International Journal of Theoretical and Applied Finance
Vol. 18, No. 7 (2015) 1550043 (26 pages)

DOI: 10.1142/S0219024915500430

World Scientific
www.worldscientific.com

COUPLED NETWORK APPROACH TO PREDICTABILITY OF FINANCIAL MARKET RETURNS AND NEWS SENTIMENTS

CHESTER CURME* and H. EUGENE STANLEY†

Center for Polymer Studies and Department of Physics
Boston University, 590 Commonwealth Avenue
Boston, MA 02215, USA
**ccurme@bu.edu*
†*hes@bu.edu*

IRENA VODENSKA

Administrative Sciences Department
Metropolitan College, Boston University
808 Commonwealth Avenue
Boston, MA 02215, USA
vodenska@bu.edu

Received 1 April 2015
Accepted 12 June 2015
Published 28 October 2015

We analyze the network structure of lagged correlations among daily financial news sentiments and returns of financial market indices of 40 countries from 2002 to 2012. Using a spectral method, we decompose the network into bipartite sub-structures, and show that these sub-structures are relevant to the performance of prediction models, bridging concepts from network theory and time series analysis. Our results suggest that, at the daily level, endogenous influences among financial markets overwhelm exogenous influences of news outlets, and that changes in financial news sentiments respond to market movements more substantially than they drive them.

Keywords: News sentiments; time-series analysis; logistic regression; singular value decomposition.

1. Introduction

Recent history has revealed the degrees to which the well-being of individuals and entire economies are tied to the state of the financial sector, directing much scientific attention at the drivers of financial market fluctuations. Fama (1970) developed the efficient market hypothesis, which suggests that all available information is reflected in the current price of financial assets, and it is therefore not possible to predict future values of an asset using only past records. When considering the assets comprising major global stock indices, relevant information may be encoded

in a variety of forms, including news and analyst reports. Weak forms of the efficient market hypothesis may additionally allow that the returns of other major indices or assets offer relevant information.

The latter phenomenon has been documented for several decades. Becker *et al.* (1990) observed that daily returns of the S&P 500 explain 7%–25% of fluctuations in the Nikkei Index returns the next day. Using simple trading strategies, the authors were able to correctly predict upward movements of the Nikkei with accuracies ranging from 72% to 81%, and downward movements with accuracies ranging from 59% to 75%. The author's simulations conclude that accounting for transaction costs, however, is sufficient to eliminate any excess profits potentially obtained from such strategies. So although predictive information might be encoded among the returns of markets with different operating hours, this information is typically not actionable, in the sense that one could consistently translate the information into a profit. A variety of studies have found similar international return and volatility spillover effects (see, in particular, Brailsford 1996, Ghosh *et al.* 1999, Hamao *et al.* 1990, Sandoval 2014, Vandewalle *et al.* 2000). Diebold & Yilmaz (2009) report that certain measures of return spillover effects have been increasing steadily since the early 1990s.

While the returns of global indices may be readily calculated and incorporated into statistical models, the impact of exogenous news is more difficult to quantify. Some have approached the problem by quantifying "news" as the difference between announced national macroeconomic fundamentals and surveyed expectations (Anderson *et al.* 2003, Anderson *et al.* 2007, Balduzzi *et al.* 2001). This approach has been central to the studies of economic efficiency. At the level of individual firms, for example, researchers have identified persistent anomalous drifts in stock prices for months following announcements of unexpectedly high earnings (see, in particular, Ball & Brown 1968, Chordia *et al.* 2009). To capture relevant news items beyond the announced financial and macroeconomic figures, however, usually requires the quantification of information from text-based sources. In recent decades, the automated forecasting of financial markets using relevant text-based information has advanced tremendously, following the growing abundance of online text data in the form of news and social media outlets. Piškorec *et al.* (2014) quantify the cohesiveness of financial news according to the co-occurrence of keywords in online news streams, and find that this cohesiveness largely responds to fluctuations in market volatility. A more common approach is sentiment analysis (see, in particular, Godbole *et al.* 2007, Zhang & Skiena 2010), in which documents are distilled to numbers that characterize the author's opinion with respect to an asset, market, or other item or event of interest. Developments in this area have enabled the statements of analysts, reporters, and individuals in online investment communities to be parsed and interpreted by forecasting algorithms at increasing speeds.

The information encoded in such sentiment analyses both reflects and influences the decisions of investors, which collectively may shape the gains and losses of

financial markets worldwide. To disentangle the directionality of these relationships, here we investigate the interactions among financial markets and news sentiment data for 40 countries for the period from 2002 through 2012. Through the consideration of lagged correlation-based networks, we explore the extent to which news leads financial market movements, and to which markets lead news. Using tools from linear algebra, we abstract away from the level of individual countries in order to identify large-scale flows of information among geographic regions. We find that, at a time resolution of one day, and both at the level of individual nodes and when considering the network's larger-scale structure, financial markets anticipate news much more substantially than news items anticipate market movements. Finally, we use logistic regression models to show that the structures in the lagged networks are indicative of some degree of predictability, and demonstrate how consideration of the network's community structure can be useful in building more robust predictive models. Our results bolster previous studies of international return spillover effects among markets, and offer a novel picture of the interactions among market returns and financial news sentiments.

In Sec. 2, we introduce the data sources, provide summary statistics, and explain our procedure for de-trending the data to guard against spurious results due to serial correlation. In Sec. 3, we describe our methodology for constructing networks of lagged correlations among news sentiment signals and market returns, interpret the results of our method, and summarize the community structures embedded in the directed network. In Sec. 4, we show how consideration of these community structures can be useful in building more robust predictive models. We offer concluding remarks and propose extensions of the work in Sec. 5.

2. Data and Summary Statistics

We obtain daily news signals for each country from the Thomson Reuters *MarketPsych* indices (MarketPsych 2013). The *MarketPsych* signals are computed using textual news from Reuters as well as various third-party sources. Text is also sourced from blogs, microblogs, and other social media. The indices are constructed using a proprietary language framework to quantify finance-specific emotions and forecasts, in addition to opinions on more specific topics (regarding inflation, fiscal policy, risk of default, etc.). These indices are used by Thomson Reuters and its clients for a variety of purposes, including the development of quantitative trading strategies, risk analysis, and volatility forecasting. For the purposes of this study, we make use of the general "sentiment" indices, which measure "overall positive references, net of negative references" (MarketPsych 2013) in financial news for a given country and takes a value in the range $[-1, 1]$.

First, we clean the data for missing values, replacing them by the sentiment value at one day prior. We discard any country with more than 1% missing values from the analysis. We thus obtain news sentiment signals $S_{i,t}$, for 40 countries indexed $i = 1, \ldots, 40$. The full list of 40 countries studied here is provided in Table 1.

Table 1. Summary statistics, including sample means and standard deviations, for the (detrended) returns $\tilde{r}_{i,t}$ and news sentiment signals $\tilde{s}_{i,t}$ in the period January 8, 2002 to December 31, 2012.

Country	Index	$\langle \tilde{r}_{i,t} \rangle$	$\sigma_{\tilde{r}}$	$\langle \tilde{s}_{i,t} \rangle$	$\sigma_{\tilde{s}}$
Argentina	MERVAL	7.33×10^{-5}	1.91×10^{-2}	-3.10×10^{-5}	9.43×10^{-2}
Australia	AS51	2.01×10^{-5}	1.09×10^{-2}	2.10×10^{-5}	6.55×10^{-2}
Austria	ATX	1.43×10^{-5}	1.61×10^{-2}	-7.95×10^{-5}	1.61×10^{-1}
Belgium	BEL20	3.16×10^{-5}	1.40×10^{-2}	-1.20×10^{-4}	1.17×10^{-1}
Brazil	IBOV	4.44×10^{-5}	1.85×10^{-2}	-5.34×10^{-5}	8.18×10^{-2}
Chile	IPSA	8.21×10^{-5}	1.06×10^{-2}	9.34×10^{-5}	1.72×10^{-1}
China	SHSZ300	4.99×10^{-5}	1.72×10^{-2}	1.42×10^{-5}	4.85×10^{-2}
Colombia	IGBC	5.94×10^{-5}	1.37×10^{-2}	-1.24×10^{-4}	1.26×10^{-1}
Denmark	KFX	3.15×10^{-5}	1.36×10^{-2}	1.24×10^{-4}	1.75×10^{-1}
Finland	HEX25	4.83×10^{-5}	1.51×10^{-2}	-1.70×10^{-4}	2.06×10^{-1}
France	CAC	3.99×10^{-5}	1.59×10^{-2}	-2.37×10^{-5}	5.66×10^{-2}
Germany	DAX	5.12×10^{-5}	1.62×10^{-2}	-2.23×10^{-5}	6.34×10^{-2}
Greece	ASE	7.44×10^{-5}	1.76×10^{-2}	-1.09×10^{-4}	1.20×10^{-1}
Hong Kong	HSI	4.87×10^{-5}	1.57×10^{-2}	4.58×10^{-5}	1.41×10^{-1}
Hungary	BUX	-1.47×10^{-5}	1.67×10^{-2}	-1.81×10^{-4}	1.84×10^{-1}
Indonesia	JCI	1.24×10^{-5}	1.45×10^{-2}	-6.33×10^{-5}	1.03×10^{-1}
Ireland	ISEQ	6.21×10^{-6}	1.54×10^{-2}	-3.62×10^{-5}	9.14×10^{-2}
Israel	TA-25	2.58×10^{-5}	1.26×10^{-2}	-3.17×10^{-5}	4.47×10^{-2}
Italy	FTSEMIB	4.34×10^{-5}	1.58×10^{-2}	-8.96×10^{-5}	6.99×10^{-2}
Japan	NKY	4.43×10^{-5}	1.54×10^{-2}	1.32×10^{-5}	7.28×10^{-2}
Malaysia	FBMKLCI	1.89×10^{-5}	7.81×10^{-3}	-2.73×10^{-5}	1.38×10^{-1}
Mexico	MEXBOL	1.41×10^{-5}	1.33×10^{-2}	-9.46×10^{-6}	8.03×10^{-2}
Netherlands	AEX	3.95×10^{-5}	1.61×10^{-2}	-1.08×10^{-4}	1.61×10^{-1}
New Zealand	NZSE50FG	1.81×10^{-5}	7.13×10^{-3}	1.05×10^{-5}	1.26×10^{-1}
Norway	OBX	3.21×10^{-6}	1.75×10^{-2}	-6.08×10^{-5}	1.36×10^{-1}
Pakistan	KSE100	1.22×10^{-5}	1.38×10^{-2}	-3.32×10^{-5}	6.70×10^{-2}
Peru	IGBVL	3.58×10^{-5}	1.57×10^{-2}	-3.54×10^{-5}	1.74×10^{-1}
Philippines	PCOMP	2.12×10^{-5}	1.29×10^{-2}	-1.01×10^{-4}	1.18×10^{-1}
Poland	WIG	1.90×10^{-6}	1.31×10^{-2}	-9.02×10^{-5}	1.47×10^{-1}
Portugal	PSI20	1.74×10^{-5}	1.18×10^{-2}	-2.22×10^{-4}	1.72×10^{-1}
Russia	INDEXCF	-5.58×10^{-5}	2.28×10^{-2}	-2.37×10^{-5}	5.88×10^{-2}
Saudi Arabia	SASEIDX	6.59×10^{-5}	1.71×10^{-2}	-4.32×10^{-5}	9.95×10^{-2}
South Africa	TOP40	2.02×10^{-5}	1.41×10^{-2}	-1.06×10^{-4}	8.09×10^{-2}
Spain	IBEX	5.01×10^{-5}	1.57×10^{-2}	-6.63×10^{-5}	8.27×10^{-2}
Sweden	OMX	8.51×10^{-5}	1.56×10^{-2}	-1.27×10^{-4}	1.42×10^{-1}
Switzerland	SMI	-9.55×10^{-6}	1.27×10^{-2}	-1.16×10^{-4}	1.14×10^{-1}
Thailand	SET	5.56×10^{-5}	1.39×10^{-2}	2.55×10^{-5}	1.17×10^{-1}
United Kingdom	UKX	2.06×10^{-5}	1.31×10^{-2}	-1.14×10^{-5}	4.14×10^{-2}
United States	SPX	2.96×10^{-5}	1.33×10^{-2}	-1.71×10^{-5}	3.05×10^{-2}
Venezuela	IBVC	7.62×10^{-5}	1.38×10^{-2}	-1.90×10^{-4}	1.24×10^{-1}

In addition to the news sentiment data, we simultaneously study the returns of major stock indices in each country. We obtain closing prices $P_{i,t}$ for major stock indices of each country i on each trading day t from Bloomberg. We then transform the prices $P_{i,t}$ to logarithmic returns

$$r_{i,t} \equiv \log(P_{i,t}) - \log(P_{i,t-1}). \tag{2.1}$$

We provide the full list of stock indices considered in Table 1.

We aim to measure one-day lagged relationships among the news sentiment signals and index returns. Many of the news sentiment signals exhibit a large degree of autocorrelation. In addition, the return signals from the markets of certain developing countries exhibit a non-negligible degree of autocorrelation at a lag of one day. To isolate the influences of exogenous factors from the endogenous structure of each time series, we first difference the news sentiment time series to obtain 40 time series $s_{i,t} \equiv S_{i,t} - S_{i,t-1}$. We then apply an AR(1) filter to the signals $s_{i,t}$ and $r_{i,t}$. This is equivalent to applying an ARIMA$(1, 1, 0)$ filter to the original news sentiment signals $S_{i,t}$ and to the logarithmic prices $\log(P_{i,t})$. We find that alternative procedures, such as incorporating moving average terms instead of autoregressive terms, or not differencing the news sentiment signals, fail to adequately remove the effects of autocorrelation from the data.

Specifically, we subtract the influences of such autocorrelation features from our signals using one-step rolling forecasts. For each point $s_{i,t}$ in each news sentiment time series, for example, we fit a local regression (Shumway & Stoffer 2011)

$$s_{i,t} = \beta_0 + \beta_1 s_{i,t-1}, \tag{2.2}$$

using the previous 100 days of data — i.e. using the values of $\{s_{t-1}, s_{t-2}, s_{t-3}, \ldots, s_{t-100}\}$ on the left-hand side of the equation. We then subtract the out-of-sample sentiment predicted from the regression from the observed sentiment at week t to obtain our fully detrended time series

$$\tilde{s}_{i,t} \equiv s_{i,t} - (\beta_0 + \beta_1 s_{i,t-1}), \tag{2.3}$$

which are the residuals from one-step rolling forecasts of our autoregressive model. This method of de-trending, in which we make use of data only from days $t' < t$ in order to adjust the value of the time series at time t, is preferred in this case over other local regression methods, many of which use a symmetric window around t. Because we will be making predictions, we explicitly avoid contaminating our processed data at time t with data from times $t' > t$.

We implement the exact same procedure on the returns $r_{i,t}$ in order to construct the detrended time series $\tilde{r}_{i,t}$. The signals $\tilde{s}_{i,t}$ and $\tilde{r}_{i,t}$ were obtained for a total of 40 countries over a period ranging from January 8, 2002 to December 31, 2012. Summary statistics, including the first two moments of $\tilde{r}_{i,t}$ and $\tilde{s}_{i,t}$ for each country and index considered, are provided in Table 1.

3. Analysis of Lagged Correlations

3.1. *Methodology*

We study Pearson correlations among all signals at a lag of one day. Although the market return data only exists at most between Monday and Friday of each week, the news sentiment data is available seven days per week. We adopt a lagging scheme that maintains a constant time series length T for all relationships studied, but ensures that each term in the Pearson product-moment sum includes signals that

are separated by the minimum possible nonzero time lag at a resolution of one day. Our procedure is given in detail in Appendix A. In Appendix C, we also consider synchronous correlations among news sentiment signals and market returns, and examine the topological structure of the synchronous correlation matrix.

For each of the four possible categories of relationships — market–market, news–news, news-market, and market-news — we assemble the time series as columns in a matrix $X^{(t)}$. We then shift the time series by one day, as detailed in Appendix A, and assemble them as columns in a matrix $X^{(t+1)}$. We construct the lagged correlation matrix, the entries of which are given by

$$L_{i,j} = \frac{1}{T-1} \sum_{k=1}^{T} \frac{(X_{i,k}^{(t)} - \langle X_i^{(t)} \rangle)(X_{j,k}^{(t+1)} - \langle X_j^{(t+1)} \rangle)}{\sigma_i \sigma_j}. \tag{3.1}$$

To study the structure of this matrix, we aim to filter its elements into a network of directed relationships. A common approach to constructing correlation-based networks is to filter edges according to a topological constraint, as in the minimal spanning tree (MST) (Mantegna 1999). This approach generally relies on a symmetric correlation between any two nodes. It therefore does not readily extend to the study of lagged correlation networks, in which the correlations are asymmetric: in general, $L_{i,j} \neq L_{j,i}$. More generally, such topological methods of filtering a correlation matrix into a network, which depend heavily on the ranking of the measured correlation coefficients, are less robust to statistical uncertainty than other methods, such as applying a threshold to the matrix (Curme *et al.* 2015). This is especially important when studying lagged correlations, which tend to be much lower in magnitude than synchronous correlations.

We could apply a simple thresholding procedure, choosing a static threshold based on statistical confidence — i.e. a correlation coefficient that has a probability less than p of being generated by uncorrelated variables. But this threshold will vary with the distribution of the signals under consideration, many of which are known to be non-normal (Mantegna & Stanley 2000). To this end, we apply a bootstrapping procedure (Curme *et al.* 2015 in which the rows of the matrix $X^{(t)}$ are shuffled repeatedly in order to construct a distribution for the sample correlation coefficient as measured using uncorrelated signals of the same distribution as the data. We then apply a uniform statistical threshold of $p = 0.01$, with false discovery rate (FDR) correction for multiple comparisons (Benjamini & Hochberg 1995), to obtain thresholds of measured correlation coefficients that vary for each time series pair. Thus, we construct the four different $X^{(t)}$ and $X^{(t+1)}$ matrices described above, perform $100 \times N^2 = 100 \times (80)^2 = 640000$ independent shufflings of the data, construct the distribution for the measured correlation coefficient under the null hypothesis of uncorrelated variables, and accept into our directed network any pair that has a probability $p < 0.01$ of being generated by uncorrelated variables after FDR correction. Further details of this procedure, including the implementation of the FDR correction, are given in Appendix B.

This procedure yields four networks of statistically-validated directed links. In the subsequent portions of the paper, we will both analyze the structure of these networks, and explore their utility as a feature-selection tool in developing prediction models.

We note that special care must be taken when interpreting the lagged relationships described above. A validated link from the United States to Japan, for example, suggests that market movements or changes in sentiments in the U.S. may impact those in Japan on the following day. Due to the location of the international dateline, this timescale may be shorter than the timescale represented by a validated link from Japan to the U.S. We adopt this approach due to its simplicity, although more nuanced approaches are certainly possible, particularly with intra-day data.

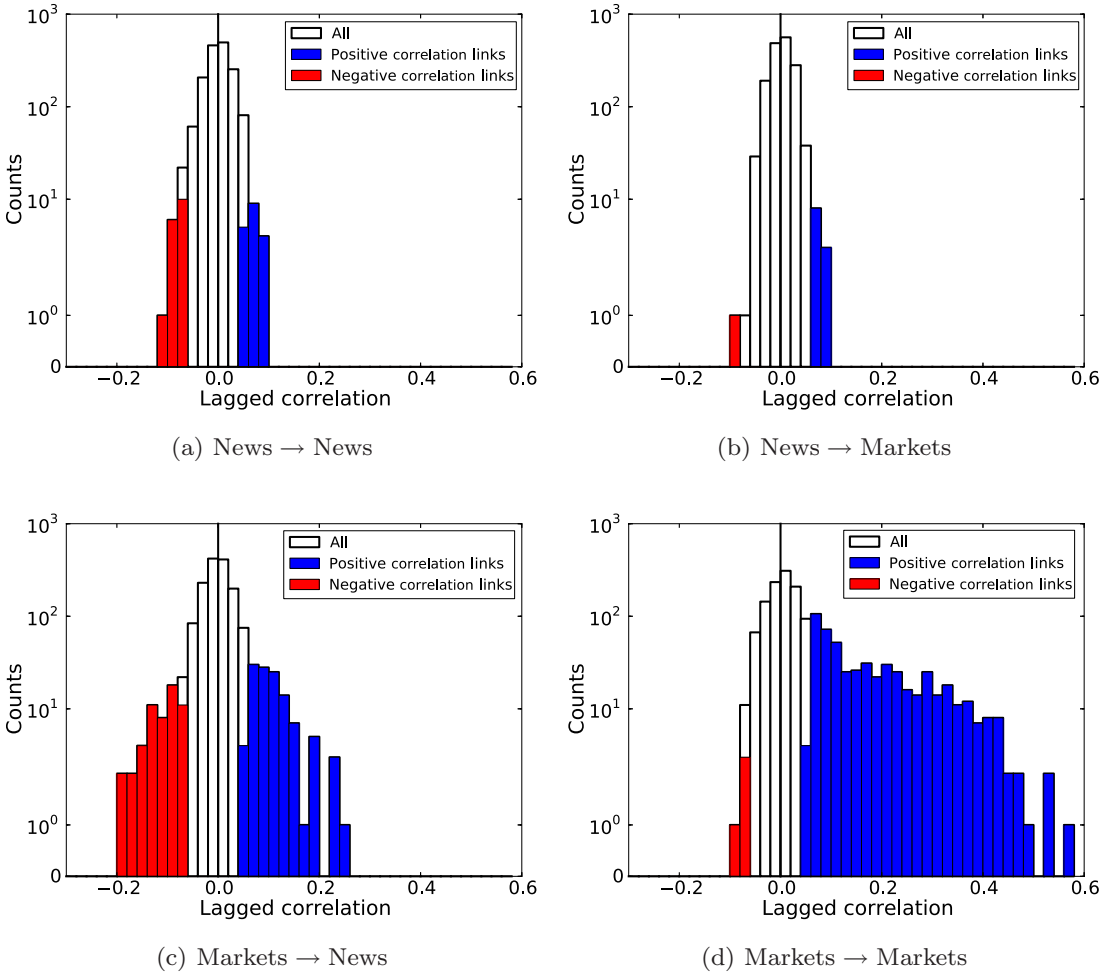
3.2. *Results*

In Fig. 1, we display histograms of measured lagged correlation coefficients separately for relationships among news sentiment signals, among market returns, and between news and markets. The histograms are shaded according to the number of links that are validated according to the statistical validation procedure described above. The corresponding subgraphs of the validated lead-lag relationships are displayed in Fig. 2, where we preserve the geographical location of each node. We distinguish positive and negative correlations by the colors of the links.

We find that the greatest number of validated links are between financial markets, with 534 links of positive correlation and 4 links of negative correlation. There is also a substantial number of links leading from markets to news sentiments, as we validate 118 links of positive correlation and 56 links of negative correlation. By contrast, we find far fewer entities, among both news sentiments and market returns, that are lead by news. In this sense, we find that the system is primarily driven by market movements, which complements our study of synchronous correlations where we find that the markets compose the base of the MST (see Appendix C). A comparison of the distributions of correlation coefficients in which news leads markets to those in which markets lead news, as displayed in Fig. 1, again suggests that the stronger relationships are those in which the markets anticipate news sentiment.

At the level of individual lead-lag relationships, then, we find that the strongest correlations are those that are driven by market movements. To analyze the higher-level structure of the networks, we make use of a well-known clustering algorithm involving a spectral decomposition of the adjacency matrix A, where $A_{i,j} = 1$ if a link exists from i to j, and 0 otherwise. Here, we consider the full $N \times N = 80 \times 80$ adjacency matrix that is the union of the graphs displayed in Fig. 2.

Clustering in networks is commonly studied using a spectral decomposition of the underlying adjacency matrix. In the case of a symmetric matrix with undirected links, as in a network defined by synchronous correlations, an eigendecomposition of the matrix A or its Laplacian can reveal groups of nodes that cluster together, in



(a) News → News

(b) News → Markets

(c) Markets → News

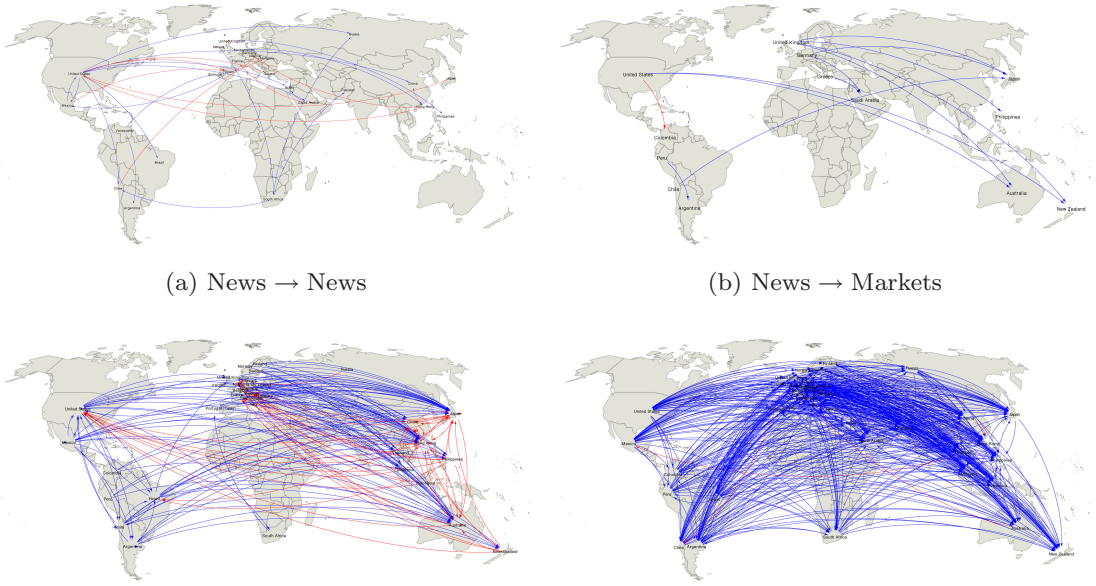
(d) Markets → Markets

Fig. 1. (Color online) Histograms of lagged correlation coefficients (a) among news sentiment signals, (b) in which news anticipate market movements, (c) in which market movements anticipate news, and (d) among market movements. Shading indicates positive (blue) and negative (red) coefficients of pairs that are filtered into the statistically validated network. Note that the distributions are presented on a semi-logarithmic scale, exaggerating the positive and negative tails.

the sense of sharing many links (Chung 1997). The interpretation of the eigenvectors and eigenvalues is less straightforward in directed networks, as the adjacency matrix A is asymmetric and we will generally obtain complex eigenvalues and eigenvectors. The singular value decomposition (SVD), however, has been shown to be a simple method to reveal clustering in even directed graphs (Drineas *et al.* 2004). The SVD is a matrix factorization of the form

$$A = U\Sigma V^{\dagger}, \tag{3.2}$$

where in the special case of an $N \times N$ matrix A, U is an $N \times N$ unitary matrix composed of the eigenvectors of AA^T, and $V^{\dagger}$ is the conjugate transpose of an $N \times N$ unitary matrix V, whose columns are composed of the eigenvectors of $A^T A$. Σ is an $N \times N$ diagonal matrix with entries σ_n that are the real square roots of



(a) News → News

(b) News → Markets

(c) Markets → News

(d) Markets → Markets

Fig. 2. (Color online) Plots of each subgraph of the statistically validated network. (a) Shows lagged relationships among news sentiment signals; (b) shows lagged relationships from news signals to market returns; (c) shows lagged relationships from market returns to news signals, and (d) shows lagged relationships among market returns. Blue color indicates validated links of positive correlation; red color indicates validated links of negative correlation. Network visualizations are prepared with the Cytoscape software framework (Shannon *et al.* 2003).

the eigenvalues of U or V. The columns of U and V are known as the left- and right-singular vectors of A, respectively, and the diagonal entries σ_n of Σ are known as the singular values of A.

In the case of directed networks, it has been shown that the SVD of the adjacency matrix A can reveal bipartite subgraphs of the network (Taylor *et al.* 2011). Informally, each entry (i, j) of AA^T is the number of nodes k to which there is an edge from both i and j, i.e. the number of common successor nodes between i and j. The eigenvectors of this matrix then represent groups of nodes that share common successors. Similarly, each entry (i, j) of $A^T A$ is the number of nodes k from which there is an edge to both i and j, i.e. the number of common predecessor nodes between i and j. The eigenvectors of this matrix then represent groups of nodes that share common predecessors.

Taylor *et al.* (2011) prove, in idealized cases of networks composed entirely of fully-connected nonoverlapping bipartite structures, that each pair of left- and right-singular vectors corresponds to a bipartite subgraph: the nonzero entries of the left-singular vector are nodes in one layer of the bipartite structure; the nonzero entries of the right-singular vector are nodes in the second layer of the structure, and edges are drawn from the nodes in the left-singular vector to the nodes in the right-singular vector. Furthermore, each singular value gives the geometric mean of the number of nodes represented in the corresponding left- and right-singular

Table 2. Largest five components of the first three left- and right-singular vector pairs. Entries refer to market indices, unless otherwise specified as news.

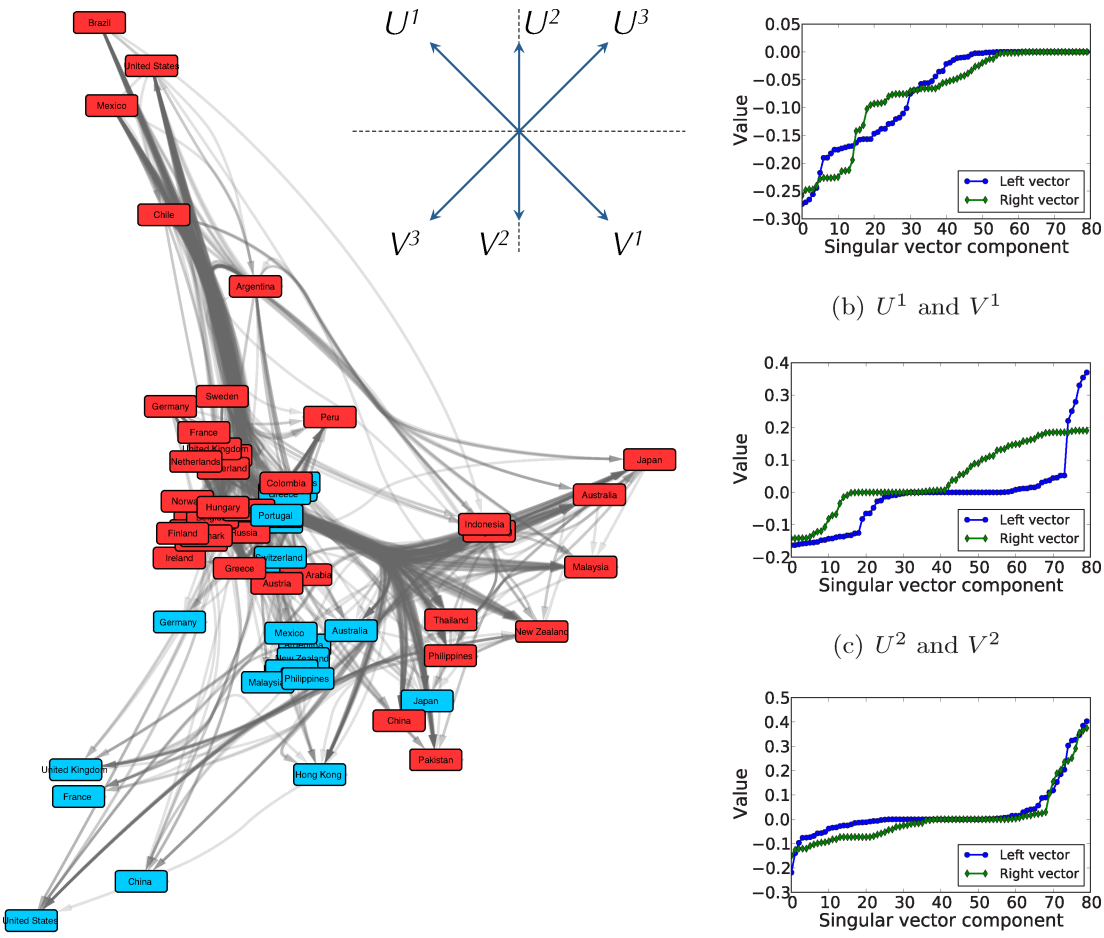
σ_1	U^1	V^1
21.9	United States	New Zealand
	Mexico	Philippines
	Brazil	Australia
	Chile	Japan
	Argentina	Malaysia
σ_2	U^2	V^2
9.74	United States	France
	Mexico	United Kingdom
	Brazil	Sweden
	Chile	Finland
	Saudi Arabia	Belgium
σ_3	U^3	V^3
6.86	Japan	China News
	Australia	United States News
	Philippines	United Kingdom News
	Hong Kong	Hong Kong News
	Malaysia	Japan News

vectors. This holds exactly for the highly-idealized situation described above, but is fairly robust in the presence of noise, such as missing edges or overlapping bipartite structures (Taylor *et al.* 2011).

To describe the large-scale flows in the statistically-validated lagged correlation network, we study the SVD of the full adjacency matrix A. In Table 2, we display the largest five components in magnitude of selected left- and right-singular vectors U^n and V^n of A. Included are the top three singular vector pairs in terms of their corresponding singular value σ_n. Plots of all entries of the first three pairs of left- and right-singular vectors are included in Fig. 3. In Fig. 3, we also plot the full directed network, arranging the positions of nodes according to their entries in the first three singular vector pairs.

We find several approximately bipartite substructures that are embedded in the network. The most prominent consists of financial markets in the Western world — the U.S., Brazil, and Mexico for example — that anticipate the next-day returns of east Asian indices. This is consistent with previous findings (Sandoval 2014), and undoubtedly has much to do with the location of the international dateline. The second singular vector pair indicates that these western markets also have a degree of influence on the next-day returns of European markets.

The third singular vector pair supports our observation that the relation between financial markets and news is asymmetric, as financial markets anticipate news sentiments much more substantially than news sentiments lead market returns. We find that the largest entries in the left-singular vector are entirely composed of



(a) Complete directed network

(b) U^1 and V^1

(c) U^2 and V^2

(d) U^3 and V^3

Fig. 3. (Color online) (a) Display of the complete directed network, showing all links among markets (red) and news sentiment signals (blue). Nodes are arranged according to their entries in the first three left- and right-singular vectors. Specifically, we associate a vector in the plane $\mathbb{R}^2$ to each of the singular vectors U^1, U^2, U^3, V^1, V^2, and V^3 (inset). Each node position in the plane is then a weighted sum of these six 2-vectors, where the weight of each 2-vector is equal to the magnitude of the node's entry in the corresponding singular vector. Edges are bundled according to the algorithm in Holten and van Wijk (2009) to highlight the larger scale flows among groups of nodes. In (b)–(d), we plot the sorted components of the first three pairs of left- and right-singular vectors. For each vector, the largest entries in magnitude tend to be of the same sign. Network visualizations are prepared with the Cytoscape software framework (Shannon *et al.* 2003).

financial markets, largely from Asia, and the largest entries of the right-singular vector are entirely composed of news sentiment signals.

4. Relation Between the Structure of the Statistically-Validated Network and Prediction Model Performance

We further investigate the predictability of node signals within the statistically-validated lead-lag network. We first divide our data into a training set from 2002 to

the end of 2010, and a testing set from 2011 to the end of 2012. We construct the statistically-validated network, using the methodology described above, with only the training subset of the data.

We then employ the networks as a feature-selection step in the training of a classifier. We aim to predict the sign (+1 or −1) of the signals $\tilde{r}_{i,t}$ and $\tilde{s}_{i,t}$, using both the most recent previous index returns and news sentiment data. For each node, we exclude days of sign zero from the training and test sets, allowing us to train a genuinely binary classifier.

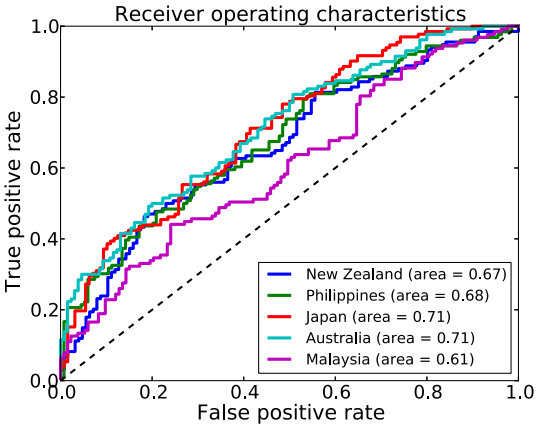
When modeling a given node i, we use all nodes j as inputs for which there is an edge from j to i in the statistically-validated network constructed from the training data. The number of inputs to each logistic regression, therefore, is equal to the in-degree of the desired node. For each node, we assemble the lagged input signals $\tilde{r}_{i,t}$ and $\tilde{s}_{i,t}$ as columns in a matrix X. Signals are lagged as in Sec. 3.1, and standardized to Z-scores by subtracting the mean and scaling by the standard deviation of the training set of each column. We then fit a logistic regression using the training data from 2002 through 2010, and test on data from 2011–2012. For a row vector $\mathbf{x}$ of X, the logistic regression models the probability for an upward movement in $\tilde{r}_{t+1}$ for a desired market as

$$\Pr(\tilde{r}_{t+1} > 0 \,|\, \mathbf{x}) = \frac{e^{\beta_0 + \boldsymbol{\beta}\cdot\mathbf{x}}}{1 + e^{\beta_0 + \boldsymbol{\beta}\cdot\mathbf{x}}}, \tag{4.1}$$

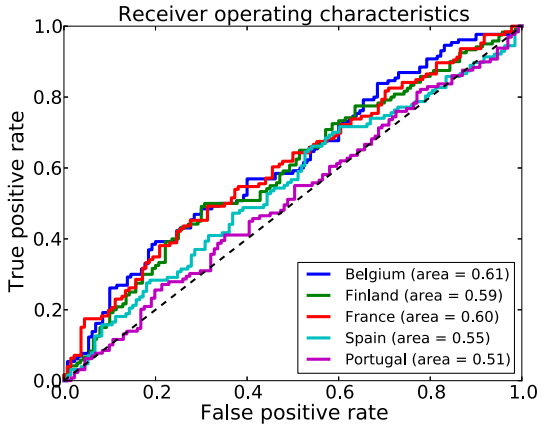
where $\boldsymbol{\beta}$ is a vector of coefficients to be fit with maximum likelihood estimation. If this probability is greater than some threshold, the model predicts an upward movement; otherwise the model predicts a downward movement. We predict news sentiment signals $\tilde{s}_{i,t}$ in exactly the same way. No regularization is used when fitting $\boldsymbol{\beta}$.

We evaluate the performance of each model on the test data by constructing its receiver operating characteristic (ROC) curve, which is generated by varying the threshold probability for an upward movement and computing the corresponding rates of true and false positives. The ROC curve is widely used in measuring the ability of a classifier to discriminate between two classes of events — in this case, upward and downward movements of the signals $\tilde{r}_{i,t}$ and $\tilde{s}_{i,t}$. The performance of each model can be quantified using the area under the curve (AUC) of the corresponding ROC curve. The AUC exhibits a number of desirable properties, including its invariance to the proportions of positive and negative events in the data (Bradley 1997).

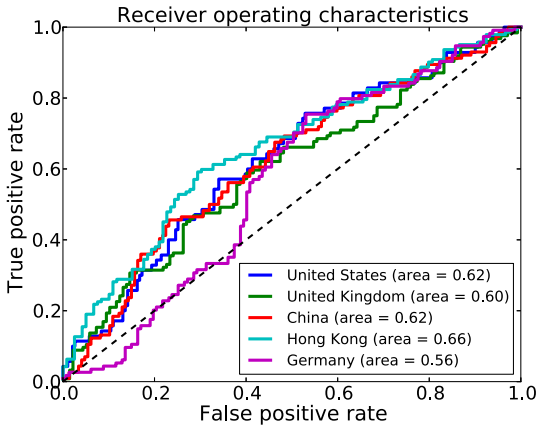
In Fig. 4, we plot some sample ROC curves for 15 of the logistic regression models. In particular, we repeat the SVD on the adjacency matrix for the network constructed from the training data, and plot the ROC curves for the largest five entries of the right singular vectors $|V^1|$, $|V^2|$, and $|V^3|$ (note the large overlap of these entries with those from Table 2, which was constructed from the full data set). The notation $|V^i|$ indicates the vector of absolute values of the entries of V^i. We find that these models perform reasonably well on the test data.



(a) Largest five entries in V^1



(b) Largest five entries in V^2



(c) Largest five entries in V^3

Fig. 4. ROC curves for the performance of the logistic regression model in predicting (a) daily returns $\tilde{r}_{i,t}$ of the stock indices in the top five entries of $|V^1|$ and (b) $|V^2|$, and (c) sentiment scores $\tilde{s}_{i,t}$ for the news signals in the top five entries of $|V^3|$. Each ROC curve is generated by varying the threshold probability for the prediction of a positive return. We provide the area under each curve in the legend.

We compare the performance of all logistic regressions, using only the inputs as defined by the validated network, with the performance of models that use all 80 nodes as inputs in the vector $\mathbf{x}$. In Fig. 5, we show the distributions of differences in AUCs between these two sets of models, finding that in nearly all cases the feature selection step represented by constraining inputs according to the validated network provides for significant gains in accuracy in the test data. The network is thus highlighting persistent relationships among nodes and excluding noisy inputs that may confound predictive models.

Finally, we explore the extent to which information on the predictive relationships among nodes is encoded in the wiring diagram of the validated network's adjacency matrix. In Fig. 6, we plot the AUC for all markets against the magnitude

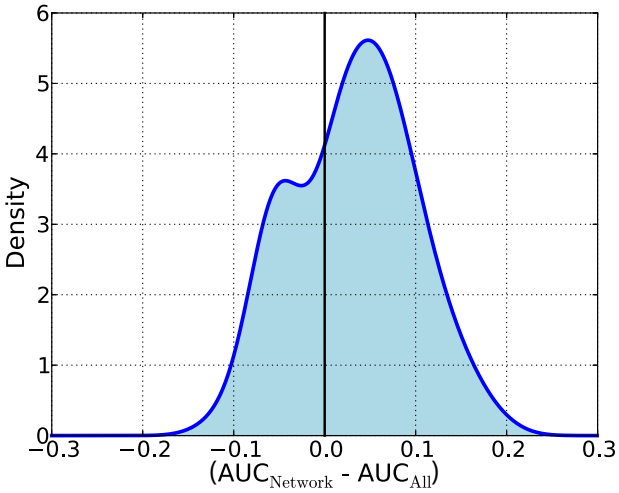
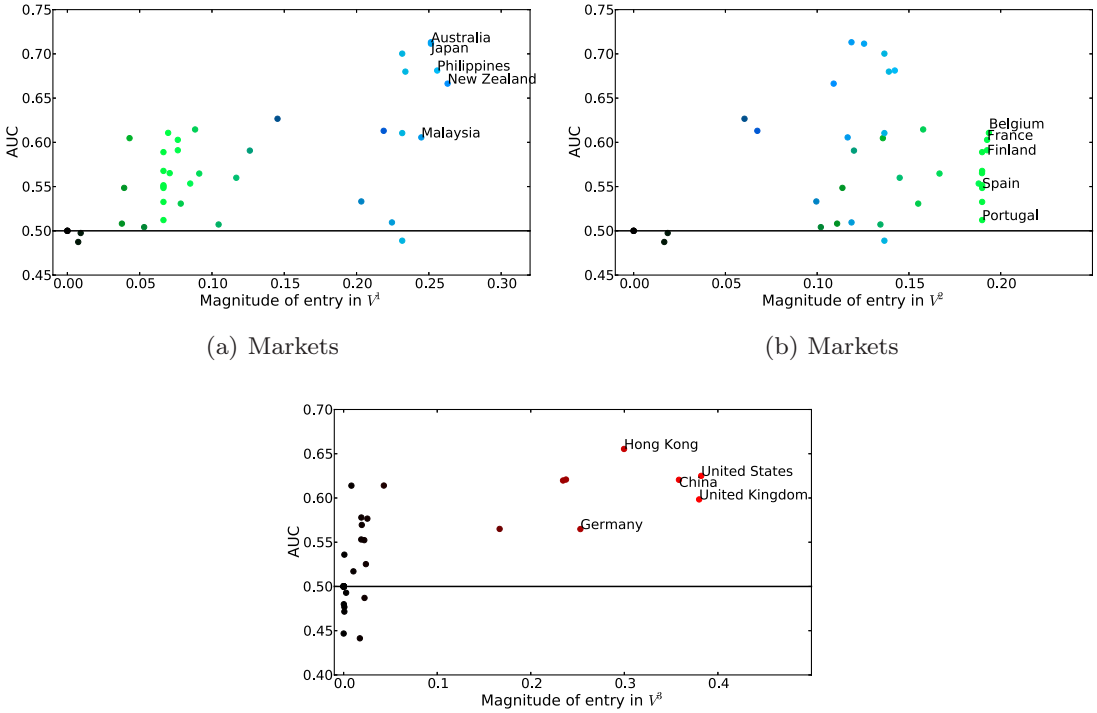


Fig. 5. Pairwise differences in AUCs between models with inputs defined by the validated network, $\mathrm{AUC}_{\mathrm{Network}}$, and those using all possible inputs, $\mathrm{AUC}_{\mathrm{All}}$. The distribution of AUC differences is shown for all news sentiment and market return signals, and is represented using a Gaussian kernel density estimate, with a bandwidth calculated using Silverman's rule of thumb. The median of this distribution differs significantly from zero according to a nonparametric Wilcox test ($V = 2407$, $p < 0.001$), suggesting that the networks constructed using the training data uncover persistent lead-lag relationships, and that restricting model inputs to the nodes defined by these networks offer improved model performance.

of the entry of each market in the right-singular vectors V^1 and V^2. Similarly, we plot the AUC for all news sentiments against the magnitude of the entry of each node in the right-singular vector V^3. We find that the majority of market indices cannot be reliably predicted using data at a time horizon of one day, in accordance with the efficient market hypothesis (Fama 1970). However, there does exist a group of nodes that exhibits a considerable degree of predictability, and these are precisely the nodes identified in the first right-singular vector V^1 of the adjacency matrix of the full network. Similarly, the most predictable signals among news sentiments are those with the highest entries (in magnitude) in the right-singular vector V^3, as shown in Fig. 6(c). These numerical demonstrations suggest that the SVD of the lagged correlation network's adjacency matrix may be a plausible method for identifying predictable subsets of nodes in networks built according to lagged correlations.

We also investigate the extent of the information encoded in the left-singular vectors. To this end, for the top five entries in each right-singular vector, we add inputs sequentially to each model, and compute the out-of-sample AUC. We compare the effect of two schemes: in the first scheme, when modeling node i, we choose each additional input at random from all nodes j for which there is an edge from j to i in the validated network. In the second scheme, we choose each additional input in the order of their ranking in the corresponding left-singular vector. In Fig. 7, we plot the mean AUC for the top five entries of each right-singular vector against the number of inputs in each model. We find that, when modeling the signals of nodes



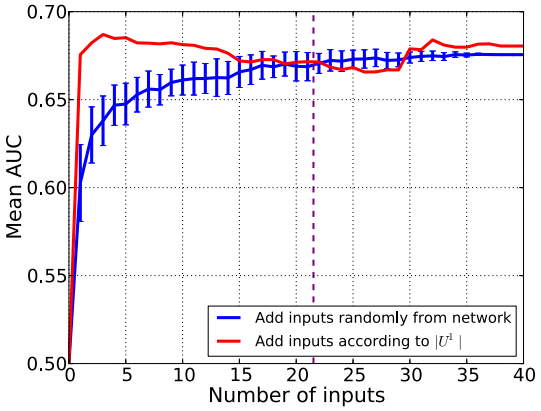
(a) Markets

(b) Markets

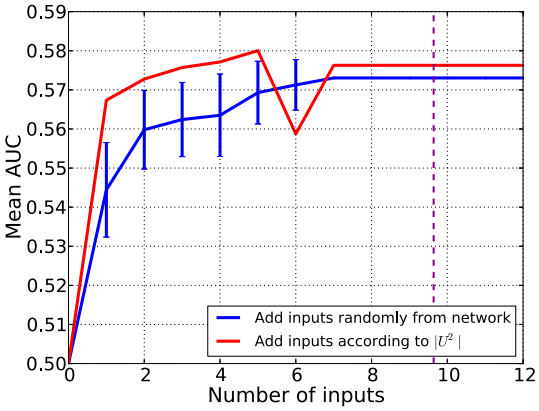
(c) News

Fig. 6. (Color online) AUC for all markets against (a) the magnitude of the entry of the corresponding element of V^1, and (b) the magnitude of the entry of the corresponding element of V^2. Points are shaded blue according to the magnitude of the entry in V^1, and green according to the magnitude of the entry in V^2. In (c) we plot the AUC for all news against the magnitude of the corresponding entry in V^3 (additionally shaded in red). We observe that the right-singular vectors of the adjacency matrix identify subsets of predictable nodes.

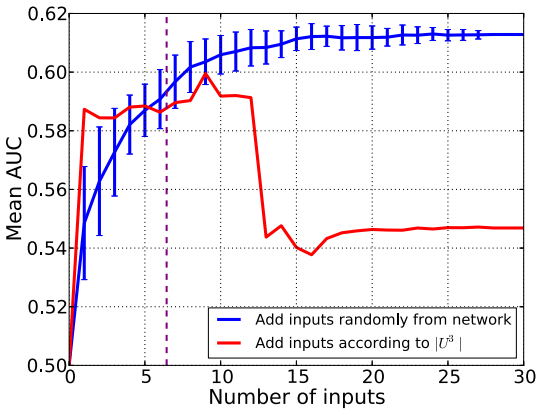
highlighted in the right-singular vectors, the corresponding nodes highlighted in the left-singular vectors tend to represent the most important inputs to the model. In the case of the nodes in V^1 and V^2, for a node of in-degree k_{in}, the largest k_{in} components of U^1 or U^2 using as inputs will, on average, result in better model performance than using the inputs selected by the network. The effect is weaker for the nodes in V^3, although choosing nodes from the largest components of the left-singular vector U^3 still yields comparable model performance to choosing them from the underlying network, up to the singular value corresponding to this singular vector pair (which, as in Taylor *et al.* 2011 approximates the geometric mean of the number of nodes involved in the large-scale flow). Whereas the right-singular vectors identify subsets of predictable nodes, then, the corresponding left-singular vectors seem to identify the most important inputs to these nodes, with respect to the performance of our prediction models. We therefore find that the network's adjacency matrix alone can offer nontrivial insights into global flows of information. In Appendix D, we verify that this finding holds on synthetic data with the same correlation structure as the empirical data.



(a) Elements of V^1



(b) Elements of V^2



(c) Elements of V^3

Fig. 7. (Color online) AUC, as averaged among the top five nodes in $|V^1|$ (markets), $|V^2|$ (markets), and $|V^3|$ (news), for each additional model input. When the number of model inputs exceeds the in-degree of a node, we cease adding inputs. In blue, we plot the mean AUC when randomly adding input nodes from the validated network, as averaged over 50 iterations. In red, we plot the mean AUC when input nodes are added in order of their magnitudes in the corresponding left-singular vectors, regardless of the presence or absence of a link in the validated network. Dashed vertical lines mark the singular value associated with each singular vector pair, approximating the number of nodes involved in the large-scale flow. We find that the most important inputs to nodes with large weight in the first three right-singular vectors are nodes with large weights in the corresponding left-singular vectors.

5. Conclusions

In summary, we have studied the structure of lagged correlation-based networks that are derived from a collection of index returns and news sentiment data of 40 countries. Although the methods used to build the networks have no *a priori* information about whether a time series describes news sentiment or market returns, we find that these two classes of nodes play vastly different roles in the structure of the networks. In particular, the dynamics of the system seem to be most strongly driven by the financial markets, as these nodes are the sources of the strongest

correlations in the system. We find that, at a time resolution of one day, market movements seem to anticipate news sentiments much more substantially than news sentiments anticipate market movements.

The networks considered here not only reveal information about the structure of the system; they also serve to identify nodes that exhibit some degree of predictability, as quantified with the out-of-sample performance of simple logistic regression models. We note that the most predictable markets, in East Asia, naturally follow market movements in the Western World due to the location of the international dateline. In addition, although these lagged relationships are persistent, they may not be actionable, as the trading hours of different markets do not necessarily overlap.

The SVD of the adjacency matrix of the lagged correlation network reveals pairs of group of nodes, and associates a directionality to the pair, in the sense that the group of nodes identified in the left-singular vector tends to lead the group of nodes identified in the right-singular vector. This simple transformation can be useful in large directed networks, where we may abstract away from individual nodes in order to identify larger-scale flows. In the context of correlation-based networks, we have found some evidence that the large-scale structures identified with this method correspond to groups of predictable nodes and their important inputs, as quantified using out-of-sample tests. Although we do not suggest that these methods could outperform conventional feature-selection algorithms, such as regularization, the results support the idea that the structures we find are representative of genuine flows of information among global markets and news outlets. According to this analysis at a daily granularity, we find that the directionality is decidedly from markets to news, and not the reverse.

A possible application of similar analyses in the context of lagged-correlation networks would be a "recommender system" for exogenous inputs in time series models. A preliminary feature-selection, such as the construction of a statistically-validated network, is always subject to false negatives or false positives. A simple SVD allows one to "recommend" inputs for a model according to the inputs of other nodes — other time series — that otherwise share similar inputs. As demonstrated here, incorporating such inputs can potentially improve performance, though the limitations of this approach are evident in Fig. 7(c). This approach could also be refined with more sophisticated recommender systems, although we make no claims about the statistical basis for the functioning of these systems.

The use of Pearson correlation is certainly a limitation of this work, as we can provide no evidence for "predictive causality", in the sense of Granger (1969). We note that in our approach, we detrend all time series for autocorrelation in order to control for the endogenous structure of each time series. This work could be extended through the incorporation of more nuanced time series analyses. We could additionally control for other exogenous factors, such as fluctuations in exchange rates; our preliminary analyses suggest, however, that the influence of daily fluctuations in exchange rates would only minimally impact our conclusions.

This work could also be expanded to analyses of intra-day data. One could construct a different statistically-validated network for every pair of consecutive hours or minutes in the day, for instance (Curme *et al.*, unpublished results). This would allow one to trace the flows of information during each 24 h period. Finer levels of time horizon could also reveal more detailed interactions between world news and the returns of major financial markets, and could perhaps better capture the influences of news on market movements.

Acknowledgments

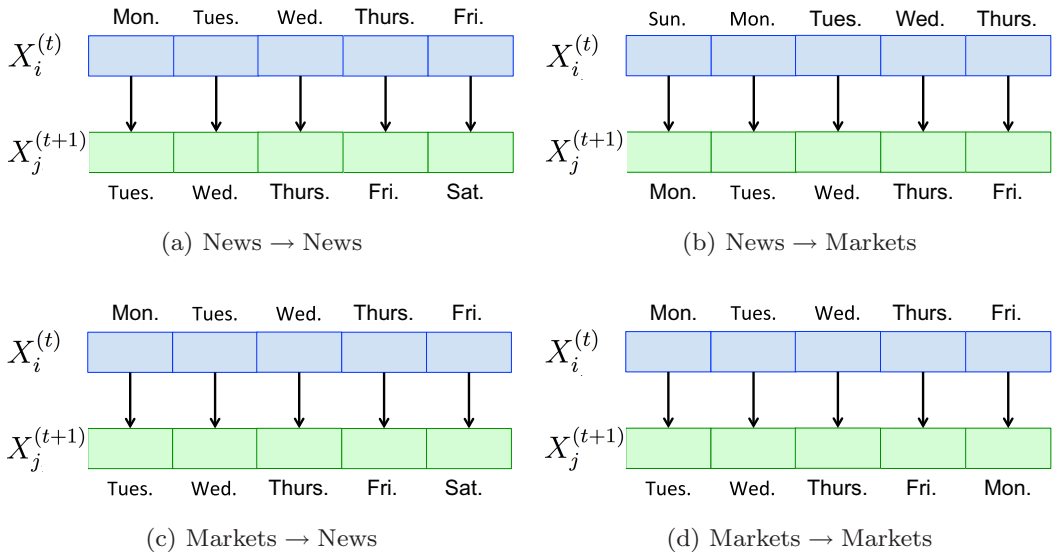
We thank Thomson Reuters for providing the MarketPsych Indices (TRMI) data, the European Commission FET Open Project "FOC" 255987 and "FOC-INCO" 297149, and NSF Grant "SES 1452061".

Appendix A. Lagging Procedure

In this work, we study Pearson correlations among news sentiment signals and market returns at one-day lag. While the news sentiment data is available seven days per week, the market return data only exists at most between Monday and Friday of each week. To account for this difference, we adopt the following scheme, which is shown in Fig. A.1.

- For correlations between financial markets, we include products between returns on Friday and those on the following Monday in the Pearson product-moment sum, using all available data.
- For correlations between news data and subsequent market movements, we relate news sentiment data between Sunday and Thursday of each week with market data from Monday to Friday.
- For correlations between market movements and subsequent news data, we relate market data from Monday to Friday with news sentiment data between Tuesday and Saturday of each week.
- For correlations between news sentiment data, we relate news sentiments between Monday and Friday of each week with those from Tuesday to Saturday of each week. This method allows for a comparison between the effects of market returns and news sentiment signals on subsequent news sentiments.

This scheme maintains a five day week, and therefore a constant time series length T, for all relationships studied. We also use all available market data. An alternative scheme is to simply synchronize all time series, removing data from Saturdays and Sundays, as is done in Appendix C. We would then simply correlate each time series against time series that have been shifted by one day. We have checked to confirm that this change only weakly impacts the results. We find no systematic differences between the distributions of the signals $\tilde{s}_{i,t}$ during the week and on the weekends.



(a) News → News (b) News → Markets

(c) Markets → News (d) Markets → Markets

Fig. A.1. Diagram of lagging procedure for measuring lagged correlations. We maintain a five day week, and therefore a constant time series length T, for all four classes of relationships. This scheme uses all available market data, but only includes terms that are spaced exactly one day apart when possible.

Appendix B. Statistical Validation of Directed Links

We aim to filter the lagged correlation coefficients in L according to a threshold of statistical significance. In this high dimensional setting, composed of signals that are by no means normally distributed, it can be difficult to infer the joint probability distribution of the data (Tumminello *et al.* 2007). We will thus apply a bootstrapping procedure (Efron & Tibshirani 1993) in order to determine the statistical significance of each entry of L separately, and filter L according to a statistical threshold. Although this threshold is uniform among all measured lagged correlations, the lagged correlation coefficient corresponding to this threshold will vary with the distributions of each pair of signals under consideration. See Curme *et al.* (2014) for an analysis of this method when applied to intraday stock returns.

According to this procedure, the rows of the matrix $X^{(t)}$ are shuffled repeatedly in order to construct a distribution for the sample correlation coefficient as measured using uncorrelated signals of the same distribution as the data. Upon each shuffling, we create 40 surrogated time series, recalculate the lagged correlation matrix, and compare this "surrogate" lagged correlation matrix $\widetilde{L}$ to the empirical matrix L. This is done separately for each scenario under consideration (e.g. news time series in $X^{(t)}$ and market returns in $X^{(t+1)}$, or market returns in $X^{(t)}$ and news time series in $X^{(t+1)}$, etc.). We then construct the matrices U and D, where $U_{m,n}$ is the number of shufflings for which $\widetilde{L}_{m,n} \geq L_{m,n}$, and $D_{m,n}$ is the number of shufflings for which $\widetilde{L}_{m,n} \leq L_{m,n}$.

From matrix U, we associate a one-tailed p-value with all positive correlations as the probability of observing a correlation that is equal to or higher than the

empirically-measured correlation, under the null hypothesis of uncorrelated signals. From D we may similarly associate a one-tailed p-value for all negative correlations. We choose our statistical threshold to be $p = 0.01$. Because we are performing many statistical inferences simultaneously, however, we must correct our p-values to account for multiple comparisons. We use the FDR (Benjamini & Hochberg 1995) protocol to correct all N^2 p-values. According to this correction, the p-values from each individual test are arranged in increasing order ($p_1 < p_2 < \cdots < p_{N^2}$), and the threshold is defined as the largest k such that $p_k < k\ 0.01/N^2$. In this case, for $N = 80$ nodes, we must construct $100N^2 = 640000$ independently shuffled surrogate time series. We may then interpret $U_{m,n}/(100N^2)$ as the p-value for the positive one-tailed test, and $D_{m,n}/(100N^2)$ as the p-value for the negative one-tailed test. Directly from the matrices U, and D, then, our threshold is the largest integer k such that U or D has exactly k entries fewer than or equal to k. From this threshold, we can filter the links in L to construct the FDR network (Tumminello *et al.* 2011).

Appendix C. Analysis of Synchronous Correlations

C.1. *Methodology*

In this section, we analyze the synchronous (same-day) relationships among the market returns and news sentiment signals. For this purpose, we synchronize the signals and assemble them as $N = 80$ columns in a matrix X. We then construct the correlation matrix C of the columns of X. Each element of C is given by the Pearson correlation

$$C_{i,j} = \frac{1}{T-1} \sum_{t=1}^{T} \frac{(X_{i,t} - \langle X_i \rangle)(X_{j,t} - \langle X_j \rangle)}{\sigma_i \sigma_j}, \tag{C.1}$$

where X_i is the ith column of X, $X_{i,t}$ is row t of column i of X, T is the number of rows of X, and $\langle X_i \rangle$ and σ_i are the mean and sample standard deviation of X_i, respectively.

To study the structure of the correlation matrix C, we next construct the "distance" matrix D (Mantegna & Stanley 2000). Each element of D is given by

$$D_{i,j} = \sqrt{2(1 - C_{i,j})} \tag{C.2}$$

and can be understood as a distance in the following sense. Each column X_i can be normalized to $\tilde{X}_i \equiv (X_i - \langle X_i \rangle)/(\sqrt{T-1}\sigma_i)$, so that $\tilde{X}_i$ is a unit vector. It is then readily seen that $C_{i,j}$ is the dot-product $\tilde{X}_i \cdot \tilde{X}_j$, and $D_{i,j}$ is the distance $\|\tilde{X}_i - \tilde{X}_j\|$.

The hierarchical structure and clustering represented in the matrix D can be visualized using the MST (Mantegna & Stanley 2000). If each time series X_i of our data is considered a node in a graph, and an edge between any two X_i and X_j is weighted by the distance $D_{i,j}$, then the MST is the tree structure that links all of the nodes and minimizes the sum of the edge weights. The MST is commonly constructed using Kruskal's algorithm (Kruskal 1956). Alternative specifications of the distance $D_{i,j}$ invoke the absolute value or square of the correlation $C_{i,j}$, so

that perfectly anti-correlated series will have a small distance (Mantegna 1999). We choose the function $D_{i,j}$ as we do because (i) it fulfills the three axioms of a metric distance (Mantegna 1999), and (ii) a large majority of the correlations studied are positive, so we verify that the resulting MST is unaffected by this choice.

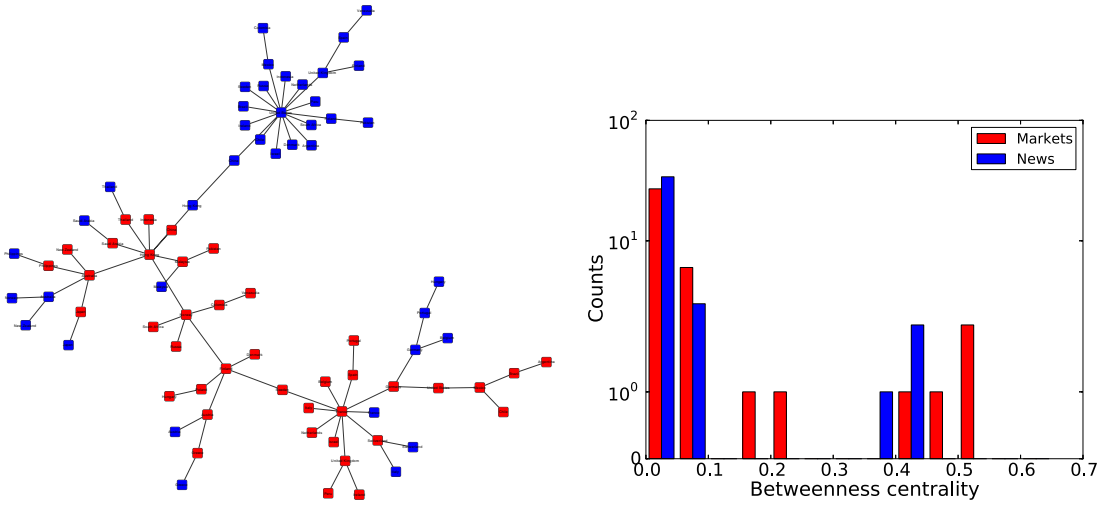
C.2. *Results*

We plot the MST of the data X in Fig. C.1(a), and observe a structure in which the "backbone," or highest-level organization is defined by the financial markets. The lowest-level of the hierarchy, or "leafs" of the tree, are commonly the news sentiment signals. This is corroborated by Fig. C.1(b), which displays histograms of the betweenness centrality for the financial market nodes and news sentiment nodes separately. The betweenness centrality of a node n is given by (Freeman 1977)

$$g(n) = \sum_{m \neq n \neq p} \frac{\sigma_{mp}(n)}{\sigma_{mp}}, \tag{C.3}$$

where σ_{mp} is the total number of shortest paths from node m to node p, and $\sigma_{mp}(n)$ is the number of those paths that pass through node n.

Furthermore, the news sentiment signal nodes are in most cases linked to their corresponding market. We thus find that the strongest correlations are among financial markets, which compose the highest-level of the hierarchy, with weaker correlations between news sentiments and the corresponding market. This may be because



(a) Minimum Spanning Tree

(b) Associated betweenness centrality

Fig. C.1. (Color online) (a) Plot of the MST of the synchronous correlations. Financial markets are colored red; news sentiment signals are colored blue. (b) Histogram of the betweenness centrality of financial markets and news sentiment separately. We find that the strongest correlations in the system are among financial markets, and between the news sentiment signals of a country and the same country's market returns. The notable exception is the node corresponding to news sentiment signals from the United States, which is strongly correlated with news from a host of other countries and so represents a hub in the network.

traders around the world observe other market movements and often take actions based on the dynamics of these markets. This may represent a signature of quasi-technical analysis, making trading decisions based on other market patterns. Moreover, we find that US news sentiments are strongly correlated with news sentiments in countries around the world. Our results are consistent with the idea that traders turn to local news for market information, while their actions (market movements) are reported in both local and international news.

Appendix D. Tests with Synthetic Data

In this section, we test the efficacy of the "recommender system" for time series model features using synthetic data. There are two broad purposes to such a study. First, we verify that our conclusions are not strictly dependent on the particular real-world dataset that we choose, and that our findings extend to other datasets satisfying a particular set of properties. Second, the use of synthetic data allows us to determine what that set of properties is, so that we may understand the scope and limitations of our methodology.

To this end we generate many simulated time series with the same underlying correlation network as the real-world data. By varying the strength of the correlations, we examine the range over which our method — selecting model inputs according to their ranking in the corresponding left-singular vector of the adjacency matrix — outperforms the "null" model of simply choosing inputs according to the adjacency matrix alone.

We generate N simulated time series of length T in an iterative fashion. The state of the system at time t can be described by an N-dimensional vector $\mathbf{x}_t$, which is updated according to the state at time $t-1$ as a vector-autoregressive process

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t. \tag{D.1}$$

Here, $\mathbf{B}$ is a matrix of fixed coefficients and $\boldsymbol{\epsilon}_t$ is an N-vector of error terms. We specify $\mathbf{B}$ and the distribution of $\boldsymbol{\epsilon}_t$ so that the resulting time series have a lagged correlation matrix L and synchronous correlation matrix Σ that is in agreement with empirical data. In particular, through the matrix $\mathbf{B}$, we will embed the same underlying lagged correlation network as was recovered from the empirical data. Scaling these correlations by a factor α allows us to test how a varying signal-to-noise ratio influences our results.

We use as our estimate of $\mathbf{B}$ the ordinary least squares (OLS) result

$$\mathbf{B} = (X^T X)^{-1} X^T Y. \tag{D.2}$$

Here, X is a $T \times N$ matrix, entry (t, i) of which gives the value of the time series of node i at time t. Similarly, Y is a $T \times N$ matrix, entry (t, i) of which gives the value of the time series of node i at time $t+1$. If these time series have zero mean and unit variance, we recognize the quantity $X^T X$ as $T\Sigma$, proportional to the synchronous

correlation matrix. Further, we recognize the quantity $X^T Y$ as TL, proportional to the lagged correlation matrix. We therefore fix

$$\mathbf{B} = \alpha \Sigma^{-1} L. \tag{D.3}$$

We take Σ to be the empirical synchronous correlation matrix of the system, and L to be the weighted adjacency matrix for the validated lagged correlation network: that is, each entry (i, j) of this matrix has a value equal to the lagged correlation between nodes i and j, if a link was validated from node i to node j, and zero otherwise. Further, we set the distribution of the error terms $\boldsymbol{\epsilon}_t$ to be multivariate normal with correlation matrix Σ. The factor α allows us to control the strength of the lagged correlations in the underlying network.

In this way, we may construct N time series of length T, and find its associated lagged correlation network as before, using FDR correction for multiple comparisons. Because our signals are homogeneously and normally distributed, we filter our network according to a Gaussian threshold correlation corresponding to $p < 0.01$. This simplification allows us to generate large numbers of these systems in a reasonable amount of time. We find that, for $\alpha = 1$, the properties of the resulting system — namely, its synchronous correlation matrix, lagged correlation matrix, and validated adjacency matrix — match closely our empirical results.

For a given value of α, we generate 500 of these systems, each of which has $N = 80$ nodes and $T = 400$. We compute the SVD of the resulting adjacency matrix, and train logistic regression models in which we attempt to classify the sign ($+1$ or -1) of the signal a given node at each time t. For these nodes we again choose the largest five entries of the first right-singular vector. We will compare the success of these models (measured by the AUC of the corresponding ROC curve) in two cases, just as before. In case (i), when predicting the sign of node j at time t, we use as model features all nodes i at time $t - 1$ for which there is a link from i to j in the validated network. In case (ii), we use the s largest entries of the first left-singular vector as our nodes i, where s is the first singular value of the adjacency matrix.

We then continue our time series for another $T = 100$ time-steps, and measure the AUC of each model in each of cases (i) and (ii) on this held-out data set. In Fig. D.1, we show differences in the measured AUCs for varying values of α. We find that for low values of α, there is no difference between cases (i) and (ii). That is, the lagged correlations in the system are so weak that both methods perform equally poorly. For values of α in the range from roughly 0.5 to 1.5, case (ii) outperforms case (i) by 1%–2%. In this regime, we find that consideration of the network's bipartite community structure can increase the accuracy of our predictions. Note that, if we characterize the strengths of the lagged correlations by what we find in the empirical data (corresponding to $\alpha = 1$), we achieve a near-optimum gain in accuracy. For α larger than 1.5, however, one is much better-off choosing model inputs from the adjacency matrix alone. In this regime, the signal-to-noise ratio is

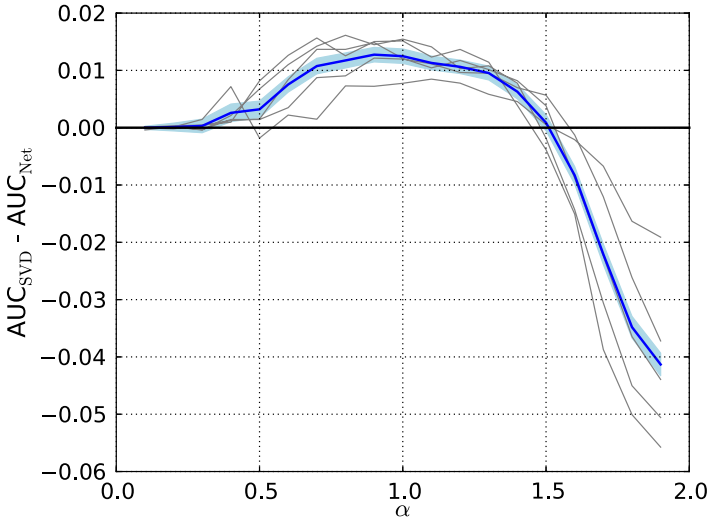


Fig. D.1. (Color online) Pairwise differences in AUCs between logistic regression models with inputs given by case (ii), $\mathrm{AUC}_{\mathrm{SVD}}$, and inputs given by case (i), $\mathrm{AUC}_{\mathrm{Net}}$, for varying α. In grey, we show characteristic trajectories of each of the five largest entries of the first right-singular vector. In blue, we show results as averaged over each of these entries.

sufficiently strong that we find a low rate of false positives and negatives in the validated links, so that our recommender system has little to offer.

References

T. G. Anderson, T. Bollerslev, F. X. Diebold & C. Vega (2003) Micro effects on macro announcements: Real-time price discovery in foreign exchange, *American Economic Review* **93**, 36–82.

T. G. Anderson, T. Bollerslev, F. X. Diebold & C. Vega (2007) Real-time price discovery in global stock, bond, and foreign exchange markets, *Journal of International Economics* **73**, 251–277.

P. Balduzzi, E. J. Elton & T. C. Green (2001) Economic news and bond prices: Evidence from the U.S. treasury market, *Journal of Financial and Quantitative Analysis* **36**, 523–543.

R. Ball & P. Brown (1968) An empirical evaluation of accounting income numbers, *Journal of Accounting Research* **6**, 159–178.

K. G. Becker, J. E. Finnerty & M. Gupta (1990) The intertemporal relation between U.S. and Japanese stock markets, *The Journal of Finance* **45**, 1297–1306.

Y. Benjamini & Y. Hochberg (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*: *Series B* (*Methodological*) **57**, 289–300.

T. J. Brailsford (1996) Volatility spillovers across the Tasman, *Australian Journal of Management* **21**, 13–27.

A. Bradley (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition* **30**, 1145–1159.

T. Chordia, A. Goyal, G. Sadka, R. Sadka & L. Shivakumar (2009) Liquidity and the post-earnings-announcement drift, *Financial Analysts Journal* **65**, 18–32.

F. Chung (1997) *Spectral Graph Theory*. USA: American Mathematical Society.

C. Curme, M. Tumminello, R. N. Mantegna, H. E. Stanley & D. Y. Kenett (2015) Emergence of statistically-validated financial intraday lead-lag relationships, *Quantitative Finance* **15**, 1375–1386.

C. Curme, M. Tumminello, R. N. Mantegna, H. E. Stanley & D. Y. Kenett (unpublished results), How lead-lag correlations affect the intra-day pattern of collective stock dynamics.

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

P. Drineas, A. Frieze, R. Kannan, S. Vempala & V. Vinay (2004) Clustering large graphs via the Singular Value Decomposition, *Machine Learning* **56**, 9–33.

B. Efron & R. Tibshirani (1993) *An Introduction to the Bootstrap*, Vol. 57. CRC Press, New York.

E. F. Fama (1970) Efficient capital markets: A review of theory and empirical work, *The Journal of Finance* **25**, 383–417.

L. Freeman (1977) A set of measures of centrality based on betweenness, *Sociometry* **40**, 35–41.

A. Ghosh, R. Saidi & K. H. Johnson (1999) Who moves the Asia-Pacific stock markets — US or Japan? Empirical evidence based on the theory of cointegration, *The Financial Review* **34**, 159–169.

N. Godbole, M. Srinivasaiah & S. Skiena (2007) Large-scale sentiment analysis for news and blogs, *International Conference on Weblogs and Social Media* (Boulder, CO, March 26–28, 2007).

C. W. J. Granger (1969) Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* **37**, 424–438.

Y. Hamao, R. W. Masulis & V. Ng (1990) Correlations in price changes and volatility across international stock markets, *The Review of Financial Studies* **3**, 281–307.

D. Holten & J. J. van Wijk (2009) Force-directed edge bundling for graph visualisation, *Eurographics/IEEE-VGTC Symposium on Visualization* **28**.

J. B. Kruskal (1956) On the shortest spanning subtree of a graph and the traveling salesman problem, *Proceedings of the American Mathematical Society* **7**, 48–50.

R. N. Mantegna (1999) Hierarchical structure in financial markets, *The European Physical Journal B* **11**, 193–197, http://arxiv.org/pdf/cond-mat/9802256.pdf.

R. N. Mantegna & H. E. Stanley (2000) *An Introduction to Econophysics*: *Correlations and Complexity in Finance*. Cambridge: Cambridge University Press.

Thomson Reuters MarketPsych Indices (TRMI). MarketPsych (2013), https://www.marketpsych.com/data/.

M. Piškorec, N. Antulov-Fantulin, P. K. Novak, I. Mozetič, M. Grčar, I. Vodenska & T. Šmuc (2014) Cohesiveness in financial news and its relation to market volatility, *Scientific Reports* **4**, 5038.

L. Sandoval (2014) To lag or not to lag? How to compare indices of stock markets that operate at different times, *Physica A* **403**, 227–243.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski & T. Ideker (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks, *Genome Research* **13**, 2498–504.

R. H. Shumway & D. S. Stoffer (2011) *Time Series Analysis and its Applications with R Examples*. New York: Springer.

A. Taylor, J. K. Vass & D. J. Higham (2011) Discovering bipartite substructure in directed networks, *Journal of Computer Mathematics* **14**, 72–86.

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

M. Tumminello, S. Miccichè, F. Lillo, J. Piilo & R. N. Mantegna (2011) Statistically validated networks in bipartite complex systems, *PLoS One* **6**, e17994.

N. Vandewalle, Ph. Boveroux & F. Brisbois (2000) Domino effect for world market fluctuations, *The European Physics Journal B* **15**, 547–549.

W. Zhang & S. Skiena (2010) Trading strategies to exploit news sentiment, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (Washington, D.C., May 23–26, 2010), 375–378 (The AAAI Press, 2010).