



Lead–lag detection and network clustering for multivariate time series with an application to the US equity market

Stefanos Bennett^{1,2} · Mihai Cucuringu^{1,2,3} · Gesine Reinert^{1,2}

Received: 12 December 2021 / Revised: 30 June 2022 / Accepted: 15 September 2022 /
Published online: 1 November 2022
© The Author(s) 2022

Abstract

In multivariate time series systems, it has been observed that certain groups of variables partially lead the evolution of the system, while other variables follow this evolution with a time delay; the result is a lead–lag structure amongst the time series variables. In this paper, we propose a method for the detection of lead–lag clusters of time series in multivariate systems. We demonstrate that the web of pairwise lead–lag relationships between time series can be helpfully construed as a directed network, for which there exist suitable algorithms for the detection of pairs of lead–lag clusters with high pairwise imbalance. Within our framework, we consider a number of choices for the pairwise lead–lag metric and directed network clustering model components. Our framework is validated on both a synthetic generative model for multivariate lead–lag time series systems and daily real-world US equity prices data. We showcase that our method is able to detect statistically significant lead–lag clusters in the US equity market. We study the nature of these clusters in the context of the empirical finance literature on lead–lag relations, and demonstrate how these can be used for the construction of predictive financial signals.

Keywords High-dimensional time series · Unsupervised learning · Lead–lag · Clustering · Financial markets · Directed networks · Flow imbalance

Editor: Joao Gama.

✉ Stefanos Bennett
stefanos.bennett@stats.ox.ac.uk

Mihai Cucuringu
mihai.cucuringu@stats.ox.ac.uk

Gesine Reinert
reinert@stats.ox.ac.uk

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

² The Alan Turing Institute, London, UK

³ Mathematical Institute, University of Oxford, Oxford, UK

1 Introduction

Multivariate time series are ubiquitous in a wide range of domains, such as the physical sciences, medicine, and economics. Often, multivariate systems describing multiple processes or quantities are thought to exhibit lead–lag relationships (Podobnik et al., 2010). In this work, time series A is said to lead time series B if A 's past values are more strongly associated with B 's future values than A 's future values are with B 's past values. The study of lead–lag relationships in multivariate time series systems is of interest in fields such as earth science (Harzallah & Sadourny, 1997), biology (Runge et al., 2019) and economics (Wang et al., 2017a; Sornette & Zhou, 2005). For example, Harzallah and Sadourny (1997) study the lead–lag relationship between the Indian summer monsoon and a number of climate variables such as snow cover, sea surface temperature and geopotential height across a grid of locations on the Earth's surface. Wang et al. (2017a) examine the lead–lag dependence between the spot and futures markets for a Chinese stock market index.

In this paper, we examine systems of lead–lag relationships in time series data through the lens of directed network analysis. By constructing a network based on pairwise lead–lag metrics between variables, we are able to study overall properties of the web of lead–lag relationships via the tools of network analysis. Our specific interest lies in discovering clusters of different variables that exhibit strong lead–lag behaviour. To this end, we employ unsupervised directed network clustering and leverage recently developed algorithms (Cucuringu et al., 2020) that identify clusters with high imbalance in the flow of weighted edges between pairs of clusters.

While we expect our unsupervised learning 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, often with a low signal-to-noise ratio (Cont, 2001). By using pairwise lead–lag detection and network analysis tools, we aim to extract clusters that capture the latent lead–lag relationships which may be present in such complex systems. Furthermore, persistent historical clusterings can be utilised for the challenging task of returns forecasting. As a result, our unsupervised learning method may prove to be a valuable component in certain financial forecasting pipelines. Beyond financial markets, this approach may lead to insights into the nature of lead–lag relationships in climate (Harzallah & Sadourny, 1997), social (Lin et al., 2013), biological (Runge et al., 2019) or economic systems (Iyetomi et al., 2020; Camilleri et al., 2019).

1.1 Problem description

In the context of multivariate time series systems, the problem of lead–lag detection consists in identifying random variables that lead or lag other random variables. There are a number of ways to mathematically define and extract the pairwise relationship between time series. Different lead–lag definitions are compared using a-priori considerations in Sect. 3.1 and synthetic experiments in Sect. 4.

Once we have chosen a metric to capture lead–lag relations, we can represent the uncovered relations using a directed weighted network. The nodes of our network correspond to different time series variables. A directed edge $A \rightarrow B$ exists between nodes A and B if time series A leads time series B . The weight of this edge is given by the magnitude of the

pairwise lead–lag metric, thus encoding the strength of the relation. 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 (Newman, 2018). Does there exist a clustering of nodes such that node similarity is, on average, stronger within clusters than between clusters? In the context of a directed network encoding lead–lag relations, the question of community detection can be framed in terms of identifying clusters that exhibit high pairwise cut imbalance, as follows. We regard the flow along a directed weighted edge $A \rightarrow B$ as a measure of the extent to which A leads B . In a directed graph with adjacency matrix A , the cut associated to two subsets of nodes \mathcal{A} and \mathcal{B} , is given by $Cut(\mathcal{A}, \mathcal{B}) = \sum_{i \in \mathcal{A}, j \in \mathcal{B}} A_{ij}$, and we refer to the difference $Cut(\mathcal{A}, \mathcal{B}) - Cut(\mathcal{B}, \mathcal{A})$ as the *cut imbalance*. A high cut imbalance between communities \mathcal{A} and \mathcal{B} indicates 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, taken in pairs, are mostly composed of either leaders or laggards. In Sect. 3, we describe a Hermitian-based directed network clustering algorithm that is suited for this task following (Cucuringu et al., 2020).

The application domain studied in this paper is that of financial time series. In this domain, 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 market. In particular, we are interested in four questions. Does there exist a statistically significant cluster structure in the US equity market? What is the nature of the data-driven clustering? How does the data-driven cluster structure relate to previously discovered lead–lag mechanisms? Can we leverage our clustering for downstream forecasting purposes?

1.2 Key contributions

Our primary contribution is the introduction of a principled method, which, to the best of our knowledge, is the first to address the problem of unsupervised clustering of leading and lagging variables in multivariate time series systems. We validate different components of our method on synthetic and real data sets. Our secondary contribution consists of an evaluation of novel pairwise lead–lag metrics using a new benchmark data generating process for multivariate time series systems with clustered lead–lag structure. Thirdly, the application of our method to US equity data provides insights into the structure of the US equity market. To the best of our knowledge, our work presents the first data-driven clustering of lead–lag networks in a financial market context. Finally, we construct a novel statistically significant trading signal for the US equity market—thus demonstrating how our method can be employed to extract valuable signals in the high-dimensional, low signal-to-noise data setting.

1.3 Paper outline

We discuss existing literature related to our work in Sect. 2. Section 3 describes our approach to solving the lead–lag extraction and clustering problems. In Sect. 4, we validate our method on synthetic data sets. We present the results of applying our algorithm to a universe of US equities in Sect. 5. In Sect. 6, we illustrate the use of our methodology in a financial forecasting application. Finally, we summarise our main findings in Sect. 7.

2 Related work

There exists substantial evidence of lead–lag relations at the scale of monthly, weekly and daily financial returns (Lo & MacKinlay, 1990; Badrinath et al., 1995; Brennan et al., 1993; Chordia & Swaminathan, 2000; Menzly & Ozbas, 2010; Cohen & Frazzini, 2008), as well as at higher frequencies (Huth, 2012; Wang et al., 2017a; Curme et al., 2015b, a). In addition, a number of studies have considered lead–lag relations from the point of view of networks (Curme et al., 2015a; Fiedor, 2014; Výrost et al., 2015; Liao et al., 2014; Sandoval, 2014; Billio et al., 2012; Wang et al., 2017b; Wu et al., 2010). Commonly studied questions in this financial lead–lag network literature concern the cluster structure of the lead–lag network (Sandoval, 2014; Billio et al., 2012; Liao et al., 2014; Wang et al., 2017b; Xia et al., 2018; Biely & Thurner, 2008). A number of papers consider the relative influence of different industry sectors within the lead–lag network (Biely & Thurner, 2008; Liao et al., 2014; Xia et al., 2018). The influence of various sub-sectors within the lead–lag network of financial institutions is also a particular question of concern (Billio et al., 2012; Wang et al., 2017b; Sandoval, 2014). For example, Billio et al. (2012) relate the lead–lag network to the systemic exposure of financial firms and sub-sectors, in order to understand their respective financial drawdowns during crisis periods. In addition, the effect of geography-based clusters has also been investigated (Sandoval, 2014).

A second commonly studied problem in the financial lead–lag literature is that of ranking. A number of lead–lag network papers focus on how network tools may be used to identify financial instruments that exhibit stronger tendencies to lead other instruments (Liao et al., 2014; Billio et al., 2012; Wu et al., 2010; Basnarkov et al., 2019; Stavroglou et al., 2017). For example, Wu et al. (2010) and Basnarkov et al. (2019), apply the PageRank algorithm (Google, 2012) to the lead–lag network in order to extract an ordering of equities in terms of their influence on the future values of other equities.

In addition to the literature on financial *lead–lag* correlation networks, there is also substantial literature on *synchronous* correlation networks (Tumminello et al., 2010; Namaki et al., 2011; Sandoval & Franca, 2012; Marti et al., 2019). The reader is referred to Marti et al. (2019) 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 (Sandoval, 2014; Billio et al., 2012; Liao et al., 2014; Sandoval, 2014; Wang et al., 2017b; Xia et al., 2018; Biely & Thurner, 2008) are only able to capture the influence of predefined groups, which are given, for instance, by industry sector (Biely & Thurner, 2008; Liao et al., 2014; Xia et al., 2018) or geography (Sandoval, 2014), 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 (Marti et al., 2019) that apply data-driven clustering to synchronous correlation networks, as well as the number of papers that apply data-driven ranking methods to lead–lag networks (Liao et al., 2014; Billio et al., 2012; Wu et al., 2010; Basnarkov et al., 2019; Stavroglou et al., 2017). Furthermore, our work is the first to show that clustered lead–lag network structure can be successfully used for downstream out-of-sample prediction tasks.

3 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 (Cucuringu et al., 2020)] between cluster pairs. The third step quantifies the *leadingness* of each cluster.

There are a number of choices for each of these components in our pipeline. In this section, we describe metrics that can be used to quantify lead–lag relations between pairs of time series, and available directed network clustering methods.

To introduce notation, let X_t^i denote the random value of the time series variable $i \in \{1, \dots, p\}$ at time $t = 0, \dots, T$. Further, define the first differences $Y_t^i = X_t^i - X_{t-1}^i$ for $i \in \{1, \dots, p\}, t = 0, \dots, T$.

In our application domain of US equities, X_t^i denotes the logarithm of the closing price for stock $i \in \{1, \dots, p\}$ on day $t = 0, \dots, T$. Hence Y_t^i provides the corresponding log-return for equity i from day $t - 1$ to t . It is suitable to use log-returns for analysis as they exhibit closer to stationary properties, and log-returns are more mathematically tractable than linear or percentage returns in the computation of multi-horizon returns (Campbell et al., 1997).

3.1 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 challenging. Here we present some options.

3.1.1 Lead–lag metrics based on a functional of the cross-correlation

A commonly used approach to defining a lead–lag metric is to use a functional of a sample cross correlation function (ccf) between two time series. The general form of a *sample cross-correlation function* between time series i and j evaluated at lag $l \in \mathbb{Z}$ is given by

$$\text{CCF}^{ij}(l) = \text{corr}\left(\{Y_{t-l}^i\}, \{Y_t^j\}\right), \quad (1)$$

where corr denotes a choice of sample correlation function. The corresponding *lead–lag metric*, a measure of the extent to which i leads j , is then obtained by

$$S_{ij} = F(\text{CCF}^{ij}), \quad (2)$$

where F is a suitable functional.

In this paper, we consider four choices for the sample correlation function corr , namely Pearson linear correlation, Kendall rank correlation (Kendall, 1938), distance correlation (Székely et al., 2007), and mutual Information based on discretised time series values (Fiedor, 2014). The four different sample correlation functions are able to detect different dependencies. Pearson correlation is able to detect linear dependencies, Kendall rank correlation is able to detect monotonic non-linear dependencies, while distance correlation

and mutual Information are able to detect general non-linear dependencies. The drawback of non-linear sample correlation functions is that they have lower power in the case of a true linear relationship.

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

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

$$S_{ij} = CCF^{ij}(1) - CCF^{ij}(-1),$$

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

The **ccf-lag1** method used with Pearson correlation is a crude lead-lag indicator (Campbell et al., 1997). This lead-lag indicator is only designed to take into account positive cross-correlation. Indeed, like the signatures-based method described further below in Sect. 3.1.2, it 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 (Campbell et al., 1997).

The **ccf-auc** method accounts for both positive and negative associations across multiple lags $l \in \{-L, \dots, L\}$. The maximum lag L can 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 the lag aggregation methodology of Wu et al. (2010).¹

Overall, we consider eight possible choices for lead-lag metrics based on functionals of the cross-correlation. This stems from four possible choices for correlation (Pearson, Kendall, distance correlation and mutual information) and two 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 such 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 (Malkiel & Fama, 1970) implies that linear return cross-correlations are too low to be used to construct trading systems that have expected returns in excess of market equilibrium expected returns. On the other hand, a stylised feature of financial returns is volatility clustering (Cont, 2001); 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

¹ We have also considered using a maximum aggregation approach, and have found similar qualitative results to the averaging-based approach presented in this paper; however, the maximum aggregation approach tends to perform slightly worse than the averaging-based approach.

relationships can be detected when taking into account volatility (Billio et al., 2012). Thus, when comparing the time-dependence in returns between two assets, we should allow for non-linear effects (Fiedor, 2014). In addition, the functional cross-correlation approach easily permits the use of correlation metrics that are robust to outliers. Since financial time series exhibit heavy tails (Cont, 2001), robustness constitutes an important feature for a lead–lag extraction method. In general, the functional cross-correlation component and, consequently, the entire pipeline will be robust to outliers if the choice of correlation metric is robust to outliers. For example, ordinal association correlation metrics such as Kendall correlation guarantee robustness to outliers.

The linear Granger causality approach that is often considered in financial lead–lag studies (Shojaie & Fox, 2021; Skoura, 2019) can be viewed as an extension of our functional linear cross-correlation-based approach that takes into account auto-correlation 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. Following the vector auto-regressive modelling example of Skoura (2019), bi-variate modelling can be used to determine the existence and direction of a lead–lag relation between two pairs of time series. Thus, a vector auto-regressive modelling approach can be used to derive a lead–lag metric and therefore be used as the lead–lag extraction component of our model. For the purposes of demonstrating our lead–lag extraction and clustering method, simpler functional cross-correlation approaches will suffice. Since the combination of data auto-correlation and co-movement can produce lead–lag associations between time series variables using our method, one must be careful not to interpret resulting lead–lag associations as apparent causal influence estimates.

We contrast our approach, which is based on correlation networks, with causality-based approaches that attempt the more difficult problems of recovering the casual network underlying a multivariate time series system (Runge et al., 2019) and quantifying its causal influences (Janzing et al., 2013). For example, whereas Runge et al. (2019) attempt to estimate the causal network underlying the time-lagged dependency structure in a given multivariate time series system, our aim is estimating and clustering the association-network for the multivariate time series system. Association-based approaches are more common in the financial network lead–lag literature (Marti et al., 2019), since financial time series have very noisy returns and exhibit weak lead–lag effect sizes due to the informational efficiency of the market (Malkiel & Fama, 1970). These characteristics of financial returns make the problem of accurately estimating a lead–lag correlation network (let alone the causality network) challenging in itself.

3.1.2 Lead–lag metric based on signatures

The approach of using a functional of the cross-correlation function relies on the user to specify the choice of functional; this choice is not obvious in many cases. In particular, it is difficult to gauge the number of lags to incorporate into our lead–lag metric a-priori. An alternative approach draws on the idea of signatures from rough path theory (Levin et al., 2016), in order to construct a pairwise lead–lag metric. The signature of a continuous path with finite 1-variation (Levin et al., 2016) $X : [a, b] \rightarrow \mathbb{R}^d$, denoted by $S(X)_{a,b}$, is the collection of all the iterated integrals of X , namely $S(X)_{a,b} = (1, S(X)_{a,b}^1, \dots, S(X)_{a,b}^d, S(X)_{a,b}^{1,1}, S(X)_{a,b}^{1,2}, \dots)$, where the iterated integrals are given by

$$S(X)_{a,t}^{i_1, \dots, i_k} = \int_{a < t_k < t} \dots \int_{a < t_1 < t_2} dX_{t_1}^{i_1} \dots dX_{t_k}^{i_k}.$$

Based on the proposal in Levin et al. (2016), the signatures-based pairwise measure of the lead–lag relation between two stocks i and j over the time period $[t - m, t]$ is given by

$$S_{ij}(t - m, t) = \iint_{t-m < u < v < t} (dX^i(u)dX^j(v) - dX^j(u)dX^i(v)). \quad (3)$$

This is the difference in the cross-terms of the second level of the time series signature of the log-prices. Theoretical results in rough path theory (Levin et al., 2016) have established that a signature is essentially unique to the path it describes, and that the truncated signature (i.e. the lower order terms) can efficiently describe the path. Chevyrev and Kormilitzin (2016) provide an interpretation of the signature lead–lag metric (3). The signature lead–lag metric is positive and grows larger whenever increases (resp. decreases) in X^i are followed by increases (resp. decreases) in X^j . If the relative moves of X^i and X^j are in the opposite directions, then the signature lead–lag measure is negative. Note that a downside of this method is that it is not able to tell the difference between

1. $i \rightarrow j$ with negative association,
2. $i \leftarrow j$ with positive association.

As a result, we do not expect the method to perform well when there is significant negative association in the lead–lag data generating process.

When analysing price data observed at discrete time points, we transform the data stream into a piecewise linear continuous path and calculate the second order signatures (Reizenstein & Graham, 2018). From this, we may calculate the lead–lag relation using the difference in second order signature cross-terms (3). We refer the reader to Gyurkó et al. (2014) for additional details on signatures and their application in a financial context, along with an interpretation in terms of second order areas and interplay with lead–lag relationships. In practice, when comparing the signature lead–lag metrics across different pairs of time series, we recommend the normalisation of the price data prior to computation of the lead–lag metric, since the absolute value of the metric is increasing in the volatility of the underlying price series.

3.1.3 Alternative lead–lag metrics

The lead–lag extraction approaches mentioned in this section are by no means exhaustive. Indeed, alternative methods can be found within the financial time series lead–lag literature (Wang et al., 2017a). Furthermore, the functional cross-correlation framework presented in this paper is agnostic to the choice of the 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 (Marti et al., 2016), maximal information coefficient (Reshef et al., 2011) or maximum mean discrepancy (Gretton et al., 2012).

The detection of lead–lag relations can be attempted in the frequency-domain as well as the time-domain (Skoura, 2019). Wavelet techniques do not rely on time series stationarity and have been shown to provide a more nuanced understanding of lead–lag relations when used in conjunction with time-domain analysis (Skoura, 2019). However, the wavelet

coherence approach (Skoura, 2019) does not straightforwardly provide a single lead–lag metric between two time series since this would require a method for aggregating across wavelet location and scale parameters. Further, the wavelet coherence method also takes into account synchronous correlation between two time series: this is not desirable for the computation of a lead–lag relation metric. Further work is required to develop a single lead–lag metric based on wavelet coherence that could be used in our lead–lag extraction and clustering pipeline.

3.2 Algorithms for clustering directed networks

Let S_{ij} denote the user-defined lead–lag metric that quantifies how much time series variable i leads j . The value S_{ij} can be positive or negative, and satisfies $S_{ij} = -S_{ji}$. The lead relationships between all pairs of time series is encoded by the asymmetric matrix $A_{ij} = \max(S_{ij}, 0)$. 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.

Note that as a pre-processing step for any of the clustering methods described below, it is possible to filter the pairwise measurements S_{ij} when constructing the network A . For example, Curme et al. (2015a) apply significance thresholding, whereby an edge exists between two nodes only if the corresponding lead–lag metric is sufficiently large in magnitude.

3.2.1 Naive symmetrisation clustering

Popular undirected network clustering methods, such as spectral clustering (Shi & Malik, 2000), cannot be immediately applied to directed networks, since directed networks with asymmetric adjacency matrices have complex spectra. Traditional approaches for directed network clustering have applied spectral analysis to a symmetrised version of the directed network adjacency matrix (Sussman et al., 2012; Pentney and Meila, 2005). We consider a commonly used naive symmetrisation-based directed clustering method as a baseline (Satuluri & Parthasarathy, 2011). This naive method applies a standard spectral clustering algorithm (Shi & Malik, 2000) 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.

3.2.2 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 (2011) propose the degree-discounted bibliometric symmetrisation that is able to take into account edge direction information. In the degree-discounted bibliometric symmetrisation, spectral clustering is applied to the adjacency matrix

$$\tilde{A} = D_o^{-1/2} A D_i^{-1/2} A^T D_o^{-1/2} + D_i^{-1/2} A^T D_o^{-1/2} A D_i^{-1/2},$$

where D_i is the diagonal matrix of weighted in-degrees and D_o is the diagonal matrix of weighted out-degrees. The degree-discounted bibliometric symmetrisation applies degree-discounting to a symmetrisation that sums the number of common in- and out-links between two pairs of nodes. Therefore, clusters produced by this method are expected to group together nodes that have a relatively large number of parent (sender) and children (receiver) nodes in common (Satuluri & Parthasarathy, 2011). Degree-discounting is a technique that has been found to work well in tasks on graphs with highly skewed degree distributions.

3.2.3 DI-SIM co-clustering

Rohe et al. (2016) 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 the 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. This co-clustering identifies two partitions of nodes: one partition groups together vertices with similar sending behaviour, while the other partition groups together vertices with similar receiving behaviour. In this paper, we denote the clustering obtained using the left singular vectors as **DI-SIM-L** and the clustering obtained using the right singular vectors as **DI-SIM-R**. We consider both choices of clustering in our experiments.

3.2.4 Hermitian clustering

The Hermitian clustering procedure (Cucuringu et al., 2020) 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 $|\lambda_1| \geq \dots \geq |\lambda_p|$. The eigenvector associated with λ_j is denoted by $g_j \in \mathbb{C}^p$ where, in Euclidean norm, $\|g_j\| = 1$ for $1 \leq j \leq p$.

Algorithm 1 describes the procedure for clustering the directed network G . In our implementation, we set the number of top eigenvectors used to $l = k$.

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 \tilde{A}
 $\{(\lambda_1, g_1), (\lambda_2, g_2), \dots, (\lambda_l, g_l)\}$ satisfying $\|g_j\| = 1$ and $|\lambda_j| > \epsilon$, $\forall j \in \{1, \dots, l\}$
 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
-

Note that in practice, for scalability purposes, one can bypass the computation of the entire $n \times n$ matrix P in order to directly cluster using the embedding given by the top l eigenvectors.

Cucuringu et al. (2020) study the performance of the algorithm theoretically and experimentally under data generated from a directed version of a 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 that detect clusters based purely on the 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 (Cucuringu et al., 2020); we refer to the resulting algorithm as the **Hermitian RW** algorithm.

3.3 Alternative clustering algorithms

State-of-the-art modularity clustering algorithms such as the Leiden algorithm (Traag et al., 2019) may be used on directed graphs using a directed modularity metric (Dugué & Perez, 2015). Dugué and Perez (2015) optimise a modularity metric that compares the number of edges within clusters to the expected number of edges under a null model. However, such modularity-based algorithms, which return clusters based on edge density, are not suited to our goal of uncovering clusters of leading and lagging variables based on flow imbalance between clusters.

An adaptation of the Hermitian clustering method has been proposed in Laenen and Sun (2020). The method presented in Laenen and Sun (2020) aims to discover a clustering that maximises a normalised flow metric between communities using the spectrum of a normalised Hermitian Laplacian matrix. As such, it could be used as an alternative to the Hermitian RW clustering algorithm considered in this paper. Also recently, Underwood et al. (2020) proposed an algorithm for clustering weighted directed networks that employs motif-based spectral clustering to uncover flow imbalance relationships between pairs of clusters.

Lastly, we draw attention to a recent approach introduced in He et al. (2021) that extends the Hermitian-based clustering algorithm (Cucuringu et al., 2020). This recent method departs from standard approaches in the literature, and treats edge directionality not as a nuisance but rather as the main latent signal. It does so by introducing a graph neural network framework for obtaining node embeddings for directed networks in a self-supervised manner, while accounting for node-level covariates.

3.4 The leadingness metric

We introduce the concept of 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 net of the normalised weights between all edges directed from one cluster to another. The skew-symmetric matrix that encodes this information is dubbed the *meta-flow* matrix, which we denote by F . Mathematically

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

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

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

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 (Huber, 1962). 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 largest numeric label $k - 1$. In this paper, all data-driven clustering results will be presented using this labelling. The ROWSUM RANKING (Huber, 1962; Gleich & Lim, 2011) algorithm is an instance of a ranking method that recovers a latent ordering of variables given variable pairwise comparisons. There exists a rich literature on ranking from pairwise comparisons. The goal in this literature is to infer the strength ℓ_i , $i = 1, \dots, p$ or ranking of p items given a (potentially incomplete) set of pairwise comparisons which encode a noise proxy for $\ell_i - \ell_j$. Alternative ranking algorithms that could be employed for defining the leadingness of a cluster include (Fogel et al., 2016; Cucuringu, 2016; De Bacco et al., 2018; d’Aspremont et al., 2021; Bradley & Terry, 1952; Page et al., 1998), as well as Chau et al. (2020) for rankings that incorporate any available node level covariates.

3.5 Algorithmic complexity of the method

Let us denote by ψ the cost of the pairwise lead–lag metric of choice. The cost of the lead–lag network construction step amounts to $O(p^2\psi)$, where p is the number of time series. For example, for the linear Pearson correlation $O(\psi) = O(TL)$, where L is the number of lags, and T is the sample size. The cost of a spectral clustering algorithm for k clusters, such as Hermitian clustering (Cucuringu et al., 2020), is $O(kp^2) < O(p^3)$. Therefore, the overall complexity amounts to $O(p^2TL + kp^2)$.

In the large p setting, the above pipeline can become computationally prohibitive. One approach to alleviate this amounts to subsampling m pairs of time series out of the $\binom{p}{2}$ choices. This will lead to a comparison lead–lag matrix with only m nonzero entries; for example, the choice of sampling each edge with probability $\frac{\log p}{p}$ renders $m = O(p \log p)$. Since computing the leading eigenvectors of a sparse matrix via an iterative power method-based approach can be performed in a running time that is linear in the number of nonzero entries in the matrix, this step takes $O(p \log p)$ time. Thus, the approximate method is almost linear in the number of edges in the comparison graph. If the underlying pairwise comparison graph is weakly connected, which in practice will be the case because correlations will not be zero, then a choice of sampling probability of $O(\frac{\log p}{p})$ results in a pairwise comparison graph that is weakly connected with high probability. In such a situation we would expect the clustering in the sampled network to be a reasonable reflection of the clustering in the true network. We refer the reader to Batson et al. (2013) for spectral

algorithms and theoretical considerations of the closely related graph sparsification literature, and to Hu and Lau (2013) for a survey and taxonomy of graph sampling techniques.

4 Synthetic data experiments

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

4.1 Synthetic data generating process

We introduce five 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

$$\begin{aligned} Z_t &\stackrel{i.i.d.}{\sim} F_Z \forall t \in \{1, \dots, T\}, \quad Z_t := 0 \quad \forall t \leq 0, \\ Y_t^i &= g_{l_i}(Z_{t-l_i}) + \epsilon_t^i, \quad \epsilon_t^i \stackrel{i.i.d.}{\sim} N(0, \sigma_\epsilon^2) \forall t \in \{1, \dots, T\}, i \in \{1, \dots, p\}, \end{aligned} \quad (5)$$

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 factor-based form of the synthetic data generation is motivated by our application to US equities (Fama & French, 1993; Jegadeesh & Titman, 1995). For instance, early work by Jegadeesh and Titman (1995) studies a lagged factor model in the context of lead–lag effects. See also Sect. 5 for a discussion of hypothesised clustered lead–lag return structures 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.

The five particular forms of (5) that we will consider are as follows.

1. Linear

$$F_Z = N(0, 1) \quad \text{and} \quad Y_t^i = Z_{t-l_i} + \epsilon_t^i. \quad (6)$$

2. Cosine

$$F_Z = U(-\pi, \pi) \quad \text{and} \quad Y_t^i = \frac{1}{\sqrt{\pi}} \cos(l_i \cdot Z_{t-l_i}) + \epsilon_t^i. \quad (7)$$

3. Legendre

$$F_Z = U(-1, 1) \quad \text{and} \quad Y_t^i = P_{l_i+1}^L(Z_{t-l_i}) + \epsilon_t^i. \quad (8)$$

4. Hermite

$$F_Z = N(0, 1) \quad \text{and} \quad Y_t^i = \frac{1}{\sqrt{l_i!}} P_{l_i+1}^H(Z_{t-l_i}) + \epsilon_t^i. \tag{9}$$

5. *Heterogeneous*

$$Z \in \mathbb{R}^K, F_Z = N_{K \times K}(0, I_{K \times K}) \quad \text{and} \quad Y_t^i = Z_{t-l_i}^{f_i} + \epsilon_t^i. \tag{10}$$

Here, P_l^L and P_l^H are respectively the Legendre polynomial of degree l and the Hermite polynomial of degree l . In the heterogeneous case, the superscript $f_i \in \{1, \dots, K\}$ indicates on which component of the multivariate factor Z the time series i depends.

In these five data generating process scenarios, by design, the cross-covariance at lag $k \in \mathbb{N}$ between any two time series $i, j \in \{1, \dots, p\}$ is

$$\mathbb{E} \left[(Y_{t-k}^i - \mathbb{E}[Y_{t-k}^i])(Y_t^j - \mathbb{E}[Y_t^j]) \right] = 0$$

whenever $k \neq l_j - l_i$ due to the independence of z_t across time. In the linear data generating case, setting (6), when $k = l_j - l_i$, then we have that $\mathbb{E} \left[(Y_{t-k}^i - \mathbb{E}[Y_{t-k}^i])(Y_t^j - \mathbb{E}[Y_t^j]) \right] = \mathbb{E}[(Z_{t-l_i})^2] \geq 0$. 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 $i \rightarrow j$ (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 settings (7)–(9) engender additional challenges for our lead–lag extraction method. Due to the respective orthogonality of the cosine functions $\{\cos(mx)\}_{m \in \mathbb{N}}$, Legendre polynomials $\{P_m^L(x)\}_{m \in \mathbb{N}}$ and Hermite polynomials $\{P_m^H(x)\}_{m \in \mathbb{N}}$, 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 or cross-covariance methods to perform poorly in these settings. Non-linear lead–lag metrics are required in order to detect a non-linear dependence of time series j on time series i at lag $k = l_j - l_i$.

The heterogeneous data generating process setting adds a further independence condition on the relationship between two time series. In this case, the cross-covariance at lag k is nonzero if and only if both $k = l_j - l_i$ and $f_i = f_j$ are satisfied. The additional factor component equality condition implies that time series i and time series j share the same cluster if and only if $l_i = l_j$ and $f_i = f_j$.

In our simulation studies, we consider the performance of different configurations of our method as the noise level σ of our 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 $i = 1, \dots, 100$
 - *Cosine*: $l_i = \lfloor \frac{i-1}{10} \rfloor + 1$ for $i = 1, \dots, 100$

- *Legendre and Hermite*: $l_i = \lfloor \frac{i-1}{10} \rfloor + 2$ for $i = 1, \dots, 100$
- *Heterogeneous*: $f_i = \lfloor \frac{i-1}{50} \rfloor$ for $i = 1, \dots, 100$ while $l_i = \lfloor \frac{i-1}{5} \rfloor$ for $i = 1, \dots, 50$ and $l_i = \lfloor \frac{i-51}{5} \rfloor$ for $i = 51, \dots, 100$.

The lag and factor structure implies that there are 10 clusters in the Linear, Cosine, Legendre and Hermite settings, while in the Heterogeneous setting there are 20 clusters. 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: $L = 5$
- *DI-SIM co-clustering*: the regularisation parameter is set equal to the average row sum of the adjacency matrix (Rohe et al., 2016) 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.
- *Naive, Bibliometric and Hermitian RW clustering*: the number of eigenvectors used in the respective spectral clustering projections is set equal to the ground truth number of clusters.

4.2 Performance metrics

We employ different 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, 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, we calculate the Adjusted Rand Index (ARI) between the ground-truth clustering and the clustering recovered by our method. The Adjusted Rand Index (Hubert & Arabie, 1985; Gates & Ahn, 2017) is a popular metric of success for a clustering algorithm which calculates its propensity to allocate of pairs of nodes that belong to the same (resp. different) cluster in the ground truth partition to the same (resp. different) cluster in the partition recovered by the algorithm.

4.3 Results

4.3.1 Marginal results over lead–lag extraction and clustering

We present the results for the lead–lag metric and clustering stages separately. In this section, we present results for the linear and cosine synthetic data generating settings; results for the other synthetic data generating settings can be found in Appendix Sections A.1 and A.2. For each experimental setting, we have generated 48 samples from the synthetic data generating process and applied our method to each one.

We display the average value and confidence interval for the lead–lag component detection classification accuracy over the 48 samples in the linear setting in Fig. 1 and the cosine setting in Fig. 2. The confidence interval is a 95% Gaussian for the classification accuracy computed on a sample from the data-generating process.

Figure 1 shows that the proposed lead–lag detection components are able to detect linear lead–lag associations and that their performance decreases to random chance performance

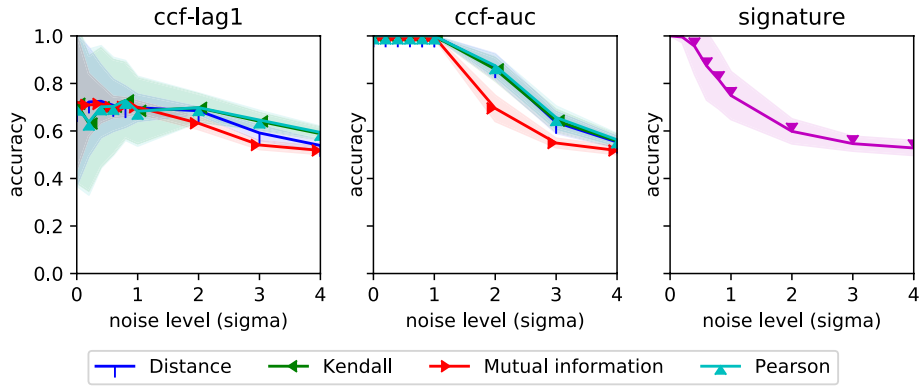


Fig. 1 Average and confidence interval for the classification accuracy by lead–lag detection method in the linear setting

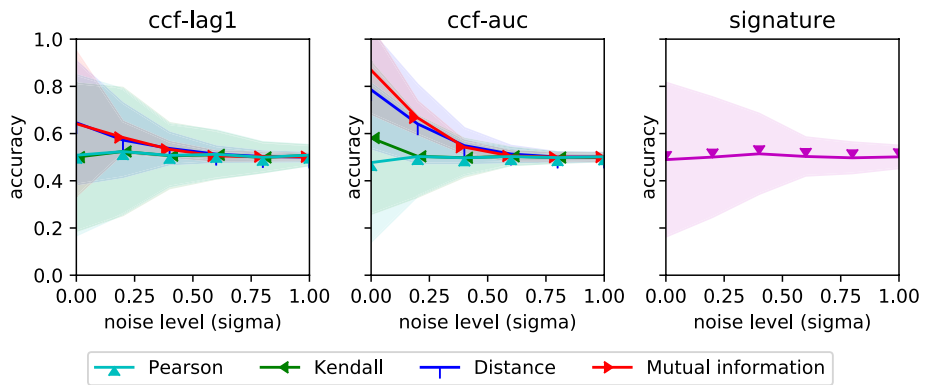


Fig. 2 Average and confidence interval for classification accuracy by lead–lag detection method in the cosine setting

as the level of noise in the synthetic data experiment increases. The **ccf-auc** and signature methods work best in this setting. 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.

The performance of the methods decreases in the cosine setting: the noise level at which the performance of all methods drops to that of random chance is about $\sigma = 0.5$ (compared with $\sigma = 4$ in the linear setting). In particular, the **ccf-lag1** and signature methods perform 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 are shown in Figs. 3 and 4.

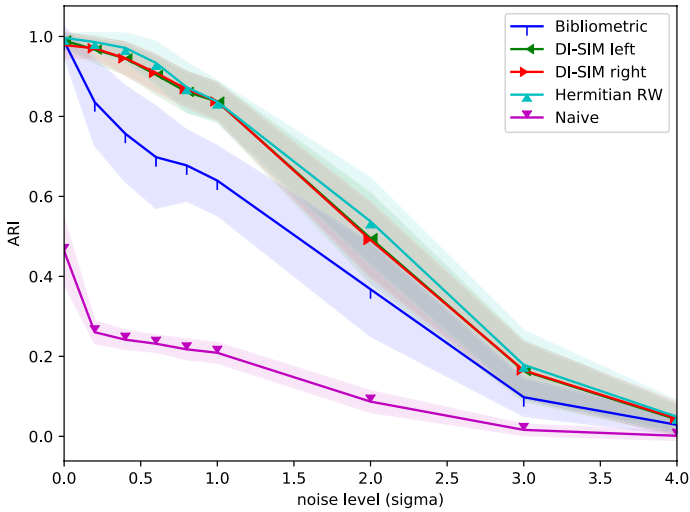


Fig. 3 Average and confidence interval for the ARI by clustering method in the linear setting

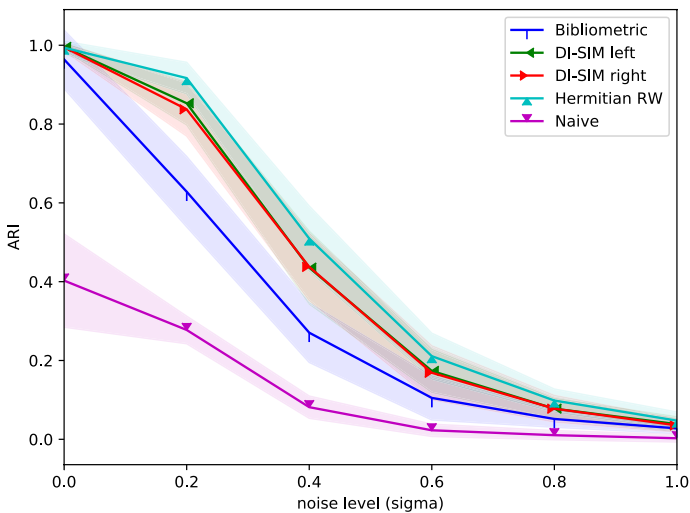


Fig. 4 Average and confidence interval for the ARI by clustering method in the cosine setting

We observe 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. As expected, the performance of our methods decrease as σ increases; the performance in the cosine setting decreases faster than 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 with high imbalance (Cucuringu et al., 2020) 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

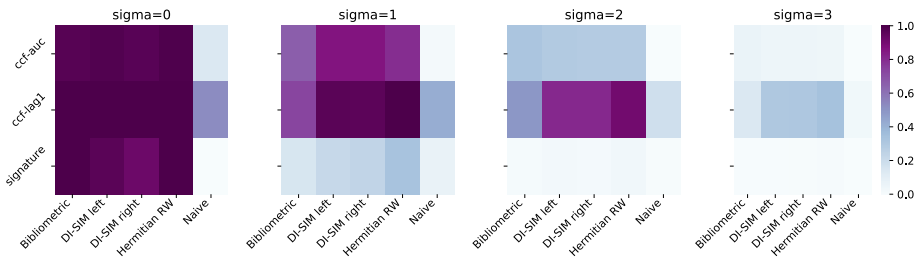


Fig. 5 Average ARI by lead-lag and clustering method in the linear setting

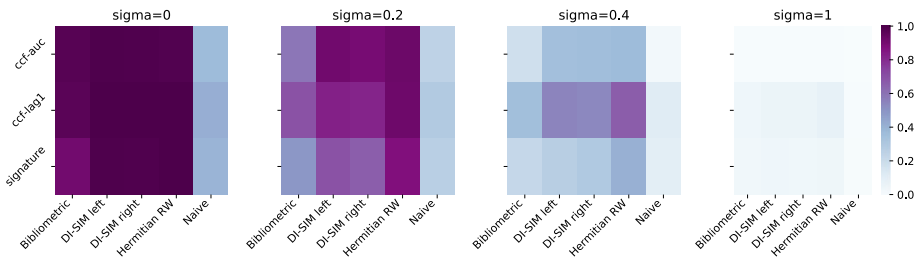


Fig. 6 Average ARI by lead-lag extraction and clustering method in the cosine setting

relatively poor performance of the naive method, which relies solely on the magnitude and not the direction of the edges.

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 pipeline 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. We are able to successfully cluster in this case since $\max_{i \in \{1, \dots, p\}} \min_{j \in \{1, \dots, p\}} |l_i - l_j| = 1$, which is less than or equal to the number of lags considered by the cross-correlation function methods.

Our experimental observations are robust to the other synthetic data generating processes reported in Appendix A. Similar results are also observed when performing simulation studies for a smaller number of time series and smaller sample sizes.

4.3.2 Interaction of lead-lag and clustering components

In this section, we investigate the joint dependence of the pipeline on the lead-lag and clustering components. The performance of the pipeline, measured by ARI averaged over the different Monte Carlo repetitions, is shown for linear and cosine synthetic data settings in Figs. 5 and 6; the other synthetic data settings are presented in Appendix A.3. For each synthetic data setting, we select a range of noise levels σ that are representative of the different levels of overall ARI significance. In these figures, for the pipelines using **ccf-lag1** and **ccf-auc** components, we show the ARI averaged across the 4 different choices of sample correlation function described in Sect. 3.1.1.

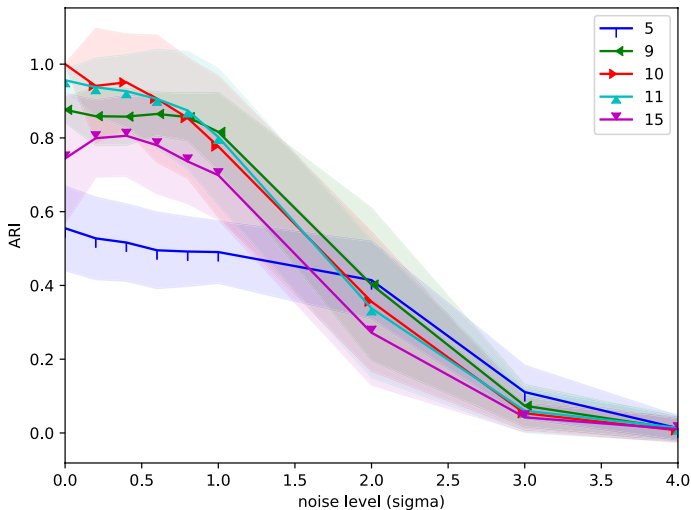


Fig. 7 Average and confidence interval for the ARI by different levels of the hyperparameter corresponding to the number of clusters in the linear setting

We observe that the pipelines that use **ccf-lag1** or **ccf-auc** lead-lag extraction components with **DI-SIM** or **Hermitian RW** as the clustering component tend to perform best. For small values of σ , the performance of each of these pipelines tends to be quite similar. For larger σ values, the relative performance difference between the different lead-lag extraction components tends to increase, with the performance of the DI-SIM and Hermitian RW components within a pipeline using **ccf-lag1** or **ccf-auc** appearing to be quite correlated. Eventually, the performance of every pipeline drops to 0 as σ increases.

4.3.3 Ablation study: varying hyperparameter corresponding to the number of clusters

We perform an ablation study to examine the sensitivity of the pipeline to the hyperparameter controlling the number of clusters returned by the clustering algorithm. In Fig. 7, we present the results for the typical linear synthetic data generating setting using a pipeline of **ccf-auc** with distance correlation and Hermitian RW clustering. Results for the cosine, Legendre and Hermite data settings are shown in the Appendix A.4.

The true underlying number of clusters in the linear data setting is 10 (see Sect. 4). In Fig. 7, we see that the performance of the pipeline is robust to small variations in the hyperparameter corresponding to the number of clusters around the true underlying number of clusters. Further, we find that using a large hyperparameter value for the number of clusters results in a large decay in the ARI of the pipeline.

4.3.4 Summary of synthetic data experiment results

To summarise this section, we have validated our pipeline on five synthetic data generating processes. While the choice of particular correlation components should be driven by the application in mind, 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.

5 US equity data experiment

It is well known that US equity returns exhibit a cross-sectional factor structure (Fama & French, 1993). Some of the prominent factors, for example the factors representing industry membership, can exhibit cluster membership. This induces a clustering structure in the synchronous cross-sectional equity returns (Farrell, 1974). 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, our method is applied to construct and cluster a lead–lag network on a US equity universe, and investigate the resulting data-driven clustering. On the basis of a-priori considerations and performance under the synthetic data experiments, a lead–lag metric that computes distance correlation (Székely et al., 2007) 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 will use **ccf-auc** with lags $l \in \{-5, \dots, 5\}$ with the distance correlation as our lead–lag metric, and Hermitian RW clustering as our clustering step. This method has the potential to capture non-linear lead–lag relations between returns on the scale of up to a week. The range of lag values is a hyperparameter of our method and in general, can be chosen using a-priori considerations or empirically selected using cross validation on a downstream loss. In our case, we set the range of lag values to $l \in \{-5, \dots, 5\}$, which allows us to capture lead–lag relations on daily and weekly scales (see Sect. 2). We set the number of clusters, a hyperparameter of our algorithm, to 10 in order to facilitate comparison with the industry-sector clustering of equities.

5.1 Data description

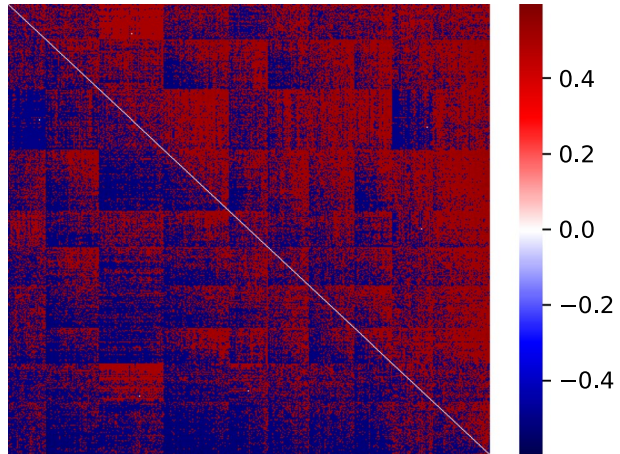
We consider the universe of 5325 NYSE equities spanning from 04-01-2000 to 31-12-2019 from Wharton’s CRSP database (Wharton Research Data Service, 2020)—restricting our attention to equities trading on the same exchange to avoid spurious lead–lag effects due to non-synchronous trading (Campbell et al., 1997). 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 (largest 500 equities in average volume) 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 (Campbell et al., 1997). Any remaining missing prices are forward-filled prior to the calculation of log-returns.

5.2 Data analysis

5.2.1 Illustration of US equity lead–lag matrix

Figure 8 shows a sorted skew-symmetric lead–lag matrix encoding the measurement between each pair of stocks. 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

Fig. 8 Heatmap of the double-sorted lead–lag $p \times p$ matrix $A - A^T$. The rows and columns of the matrix index the $p = 434$ equities, and are categorised by cluster membership [labelled by the leadingness metric (4)]. Within each cluster, we sort the equities by their respective row-sum in $A - A^T$, a proxy for their individual leadingness



the horizontal axis. Similarly, negative values indicate that the horizontal axis stock leads the vertical axis one. The skew-symmetric matrix $A - A^T$ depicted in Fig. 8 is double-sorted by the leadingness metric (4) for each cluster and then, within each cluster, by the rowsum $\sum_{j=1}^p [A_{ij} - A_{ji}]$ of each equity i that is a member of the cluster. A block structure is apparent, with the last block being a highly lagging cluster.

5.2.2 Statistical significance testing for lead–lag clusters

We test whether there is a statistically significant time dependence in daily US equity returns using a permutation test on the spectrum of the Hermitian adjacency 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 $\{1, \dots, T\}$, $\sigma \in S_T$. 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 matrices $\{\tilde{A}_\sigma\}_{\sigma \in S_T}$ computed using row-permuted returns matrices $Y_{\sigma(t),j}, t = 1, \dots, T, j = 1, \dots, p$. Since lead–lag cluster structure is associated with the largest eigenvalues of the Hermitian matrix \tilde{A} (Cucuringu et al., 2020), 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.

It is not possible to resolve the identification issue between these three cases using our method. 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 (Cucuringu et al., 2020), a rejection of the null *suggests* that there is significant

Table 1 Number of equities in each SIC industry sector

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

cluster structure in the lead–lag matrix. Our statistically significant results when using our method for downstream prediction tasks (which relies solely on cross-equity prediction and not auto-correlation) in Sect. 6 provide further evidence for significant clustered lead–lag structure in the US equities.

5.2.3 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 in the empirical finance lead–lag literature.

1. Sector membership induces clustered lead–lag effects. Biely and Thurner (2008) 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 (Chordia & Swaminathan, 2000; Campbell et al., 1997). 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 (Lo & MacKinlay, 1990). This market capitalisation mechanism can produce lead–lag effects partly via non-trading effects and partly via other channels (Campbell et al., 1997). Conrad et al. (1991) 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) (Wharton Research Data Service, 2020) code for the firm corresponding to each equity in order to assign the equity to an industry. Table 1 counts the number of equities that are a member of each SIC sector. Most sectors have a relatively large number of equities, with *Agriculture, Forestry and Fisheries* and *Services* being quite small.

For comparison, the number of equities in each of the Hermitian RW clusters is shown in Table 2. The Hermitian RW algorithm leads to clusters of approximately equal size.

Table 2 Number of equities in each Hermitian RW cluster

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|--|----|----|----|----|----|----|----|----|----|----|
| | 37 | 49 | 57 | 58 | 35 | 35 | 42 | 34 | 32 | 55 |

Cluster ID is shown in the top row

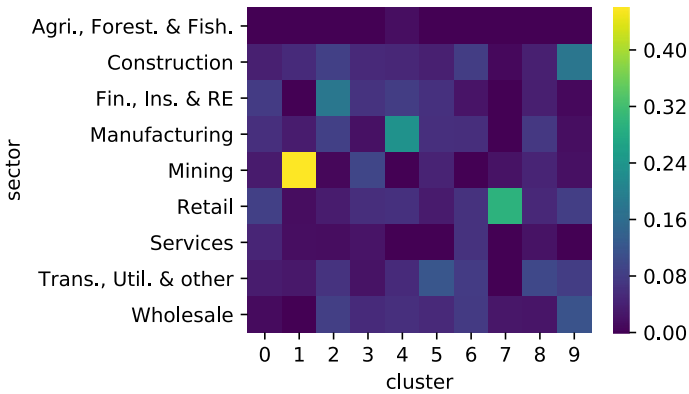


Fig. 9 The Jaccard similarity coefficient between the Hermitian RW clusters and industry clusters (SIC)

Figure 9 displays the Jaccard similarity between each pair of Hermitian RW and industry clusters. Overall, given the low values of the Jaccard similarity coefficients, the clustering seems to recover a structure that goes beyond simple industry sectors.

However, there does appear to be some association between certain SIC sectors and Hermitian RW clusters. We observe 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 and Thurner (2008) that the finance and energy sectors have strong participation in the significant eigenvalues of the lead–lag matrix.² Xia et al. (2018) 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 Fig. 10. For example, we see that one of the strongest flows is from cluster 4 to 9—which are associated with Manufacturing and Construction respectively.

Figure 11 displays a histogram of the edge weights of two meta-flow networks: one corresponding to Hermitian RW clustering and the other corresponding to SIC clustering. We see that the distribution of meta-flow network edge weights obtained through the Hermitian RW clustering appears to be shifted to the right of the distribution of edge weights for the industry-based clustering. Since edge weights in the meta-flow network measure flow imbalance between pairs of clusters, this suggests that the data-driven Hermitian RW clustering results in larger flow between pairs of clusters than an

² While Biely and Thurner (2008) use GICS sector classification in their analysis, the GICS Energy sector has substantial overlap with the Mining SIC sector.

Fig. 10 Meta-flow network for Hermitian RW clusters; clusters are represented by nodes and larger edge weights are depicted by bolder colours and thicker lines

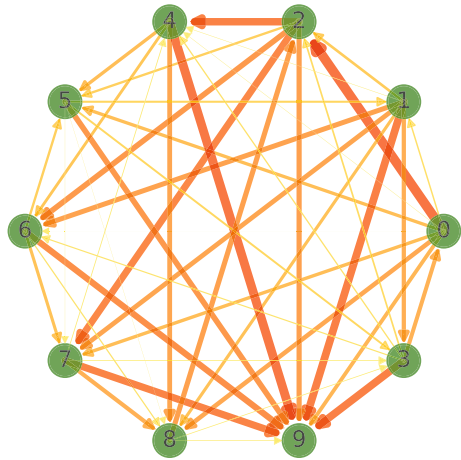
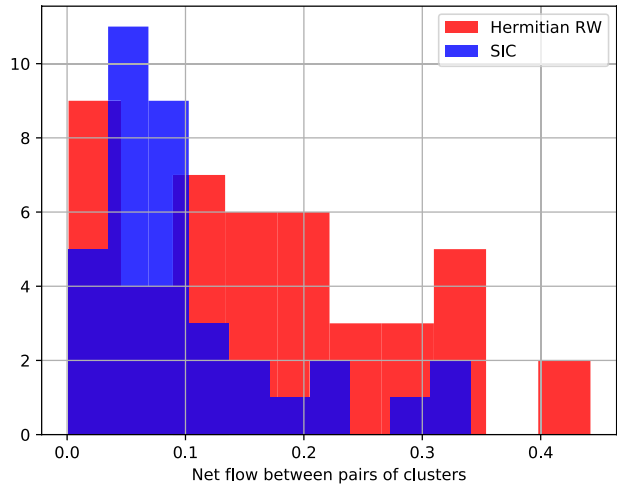


Fig. 11 Histogram of Hermitian RW and SIC clustering meta-flow network edge weights. The edge colours are layered in a semi-transparent fashion



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

Figures 12 and 13 display the average daily dollar volume and market capitalisation averaged across all stocks in a given cluster. We observe that the leading clusters (clusters labelled 0–3) do not appear to have larger average daily dollar volume or market capitalisation.

In order to examine 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 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

Fig. 12 Average daily dollar volume by Hermitian RW cluster

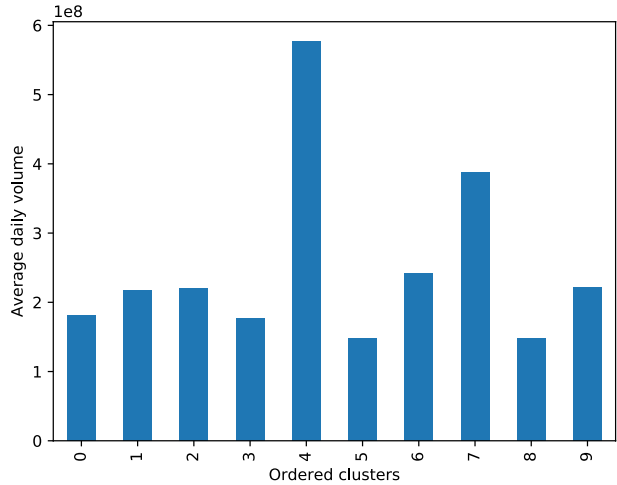
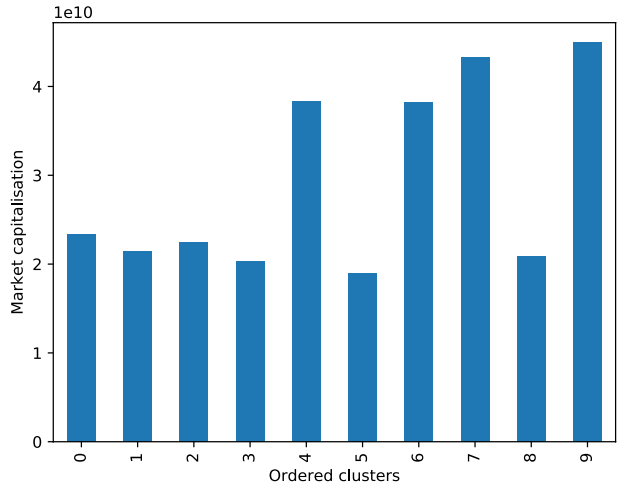


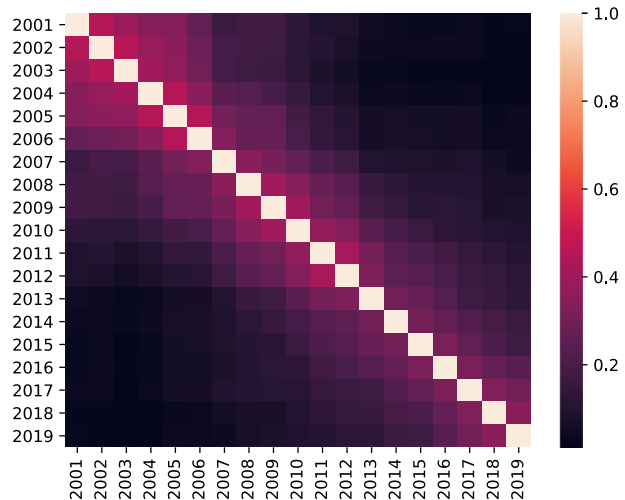
Fig. 13 Average market capitalisation by Hermitian RW cluster



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 three previously hypothesised mechanisms outlined in Sect. 5.2.3. Our proposed method may prove to be useful in the exploration of novel lead–lag mechanisms in the empirical finance community.

Fig. 14 Adjusted Rand index between clusters computed on yearly snapshots of data



5.3 Time-variation in clusters

To investigate the time-variation in the clustering obtained from our method, we recompute the clustering year-by-year using only data from the retrospective year to do so.³ 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 Fig. 14.

The relatively low ARI values between pairs of clusters indicates some—albeit low—persistence in year-to-year lead–lag structure. Biely and Thurner (2008) find that there is significant persistence in lead–lag structures across time. Xia et al. (2018) agree with our observations and find that the lead–lag phenomenon between two stocks is not constant but emerges during certain periods. They find that, on average, individual lead–lag relationships tend to last for around a year.

Further, we see that higher ARI values occur in earlier years: this suggests that there is a decrease in persistence between clusterings as time increases. Nevertheless, in Sect. 6, we show that there is sufficient persistence in the lead–lag cluster relationships in order for a dynamically updated clustering to be useful for forecasting purposes on a daily scale.

5.4 Limitations and implications of the empirical analysis

Our novel lead–lag extraction and clustering method yields clusters that cannot be explained by three previously considered mechanisms for lead–lag structure in US equity markets. Below, we discuss the limitations of our empirical analysis and its implications for understanding lead–lag structure in US equity markets.

First, a caveat of our empirical analysis is the instability of the lead–lag structure across time. In Sect. 5.3, we observe that the lead–lag structure does not exhibit high overall persistence. Since the lead–lag structure is not stable year-to-year, it is possible that the

³ If an equity time series does not have sufficient data during a year then its missing entries in that year's lead–lag matrix are set to 0.

lead–lag results can be partially explained by the three mechanisms on a subset of the data. However, we have repeated our empirical clustering analysis on the relatively stable range⁴ 2000–2006 and have found that the industry-based clustering is unable to fully explain the resulting data-driven clustering on this subset of the data. Furthermore, we have repeated the Spearman correlation analysis that was described in Sect. 5.2.3 using yearly snapshots of data. Appendix Figs. 27 and 28 display the Spearman correlation between an equity's tendency to be a leader (which is given by its lead–lag matrix row-sum) and its market capitalisation or trading volume. As these figures suggest, the association between an equity's tendency to be a leader and its market capitalisation or trading volume is not stable throughout time. There appear to be some periods when the sign of the association is consistent with the positive association predicted by the trading volume and market capitalisation lead–lag mechanisms. Nevertheless, the general sign and transience of the association across time does not support trading volume and market capitalisation as mechanisms which can explain the observed lead–lag structure.

A second caveat for the interpretation of our results concerns the relevancy of the market capitalisation mechanism. As explained in Sect. 5.1, we have restricted our attention to large capitalisation equities in order to avoid non-synchronous trading effects. Therefore, any interpretation of our empirical results must be conditioned by the large capitalisation of our equity universe. In particular, the market capitalisation mechanism may not be relevant under the condition that we restrict attention to the largest equities. In addition, previous papers that have found that smaller cap equities are able to lead larger cap equities if these smaller cap equities receive more news coverage (Scherbina & Schlusche, 2015). Thus, the hypothesised market capitalisation mechanism can be modulated by other information diffusion channels. This implies that the market capitalisation mechanism does not necessarily manifest itself in a positive association between market capitalisation and the tendency of an equity to be a leader.

Thirdly, the lead–lag literature contains other mechanisms that could potentially explain our results (Badrinath et al., 1995; Brennan et al., 1993; Menzly & Ozbas, 2010; Cohen & Frazzini, 2008). For example, cross-firm information flows through supplier networks have been hypothesised as lead–lag mechanisms (Menzly & Ozbas, 2010; Cohen & Frazzini, 2008). Testing these and other hypothesised mechanisms as sources for our observed lead–lag results remains further work.

Finally, given the novelty of our method and the fact that the resulting lead–lag structure cannot be explained through the three hypotheses that we have tested, our method may prove to be useful in the exploration of new mechanisms. The use of non-linear lead–lag metrics and effective algorithms for clustering directed networks (such as the distance correlation lead–lag metric and Hermitian RW algorithm) may illuminate lead–lag structures in US equity markets that cannot be explained by existing lead–lag mechanisms in the empirical finance literature.

6 Financial forecasting application

A difficulty in the modelling of high-dimensional systems is the identification of a suitable group of variables that can be used as predictors for other variables. This is related to the problem of variable selection in high-dimensional predictive modelling. On the one hand, the selection of too few conditioning variables can result in poor predictive power due to

⁴ Cf relatively large values of ARI displayed during 2000–2006 in Fig. 14.

not capturing the temporal dependence between the response variable and relevant omitted variables. On the other hand, conditioning on too many variables can lead to the inclusion of many irrelevant variables; this dilutes the predictive power of the model (Runge et al., 2019).

In general, our unsupervised learning method can be used as a preliminary step to inform the choice, or design of, potential target and feature variables in a predictive model. This is achieved by using clusters with large net inflows (lagging clusters) to guide the choice of *target variables*, and clusters with large net outflows (leading clusters) guide the choice of *feature variables*. For example, in the latent variable synthetic data generating model presented in Sect. 4, the method identifies clusters of variables sharing the same lagged dependence on the latent variable z . By averaging the time series variables within each cluster, $\frac{1}{|C_l|} \sum_{j \in C_l} Y_t^j$, $\forall i \in \{1, \dots, k\}$, the leading latent function $g_1(Z_t)$ at time t (the average of time series values in the most leading cluster) and the lagged latent functions $g_l(Z_{t-l})$, $l = 2, \dots, k$ (the average of time series values in lagging clusters) can be recovered for each $t = 1, \dots, T$ thanks to the reduction in observation noise resulting from the averaging procedure. By fitting models that capture the relations between the average value of the lagging clusters (target variable) and the average value of the leading cluster (feature variable), the latent variable dynamics can be captured, allowing the user to make predictions on the subsequent values of the lagging clusters. In Sect. 6.1, we illustrate the use of our lead–lag detection and clustering method for target and feature variable extraction in a financial forecasting application.

When a downstream model is built to capture the relationships between such target and feature variables, it is likely to exhibit stronger predictive power since our method has screened potential explanatory variables. Our method identifies predictable response variables and diminishes the risk of conditioning on irrelevant variables when used in downstream predictive modelling in high-dimensional time series systems. This approach to identifying groups of target and feature variables is useful for the application of returns forecasting in the US equity universe, since this is a highly noisy multivariate time series system where statistical lead–lag effect sizes are weak.⁵

We assess the predictive power of our lead–lag extraction and spectral clustering approach by evaluating the out-of-sample performance of a trading signal that was constructed using our method. The risk-adjusted returns of our trading signal will be evaluated using the Sharpe ratio. In order to test whether the signal’s Sharpe ratio is significantly different to 0, we use a hypothesis test (Opdyke, 2007) that holds asymptotically under the general conditions of stationary and ergodic signal returns.

Our approach to quantifying the predictive performance of our method by studying the risk-adjusted performance of a portfolio constructed using our method is common in the quantitative finance literature (Asness et al., 2013). The task of constructing a statistically significant trading signal using only publicly available price data in a highly liquid market such as the US equity market is a challenging task due to the informational efficiency of such markets (Malkiel & Fama, 1970). The weak-form of the Efficient Markets Hypothesis states (Malkiel & Fama, 1970) that markets fully reflect all historical price data; this implies that it is not possible to make economic profits in excess of market equilibrium profits by trading on the basis of such historical price data. The number of empirical studies (Malkiel & Fama, 1970) in strong support of the weak-form of the Efficient Markets

⁵ due to the Efficient Markets Hypothesis (Malkiel & Fama, 1970).

Hypothesis underlines the informational efficiency of US equity markets and hence the challenge of constructing a statistically significantly profitable trading signal.

Similarly, Curme et al. (2015b) argue for the use of lead–lag networks to guide variable selection for downstream financial forecasting tasks. However, our results are stronger as we test the performance of our lead–lag network method for variable subset selection in a rolling out-of-sample evaluation.

6.1 Signal construction

We keep the trading signal relatively simple in order to effectively assess the predictive performance of the underlying signal derived from our lead–lag extraction and clustering methodology. Our trading signal forecasts lagging cluster returns using smoothed leading cluster returns. In order to evaluate the out-of-sample performance of our method, we compute the clustering C_1, \dots, C_k and flow graph F on a rolling basis using a 2-month update period and yearly look-back window. Further, using the same update frequency and yearly look-back window, we fit a separate linear model for each pair of clusters. In particular, for each ordered pair of clusters $i, j \in \{1, \dots, k\}$, we fit a linear model to forecast the mean daily return for lagging cluster j

$$Y_t^{(j)} = \frac{1}{|C_j|} \sum_{n \in C_j} Y_t^{(n)},$$

using an exponentially weighted moving average of the mean returns for cluster i as the covariate (input variable to the linear regression)

$$X_t^{(i)} = \frac{1}{|C_i|} \sum_{n \in C_i} \sum_{l=1}^t (1 - \alpha)^{l-1} Y_{t-l}^{(n)}.$$

The choice of exponential parameter $\alpha = 0.4$ assigns 92% of the total weight of the exponential sum $\sum_{l=1}^{\infty} (1 - \alpha)^{l-1}$ to the first 5 lags $l = 1, \dots, 5$. Thus, the exponential moving average mainly captures lead–lag effects on the scale of approximately up to 1 week, while emphasising higher-frequency daily lead–lag effects. The coefficient θ_{ij} of the linear model $Y^{(j)} = \theta_{ij} X_t^{(i)}$ is fitted using ordinary least squares.⁶

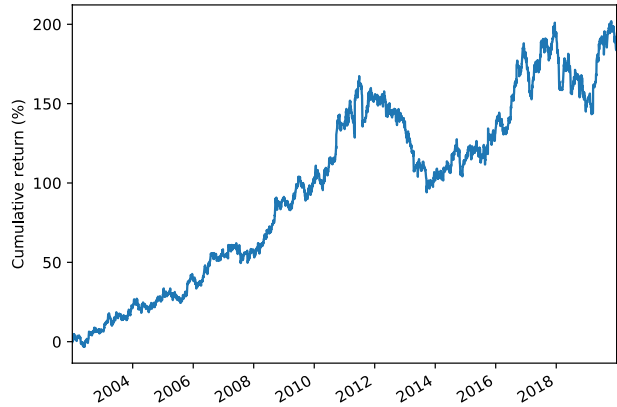
For every day $t = 1, \dots, T$, we compute the predictive signal from cluster i to j for each ordered pair $i, j \in \{1, \dots, k\}$ of clusters

$$\hat{Y}_t^{(j)} = \theta_{ij} X_t^{(i)}.$$

These predictive signals are aggregated using a thresholded flow graph \tilde{F} where $\tilde{F}_{ij} = \mathbb{1}\{F_{ij} > c\}$ where c is the 90% quantile of the edge weights of the flow graph F . Thus, the flow graph ensures that only the cluster-to-cluster relationships that have shown the

⁶ Note that unbiasedness and consistency do not hold in general for this ordinary least squares estimation due to network effects within the residual error structure.

Fig. 15 Cumulative return for the financial forecasting signal; the signal is scaled to target a 10% yearly volatility



greatest historical flow are included in the construction of the signal. Mathematically, the predictive signal S_t for cluster $j \in \{1, \dots, k\}$ is given by

$$S_t(j) = \text{sign}\left(\sum_{n=1}^k \tilde{F}_{ij} \hat{Y}_t^{(j)}\right).$$

The signal for a specific equity $m \in \{1, \dots, p\}$ on day t is set to be the signal for its cluster C_m i.e. $S_t(C_m)$.

Finally, the signals for each equity are normalised by a 21-day historical rolling estimator of the overall signal’s volatility. This rolling normalisation ensures that the overall position size is dynamically adjusted to target a constant 10% annual volatility. Assuming that the Sharpe ratio of our signal is constant throughout time, this procedure can be seen as targeting an optimal Kelly criterion (Thorp, 2011) for the signal on a rolling basis. Further, volatility normalisation tends to bring our daily trading returns closer to stationarity while decreasing their absolute skew and kurtosis; this makes the analysis of our trading returns more reliable.

6.2 Evaluation metric

We evaluate the performance of the signal constructed using our method by its risk-adjusted return. Since the return of the signal on equity m at time t is given by $S_t(C_m) \cdot Y_t^{(m)}$, the total return of the signal at time t is given by

$$\sum_{m=1}^p S_t(C_m) \cdot Y_t^{(m)}$$

The metric that we use for the risk-adjusted return is the Sharpe ratio (Campbell et al., 1997) of the total signal return.

6.3 Results

The cumulative total return of the signal across time is displayed in Fig. 15.

The trading signal results in an annualised Sharpe ratio of 0.62 with a corresponding significant one-sided p -value of $p < 0.004$ (Opdyke, 2007). We compare this with the

Sharpe ratio of 0.40 for the S &P500 market return on the same period. Further, the trading signal exhibits a low correlation (0.04) with the market return. This suggests that the trading signal cannot be explained by market equilibrium returns. The mean daily return of the trading signal is 2.4 basis points.⁷

We observe in Fig. 15 that there is a decay in the performance of the signal after 2012; this can be compared with the reduction in clustering persistence observed after 2012 in Fig. 14, and with the observation in the work of Curme et al. (2015a) that the informational efficiency of the market appears to increase in 2012 relative to earlier years.

Ablation study

We conduct an ablation study in order to test the importance of the lead–lag clustering structure on the observed performance of the trading signal. Specifically, under the null hypothesis that there is no lead–lag cluster structure in US equity returns, the clustering for the US equities is drawn uniformly at random from the set of permutations on cluster labels. Therefore, under the hypothesis of lead–lag cluster structure, the Sharpe ratio of the trading signal described in Sect. 6.1 should be consistent with the distribution over the Sharpe ratios of trading signals that are computed with permuted cluster labels. We use 200 Monte Carlo samples from the null distribution that computes the Sharpe ratio of the same trading signal pipeline described in Sect. 6.1 but with any clustering in this pipeline drawn uniformly at random from S_p . Under the null hypothesis, the Monte Carlo probability that the Sharpe ratio is greater than or equal to the observed Sharpe ratio of 0.62 is 1/201. We thus reject the null hypothesis with p -value $p < 0.005$, and conclude that the lead–lag cluster structure is significant in the construction of the predictive trading signal.

For comparison, we also implement a LASSO-VAR model (Friedman et al., 2010) that fits a multivariate linear model for each equity's next-day return using the 5 previous lagged returns across all equities. Rolling validation and volatility normalisation was performed as described in Sect. 6.1. Specifically, we use an update period of 2 months and a yearly look-back window to fit the LASSO-VAR model on a rolling basis. At each update period, the L1 regularisation hyperparameter is selected using 5-fold cross validation on the rolling year's worth of data across the grid of values $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The signals for each equity are normalised by a 21-day historical rolling estimator of the overall signal's volatility. The LASSO-VAR signal yields a Sharpe ratio of 0.27 with a one-sided p -value $p > 0.13$; the relatively poor performance of the LASSO-VAR baseline model highlights the informational efficiency of the US equity market as well as the difficulty of variable selection in equity returns forecasting problems. Further, the signal constructed using our cluster-based method has a low correlation of -0.005 with the LASSO-VAR signal indicating that the predictive signals of our method cannot be captured using a simple baseline LASSO-VAR model.

A caveat to our results is that we do not take into account transaction costs when calculating the profit of our signal. These may be significant in practice given the basis point size of the average daily returns. On the other hand, the turnover of the trading signal, which is based on a weekly smoothing of lagged returns, is relatively low. Regardless of the economic significance of the signal, it is clear that the clustered lead–lag structure is statistically strong enough to be used as a predictive signal for equity returns.

⁷ Cf a mean daily market return of 3.0 basis points.

7 Conclusion

We propose a methodology for the problem of data-driven detection of leading and lagging clusters of time series. Our unsupervised learning method can capture 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 is statistically significant but that cannot be explained by three prominent lead–lag hypotheses in the empirical finance literature; this suggests that our methodology is a useful tool for the exploration of novel lead–lag mechanisms in the discipline of empirical finance. Furthermore, we find that our method can be employed for challenging downstream forecasting tasks in noisy, high-dimensional settings. In particular, we show how our method can be used for the construction of a statistically significant, parsimonious trading signal in the US equity market.

In addition to the financial domain, the applicability of our proposed methodology extends to other areas—such as economics, medicine and earth sciences—that are characterised by large multivariate time series data which exhibit a latent lead–lag structure. 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.

Appendices

A Additional numerical experiments

A.1 Synthetic data experiment: lead–lag results

Figures 16 and 17 display the lead–lag metric classification accuracy for the Legendre (8) and Hermite (9) synthetic data generating settings, respectively. We observe that the **ccf-*auc*** method with the distance correlation performs best in these non-linear settings.

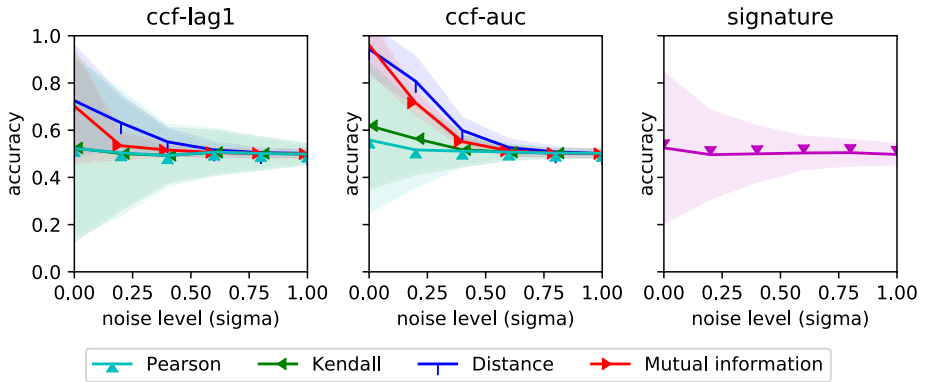


Fig. 16 Average and confidence interval for classification accuracy by lead-lag detection method in the Legendre setting (8)

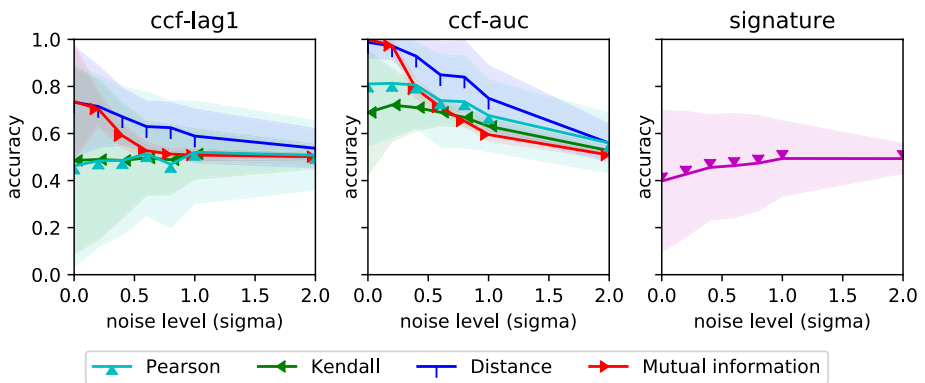


Fig. 17 Average and confidence interval for classification accuracy by lead-lag detection method in the Hermite setting (9)

A.2 Synthetic data experiment: clustering results

Figures 18, 19 and 20 display the ARI of our pipeline in the Legendre (8), Hermite (9) and Heterogeneous (10) synthetic data generating settings, respectively. The pipeline performs best on average using the Hermitian RW clustering component in these settings.

Fig. 18 Average and confidence interval for the ARI by clustering method in the Legendre setting

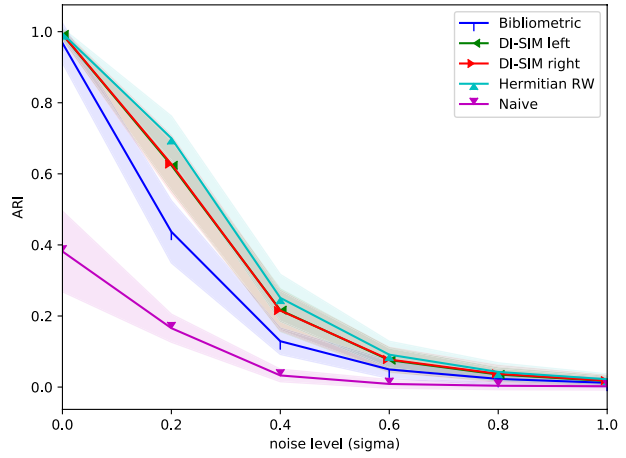


Fig. 19 Average and confidence interval for the ARI by clustering method in the Hermite setting

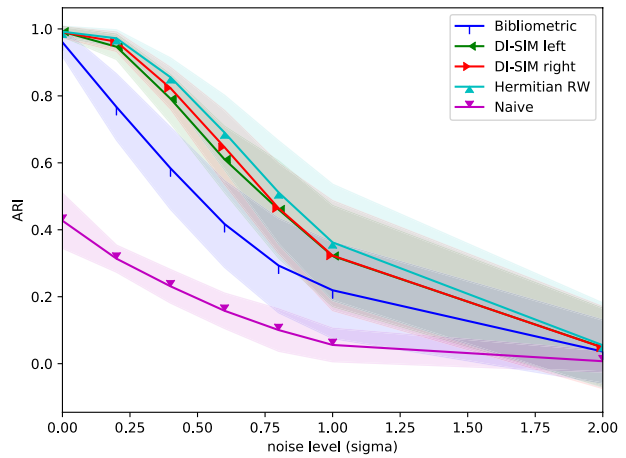
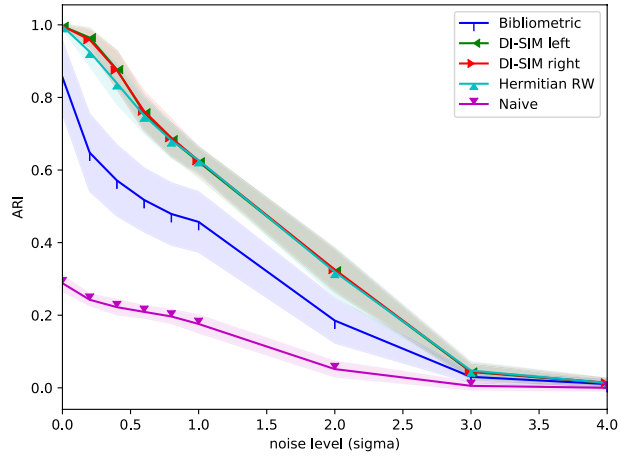


Fig. 20 Average and confidence interval for the ARI by clustering method in the heterogeneous setting



A.3 Synthetic data experiment: interaction of lead-lag and clustering components

Figures 21, 22 and 23 display the ARI of the pipeline for each choice of lead-lag extraction and clustering components.

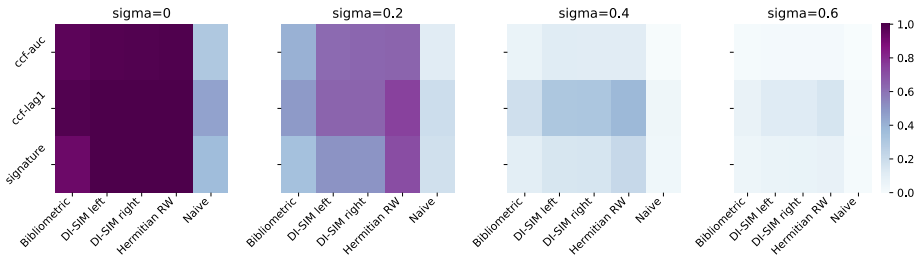


Fig. 21 Average ARI by lead-lag and clustering component in the Legendre setting

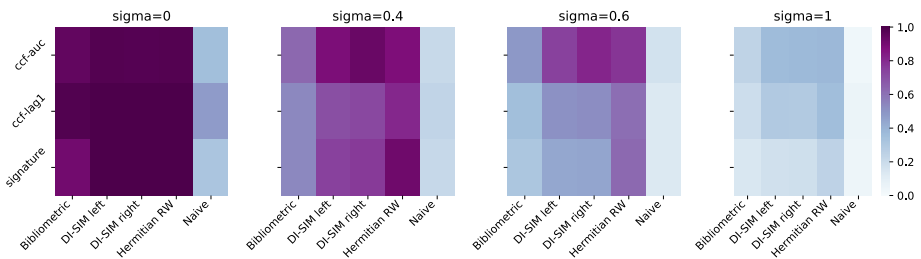


Fig. 22 Average ARI by lead-lag and clustering method in the Hermite setting

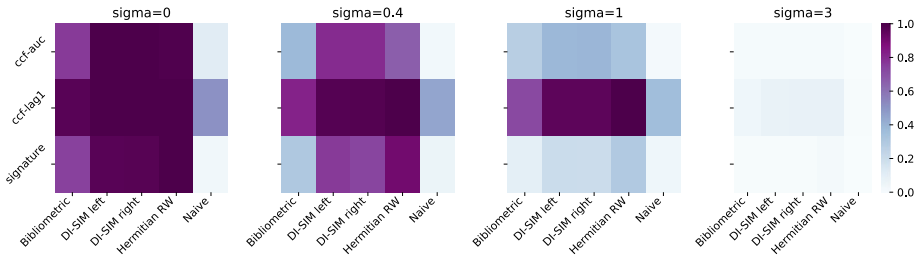


Fig. 23 Average ARI by lead-lag and clustering method in the heterogeneous setting

A.4 Synthetic data ablation study: varying the hyperparameter corresponding to the number of clusters

In Figs. 24, 25 and 26 we display the average and confidence interval for the ARI across different hyperparameter levels for the number of clusters used in the clustering component

Fig. 24 Average and confidence interval for the ARI by different levels of the hyperparameter corresponding to the number of clusters in the cosine setting

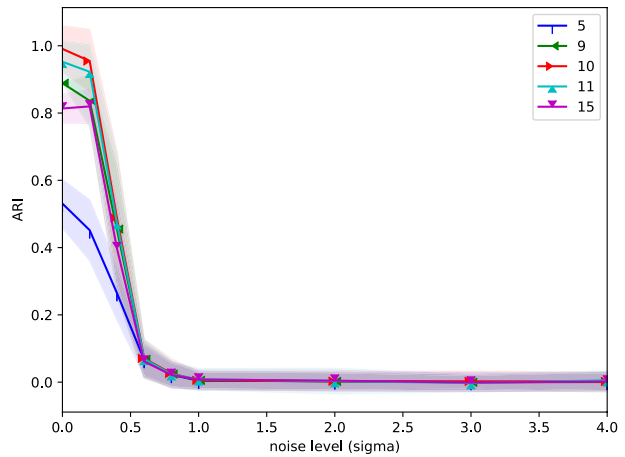


Fig. 25 Average and confidence interval for the ARI by different levels of the hyperparameter corresponding to the number of clusters in the Legendre setting

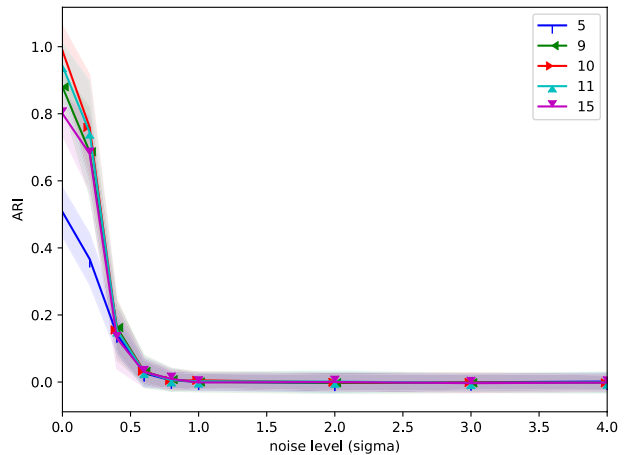
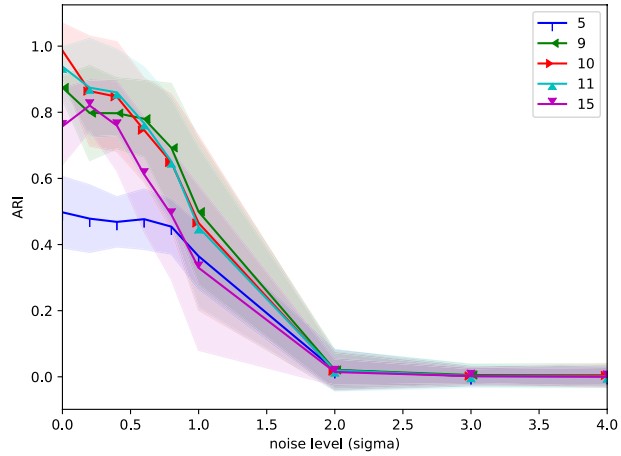


Fig. 26 Average and confidence interval for the ARI by different levels of the hyperparameter corresponding to the number of clusters in the Hermite setting



of the pipeline.

A.5 Real data experiment: time-variation in results

Figures 27 and 28 display the temporal variation in Spearman correlation between the US equity lead–lag matrix row-sums and a given characteristic (*average daily trading volume* in Fig. 27 and *market capitalisation* in Fig. 28) of each equity. We observe that there is

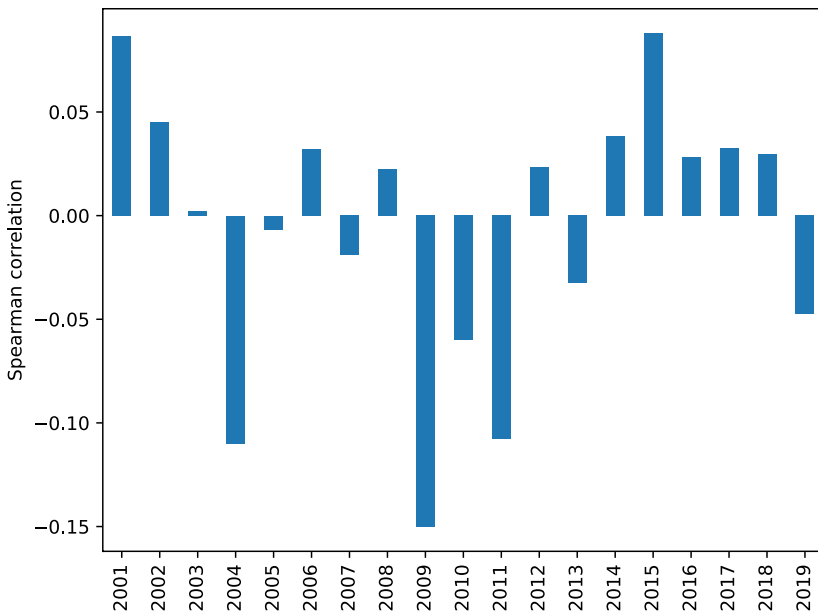


Fig. 27 Spearman correlation between the lead–lag matrix row-sums and *average daily trading volume* for each equity, using yearly snapshots of data

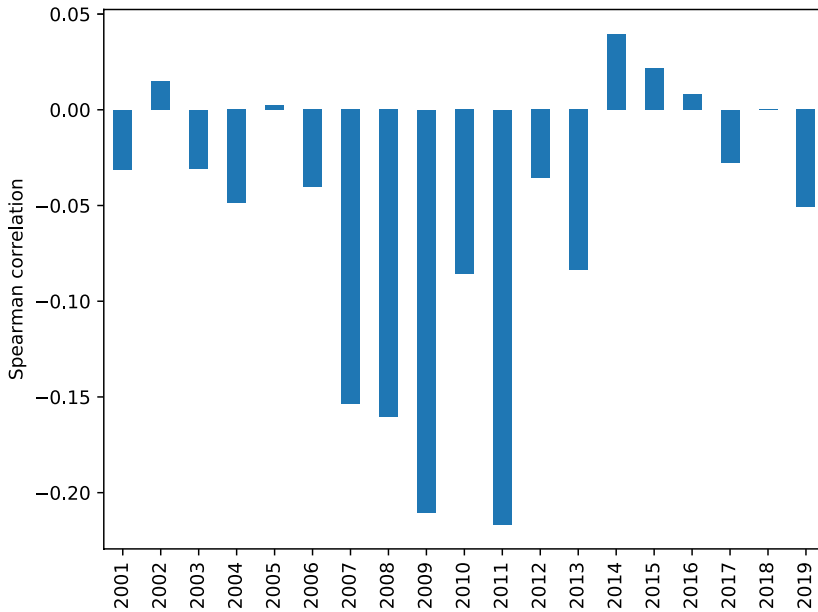


Fig. 28 Spearman correlation between lead-lag matrix row-sums and *market capitalisation* for each equity, using yearly snapshots of data

substantial temporal variation in each equity's tendency to be a leader (as measured by its lead-lag matrix row-sum) and its underlying characteristic.

Acknowledgements The authors would like to acknowledge the anonymous referees for their helpful feedback.

Author contributions All authors made substantial contributions to the work.

Funding SB is supported by the EPSRC CDT in Modern Statistics and Statistical Machine Learning (EP/S023151/1) and The Alan Turing Institute's Finance and Economics Programme. GR is funded in part by EPSRC Grants EP/T018445/1 and EP/R018472/1. All authors acknowledge support from the EPSRC Grant EP/N510129/1 at The Alan Turing Institute.

Data availability All data is available through a WRDS subscription <https://wrds-www.wharton.upenn.edu/>.

Code availability (software application or custom code) All code is available at <https://github.com/stefanosbennett/mlj-lead-lag>.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

Ethics approval The study did not involve human or animal participants.

Consent to participate The study did not involve human or animal participants.

Consent for publication The study did not involve sensitive or confidential data.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Asness, C. S., Moskowitz, T. J., & Pedersen, L. H. (2013). Value and momentum everywhere. *Journal of Finance*, 68(3), 929–985. <https://doi.org/10.1111/jofi.12021>
- Badrinath, S. G., Jayant, R. K., & Thomas, H. N. (1995). Of Shepards, Sheep and the cross-autocorrelations in equity returns. *The Review of Financial Studies*, 8(2), 401.
- Basnarkov, L., Stojkoski, V., Utkovski, Z., & Kocarev, L. (2019). Lead-lag relationships in foreign exchange markets. arXiv <https://doi.org/10.1016/j.physa.2019.122986>, arXiv:1906.10388
- Batson, J., Spielman, D. A., Srivastava, N., & Teng, S. H. (2013). Spectral sparsification of graphs: Theory and algorithms. *Communications of the ACM*, 56(8), 87–94. <https://doi.org/10.1145/2492007.2492029>
- Biely, C., & Thurner, S. (2008). Random matrix ensembles of time-lagged correlation matrices: Derivation of eigenvalue spectra and analysis of financial time-series. *Quantitative Finance*, 8(7), 705–722. <https://doi.org/10.1080/14697680701691477arxiv:abs/0609053> [physics].
- Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3), 535–559. <https://doi.org/10.1016/j.jfineco.2011.12.010>
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39, 324–345.
- Brennan, M. J., Narasimhan, J., & Swaminathan, B. (1993). Investment analysis and the adjustment of stock prices to common information source. *The Review of Financial Studies*, 6(4), 799–824.
- Camilleri, S. J., Scicluna, N., & Bai, Y. (2019). Do stock markets lead or lag macroeconomic variables? Evidence from select European countries. *The North American Journal of Economics and Finance*, 48, 170–186. <https://doi.org/10.1016/j.najef.2019.01.019>
- Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton University Press. <https://doi.org/10.1515/9781400830213-004>
- Chau, S. L., Cucuringu, M., & Sejdinovic, D. (2020). Spectral ranking with covariates. arXiv preprint [arXiv:2005.04035](https://arxiv.org/abs/2005.04035)
- Chevvyrev, I., & Kormilitzin, A. (2016). A primer on the signature method in machine learning. arXiv [arXiv:1603.03788v1](https://arxiv.org/abs/1603.03788v1)
- Chordia, T., & Swaminathan, B. (2000). Trading volume and cross-autocorrelations in stock returns. *The Journal of Finance*, LV(2), 913–935.
- Cohen, L., & Frazzini, A. (2008). Economic links and predictable returns. *Journal of Finance*, 63(4), 1977–2011. <https://doi.org/10.1111/j.1540-6261.2008.01379.x>
- Conrad, J., Gultekin, M., & Kaul, G. (1991). Asymmetric predictability of conditional variances. *The Review of Financial Studies*, 4(4), 597–622.
- Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2), 223–236. <https://doi.org/10.1080/713665670>
- Cucuringu, M. (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), 58–79.
- Cucuringu, M., Li, H., Sun, H., & Zanetti, L. (2020). Hermitian matrices for clustering directed graphs: Insights and applications. AISTATS pp 1–19. [arXiv:1908.02096](https://arxiv.org/abs/1908.02096)
- Curme, C., Tumminello, M., Mantegna, R. N., Stanley, H. E., & Kenett, D. Y. (2015a). Emergence of statistically validated financial intraday lead-lag relationships. *Quantitative Finance*, 15(8), 1375–1386. <https://doi.org/10.1080/14697688.2015.1032545arXiv:1401.0462>
- Curme, C., Tumminello, M., Mantegna, R. N., & Stanley, H. E. (2015b). *How lead-lag correlations affect the intraday pattern of collective stock dynamics*. Office of Financial Research Working Paper Series <https://doi.org/10.2139/ssrn.2648490>
- d'Aspremont, A., Cucuringu, M., & Tyagi, H. (2021). Ranking and synchronization from pairwise measurements via SVD. *Journal of Machine Learning Research*, 22(19), 1–63.

- De Bacco, C., Larremore, D. B., & Moore, C. (2018). A physical model for efficient ranking in networks. *Science Advances*, 4(7), 1–10.
- Dugué, N., & Perez, A. (2015). Directed Louvain: Maximizing modularity in directed networks. In *HAL archives ouvertes* (pp. 0–14). <https://hal.archives-ouvertes.fr/hal-01231784>
- Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56. <https://doi.org/10.2469/dig.v36.n3.4225>
- Farrell, J. (1974). Analyzing covariation of returns to determine homogeneous stock groupings. *Journal of Business*, 47(2), 186–207.
- Fiedor, P. (2014). Information-theoretic approach to lead–lag effect on financial markets. *European Physical Journal B*. <https://doi.org/10.1140/epjb/e2014-50108-3arXiv:1402.3820>
- Fogel, F., d'Aspremont, A., & Vojnovic, M. (2016). Spectral ranking using seriation. *Journal of Machine Learning Research*, 17(88), 1–45.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22. [arxiv:0908.3817](https://arxiv.org/abs/0908.3817)
- Gates, A. J., & Ahn, Y. Y. (2017). The impact of random models on clustering similarity. *Journal of Machine Learning Research*, 18(87), 1–28.
- Gleich, D. F., & Lim, L. H. (2011). Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM* (pp. 60–68).
- Google. (2012). The PageRank citation ranking: Bringing order to the web January. In *Proceedings of the 2012 IEEE international symposium on workload characterization. IISWC* (Vol. 2012, pp. 111–112). <https://doi.org/10.1109/IISWC.2012.6402911>
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13, 723–773.
- Gyurkó, L. G., Lyons, T., Kontkowski, M., & Field, J. (2014). Extracting information from the signature of a financial data stream (pp 1–22). arXiv [arXiv:1307.7244](https://arxiv.org/abs/1307.7244)
- Harzallah, A., & Sadourny, R. (1997). Observed lead–lag relationships between Indian summer monsoon and some meteorological variables. *Climate Dynamics*, 13(9), 635–648. <https://doi.org/10.1007/s003820050187>
- He, Y., Reinert, G., & Cucuringu, M. (2021). Digragc: Digraph clustering with flow imbalance. arXiv [arxiv: 2106.05194](https://arxiv.org/abs/2106.05194) [stat.ML]
- Hu, P., & Lau, W. C. (2013). A survey and taxonomy of graph sampling. [arxiv:1308.5865](https://arxiv.org/abs/1308.5865)
- Huber, P. J. (1962). Pairwise comparison and ranking: Optimum properties of the row sum procedure. *The Annals of Mathematical Statistics*, 34, 511.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1), 193–218.
- Huth, N. (2012). High frequency lead/lag relationships empirical facts. *Journal of Empirical Finance*, 26(March 2014), 41–58.
- Iyetomi, H., Aoyama, H., Fujiwara, Y., Souma, W., Vodenska, I., & Yoshikawa, H. (2020). Relationship between macroeconomic indicators and economic cycles in US. *Scientific Reports*, 10(1), 1–12.
- Janzing, D., Balduzzi, D., Grosse-Wentrup, M., & Schölkopf, B. (2013). Quantifying causal influences. *Annals of Statistics*, 41(5), 2324–2358. <https://doi.org/10.1214/13-AOS1145arXiv:1203.6502>
- Jegadeesh, N., & Titman, S. (1995). Overreaction, delayed reaction, and contrarian profits. *The Review of Financial Studies*, 8(4), 973–993.
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1), 81–93.
- Laenen, S., & Sun, H. (2020). Higher-order spectral clustering of directed graphs. In *Advances in neural information processing systems 2020 (NeurIPS)*. [arXiv:2011.05080](https://arxiv.org/abs/2011.05080)
- Levin, D., Lyons, T., & Ni, H. (2016). Learning from the past, predicting the statistics for the future, learning an evolving system (pp 1–40). arXiv [arXiv:1309.0260](https://arxiv.org/abs/1309.0260)
- Liao, C., Huang, Y., Shi, X., & Jin, X. (2014). Mining influence in evolving entities: A study on stock market. In: *DSAA 2014—Proceedings of the 2014 IEEE international conference on data science and advanced analytics* (pp. 244–250). <https://doi.org/10.1109/DSAA.2014.7058080>
- Lin, Z., Ding, W., Yan, G., Yu, C., & Giua, A. (2013). Leader–follower formation via complex Laplacian. *Automatica*, 49, 1900–1906.
- Lo, A. W., & MacKinlay, A. C. (1990). When are contrarian profits due to stock market overreaction. *The Review of Financial Studies*, 3(2), 175–205.
- Malkiel, B. G., & Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*. <https://doi.org/10.2307/2325488>
- Marti, G., Andler, S., & Nielsen, F., & Donnat, P. (2016). Exploring and measuring non-linear correlations: Copulas, lightspeed transportation and clustering. arXiv [arXiv:1610.09659](https://arxiv.org/abs/1610.09659)

- Marti, G., Nielsen, F., & Bińkowski, M., & Donnat, P. (2019). A review of two decades of correlations, hierarchies, networks and clustering in financial markets (pp. 1–34). arXiv [arXiv:1703.00485](https://arxiv.org/abs/1703.00485)
- Menzly, L., & Ozbas, O. (2010). Market segmentation and cross-predictability of returns. *Journal of Finance*, 65(4), 1555–1580. <https://doi.org/10.1111/j.1540-6261.2010.01578.x>
- Namaki, A., Shirazi, A. H., Raei, R., & Jafari, G. R. (2011). Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and its Applications*, 390(21–22), 3835–3841. <https://doi.org/10.1016/j.physa.2011.06.033>
- Newman, M. (2018). *Networks* (2nd ed.). Oxford University Press.
- Opdyke, J. D. (2007). Comparing sharpe ratios: So where are the p -values? *Journal of Asset Management*, 8(5), 308–336. <https://doi.org/10.1057/palgrave.jam.2250084>
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th international world wide web conference* (pp 161–172).
- Pentney, W., & Meila, M. (2005). Spectral clustering of biological sequence data. *Proceedings of the National Conference on Artificial Intelligence*, 2, 845–850.
- Podobnik, B., Wang, D., Horvatic, D., Grosse, L., & Stanley, H. (2010). Time-lag cross-correlations in collective phenomena. *EPL*, 90, 68001. <https://doi.org/10.1209/0295-5075/90/68001>
- Reizenstein, J. & Graham, B. (2018). The iisignature library: Efficient calculation of iterated-integral signatures and log signatures (pp. 1–18). arXiv:1802.08252
- Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large datasets. *Science*, 334(6062), 1518–1524. <https://doi.org/10.1126/science.1205438.Detecting>
- Rohe, K., Qin, T., & Yu, B. (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), 12679–12684. <https://doi.org/10.1073/pnas.1525793113>
- Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting causal associations in large nonlinear time series datasets. *Science Advances*, 5(11), 1–15.
- Sandoval, L. (2014). Structure of a Global Network of financial companies based on transfer entropy. *Entropy*, 16(8), 4443–4482. <https://doi.org/10.3390/e16084443>
- Sandoval, L., & Franca, I. D. P. (2012). Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and its Applications*, 391(1–2), 187–208. <https://doi.org/10.1016/j.physa.2011.07.023arXiv:1102.1339>
- Satuluri, V., & Parthasarathy, S. (2011). Symmetrizations for clustering directed graphs. In *ACM international conference proceeding series* (pp 343–354). <https://doi.org/10.1145/1951365.1951407>
- Scherbina, A. D., & Schlusche, B. (2015). Cross-firm information flows and the predictability of stock returns. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2263033>
- Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/ICIP.2014.7025680>
- Shojaie, A., & Fox, E. B. (2021). Granger causality: A review and recent advances. arXiv [arXiv:2105.02675](https://arxiv.org/abs/2105.02675)
- Skoura, A. (2019). Detection of lead–lag relationships using both time domain and time-frequency domain; An application to wealth-to-income ratio. *Economies*, 7(2), 28–60. <https://doi.org/10.3390/economies7020028>
- Sornette, D., & Zhou, W. X. (2005). Non-parametric determination of real-time lag structure between two time series: The ‘optimal thermal causal path’ method. *Quantitative Finance*, 5(6), 577–591. <https://doi.org/10.1080/14697680500383763>
- Stavroglou, S., Pantelous, A., Soramaki, K., & Zuev, K. (2017). Causality networks of financial assets. *The Journal of Network Theory in Finance*, 3(2), 17–67. <https://doi.org/10.21314/jntf.2017.029>
- Sussman, D. L., Tang, M., Fishkind, D. E., & Priebe, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499), 1119–1128. <https://doi.org/10.1080/01621459.2012.699795arXiv:1108.2228>
- Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6), 2769–2794. <https://doi.org/10.1214/009053607000000505>
- Thorp, E. O. (2011). The Kelly criterion in blackjack sports betting, and the stock market. In *The Kelly capital growth investment criterion* (Chapter 9). World Scientific Book [https://doi.org/10.1016/s1872-0978\(06\)01009-x](https://doi.org/10.1016/s1872-0978(06)01009-x)
- Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9, 1–12. <https://doi.org/10.1038/s41598-019-41695-zarXiv:1810.08473>

- Tumminello, M., Lillo, F., & Mantegna, R. N. (2010). Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior and Organization*, 75(1), 40–58. <https://doi.org/10.1016/j.jebo.2010.01.004>arXiv:0809.4615
- Underwood, W. G., Elliott, A., & Cucuringu, M. (2020). Motif-based spectral clustering of weighted directed networks. *Applied Network Science*, 5(62), 1–14.
- Výrost, T., Lyócsa, Š, & Baumöhl, E. (2015). Granger causality stock market networks: Temporal proximity and preferential attachment. *Physica A: Statistical Mechanics and its Applications*, 427, 262–276. <https://doi.org/10.1016/j.physa.2015.02.017>
- Wang, D., Tu, J., Chang, X., & Li, S. (2017). The lead–lag relationship between the spot and futures markets in China. *Quantitative Finance*, 17(9), 1447–1456. <https://doi.org/10.1080/14697688.2016.1264616>
- Wang, G. J., Xie, C., He, K., & Stanley, H. E. (2017). Extreme risk spillover network: Application to financial institutions. *Quantitative Finance*, 17(9), 1417–1433. <https://doi.org/10.1080/14697688.2016.1272762>
- Wharton Research Data Service (2020) Center for Research in Security Prices (CRSP)
- Wu, D., Ke, Y., & Yu, J. X., Chen, L.(2010). Detecting leaders from correlated time series. In *International conference on database systems for advanced applications 5981 LNCS* (pp. 352–367). https://doi.org/10.1007/978-3-642-12026-8_28
- Xia, L., You, D., Jiang, X., & Chen, W. (2018). Emergence and temporal structure of lead–lag correlations in collective stock dynamics. *Physica A: Statistical Mechanics and its Applications*, 502, 545–553. <https://doi.org/10.1016/j.physa.2018.02.112>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.