Co-trading networks for modelling dynamic interdependency structures and estimating high-dimensional covariances in US equity markets

Yutong Lu[1], Gesine Reinert[1,3], and Mihai Cucuringu[1,2,3,4]

[1]Department of Statistics, University of Oxford, Oxford, UK

[2]Mathematical Institute, University of Oxford, Oxford, UK

[3]The Alan Turing Institute, London, UK

[4]Oxford-Man Institute of Quantitative Finance, University of Oxford, Oxford, UK

January 31, 2023

Abstract

The time proximity of trades across stocks reveals the topological structure of the equity market in the United States. In this article, we investigate how such cross-stock trading behaviors, which we name as *co-trading*, shape the market structures and affect stock price co-movements. By leveraging a co-trading-based pairwise similarity measure, we propose a novel method to construct (dynamic) networks of stocks. Our empirical studies use high-frequency limit order book data from 2017-01-03 to 2019-12-09. By applying the spectral clustering algorithm on co-trading networks, we uncover economically meaningful clusters of stocks. Beyond the static Global Industry Classification Standard sectors, our data-driven clusters capture the time evolution of the dependency among stocks. Furthermore, we demonstrate statistically significant positive relations between low-latency co-trading and return covariance. With the aid of the co-trading network, we develop a robust estimator for high-dimensional covariance matrix estimation, which yields superior economic value on portfolio allocation. The mean-variance portfolios based on our covariance estimates achieve both lower volatility and higher Sharpe ratios than standard benchmarks.

Keywords: Market microstructure; Co-occurrence analysis; Network; Machine learning; Robust covariance estimation; Portfolio allocation

Contents

1 Introduction

The pioneering work of Kyle (1985) posits the price formation at high-frequency level as the interaction among market participants. In his model, market makers monitor the aggregated order flows after informed and liquidity traders submit their orders, and then set their fair prices. With the development and boom of high-frequency trading (HFT) strategies, the interplay becomes more aggressive and sophisticated. Recent works (Brunnermeier and Pedersen (2005); Van Kervel and Menkveld (2019); Hirschey (2021); Yang and Zhu (2020)) find that HFT traders actively detect activities of other participants in the market, and fiercely trade against them. Furthermore, these interactions can span across different stocks (Hasbrouck and Seppi (2001); Bernhardt and Taub (2008); Capponi and Cont (2020)). We investigate concurrent (almost instantaneous) trading across multiple stocks, a phenomenon which we refer to as *co-trading* behavior, at a very granular level by directly considering individual trades, and zooming-in around their local neighborhoods.

In this paper, we propose a novel method that constructs *co-trading networks*, in order to model the complex structures of co-trading activities in equity markets. Constructed from limit order books, our co-trading network bridges the trading behaviors at a granular level with the dynamic topological structures of the market and price co-movements among individual stocks. Moreover, by making use of the co-trading network, we develop a robust estimator for high-dimensional covariance matrices, and demonstrate its conspicuous economic value on portfolio allocation.

The network construction starts with a pairwise similarity measure between stocks. Inspired by the idea of trade co-occurrence originating from Lu, Reinert, and Cucuringu (2022), we define the pairwise similarity as the normalized count of times that trades, for a given pair of stocks, arrive concurrently. We name this measure as a *co-trading score*, since it embeds the intuition that stocks frequently traded together are closely related. Concatenating co-trading scores between every pair of stocks, we obtain the co-trading matrix, which serves as the representation of the proposed network of equity markets.

Utilizing the developed algorithms and tools for network analysis, we provide empirical evidence that co-trading networks capture meaningful patterns of the market. By visualizing the co-trading network, aggregated over the entire period of study, with information filtering (Rosario N Mantegna (1999)), we observe that stocks from the same sector groups tend to have strong co-trading relations. It echos the empirical phenomena driven by the existence of sector structures in the market, where stocks in the same sector groups are likely to appear in the same portfolios and their prices tend to move together. Additionally, we uncover clusters of stocks in the co-trading network. To detect these communities, we apply spectral clustering on co-trading matrices to group stocks with similar co-trading behaviors into clusters. For comparison purpose, we select the Global Industry Classification Standard (GICS) as sector labels. Our empirical results show substantial overlap between the data-driven clusters and GICS sectors, which confirms that the co-trading networks accommodate the sector structures as expected. Moreover, we leverage the uncovered networks to study the influence of both individual stocks and sectors on the market. By using tools such as eigenvector centrality (Bonacich (1972), Bonacich (1987)), we identify that large technology companies and financial institutions, such as Microsoft, Apple JPMorgan, etc., present stronger co-trading relations with others, thus have higher impact on the structure of the market. Despite the alignment with GICS sectors, the co-trading network also contains information beyond sectorial structure. The data-driven clusters also group together closely related stocks from different sectors.

Apart from the aforementioned static graph, the construction of co-trading matrices is flexible at various frequency. It is informative to analyse the time evaluation of co-trading networks and explore the dynamic of market structures. In this study, we focus on network time series at daily level, while it can be easily generalized to intraday, monthly and so forth. Our empirical findings from daily co-trading networks indicates that the co-trading relations across different sectors increase from 2017 to 2019. Using spectral clustering, we detect clusters at a daily level and compare them with GICS sectors. By plotting the similarity across time, we observe a downward tend with fluctuations. In addition, we also compare the similarity among daily clusters. The variation in clusters also increases as the spread of co-trading beyond sector groups. By applying the spectral clustering algorithm for change-point detection based on the temporal similarity heatmap, we uncover three distinct regimes over the period of study.

To exploit the association between co-trading behaviors and price co-movements, we conduct network regression analysis. We build realized covariance matrices from intraday returns as proxies for co-movements among stocks. By regressing covariance matrices against co-trading networks on a daily basis, we reveal positive associations between the two types of matrices. Furthermore, on 98.51% of the days in the study, the positive regression coefficients are statistically significant. Further controlling for GICS sectors, we conduct multiple regressions by adding a fixed sector network as an independent variable. Even with the presence of sectors, the positive and significant relation still holds. Therefore, the co-trading behaviors at the high-frequency level are positively correlated with the return covariance, and have explainability power on price co-movements that goes beyond the commonly adopted sectors.

With the aid of co-trading matrices, we propose a robust estimator for the high-dimensional covariance matrix of stock returns at daily frequency. By assuming stock returns follow a linear factor structure (Ross (1976)), we decompose the covariance matrix as the sum of factor covariance and idiosyncratic covariance. Then, we impose a block structure on the diagonal of the idiosyncratic covariance matrix, such that elements outside the blocks are set to zero. Our approach extends the work of Ait-Sahalia and Xiu (2017), which use principle component analysis to derive latent factors and form blocks using GICS sectors. However, the sector blocks remain static over time. As we have shown, the market structures are time-varying, and thus static sector memberships are not enough to capture the similarity of stocks at higher frequency. Therefore, we update the diagonal block structure daily with data-driven clusters derived from co-trading networks. To evaluate the performance, we construct mean-variance portfolios, subject to various leverage constraints. Our covariance estimators, based on dynamic clusters, generally outperform the baselines, using fixed GICS sectors, by achieving lower volatility and higher Sharpe ratios. Our best-performing portfolio achieves a Sharpe ratio of 1.40, which is 0.43 higher than the corresponding baseline, and 0.57 higher than the market over the same period.

The remainder of this paper is organized as follows. Section 2 outlines our contributions to the existing literature. Section 3 introduces the definition of co-trading score and the construction of co-trading networks. In Section 4, we begin our empirical studies with conducting exploratory analysis on a static co-trading network and detecting clusters. Next, we study the dynamics of the daily co-trading matrices in Section 5. Subsequently, we explore the relation between daily co-trading networks and covariance matrices in Section 6 and propose a co-trading based covariance estimator in Section 7. Finally, in Section 9, we conclude and discuss our limitations and potential research directions.

2 Literature review

This study sits at the confluence of three strands of literature. Firstly, our research enriches the network analysis and the modeling of complex inter-dependency relations in financial markets. Network analysis has been proven to be effective in studying inter-dependency relations in complex systems. In particularly, there is a large literature on networks in financial markets (Bardoscia et al. (2021); Marti et al. (2021)). Previous research developed various methodology to build financial networks. In 1999, the influential paper of Rosario N Mantegna (1999) first built a network from a distance measure based on correlations of stock returns, filtered with a minimum spanning tree (MST). Since then, many works followed by constructing networks with diverse distance/similarity measures, such as Granger causality (Billio et al. (2012)), mutual information (Fiedor (2014)), co-jumps of stock prices (Ding et al. (2021)), and so forth. Additionally, multiple methods have been used to replace MST for information filtering, including random matrix theory (Plerou et al. (2000)), Potts super-paramagnetic transitions (Kullmann, Kertesz, and R. Mantegna (2000)), planar maximally filtered graph (Tumminello et al. (2005)), threshold-filter method (W.-Q. Huang, Zhuang, and Yao (2009), Namaki et al. (2011)), etc.

Concerning networks of financial markets, preceding studies uncovered the topological structures of markets through community detection. For example, the pioneering work of Rosario N Mantegna (1999) applied hierarchical clustering and discovered hierarchical structure stock portfolios. Moreover, the networks can be constructed as a time series (McDonald et al. (2005); Nie (2017)). Recent work by Bennett, Cucuringu, and Reinert (2022) built lead-lag networks with different similarity measures, tested multiple clustering algorithms and studied the time-varying lead-lag structures in the market.

To this field, our contribution is proposing an original similarity measure, directly derived from very granular records of trades, with explicit interpretation as how frequently two stocks are traded together, with the final goal of constructing networks. In addition, we detect dynamic clusters and provide a comprehensive comparison with GICS sectors.

Secondly, this research contributes to the studies of market microstructure, especially interplay among trading activities. In 1985, Kyle (1985) posits a famous two-period model of high-frequency price formation by solving the equilibrium of the game between liquidity takers and market makers. Further, various studies show that, high-frequencies strategies can be more aggressive. Hirschey (2021) claims that high-frequency traders (HFTs) can predict order flows form other market participants and trade in front of them. Van Kervel and Menkveld (2019) provides empirical evidence that HFTs can detect the trading activities of institutional traders, and adjust their own strategies to speculate. Moreover, HFTs even actively explore the market by initiating small trades and watch the response of others(Clark-Joseph (2013)). Researchers also propose theoretical models Grossman and Miller (1988); Brunnermeier and Pedersen (2005); Yang and Zhu (2020)) for the interactions between HFTs and other traders.

The interactions can span across different stocks on the market. Bernhardt and Taub (2008) states that the strategical interplay among speculators is often concurrent and across many stocks. There is vast literature on cross-impact (Pasquariello and Vega (2015); Benzaquen et al. (2017); Schneider and Lillo (2019)), showing order flow of a stocks can affect prices of other stocks. Recent work of Lu, Reinert, and Cucuringu (2022) classifies trades of stocks by whether they concurrently arrive with other trades for the same or different, or both same and different stocks, and investigate price impact using order imbalance from different groups of trades. They discover that the time proximity of trade arrivals explains stock returns.

This study extends Lu, Reinert, and Cucuringu (2022) to the network setting by considering interactions between the trades of every pair of stocks. Rather than studying the price impact of order imbalance on one stock at a time, we construct co-trading networks of all stocks together and explain the impact of trading interactions at market microstructure level on macroscopic price co-movements.

Finally, our study adds to the growing body of financial econometrics literature on robust estimation of high-dimensional covariance matrices of stock returns. An invertible and well-behaved covariance matrix is essential for portfolio allocation with mean-variance optimization (Markowitz (1952)). However, when the number of sampled timestamps is small relative to the size of the panel of stocks, sample covariance matrices are singular or ill-conditioned. To overcome this issue, previous studies develop different streams of regularized estimation methods, including thresholding (Bickel and Levina (2008b); Bickel and Levina (2008a)), shrinkage (Ledoit and Wolf (2003); Ledoit and Wolf (2004); Chen et al. (2010)), etc. These regularization techniques are based on structural assumptions; for example, thresholding estimators assume the covariance matrices are sparse. In particular for stocks, a large literature of asset pricing studies revealed linear factor structures of equity returns (Sharpe (1964); Ross (1976); Fama and French (1992); Fama and French (1993); Fama and French (2015)). Based on factor models, the covariance matrix of returns can be decomposed as sum of a low-rank factor component and a residual idiosyncratic component. Various lines of work have then imposed different types of sparsity assumptions on the residual component which represents the covariance of idiosyncratic risk of stocks. Fan, Liao, and H. Liu (2016) provides a through overview of factor-based robust covariance estimation. By using observable and latent factors, respectively, the works of Fan, Furger, and Xiu (2016) and Ait-Sahalia and Xiu (2017) impose block structures on the idiosyncratic component, where stocks are sorted by their GICS sector membership, thus forcing the residual covariance of stocks in different sectors to be zero. Consequently, they conclude that incorporating clusters with economic interpretation benefits the covariance estimation task.

Our method for robust estimation contributes to this body of literature, and can be construed as a direct extension of the two aforementioned articles, by taking into account very granular high-frequency data that encodes higher-order relationships on the co-trading behaviour. Instead of employing static GICS sectors, we perform time-varying data-driven clustering. By capturing the dynamic dependency relations between the universe of stocks, our proposed method outperforms baselines in portfolio allocation tasks.

3 Construction of co-trading networks

In this section, we first propose pairwise co-trading scores to measure the similarity between two stocks, using the *co-occurrence of trades* methodology, proposed by Lu, Reinert, and Cucuringu (2022). We then leverage these scores to build affinity matrices and construct co-trading networks.

3.1 Co-occurrence of trades

We first introduce notations and define co-occurrence of trades. Here, we define each trade as a 4-tuple. Let $x_k = (\tau_k, s_k, d_k, q_k)$ denote the information of the k^{th} trade, where we capture the following trade statistics

- τ_k is the time when the trade occurs;
- s_k indicates name of the stock for which the trade is executed;
- $d_k \in \{buy, sell\}$ indicates whether the trade is buyer- or seller- initiated;
- q_k is the volume of the trade.

Then, for every trade x_k, we define a δ-neighbourhood of trades which are close in time

$$\mathcal{N}^{\delta}_{x_k} = \{x_a | a \neq k \text{ and } \tau_a \in (\tau_k - \delta, \tau_k + \delta)\},$$

where δ is a predefined threshold corresponding to time.

Following the definition of trade co-occurrence from Lu, Reinert, and Cucuringu (2022), we say that trade x_k *co-occurs* with x_l at level δ, denoted by $x_k \overset{\delta}{\sim} x_l$, if and only if $x_l \in \mathcal{N}^{\delta}_{x_k}$. Figure 1 visualizes the definition, where x_k co-occurs with x_l and x_m, but does not co-occur with x_n. Trade x_n co-occurs with x_m, and both trade x_l and trade x_m co-occur with trade x_k, but they do not co-occur with each other.

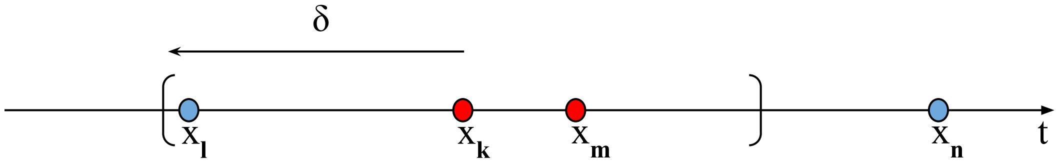

Figure 1: Illustration of trade co-occurrence. This figure visualizes the idea of co-occurrence of trades. With a pre-defined neighbourhood size δ, trade x_l arrives within the δ-neighbourhood of trade x_k, and thus they co-occur. In contrast, trade x_n locates outside x_k's δ-neighbourhood, and thus the two trades do not co-occur. Trade x_n co-occurs with x_m, and both trade x_l and trade x_m co-occur with trade x_k, but they do not co-occur with each other.

The co-occurrence relation is symmetric, that is, given δ, if $x_k \overset{\delta}{\sim} x_l$, then $x_l \overset{\delta}{\sim} x_k$. Notice that co-occurrence of trades is not a equivalence relation. When it is clear from the context, we omit the level δ of the co-occurrence when referring to co-occurrence.

3.2 Pairwise co-trading score

Motivated by the intuition that stocks are more inter-dependent if they co-trade together more often, we propose pairwise co-trading scores to measure the similarity between stocks. For a pair of stocks, we calculate the similarity score by counting selected types of co-occurred trades and normalize by the total number of trades of both stocks.

The formal definitions are as follows. For stock i on day t, the set of all trades, with direction $d^i \in \{buy, sell, all\}$, is denoted by

$$S_t^{i,d^i} = \{x_a | \tau_a \in [t_{start}, t_{end}], s_a = i, d_a = d^i\},$$

where $d^i = all$ denotes all trades without distinguishing between buy and sell. Then, if given another set S_t^{j,d^j}, we count the number of trades for stock j, which co-occur with trades in S_t^{i,d^i}, denoted as

$$L_{t,j\to i}^{d^j \to d^i} = \sum_{x_k \in S_t^{i,d^i}} |\{x_a \in \mathcal{N}_{x_k}^{\delta} | s_a = j, d_a = d^j\}|,$$

where $|\cdot|$ denotes the cardinality of a set.

The pairwise co-occurrence count index $c_{t,i,j}^{\delta,d^i,d^j}$ is a scaled count of the number of trades for stock i with direction d^i, and trades for stock j with direction d^j, which co-occur on day t. Formally, it is defined as

$$c_{t,i,j}^{\delta,d^i,d^j} := \frac{L_{t,i\to j}^{d^i\to d^j} + L_{t,j\to i}^{d^j\to d^i}}{\sqrt{|S_t^{i,d^i}|}\sqrt{|S_t^{j,d^j}|}}.$$

These pairwise co-occurrence indices have three useful properties. Firstly, they are non-negative and higher values indicates stronger co-occurrence relations. Secondly, the indices are scaled, so that they can be used to compare relations across pairs of stocks. Thirdly, the indices are symmetric. Additionally, the indices are defined using sets of trades which are filtered based on different conditions so that they are flexible and can easily be generalized to a customized set of orders.

3.3 Co-trading matrices and networks

Using the pairwise co-occurrence measures, we build the corresponding daily $N \times N$ co-occurrence matrix, denoted as $\mathbf{C}_t^{\delta,d^i,d^j}$, having entries

$$(\mathbf{C}_t^{\delta,d^i,d^j})_{i,j} = c_{t,i,j}^{\delta,d^i,d^j}.$$

Built from daily matrices, co-occurrence matrices over a longer time period T, such as months and years, are simply calculated by averaging the daily co-occurrence matrices;

$$\mathbf{C}_{\{T\}}^{\delta,d_i,d_j} = \frac{1}{|T|}\sum_{t\in T} C_t^{\delta,d_i,d_j}.$$

Taking advantage of the symmetric co-occurrence matrices, we build co-occurrence networks to represent complex structures in the stock market. We consider dynamic co-trading networks of stocks, $G_t = (V_t, E_t)$, $t \in T$, where each vertex $v_{i,t} \in V_t$ represents a certain stock at time t and a weighted edge $e_{i,j}(t) \in E$ denotes a type of co-occurrence relation between two stocks at time t; the weight is the corresponding co-trading score. We use the co-trading matrices as representations of each co-trading network. The empirical study in the following sections focuses on the co-trading matrices without taking into account the directions of trades. In line with Lu, Reinert, and Cucuringu (2022), we choose $\delta = 500$ milliseconds as the neighborhood size, in order to determine co-occurring trades. Therefore, we omit the superscripts for brevity and only keep the subscript for time index, e.g. $\mathbf{C}_{2017-2019}$ stands for the co-trading network aggregated from 2017 to 2019.

4 Empirical network analysis

In this section, we construct co-trading networks from the empirical data. We provide a visualization of the networks, and show that they reflect empirical phenomena which have been observed in equity markets. Furthermore, we demonstrate that co-trading networks capture inter-dependency of stocks and time-varying topological structures of these markets.

4.1 Data

The empirical studies in this paper are based on 457 stocks in US equity markets from 2017-01-03 to 2019-12-09. We acquire limit order book data from the LOBSTER database (R. Huang and Polak (2011)), which keeps track of submissions, cancellations and executions of limit orders for stocks traded in NASDAQ. Each record has a timestamp with precision up to 10^{-9} seconds and indicates the price, size and direction of the respective order book event. A trade occurs when a market/marketable order arrives and consumes existing limit orders, and thus can be inferred by order executions. Here we denote a trade as 'buy' if the limit order denoted a willingness to sell, and as 'sell' if the limit order denoted a willingness to buy. To derive trades, we filter out events other than executions. Then we aggregate records with exactly the same timestamp and direction, as they are likely to be caused by one large marketable order. The direction of a trade is opposite to those of the executed limit orders it is matched against.

In addition, we obtain per-minute price data from LOBSTER and daily stock prices from the Center for Research in Security Prices (CRSP) database. Apart from prices, we label stock sectors using the Global Industry Classification Standard (GICS) drawn from Compustat database. The GICS decomposition classifies stocks into 11 sectors of varying sizes.

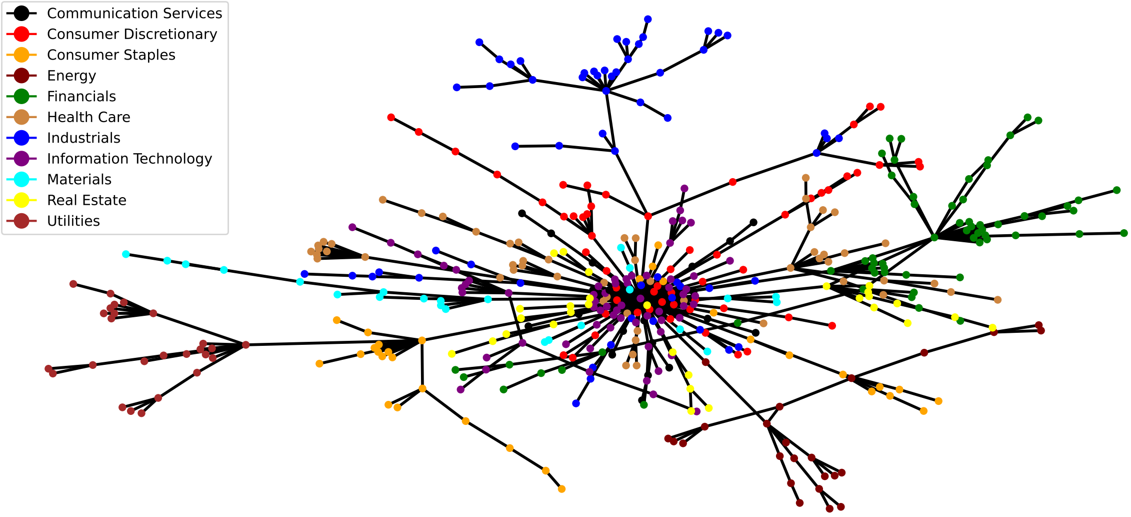

Figure 2: A co-trading network of the period from 2017-01-03 to 2019-12-09.
The figure presents the co-trading network of 457 stocks represented by the co-trading matrix, without trading directions, of the entire period of study, that is $\mathbf{C}_{2017-2019}$. Each node represents a stock, whose color indicates its GICS sector. For visualization, we use a maximum spanning tree (Rosario N Mantegna (1999)) to filter out edges by maximizing sum of co-trading scores while keeping a connected network with 456 edges.

4.2 An example of a co-trading network

To provide a general view of co-trading, we construct a network-based matrix $\mathbf{C}_{2017-2019}$, capturing all co-trading relations of the entire period of study from 2017-01-03 to 2019-12-09, without differentiating directions of trades.

To visualize the network, we follow Rosario N Mantegna (1999) to filter edges with a minimum spanning tree (MST), because the co-trading matrix is dense. To be specific, a MST of a graph with N nodes is a connected subgraph with only $N-1$ edges such that the sum of the negations of edge weights is minimized, which can be found by Kruskal's algorithm (Kruskal (1956)). Figure 2 shows the MST of the co-trading network based on all trades of all stocks. The graph vertices represent individual stocks and their colors indicate the GICS sectors. It is noteworthy that companies within the same sector groups are often on the same branches, suggesting that stocks may frequently be co-traded in sector baskets. This reconciles with the known fact that stocks in the same sector tend to move together, due to common membership in the same index traded funds (Harford and Kaul (2005)) and similar exposure to the same factors. Hence, this co-trading network captures meaningful patterns of the cross-sectional structure of stocks.

Furthermore, the network visualization reveals a dense cluster at the center of the plot. The stocks close to the centroid co-trade with more stocks and are, on average, more inter-related with the rest of the network. To further investigate the cluster, we use eigenvector centrality (Bonacich (1972), Bonacich (1987)) to measure the influence of each node on the network. For each stock i, we define the eigenvector centrality as the ith element of the eigenvector corresponding to the dominant eigenvalue of the co-trading matrix. The larger the centrality, the more "influential" the stock.

We list the 10 stocks with the highest eigenvector centrality in Table 1. Microsoft is the most influential stock followed by Apple and JPMorgan Chase. Moreover, among the top 10 stocks, 60% are technology companies, including Facebook, and 40% are financial institutions. The co-trading matrix can not only model individual stocks, but it can also model relations among sectors. Following Bennett, Cucuringu, and Reinert (2022), we build a meta-flow network of sectors as follows. We group stocks by their sector labels, use the sectors as nodes in the meta-flow network, and use the average co-trading scores of stocks in each pair of sectors to be the sector co-trading scores, serving as edge weights in the meta-flow network. Figure 3 pictures the fully connected sector network, where the edge width indicates the strength of co-trading. The strongest co-trading relation is between Information Technology and Communication Services. By comparing edge widths, we can identify which sectors are more closely co-traded with a given sector. For example, the Real Estate sector has a strong co-trading relation with the Financials sector, while it does not show a strong relationship with the other sectors. Moreover, we document centrality of each sector in Table 2, and find that Information Technology, Financials and Communication Services are the most influential sectors, while Real Estate and Utilities have the lowest eigenvalue centrality.

4.3 Temporal evolution of co-trading networks

In Figure 4, we plot networks corresponding to the month of January, for each of 2017, 2018 and 2019, with colours representing GICS sectors. After thresholding, we only preserve 1% of edges, selected such that we keep the edges with the highest weight. The co-trading structures in the US market change over the sample period. At the start of the period, there are no strong co-trading relations across GICS sectors. However, towards the end of the period, an increasing amount of edges connect stocks from different sectors, thus providing supporting evidence for the importance of data-driven clustering that goes

Table 1: Top 10 stocks by eigenvector centrality.
This table lists 10 stocks ranked by their eigenvector centrality, as well as their company names and GICS sectors.

Ticker	Centrality	Company	Sector
MSFT	0.12	Microsoft Corp.	Information Technology
AAPL	0.11	Apple Inc.	Information Technology
JPM	0.11	JPMorgan Chase & Co.	Financials
BRK.B	0.10	Berkshire Hathaway	Financials
TXN	0.10	Texas Instruments	Information Technology
FB	0.09	Facebook, Inc.	Communication Services
V	0.09	Visa Inc.	Information Technology
AXP	0.09	American Express Co	Financials
PYPL	0.09	PayPal	Information Technology
CSCO	0.09	Cisco Systems	Information Technology

Table 2: Centrality of sectors in the meta-flow network.
This table shows the eigenvector centrality of each GICS sector in the meta-flow network. The 'Rank' column reports the rank, in descending order, of each sector in terms of their centrality.

	Centrality	Rank
Information Technology	**0.39**	**1**
Financials	**0.35**	**2**
Communication Services	**0.34**	**3**
Industrials	0.33	4
Health Care	0.33	5
Consumer Staples	0.31	6
Consumer Discretionary	0.29	7
Materials	0.26	8
Energy	0.26	9
Utilities	0.20	10
Real Estate	0.19	11

beyond the usual sector-based decompositions.

5 Clustering analysis

So far, we have observed the presence of clusters, or communities, in the co-trading network of stocks. In this section, we use an unsupervised clustering method to classify every stock, with the aim to group similar stocks into the same cluster. Starting with the network in Figure 2, we show that the co-trading matrix not only captures sector structure, but also incorporates associations beyond sectors. Furthermore, we show that the co-trading matrices contain information on the dynamic of market structures, by

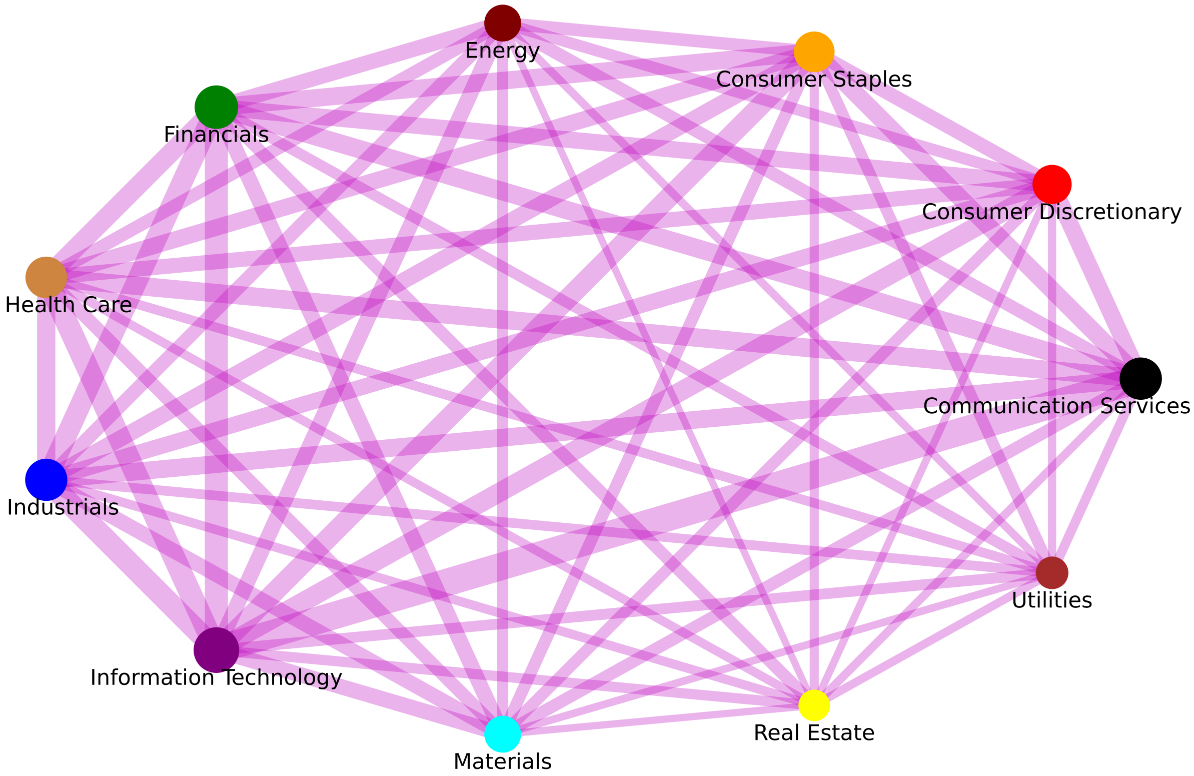

Figure 3: Meta-flow network of GICS sectors.
This figure illustrates the fully connected network of GICS sectors based on the co-trading matrix $C_{2017-2019}$. The edges are calculated by grouping stocks by their GICS sectors and averaging the co-trading scores of sector groups.

studying the relations between generated clusters and GICS sectors, making pairwise comparison between clusters derived from different co-trading matrices, and studying the changes of clusters over time.

5.1 Methodology and evaluation metrics

Spectral clustering is a family of algorithms (Shi and Malik (2000), Ng, Jordan, and Weiss (2002), Cucuringu, Davies, et al. (2019), Cucuringu, Li, et al. (2020)), built upon spectral graph theory, to detect communities or clusters in networks. For details, Von Luxburg (2007) provides a comprehensive survey on spectral methods and their theoretical backgrounds. Given a co-trading matrix, we apply a spectral clustering algorithm, outlined in Appendix A, to identify clusters in our universe of stocks. The number of clusters is a hyper-parameter of this algorithm, which we need to determine beforehand. Although the spectral clustering depends on random initialization, in Appendix B, we show that the algorithm is robust in our empirical setting.

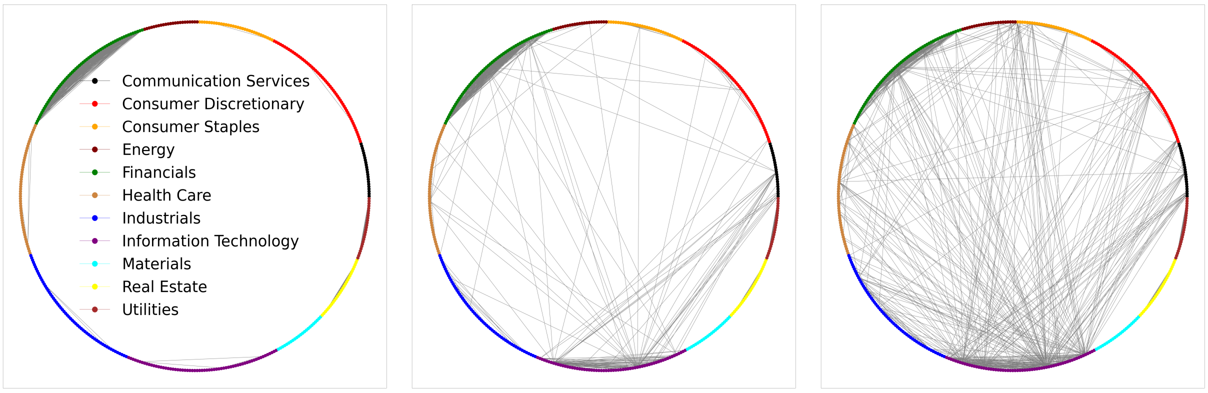

Figure 4: Thresholded temporal Co-trading Networks.
These three networks are based on co-occurrence of all trades regardless of trading directions. From left to right, are networks of 2017-01, 2018-01 and 2019-01. Colours stand for the GICS sector each stock belongs to.

The Adjusted Rand index (ARI) (Hubert and Arabie (1985)) is our measure of similarity between clusters. The valid range of ARI is $[0, 1]$, where 1 is achieved if and only if two clusters perfectly match. Notice that ARI can take negative values, which means the two clusters are not similar at all. In general, a higher ARI indicates that two clusters are more similar to each other, and vice versa. A value close to 0 indicates that points are assigned into clusters randomly.

5.2 Clusters v.s. GICS

Analyzing the clusters, we discover that co-trading networks capture sectors in the US stock market, which supports our observations in Section 4.2.

Figure 5 shows the commonality between GICS sectors and clusters detected in the co-occurrence network corresponding to the matrix $\mathbf{C}_{2017-2019}$. By setting the number of clusters to 11, which is also the number of GICS sectors after the classification change in 2018, we observe that the unsupervised clustering method can recover the sector groups to a good standard. This is especially the case for companies in the Financial, Utilities, Real Estate and Energy sectors, when there are hardly any mismatches. There are also clear clusters for stocks in the sectors Materials, Health Care, Industrials and Consumer Staples, with small amounts of disagreements. Comparatively, the structure of the Information Technology sector is strikingly different, with one major cluster containing most of the stocks and most of the other stocks well spread out across other clusters. Moreover, the Consumer Discretionary sector stocks concentrate in two clusters, one of which also

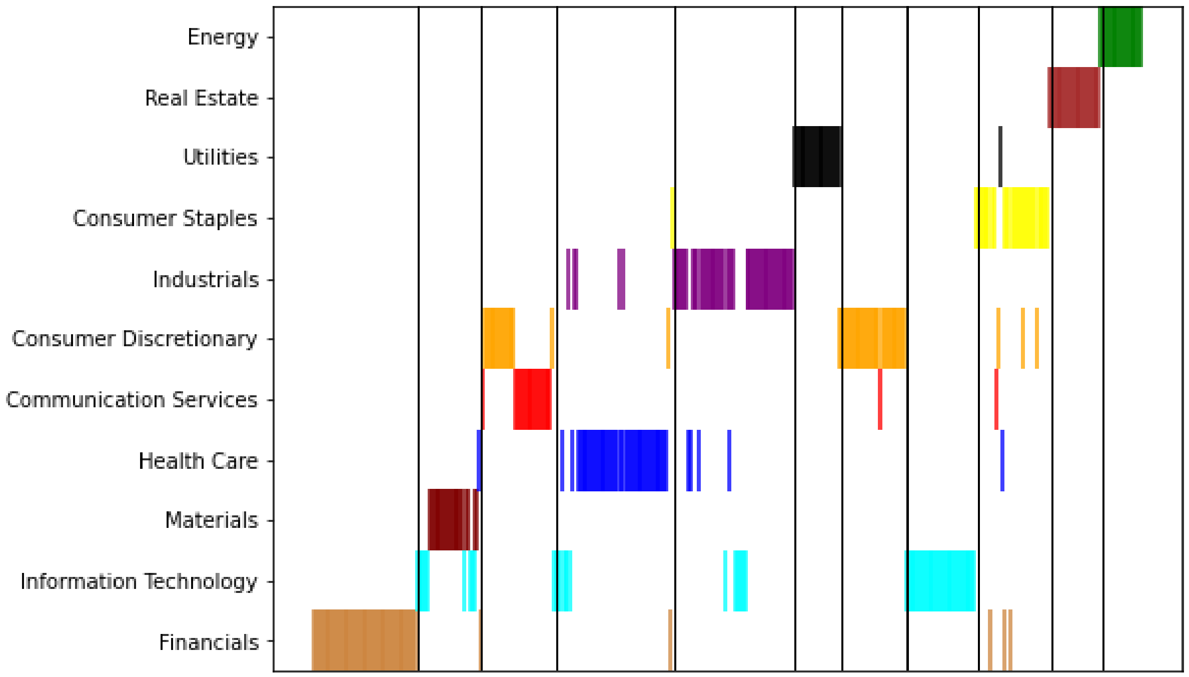

Figure 5: Clusters v.s. sectors.
This plot visualizes the overlap between the data-driven clusters, derived from $\mathbf{C}_{2017-2019}$ by spectral clustering, and the GICS sectors. The horizontal axis indexes stocks grouped by data-driven clusters, separated by vertical lines. The vertical axis indicates GICS sector labels. The colored area of a sector is indicative of its size. The Financial stocks are well grouped in Cluster 1, while the Information Technology stocks are spread over multiple clusters, with a strong presence in Cluster 8.

contains the entire sector of Communication Services.

5.3 Temporal evolution of clusters and market regime detection

In addition to the uncovered long-term agreement between data-driven clusters and GICS sectors, we are also interested in assessing and quantifying the extent to which the cross-sectional trading behaviors deviate from sectors within a short period of time. Therefore, we also proceed with clustering stocks on a daily basis, and then compare the similarity between clusters and sectors, via the same adjusted Rand index. Since the true number of communities in the stock market is unknown, we consider multiple values for the number of clusters, namely 11 (the same number as the number of GICS sectors), 15, 20 and 50.

Figure 6 shows the time series of ARIs between daily clusters and GICS sectors. For a small number of clusters, we observe meaningful levels of similarity between the data-driven clusters and economic sectors. In contrast, when the number is large as 50, the similarities are constantly low. As empirical observations indicate that market partici-

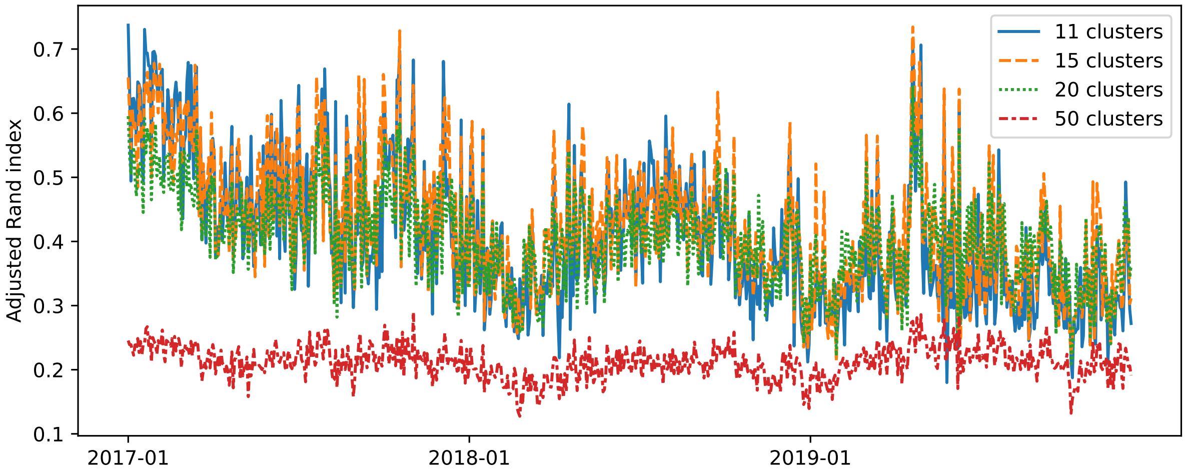

Figure 6: Adjusted Rand index between daily clusters and GICS sectors.
This figure sketches dynamics similarity between data-driven clusters and GICS sectors. For every day, we derive data-driven clusters and calculate the adjusted Rand index with GICS, with predefined numbers of clusters $\{11, 15, 20, 50\}$. We then plot the ARIs over time from 2017-01-03 to 2019-12-09, so that each line represents one choice of number of clusters.

Table 3: Statistics of daily ARI.
This table shows summary statistics of daily ARIs between dynamic clusters and fixed GICS sectors plotted in Figure 6. For different number of clusters, the table reports the mean and standard deviation of ARIs from 2017-01-03 to 2019-12-09, as well as the signal-to-noise ratio.

	Clusters			
	11	15	20	50
Mean	0.41	0.43	0.40	0.21
Standard deviation	0.11	0.10	0.07	0.03
Signal-to-noise ratio	3.73	4.30	5.71	7.00

pants tend to trade sectors, we conclude that the true number of clusters is relatively small. Moreover, in Table 3 we summarize the mean and standard deviation of daily ARIs, for each choice of cluster numbers, over the entire period, echoing our visual findings. In addition, although the average ARI values of 11, 15 and 50 clusters are comparable, the variation drops by 0.03 as the number of clusters increases from 15 to 20, which is more conspicuous than the increase from 11 to 15. Based on the signal-to-noise ratio and taking the ARI into account, we settle to further investigate the case of 20 clusters.

Focusing on the temporal ARI curve of 20 clusters, we highlight two empirical findings. Firstly, at the daily frequency, co-trading behaviors align well with economic sectors, with frequent fluctuations. Secondly, we observe a downward trend in similarities, hinting that the sector structures become less prominent from 2017 to 2019. This reinforces the observation from Figure 4 concerning the growth of strong co-trading relations across GICS sectors. As expected, our co-trading networks embed dynamic structures in stock markets beyond static economic sectors.

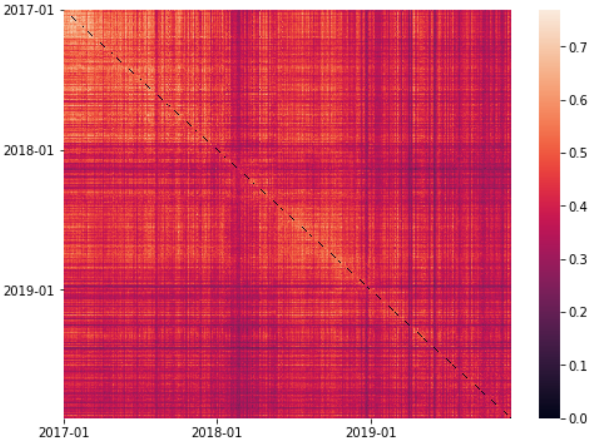

Figure 7: Similarity of daily clusters.
This heatmap shows the similarity between days. Every day, we group stocks into 20 clusters by applying spectral clustering on the daily co-trading matrix. Then, for each pair of days, we calculate the ARI between the clusters.

In addition to comparing daily clusters with the benchmark, we also measure the similarity between clusters corresponding to every pair of days. Figure 7 shows a heatmap of all such pairwise ARIs, based on 20 clusters. It is noteworthy that the colors along the diagonal get darker from upper left to lower right corner, which indicates that the market structures become unstable over time. Moreover, we distinguish three blocks along the main diagonal. For a quantitative regime detection, we apply the spectral clustering method on the heatmap, and clearly identify three clusters of trading days, shown in Figure 8.

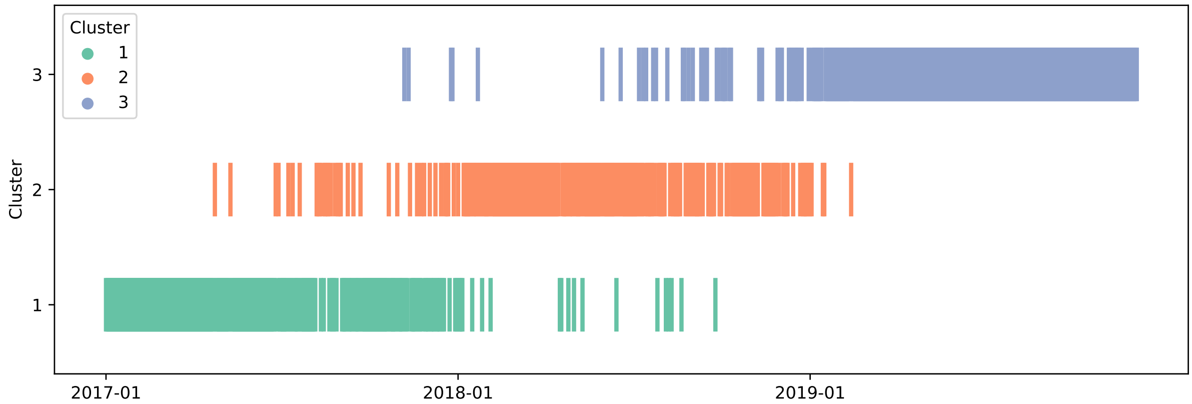

Figure 8: Regimes detected by spectral clustering.
This figure shows the clusters of days, derived by applying spectral clustering on the heatmap of ARI in Figure 7 to classify days into three regimes. The horizontal axis indicates dates and the vertical axis indicates which cluster each day belongs to.

6 Co-trading and covariance

In this section, we connect the co-trading behavior to co-movement in stock prices. A direct measure of price co-movement is the covariance matrix. On every day t, we define a *realized covariance matrix*, denoted as $\hat{\mathbf{\Sigma}}_t^R$, as a proxy of the true covariance $\mathbf{\Sigma}_t$, based on logarithmic returns of stocks. We conduct a regression analysis and present significant associations between co-trading and realized covariance matrices.

6.1 The realized covariance matrix

For each day t, we define the *realized covariance matrix* following the three steps below. Since we only use intraday data, the index t is omitted in this part to avoid ambiguity. Firstly, we split the normal trading period from 9:30 to 16:00 into equally spaced and non-overlapping intervals of length Δ. We denote $Int = \{\tau_1, \tau_2, \ldots, \tau_m\}$ as the set of left end-points of each interval. Secondly, we define the logarithmic return of all stocks, $\mathbf{r}_\tau \in \mathbb{R}^N$, for each period, indexed by the left point of the interval τ, as

$$\mathbf{r}_\tau = \log(\mathbf{P}_\tau) - \log(\mathbf{P}_{\tau-\Delta}),$$

where $\mathbf{P}_\tau \in \mathbb{R}^N$ are mid prices at time τ. Finally, we build the realized covariance matrix as

$$\hat{\mathbf{\Sigma}}^R = \sum_{\tau \in Int} \mathbf{r}_\tau \mathbf{r}_\tau^T.$$

For the empirical study in the following sections, we construct daily realized covariance matrices with sampling frequency of 5 minutes, that is $\Delta = 5$ min, in line with Andersen et al. (2001); L. Y. Liu, Patton, and Sheppard (2015).

6.2 Network regression analysis

In order to assess significance of the relationships between co-trading and covariance matrices, we perform a network regression on each day t

$$\hat{\mathbf{\Sigma}}_t^R = \gamma_t \mathbf{C}_t + \mathbf{E}_t, \tag{1}$$

where $\gamma_t \in \mathbb{R}$ is the regression coefficient, and $\mathbf{E}_t \in \mathbb{R}^{N \times N}$ is the residual matrix. As there are cross-sectional autocorrelations among stocks, it is not appropriate to assume independence in the residuals. Instead, to draw inference on the regression coefficients, we use the quadratic assignment procedure (QAP) (Mantel (1967), Krackardt (1987), Krackhardt (1988)), which simulates the distribution of test statistics by simultaneously permuting pairs of row and column of the explanatory matrix at random. Table 4 reports the regression coefficients. Positive relations between covariance and co-trading matrices appear in all days over the sample period, with mean and median of daily regression coefficients equal 5.10 and 4.48, respectively. Moreover, 98.51% of these positive relations are statistically significant at the 5% significance level.

Table 4: Simple network regression.
This table reports the mean, median and standard deviation of daily network regression coefficients in (1), as well as their p-values. The p-values are obtained using theQAP with 2000 simulations. We run one regression for each trading day from 2017-01-03 to 2019-12-09.

	γ_t	p-value
Mean	5.10	0.004
Median	4.48	0.000
Standard deviation	2.86	0.034
Percentage positive	100	-
Percentage significant	98.51	-

In the next step, we assess whether co-trading networks explain cross-sectional variation in covariance beyond GICS sectors. To account for the sector structure in the regression, we introduce a static network of GICS sectors, as follows

$$\mathbf{S}_{i,j} = \begin{cases} 1, & \text{if stock } i \text{ and stock } j \text{ belong to the same cluster} \\ 0, & \text{otherwise.} \end{cases}$$

Then, controlling for the sector network, we perform the following regression

$$\hat{\mathbf{\Sigma}}_t^R = \gamma_t^C \mathbf{C}_t + \gamma_t^S \mathbf{S} + \mathbf{E}_t, \tag{2}$$

where $\gamma_t^C, \gamma_t^S \in \mathbb{R}$ are regression coefficients. We conduct a QAP in a multiple regression setting (MRQAP) using double semi-partialing (SDP), an approach proposed by Dekker, Krackhardt, and Snijders (2007).

The regression coefficients for the both co-trading and sector networks are shown in Table 5; they are highly significant, indicating that there is a relationship between the co-trading networks and the realized covariance matrix. It is noteworthy that the conclusion holds even when accounting for sectors. The mean and median of daily γ_t^C are 3.61 and 3.04, with 99.73% of positive values. During the whole sample period, 89.81% of the coefficients are statistically significant. As expected, sector networks are also positively correlated with price co-movements and significantly associated to the covariance structure of the stocks. With GICS sectors included in regressions, the coefficients of daily co-trading matrices are lower, since co-trading matrices evidently incorporate sector structures as well. Still, even then co-trading matrices explain some of the covariance of stocks beyond GICS sectors.

7 High-dimensional covariance estimation and application to portfolio allocation

The realized covariance matrices we study in Section 6 pertain to 457 stocks, but are constructed from 78 samples (corresponding to 5 minute buckets) for each stock. The estimations fall into a high-dimensional setting where the number of samples is smaller

Table 5: Multiple network regression.
This table reports the mean, median and standard deviation of daily network regression coefficients in (2), as well as their p-values. We perform one regression for each trading day from 2017-01-03 to 2019-12-09. For inference, we use MRQAP with SDP, in line with Dekker, Krackhardt, and Snijders (2007) and simulate 2000 runs to calculate p-values.

	γ_t^C	p-value	γ_t^S	p-value
Mean	3.61	0.029	0.24	0.001
Median	3.04	0.000	0.20	0.000
Standard deviation	2.52	0.116	0.16	0.009
Percentage positive	99.73	-	99.86	-
Percentage significant	89.81	-	99.56	-

than the dimension of the covariance matrix, resulting in singular estimates. However, it is important to have invertible and well-behaved estimates in practice, for tasks such as mean-variance portfolio allocation.

Fortunately, we have proven significant associations between realized covariance and co-trading matrices. With the aid of data-driven co-trading clusters, we develop a method for robust estimation of high-dimensional covariance matrices. This approach is motivated by the work of Ait-Sahalia and Xiu (2017), and can be construed as an extension of it that considers higher-order interactions by leveraging very granular high-frequency data.

7.1 Robust covariance matrix estimation

As in Section 6, we estimate covariance matrices on every day t, noting that the approach easily generalizes to any length of time period. For simplicity, we omit the subscript t in this subsection.

Assuming that the intraday logarithmic stock returns, $\mathbf{r}_\tau \in \mathbb{R}^N$, at time τ, follow a linear K-factor model, they can be decomposed as

$$\mathbf{r}_\tau = \beta \mathbf{f}_\tau + \mathbf{u}_\tau,$$

where $\mathbf{f}_\tau \in \mathbb{R}^K$ is a vector of latent factors, $\beta \in \mathbb{R}^{N\times K}$ is a static loading matrix, and $\mathbf{u}_\tau \in \mathbb{R}^K$ is the idiosyncratic component assumed to be independent of $\mathbf{f}_\tau$. Therefore, the covariance matrix of $\mathbf{r}_\tau$ can be decomposed as

$$\mathbf{\Sigma} = \beta \mathbf{\Sigma}^f \beta^T + \mathbf{\Sigma}^u, \tag{3}$$

where $\mathbf{\Sigma}^f = \sum_{\tau \in Int} \mathbf{f}_\tau \mathbf{f}_\tau^T$ is the low-rank covariance matrix of factors evaluated on the left endpoints of the intervals, and $\mathbf{\Sigma}^u = \sum_{\tau \in Int} \mathbf{u}_\tau \mathbf{u}_\tau^T$ is the idiosyncratic covariance. Ait-Sahalia and Xiu (2017) prove that, under additional assumptions, one can use the eigenvalues and eigenvectors to approximate the factor and idiosyncratic covariance matrices, and that the approximation errors are bounded. That is, one may write

$$\beta \mathbf{\Sigma}^f \beta^T \approx \sum_{k=1}^{K} \lambda_k \mathbf{v}_k \mathbf{v}_k^T,$$
$$\mathbf{\Sigma}^u \approx \sum_{k=K+1}^{N} \lambda_k \mathbf{v}_k \mathbf{v}_k^T,$$

where λ_k is the k-th largest eigenvalue of $\mathbf{\Sigma}$ and $\mathbf{v}_i$ denotes its corresponding eigenvector. Thus, we can write the sample covariance matrix as

$$\hat{\mathbf{\Sigma}}^R = \sum_{k=1}^{K} \hat{\lambda}_k \hat{\mathbf{v}}_k \hat{\mathbf{v}}_k^T + \hat{\mathbf{\Sigma}}^u, \tag{4}$$

where $\hat{\lambda}_k$ and $\hat{\mathbf{v}}_k$ are the k-th largest eigenvalue and eigenvector of the sample covariance matrix, and $\hat{\mathbf{\Sigma}}^u = \sum_{k=K+1}^{N} \hat{\lambda}_k \hat{\mathbf{v}}_k \hat{\mathbf{v}}_k^T$ is the approximation of the idiosyncratic covariance.

We propose a robust covariance estimator achieving positive definiteness by imposing a block structure on the idiosyncratic part of the sample covariance matrix, $\hat{\mathbf{\Sigma}}^u$, based on the data-driven clusters derived from co-trading matrices. We first threshold the sample covariance matrix to obtain a sparse matrix $\hat{\mathbf{\Gamma}}^u$, whose element corresponding to the pair of stocks i and j is

$$\hat{\mathbf{\Gamma}}^u_{i,j} = \begin{cases} \hat{\mathbf{\Sigma}}^u_{i,j}, & \text{if stock } i \text{ and stock } j \text{ belong to the same cluster} \\ 0, & \text{otherwise.} \end{cases}$$

Our proposed covariance estimator, incorporating the co-trading clusters, is given by

$$\hat{\mathbf{\Sigma}}^{Cluster} = \sum_{k=1}^{K} \hat{\lambda}_k \hat{\mathbf{v}}_k \hat{\mathbf{v}}_k^T + \hat{\mathbf{\Gamma}}^u.$$

The estimator is positive definite if the size of every cluster is smaller than the rank of the singular sample covariance matrix, which is usually the number of samples used to calculate it. To be specific, with a sampling frequency of 5 minutes, the largest cluster should contain no more than 78 stocks. In contrast to the paper Ait-Sahalia and Xiu (2017), which uses GICS sectors as a fixed "cluster", we can tune the number of clusters to guarantee positive definiteness given any universe of stocks. Moreover, thresholding with co-trading clusters embeds the dynamic of the dependency structure in stock markets, which is reasonable for covariance estimation at daily or higher frequency.

7.2 Mean-variance portfolio construction

To demonstrate economic value of the proposed robust covariance estimates, we develop a daily rebalanced mean-variance strategy, which opens positions at market open and liquidates at market close, for each trading day over the period of study. Based upon the assumption that $\mathbb{E}[\hat{\mathbf{\Sigma}}_t|\hat{\mathbf{\Sigma}}_{t-1}] = \hat{\mathbf{\Sigma}}_{t-1} \in \mathbb{R}^{N\times N}$, we derive mean-variance portfolio weights $\mathbf{w}_t \in \mathbb{R}^N$, on day t, by solving the following constrained optimization

$$\begin{aligned} \min_{w_t} \quad & \mathbf{w}_t^T \hat{\mathbf{\Sigma}}_{t-1} \mathbf{w}_t \\ \text{`s.t.} \quad & \mathbf{w}_t^T \vec{\mathbf{1}} = 1 \\ & ||\mathbf{w}_t||_1 \leq l, \end{aligned} \tag{5}$$

where $\vec{\mathbf{1}}$ denotes the all-ones vector, and $l \geq 0$ restricts the level of leverage. When $l = 1$, leverage is not allowed and all weights are non-negative. In contrast, when l is ∞, short-sell is unrestricted and the optimization leads to the global minimum variance portfolio (GMV) (Jagannathan and Ma (2003); Bollerslev, Patton, and Quaedvlieg (2018)). Note that it is possible for $\hat{\mathbf{\Sigma}}_{t-1}$ to be numerically singular. To avoid this issue, we do not trade on days with ill-behaved covariance estimates, and setting $\mathbf{w}_t = \vec{\mathbf{0}}$ if the condition number of $\hat{\mathbf{\Sigma}}_{t-1}$ is greater than 1×10^9. Finally, the daily portfolio return is calculated as

$$r_t^{mv} = \mathbf{w}_t^T \mathbf{r}_t,$$

where $\mathbf{r}_t \in \mathbb{R}^N$ is a vector of stock logarithmic open-to-close returns.

The evaluation metrics of portfolio performance are annualized volatility and Sharpe ratio. The annualized volatility is calculated as

$$\sigma^{mv} = stdev(r_t^{mv}) \times \sqrt{252},$$

where $stdev(\cdot)$ is the standard deviation function. The annualized Sharpe ratio (Sharpe (1994)) is defined as

$$sr^{mv} = \frac{\overline{r^{mv}} \times 252}{\sigma^{mv}},$$

where $\overline{r^{mv}}$ denotes the average daily return of the portfolio. According to the definition, a high Sharpe ratio indicates lower portfolio volatility adjusted by mean return.

7.3 Analysis of portfolio performance

In the following empirical portfolio analysis, we experiment with multiple values of the parameters corresponding to the number of data-driven clusters, number of factors for covariance estimation, and leverage constraints for optimization. For each set of parameters, we construct daily portfolio, and perform a backtest over the period of study. Since we calculate portfolio weights traded on day t from covariance estimates from day $t-1$, all of our tests are out-of-sample.

The annualized volatility of the portfolios is shown in Figure 9. We conclude that a larger number of clusters, which imposes higher level of sparsity on the residual covariance components, leads to more robust covariance estimation, and lower variation in portfolio returns. With short-sell prohibited, the portfolio risks are similar. As leverage constraint relax, the GICS portfolios with 11 fixed clusters and the portfolios corresponding to 15 co-trading clusters exhibit increasing annualized volatility. In contrast, when we consider 20 and 50 co-trading clusters, the annualized portfolio risks decrease, and stabilize at around 8%, regardless of the number of factor.

Figure 9 thus indicates that portfolios using our method of robust covariance estimation can lead to superior performance compared to portfolios built upon GICS sector-based covariance estimates. For further comparison after adjusting for returns, we add Sharpe ratios for selected portfolios in Table 6. In alignment with with our findings from Figure 9, data-driven clusters outperform fixed GICS sectors for robust covariance matrix estimation. However, increasing the number of clusters too much deteriorates the Sharpe ratio. It is noteworthy that, with 50 clusters, the idiosyncratic covariance matrices are

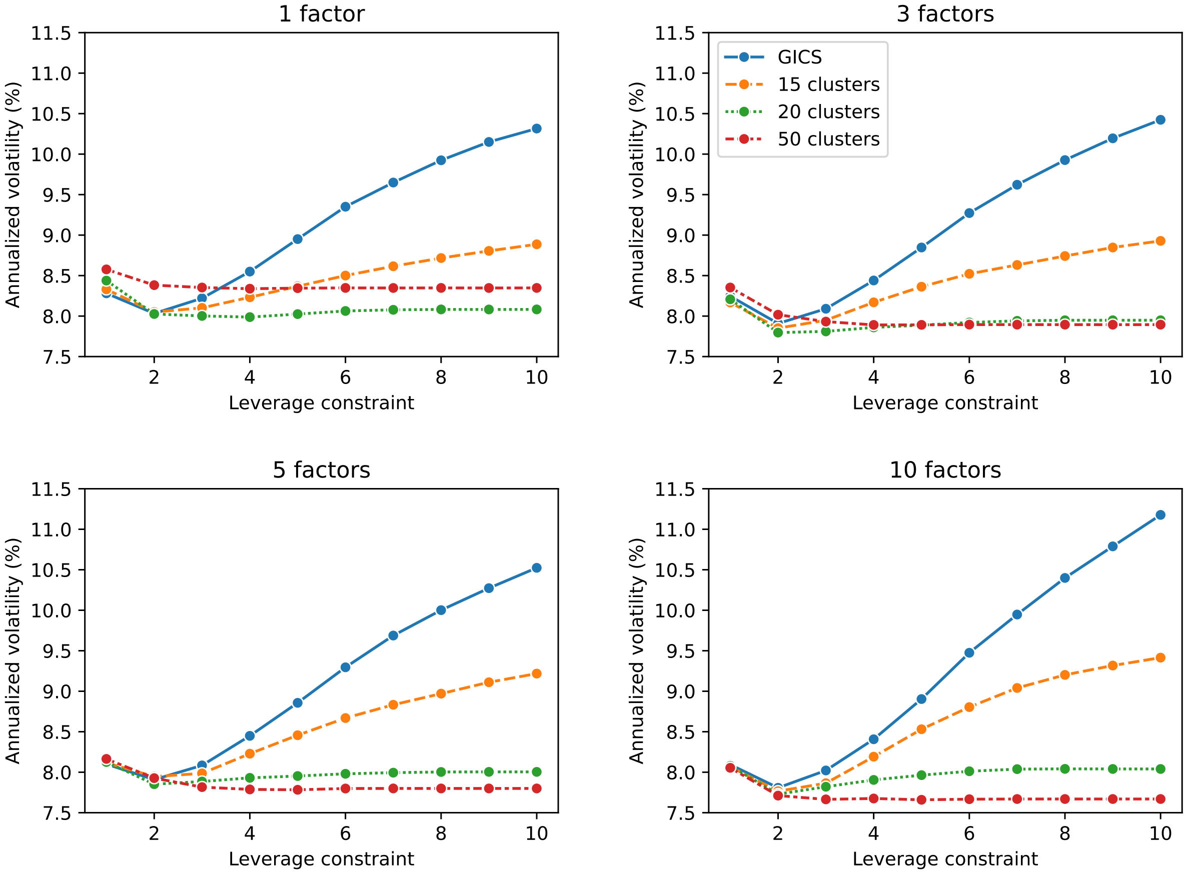

Figure 9: Annualized volatility of mean-variance portfolios.
These figures show the annualized volatility of mean-variance portfolios constructed by solving (5) based on different covariance matrices estimates. The out-of-sample backtests span the period from 2017-01-03 to 2019-12-09. Every sub-figure corresponds to one choice of latent factor numbers while decomposing sample covariance matrices in (3). In each sub-figure, every curve plots portfolio volatility, for a choice of blocks structure imposed on the residual covariance, along leverage constraints.

overly sparse and the covariance among stocks is underestimated. Thus the portfolio optimization tends to select stocks with lower volatility instead of diversifying, and results in low level of returns. From the numbers of clusters considered here, the highest Sharpe ratio of 1.40 is attained by the GMV for 20 clusters with 10 latent factors.

Table 6: Annualized Sharpe ratio of mean-variance portfolios.
This table documents the annualized Sharpe ratios of mean-variance portfolios constructed by solving (5) based on different covariance matrices estimates. The out-of-sample backtests span the period from 2017-01-03 to 2019-12-09. The 'Factor' column specifies the number of latent factors while decomposing the sample covariance matrices. The 'Leverage' column indicates leverage constraints in (5), where ∞ means that short-sell is unrestricted. We use 15, 20 and 50 clusters, together with GICS sectors as the baseline, while imposing diagonal block structure on the residual covariance matrices.

Factor	Leverage	Cluster			
		GICS	15	20	50
1	1	0.01	0.13	0.09	-0.06
	3	0.57	0.61	0.53	0.12
	5	0.75	0.89	0.69	0.12
	7	0.71	0.98	0.69	0.12
	∞	0.45	**1.07**	0.69	0.12
3	1	0.35	0.48	0.44	0.36
	3	0.54	0.74	0.86	0.51
	5	0.95	1.02	1.04	0.54
	7	0.92	1.13	1.11	0.54
	∞	0.59	**1.19**	1.12	0.54
5	1	0.38	0.52	0.51	0.47
	3	0.63	0.81	0.88	0.62
	5	1.04	1.11	1.16	0.70
	7	1.06	1.28	1.23	0.71
	∞	0.69	**1.29**	1.23	0.71
10	1	0.44	0.50	0.45	0.42
	3	0.64	0.89	1.04	0.58
	5	1.12	1.16	1.34	0.69
	7	1.27	1.28	1.39	0.70
	∞	0.97	1.18	**1.40**	0.70

Zooming into the trajectory of portfolio gains, we sketch the cumulative returns of the best GMV portfolios for the data-driven clusters we propose, along with the GICS sector as a baseline, respectively, in Figure 10. Furthermore, we add the S&P 500 index (SPY ETF) as a proxy for the market. Clearly, trading SPY from open to close every day is not a profitable trading strategy. Apart from that, the portfolio derived from our method outperforms the GICS baseline and the market benchmark, by achieving comparable profits while bearing much smaller fluctuations and shorter drawdown periods.

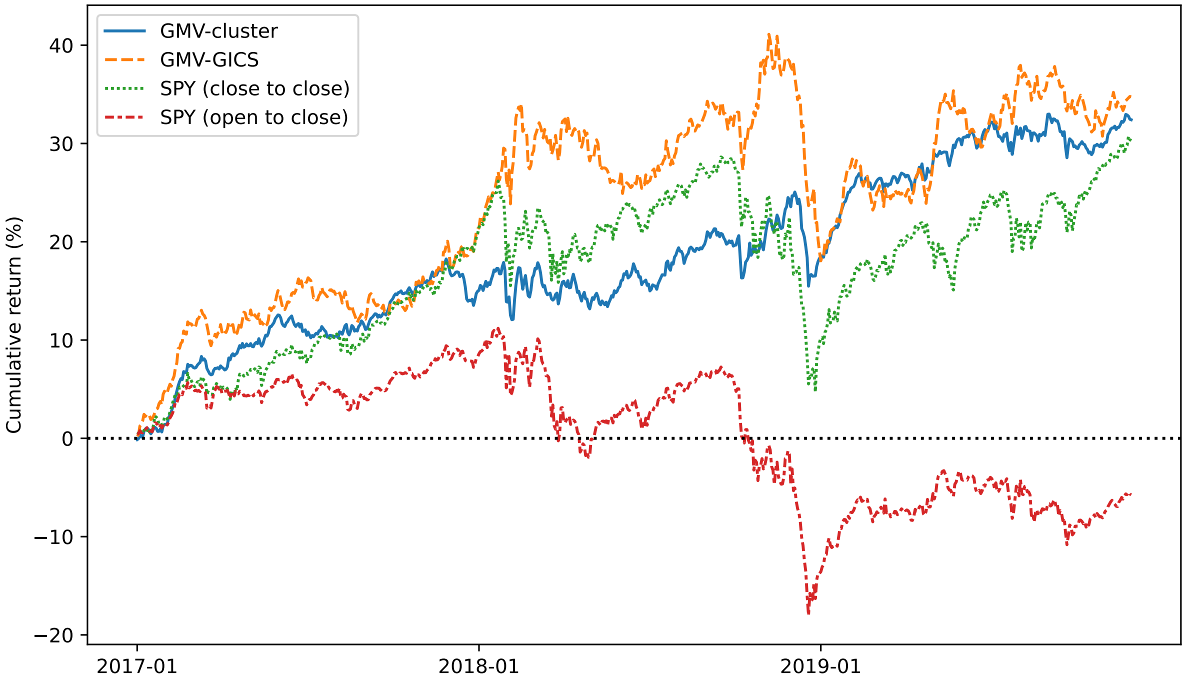

Figure 10: Cumulative returns of portfolios.
This figure plots cumulative returns of four portfolios from 2017-01-03 to 2019-12-09. The portfolios include (1) 'GMV-cluster': the global minimum variance portfolio based on robust covariance matrices estimated using our method with 10 factors and 20 clusters; (2) 'GMV-cluster': the GMV portfolio based on robust covariance matrices estimated using 10 factors and GICS sectors; (3) 'SPY (close to close)': the SPDR S&P 500 ETF Trust which tracks the S&P 500 Index; (4) 'SPY (open to close)': the SPY ETF, following our strategy, with positions only during normal trading hours.

8 Robustness

In this section, we briefly discuss the robustness of the spectral clustering algorithm. Moreover, we also report on the results when calculating co-trading scores based on traded volume instead of the number of trades. Details of robustness checks are provided in the appendix.

8.1 Random initialization of spectral clustering algorithm

The spectral clustering method applies the K-means algorithm (MacQueen (1967)) on the spectral domain of the co-trading matrices. An issue of K-means is that the clustering results may be sensitive to random initialization. In order to mitigate the sensitivity, we adopt K-means++ (Arthur and Vassilvitskii (2006)) method for initialization. To further empirically examine the robustness, we repeat the experiment of clustering on daily co-trading networks in Section 5.3 for 100 times with different random seeds. Then we compare the daily means of ARIs between clusters from each pair of experiments. Our results ensure that the spectral clustering method we use is robust on finding clusters in co-trading networks. Further details are in Appendix B.

8.2 Co-trading measured in volume

Instead of incorporating the number of transactions, we also explore the possibility of defining co-trading in terms of trading volumes. Details are included in Appendix C. By comparing ARI between corresponding clusters and GICS sectors, we observe the same patterns as for volume measured co-trading matrices. According to the portfolio performance, described in Section 7, the Sharpe ratios of GMV portfolios corresponding to the volume measured co-trading matrices surpass the GICS benchmarks. Overall our findings are robust under the volume measure in the sense that they show similar patterns.

Using count of trades appears to be more appropriate than volume in measuring co-trading. This finding echos previous research (Chan and Lakonishok (1995); Chordia and Subrahmanyam (2004)) which shows that, since institutions tend to split large orders to high their liquidation purpose and reduce market impact, the number of transactions outperforms the volume in measuring price impact. We confirm that using count of trades better captures patterns in price co-movement.

9 Conclusion and future research directions

We introduce and construct co-trading networks to model the dependency structure of stocks arising from the interplay of cross-stock trading among market participants, in response of trade arrivals on the market. Using a spectral clustering algorithm, we uncover clusters which capture well temporally evolving structures within markets, containing information beyond industry sectors. Our empirical studies, focusing on daily co-trading during 2017-01-03 to 2019-12-09, reveal that cross-stock trading behaviors are time-varying, and co-trading relations across different GICS sectors become apparent.

These two observations indicate that solely relying on sectors is not sufficient for capturing co-trading behavior, and they motivate the construction and analysis of time series of co-trading networks built from very granular high-frequency data.

Taking a further step, we establish that strong co-trading relations can lead to a high level of co-movements in stocks prices. We use realized covariance matrices to measure the co-movements of prices. Through network regression analysis, we document a significant positive relation between co-trading and covariance matrices. Even when adding a network of sectors as a control variable in the regression, the conclusion remains valid. This conclusion bridges the gap between cross-stock trading activities at the microstructure level and the macroscopic covariance of stock returns.

Employing dynamic co-trading networks and data-driven clusters, we develop a robust co-variance estimator for stock covariance, in a situation where the number of samples for estimating the covariance is smaller than the number of stocks. Our method outputs well-behaved estimates from sample covariance matrices, by incorporating contemporaneous information of stock clusters. As a result, a mean-variance portfolio constructed with our robust estimates achieves lower volatility and higher Sharpe ratios in comparison with baseline methods and market returns.

Our concept of co-trading provides a general framework to investigate the interaction of trade flow corresponding to different stocks; for example, one could take into account the directions of trades, as defined in Section 3, when analyzing the co-trading behaviour. It would also be worthwhile to further investigate how trades with same (resp., opposite) directions contribute to the positive (resp., negative) components of covariance and correlation of stocks. Additionally, in this study we assume that the co-trading score is symmetric; however, asymmetric scores may also be of interest, and the pairwise relationships could be modeled and clustered using methodology from the directed graph clustering literature (Cucuringu, Li, et al. (2020)). Intuitively, shocks on companies with large market cap can have impact on small cap stocks, but the converse is often not true. With a simple adjustment to the current co-trading scores, this asymmetry could be embedded into the network construction, allowing for the investigation of asymmetric spillover effects. Furthermore, it would be interesting to leverage the co-trading network time series for forecasting tasks concerning market structures, returns, and covariances, possibly combined with models ranging from parsimonious to deep learning.

References

[1] Yacine Ait-Sahalia and Dacheng Xiu. “Using principal component analysis to estimate a high dimensional factor model with high-frequency data”. In: *Journal of Econometrics* 201.2 (2017), pp. 384–399.

[2] Torben G Andersen et al. “The distribution of realized stock return volatility”. In: *Journal of Financial Economics* 61.1 (2001), pp. 43–76.

[3] David Arthur and Sergei Vassilvitskii. *k-means++: The advantages of careful seeding*. Tech. rep. Stanford, 2006.

[4] Marco Bardoscia et al. “The physics of financial networks”. In: *Nature Reviews Physics* (2021), pp. 1–18.

[5] Stefanos Bennett, Mihai Cucuringu, and Gesine Reinert. “Lead-lag detection and network clustering for multivariate time series with an application to the US equity market”. In: *arXiv preprint arXiv:2201.08283* (2022).

[6] Michael Benzaquen et al. “Dissecting cross-impact on stock markets: An empirical analysis”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2017.2 (2017), p. 023406.

[7] Dan Bernhardt and Bart Taub. “Cross-asset speculation in stock markets”. In: *The Journal of Finance* 63.5 (2008), pp. 2385–2427.

[8] Peter J Bickel and Elizaveta Levina. “Covariance regularization by thresholding”. In: *The Annals of Statistics* 36.6 (2008), pp. 2577–2604.

[9] Peter J Bickel and Elizaveta Levina. “Regularized estimation of large covariance matrices”. In: *The Annals of Statistics* 36.1 (2008), pp. 199–227.

[10] Monica Billio et al. “Econometric measures of connectedness and systemic risk in the finance and insurance sectors”. In: *Journal of Financial Economics* 104.3 (2012), pp. 535–559.

[11] Tim Bollerslev, Andrew J Patton, and Rogier Quaedvlieg. “Modeling and forecasting (un) reliable realized covariances for more reliable financial decisions”. In: *Journal of Econometrics* 207.1 (2018), pp. 71–91.

[12] Phillip Bonacich. “Factoring and weighting approaches to status scores and clique identification”. In: *Journal of Mathematical Sociology* 2.1 (1972), pp. 113–120.

[13] Phillip Bonacich. “Power and centrality: A family of measures”. In: *American Journal of Sociology* 92.5 (1987), pp. 1170–1182.

[14] Markus K Brunnermeier and Lasse Heje Pedersen. “Predatory trading”. In: *The Journal of Finance* 60.4 (2005), pp. 1825–1863.

[15] Francesco Capponi and Rama Cont. “Multi-asset market impact and order flow commonality”. In: *Available at SSRN* (2020).

[16] Louis KC Chan and Josef Lakonishok. “The behavior of stock prices around institutional trades”. In: *The Journal of Finance* 50.4 (1995), pp. 1147–1174.

[17] Yilun Chen et al. “Shrinkage algorithms for MMSE covariance estimation”. In: *IEEE Transactions on Signal Processing* 58.10 (2010), pp. 5016–5029.

[18] Tarun Chordia and Avanidhar Subrahmanyam. “Order imbalance and individual stock returns: Theory and evidence”. In: *Journal of Financial Economics* 72.3 (2004), pp. 485–518.

[19] Adam Clark-Joseph. “Exploratory trading”. In: *Unpublished job market paper. Harvard University, Cambridge, MA* (2013).

[20] Mihai Cucuringu, Peter Davies, et al. “SPONGE: A generalized eigenproblem for clustering signed networks”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1088–1098.

[21] Mihai Cucuringu, Huan Li, et al. “Hermitian matrices for clustering directed graphs: insights and applications”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 983–992.

[22] David Dekker, David Krackhardt, and Tom AB Snijders. “Sensitivity of MRQAP tests to collinearity and autocorrelation conditions”. In: *Psychometrika* 72.4 (2007), pp. 563–581.

[23] Yi Ding et al. “Stock co-jump networks”. In: *Available at SSRN* (2021).

[24] Eugene F Fama and Kenneth R French. “A five-factor asset pricing model”. In: *Journal of Financial Economics* 116.1 (2015), pp. 1–22.

[25] Eugene F Fama and Kenneth R French. “Common risk factors in the returns on stocks and bonds”. In: *Journal of Financial Economics* 33.1 (1993), pp. 3–56.

[26] Eugene F Fama and Kenneth R French. “The Cross-Section of Expected Stock Returns”. In: *The Journal of Finance* 47.2 (1992), pp. 427–465.

[27] Jianqing Fan, Alex Furger, and Dacheng Xiu. “Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data”. In: *Journal of Business & Economic Statistics* 34.4 (2016), pp. 489–503.

[28] Jianqing Fan, Yuan Liao, and Han Liu. “An overview of the estimation of large covariance and precision matrices”. In: *The Econometrics Journal* 19.1 (2016), pp. C1–C32.

[29] Paweł Fiedor. “Information-theoretic approach to lead-lag effect on financial markets”. In: *The European Physical Journal B* 87.8 (2014), pp. 1–9.

[30] Sanford J Grossman and Merton H Miller. “Liquidity and market structure”. In: *The Journal of Finance* 43.3 (1988), pp. 617–633.

[31] Jarrad Harford and Aditya Kaul. “Correlated order flow: Pervasiveness, sources, and pricing effects”. In: *Journal of Financial and Quantitative Analysis* 40.1 (2005), pp. 29–55.

[32] Joel Hasbrouck and Duane J Seppi. “Common factors in prices, order flows, and liquidity”. In: *Journal of Financial Economics* 59.3 (2001), pp. 383–411.

[33] Nicholas Hirschey. “Do high-frequency traders anticipate buying and selling pressure?” In: *Management Science* 67.6 (2021), pp. 3321–3345.

[34] Ruihong Huang and Tomas Polak. “Lobster: Limit order book reconstruction system”. In: *Available at SSRN 1977207* (2011).

[35] Wei-Qiang Huang, Xin-Tian Zhuang, and Shuang Yao. “A network analysis of the Chinese stock market”. In: *Physica A: Statistical Mechanics and its Applications* 388.14 (2009), pp. 2956–2964.

[36] Lawrence Hubert and Phipps Arabie. “Comparing partitions”. In: *Journal of classification* 2.1 (1985), pp. 193–218.

[37] Ravi Jagannathan and Tongshu Ma. “Risk reduction in large portfolios: Why imposing the wrong constraints helps”. In: *The Journal of Finance* 58.4 (2003), pp. 1651–1683.

[38] David Krackardt. “QAP partialling as a test of spuriousness”. In: *Social networks* 9.2 (1987), pp. 171–186.

[39] David Krackhardt. “Predicting with networks: Nonparametric multiple regression analysis of dyadic data”. In: *Social networks* 10.4 (1988), pp. 359–381.

[40] Joseph B Kruskal. “On the shortest spanning subtree of a graph and the traveling salesman problem”. In: *Proceedings of the American Mathematical society* 7.1 (1956), pp. 48–50.

[41] L Kullmann, J Kertesz, and RN Mantegna. “Identification of clusters of companies in stock indices via Potts super-paramagnetic transitions”. In: *Physica A: Statistical Mechanics and its Applications* 287.3-4 (2000), pp. 412–419.

[42] Albert S Kyle. “Continuous auctions and insider trading”. In: *Econometrica: Journal of the Econometric Society* (1985), pp. 1315–1335.

[43] Olivier Ledoit and Michael Wolf. “A well-conditioned estimator for large-dimensional covariance matrices”. In: *Journal of Multivariate Analysis* 88.2 (2004), pp. 365–411.

[44] Olivier Ledoit and Michael Wolf. “Improved estimation of the covariance matrix of stock returns with an application to portfolio selection”. In: *Journal of Empirical Finance* 10.5 (2003), pp. 603–621.

[45] Lily Y Liu, Andrew J Patton, and Kevin Sheppard. “Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes”. In: *Journal of Econometrics* 187.1 (2015), pp. 293–311.

[46] Yutong Lu, Gesine Reinert, and Mihai Cucuringu. “Trade co-occurrence, trade flow decomposition, and conditional order imbalance in equity markets”. In: *arXiv preprint arXiv:2209.10334* (2022).

[47] J MacQueen. “Classification and analysis of multivariate observations”. In: *5th Berkeley Symp. Math. Statist. Probability.* University of California Los Angeles LA USA. 1967, pp. 281–297.

[48] Rosario N Mantegna. “Hierarchical structure in financial markets”. In: *The European Physical Journal B-Condensed Matter and Complex Systems* 11.1 (1999), pp. 193–197.

[49] Nathan Mantel. “The detection of disease clustering and a generalized regression approach”. In: *Cancer research* 27.2_Part_1 (1967), pp. 209–220.

[50] Harry Markowitz. “Portfolio selection”. In: *The Journal of Finance* 7.1 (1952), pp. 77–91. ISSN: 00221082, 15406261. (Visited on 01/21/2023).

[51] Gautier Marti et al. “A review of two decades of correlations, hierarchies, networks and clustering in financial markets”. In: *Progress in Information Geometry* (2021), pp. 245–274.

[52] Mark McDonald et al. “Detecting a currency’s dominance or dependence using foreign exchange network trees”. In: *Physical Review E* 72.4 (2005), p. 046106.

[53] Ali Namaki et al. “Network analysis of a financial market based on genuine correlation and threshold method”. In: *Physica A: Statistical Mechanics and its Applications* 390.21-22 (2011), pp. 3835–3841.

[54] Andrew Y Ng, Michael I Jordan, and Yair Weiss. “On spectral clustering: Analysis and an algorithm”. In: *Advances in Neural Information Processing Systems*. 2002, pp. 849–856.

[55] Chun-Xiao Nie. “Dynamics of cluster structure in financial correlation matrix”. In: *Chaos, Solitons & Fractals* 104 (2017), pp. 835–840.

[56] Paolo Pasquariello and Clara Vega. “Strategic cross-trading in the US stock market”. In: *Review of Finance* 19.1 (2015), pp. 229–282.

[57] Vasiliki Plerou et al. “A random matrix theory approach to financial cross-correlations”. In: *Physica A: Statistical Mechanics and its Applications* 287.3-4 (2000), pp. 374–382.

[58] Stephen A Ross. “The arbitrage theory of capital asset pricing”. In: *Journal of Economic Theory* 13.3 (1976), pp. 341–360.

[59] Michael Schneider and Fabrizio Lillo. “Cross-impact and no-dynamic-arbitrage”. In: *Quantitative Finance* 19.1 (2019), pp. 137–154.

[60] William F Sharpe. “Capital asset prices: A theory of market equilibrium under conditions of risk”. In: *The Journal of Finance* 19.3 (1964), pp. 425–442.

[61] William F Sharpe. “The sharpe ratio”. In: *Journal of Portfolio Management* 21.1 (1994), pp. 49–58.

[62] Jianbo Shi and Jitendra Malik. “Normalized cuts and image segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), pp. 888–905.

[63] Michele Tumminello et al. “A tool for filtering information in complex systems”. In: *Proceedings of the National Academy of Sciences* 102.30 (2005), pp. 10421–10426.

[64] Vincent Van Kervel and Albert J Menkveld. “High-frequency trading around large institutional orders”. In: *The Journal of Finance* 74.3 (2019), pp. 1091–1137.

[65] Ulrike Von Luxburg. “A tutorial on spectral clustering”. In: *Statistics and Computing* 17.4 (2007), pp. 395–416.

[66] Liyan Yang and Haoxiang Zhu. “Back-running: Seeking and hiding fundamental information in order flows”. In: *The Review of Financial Studies* 33.4 (2020), pp. 1484–1533.

A Spectral Clustering

In this section, we describe the spectral clustering algorithm used to cluster stocks based on the co-trading matrices in Section 5.1.

We represent an undirected graph $G = (V, E)$, where $V = \{v_1, v_2, ..., v_N\}$ is a collection of N vertices which are data points and E is a set of edges, by its (weighted) adjacency matrix $A \in \mathbb{R}^{N \times N}$. The degree matrix of A, denoted as D, is the diagonal matrix with entries

$$D_{ii} = \sum_{j=1, j \neq i}^{N} A_{ij}.$$

The graph Laplacian L is defined as

$$L = D - A.$$

We use the degree matrix to normalize L, and the normalized version is

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}},$$

which contains which contains the information of graph connectivity. Then we perform K-means clustering on the matrix of eigenvectors corresponding to the smallest K eigenvalues of L_{sym}. Here, K is the pre-selected number of clusters. The procedures are summarized in Algorithm 1.

Algorithm 1 Spectral Clustering

```
    Input: An $N \times N$ similarity matrix $A$, number of clusters $K$.
    Output: Clusters $C_1, C_2, ..., C_K$
procedure SPECTRAL_CLUSTERING($A, K$)
    Compute normalized Laplacian $L_{sym}$
    Compute the eigenvectors $v_1, v_2, ..., v_K$ corresponding to K smallest eigenvalues of $L_{sym}$
    Construct matrix $Q \in \mathbb{R}^{N \times K}$ with $v_1, v_2, ..., v_K$ as columns
    Form matrix $\tilde{Q} \in \mathbb{R}^{N \times K}$ by normalizing row vectors of Q to norm 1
    Apply K-means clustering, with k-means++ (Arthur and Vassilvitskii (2006)) for random initialization, to assign rows of $\tilde{Q}$ to clusters $C_1, C_2, ..., C_K$
```

B Random Initialization

Figure 11 plots the average daily ARIs against time. We observe that the values of average daily ARIs are high, for all numbers of clusters, with acceptable level of variation over the entire period of study. Hence, our clustering analysis on co-trading matrices is robust to random initialization.

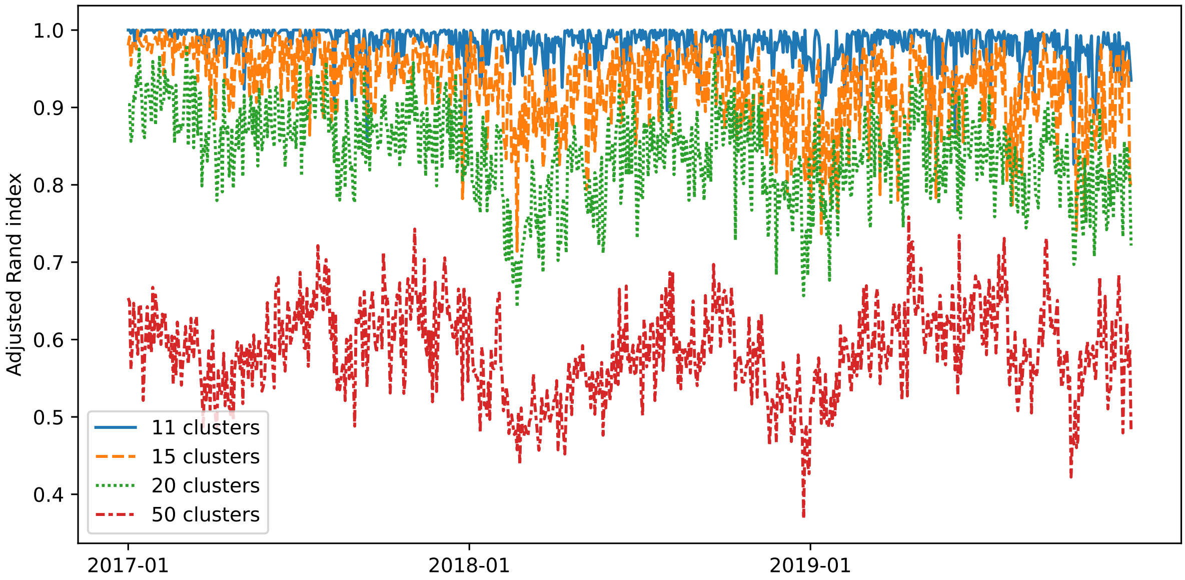

Figure 11: Robustness of applying spectral clustering on daily co-trading networks. This figure plots the daily means of ARIs between each pair of clusters obtained by running the spectral clustering method 100 times on the same co-trading matrix every day, with different random initialization. Each line corresponds to a value of number of clusters, chosen from 11, 15, 20 and 50.

C Volume Measure

In this section, we define co-trading in terms of volume of trades. As an analogue to Section 3.2, we begin with define

$$V^{d^j \to d^i}_{t,j \to i} = \sum_{x_k \in S_t^{i,d^i}} \; \sum_{x_l \in \{x_a \in \mathcal{N}^{\delta}_{x_k} | s_a = j, d_a = d^j\}} q_l,$$

where S_t^{i,d^i} is the set of all filtered trades and q_l is the size of trade x_l.

Then, the pairwise volume co-trading score between stock i and stock j on day t, using volumes of co-occurred trades for stock i and j with direction d^j and d^i, respectively, is defined as

$$c^{\delta, d^i, d^j}_{t,i,j} := \frac{V^{d^i \to d^j}_{t,i \to j} + V^{d^j \to d^i}_{t,j \to i}}{\sqrt{\sum_{x_l \in S_t^{i,d^i}} q_l} \sqrt{\sum_{x_m \in S_t^{j,d^j}} q_m}}.$$

Similarly, incorporating volumes, a pairwise co-occurrence count index is determined by summing up volumes of co-occurred trades of a pair of stocks and normalizing with their total volumes. In line with the article, we set δ is set to 500 milliseconds.

Finally, we concatenate the pairwise co-trading scores to be the co-trading matrix in volume measure. We repeat the analysis and summarize the results in Table 7. The volume measured co-trading matrices are robust, however, underperform those measured in count of trades.

Table 7: Summary of analysis on volume based co-trading matrices.
This table reports the annualized Sharpe ratios of GMV portfolios constructed by solving (5) based on different covariance matrices estimates. The out-of-sample backtests span the period from 2017-01-03 to 2019-12-09. The 'Factor' column specifies the number of latent factors while decomposing the sample covariance matrices. The 'GICS' and 'Count-20' columns indicate GMV portfolios corresponding to GICS and clustering count based co-trading matrices with 20 clusters as benchmarks. For volume based co-trading matrices, we use 15, 20 and 50 clusters while imposing diagonal block structure on the residual covariance matrices.

Factor	Cluster				
	GICS	Count-20	15	20	50
1	0.45	0.69	0.10	0.65	0.09
3	0.59	1.12	0.36	1.00	0.63
5	0.69	1.23	0.20	0.99	0.80
10	0.97	1.40	0.06	1.07	0.83