstructure and other low-frequency noises are not detrimental and should be included, as they contribute to the observed associations and allow us to characterize the network of relationships among stocks, which is the ultimate focus of our analysis.

More recently, a promising line of research has emerged in the frequency domain Barigozzi and Farnè (2024); Krampe and Paparoditis (2025), where high-dimensional procedures are developed for estimating large spectral density matrices. These contributions provide an alternative route to network recovery based on frequency-domain conditional dependence, which is complementary to the perspective adopted

here. Once we have the fused matrix, we need to choose a clustering algorithm to detect communities of actions that share similar patterns. The next section will present the chosen algorithm and justify its choice.

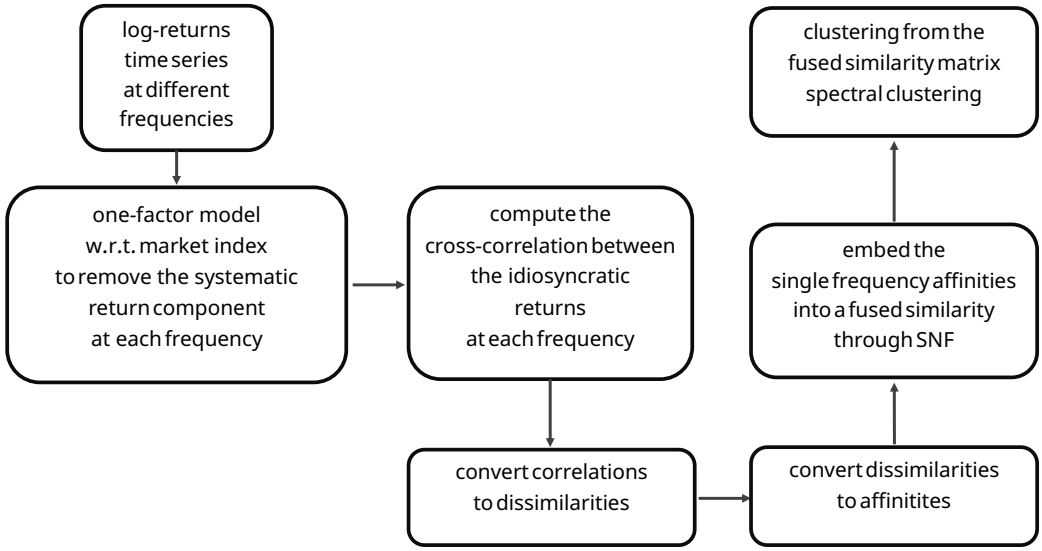


Figure 2.8: Schematic representation of the procedure used to embed different frequencies

2.5 Detecting communities via spectral clustering

The fused similarity matrix produced by the SNF algorithm represents an affinity network rather than a collection of feature vectors in a Euclidean space. Each entry S_{ij} measures the strength of the relationship between samples i and j and results from nonlinear similarity transformations, normalization over local neighborhoods, and an iterative diffusion process across heterogeneous data types. Consequently, S encodes the topological structure of sample-to-sample relationships, not dissimilarities in a metric space. Therefore applying clustering algorithms completely based on Euclidean dissimilarities, such as k-medoids or hierarchical clustering, is conceptually unjustified, as these methods presuppose a globally consistent geometric

structure where cluster membership reflects spatial proximity.

In contrast, the affinity matrix S defines a weighted graph $G = (V, E, W)$ whose edges represent pairwise similarities among samples. This graph structure can, in principle, be analyzed using any network-based community detection method (Fortunato (2010)), for instance modularity maximization Newman (2006) or stochastic block modeling Peixoto (2019).

Spectral clustering Ng et al. (2002); Von Luxburg (2007) naturally occupies an intermediate position between these two domains: it treats S as a graph by constructing its Laplacian, but then relaxes the discrete partitioning problem through an eigen-decomposition of L to obtain a continuous embedding in which Euclidean dissimilarities reflect the underlying graph connectivity. Moreover, the SNF diffusion process itself can be interpreted as a Laplacian-based propagation of similarity information across networks, making spectral clustering not only compatible but conceptually aligned with the way SNF integrates data across modalities. For these reasons, spectral clustering represents a theoretically and geometrically appropriate choice for identifying coherent sample groups in the fused similarity network.

As already suggested above, the similarity matrix effectively represents a similarity graph and spectral clustering is based on the eigen decomposition of the Laplacian of the similarity graph, in this particular case the symmetric normalized Laplacian

$$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}},$$

with I being the identity matrix and D being the diagonal degree matrix of the similarity matrix W

$$D_{ii} = \sum_{j=1}^{N} W_{ij}, \qquad D_{ij} = 0 \;\; (i \neq j).$$

The algorithm then proceeds to compute the first k eigenvectors corresponding to the k smallest eigenvalues of the normalized Laplacian L and builds a new matrix U whose columns are the the k eigenvectors extracted above:

$$U = \left[u_1, u_2, \ldots, u_k\right] \in \mathbb{R}^{N \times k}$$

Finally the rows of U are normalized to unit length:

$$Y_{ij} = \frac{U_{ij}}{\sqrt{\sum_{j=1}^{k} U_{ij}^2}}.$$

Finally, the clusters are extracted by applying k-means to the rows of Y. This discretization step assigns each sample to one of the k clusters based on its position in the spectral embedding.

A central issue in spectral clustering is the determination of the appropriate number of clusters, K, to retain from the eigen-decomposition of the graph Laplacian. Two main strategies are commonly employed: the eigen-gap heuristic and the rotation cost criterion.

The eigen-gap heuristic Ng et al. (2002) relies on inspecting the ordered eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ of the normalized Laplacian. In theory, when the similarity graph consists of K disconnected components, the first K eigenvalues are exactly zero, and the $(K+1)$-th eigenvalue marks a positive jump. In empirical settings with noisy or weighted graphs, one looks for the largest gap between consecutive eigenvalues, that is, $\lambda_{k+1} - \lambda_k$, and interprets the corresponding index k as the optimal number of clusters. While this method is simple and grounded in the spectral theory of graphs, it is often unstable in real-world data where eigenvalue spectra decay smoothly, making it difficult to identify a distinct gap.

An alternative approach is provided by the rotation cost criterion Zelnik-Manor and

Perona (2004), which evaluates the stability of cluster assignments after rotating the continuous spectral embedding towards a discrete partition. In spectral clustering, the eigenvectors of the Laplacian define a continuous $n \times K$ embedding, which is then discretized (typically by k-means) to assign samples to clusters. The rotation cost quantifies the alignment error between the continuous embedding and its closest discrete counterpart, Intuitively, it measures how "well-separated" the eigenvector structure is when projected onto a simplex of cluster indicators. Lower rotation costs indicate a clearer and more stable partitioning of the data. Compared to the eigengap heuristic, the rotation cost provides a more data-driven and robust measure of cluster separability, especially when the graph spectrum lacks clear discontinuities or when clusters have unbalanced sizes and densities. This is precisely the situation encountered in our data, where the fused similarity network exhibits a smoothly decaying spectrum and heterogeneous community sizes. An interesting alternative worth mentioning (Di Nuzzo and Farnè (2025)) employs nuclear-norm penalization of the Laplacian matrix to estimate the number of clusters. This method represents a novel, promising alternative direction for determining K. While we explored alternative methods (the corresponding results are not shown here), these analyses generally produced unstable results. By contrast, the rotation cost criterion yielded partitions that were more stable and more consistent with empirical evidence, and is therefore adopted to determine the number of clusters in the subsequent analyses.

2.6 Empirical Analysis

So far we have discussed the main gaps we identified in the standard methodology and motivated the use of the proposed solutions but, in order to test whether these methodological improvements translate into better empirical performances, we should evaluate them on real market data. For this reason we performed an

extended empirical analysis covering all the aspects outlined above.

The empirical analysis is based on OHLCV (Open, High, Low, Close, Volume) data at frequencies ranging from 1 minute to daily for both a set of stocks and the SPY (SPDR S&P 500 ETF Trust) used as a proxy for the S&P 500 index for a time period of 20 years: from 01/01/2003 to 31/12/2023. The stocks were selected from the stocks included in the S&P 500 index (at the last update before 31/12/2023) for the whole 20 years period, for a final set of 310 stocks. Additionally the GICS sector classification (sourced from Wikipedia) was used to validate the clustering performances. It is a classification of stocks into 11 sectors, representing the main industrial sectors of activity of the companies. These data represent the three main aspects relative to stock movements over time: price, volume and volatility. All the OHLCV data at different frequencies went through a common pre-processing step to handle NAs and missing values. The missing values were replaced with the last observed value in the time series. This is a common practice when dealing with high-frequency data for which missing values are pretty common, especially for less liquid stocks.

The empirical analysis presented in this chapter represents, to the best of our knowledge, the most comprehensive examination of multi-data, multi-frequency correlation structures in the literature. Our dataset comprises 310 stocks, for which we compute pairwise correlations across three distinct dependence measures and six time scales. Each correlation matrix contains approximately $48,000$ unique pairwise correlation coefficients, and we perform this analysis over 143 rolling time windows using multiple estimation methodologies. For return-based correlations, we remove the market factor at each frequency using non-overlapping short windows, employing two different regression methodologies, resulting in hundreds of thousands of performed regressions. In total, our analysis encompasses several hundred million

estimated pairwise correlation coefficients, allowing us to comprehensively characterize the evolution of the world's most important equity market over the past 20 years, including two of the most extensively studied crises: the subprime mortgage crisis and the COVID-19 pandemic.

The information contained in these correlation structures is jointly explored along both the cross-variable and cross-frequency dimensions through SNF and spectral clustering. We perform clustering analyses at multiple levels: on single-frequency and single-measure matrices, on cross-frequency fused networks that integrate information across temporal scales, and on cross-variable fused networks that combine different measures. This multi-level approach yields tens of thousands of clustering solutions, enabling us to comprehensively characterize the multi-scale dependence structure of equity markets and to evaluate the stability and robustness of the identified groupings across different market conditions, estimation procedures, and temporal resolutions.

The three aspects to cover in the analysis are related to robustness, joint information from different data types and joint information from different frequencies. The last two aspects are more specific to the type of information one wants to extract while the first one is more of a general aspect that can be applied in any type of correlation based financial analysis independently of the filtering or clustering methodology used. For this reason we will first discuss some results related to robustness on single frequency returns by showing the results of using both frameworks, the standard one and the novel one we introduced, on dissimilarities from both robust and non-robust methodologies and then, when addressing the remaining two aspects, we also present and compare the results obtained under both robust and non-robust dissimilarities. The results for all methodologies reported below are derived using a uniform analysis approach, based on a rolling window applied to daily data with window length $T = 1000$ and window shift $h = 30$.

The de-facto standard for validating clustering results in these studies is to check the overlap between detected communities and economic sector classifications. This validation approach addresses a fundamental challenge: there is no objective ground truth for what constitutes a "correct" clustering of stocks and therefore the communities we seek to identify from correlation networks are latent structures whose validity cannot be directly verified. However, economic theory provides clear expectations about which companies should cluster together. Firms operating in the same sector, sharing similar factor exposures and regulatory environments, face common economic fundamentals that should manifest as correlated behavior in their time series. Therefore, if detected communities show strong alignment with sector classifications, this suggests the clustering algorithm has successfully identified economically meaningful information embedded in the correlation structure. Conversely, communities that are completely disconnected from sectoral boundaries would be difficult to interpret and economically implausible, likely reflecting statistical artifacts rather than genuine underlying relationships. While sector alignment has clear limitations, relying on business models that may no longer fully reflect modern, service-driven economies, it is still an important benchmark since it connects clustering results to familiar economic intuition about how firms relate in the real economy.

In this work we employ the Global Industry Classification Standard (GICS) 11-sector taxonomy, developed jointly by MSCI and Standard & Poor's. GICS-11 assigns companies to sectors based on their main business activity, determined through a combination of revenue sources, earnings profiles, and market perception.

Following the standard practice in the literature, we quantify the overlap between clusters and economic sectors using the Adjusted Rand Index (ARI). Previous studies show that best-in-class methods achieve ARI values in the range 0.3–0.5 for daily

data, while ARIs are generally much lower at higher intraday frequencies, reflecting the increased difficulty of recovering sector structure at finer temporal resolutions. In the following sections, we use the ARI as the primary validation measure, and for our best-performing clustering method, we complement it with additional, original validation approaches. The empirical analysis, as outlined above, is structured as a sequence of interdependent processing stages, where each step builds on the output of the previous one. Within each stage, estimations are conducted on a rolling-window basis. Because computations across windows are conditionally independent, they are parallelized whenever possible. Thus, while the overall analytical workflow follows a sequential pipeline, the implementation exploits parallel processing at the window level to improve computational efficiency and reduce overall runtime.

2.6.1 Robustness Consistently Pays Off

As a first step we applied the standard framework, outlined in figure 2.1, on returns at different frequencies by using both robust and non-robust methodologies. The results are shown in figures 2.9 and 2.10 which shows the ARIs between the clusters extracted by applying the standard method on both robust and non-robust idiosyncratic returns. In this case the number of clusters has been fixed to 11, the number of GICS sectors. Both figures provide pretty strong evidence in favor of using robust methods and, interestingly, the higher the frequency the bigger the improvement in terms of ability to retrieve the underlying economic sector when using robust methods instead of standard non-robust ones.

Moving on to our methodology, figures 2.11 and 2.12 show the results of applying our novel framework, based on spectral clustering, on returns at different frequencies by using both robust and non-robust methodologies. The results are consistent with the standard methodology in the fact that robust idiosyncratic returns provide better results in general with significant improvements for higher frequencies. These



Figure 2.9: time series of Adjusted Rand Index between clusters extracted by applying AL hierarchical clustering clustering to the affinity matrices extracted by computing the Pearson correlation between OLS idiosyncratic returns (in blue) and Tyler correlations between LTS idiosyncratic returns (in green) on returns sampled at different frequencies.

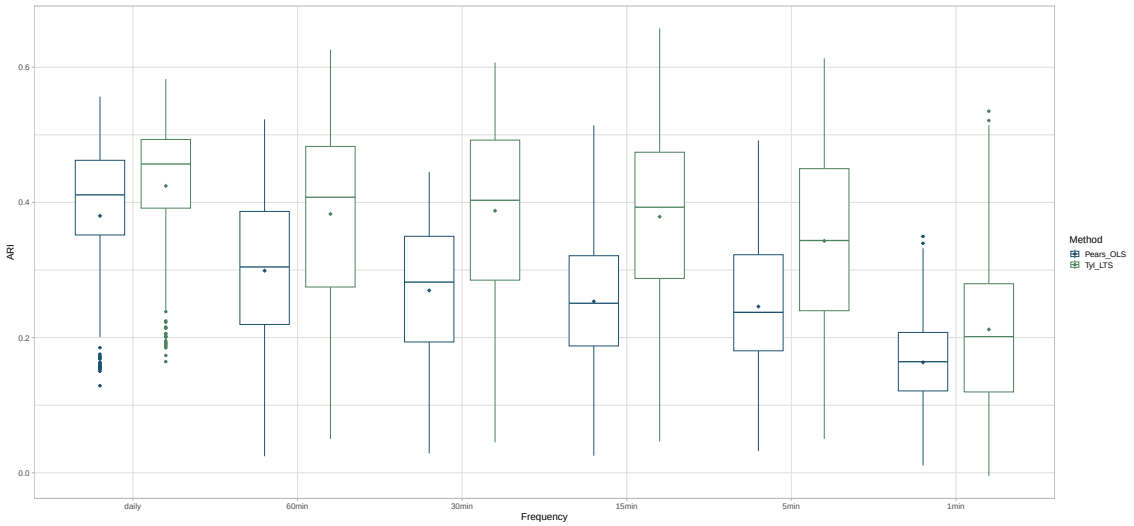


Figure 2.10: box plots of Adjusted Rand Index between clusters extracted by applying AL hierarchical clustering to the affinity matrices extracted by computing the Pearson correlation between OLS idiosyncratic returns (in blue) and Tyler correlations between LTS idiosyncratic returns (in green) on returns sampled at different frequencies.

findings suggest that as the frequency increases and noise becomes more dominant, relying on robust methods is crucial to capture the true underlying relations between the assets.



Figure 2.11: time series of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Pearson correlation between OLS idiosyncratic returns (in blue) and Tyler correlations between LTS idiosyncratic returns (in green) on returns sampled at different frequencies.

Since the next step of the analysis is concerned with the integration of different data sources when detecting the communities, it is worth investigating the impact of robust methods also individually on the alternative data sources, namely volumes and volatilities, since these data might have different features compared to returns and therefore respond differently to robust estimation methods. In this case robustness concerns only the correlation computation since we work directly with the original data and not with extracted idiosyncratic components. Therefore we applied our novel framework to both volumes and volatilities individually on each frequency and the results, in figures 2.13 and 2.14, show a marginal improvement for the volumes but a massive one for volatilities which is even stronger compared to the improvement we found on idiosyncratic returns. This could be due to the

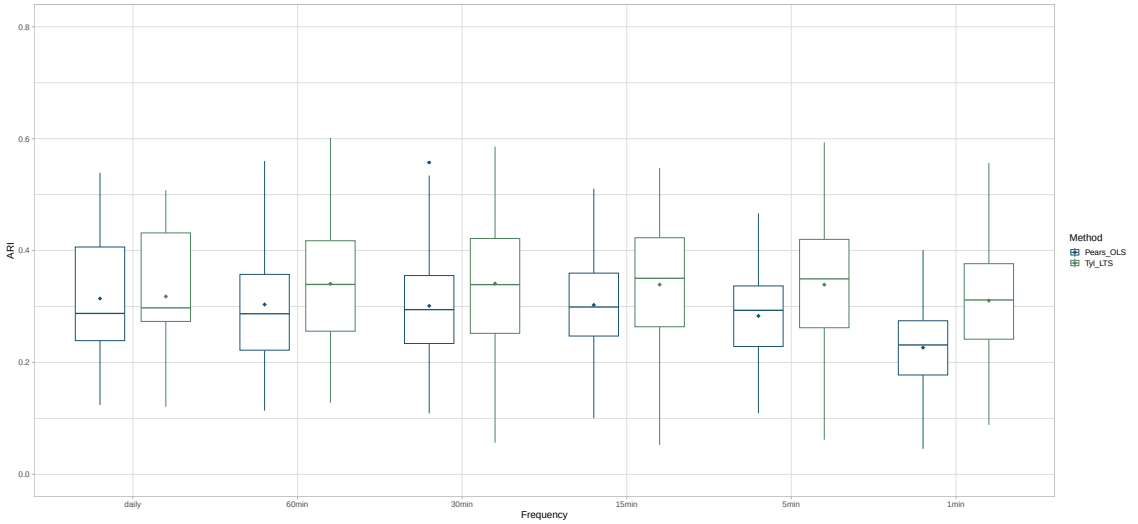


Figure 2.12: box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Pearson correlation between OLS idiosyncratic returns (in blue) and Tyler correlations between LTS idiosyncratic returns (in green) on returns sampled at different frequencies.

fact that volatility time series tend to be noisier and exhibit more extreme values compared to other measures, making make robust methods more effective.

2.6.2 Uncovering Enhanced Structures via Horizontal Data Integration

Having established the advantages of robust methods over non-robust ones, we now turn to the results obtained from jointly analyzing the different data types instead of focusing solely on returns. First we will show the results for the different frequencies using robust methodologies for all the data types and then, as a further argument in favor of robustness we will show the differences between fused data from both robust and non-robust methods. Figures 2.15 and 2.16 show the results for the clusters extracted from returns and from the fused matrix embedding information from returns, volatility and volumes. There is a clear improvement on the ability to retrieve the economic sectors at each single frequency, proving that volatilities and volumes

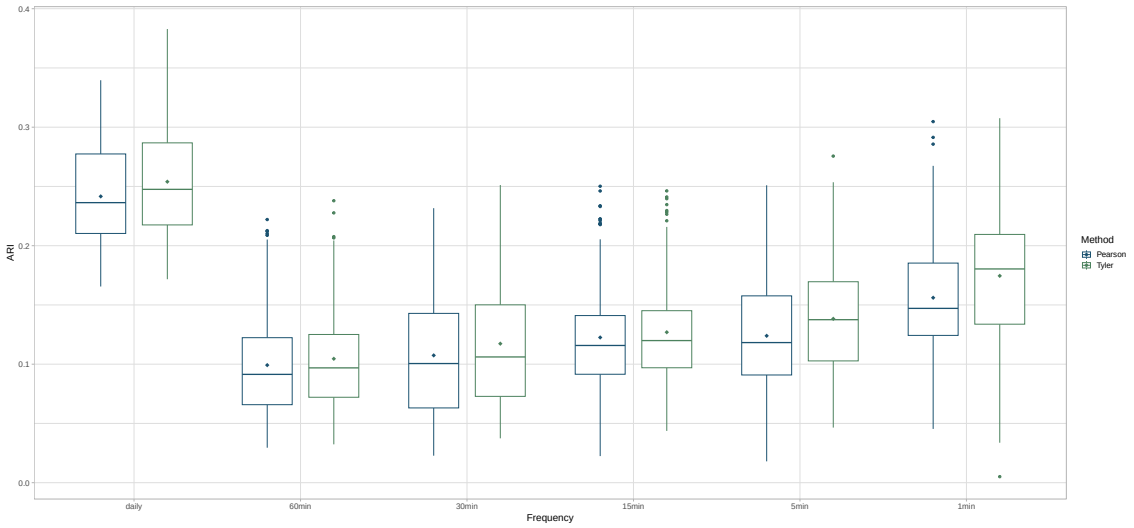


Figure 2.13: box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Pearson correlation (in blue) and Tyler correlations (in green) between volumes sampled at different frequencies.

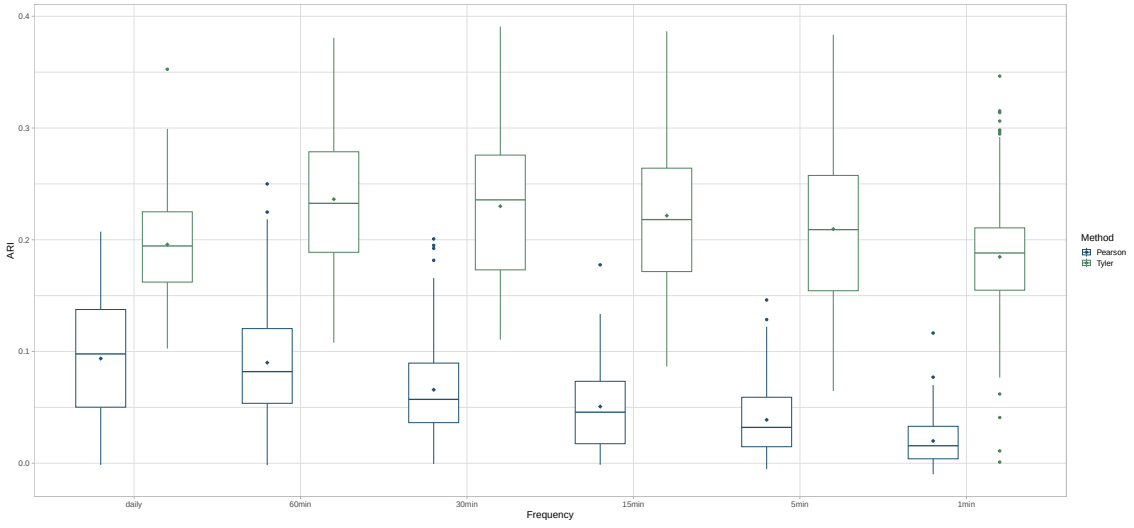


Figure 2.14: box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Pearson correlation (in blue) and Tyler correlations (in green) between volatilities sampled at different frequencies.

indeed provide complementary information with respect to returns and can therefore improve over the standard method at any frequency. We have already shown the advantage of using robust methodologies over standard ones on single data types at any frequency and reached the general finding that robust methodologies, being less sensitive to noise, perform significantly better at higher frequencies. The next step is to check the importance of robustness also in the case of communities extracted from a single matrix embedding the affinities extracted from returns, volumes and volatilities. Intuitively one would expect that merging together affinities from noisy correlation estimates would lead to an ever more noisy final matrix, resulting in worse overall performances. As shown in figure 2.17 this is exactly what happens when using non-robust fused matrices to extract the communities with respect to robust ones.

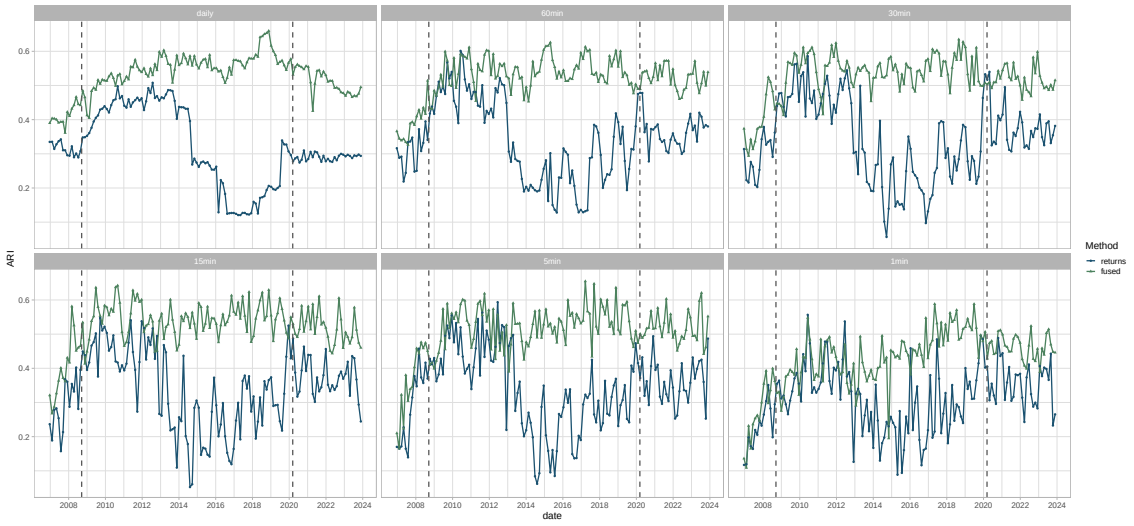


Figure 2.15: time series of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Tyler correlation on LTS idiosyncratic returns (in blue) and on the fused matrix from Tyler correlations computed on LTS idiosyncratic returns, volumes and volatilities (in green) sampled at different frequencies.

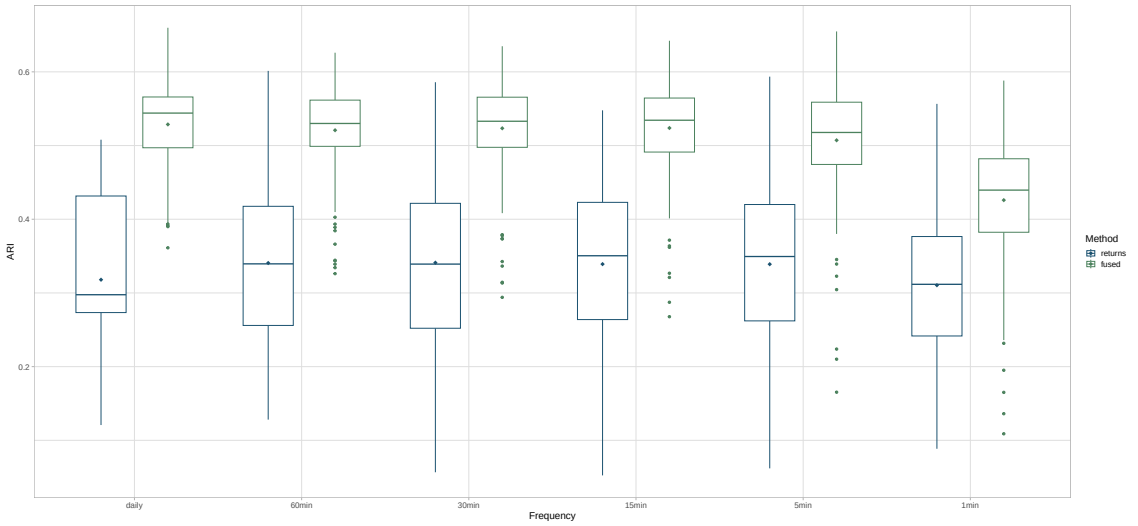


Figure 2.16: box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Tyler correlation on LTS idiosyncratic returns (in blue) and on the fused matrix from Tyler correlations computed on LTS idiosyncratic returns, volumes and volatilities (in green) sampled at different frequencies.

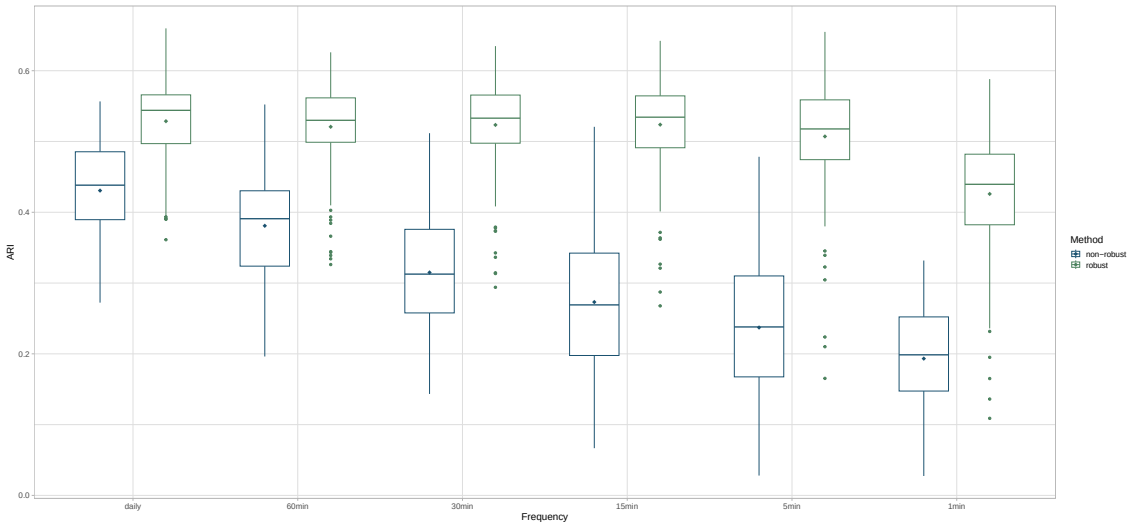


Figure 2.17: box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the fused matrices from Pearson correlations computed on OLS idiosyncratic returns, volumes and volatilities (in blue) and Tyler correlations computed on LTS idiosyncratic returns, volumes and volatilities (in green) sampled at different frequencies.

2.6.3 Vertical Data Integration Reveals the Most Coherent Group Structures

Now we move onto the main result of the chapter coming from the joint consideration of returns at different frequencies.

Figures 2.18 and 2.21 show the adjusted rand indexes with respect to economic sectors for both fixed and varying clusters, extracted from both individual frequencies and the similarity matrix embedding information from different frequencies. Both clustering relied on robust methods for returns detrending and correlation computation. The results here are pretty striking, with the clustering extracted from the fused matrix embedding information from different frequencies showing a remarkable overlap with economic sectors in both the fixed and varying clusters cases showing that returns across different frequencies convey unique information that, when jointly analyzed, significantly strengthens the recovery of economic sectors.

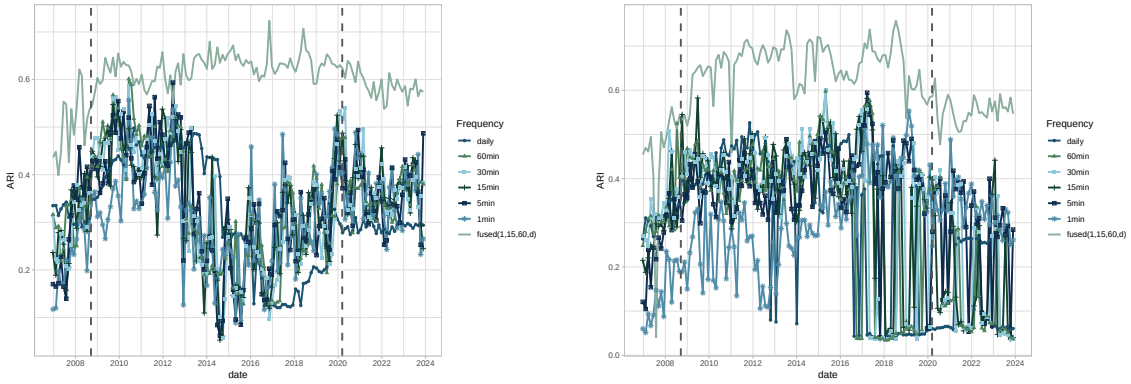


Figure 2.18: time series of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Tyler correlation on LTS idiosyncratic returns individually for each and on the fused matrix from Tyler correlations computed on LTS idiosyncratic returns at different frequencies for fixed clusters (left) and varying clusters (right).

As pointed out above, checking the overlap with economic sectors is the de facto standard validation procedure for these methods, under the assumption that a higher

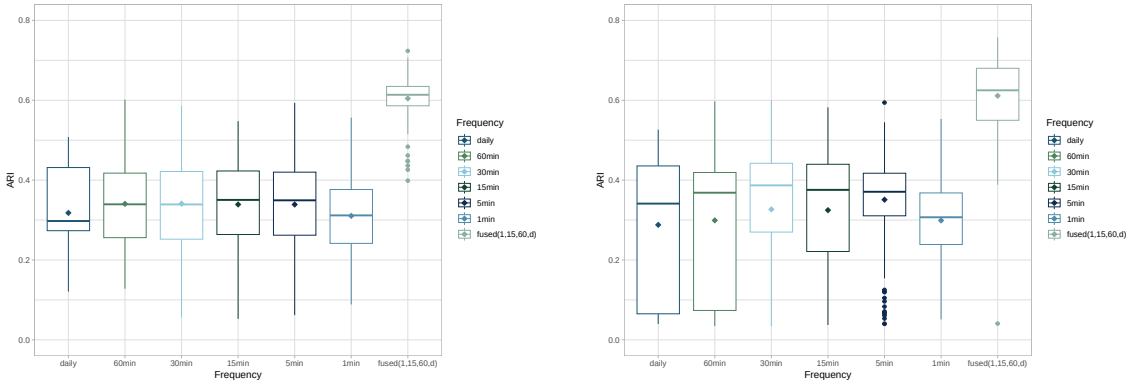


Figure 2.19: time series of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Tyler correlation on LTS idiosyncratic returns individually for each and on the fused matrix from Tyler correlations computed on LTS idiosyncratic returns at different frequencies for fixed clusters (left) and varying clusters (right).

degree of overlap reflects a better extraction of economically meaningful information. However, since we consider it to capture only part of the overall performance landscape, we supplemented it with further indicators comparing the time series of stocks within each cluster to the market average, specifically information ratios and realized variances computed as described below.

Cluster-wise performances: Information Ratio. As a further measure we implemented a framework based on the Information Ratio (IR), an indicator used in finance to measure the risk-adjusted returns of a portfolio with respect to a benchmark. The IR measures the excess return per unit of active risk and is defined as

$$\mathrm{IR} = \frac{\mathrm{E}(r_p - r_b)}{\sqrt{\mathrm{Var}(r_p - r_b)}},$$

where r_p denotes the return of the portfolio, while r_b is the return of the benchmark, and $\sqrt{\mathrm{Var}(r_p - r_b)}$ is the standard deviation of the active returns, also defined as the tracking error. Therefore the higher the IR, the better the performances of the portfolio with respect to the benchmark. We build upon this measure by implementing the following framework. Let $r_{i,t}$ denote the return of asset i at time

t within a given time window of length T, and let the assets be grouped either by detected clusters or by economic sectors. Denote by C the set of clusters or sectors, and by N_c the number of assets in cluster/sector $c \in C$. For each cluster/sector c, we construct an equally weighted portfolio:

$$r_{c,t} = \frac{1}{N_c} \sum_{i \in c} r_{i,t}.$$

The IR of cluster/sector c with respect to the market portfolio $r_{m,t}$ is then

$$\mathrm{IR}_c = \frac{\mathbb{E}[r_{c,t} - r_{m,t}]}{\sqrt{\mathrm{Var}(r_{c,t} - r_{m,t})}}$$

To obtain a single measure for the time window, the IRs are aggregated using a volume-weighted average. Let V_c be the total traded volume of cluster/sector c over the time window. The aggregated Information Ratio is defined as

$$\mathrm{IR}_{\mathrm{agg}} = \frac{\sum_{c \in C} V_c \, \mathrm{IR}_c}{\sum_{c \in C} V_c}.$$

This procedure produces a weighted average Information Ratio for each time window, for both clusters and sectors, providing a principled way to evaluate the efficiency of the stock groupings.

Cluster-wise variability: Realized volatilities. To complete the picture, we also measure the internal cluster variability using a procedure analogous to the one employed for the Information Ratios described above. This measure is based on the realized volatilities of cluster-wise equally weighted portfolios, computed from cluster-wise realized covariance matrices derived from 5-minute returns. Formally, let $\{r_t\}_{t=1}^{T}$ denote a c-dimensional time series of intraday log-returns for c assets in

a given cluster or sector. The realized covariance matrix ($\widehat{\Sigma}$) is defined as

$$\widehat{\Sigma} = \sum_{t=1}^{T} r_t r_t^\top,$$

where $r_t \in \mathbb{R}^c$ is the vector of returns at time interval t. The diagonal entries of $\widehat{\Sigma}$ correspond to the realized variances of individual assets, while the off-diagonal entries represent the realized covariances between pairs of assets. For each cluster/sector c with N_c assets, we construct an equally weighted portfolio and compute its realized variance using the cluster-wise covariance matrix:

$$RV_c = \frac{1}{N_c^2} \mathbf{1}^\top \widehat{\Sigma}_c \mathbf{1}$$

where $\mathbf{1}$ is an N_c-dimensional vector of ones, and $\widehat{\Sigma}_c$ is the realized covariance matrix of cluster/sector c. The corresponding realized volatility is then

$$\sigma_c = \sqrt{RV_c}.$$

To obtain a single aggregated measure of internal variability for the time window, the realized volatilities are combined using a volume-weighted average:

$$\sigma_{\text{agg}} = \frac{\sum_{c \in C} V_c \, \sigma_c}{\sum_{c \in C} V_c}.$$

As with the Information Ratio, market average returns are used as a benchmark, and GICS economic sectors serve as an alternative grouping alongside the cluster partitions, which can have a fixed or time-varying number of clusters. This framework provides a consistent way to evaluate both the efficiency and internal coherence of the stock groupings. Figure 2.20 shows the information ratios, average extra returns and standard deviation of the average extra returns with respect to the mar-

ket average for the three groupings considered, namely: economic sectors, cluster numbers fixed to 11 and varying cluster numbers over time windows. The results show that, while both sectors and fixed clusters are somewhat neutral with respect to the market the varying clusters show positive information ratios with significantly smaller standard deviations. These results suggest that financial markets exhibit an intrinsic tendency to self-organize efficiently into economically meaningful groups. However, this efficient self-organization becomes apparent only when information from multiple frequencies is jointly considered.

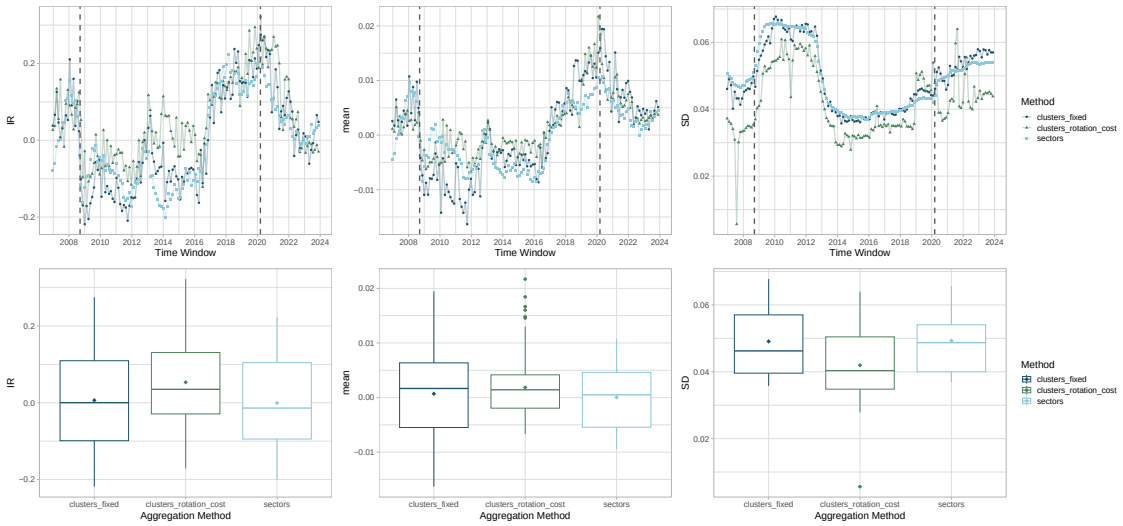


Figure 2.20: Information ratios of the individual cluster-wise/sector-wise returns with respect to market average return (left). average individual cluster-wise/sector-wise extra returns with respect to market average (center). standard deviation of the individual cluster-wise/sector-wise extra returns with respect to market average (right). All statistics are aggregated by taking the weighted mean, weighting each cluster by the total traded volume of its constituents over the period.

The results for all the methods and frequencies considered are summarized in table 2.2 which outlines the superiority of the proposed methods over the current state of the art regarding both robustness and data integration. Now we turn our attention to a more intuitive validation method applied to the best performing method we have, namely the one combining cross-frequency integration and robust methods.

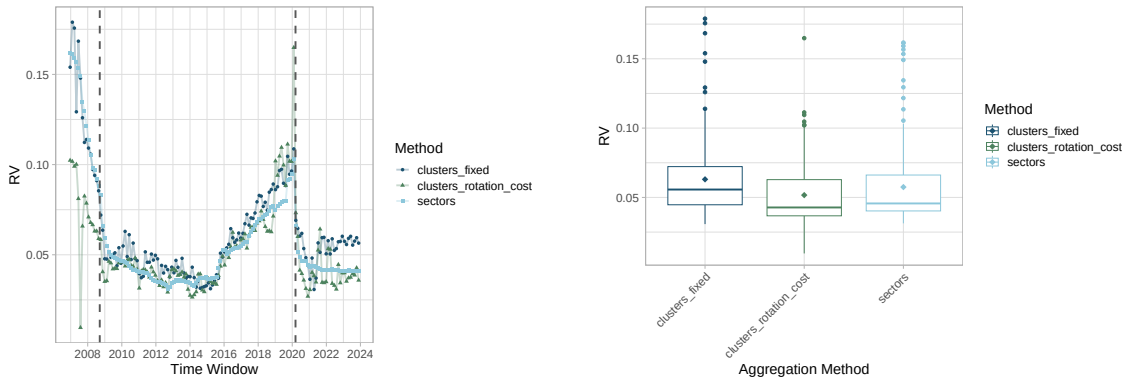


Figure 2.21: Left: time series of the percentage increase of Realized volatilities of the individual cluster-wise/sector-wise portfolios with respect to the realized volatility of the market portfolio. Right: box plots of the percentage increase of Realized volatilities of the individual cluster-wise/sector-wise portfolios with respect to the realized volatility of the market portfolio. All statistics are aggregated by taking the weighted mean, weighting each cluster by the total traded volume of its constituents over the period.

Acronym	Frequency	Data type	Clustering	ARI with sectors				Information Ratio				Volatility				N clusters			
				Mean	Median	SD	IQR	Mean	Median	SD	IQR	Mean	Median	SD	IQR	Mean	Median	SD	IQR
Baseline	daily	returns	Sector					−0.001	−0.014	0.113	0.199	0.057	0.046	0.03	0.026				
Baseline	daily	returns	clust_fixed	0.314	0.288	0.096	0.167	0.041	0.023	0.142	0.189	0.053	0.045	0.025	0.027				
Baseline	daily	returns	clust_rotation	0.1	0.095	0.06	0.034	−0.001	−0.010	0.30	0.324	0.034	0.007	0.079	0.090	8.29	2	15.5	1
Baseline	60min	returns	Sector																
Baseline	60min	returns	clust_fixed	0.304	0.287	0.103	0.135	0.030	0.016	0.132	0.175	0.052	0.045	0.025	0.029				
Baseline	60min	returns	clust_rotation	0.073	0.082	0.049	0.049	−0.083	−0.106	0.276	0.357	0.074	0.009	0.032	0.007	3.07	2	4.1	1
Baseline	15min	returns	Sector																
Baseline	15min	returns	clust_fixed	0.303	0.299	0.085	0.112	0.016	0.002	0.129	0.188	0.055	0.025	0.049	0.029				
Baseline	15min	returns	clust_rotation	0.104	0.073	0.073	0.087	−0.057	−0.057	0.288	0.32	0.049	0.010	0.082	0.032	13.4	3	19.8	3
Baseline	1min	returns	Sector																
Baseline	1min	returns	clust_fixed	0.225	0.231	0.075	0.097	0.054	0.049	0.134	0.177	0.057	0.051	0.028	0.03				
Baseline	1min	returns	clust_rotation	0.157	0.185	0.079	0.103	0.055	0.032	0.16	0.211	0.175	0.155	0.159	0.088	42.1	47	14.5	4
Baseline	fused	returns	Sector					−0.001	−0.014	0.113	0.199	0.057	0.046	0.03	0.026				
Baseline	fused	returns	clust_fixed	0.552	0.558	0.08	0.085	0.019	0.015	0.131	0.202	0.055	0.055	0.027	0.029				
Baseline	fused	returns	clust_rotation	0.476	0.528	0.191	0.171	0.015	0.008	0.155	0.18	0.048	0.042	0.027	0.028	8.09	9	2.82	1
A	daily	returns	Sector					−0.001	−0.014	0.113	0.199	0.057	0.046	0.03	0.026				
A	daily	returns	clust_fixed	0.318	0.298	0.108	0.158	0.035	0.022	0.15	0.177	0.051	0.041	0.027	0.028				
A	daily	returns	clust_rotation	0.288	0.198	0.341	0.37	−0.047	−0.034	0.145	0.144	0.078	0.074	0.058	0.101	30.3	47	22.2	47
A	60min	returns	Sector																
A	60min	returns	clust_fixed	0.341	0.339	0.11	0.162	0.017	0.010	0.112	0.159	0.049	0.046	0.021	0.028				
A	60min	returns	clust_rotation	0.299	0.308	0.17	0.346	−0.020	−0.025	0.188	0.222	0.089	0.085	0.071	0.109	31.6	45	20.7	44
A	15min	returns	Sector																
A	15min	returns	clust_fixed	0.339	0.35	0.113	0.159	0.017	−0.006	0.138	0.139	0.046	0.045	0.021	0.021				
A	15min	returns	clust_rotation	0.325	0.375	0.155	0.218	0.005	−0.014	0.173	0.22	0.103	0.105	0.077	0.050	34.8	44	18.5	5
A	1min	returns	Sector																
A	1min	returns	clust_fixed	0.31	0.312	0.098	0.135	0.042	0.033	0.158	0.196	0.055	0.051	0.025	0.03				
A	1min	returns	clust_rotation	0.299	0.307	0.107	0.120	0.045	0.039	0.128	0.188	0.157	0.147	0.082	0.002	45.2	45	2.05	3
A	fused	returns	Sector					−0.001	−0.014	0.113	0.199	0.058	0.048	0.03	0.026				
A	fused	returns	clust_fixed	0.505	0.514	0.051	0.049	0.005	0.000	0.128	0.208	0.040	0.048	0.009	0.018				
A	fused	returns	clust_rotation	0.611	0.625	0.091	0.15	0.055	0.035	0.105	0.16	0.026	0.020	0.017	0.034	9.07	9	1.5	1
B	fused	fused	Sector					−0.001	−0.014	0.113	0.199	0.058	0.046	0.03	0.026				
B	fused	fused	clust_fixed	0.55	0.554	0.065	0.083	0.011	−0.002	0.129	0.209	0.055	0.055	0.029	0.027				
C	fused	fused	clust_rotation	0.496	0.533	0.157	0.136	0.026	0.016	0.146	0.182	0.053	0.046	0.029	0.023	8.64	9	2.448	2
C	fused	fused	Sector					−0.001	−0.014	0.113	0.199	0.058	0.046	0.03	0.026				
D	fused	fused	clust_fixed	0.552	0.575	0.071	0.065	0.005	−0.020	0.125	0.219	0.055	0.057	0.031	0.027				
D	fused	fused	clust_rotation	0.423	0.501	0.188	0.347	0.026	0.016	0.152	0.200	0.048	0.046	0.03	0.026	8.31	8	3.741	4
D	fused	fused	Sector					−0.001	−0.014	0.113	0.199	0.058	0.046	0.03	0.026				
D	fused	fused	clust_fixed	0.552	0.558	0.065	0.055	0.024	0.010	0.121	0.19	0.057	0.059	0.031	0.026				
D	fused	fused	clust_rotation	0.274	0.185	0.212	0.444	−0.066	−0.042	0.219	0.329	0.032	0.027	0.026	0.037	5.53	5	3.110	3.110

Table 2.2: Statistical analysis results across different configurations

2.6.4 Meta-Clusters and Co-Occurrence Networks: A Higher-Level Interpretation and Validation Framework

Having established that vertical data integration yields the most coherent and stable community structures, we now introduce a further methodological innovation of this thesis, consisting of a higher-level procedure aimed at validating and interpreting the clustering results from a broader perspective. To this end, we propose a two-step approach based on meta-clusters and co-occurrence frequency networks. The method first identifies meta-clusters by clustering the community labels obtained across multiple configurations, thus capturing the most persistent structural patterns. In the second step, statistically validated co-occurrence networks are constructed to represent the frequency with which assets are assigned to the same meta-cluster. This novel framework adds an additional layer of validation and interpretability, emphasizing the core relationships that consistently emerge across time periods. We begin by carefully describing the procedure used to extract both the meta-clusters and the subsequent co-occurrence frequency networks, and then discuss the results obtained.

Meta-cluster extraction. The first step of the procedure consists in extracting meta-clusters from the cluster labels of the various time windows to analyze the overlap between the meta-clusters and meaningful economic periods. In particular, from set of cluster labels across time windows $\mathcal{D} = \{L_1, L_2, \ldots, L_T\}$ we compute the variation of information (VI) Meilă (2007) between each combination of cluster labels for all the time windows to obtain a dissimilarity matrix $D(L)$ between each time window L and then we use the dissimilarity matrix $D(L)$ as input to a Ward clustering Ward Jr (1963) algorithm to extract the meta-clusters $M(L_t) \in \{1, 2, \ldots, K\}$ to obtain K meta-clusters:

$$M = \{m_1, m_2, \ldots, m_K\}$$

Therefore this procedure allows us to extract meta clusters in which the clustered elements are the cluster labels L over time and check the overlap of the meta-clusters with meaningful economic periods.

Statistically validated co-occurrence networks. The second step consists of building statistically validated co-occurrence networks for each meta-cluster. More in detail, for each one of the previously extracted meta clusters m, we create two subsets of the original cluster labels data-frame, $\mathcal{S}_{\Updownarrow}$ and $\mathcal{S}_{-\Updownarrow}$ consisting of all the cluster label windows included in the meta-cluster m and another including all cluster labels but the ones belonging to the meta-cluster m:

$$\mathcal{S}_m = \{L_t \in \mathcal{D} \ : \ M(L_t) = m\}, \qquad \mathcal{S}_{-m} = \{L_t \in \mathcal{D} \ : \ M(L_t) \neq m\}$$

Then, from the two subsets $\mathcal{S}_{\Updownarrow}$ and $\mathcal{S}_{-\Updownarrow}$, we build two co-occurrence frequency matrices in which the elements are the original stocks of the sample while the entries are the frequencies with which two stocks appear in the same cluster in the two subsets:

$$F_{ij}^{(m)} = \frac{1}{|\mathcal{S}_m|} \sum_{L_t \in \mathcal{S}_m} \mathbf{1}(L_t(i) = L_t(j)) \qquad F_{ij}^{(-m)} = \frac{1}{|\mathcal{S}_{-m}|} \sum_{L_t \in \mathcal{S}_{-m}} \mathbf{1}(L_t(i) = L_t(j))$$

Since the entries of the co-occurrence frequency matrices are in fact frequencies of co-occurrence of the same pair of elements, we can interpret the entries $F_{ij}^{(m)}$and $F_{ij}^{(-m)}$ as empirical success probabilities of a Bernoulli trial over N_m and N_{-m} trials respectively:

$$\hat{p}_{ij}^{(m)} = F_{ij}^{(m)} = \frac{x_{ij}^{(m)}}{N_m} \qquad \hat{p}_{ij}^{(-m)} = F_{ij}^{(-m)} = \frac{x_{ij}^{(-m)}}{N_{-m}}$$

where $x_{ij}^{(m)}$ and $x_{ij}^{(-m)}$ are, respectively, the count of windows in $\mathcal{S}_m$ and $\mathcal{S}_{-m}$ where stocks i and j co-occur. We can then use the probabilities $\hat{p}_{ij}^{(m)}$ and $\hat{p}_{ij}^{(-m)}$ to run a z-test for the difference between two independent proportions:

$$z_{ij} = \frac{\hat{p}_{ij}^{(m)} - \hat{p}_{ij}^{(-m)}}{\widetilde{\mathrm{SE}}_{ij}}$$

where $\mathrm{SE}_{ij} = \sqrt{\sigma_{ij}^{(m)2} + \sigma_{ij}^{(-m)2}}$ and $\sigma_{ij}^{(m)2} = \frac{\hat{p}_{ij}^{(m)}\left(1-\hat{p}_{ij}^{(m)}\right)}{N_m}$, $\sigma_{ij}^{(-m)2} = \frac{\hat{p}_{ij}^{(-m)}\left(1-\hat{p}_{ij}^{(-m)}\right)}{N_{-m}}$
This procedure allows us to obtain a 0/1 matrix B in which 1s represent significant differences while 0s represent non-significant differences in the co-occurrence frequencies between the two frequency matrices $F^{(m)}$ and $F^{(-m)}$:

$$B_{ij} = \begin{cases} 1 & |z_{ij}| > z^* \\ 0 & \text{otherwise} \end{cases}$$

Now we build a new matrix $\tilde{D}$ whose elements are the element wise differences between the two adjacency matrices $F^{(m)} - F^{(-m)}$ in the case where the difference is meaningful and 0 otherwise:

$$\tilde{D_{ij}} = S_{ij}\left(\hat{p}_{ij}^{(m)} - \hat{p}_{ij}^{(-m)}\right)$$

This new matrix $\tilde{D}$ can be viewed as the adjacency matrix representing a signed weighted network whose edges are the meaningful differences in the co-occurrence frequencies between the two subset frequency networks $F^{(m)}$ and $F^{(-m)}$. The last step then consists of aggregating the nodes by GICS-11 sector and then building a summary adjacency matrix A in which nodes represent economic sectors and edges represent the average value of the edges between nodes in the previous network $\tilde{D}$. This procedure leads us to K aggregated co-occurrence frequency networks whose

edges represent statistically validated average over/under expressions between economic sectors for the time windows included in a meta-cluster m with respect to the rest of the time windows m^{-}. in this way we can directly analyze the evolution of the between-sectors relationships over time.

The results of this procedure, for the case of robust cross-frequency SNF, are shown in figure 2.22. In the top figure we can observe a striking overlap between meta clusters and meaningful economic periods while in the bottom part of the figure we can observe the differences between the co-occurrence frequency networks extracted from these meta clusters. There are some interesting patterns worth investigating. The first meta cluster, roughly overlapping with the subprime crisis period, shows a strong over expression of connections of the Real Estate sector with most of the other sectors in the graph, highlighting its central role in the crisis. Meta cluster 2, corresponding approximately to the aftermath of the subprime crisis and the European debt crisis, shows a persistence of the over expression of the Real Estate links, even though with smaller values and, interestingly, a strong positive co-occurrence frequency between energy and utilities, likely reflecting the impact of the oil price crisis of 2014. Moving on meta cluster 3, representing a relatively stable economic period, we observe mostly negative co-occurrences with respect to the rest, likely indicating that in stable period clusters tend to overlap more with economic sectors and therefore create less cross-sectoral clusters. In meta-cluster 4, then, corresponding to the first phase of the COVID-19 pandemic period, we observe an over-expression of co-occurrences between Utilities and Financials, Consumer Staples, Consumer Discretionary, and Industrials. This pattern likely indicates a liquidity-driven regime in which massive policy interventions and market stress temporarily erased sectoral distinctions, causing defensive and cyclical sector stocks to move in unison. Finally, meta-cluster 5 shows is a strong over-expression of co-occurrences between Materials, Consumer Discretionary, and Industrials, likely reflecting the broad recovery

and reactivation of the real economy following the pandemic. A pronounced over-expression between Financial and Health Care likely reflecting the central role of the Health Care sector in the post-crisis adjustment phase and its increasing relevance for financial markets. Finally, the strong over-expression between Energy and Consumer Staples, Consumer Discretionary, and Industrials likely captures the impact of the energy crisis triggered by the Russian invasion of Ukraine, which intensified inter-sectoral linkages through price and supply-chain shocks.

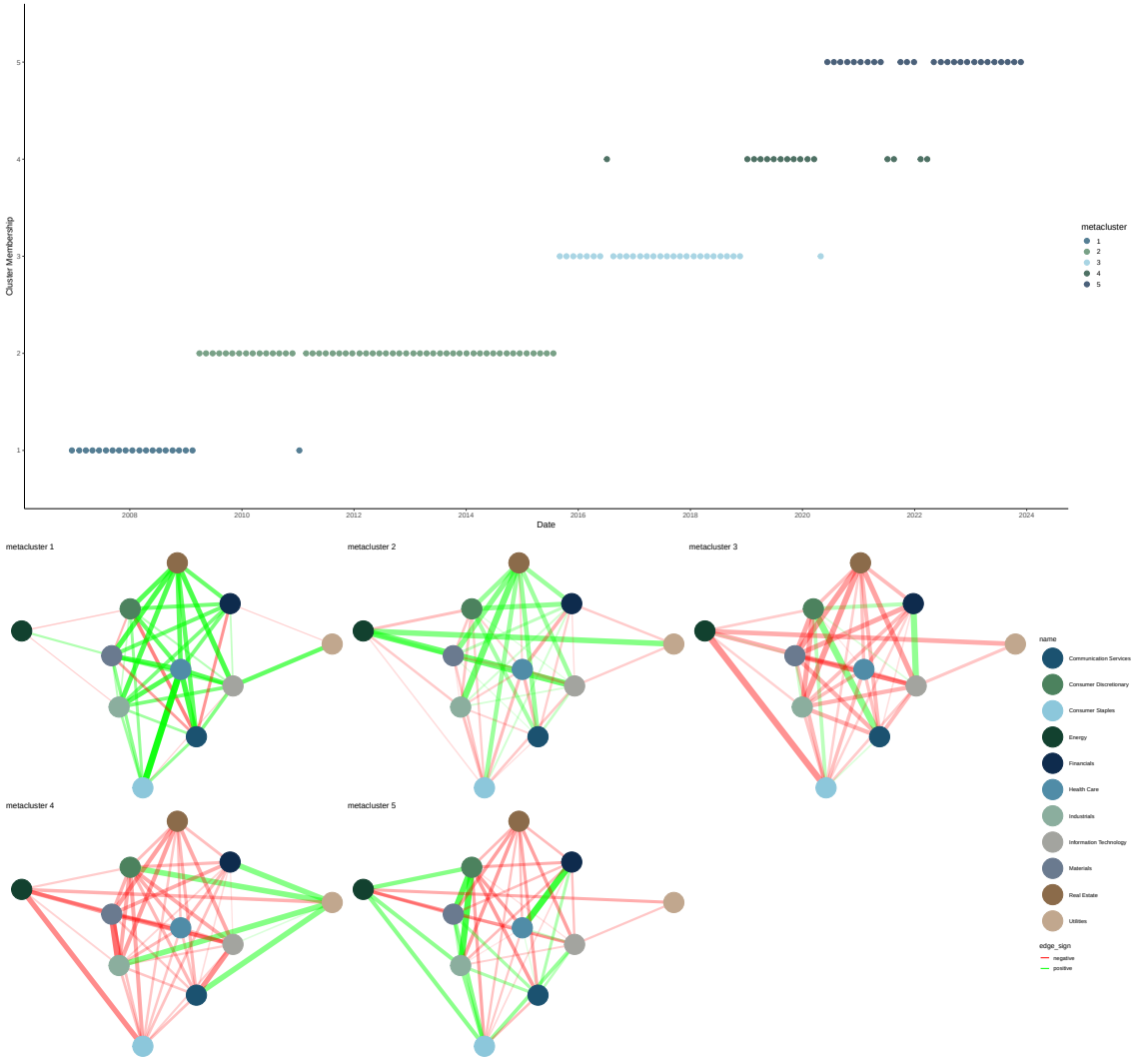


Figure 2.22: Top: time series of the meta cluster labels over time. Bottom: sparse co-occurrence frequency graphs for the meta clusters above.

2.7 Concluding remarks

This chapter has addressed two fundamental yet insufficiently explored aspects of stock correlation networks through highly relevant methodological advances supported by an extensive multi-frequency empirical investigation. The joint implementation of robust estimation and heterogeneous information integration reveals that financial markets possess an intrinsic capacity for efficient self-organization, a property that becomes empirically observable only when multiple frequencies are jointly considered within appropriate robust statistical frameworks.

Several promising directions emerge for future research. First, while our robustness contribution has demonstrated substantial improvements through robust correlation estimation, future investigations may extend this framework by applying it to alternative community detection algorithms and exploring different robust statistical methodologies beyond those considered here.

Second, the data integration framework we have implemented exhibits remarkable generalizability, as it can be virtually applied to any set of time series describing various aspects of a set of stocks. A particularly promising extension involves the integration of individual stock sentiment data, which would capture latent market dimensions not fully evident from traditional information sources. Such incorporation could further enhance community detection performances and reveal behavioral patterns that complement the fundamental economic relationships identified in this chapter.

Finally, a more applied research direction involves developing comprehensive asset allocation frameworks that build upon the two main advancements presented in this chapter. Such frameworks would apply robust cross-frequency integration for practical portfolio construction, potentially incorporating rebalancing strategies that respond to evolving community structures. Careful consideration of transac-

tion costs, liquidity constraints, and implementation frictions would be essential for translating our findings into implementable investment strategies.

Collectively, the methodological advances and empirical findings presented in this chapter not only address critical gaps in the stock correlation networks literature but also establish a robust foundation for future theoretical innovations and practical applications in financial network analysis and asset allocation frameworks.

Chapter 3

Stochastic Block Models for Large-Scale Financial Network Analysis

3.1 Introduction

The previous chapter's focus was broadening the scope of correlation-based network analysis, by jointly examining multiple views of the temporal dynamics among stocks and showing how different representations of co-movement patterns reveal distinct yet complementary aspects of market structure.

In this chapter, we turn to a complementary perspective and examine how generative models of network structure, specifically the Stochastic Block Model (SBM), can be applied to financial time series. The SBM describes networks as outcomes of probabilistic mechanisms governing interactions between groups of nodes, allowing community structure to arise directly from an underlying generative process. To apply this framework in financial contexts where, as already discussed, the network is not directly observable we adopt a recently developed nonparametric Bayesian

methodology that reconstructs the latent network jointly with its block structure from the observed time series. This approach avoids the need for arbitrary thresholds or fixed community numbers and does not impose external topological constraints, enabling topological structures to emerge in a fully data-driven way.

We contribute to the existing literature by providing the first comprehensive, large-scale study of stochastic block models applied to financial time series. Implementing the generative reconstruction-community detection algorithm across twenty years of daily returns for continuously traded S&P 500 constituents, we track the evolution of market mesoscopic organization through economic cycles and crisis periods. This extensive empirical investigation establishes that generative network approaches can successfully capture meaningful market structure at the scale required for practical financial applications.

The second second contribution of this chapter addresses a fundamental limitation that has constrained network reconstruction research, particularly in the financial time series domain: without access to the true underlying network, assessing reconstruction quality remains inherently problematic. We introduce a novel graph-clustering-based validation framework that circumvents this well-known challenge, enabling systematic evaluation of reconstructed networks through their temporal coherence and alignment with economic fundamentals rather than comparison against unavailable ground truth. The empirical analysis reveals clear temporal signatures of abrupt transitions associated with major economic events such as the 2008 financial crisis and the COVID-19 shock, highlighting the ability of the reconstruction algorithm to capture meaningful changes in market structure.

The remainder of this chapter is organized as follows. Section 3.2 provides a detailed description of the reconstruction-community detection algorithm and its theoretical foundations. Section 3.3 introduces the graph-clustering-based framework for reconstruction evaluation. Section 3.4 presents the comprehensive empirical analysis

supporting our findings, and Section 3.5 concludes.

3.2 Joint reconstruction-community detection

As briefly mentioned, the framework we are using here jointly performs network reconstruction and community detection, leveraging a synergistic effect that allows the detected communities to inform the edge reconstruction and, the edge presence/absence to inform the detection communities through an iterative procedure. The iterative procedure starts from an empty network structure and then alternates between two main steps. First, given the current community structure, it updates the network edges and their weights by evaluating which modifications decrease the total description length. Second, given the current network structure, it updates the community assignments using MCMC moves on the block partition. This iterative procedure continues until the description length converges, with the community structure informing which edges are more plausible (edges within communities have higher prior probability) and the reconstructed edges informing the community detection (nodes with similar connection patterns are grouped together).

Since this approach relies on the stochastic block model (SBM) as its underlying generative mechanism for identifying communities, we begin by introducing this class of models in greater detail and then we will move onto describing in detail the joint reconstruction-community detection procedure.

The Stochastic Block Model (SBM) Holland et al. (1983), briefly mentioned in chapter 1, is a generative probabilistic model for networks that partitions nodes into communities based on their connectivity patterns. In its classical formulation, the SBM assumes that the probability of an edge between any two nodes depends solely on the block memberships of those nodes. Formally, given a network with N nodes divided into K blocks, each node i is assigned to a block $c_i \in \{1, \ldots, K\}$, and the

probability of an edge between nodes i and j is determined by a $K \times K$ matrix of connection probabilities ω_{rs}, where $r = c_i$ and $s = c_j$. The adjacency matrix A is then generated according to:

$$P(A_{ij} = 1 \mid c_i = r, c_j = s) = \omega_{rs} \tag{3.1}$$

This classical formulation of the SBM suffers from a significant limitation since it assumes that all nodes within a block have the same expected degrees. This homogeneity assumption is violated in most real-world networks, which typically exhibit substantial degree heterogeneity even within the same community.

To address this limitation, Karrer and Newman (2011) introduced the Degree-Corrected Stochastic Block Model (DC-SBM), which incorporates node-specific parameters to account for degree heterogeneity within blocks. In the DC-SBM, each node i is assigned both a block membership c_i and a degree-correction parameter θ_i, which modulates the node's propensity to form connections. The edge probability becomes:

$$P(A_{ij} = 1 \mid c_i = r, c_j = s, \theta_i, \theta_j) \propto \theta_i \theta_j \omega_{rs} \tag{3.2}$$

This extension significantly improves the model's ability to detect meaningful community structure in empirical networks, particularly in networks with hub nodes[1]. A number of algorithms have been proposed for the SBM esimation, including expectation–maximization (EM) (Karrer and Newman, 2011) and belief propagation (Decelle et al., 2011). Despite their computational appeal, these methods exhibit well-documented deficiencies: EM and maximum-likelihood approaches systematically overfit by favoring fine-grained partitions (Bickel and Chen, 2009; Peixoto, 2013), while variational approximations tend to underestimate posterior uncertainty

[1] In network science hubs are nodes with exceptionally high degree, meaning that they have a substantially larger number of connections compared to the average node in the network

and yield biased estimates of the number of blocks (Blei et al., 2017). Another important point is that both EM and variational inference suffer from strong sensitivity to initialization and can converge to suboptimal local maxima in the complex likelihood shape (Celisse et al., 2012).

In the context of SBMs, the nonparametric Bayesian framework based on the Minimum Description Length (MDL) principle(Peixoto (2013, 2014)) provides an efficient and principled solution that prevents overfitting and infers the number of communities from the data.

A further challenge in applying SBMs to real networks is the resolution limit problem. Indeed, when using standard Bayesian inference with noninformative priors, the maximum number of detectable groups scales as $K_{\max} = O(\sqrt{N})$, where N is the number of vertices in the network Peixoto (2013). This limitation prevents the detection of small but statistically significant communities in large networks. Peixoto (2014) provided a solution to this issue by introducing the Hierarchical Stochastic Block Model (HSBM), which replaces noninformative priors with a hierarchy of priors and hyperpriors. In the HSBM, the network is not characterized by a single partition into blocks, but rather by a nested hierarchy of partitions at multiple scales. The blocks at level ℓ are themselves partitioned into blocks at level $\ell + 1$, forming a tree-like hierarchical structure.

Formally, let $\mathbf{c}^{(\ell)}$ denote the block assignment at level ℓ, where $\mathbf{c}^{(0)}$ represents the finest partition of the original nodes. The matrix $\mathbf{e}^{(\ell)}$ of edge counts between blocks at level ℓ is modeled as being generated by an SBM at level $\ell + 1$. This recursive construction continues until reaching a top level with a small number of blocks. The joint posterior distribution incorporates all levels simultaneously:

$$P(\{\mathbf{c}^{(\ell)}\}, \{\mathbf{e}^{(\ell)}\} \mid \mathbf{A}) \propto P(\mathbf{A} \mid \mathbf{c}^{(0)}, \mathbf{e}^{(0)}) \prod_{\ell=0}^{L-1} P(\mathbf{e}^{(\ell)} \mid \mathbf{c}^{(\ell+1)}, \mathbf{e}^{(\ell+1)}) P(\mathbf{c}^{(\ell+1)}) \tag{3.3}$$

The hierarchical formulation removes the resolution limit by allowing the detection of communities at multiple scales, from fine-grained local structures to coarse-grained global organization. Importantly, the hierarchy itself is inferred from the data rather than imposed a priori, and the method employs the Minimum Description Length (MDL) principle to prevent overfitting by balancing model complexity against goodness of fit. Moreover Peixoto (2017) developed a microcanonical formulation of the SBM that imposes hard constraints on network generation. Rather than sampling edges independently with specified probabilities, the microcanonical SBM generates networks with exact prescribed degree sequences and block edge counts. This formulation enables more efficient inference algorithms and facilitates the construction of deeper Bayesian hierarchies. It follows a nonparametric Bayesian framework based on the posterior distribution:

$$P(\mathbf{c}, \mathbf{e}, \mathbf{k} \mid \mathbf{A}) = \frac{P(\mathbf{A} \mid \mathbf{c}, \mathbf{e}, \mathbf{k}) P(\mathbf{e}, \mathbf{k} \mid \mathbf{c}) P(\mathbf{c})}{P(\mathbf{A})} \tag{3.4}$$

where $\mathbf{k}$ denotes the degree sequence. Under the MDL principle, the optimal model parameters correspond to those that minimize the description length $\mathcal{D} = -\ln P(\mathbf{c}, \mathbf{e}, \mathbf{k} \mid \mathbf{A})$, which balances the complexity of describing the model against the complexity of describing the data given the model. Although the SBM in all its variants is designed to operate on a fully observed network, in many empirical applications the underlying structure is not directly accessible. Instead one typically observes indirect measurements, such as time series of node-level variables whose dynamics are shaped by an unobserved interaction network. In such settings the latent network W must be inferred from the available data rather than taken as given. This task is referred to as network reconstruction in the network science literature. Formally, consider a data matrix $\mathbf{X}$ of dimensions $N \times T$, where N is the number of nodes and T is the number of observations for each node. In the formulation we

are adopting the likelihood assumes independent and identically distributed observations:

$$P(\mathbf{X} \mid \mathbf{W}, \boldsymbol{\theta}) = \prod_{t=1}^{T} P(\mathbf{X}_t \mid \mathbf{W}, \boldsymbol{\theta}) \tag{3.5}$$

where $\mathbf{W}$ represents the weighted adjacency matrix and $\boldsymbol{\theta}$ denotes additional node-specific parameters. The reconstruction problem is formulated as a maximum a posteriori (MAP) estimation:

$$(\hat{\mathbf{W}}, \hat{\boldsymbol{\theta}}) = \max_{\mathbf{W},\boldsymbol{\theta}} P(\mathbf{W}, \boldsymbol{\theta} \mid \mathbf{X}) = \max_{\mathbf{W},\boldsymbol{\theta}} P(\mathbf{X} \mid \mathbf{W}, \boldsymbol{\theta}) P(\mathbf{W}, \boldsymbol{\theta}) \tag{3.6}$$

The prior $P(\mathbf{W}, \boldsymbol{\theta})$ implements regularization to prevent overfitting. Here we use a nonparametric regularization scheme based in the Minimum Description Length (MDL) principle introduced in Peixoto (2025a). The key innovation is to incorporate a hierarchical SBM as part of the prior for the latent network structure, combined with a weight quantization scheme that discretizes edge weights into categories.

The weight quantization works by partitioning the range of observed weights into Q discrete categories. Each edge weight W_{ij} is assigned to the category whose representative value best explains the observed data, with the number of categories Q and their representative values determined automatically during inference by minimizing the description length. This discretization serves dual purposes since it reduces the model complexity by limiting the number of distinct weight values, and it also introduces a natural sparsity-promoting mechanism since edges with weight zero form a natural category. The total description length combines three components:

$$\mathcal{D} = -\ln P(\mathbf{X} \mid \mathbf{W}, \boldsymbol{\theta}) - \ln P(\mathbf{W} \mid \mathbf{c}, \mathbf{q}) - \ln P(\mathbf{c}, \mathbf{q}) \tag{3.7}$$

where $\mathbf{q}$ represents the weight quantization, $\mathbf{c}$ denotes the block partition, and the second term represents the likelihood of the weighted network under the SBM

prior. The third term accounts for the prior probability of the block structure and quantization scheme. This formulation does not assume a fixed number of edges or a predetermined sparsity level and the optimal network density emerges naturally from minimizing the description length, which automatically balances goodness of fit against model complexityPeixoto (2025b).

More specifically, for stock-return time series we employ a multivariate normal model, starting from the assumption that the log-prices $\log(P_t)$ follow a random walk Granger and Cootner (1965), Palomar (2025). This is one of the simplest return models, dating back to the early foundations of price theory, and it allows returns to be represented in a particularly simple form. Under this hypothesis, the log-returns

$$r_t = \log(P_t/P_{t-1}),$$

are IID conditional on past information. Although this is a crude simple model, empirical evidence suggest that if take not too large time windows returns are rather well represented by this simple model. A standard short-time return model is therefore

$$r_t = \mu + \varepsilon_t, \qquad \varepsilon_t \sim \text{IID}(0, \Sigma),$$

so that $\mathbb{E}[r_t] = \mu$ and $\text{Var}[r_t] = \Sigma$. Formally, let R be a $T \times N$ data matrix of log-returns, where each row corresponds to an observation time t and each column corresponds to a stock. Denote by

$$\mathbf{r}_t = (r_{t1}, \ldots, r_{tN})^T,$$

the N-dimensional vector of returns at time t. Assuming the normality of the stochastic fluctuation around the mean, that is ε_t, it follows that

$$\mathbf{r}_t \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ is the vector of mean returns and $\boldsymbol{\Sigma}$ is the covariance matrix shared across all time steps. The previous normality assumption is rather strong, given the well-known heavier-than-normal tail behavior of stock returns. However, this assumption allows us to keep the analysis and computations tractable. In future work, it will be interesting to investigate the model under more general elliptically symmetric distributions for the IID shocks. The precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ encodes conditional independence relationships. Thus, nonzero entries $\Omega_{ij} \neq 0$ correspond to edges in the underlying network, representing direct interactions between variables i and j. The network reconstruction problem then consists of determining which entries of $\boldsymbol{\Omega}$ are nonzero and estimating their magnitudes.

Given T observations $R = (\mathbf{r}_1, \ldots, \mathbf{r}_T)^T$, the likelihood under the multivariate normal model is

$$P(R \mid \boldsymbol{\Omega}, \boldsymbol{\mu}) = \prod_{t=1}^{T} (2\pi)^{-N/2} |\boldsymbol{\Omega}|^{1/2} \exp\left(-\frac{1}{2} (\mathbf{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Omega} (\mathbf{r}_t - \boldsymbol{\mu}) \right).$$

The reconstruction approach parameterizes $\boldsymbol{\Omega}$ in terms of a weighted adjacency matrix $\mathbf{W}$ and node parameters $\boldsymbol{\theta}$, introducing sparsity through the hierarchical SBM prior. Specifically, the precision matrix $\boldsymbol{\Omega}$ is constructed from the weighted adjacency matrix $\mathbf{W}$, where the sparsity is determined by the inferred network structure, with $\Omega_{ij} \neq 0$ only when $W_{ij} \neq 0$

The procedure alternates between updating the network structure and updating the node parameters, with each modification evaluated according to its impact on the total description length. This results in a reconstructed network that represents the most parsimonious explanation of the observed dependencies in the data, with the number and strength of edges automatically determined by the MDL principle rather than requiring manual tuning of regularization parameters.

For our specific setting, each node represents a stock, and each observation corre-

sponds to the log-return at a specific time point. The reconstructed network reveals dependencies between asset returns that persist after controlling for indirect effects mediated through other assets. Additionally, the hierarchical SBM uncovers community structure among the assets, identifying groups of stocks with similar patterns of co-movement.

3.3 Graph Clustering and Network Reconstruction Evaluation

The reconstruction framework described in the previous section provides a principled approach to inferring both the latent topology of the financial network and its community organization directly from the data. However, assessing the quality and consistency of the reconstructed networks remains a non-trivial task, as no true reference network structure exists for such financial networks. Unlike controlled simulations, where the underlying network is known, empirical financial data only provide indirect observations of asset interdependencies through co-movement patterns over time. This absence of a ground truth calls for alternative strategies capable of evaluating reconstruction quality in a meaningful and data-driven manner.

Traditionally, financial network studies rely on meso-scale validation, comparing the detected communities with economic sector classifications to assess the economic interpretability of the inferred modules. We already expressed our views on this evaluation method that, while confirming that reconstructed groupings align with known fundamentals, provides only a partial view of the problem. Two networks may exhibit similar community partitions yet differ substantially in their edge configurations, density, or connectivity profiles. To complement this limited perspective, we propose a graph-clustering-based evaluation framework, similar in spirit to the

meta clustering-based framework of the previous chapter, that operates directly on the reconstructed networks enabling the assessment of structural consistency and temporal organization.

The central idea of this approach is that a well-reconstructed network should display structural continuity during periods of market stability and undergo identifiable reconfigurations when the market transitions between regimes.

To this end we analyze the temporal sequence of networks reconstructed via the Bayesian procedure and quantify their pairwise structural similarity. Following Felippe et al. (2024), we employ three complementary similarity measures: i) the basic network normalized mutual information (NMI), capturing overall structural resemblance; ii) the degree-corrected NMI, which accounts for heterogeneity in node degree distributions and iii) the meso-scale NMI, which assesses similarity based on group-level organization, using in this case economic sectors as a reference.

We will briefly describe the intuition behind these measures starting from the basic NMI[2], moving then to the DC-NMI and concluding with the meso-scale NMI. Let G_1 and G_2 be two undirected, unweighted graphs with the same set of vertices V, with edge sets E_1 and E_2, and let $M = \binom{N}{2}$ denote the number of possible edges. The first measure is a micro-scale NMI obtained by representing each graph as a binary vector of length M, where each component encodes the presence or absence of an edge. The marginal edge probabilities are defined as:

$$p_1 = \frac{|E_1|}{M}, \qquad p_2 = \frac{|E_2|}{M},$$

and $E_{12} = |E_1 \cap E_2|$ is the number of common edges, with joint probability $p_{12} = E_{12}/M$. The two graphs therefore induce a 2×2 joint distribution (contingency

[2]here NMI stands for Network Mutual Information. This should not be confused with the Normalized Mutual Information used in clustering quality evaluation, also commonly referred to as NMI.

table):

$$P_{12}(1,1) = p_{12}, \quad P_{12}(1,0) = p_1 - p_{12}, \quad P_{12}(0,1) = p_2 - p_{12}, \quad P_{12}(0,0) = 1 - p_1 - p_2 + p_{12}.$$

Writing the binary entropy [3] as:

$$H_b(p) = -p \log p - (1-p) \log(1-p)$$

and the Shannon entropy as:

$$H_s(P_{12}) = - \sum_{x,y \in \{0,1\}} P_{12}(x,y) \log P_{12}(x,y)$$

the mutual information between the two edge encodings is given by:

$$I(G_1, G_2) = H_b(p_1) + H_b(p_2) - H_s(P_{12}).$$

Then, in order to obtain a bounded similarity measure in $[0,1]$, the authors define the normalized NMI as

$$\mathrm{NMI}(G_1, G_2) = \frac{2I(G_1, G_2)}{H_b(p_1) + H_b(p_2)}$$

Therefore $\mathrm{NMI}(G_1, G_2)$ captures the extent to which the presence or absence of edges in one graph reduces the uncertainty about edge occurrences in the other.

However, the simple NMI does not take degree heterogeneity into account, since two graphs may share nearly identical degree sequences yet differ substantially in how edges are distributed across nodes. Therefore, to capture similarity from a local neighborhoods perspective, the authors introduce a degree-corrected version of the

[3] $H_b(p)$ denotes the binary entropy of a Bernoulli random variable with success probability p, defined as $H_b(p) = -p \log_2 p - (1-p) \log_2(1-p)$. We use base 2 logarithms so that all quantities are expressed in bits.

NMI. For each node i, let $k_1(i)$ and $k_2(i)$ denote its degrees in G_1 and G_2, and let $k_{12}(i)$ be the number of neighbors shared by i in the two graphs. The normalized probabilities are defined as:

$$p_1(i) = \frac{k_1(i)}{N-1}, \qquad p_2(i) = \frac{k_2(i)}{N-1}, \qquad p_{12}(i) = \frac{k_{12}(i)}{N-1}$$

and the corresponding joint distributions $P_{12}(i)$ defined analogously to the simple NMI. The node-wise mutual information is:

$$I_i = H_b\big(p_1(i)\big) + H_b\big(p_2(i)\big) - H_s\big(P_{12}(i)\big)$$

then, averaging over all nodes yields:

$$I_{\deg}(G_1, G_2) = \frac{1}{N}\sum_{i=1}^{N} I_i$$

Therefore the degree-corrected NMI is defined as:

$$\text{DC-NMI}(G_1, G_2) = \frac{2I_{\deg}(G_1, G_2)}{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\big[H_b(p_1(i)) + H_b(p_2(i))\big]}.$$

This measure quantifies similarity in terms of local neighborhoods rather than individual edges, thereby taking into account degree variation across nodes.

Lastly, the third measure addresses mesoscale similarity by comparing coarse-grained representations obtained from a partition $c = \{c_1, \ldots, c_N\}$ of the nodes into K groups. For a graph G_ℓ ($\ell = 1, 2$), let $m_\ell(r, s)$ be the number of edges between groups r and s. The coarse-grained representation $\tilde{G}_\ell^{(c)}$ is a multiset where each element (r, s) with $r \leq s$ appears with multiplicity $m_\ell(r, s)$. The overlap between

the two coarse-grained representations is quantified by:

$$E_{12}^{(c)} = \sum_{r=1}^{K} \sum_{s=r}^{K} \min\{m_1(r,s), m_2(r,s)\}.$$

The entropy of each coarse-grained representation is:

$$H(\tilde{G}_\ell^{(c)}) = \log \left(\left(\binom{K}{2} + K \right) \right)$$

where $\left(\!\binom{n}{k}\!\right)$ denotes the multiset coefficient. The joint entropy is bounded by:

$$H(c) \leq \log \left(\left(\binom{K}{2} + K \right) \right) + \log \left(\left(\binom{K}{2} + K \right) \right) + \log \left(\left(\binom{K}{2} + K \right) \right).$$

The mesoscale mutual information is then:

$$I_{\text{meso}}^{(c)}(G_1, G_2) = H(\tilde{G}_1^{(c)}) + H(\tilde{G}_2^{(c)}) - H(c),$$

and the corresponding normalized measure is:

$$\text{MesoNMI}^{(c)}(G_1, G_2) = \frac{2I_{\text{meso}}^{(c)}(G_1, G_2)}{H(\tilde{G}_1^{(c)}) + H(\tilde{G}_2^{(c)})}.$$

This quantity captures similarity in mixing patterns between the groups defined by the chosen partition c, thereby quantifying resemblance at a specific structural resolution. In this work we use the GICS-11 economic sectors as the defined partition.

Taken together, these three measures cover the similarity between graphs from three complementary perspectives, providing a complete overview of the similarity between graphs. The method proceeds by converting each similarity value $s_{i,j}$ between networks i and j to a dissimilarity measure as $d_{i,j} = 1 - s_{i,j}$, producing a dissimilarity matrix that summarizes structural dissimilarities over time. To uncover temporal patterns in these evolving network structures, we apply Ward's hierarchical

clustering to the dissimilarity matrix. This procedure groups networks with similar topological properties into clusters. The resulting clusters reveal periods of structural persistence, when network topology remains relatively stable, as well as abrupt transitions. In this way, the graph-clustering-based evaluation provides a dynamic, network-based complement to sector-based validation. The framework thus serves a dual purpose: it acts as both a validation tool and an interpretative lens, offering a complete overview of how network structure reflects the changing market dynamics.

3.4 Empirical analysis

Having outlined the methodological foundations of the Bayesian network reconstruction and the graph-clustering-based evaluation framework, we now turn to their empirical implementation on financial data. The goal of this section is to assess how effectively the proposed approach captures the evolving structure of financial markets and to examine the interpretability of the resulting networks in light of economic fundamentals.

In practice, the Bayesian reconstruction framework is applied to time series of asset returns, producing both a full-sample reconstruction and a sequence of time-resolved networks. The full-sample analysis serves as an initial exploration of the global dependency structure, allowing us to characterize the overall topology of the financial market, its community organization, and the sectoral composition of the detected groups. This static perspective also provides a benchmark against which the subsequent dynamic analysis can be interpreted.

We then extend the reconstruction to a rolling-window setting, tracing the evolution of network topology over time. These temporally resolved networks are analyzed using the graph-clustering-based evaluation framework introduced in the previous section, which enables the identification of persistent configurations and the detec-

tion of regime shifts in market organization. By combining rolling window reconstruction and graph-clustering-based evaluation, we obtain a comprehensive understanding of how the underlying dependency network evolves across different market conditions. In doing so, we demonstrate that the proposed framework yields robust, interpretable, and economically meaningful representations of the financial system's structural and temporal complexity.

Our analysis uses daily adjusted closing prices and GICS industrial sector classifications for 383 stocks from the S&P 500 index constituents, which serves as a widely accepted proxy for the whole U.S. stock market. The price data were obtained from Yahoo Finance, covering the 20-year period from January 1, 2003, to December 31, 2023, while the GICS sector classifications were sourced from Wikipedia. Since dynamic analysis tracks the stocks over time, only the stocks that were continuously traded in the period under analysis were selected. This selection leads to a strong survivor bias, but is essential for our analysis. Table 3.1 shows a comparison between the sectoral composition of the complete S&P 500 index (composition at the 31 December 2023) and our dataset. As shown, our dataset maintains a generally proportional representation across all sectors, with some minor differences. Communication Services show the most significant under-representation (2.61% in our dataset versus 4.37% in the full index), while Utilities exhibits slight over-representation (7.05% versus 5.96%). Importantly, the Financial sector, which plays a central role in market structure analysis during crisis periods, maintains nearly identical representation in our dataset (14.62%) compared to the full index (14.12%). These differences in sectoral composition, while generally minor, should be considered when interpreting sector-specific analyses. However, the overall sectoral balance of our dataset suggests that it remains broadly representative of the S&P 500 index, making it suitable for studying both market-wide structural patterns and sector-level dynamics. The joint reconstruction-community detection framework is applied directly to the log-returns

	S&P 500		Dataset	
Sectors	N	perc.	N	perc.
Communication Services	22	4.37%	10	2.61%
Real Estate	31	6.16%	26	6.79%
Utilities	30	5.96%	27	7.05%
Consumer Discretionary	52	10.34%	37	9.66%
Consumer Staples	38	7.55%	31	8.09%
Energy	23	4.57%	17	4.44%
Financials	71	14.12%	56	14.62%
Health Care	64	12.72%	53	13.84%
Industrials	79	15.71%	57	14.88%
Information Technology	65	12.92%	48	12.53%
Materials	28	5.57%	21	5.48%

Table 3.1: S&P 500 composition vs dataset composition.

time series. As already anticipated we first applied the framework to the full 20 years sample and then on different time windows following a rolling window scheme with 1000 days windows and a 30 days shift between windows. The following subsections will discuss the results in detail for both the static and the dynamic case. While the reconstruction procedure returns a weighted network to leverage information about the strength of dependencies, all analyses presented below consider the binarized network structure. Specifically, we retain only edges with nonzero weights from the reconstruction, treating the network as unweighted for community detection and topological analysis.

3.4.1 Static

As an initial investigation, we applied our methodology to the entire 20-year period to establish a base understanding of the market's overall structure. Figure 3.1 displays the reconstructed network showing nodes colored according to their sector classification (graph on the right) and positioned based on the communities identified by our algorithm.

Several notable patterns emerge from both the reconstructed network and his community structure. First there is a strong correspondence between detected communities and GICS sector classifications, indicating that our algorithm successfully captures fundamental economic relationships without requiring predetermined sector information. This alignment is particularly striking given that our methodology is entirely data-driven and makes no assumptions about the number or composition of communities. Second we observe several mixed communities that span multiple sectors, a first naive interpretation would lead to the conclusion that cross-sectoral communities represent algorithmic limitations but a closer examination of the composition of these communities leads to the opposite conclusion since these cross-sectoral groupings often reflect economic realities that transcend strict classification boundaries. As an example in figure 3.1 bottom-left we see that companies like Amazon and eBay, classified as Consumer Discretionary stocks under GICS, are grouped with Information Technology stocks by our algorithm: a grouping that arguably better reflects their business models and risk characteristics as technology-driven platforms. Third, some sectors are split into multiple communities, suggesting the presence of sub-sectoral dynamics that the broad GICS classification may not fully capture, this aligns with the hierarchical nature of the GICS framework which includes more granular industry classifications beyond the sector level used in our comparative analysis. Both bottom graphs in figure 3.1 clearly show the ability of the algorithm to detect communities at higher granularity than GICS sectors. The bottom-left graph shows two communities for the IT sector, by looking at the single companies we can see that the two comminutes roughly correspond to, respectively, IT Services and IT hardware companies while, in the bottom-right graph, we can clearly see that the Financial sector is split into three communities corresponding approximately to Insurance, Banks and Financial Services companies.

These observations demonstrate that our community detection approach not only

recovers known economic sectors but also identifies more nuanced groupings that may better reflect functional similarities between companies, exploiting both established and emergent structures and improving the understanding of the evolving nature of market relationships. A well established idea in the literature is that the more overlap there is between communities and economic sectors, the better the community detection method is in retaining economic information. Despite the limitations of this idea, directly coming from to the limitations of the strict industry sector classification discussed above, it is still important to measure this overlap and use it as a first indicator of performances for community detection algorithms.

Here we measure the overlap with economic sectors by computing the partition overlap Peixoto (2021) between community labels and GICS economic sectors. In figure 3.2 we show the partition overlap between the partition obtained through different clustering methodologies based on the dissimilarities computed from the correlations (except for the SBM) and, in the last column, the overlap with the economic sector classification for both 11 clusters (same as economic sectors) and 15 (same as the SBM). Our algorithm shows a significant overlap with both DBHT and CL hierarchical clustering and, more importantly, the hightest overlap with economic sectors in both cases.

While the static analysis provides useful insights into the overall market structure, financial systems are inherently dynamic, continuously evolving in response to new information, changing economic conditions, and shifting market sentiment. To capture this temporal dimension in the next section we implement a dynamic analysis framework.

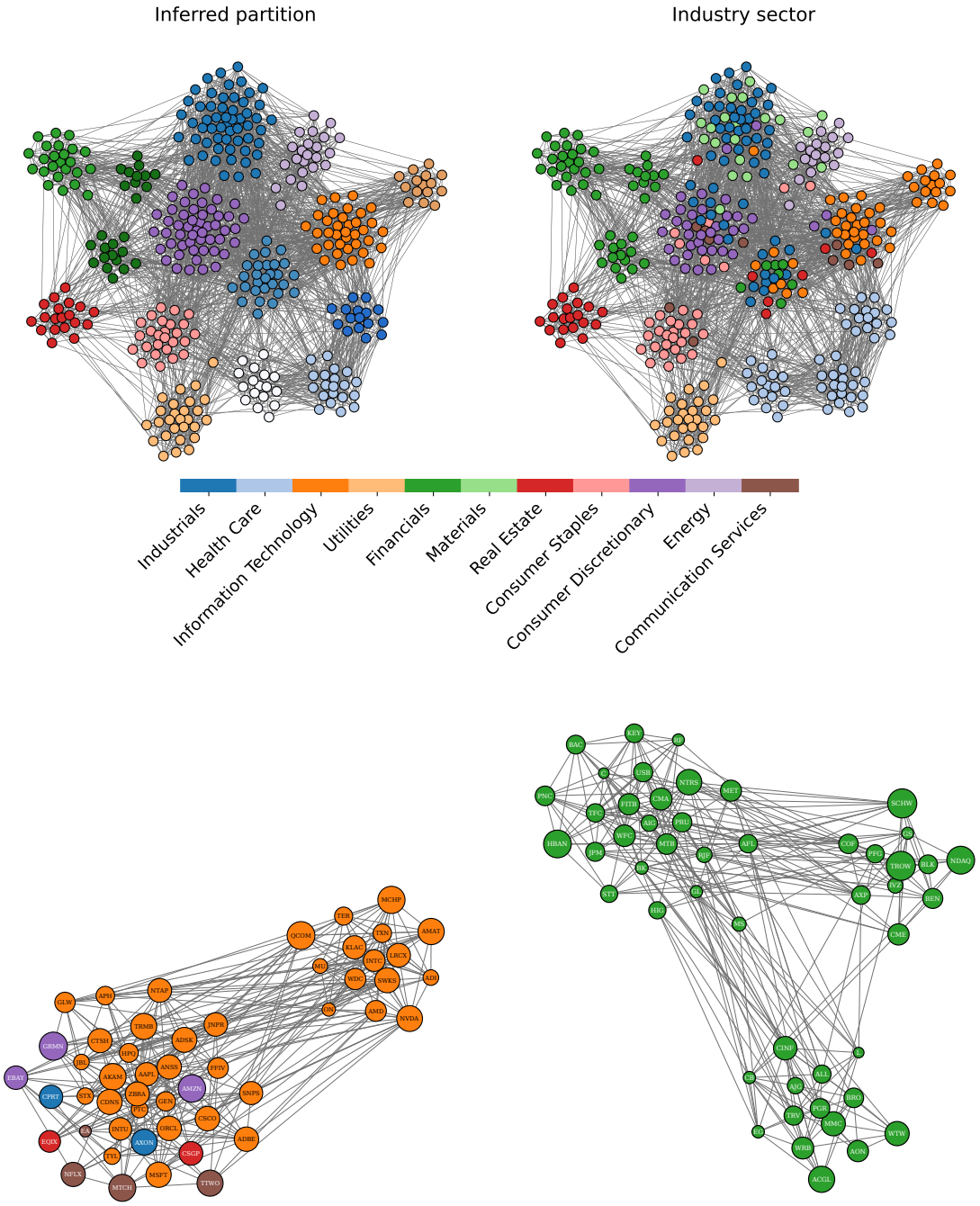


Figure 3.1: top: Network reconstructed from the full 20 years sample. Nodes are positioned according to the communities they belong to. Node color represents the communities found by the SBM (left) and the economic sectors of activity (right); bottom-left: sub graph of two communities of stocks belonging mostly to the IT sector; bottom-right: sub-graph of communities of stocks belonging mostly to the IT sector.

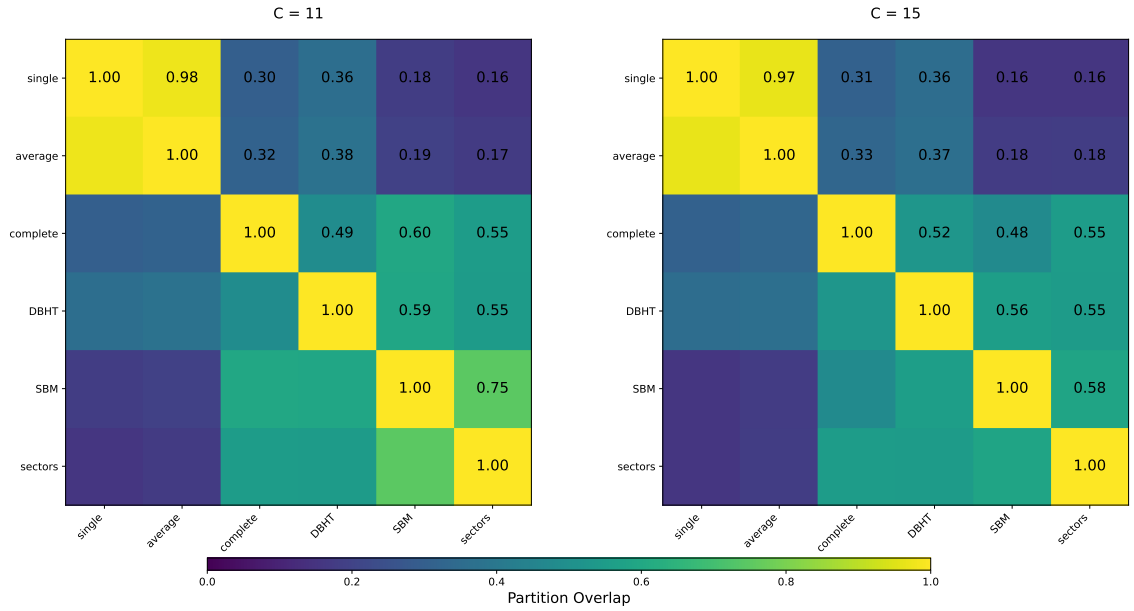


Figure 3.2: Partition overlap between the community labels of different algorithms for 11 clusters (left) and 15 clusters (right). The last column shows the overlap with industrial sectors.

3.4.2 Dynamic

To study the dynamic evolution of the market we divide the 20-year sample into partially overlapping 1000-day moving windows, with a 30-day shift between consecutive windows and, for each time window, we perform network reconstruction and community detection using our nonparametric Bayesian approach. We then analyze the evolution of network properties and community structures over time, assessing the algorithm's ability to capture economically meaningful information across different market regimes.

Community detection is an important task in network science in general, helping us understand the macroscopic structure of a system and, possibly, reduce its dimensionality to better analyze it. In the context of stock returns time series networks this task is perhaps even more important since it constitutes one of the building blocks of many network-based risk management and asset allocation frameworks. To meaningfully measure the community detection performances we compare it to

the Louvain algorithm based on a redefinition of modularity adapted to the unique features of stock returns time series correlation matrices in MacMahon and Garlaschelli (2013). We chose to use the adapted Louvain algorithm because it presents many points in common with our algorithm since it is hierarchical and does not require the number of communities as input.

The upper panel of figure 3.3 shows the partition overlap between communities found by both the adapted Louvain and our algorithm and the economic sectors. We can clearly see that the DC-SBM shows superior performances in all time windows, especially after the COVID-19 pandemic, and is generally much better on average than the Louvain algorithm. The bottom panel of figure 3.3 then shows the number of communities found in each window by the two algorithms. We can clearly see that the DC-SBM is capable of detecting much more structure with communities ranging between 5 and 12 while the Louvain algorithm is tied up on a range between 3 and 8. This is a further confirmation that our method is capable of extracting meaningful information and significantly outperforms modularity-based methods.

The detected communities reveal an interesting structure, but the method also produces a reconstructed network for each time window, which deserves as much attention—if not more—than the community structures themselves. Therefore we perform a deep analysis of these networks with particular regard to their structural evolution over time. The reconstructed networks are unweighted and undirected.

Figure 3.5 presents the Spearman correlation between a series of node-level statistics across the time-windowed networks. Notably, all four measures exhibit consistent temporal patterns, revealing three well-defined structural blocks and a less distinct block spanning from 2012 to 2020. These blocks largely align with major economic phases: the first corresponds to the period preceding the 2008 subprime crisis, the second encompasses the time frame from the subprime crisis through the European debt crisis, the third reflects a relatively stable economic phase and the final block

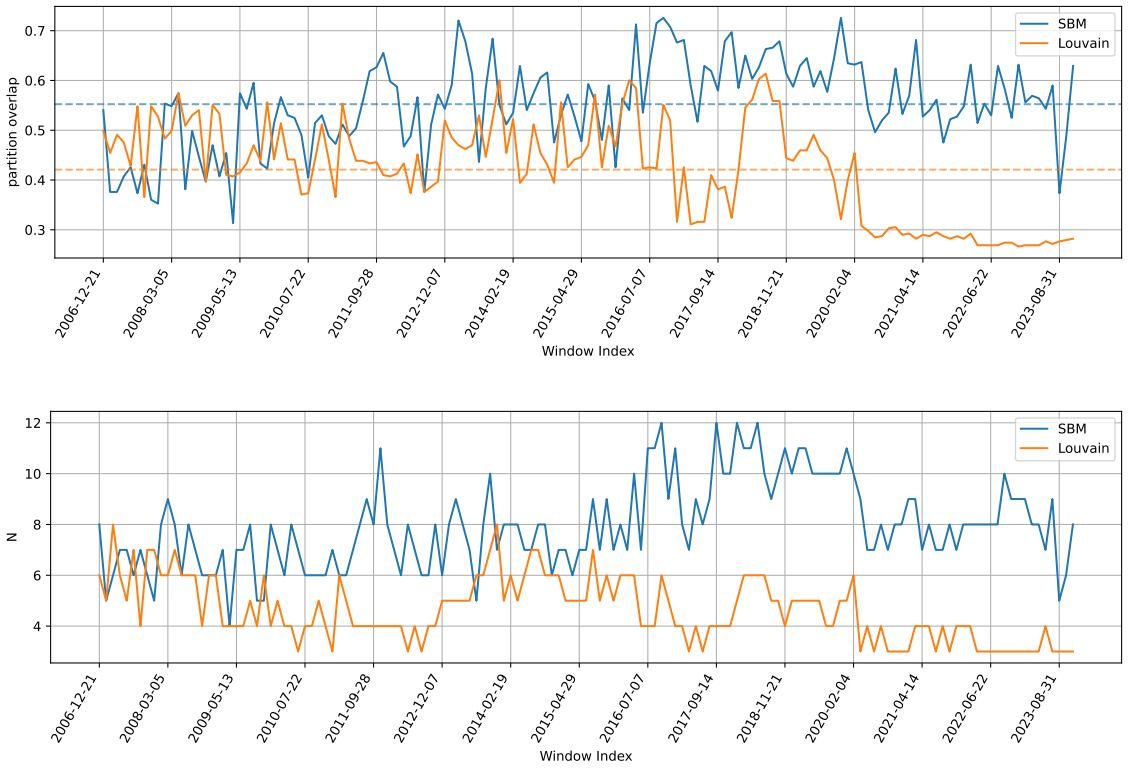


Figure 3.3: top: partition overlap between partitions found by the algorithms and economic sectors of activity, vertical dashed lines represent the temporal averages; bottom: number of communities found by the algorithms.

clearly coincides with the aftermath of the COVID-19 pandemic.
These results suggest that the reconstructed networks carry meaningful economic information in their structure, reflecting changes in market conditions over time, giving us more confidence in the method and showing that it can be a useful way to explore how financial networks evolve in correspondence of different economic periods. Another interesting thing to study is the persistence of edges over time and, to do this, we aggregated the unweighted networks over all time windows by summing up the adjacency matrices for each window and then dividing each entry of the final aggregated adjacency matrix by the total number of windows. Therefore each entry of the final graph represents the share of windows in which the edge appears in the time-window networks with respect to the total number of windows. Figure 3.4 shows the aggregated network with the most persistent edges at different threshold levels with nodes colored and positioned according to the economic sectors. We can notice some interesting patterns such as the cohesiveness of sectors like Utilities and Information Technology and most notably the high interconnectedness of the financial sector with other important sectors such as Industrials and Information Technology.

The block structure observed in the node level statistics correlation inspired us to further investigate the evolution of the reconstructed networks over time on a global scale. Therefore we applied the graph-clustering procedure introduced in the previous section to the graphs reconstructed from the different windows. Clustering the graphs reconstructed from the different windows can be viewed as a different way of clustering the time windows themselves and therefore this clustering structure can be analyzed and used as a further validation of the reconstruction algorithm, as already explained in the previous section.
Figure 3.6 shows, in the upper panel, the S&P 500 index time series colored according

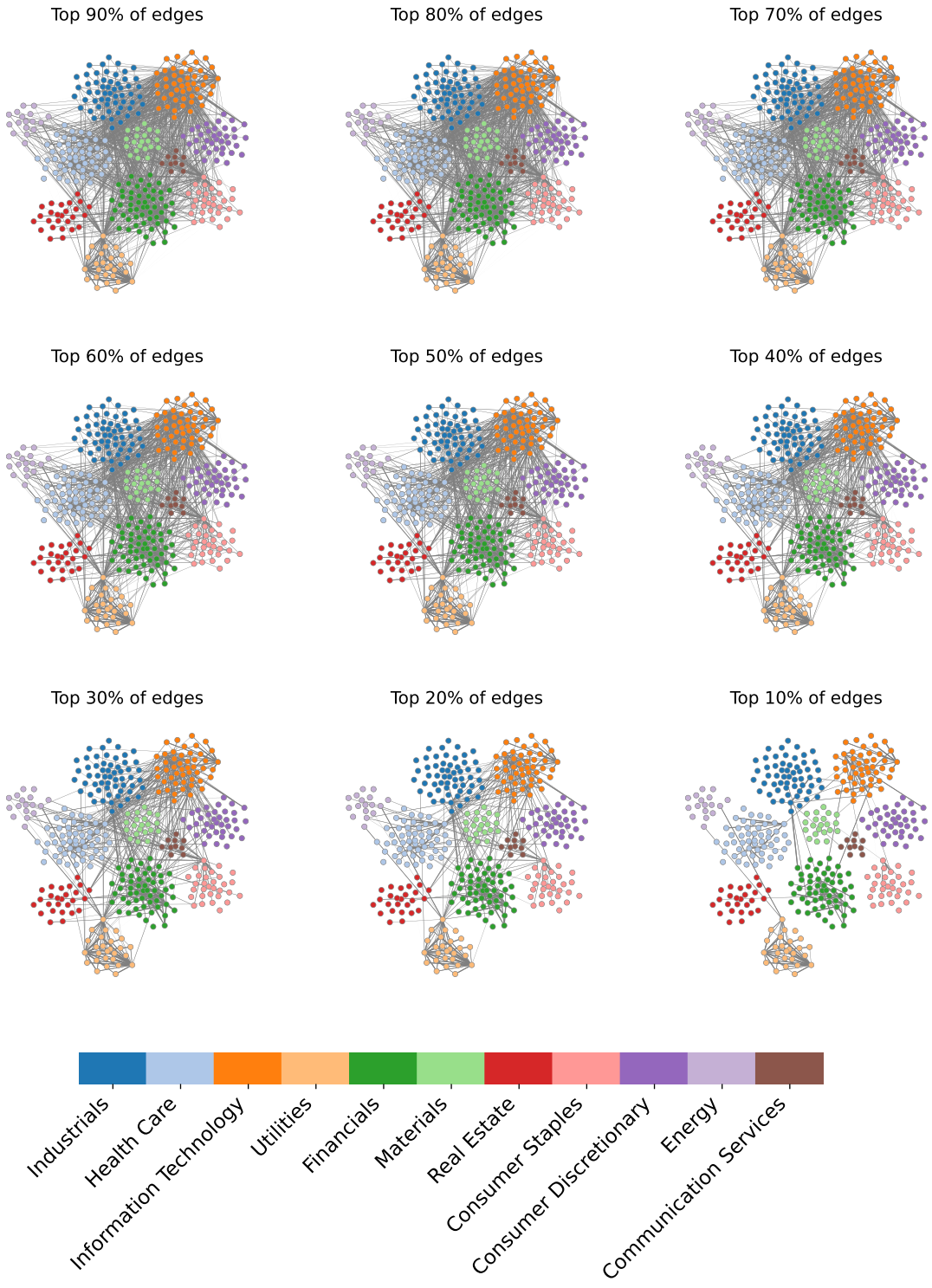


Figure 3.4: Threshold graph at different levels of the graph resulting from aggregating the (unweighted) graphs from all the time windows and applying different thresholds. The goal is to show the most persistent edges over time at the various levels.

to the cluster labels of the underlying reconstructed networks from the DC-NMI similarity measure, there is a striking overlap between particularly relevant economic periods and clusters suggesting that the reconstructed networks embed the changing market conditions in their structure.

In order to better analyze these structures we computed the aggregated networks for each temporal cluster in the same way as the full period aggregated network of figure 3.4. The results are displayed in the bottom panel of figure 3.6 and show some interesting features of the graphs, most notably the strong connection between financial and real estate sectors in the aftermath of the subprime financial crisis (cluster 2) which disappears in other clusters. This is a manifestation of the central role played by these two sectors in the build-up and subsequent explosion of the subprime mortgage bubble. Another interesting aspect is the strong connectedness between the Utilities and Consumer Staples sectors during the COVID-19 pandemic period, likely reflecting investor behavior in times of market stress. These sectors, traditionally considered defensive, were targeted by investors seeking stable and 'safe' investments amid heightened uncertainty and volatility. Such co-movement is typical during systemic disruptions, when capital flows into low-risk assets.

3.5 Concluding remarks

This chapter has demonstrated the effectiveness of the joint reconstruction-community detection method in analyzing financial market dynamics from both static and dynamic perspectives. Through the application to twenty years of S&P 500 stocks, we have shown how market structures evolve over time and respond to major economic events, revealing both sectoral patterns that align with traditional classifications and cross-sectoral relationships that may be overlooked by conventional industry groupings.

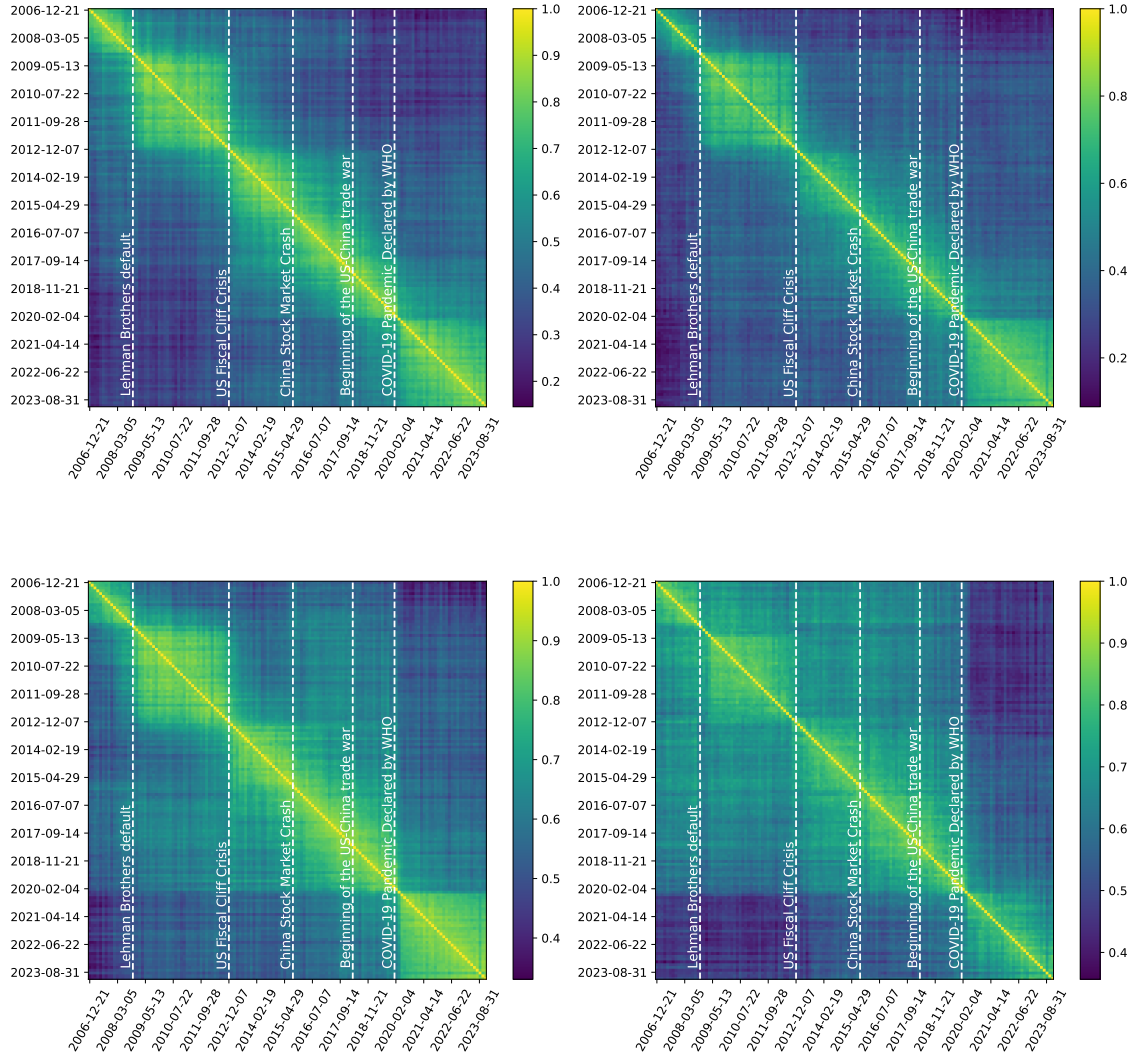


Figure 3.5: Spearman correlation between network statistics computed on graphs from different time windows: Degree (upper-left), Betweenness centrality (upper-right), Closeness centrality (lower-left), local clustering coefficient (lower-right)

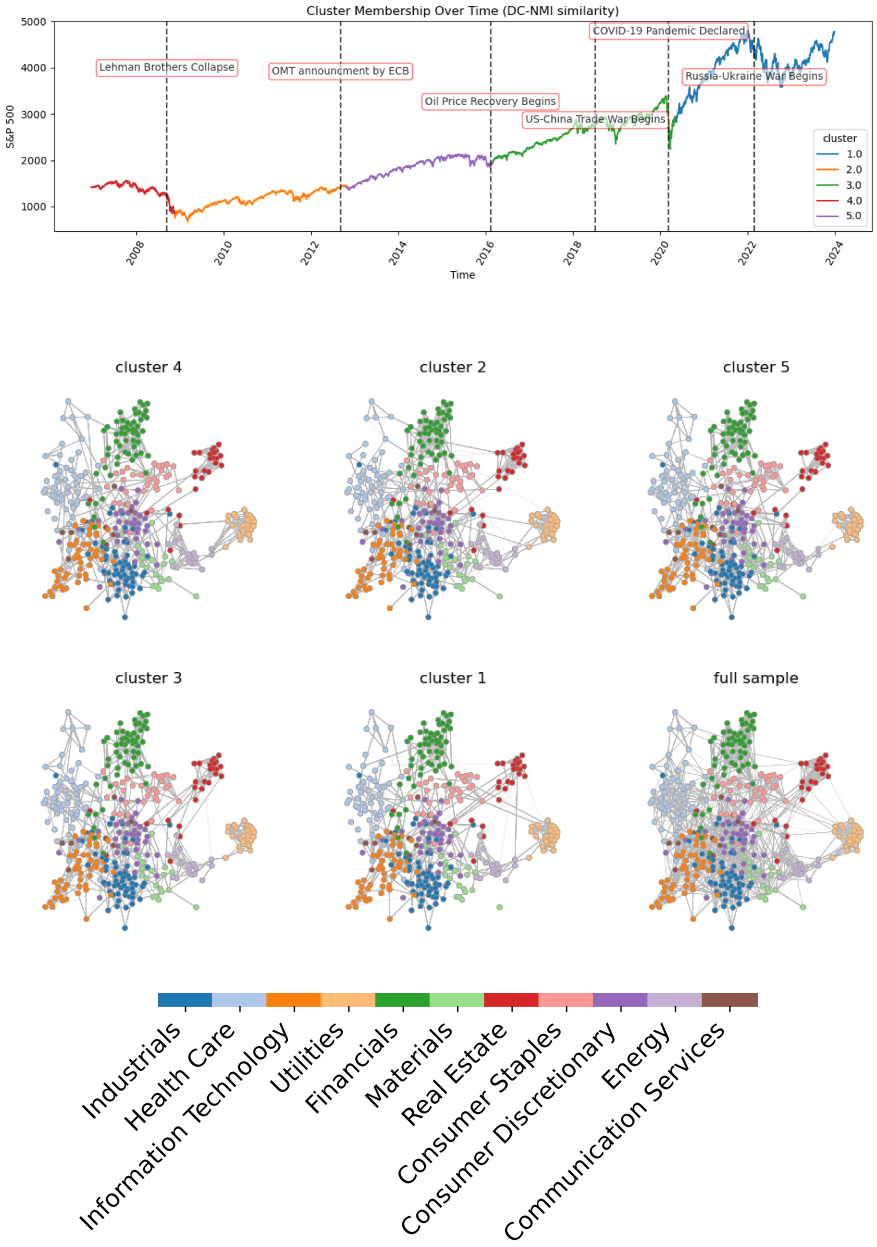


Figure 3.6: a) Value of the S&P 500 index over time. Colors represent the clusters obtained from the DC-NMI similarity. b) networks reconstructed by aggregating the reconstructed networks within each cluster and retaining the 10% most significant edges.

The novel graph clustering based evaluation methodology introduced in this work represents a significant contribution, providing a market-driven validation approach that complements traditional metrics. The strong correspondence between identified regime transitions and known economic events confirms that our reconstruction framework effectively captures meaningful economic information in network structure.

Several promising directions emerge for future research. The most straightforward extension would involve incorporating fat-tailed distributions for return modeling, which could improve reconstruction quality, particularly at higher sampling frequencies. Another particularly valuable direction would be investigating whether changes in network structure can provide early warning signals of market regime transitions or emerging sector-specific vulnerabilities.

In conclusion, this work demonstrates that generative modeling approaches offer valuable insight into the complex evolving structure of financial markets, with the graph clustering methodology providing a natural evaluation framework that aligns with the dynamic nature of financial systems.

Conclusions

This thesis provides several advances to the analysis of financial markets through the lens of network science, spanning both methodological and empirical contributions to the understanding of how financial markets evolve and respond to shocks. Chapter 1 provides a concise overview of the main theoretical and methodological foundations of financial network analysis, covering the essential concepts and established approaches in the literature. As such, it could serve as a general reference for researchers approaching this field. Chapter 2 introduces a novel, robust framework for stock correlation network analysis that combines robust estimation with multi-view data integration, revealing an intrinsic capacity for self-organization in equity markets that becomes visible only when these two elements are jointly employed. Finally, Chapter 3 explores financial markets through a generative modeling approach, demonstrating, through twenty years of S&P 500 data and a novel graph-clustering evaluation methodology, that network topology encodes meaningful economic information and faithfully tracks major market regime transitions. The results presented across the thesis reinforce the value of network science as a rigorous and versatile framework for the empirical and theoretical study of financial markets. Notwithstanding these contributions, some limitations should be acknowledged. The empirical scope is concentrated on large-cap U.S. equities, and how well the findings generalize to less liquid markets or other asset classes remains an open question. Several directions appear promising for future research. Integrating firm-level sentiment data into the community detection framework introduced in chapter 2 could bridge behavioral and structural perspectives on market organization. Developing dynamic asset allocation strategies that exploit evolving community structure represents a practically relevant extension. Finally, testing whether network topology can serve as a leading indicator of systemic stress or sector-specific fragility would

considerably broaden the practical relevance of the generative modeling approach introduced in Chapter 3.

Bibliography

Acemoglu, D., Ozdaglar, A., and Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608.

Aldasoro, I. and Alves, I. (2018). Multiplex interbank networks and systemic importance: An application to european data. *Journal of Financial Stability*, 35:17–37.

Anderberg, M. R. (2014). *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, volume 19. Academic press.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113.

Barigozzi, M. and Brownlees, C. (2019). Nets: Network estimation for time series. *Journal of Applied Econometrics*, 34(3):347–364.

Barigozzi, M. and Farnè, M. (2024). An algebraic estimator for large spectral density matrices. *Journal of the American Statistical Association*, 119(545):498–510.

Bassett, D. S. and Sporns, O. (2017). Network neuroscience. *Nature neuroscience*, 20(3):353–364.

Battiston, S., Puliga, M., Kaushik, R., Tasca, P., and Caldarelli, G. (2012). Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific reports*, 2(1):541.

Becher, C., Millard, S., and Soramaki, K. (2008). The network topology of chaps sterling.

Bhachech, J., Chakrabarti, A., Kaizoji, T., and Chakrabarti, A. S. (2022). Instability of networks: effects of sampling frequency and extreme fluctuations in financial data. *The European Physical Journal B*, 95(4):71.

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.

Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, 104(3):535–559.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308.

Bonanno, G., Lillo, F., and Mantegna, R. (2001). High-frequency cross-correlation in a set of stocks. *Quantitative Finance*, 1(1):96–104.

Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network analysis in the social sciences. *science*, 323(5916):892–895.

Borghesi, C., Marsili, M., and Micciche, S. (2007). Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 76(2):026104.

Brida, J. G., Matesanz, D., and Seijas, M. N. (2016). Network analysis of returns and volume trading in stock markets: The euro stoxx case. *Physica A: Statistical Mechanics and its Applications*, 444:751–764.

Briola, A. and Aste, T. (2022). Dependency structures in cryptocurrency market from high to low frequency. *Entropy*, 24(11):1548.

Caccioli, F., Barucca, P., and Kobayashi, T. (2018). Network models of financial systemic risk: a review. *Journal of Computational Social Science*, 1:81–114.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model.

Coletti, P. (2016). Comparing minimum spanning trees of the italian stock market using returns and volumes. *Physica A: Statistical Mechanics and its Applications*, 463:246–261.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223.

Cont, R. and Schaanning, E. (2017). Fire sales, indirect contagion and systemic stress testing. *Indirect Contagion and Systemic Stress Testing (June 13, 2017).*

Coronnello, C., Tumminello, M., Lillo, F., Micciche, S., and Mantegna, R. (2007). Economic sector identification in a set of stocks traded at the new york stock exchange: a comparative analysis. In *Noise and Stochastics in Complex Systems and Finance*, volume 6601, pages 177–188. SPIE.

Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 84(6):066106.

Di Nuzzo, C. and Farnè, M. (2025). Recovering the number of clusters from a laplacian matrix by nuclear norm penalization. *Statistical Analysis and Data Mining: An ASA Data Science Journal*, 18(4):e70031.

Diebold, F. X. and Yilmaz, K. (2009). Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal*, 119(534):158–171.

Diebold, F. X. and Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of forecasting*, 28(1):57–66.

Diebold, F. X. and Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of econometrics*, 182(1):119–134.

Duffie, D. and Zhu, H. (2011). Does a central clearing counterparty reduce counterparty risk? *The Review of Asset Pricing Studies*, 1(1):74–95.

Epps, T. W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366a):291–298.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Felippe, H., Battiston, F., and Kirkley, A. (2024). Network mutual information measures for graph similarity. *Communications Physics*, 7(1):335.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3-5):75–174.

Galbiati, M. and Stanciu-Vizeteu, L. (2015). Communities and driver nodes in the target2 payment system. *Journal of Financial Market Infrastructures*, 3(4).

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Granger, C. W. J. and Cootner, P. H. (1965). The random character of stock market prices. *Journal of the Royal Statistical Society. Series A (General)*, 128(3):447.

Greenwood, R., Landier, A., and Thesmar, D. (2015). Vulnerable banks. *Journal of Financial Economics*, 115(3):471–485.

Gualdi, S., Cimini, G., Primicerio, K., Di Clemente, R., and Challet, D. (2016). Statistically validated network of portfolio overlaps and systemic risk. *Scientific reports*, 6(1):39467.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York.

Hidalgo, C. A. and Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26):10570–10575.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

Iori, G., De Masi, G., Precup, O. V., Gabbi, G., and Caldarelli, G. (2008). A network analysis of the italian overnight money market. *Journal of Economic Dynamics and Control*, 32(1):259–278.

Jackson, M. O. et al. (2008). *Social and economic networks*, volume 3. Princeton university press Princeton.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 83(1):016107.

Kaufmann, L. and Rousseeuw, P. (1987). Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416.

Krampe, J. and Paparoditis, E. (2025). Frequency domain statistical inference for high-dimensional time series. *Journal of the American Statistical Association*, 120(551):1580–1592.

Leibon, G., Pauls, S., Rockmore, D., and Savell, R. (2008). Topological structures in the equities market network. *Proceedings of the National Academy of Sciences*, 105(52):20589–20594.

Leinonen, H. and Soramäki, K. (1999). *Optimizing liquidity usage and settlement speed in payment systems*. Number 16/1999. Bank of Finland Discussion Papers.

Lopez de Prado, M. (2016). Building diversified portfolios that outperform out-of-sample. *Journal of Portfolio Management*.

MacMahon, M. and Garlaschelli, D. (2013). Community detection for correlation matrices. *arXiv preprint arXiv:1311.1924*.

MacMahon, M. and Garlaschelli, D. (2015). Community detection for correlation matrices. *Phys. Rev. X*, 5:021006.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11:193–197.

Mantegna, R. N. and Stanley, H. E. (1999). *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). Robust statistics.

Massara, G. P., Di Matteo, T., and Aste, T. (2016). Network filtering for big data: Triangulated maximally filtered graph. *Journal of complex Networks*, 5(2):161–178.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895.

Meilă, M. and Shi, J. (2001). A random walks view of spectral segmentation. In *International workshop on artificial intelligence and statistics*, pages 203–208. PMLR.

Miccichè, S., Bonanno, G., Lillo, F., and Mantegna, R. N. (2003). Degree stability of a minimum spanning tree of price return and volatility. *Physica A: Statistical Mechanics and its Applications*, 324(1-2):66–73.

Musmeci, N., Aste, T., and Di Matteo, T. (2015). Relation between financial market structure and the real economy: comparison between clustering methods. *PloS one*, 10(3):e0116201.

Musmeci, N., Nicosia, V., Aste, T., Di Matteo, T., and Latora, V. (2017). The multiplex dependency structure of financial markets. *Complexity*, 2017(1):9586064.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14:849–856.

Onnela, J.-P., Chakraborti, A., Kaski, K., Kertesz, J., and Kanto, A. (2003a). Asset trees and asset graphs in financial markets. *Physica Scripta*, 2003(T106):48.

Onnela, J.-P., Chakraborti, A., Kaski, K., Kertesz, J., and Kanto, A. (2003b). Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 68(5):056110.

Palomar, D. P. (2025). *Portfolio Optimization: Theory and Application*. Cambridge University Press.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of business*, pages 61–65.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925–979.

Peixoto, T. P. (2013). Parsimonious module inference in large networks. *Physical review letters*, 110(14):148701.

Peixoto, T. P. (2014). Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047.

Peixoto, T. P. (2017). Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1):012317.

Peixoto, T. P. (2019). Bayesian inference of network structure from data. *Physical Review Letters*, 123(12):128301.

Peixoto, T. P. (2021). Revealing consensus and dissensus between network partitions. *Physical Review X*, 11(2):021003.

Peixoto, T. P. (2023). *Descriptive vs. inferential community detection in networks: Pitfalls, myths and half-truths*. Cambridge University Press.

Peixoto, T. P. (2025a). Network reconstruction via the minimum description length principle. *Physical Review X*, 15(1):011065.

Peixoto, T. P. (2025b). Scalable network reconstruction in subquadratic time. *Proceedings of the Royal Society A*, 481(2324):20250345.

Peralta, G. and Zareei, A. (2016). A network approach to portfolio selection. *Journal of Empirical Finance*, 38:157–180.

Podobnik, B., Horvatic, D., Petersen, A. M., and Stanley, H. E. (2009). Cross-correlations between volume change and price change. *Proceedings of the National Academy of Sciences*, 106(52):22079–22084.

Poledna, S., Martínez-Jaramillo, S., Caccioli, F., and Thurner, S. (2021). Quantification of systemic risk from overlapping portfolios in the financial system. *Journal of Financial Stability*, 52:100808.

Pozzi, F., Di Matteo, T., and Aste, T. (2013). Spread of risk across financial markets: better to invest in the peripheries. *Scientific reports*, 3(1):1665.

Raffinot, T. (2018). The hierarchical equal risk contribution portfolio. *Available at SSRN 3237540*.

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology advances*, 49:107739.

Ronchetti, E. M. and Huber, P. J. (2009). *Robust statistics*. John Wiley & Sons Hoboken, NJ, USA.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

Song, W.-M., Di Matteo, T., and Aste, T. (2012). Hierarchical information clustering by means of topologically embedded graphs. *PloS one*, 7(3):e31929.

Soramäki, K., Bech, M. L., Arnold, J., Glass, R. J., and Beyeler, W. E. (2007). The topology of interbank payment flows. *Physica A: Statistical Mechanics and its Applications*, 379(1):317–333.

Soramäki, K. and Cook, S. (2013). Sinkrank: An algorithm for identifying systemically important banks in payment systems. *Economics*, 7(1):20130028.

Squartini, T., Van Lelyveld, I., and Garlaschelli, D. (2013). Early-warning signals of topological collapse in interbank networks. *Scientific reports*, 3(1):1–9.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051.

Sun, Y., Babu, P., and Palomar, D. P. (2014). Regularized tyler's scatter estimator: Existence, uniqueness, and algorithms. *IEEE Transactions on Signal Processing*, 62(19):5143–5156.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the royal statistical society: series b (statistical methodology)*, 63(2):411–423.

Tola, V., Lillo, F., Gallegati, M., and Mantegna, R. N. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258.

Tumminello, M., Aste, T., Di Matteo, T., and Mantegna, R. N. (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426.

Tumminello, M., Coronnello, C., Lillo, F., Micciche, S., and Mantegna, R. N. (2007a). Spanning trees and bootstrap reliability estimation in correlation-based networks. *International Journal of Bifurcation and Chaos*, 17(07):2319–2329.

Tumminello, M., Di Matteo, T., Aste, T., and Mantegna, R. N. (2007b). Correlation

based networks of equity returns sampled at different time horizons. *The European Physical Journal B*, 55:209–217.

Tyler, D. E. (1987). A distribution-free m-estimator of multivariate scatter. *The annals of Statistics*, pages 234–251.

Vitali, S., Glattfelder, J. B., and Battiston, S. (2011). The network of global corporate control. *PloS one*, 6(10):e25995.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17:395–416.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337.

Wang, G.-J., Xie, C., and Chen, S. (2017). Multiscale correlation networks analysis of the us stock market: a wavelet analysis. *Journal of Economic Interaction and Coordination*, 12(3):561–594.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Wu, Z. and Leahy, R. (2002). An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 15(11):1101–1113.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of statistics*, pages 642–656.

Zelnik-Manor, L. and Perona, P. (2004). Self-tuning spectral clustering. *Advances in neural information processing systems*, 17.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI open*, 1:57–81.

List of Figures

1.1 Different filtering methodologies. 21

2.1 Schematic representation of the standard procedure 40

2.2 Comparison of monthly beta estimates (30-day windows) obtained with OLS and LTS methods for Colgate (left) and Procter & Gamble (right). 42

2.3 Time series of deviations between idiosyncratic return components extracted from one-factor models estimated via OLS and LTS methods for Colgate (left) and Procter & Gamble (right). 43

2.4 Comparison of rolling window Pearson correlations computed from residuals estimated using OLS and LTS methods for Colgate and Procter & Gamble (left) and for Equity Residential and Realty Income Corporation (right). The correlations are computed on 1000 days rolling windows with a 30 days shift between windows. 43

2.5 Comparison of rolling window correlations on OLS residuals computed from the Pearson and regularized Tyler estimators for Colgate and Procter & Gamble (left) and for Equity Residential and Realty Income Corporation (right). The correlations are computed on 1000 days rolling windows with a 30 days shift between windows. 48

2.6 Comparison of rolling window correlations computed with the Pearson estimator on OLS residuals and the regularized Tyler estimator on LTS residuals for Colgate and Procter & Gamble (left) and for Equity Residential and Realty Income Corporation (right). The correlations are computed on 1000 days rolling windows with a 30 days shift between windows. 49

2.7 Schematic representation of the procedure used to embed different data types 56

2.8 Schematic representation of the procedure used to embed different frequencies 59

2.9 time series of Adjusted Rand Index between clusters extracted by applying AL hierarchical clustering clustering to the affinity matrices extracted by computing the Pearson correlation between OLS idiosyncratic returns (in blue) and Tyler correlations between LTS idiosyncratic returns (in green) on returns sampled at different frequencies. 67

2.10 box plots of Adjusted Rand Index between clusters extracted by applying AL hierarchical clustering to the affinity matrices extracted by computing the Pearson correlation between OLS idiosyncratic returns (in blue) and Tyler correlations between LTS idiosyncratic returns (in green) on returns sampled at different frequencies. 67

2.11 time series of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Pearson correlation between OLS idiosyncratic returns (in blue) and Tyler correlations between LTS idiosyncratic returns (in green) on returns sampled at different frequencies. 68

2.12 box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Pearson correlation between OLS idiosyncratic returns (in blue) and Tyler correlations between LTS idiosyncratic returns (in green) on returns sampled at different frequencies. 69
2.13 box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Pearson correlation (in blue) and Tyler correlations (in green) between volumes sampled at different frequencies. 70
2.14 box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Pearson correlation (in blue) and Tyler correlations (in green) between volatilities sampled at different frequencies. 70
2.15 time series of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Tyler correlation on LTS idiosyncratic returns (in blue) and on the fused matrix from Tyler correlations computed on LTS idiosyncratic returns, volumes and volatilities (in green) sampled at different frequencies. 71
2.16 box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Tyler correlation on LTS idiosyncratic returns (in blue) and on the fused matrix from Tyler correlations computed on LTS idiosyncratic returns, volumes and volatilities (in green) sampled at different frequencies. 72

2.17 box plots of Adjusted Rand Index between clusters extracted by applying spectral clustering to the fused matrices from Pearson correlations computed on OLS idiosyncratic returns, volumes and volatilities (in blue) and Tyler correlations computed on LTS idiosyncratic returns, volumes and volatilities (in green) sampled at different frequencies. 72

2.18 time series of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Tyler correlation on LTS idiosyncratic returns individually for each and on the fused matrix from Tyler correlations computed on LTS idiosyncratic returns at different frequencies for fixed clusters (left) and varying clusters (right). 73

2.19 time series of Adjusted Rand Index between clusters extracted by applying spectral clustering to the affinity matrices extracted by computing the Tyler correlation on LTS idiosyncratic returns individually for each and on the fused matrix from Tyler correlations computed on LTS idiosyncratic returns at different frequencies for fixed clusters (left) and varying clusters (right). 74

2.20 Information ratios of the individual cluster-wise/sector-wise returns with respect to market average return (left). average individual cluster-wise/sector-wise extra returns with respect to market average (center). standard deviation of the individual cluster-wise/sector-wise extra returns with respect to market average (right). All statistics are aggregated by taking the weighted mean, weighting each cluster by the total traded volume of its constituents over the period. 77

2.21 Left: time series of the percentage increase of Realized volatilities of the individual cluster-wise/sector-wise portfolios with respect to the realized volatility of the market portfolio. Right: box plots of the percentage increase of Realized volatilities of the individual cluster-wise/sector-wise portfolios with respect to the realized volatility of the market portfolio. All statistics are aggregated by taking the weighted mean, weighting each cluster by the total traded volume of its constituents over the period. 78

2.22 Top: time series of the meta cluster labels over time. Bottom: sparse co-occurrence frequency graphs for the meta clusters above. 83

3.1 top: Network reconstructed from the full 20 years sample. Nodes are positioned according to the communities they belong to. Node color represents the communities found by the SBM (left) and the economic sectors of activity (right); bottom-left: sub graph of two communities of stocks belonging mostly to the IT sector; bottom-right: sub-graph of communities of stocks belonging mostly to the IT sector. 105

3.2 Partition overlap between the community labels of different algorithms for 11 clusters (left) and 15 clusters (right). The last column shows the overlap with industrial sectors. 106

3.3 top: partition overlap between partitions found by the algorithms and economic sectors of activity, vertical dashed lines represent the temporal averages; bottom: number of communities found by the algorithms. 108

3.4 Threshold graph at different levels of the graph resulting from aggregating the (unweighted) graphs from all the time windows and applying different thresholds. The goal is to show the most persistent edges over time at the various levels. 110

3.5 Spearman correlation between network statistics computed on graphs from different time windows: Degree (upper-left), Betweenness centrality (upper-right), Closeness centrality (lower-left), local clustering coefficient (lower-right) 112

3.6 a) Value of the S&P 500 index over time. Colors represent the clusters obtained from the DC-NMI similarity. b) networks reconstructed by aggregating the reconstructed networks within each cluster and retaining the 10% most significant edges. 113

List of Tables

2.1 Summary of all the data used in the analysis 57

2.2 Statistical analysis results across different configurations 78

3.1 S&P 500 composition vs dataset composition. 102