Detection and clustering of lead-lag networks for multivariate time series with an application to financial markets

Stefanos Bennett
University of Oxford
The Alan Turing Institute
stefanos.bennett@stats.ac.uk

Mihai Cucuringu
University of Oxford
The Alan Turing Institute
mihai.cucuringu@stats.ox.ac.uk

Gesine Reinert
University of Oxford
The Alan Turing Institute
reinert@stats.ox.ac.uk

ABSTRACT

In this paper, we propose a method for the detection of lead-lag clusters in multivariate time series, using a pairwise lead-lag metric and a directed network clustering algorithm. We demonstrate that the latent network of pairwise lead-lag relationships between time series can be helpfully construed as a directed network, for which there exists a suitable algorithm for the detection of pairs of lead-lag clusters with high pairwise imbalance. Our method is able to detect statistically significant lead-lag clusters in our primary domain of study, the US equity market. We study the nature of these clusters in the context of the empirical finance literature on lead-lag relations.

CCS CONCEPTS

• **Mathematics of Computing -> Graph Algorithms**;

KEYWORDS

Multivariate time series, lead-lag relationship, clustering, financial markets, flow imbalance in directed graphs.

ACM Reference Format:
Stefanos Bennett, Mihai Cucuringu, and Gesine Reinert. 2021. Detection and clustering of lead-lag networks for multivariate time series with an application to financial markets. In *MileTS '21: 7th KDD Workshop on Mining and Learning from Time Series, August 14th, 2021, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The study of lead-lag relationships in multivariate time series systems is of interest in a number of fields [37], including earth sciences [24], biology [40], and economics [46, 52]. In this paper, we examine ensembles of pairwise lead-lag relationships in time series data, through the lens of directed network analysis. Our specific interest lies in discovering clusters of variables that exhibit strong lead-lag behaviour. To this end, we leverage recently developed algorithms [14] that identify clusters with high imbalance in the flow of weighted edges between pairs of clusters.

While we expect our method to be applicable to a number of multivariate time series domains, the particular application domain of interest in this study is the analysis of lead-lag clusters in financial time series data. Large financial markets, such as the US equity market, exhibit complex non-linear behaviour [12]. By using pairwise lead-lag detection and the tools of network analysis, we aim to extract clusters that capture the latent lead-lag relationships inherent in such complex systems. This approach may lead to insights into the nature of lead-lag relationships in a system such as the US stock market. Furthermore, persistent historical clusterings can be utilised for the challenging task of returns forecasting. Our unsupervised learning method may prove to be a valuable component in high-dimensional forecasting problems.

Key contributions. Our primary contribution is the introduction of a method for the previously unaddressed problem of extracting clusters of leading and lagging time series in multivariate time series systems. Secondly, our empirical analysis yields the first data-driven lead-lag clustering of the US equity market and provides insights into its structure. Thirdly, in the Supplementary Information (SI) section A.2, we introduce a benchmark data generating process for multivariate time series systems with clustered lead-lag structure, which is used to evaluate our method[1].

2 PROBLEM SETTING

In the context of our study of multivariate time series systems, values of a time series α may be associated with future values of time series β with some strength of association $I(\alpha, \beta)$ which will be defined later. Similarly, we can define the strength of association $I(\beta, \alpha)$. We say that time series α leads time series β if $I(\alpha, \beta) > I(\beta, \alpha)$. Otherwise, we say that α lags β. There are a number of ways to mathematically define and extract the pairwise lead-lag relationship between time series. In Section 4, we propose a specific metric that is suited to our application domain and performs well in the synthetic data experiments.

Once we have chosen a metric to capture lead-lag relations, we can represent the relations using a directed weighted network. The nodes of our network correspond to different time series. A directed edge $\alpha \to \beta$ exists between nodes α and β if α leads β. The weight of this edge is given by the magnitude of the pairwise lead-lag metric. We are thus able to study the properties of lead-lag relationships using the tools of network analysis.

A key question in network analysis concerns community detection. Does there exist a clustering of nodes such that node similarity is stronger within clusters than between clusters? In the context of a directed network encoding lead-lag relations, the question of community detection can be interestingly framed in terms of identifying clusters that have high pairwise imbalance. We regard the flow along a directed weighted edge $\alpha \to \beta$ as a measure of how

[1] All code is available at https://github.com/stefanosbennett/kdd-time-series

MileTS '21, August 14th, 2021, Singapore

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

much α leads β. A high cut imbalance between communities $\mathcal{A}$ and $\mathcal{B}$ means that variables in $\mathcal{A}$ are, on average, leaders of variables in $\mathcal{B}$. Therefore, by identifying pairs of clusters with high imbalance, we segment our multivariate system into communities that are, taken in pairs, mostly composed of either leaders or laggers. In Section 4, we describe a specific directed network clustering algorithm that is suited for this task.

The application domain studied in this paper is that of financial time series. In this application, each time series corresponds to the return time series for a particular financial instrument. We investigate the lead-lag cluster structure of the US equity universe. In particular, we are interested in three questions. Does there exist a statistically significant cluster structure in US equities? What is the nature of the data-driven clustering? How does the data-driven cluster structure relate to previously discovered lead-lag mechanisms?

3 RELATED WORK

There exists substantial evidence of lead-lag relations at the scale of monthly, weekly and daily financial returns [2, 7, 9, 10, 29, 33] as well as at higher frequencies [15, 16, 26, 52]. In addition, a number of papers have considered lead-lag relations from the point of view of networks [5, 15, 20, 28, 41, 51, 53, 54]. Commonly studied questions in this financial lead-lag network literature concern the cluster structure of the lead-lag network [4, 5, 28, 41, 53, 55]. A number of papers consider the relative influence of different industry sectors within the lead-lag network [4, 28, 55]. The influence of various sub-sectors within the lead-lag network of financial institutions is also a particular question of concern [5, 41, 53]. The effect of geography-based clusters has also been investigated [41].

In addition to the literature on financial *lead-lag* correlation networks, there is also substantial literature on *synchronous* correlation networks [32, 34, 42, 50]. The reader is referred to Marti et al. [32], for an extensive review of clustering on (mostly) synchronous financial correlation networks.

Our empirical analysis is novel within the financial lead-lag literature since it is the first work to extract a data-driven clustering of the lead-lag network. In contrast, previous studies [4, 5, 28, 41, 53, 55] are only able to capture the influence of pre-defined groups[2] within the financial lead-lag network. We believe that the academic interest in our data-driven clustering approach is underscored by the plurality of papers [32] that apply data-driven clustering to the synchronous correlation network, as well as the number of papers that apply data-driven ranking methods to the lead-lag network [3, 5, 28, 47, 54].

4 METHOD

Our method is a pipeline consisting of three steps. First, we apply a pairwise lead-lag metric to capture the lead-lag relationship between each pair of time series; this results in a network of lead-lag relationships. Second, we apply a directed network clustering method to extract a partition of the multivariate system such that there is a large flow imbalance (net sum of weights of inter-cluster edges [14]) between cluster pairs. The third step leverages this clustering to identify the *leadingness* of a cluster.

[2]that are defined, for instance, by industry sector [4, 28, 55] or geography [41]

There are a number of choices for each of these components of our pipeline. In SI A.2, we evaluate different metrics that can be used to quantify lead-lag relations between pairs of time series and possible directed network clustering methods, under synthetic data experiments. On the basis of a-priori considerations and performance under the synthetic data experiments, a lead-lag metric that computes distance correlation [49] between the shifted time series and a directed clustering method that uses the spectrum of a Hermitian adjacency matrix are suitable components for the application of our method to US equity returns. We outline the implementation of our method using these components.

A lead-lag adjacency matrix. Let X_t^i denote the random value of the time series variable $i \in \{1, ..., p\}$ at time $t = 0, \ldots, T$. Further, define the first differences $Y_t^i = X_t^i - X_{t-1}^i$ for $i \in \{1, ..., p\}$, $t = 1, \ldots, T$. In our application domain of US equities, X_t^i denotes the logarithm of the closing price for stock $i \in \{0, 1, ..., p\}$ on day $t = 0, \ldots, T$. Hence Y_t^i provides the corresponding log-return for equity i from day $t-1$ to t.

First, we define the sample cross-correlation function between time series i and j evaluated at lag $l \in \mathbb{Z}$ to be

$$\mathrm{ccf}^{ij}(l) = \mathrm{corr}\left(\{Y_{t-l}^i\}, \{Y_t^j\}\right), \tag{1}$$

where corr denotes the distance correlation [49]; more choices of correlation measures are given in SI A.1.1. Next, the lead-lag metric, a measure of the extent to which i leads j, is obtained by applying a functional, which we denote **ccf-auc**, that computes the signed normalised area under the curve (auc) of the cross-correlation function (ccf). Mathematically, this amounts to

$$S_{ij} = \frac{\mathrm{sign}(I(i,j) - I(j,i)) \cdot \max(I(i,j), I(j,i))}{I(i,j) + I(j,i)}, \tag{2}$$

where $I(i,j) = \sum_{l=1}^{L} \left|\mathrm{corr}\left(\{Y_{t-l}^i\}, \{Y_t^j\}\right)\right|$ for a user-specified maximum lag L. Thus S_{ij} is designed to quantify how much time series variable i leads j. The value S_{ij} satisfies $S_{ij} = -S_{ji}$, rendering the matrix skew-symmetric. Our method is able to detect general non-linear dependencies across multiple lags $l \in \{-L, \ldots, L\}$.

Clustering the lead-lag adjacency matrix. Define the asymmetric *lead-lag adjacency matrix* matrix $A_{ij} = \max(S_{ij}, 0)$ which encodes the leading relationships between all pairs of time series. We apply the directed network clustering algorithm of Cucuringu et al. [14], to the weighted and directed network G, where each node corresponds to a time series variable and the adjacency matrix is A. This algorithm for clustering directed networks considers the spectrum of the complex matrix $\tilde{A} \in \mathbb{C}^{p \times p}$ derived from the directed network adjacency matrix as $\tilde{A} = i(A - A^T)$. Since $\tilde{A}$ is Hermitian, it has a real spectrum which can be used to extract an eigenvector embedding that is amenable to clustering. The Hermitian clustering algorithm is particularly suited to our setting of clustering lead-lag networks since we aim to extract pairs of clusters with high flow imbalance [14]. In addition, as a pre-processing step for this algorithm, we apply random-walk normalisation to the adjacency matrix $\tilde{A}$ so that the method is robust to heterogeneous degree distributions [14]; we refer to the resulting algorithm as the **Hermitian RW** algorithm. Details of this algorithm as well as alternative clustering algorithms can be found in SI A.1.2.

The leadingness metric. We introduce a *meta-flow graph* in order to capture the aggregate weighted flow between pairs of clusters. The total "*flow*" between any two clusters is given by the normalised sum of the signed weights between all edges directed from one cluster to another. The skew-symmetric matrix that encodes this information is dubbed the *meta-flow* matrix F with entries

$$F_{ij} = \frac{1}{|C_i||C_j|} \sum_{l \in C_i, m \in C_j} [A_{lm} - A_{ml}], \tag{3}$$

where C_a denotes the set of all nodes in cluster $a \in \{1, \ldots, k\}$, and $i, j \in \{1, \ldots, k\}$, $i \neq j$. The diagonal of F consists of zeros: $F_{ii} = 0$, $\forall i \in 1, \ldots, k$. We also define a metric for the *leadingness* of each cluster $i \in \{1, \ldots, k\}$, $L(i)$, as follows

$$L(i) := \frac{1}{|C_i|} \sum_{l \in C_i, m \in \{1, \ldots, p\}} [A_{lm} - A_{ml}]. \tag{4}$$

Thus, $L(i)$ averages the row-sums of the skew-symmetric matrix $A - A^T$ for nodes within the cluster i; the row sums of the lead-lag matrix provide a measure of the total tendency of the equity corresponding to the row to be a leader [25]. From this metric, we obtain a ranking of the clusters from the most leading cluster (largest row-sum value), which we will label 0, to the most lagging cluster (smallest row-sum value), which has the greatest label. The row-sums [22, 25] algorithm is an instance of a ranking method that recovers a latent ordering of variables given pairwise observations. There exists a number of alternative ranking algorithms [6, 13, 17, 21, 35] that can be used for defining the leadingness of a cluster.

5 US EQUITY DATA EXPERIMENT

It is well known that US equity returns have a cross-sectional factor structure [18]. Some of the prominent factors, for example the factors representing industry membership, exhibit cluster membership. This induces a clustering structure in the synchronous cross-sectional equity returns [19]. In addition to this synchronous clustering structure, we conjecture that there exists a clustering structure in US equities due to inter-temporal relations in equity returns. In this section, we apply our method to construct and cluster a lead-lag network on a US equity universe, and investigate the resulting data-driven clustering. We will use **ccf-auc** with lags $l \in \{-5, \ldots, 5\}$ with the distance correlation as our lead-lag metric, and Hermitian RW clustering as our clustering step. This method is able to capture non-linear lead-lag relations between returns on the scale of up to a week. We set the number of clusters, a hyper-parameter of our algorithm, to 10 in order to facilitate comparison with the industry-sector clustering of equities.

Data description. We consider a universe of 434 equities spanning from 04-01-2000 to 31-12-2019 in Wharton's CRSP database [1]. Our data gathering and pre-processing steps reduce the risk of detecting spurious lead-lag relations that are due to non-trading effects; the details of these steps can be found in SI A.3.1. The data consist of daily closing prices from which we compute daily log-returns.

Figure 1 (left) shows a sorted skew-symmetric lead-lag matrix encoding the measurement between each pair of stocks. A block structure is apparent, with the last block being a highly lagging cluster.

Testing statistical significance for lead-lag clusters using a permutation test. We test whether there is a statistically significant time

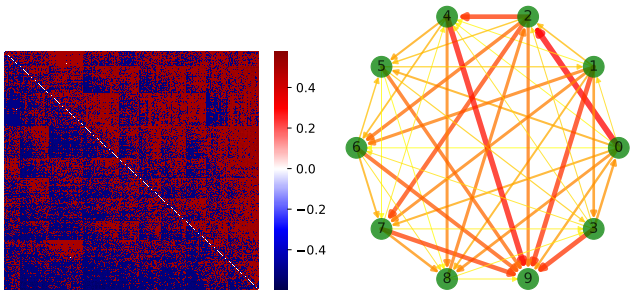

Figure 1: Left: Heatmap of the double-sorted lead-lag matrix $A - A^T$; positive entries in the matrix correspond to a leading relationship between the stock depicted on the vertical axis with respect to the stock depicted on the horizontal axis. Right: Meta-flow network for Hermitian RW clusters; clusters are represented by nodes and larger edge weights are depicted by bolder lines.

dependence in daily US equity returns using a permutation test on the spectrum of the Hermitian adjancency matrix $\tilde{A} = i(A - A^T)$. Under the null hypothesis that there is no time dependence, the ordering of the rows of the daily returns matrix $Y \in \mathbb{R}^{T \times p}$ is drawn uniformly at random from the set of all permutations on $\{1, \ldots, T\}$, $\sigma \in S_T$. Therefore, under the hypothesis of no time dependence, the spectrum of the observed lead-lag matrix should be consistent with the distribution over the spectra of row-permuted matrices $\tilde{A}_{\sigma(t),j}, t = 1, \ldots, T, j = 1, \ldots, p$. Since lead-lag cluster structure is associated with the largest eigenvalues of the Hermitian matrix $\tilde{A}$ [14], our permutation test statistic is set to be the largest eigenvalue of $\tilde{A}$. We use 200 Monte Carlo samples from the null distribution. Under the null hypothesis, the Monte Carlo probability that the largest eigenvalue is greater than or equal to the observed largest eigenvalue is 1/201. We thus reject the null hypothesis with p-value $p < 0.005$, and conclude that there is significant temporal structure in US equity markets.

Note that a rejection of the null implies either

(1) Significant auto-correlation,
(2) Significant cross-correlation,
(3) Some combination of (1) and (2).

However, since our test statistic is a summary statistic of the lead-lag matrix spectrum, which encodes cross-correlations between time series and relates to the clustering structure [14], a rejection of the null suggests that there is significant cluster structure in the lead-lag matrix.

Comparing data-driven clustering with known lead-lag mechanisms. We investigate whether our data-driven lead-lag extraction and clustering results can be explained by three potential mechanisms explored in the financial lead-lag literature.

(1) Sector membership induces clustered lead-lag effects. Bieley et al. [4] find associations between sector membership and lead-lag structure on the high-frequency scale of returns.
(2) Equities with higher trading volume are hypothesised to lead lower volume equities. The disparities in trading volume across equities can lead to non-synchronous trading lead-lag effects [8, 9]. Clustering structure may be induced

by ordering equities based on quantiles of average trading volume.

(3) Larger capitalisation equities are hypothesised to lead lower capitalisation equities [29]. This market capitalisation mechanism can produce lead-lag effects partly via non-trading effects and partly via other channels [8]. Conrad et al. [11] also find that large stocks may lead small stocks via volatility spillovers. Clustering structure may be induced by ordering equities based on quantiles of market capitalisation.

Comparison of data-driven clustering with industry membership clustering. We compute the Jaccard similarity coefficient between the data-driven Hermitian RW clustering and the clustering due to industry membership. We use the first level of the Standard Industrial Classification (SIC) [1] code for the firm corresponding to each equity in order to assign the equity to an industry. Table 1 in SI A.3.2 counts the number of equities that are a member of each SIC sector. Figure 2 displays the Jaccard similarity between each pair of Hermitian RW and industry clusters.

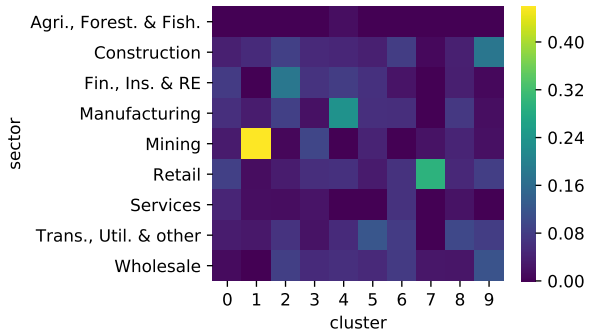

Figure 2: The Jaccard similarity coefficient between Hermitian RW and industry clusters (SIC)

Overall, noting the low value of the Jaccard similarity coefficient between the Hermitian RW and SIC industry clusters, we see that we cannot fully explain the Hermitian RW clustering in terms of the SIC industry sectors. However, there does appear to be some association between certain SIC sectors and Hermitian RW clusters.

We see that the Mining sector seems to be strongly associated with cluster 1 (the second most leading cluster). The Finance, Insurance and Real Estate sector is also associated with a relatively leading cluster (cluster 2). These observations are consistent with the findings of Biely et al. [4] that the finance and energy sectors have strong participation in the significant eigenvalues of the lead-lag matrix[3]. Xia et al. [55] also find that the financial and real estate sectors are associated with leading equities in the Chinese equity market. These associations between SIC code and Hermitian RW membership provide a partial interpretation for the links of the meta-flow network corresponding to the Hermitian RW clustering. The meta-flow network is depicted in Figure 1 (right). For example, we see that one of the strongest flows is from cluster 4 to 9 – which are associated with Manufacturing and Construction respectively.

[3]While Biely and co-authors [4] use GICS sector classification in their analysis, the GICS Energy sector has substantial overlap with the Mining SIC sector.

Figure 7 of SI A.3.2 displays a histogram of the edge weights of the two meta-flow networks: one corresponding to Hermitian RW clustering and the other corresponding to SIC clustering. The data-driven Hermitian RW clustering results in larger flow between pairs of clusters than an industry-based clustering. This demonstrates the efficacy of our method in retrieving pairs of clusters with high flow imbalance.

Comparing data-driven clustering with market capitalisation and volume-based explanations. We also find that more leading clusters (clusters labelled $0-3$) do not appear to have larger average daily dollar volume or market capitalisation[4]. These results are not consistent with the hypotheses that a cluster's tendency to lead is positively associated with the trading volume or market capitalisation of its constituents.

Therefore, the results obtained by our data-driven clustering method cannot be explained by the three previously hypothesised mechanisms outlined in this section. Our novel method may prove to be useful in the exploration of novel lead-lag mechanisms in the empirical finance community.

Time-variation in clusters. To investigate the time-variation in the clustering obtained through our method, we recompute the clustering year-by-year using only data from the retrospective year to do so. In order to compare the similarity in clusterings across time, we calculate the Adjusted Rand Index (ARI) between each pair of yearly clusterings. The results are illustrated in Figure 10 of SI A.3.4. The relatively low ARI values between pairs of clusters indicates a low persistence in year-to-year lead-lag structure. Further, higher ARI values occur in earlier years: this suggests that there is a decrease in persistence between clusterings as time increases. This agrees with the observation in the work of Curme et al. [15] that the informational efficiency of the market appears to increase in 2012 relative to earlier years.

6 CONCLUSION

We propose a method for the previously unaddressed problem of data-driven detection of leading and lagging time series clusters. Our method captures general, non-linear lead-lag correlations and leverages a state-of-the-art directed network clustering algorithm which is able to detect clusters with high flow imbalance. When applied to US equity data, our method produces a clustering that cannot be explained by three prominent lead-lag hypotheses; this suggests that our method is useful for the exploration of novel lead-lag mechanisms in the discipline of empirical finance. In ongoing work, we find that our method can be used for challenging downstream forecasting tasks in noisy, high-dimensional settings. In addition to the application domain of finance, our method may be used in domains – such as economics, medicine and earth sciences – that are characterised by large multivariate time series data. Finally, our network approach to time series, which is able to infer global clustering structure based on local pairwise interactions, can be applied to general pairwise directed interaction data between time series variables. Thus, our framework may be generalised beyond *lead-lag* interactions, in order to discover cluster structure in high-dimensional time-series systems based on *general* directed interactions.

[4]The details of our methodology and results can be found in SI A.3.3

REFERENCES

[1] (2020). Center For Research in Security Prices. Graduate School of Business. University of Chicago. Retrieved from Wharton Research Data Service.

[2] S. G. Badrinath, R. Kale Jayant, and H. Noe Thomas. 1995. Of Shepards, Sheep and the cross-autocorrelations in equity returns. *The Review of Financial Studies* 8, 2 (1995).

[3] Lasko Basnarkov, Viktor Stojkoski, Zoran Utkovski, and Ljupco Kocarev. 2019. Lead-lag Relationships in Foreign Exchange Markets. *arXiv* (2019). https://doi.org/10.1016/j.physa.2019.122986 arXiv:1906.10388

[4] Christoly Biely and Stefan Thurner. 2008. Random matrix ensembles of time-lagged correlation matrices: Derivation of eigenvalue spectra and analysis of financial time-series. *Quantitative Finance* 8, 7 (2008), 705–722. https://doi.org/10.1080/14697680701691477 arXiv:physics/0609053

[5] Monica Billio, Mila Getmansky, Andrew W. Lo, and Loriana Pelizzon. 2012. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics* 104, 3 (2012), 535–559. https://doi.org/10.1016/j.jfineco.2011.12.010

[6] R. A. Bradley and M. E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* (1952), 324–345.

[7] Michael J . Brennan, Jegadeesh Narasimhan, and Bhaskaran Swaminathan. 1993. Investment Analysis and the Adjustment of Stock Prices to Common Information Source. *The Review of Financial Studies* 6, 4 (1993), 799–824.

[8] John Y. Campbell, Andrew W. Lo, and A. Craig MacKinlay. 1997. The econometrics of financial markets. *The Econometrics of Financial Markets* (1997). https://doi.org/10.1515/9781400830213-004

[9] Tarun Chordia and Bhaskaran Swaminathan. 2000. Trading Volume and Cross-Autocorrelations in Stock Returns. *The Journal of Finance* LV, 2 (2000), 913–935.

[10] Lauren Cohen and Andrea Frazzini. 2008. Economic links and predictable returns. *Journal of Finance* 63, 4 (2008), 1977–2011. https://doi.org/10.1111/j.1540-6261.2008.01379.x

[11] Jennifer Conrad, Mustafa Gultekin, and Gautam Kaul. 1991. Asymmetric Predictability of Conditional Variances. *The Review of Financial Studies* 4, 4 (1991), 597–622.

[12] Rama Cont. 2001. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance* 1, 2 (2001), 223–236. https://doi.org/10.1080/713665670

[13] M. Cucuringu. 2016. Sync-Rank: Robust Ranking, Constrained Ranking and Rank Aggregation via Eigenvector and Semidefinite Programming Synchronization. *IEEE Transactions on Network Science and Engineering* 3, 1 (2016), 58–79.

[14] Mihai Cucuringu, Huan Li, He Sun, and Luca Zanetti. 2020. Hermitian matrices for clustering directed graphs: insights and applications. *AISTATS* c (2020), 1–19. arXiv:1908.02096 http://arxiv.org/abs/1908.02096

[15] Chester Curme, Michele Tumminello, Rosario N. Mantegna, H. Eugene Stanley, and Dror Y. Kenett. 2015. Emergence of statistically validated financial intraday lead-lag relationships. *Quantitative Finance* 15, 8 (2015), 1375–1386. https://doi.org/10.1080/14697688.2015.1032545 arXiv:1401.0462

[16] Chester Curme, Michele Tumminello, Rosario N. Mantegna, H. Eugene Stanley, and Dror Y. Kenett. 2015. How Lead-Lag Correlations Affect the Intraday Pattern of Collective Stock Dynamics. *SSRN Electronic Journal* (2015). https://doi.org/10.2139/ssrn.2648490

[17] Caterina De Bacco, Daniel B. Larremore, and Cristopher Moore. 2018. A physical model for efficient ranking in networks. *Science Advances* 4, 7 (2018).

[18] Eugene F Fama and Kenneth R French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* (1993). https://doi.org/10.2469/dig.v36.n3.4225

[19] James Farrell. 1974. Analyzing Covariation of Returns to Determine Homogeneous Stock Groupings. *Journal of Business* 47, 2 (1974), 186–207.

[20] Paweł Fiedor. 2014. Information-theoretic approach to lead-lag effect on financial markets. *European Physical Journal B* 87, 8 (2014). https://doi.org/10.1140/epjb/e2014-50108-3 arXiv:1402.3820

[21] Fajwel Fogel, Alexandre d'Aspremont, and Milan Vojnovic. 2016. Spectral ranking using seriation. *Journal of Machine Learning Research* 17, 88 (2016), 1–45.

[22] David F Gleich and Lek-heng Lim. 2011. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 60–68.

[23] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13 (2012), 723–773.

[24] A. Harzallah and R. Sadourny. 1997. Observed lead-lag relationships between Indian summer monsoon and some meteorological variables. *Climate Dynamics* 13, 9 (1997), 635–648. https://doi.org/10.1007/s003820050187

[25] Peter J Huber. 1962. Pairwise Comparison and Ranking: Optimum Properties of the Row Sum Procedure. *The Annals of Mathematical Statistics* (1962).

[26] Nicolas Huth. 2012. High Frequency Lead / lag Relationships Empirical facts. *Journal of Empirical Finance* 26, march 2014 (2012), 41–58. http://www.sciencedirect.com/science/article/pii/S0927539814000048

[27] M G Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1 (1938), 81–93.

[28] Chang Liao, Yinfei Huang, Xibin Shi, and Xin Jin. 2014. Mining influence in evolving entities: A study on stock market. *DSAA 2014 - Proceedings of the 2014 IEEE International Conference on Data Science and Advanced Analytics* (2014), 244–250. https://doi.org/10.1109/DSAA.2014.7058080

[29] Andrew W. Lo and A. Craig MacKinlay. 1990. When are Contrarian Profits Due to Stock Market Overreaction. , 175–205 pages.

[30] B. G. Malkiel and Eugen Fama. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance* 25, 2 (1970). https://doi.org/10.2307/2325488

[31] Gautier Marti, Sebastien Andler, Frank Nielsen, and Philippe Donnat. 2016. Exploring and measuring non-linear correlations: Copulas, Lightspeed Transportation and Clustering. *arXiv* (2016). arXiv:1610.09659 http://arxiv.org/abs/1610.09659

[32] Gautier Marti, Frank Nielsen, Mikołaj Bińkowski, and Philippe Donnat. 2019. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. *arXiv* (2019), 1–34. arXiv:1703.00485 http://arxiv.org/abs/1703.00485

[33] Lior Menzly and Oguzhan Ozbas. 2010. Market segmentation and cross-predictability of returns. *Journal of Finance* 65, 4 (2010), 1555–1580. https://doi.org/10.1111/j.1540-6261.2010.01578.x

[34] A Namaki, A H Shirazi, R Raei, and G R Jafari. 2011. Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and its Applications* 390, 21-22 (2011), 3835–3841. https://doi.org/10.1016/j.physa.2011.06.033

[35] L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference*. 161–172.

[36] William Pentney and Marina Meila. 2005. Spectral clustering of biological sequence data. *Proceedings of the National Conference on Artificial Intelligence* 2 (2005), 845–850.

[37] B. Podobnik, D. Wang, D. Horvatic, I. Grosse, and H. E. Stanley. 2010. Time-lag cross-correlations in collective phenomena. *EPL* 90, 68001 (2010). https://doi.org/10.1209/0295-5075/90/68001

[38] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. 2011. Detecting Novel Associations in Large Datasets. *Science* 334, 6062 (2011), 1518–1524. https://doi.org/10.1126/science.1205438.Detecting

[39] Karl Rohe, Tai Qin, and Bin Yu. 2016. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences of the United States of America* 113, 45 (2016), 12679–12684. https://doi.org/10.1073/pnas.1525793113

[40] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. 2017. Detecting causal associations in large nonlinear time series datasets. *arXiv* November (2017).

[41] Leonidas Sandoval. 2014. Structure of a Global Network of financial companies based on transfer entropy. *Entropy* 16, 8 (2014), 4443–4482. https://doi.org/10.3390/e16084443

[42] Leonidas Sandoval and Italo De Paula Franca. 2012. Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and its Applications* 391, 1-2 (2012), 187–208. https://doi.org/10.1016/j.physa.2011.07.023 arXiv:1102.1339

[43] Venu Satuluri and Srinivasan Parthasarathy. 2011. Symmetrizations for clustering directed graphs. *ACM International Conference Proceeding Series* i (2011), 343–354. https://doi.org/10.1145/1951365.1951407

[44] Jianbo Shi and Jitendra Malik. 2000. Normalized Cuts and Image Segmentation. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 22, 8 (2000). https://doi.org/10.1109/ICIP.2014.7025680

[45] Ali Shojaie and Emily B. Fox. 2021. Granger Causality: A Review and Recent Advances. *arXiv* (2021). arXiv:2105.02675 http://arxiv.org/abs/2105.02675

[46] Didier Sornette and Wei-Xing Zhou. 2011. Non-Parametric Determination of Real-Time Lag Structure between Two Time Series: The 'Optimal Thermal Causal Path' Method. *SSRN Electronic Journal* August 2004 (2011), 1–37. https://doi.org/10.2139/ssrn.576822

[47] Stavros Stavroglou, Athanasios Pantelous, Kimmo Soramaki, and Konstantin Zuev. 2017. Causality networks of financial assets. *The Journal of Network Theory in Finance* 3, 2 (2017), 17–67. https://doi.org/10.21314/jntf.2017.029

[48] Daniel L. Sussman, Minh Tang, Donniell E. Fishkind, and Carey E. Priebe. 2012. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* 107, 499 (2012), 1119–1128. https://doi.org/10.1080/01621459.2012.699795 arXiv:1108.2228

[49] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *Annals of Statistics* 35, 6 (2007), 2769–2794. https://doi.org/10.1214/009053607000000505

[50] Michele Tumminello, Fabrizio Lillo, and Rosario N. Mantegna. 2010. Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior and Organization* 75, 1 (2010), 40–58. https://doi.org/10.1016/j.jebo.2010.01.004 arXiv:0809.4615

[51] Tomáš Výrost, Štefan Lyócsa, and Eduard Baumöhl. 2015. Granger causality stock market networks: Temporal proximity and preferential attachment. *Physica A: Statistical Mechanics and its Applications* 427 (2015), 262–276. https://doi.org/10.1016/j.physa.2015.02.017

[52] Donghua Wang, Jingqing Tu, Xiaohui Chang, and Saiping Li. 2017. The lead–lag relationship between the spot and futures markets in China. *Quantitative Finance* 17, 9 (2017), 1447–1456. https://doi.org/10.1080/14697688.2016.1264616

[53] Gang Jin Wang, Chi Xie, Kaijian He, and H. Eugene Stanley. 2017. Extreme risk spillover network: application to financial institutions. *Quantitative Finance* 17, 9 (2017), 1417–1433. https://doi.org/10.1080/14697688.2016.1272762

[54] Di Wu, Yiping Ke, Jeffrey Xu Yu, Philip S. Yu, and Lei Chen. 2010. Detecting leaders from correlated time series. *International Conference on Database Systems for Advanced Applications* 5981 LNCS (2010), 352–367. https://doi.org/10.1007/978-3-642-12026-8_28

[55] Lisi Xia, Daming You, Xin Jiang, and Wei Chen. 2018. Emergence and temporal structure of Lead–Lag correlations in collective stock dynamics. *Physica A: Statistical Mechanics and its Applications* 502 (2018), 545–553. https://doi.org/10.1016/j.physa.2018.02.112

A SUPPLEMENTARY MATERIAL

A.1 Description of the method components

A.1.1 Pairwise metrics of lead-lag relationship. In a complex, non-linear system such as the US stock market, determining a suitable way to define a metric to capture lead-lag relationships is a challenging task.

Lead-lag metrics based on a functionals of the cross-correlation. A commonly used approach to defining a lead-lag metric is to use a functional of the cross-correlation function (ccf) between two time series. As in (1), the sample cross-correlation function between time series i and j evaluated at lag $l \in \mathbb{Z}$ is given by

$$\text{ccf}^{ij}(l) = \text{corr}\left(\{Y^i_{t-l}\}, \{Y^j_t\}\right),$$

where corr denotes a choice of sample correlation function. Generalising equation (2), the lead-lag metric, a measure of the extent to which i leads j, is then obtained by

$$S_{ij} = F(\text{ccf}^{ij}),$$

where F is a suitable functional. Concretely, we consider four choices for the sample correlation function corr

(1) **Pearson linear correlation**
(2) **Kendall rank correlation** [27]
(3) **Distance correlation** [49]
(4) **Mutual information** based on discretised time series values [20]

Further, we consider two choices for the functional F as follows

(1) **ccf-lag1**: computes the difference of the cross-correlation function at lag $l \in \{-1, 1\}$

$$S_{ij} = \text{ccf}^{ij}(1) - \text{ccf}^{ij}(-1),$$

(2) **ccf-auc**: computes the signed normalised area under the curve (auc) of the cross-correlation function

$$S_{ij} = \frac{\text{sign}(I(i,j) - I(j,i)) \cdot \max(I(i,j), I(j,i))}{I(i,j) + I(j,i)},$$

where $I(i,j) = \sum_{l=1}^{L} \left|\text{corr}\left(\{Y^i_{t-l}\}, \{Y^j_t\}\right)\right|$ for a user-specified maximum lag L.

The **ccf-lag1** method used with Pearson correlation is a crude lead-lag indicator [8]. This lead-lag indicator is only able to correctly determine the direction of the lead-lag relationship under a positive cross-correlation association between time series. Thus, this lead-lag indicator should be restricted to domains such as US equity returns, where cross-correlations between time series variables are predominantly positive [8].

The **ccf-auc** method accounts for both positive and negative associations across multiple lags $l \in \{-L, \ldots, L\}$. The maximum lag L can be be chosen a-priori as the maximum time lag expected in the multivariate system or by using cross-validation on some downstream validation criterion. The averaging approach **ccf-auc** presented here is similar to Wu and coauthors' [54] lag aggregation methodology.

The four different sample correlation functions are devised to detect different dependencies. Pearson correlation is able to detect linear dependencies, Kendall rank correlation is able to detect monotonic non-linear dependencies, Distance and Mutual information are able to detect general non-linear dependencies. A drawback of such non-linear sample correlation functions is that they have lower power in the case of a true linear relationship.

In total, we consider 8 possible choices for lead-lag metrics based on functionals of the cross-correlation. This comprises 4 possible choices for correlation (Pearson, Kendall, distance and mutual information) and 2 possible choices for the functional form (**ccf-lag1** and **ccf-auc**).

The functional cross-correlation approach is flexible and computationally simple. The flexibility of the framework permits the use of robust and non-linear correlation metrics. The use of non-linear correlation metrics is particularly useful for the extraction of lead-lag relationships in the financial time series domain where linear cross-correlations between returns are expected to be low. High information efficiency in US equity markets [30] implies low linear return cross-correlation. On the other hand, a stylised feature of financial returns is volatility clustering [12]; the size of the cross-correlation between the volatility of returns is expected to be larger than the cross-correlation between the raw returns themselves. A linear cross-correlation approach is unable to capture the relationship between the volatility of two instruments across time. Empirical studies have also found that stronger lead-lag relationships can be detected when taking into account volatility [5]. Thus, when comparing the time-dependence in returns between two assets, we should allow for non-linear effects [20]. In addition, the functional cross-correlation approach easily permits the use of correlation metrics that are robust to outliers. Since financial times series exhibit heavy tails [12], this is an important feature for a lead-lag extraction method.

The linear Granger causality approach that is often considered in financial lead-lag studies [45] can be viewed as an extension of a functional linear cross-correlation-based approach that takes into account autocorrelation and also filters for statistical significance. General Granger causality methods may also use non-linear functional forms to capture the association between time series. These more general methods can be used as the lead-lag extraction component of our method. However, for our purposes of demonstrating our method and using robust and non-linear lead-lag extraction, simpler functional cross-correlation approaches will suffice. Further, in the synthetic data generating processes that we consider in SI A.2 there is no association between past values of a time series and *its own* future values; thus, Granger causality approaches that take into account such association are not relevant in this setting.

Alternative lead-lag metrics. The lead-lag extraction approaches mentioned in this section are not exhaustive. Indeed, alternative methods can be found within the financial time series lead-lag literature [52]. Further, the functional cross-correlation framework presented in this paper is agnostic to the choice of correlation metric used within it. As such, it is able to draw on a wide array of non-linear correlation metrics such as Target/Forget Dependence Coefficient [31], maximal information coefficient [38] or maximum mean discrepancy [23].

A.1.2 Algorithms for clustering directed networks. Let S_{ij} denote the user-defined lead-lag metric that captures how much time series variable i leads j. The value S_{ij} can be positive or negative and satisfies $S_{ij} = -S_{ji}$. Define the asymmetric matrix $A_{ij} = \max(S_{ij}, 0)$

which encodes the leading relationships between each all pairs of time series. We apply directed network clustering algorithms to the weighted and directed network G where each node corresponds to a time series variable and the adjacency matrix is A. In this section, we present different relevant clustering methods for such directed networks. We start with more details on Hermitian random walk clustering, which is the method used in the main part of the paper.

Hermitian clustering. The Hermitian clustering procedure [14] for clustering directed networks considers the spectrum of the complex matrix $\tilde{A} \in \mathbb{C}^{p\times p}$, which is derived from the directed network adjacency matrix as $\tilde{A} = i(A - A^T)$. Since $\tilde{A}$ is Hermitian, it has p real-valued eigenvalues which we order by magnitude $|\lambda_1| \geq \ldots \geq |\lambda_p|$. The eigenvector associated with λ_j is denoted by $g_j \in \mathbb{C}^p$ where $||g_j|| = 1$ for $1 \leq j \leq p$.

Algorithm 1 describes the procedure for clustering the directed graph G. In our implementation, we set $l = k$.

Algorithm 1 Hermitian clustering algorithm

Input: A directed graph $G = (V, E)$ with Hermitian adjacency matrix $\tilde{A}$; number of clusters $k \geq 2$; $\epsilon > 0$

(1) Compute all the eigenvalue/eigenvector pairs of A $\{(\lambda_1, g_1), (\lambda_2, g_2), \ldots, (\lambda_l, g_l)\}$ satisfying $|\lambda_j| > \epsilon$
(2) $P \leftarrow \sum_{j=1}^{l} g_j g_j^T$
(3) Apply a k-means algorithm with input rows of P
(4) Return a partition of V corresponding to the output of k-means

Cucuringu et al. [14] study the performance of the algorithm theoretically and experimentally under data generated by a directed version of the stochastic block model that embeds latent structure in terms of flow imbalance between clusters. They show that the algorithm is able to discover cluster structures based on directed edge imbalance. This contrasts with previous spectral clustering methods which detect clusters based purely on edge-density of symmetrised networks. The Hermitian clustering algorithm is particularly suited to our setting of clustering lead-lag networks since we aim to extract pairs of clusters with high flow imbalance. In addition, as a pre-processing step for this algorithm, we apply random walk normalisation to the adjacency matrix $\tilde{A}$ so that the method is robust to heterogeneous degree distributions [14]; we refer to the resulting algorithm as the **Hermitian RW** algorithm.

In addition to Hermitian clustering, we also considered the following clustering algorithms for directed networks.

Naive symmetrisation clustering. Popular undirected network clustering methods such as spectral clustering [44] cannot be immediately applied to directed networks, since directed networks have complex spectra. Traditional approaches for directed network clustering have applied spectral analysis to a symmetrised version of the directed network adjacency matrix [36, 48]. We consider a commonly used naive symmetrisation-based directed clustering method as a baseline [43]. This naive method applies a standard spectral clustering [44] algorithm to the undirected network with adjacency matrix $\tilde{A} = A + A^T$. In this paper, the spectral clustering algorithm applied to the derived undirected networks uses k-means clustering on a projection onto the first k non-trivial eigenvectors of the random-walk normalised graph Laplacian (we drop the first eigenvector since for connected networks it is always the unit vector). The value of k, corresponding to the desired number of clusters, is a hyperparameter of the algorithm.

Bibliometric symmetrisation clustering. Naive symmetrisation methods produce a clustering that only takes into account edge density and not edge direction. As a result, they are unable to target clusterings with high pairwise flow imbalance between clusters. Satuluri and Parthasarathy [43] propose the degree-discounted bibliometric symmetrisation that is able to take into account edge direction information. Clusters produced by this method are expected to group together nodes that have a relatively large number of parent and children nodes in common [43].

DI-SIM co-clustering. Rohe et al. [39] propose a co-clustering algorithm for directed networks. The co-clustering algorithm first computes a regularised graph Laplacian using A; this initial step is performed so that algorithm may deal with heterogeneous and sparse data. Then, co-clustering is performed by applying k-means on the k-largest of each of the left and right normalised singular vectors of the Laplacian; in the plots, we shall abbreviate these two methods by **DI-SIM-L** and **DI-SIM-R**, respectively.

A.2 Synthetic data experiment

The purpose of this section is to validate our method in synthetic experiments in which the ground truth lead-lag relationships and clusters are known. It will also give an indication of the relative performance of each of our lead-lag metrics and clustering components under different data generating settings.

A.2.1 Synthetic data generating process. We introduce two different lagged latent variable synthetic generating processes to test our method. The general form of these synthetic generating processes is a latent variable model whereby the lagged dependence on the latent variable z induces the clustering amongst the different times series $\{y_t^i\}$. Mathematically, the synthetic data generating processes take the form

$$z_t \overset{i.i.d.}{\sim} F_z \; \forall t \in \{1, \ldots, T\}, \; z_t := 0 \; \forall t \leq 0,$$
$$y_t^i = g_{l_i}(z_{t-l_i}) + \epsilon_t^i, \; \epsilon_t^i \overset{i.i.d.}{\sim} N(0, \sigma^2) \quad \forall t \in \{1, \ldots, T\}, \; i \in \{1, \ldots, p\},$$

where the lag corresponding to time series variable i is $l_i \in L$ and L is the set of lag values. The choice of the shared latent variable distribution F_z and the functional dependencies $g_l, l \in L$ on the latent variable z determines the data generating process. The two particular forms that we will consider are as follows

(1) **Linear**

$$F_z = N(0, 1),$$
$$y_t^i = z_{t-l_i} + \epsilon_t^i,$$

(2) **Cosine**

$$F_z = U(-\pi, \pi),$$
$$y_t^i = \frac{1}{\sqrt{\pi}} \cos(l_i \cdot z_{t-l_i}) + \epsilon_t^i.$$

The factor-based form of the synthetic data generation is motivated by our application to US equity markets [18]; see Section 5 for a discussion of hypothesised clustered lead-lag return structure in the US equity market. The synthetic data generating process considered in this section is a toy model that is designed to test whether our method can correctly detect and cluster time series in a factor-driven scenario.

In these two data generating process scenarios, by design, the cross-covariance at lag $k \in \mathbb{N}$ between any two time series $i, j \in \{1, \ldots, p\}$ is $\mathbb{E}\left[(y^i_{t-k} - \mathbb{E}[y^i_{t-k}])(y^j_t - \mathbb{E}[y^j_t])\right] = 0$ whenever $k \neq l_j - l_i$ due to the independence of z_t across time. In the linear data generating case, setting (1), when $k = l_j - l_i$, then we have that $\mathbb{E}\left[(y^i_{t-k} - \mathbb{E}[y^i_{t-k}])(y^j_t - \mathbb{E}[y^j_t])\right] = \mathbb{E}[(z_{t-l_j})^2] \geq 0$. This induces a linear dependence between time series i and time series j through the single non-zero value in the cross-covariance function between these two time series. Considering the whole network of lead-lag relations, we find that $i \longrightarrow j$ (i is a leader of j) if and only if $l_i < l_j$. Since multiple time series share the same lag, this network is clustered: time series i and j share the same cluster if and only if $l_i = l_j$. Our synthetic experiments test our method's ability to correctly detect lead-lag relationships and recover the underlying ground-truth clustering structure of the lead-lag network.

The non-linear data generating case, setting (2), engenders additional challenges for our lead-lag extraction method. Due to the orthogonality of the cosine functions $\{\cos(mx)\}_{m\in\mathbb{N}}$, the linear cross-covariance evaluated at lag k between two time series i and j is zero even when $k = l_j - l_i$. Thus we expect metrics based on linear cross-covariance methods to perform poorly in these settings[5]. Non-linear lead-lag metrics are required in order to detect the non-linear dependence of time series j on time series i at lag $k = l_j - l_i$.

In our simulation studies, we consider the performance of different configurations of our method as the noise level σ of the idiosyncratic error increases. The following experiment parameter choices are considered

- Number of data points per time series: $T = 250$,
- Number of time series: $p = 100$,
- The standard deviation of the idiosyncratic noise: $\sigma \in \{0, 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 4\}$,
- Latent variable lag dependence for each time series by experiment setting:
  - Linear: $l_i = \lfloor \frac{i-1}{10} \rfloor$ for $i = 1, \ldots, 100$,
  - Cosine: $l_i = \lfloor \frac{i-1}{10} \rfloor + 1$ for $i = 1, \ldots, 100$.

The lag and factor structure implies that there are 10 clusters in each setting. In each configuration of our method, we set the clustering algorithm hyperparameter corresponding to the number of clusters to be equal to the ground truth number of clusters. The remaining hyperparameter choices for the different method configuration components are

- **ccf-auc**: the maximum cross-covariance lag: $L = 5$.
- **Hermitian RW, Naive symmetrisation and Bibliometric clustering**: the number of eigenvectors used in the respective spectral clustering projections is set equal to the ground truth number of clusters.
- **DI-SIM co-clustering**: the regularisation parameter is set equal to the average row sum of the adjacency matrix [39] and the number of singular vectors used in the co-clustering is set equal to the ground truth number of clusters in each synthetic data generating setting.

A.2.2 Performance metrics. We define performance criteria to evaluate both components – the lead-lag detection component and the clustering component – of our method. In order to evaluate the lead-lag detection component of our method, we calculate the proportion of correctly classified edges in the true underlying lead-lag network (i.e. the accuracy of correctly classifying the direction of the lead-lag relationship between two time series). In order to evaluate the clustering component of our method, we calculate the Adjusted Rand Index (ARI) between the ground-truth clustering and the clustering recovered by our method.

A.2.3 Results. We present the results for the lead-lag metric and clustering stages separately. For each experimental setting, we have generated 48 samples from the synthetic data generating process and applied our method to it.

We display the average value and confidence interval for the lead-lag component detection accuracy over the 48 samples in the linear setting in Figure 3, and the cosine setting in Figure 4. The confidence interval is a 95% Gaussian for the accuracy computed on a sample from the data-generating process.

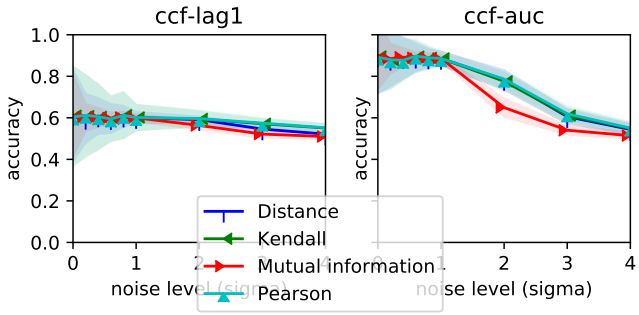

Figure 3: Average and confidence interval for accuracy by lead-lag detection method in the linear setting.

We observe from Figure 3 that the proposed lead-lag detection components are able to detect linear lead-lag associations, and that their performance decreases to random chance performance as the level of noise in the synthetic data experiment increases. The **ccf-auc** method performs better than the **ccf-lag 1** method. Within the **ccf-auc** method, the non-linear Kendall and distance correlation metrics are able to maintain similar performance to the linear metric. The outperformance of the **ccf-auc** method over the **ccf-lag1** method shows the advantage of considering a larger number of lags in the cross-correlation function when pairs of time series depend on each other through large lag values.

[5]Yet, we note that even if the ground-truth cross-covariance is zero, this does not imply that the expected value of the sample Pearson coefficient will be zero due to the denominator of the Pearson coefficient.

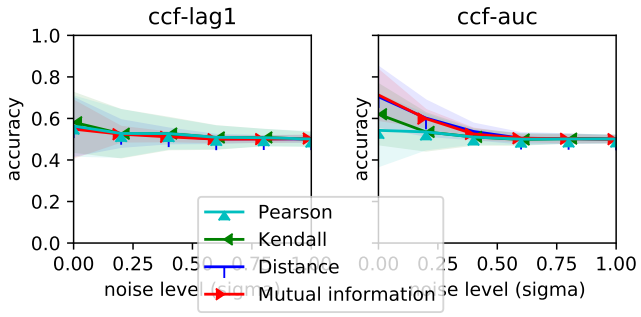

Figure 4: Average and confidence interval for accuracy by lead-lag detection method in the cosine setting.

The performance of the methods is worse in the cosine setting: the noise level at which the performance of all methods drops to that of random chance around $\sigma = 0.5$ (compared with $\sigma = 4$ in the linear setting). The **ccf-lag1** method performs poorly: this is not a surprise since this method cannot deal with negative associations. The **ccf-auc** method using mutual information or distance correlation is able to achieve the highest accuracy; this illustrates the use of methods that are able to take into account negative and non-linear associations.

In order to compare the performance of different clustering methods, we compute, for each clustering method and experimental repetition, the marginal of ARI over the different lead-lag detection metrics. The mean and confidence interval for the ARI values over the experimental repetitions is shown in Figures 5 and 6.

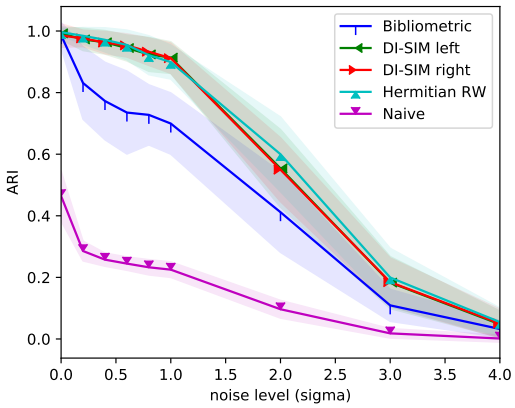

Figure 5: Average and confidence interval for the ARI by clustering method in the linear setting.

We see that the (non-naive) implementations of our method are able to recover almost perfectly (ARI of 1) the clustering in both settings (1) and (2) when σ is low. The performance of our methods decrease as σ increases – the performance in the cosine setting decreases faster than the performance in the linear setting. The Hermitian RW and the DI-SIM clustering methods perform best in the settings considered. The Hermitian RW method targets clusters

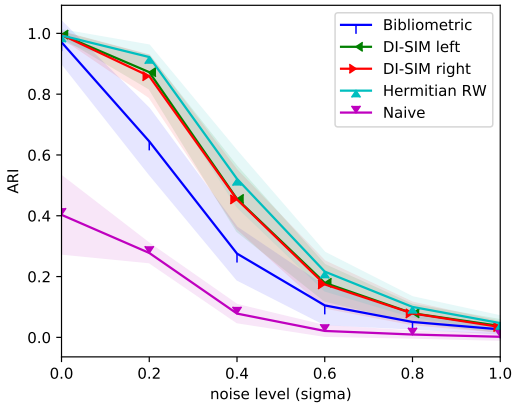

Figure 6: Average and confidence interval for the ARI by clustering method in the cosine setting.

with high imbalance [14] and is therefore particularly suited to the task of clustering time series according to directed imbalances in their lead-lag relations. The importance of edge direction is illustrated by the relatively poor performance of the naive method which relies solely on the magnitude and not the direction of the edges.

We note that even as the number of lags considered in the cross-correlation function by the **ccf-lag1** and **ccf-auc** component method (1 and 5 lags, respectively) is lower than the largest lag dependence between any two pairs of time series (e.g. $l_{100} - l_1 = 9$ in the linear setting), our overall two-stage method using these component methods is still able to leverage enough similarities in the dependence structure between the time series to correctly recover the ground-truth clustering.

To summarise this section, we have validated our pipeline on two synthetic data generating processes. While the choice of particular correlation components should be driven by the application in mind, we find that the **ccf-auc** method using distance correlation achieves relatively strong performance both in the linear and in the cosine synthetic data generating settings. The clustering component methods that were found to perform best were the DI-SIM and Hermitian RW methods.

A.3 US equity data experiment

A.3.1 US equity data preparation description. We consider the universe of 5325 NYSE equities spanning from 04-01-2000 to 31-12-2019 from Wharton's CRSP database [1] – restricting our attention to equities trading on the same exchange to avoid spurious lead-lag effects due to non-synchronous trading [8]. The data consists of daily closing prices from which we compute daily log-returns. We also compute the average daily dollar volume that is traded for each equity. We subset to the equities that have the largest average volume (ranking 500th or better) and the least number of missing values (at least 2.5 years' worth of non-missing data). This results in a data set of 434 equities. Filtering to the most traded equities with the least number of missing prices reduces the risk of spurious lead-lag

effects due to non-synchronous trading [8]. Any remaining missing prices are forward-filled prior to the calculation of log-returns.

A.3.2 Comparing industry and data-driven Hermitian RW clustering.

Table 1 shows the number of equities in each SIC sector. Most sectors have a relatively large number of equities, with *Agriculture, Forestry and Fisheries* and *Services* being quite small.

Retail	90
Manufacturing	67
Construction	66
Mining	58
Trans., Util. & other	54
Fin., Ins. & RE	46
Wholesale	43
Services	9
Agri., Forest. & Fish.	1

Table 1: Number of equities in each SIC sector.

Figure 7 compares the edge weights of the Hermitian RW and SIC clustering meta-flow networks. It illustrates that the Hermitian RW deviates from the SIC clustering and thus captures a clustering with greater net flow imbalance between pairs of clusters.

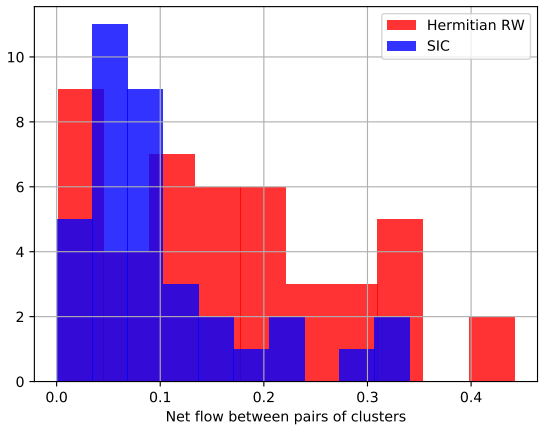

Figure 7: Histogram of Hermitian RW and SIC clustering meta-flow network edge weights. We see that the edge weights for the meta-flow network corresponding to a Hermitian RW clustering tend to be larger than the edge weights for the meta-flow network corresponding to a SIC clustering.

A.3.3 Comparing data-driven clustering with market capitalisation and volume-based explanations. Figures 8 and 9 display the average daily dollar volume and market capitalisation averaged across all stocks in a given cluster.

In order to understand the association between the tendency for an equity to lead and its daily dollar volume or market capitalisation at a sub-cluster level, we compute the Spearman correlation

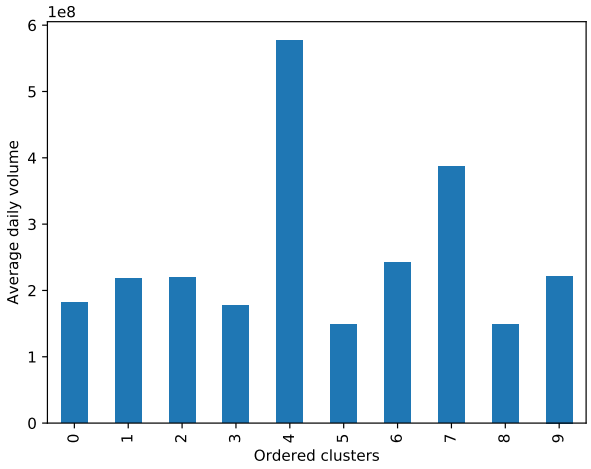

Figure 8: The average daily dollar volume for each Hermitian RW cluster.

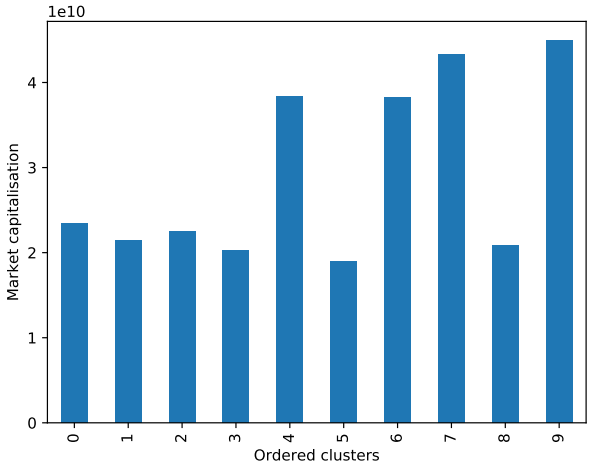

Figure 9: The average market capitalisation for each Hermitian RW cluster.

between the row-sums of the lead-lag matrix – which provides a metric for the tendency of each cluster to lead – and these equity characteristics (trading volume and market capitalisation). This results in a Spearman correlation of 0.01 and -0.15 between the lead-lag row-sums and the equity trading volume and market capitalisation, respectively.

These results are not consistent with a positive association between a cluster's tendency to lead and the trading volume or market capitalisation of its constituents. Therefore, the results obtained by our data-driven clustering method cannot be explained by the previously hypothesised mechanisms outlined in Section 5.

A.3.4 Time-variation in clusters. Figure 10 displays the adjusted Rand index between clusters which are computed on yearly snapshots of the data. The similarity between the partitions is strongest for the first 6 years of the data. This figure shows that while there

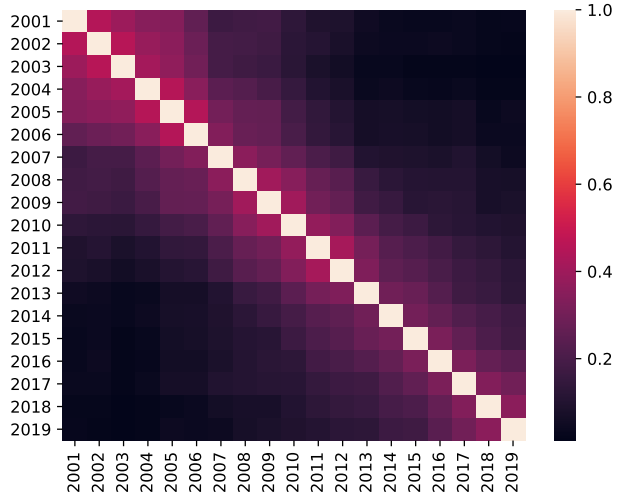

Figure 10: Adjusted Rand index between clusters computed on yearly snapshots of data.

is some similarity between the partitions across adjacent years, the partitions vary considerably over the 20 year period.