OFR WORKING PAPER
OFFICE OF FINANCIAL RESEARCH

15-15 | August 13, 2015

How Lead-Lag Correlations Affect the Intraday Pattern of Collective Stock Dynamics

Chester Curme
Boston University
chester.curme@googlemail.com

Michele Tumminello
University of Palermo, Palermo, Italy
m.tumminello@unipa.it

Rosario N. Mantegna
University of Palermo, Palermo, Italy,
and Central European University,
Budapest, Hungary
mantegna@unipa.it

H. Eugene Stanley
Boston University
hes@bu.edu

Dror Y. Kenett
Office of Financial Research
Dror.Kenett@treasury.gov

How Lead-Lag Correlations Affect the Intraday Pattern of Collective Stock Dynamics

Chester Curme[a,*], Michele Tumminello[b], Rosario N. Mantegna[c,d],
H. Eugene Stanley[a], Dror Y. Kenett[a,e,1,*]

[a] *Center for Polymer Studies and Department of Physics, Boston University, Boston, USA*
[b] *Dipartimento di Scienze Statistiche e Matematiche Silvio Vianelli, University of Palermo, Palermo, Italy*
[c] *Dipartimento di Fisica e Chimica, University of Palermo, Palermo, Italy*
[d] *Center for Network Science and Department of Economics, Central European University, Budapest, Hungary*
[e] *U.S. Department of the Treasury, Office of Financial Research*

Abstract

The degree of correlation among stock returns affects the possibility to diversify the risk of investment, and it plays a major role in financial spillover. During the last decade, the increasing level of correlation observed in financial markets has become a threat to market stability. Here, we analyze high frequency data of stock returns traded at the New York Stock Exchange in the periods 2001-03 and 2011-13. In each period we uncouple the factors contributing to the intraday pattern of synchronous correlations, including volatility, autocorrelations and lagged cross-correlations among assets. We find that intraday market dynamics have changed considerably in the last decade, and relate our findings to the dynamics of an underlying network of lead-lag relationships among equities. In particular, while in 2001-03 lagged cross-correlations contributed significantly to the intraday correlation profile, the increased degree of synchronous correlation observed in the period 2011-13 can be associated with the presence of many significant auto-correlations, especially at the end of a trading day, with a stronger coupling between auto-correlations and lagged cross-correlations of returns. The presented method of data analysis could be used to inform policy makers and financial institutions about market efficiency and risk of financial spillover, and could also be helpful for portfolio management.

Keywords: Financial markets, Correlation analysis, Complex systems, Lagged correlations
JEL: G21, D85, N26, G18

1. Introduction

Filtering information out of vast multivariate datasets is a crucial step in managing and understanding the complex systems that underlie them. These systems are composed of many components, the interactions among which typically induce larger-scale organization or structure. A major scientific challenge is to extract insights into the large-scale organization of the system using data on its individual components.

Financial markets are a primary example of a setting in which this approach has value. When constructing an optimal portfolio of assets, for example, the goal is typically to allocate resources so as

*Corresponding author email: chester.curme@googlemail.com (Chester Curme), dror.kenett@treasury.gov (Dror Y. Kenett)

[1] The views and opinions expressed are those of the individual authors and do not necessarily represent official positions or policy of the Office of Financial Research or the U.S. Treasury.

to balance the trade-off between return and risk. As has been understood at least since the work of Markowitz (1952), risk can be quantified by studying the co-movements of asset prices: placing a bet on a single group of correlated assets is risky, whereas this risk can at least in part be diversified away by betting on uncorrelated or anti-correlated assets. An understanding of the larger-scale structure of co-movements among assets can be helpful, not only in the pursuit of optimal portfolios, but also in for our ability to accurately measure marketwide systemic risks (Glasserman and Young, 2015; Kritzman and Li, 2010).

Time series obtained by monitoring the evolution of a multivariate complex system, such as time series of price returns in a financial market, can be used to extract information about the structural organization of such a system. This is generally accomplished by using the correlation between pairs of elements as a similarity measure, and analyzing the resulting correlation matrix. A spectral analysis of the sample correlation matrix can indicate deviations from a purely random matrix (Laloux et al., 1999; Plerou et al., 1999) or more structured models, such as the single index (Laloux et al., 1999). Clustering algorithms can also be applied to elicit information about emergent structures in the system from a sample correlation matrix (Mantegna, 1999). Such structures can also be investigated by associating a (correlation-based) network with the correlation matrix. One popular approach has been to extract the minimum spanning tree (MST), which is the tree connecting all the elements in a system in such a way to maximize the sum of node similarities (Mantegna, 1999; Bonanno et al., 2003; Onnela et al., 2003). Different correlation based networks can be associated with the same hierarchical tree, putting emphasis on different aspects of the sample correlation matrix. For instance, while the MST reflects the ranking of correlation coefficients, other methods, such as threshold methods, emphasize more the absolute value of each correlation coefficient. Researchers have also aimed to quantify the extent to which the behavior of one market, institution or asset can provide information about another through econometric studies (Hamao et al., 1990), partial-correlation networks (Kenett et al., 2010, 2012) and by investigating Granger-causality networks (Billio et al., 2012).

A network is defined as a set of nodes (such as people, companies, or airports), and links that connect the nodes on the basis of interaction or relationship. Such links can be structural, such as well-defined plane routes that connect airports. Alternatively, links can be functional, or derived using similarities between the activities of two nodes, such as similarities in the number of travelers in two airports. Such functional links can be derived using correlation measures, and correlation-based networks have been found to be a very important tool for investigating real world systems Tumminello et al. (2010); Kenett et al. (2010).

In the context of financial markets, the correlation matrix among asset returns is an object of central importance in measuring risk. The filtering procedures described above may reveal statistically reliable features of the correlation matrix (Laloux et al., 1999; Mantegna, 1999; Tumminello et al., 2010), improving both our understanding of the nature of co-movements among assets in financial markets and our ability to accurately measure risk. Much work has also been devoted toward developing more robust measures of correlation that incorporate dynamics (Barndorff-Nielsen and Shephard, 2004; Lundin et al., 1998), especially those dynamics described by intraday patterns in volume, price and volatility (Admati and Pfleiderer, 1988; Ederington and Lee, 1993; Andersen and Bollerslev, 1997; Allez and Bouchaud, 2011).

What is largely missing is an understanding of the drivers of these synchronous correlations, using the properties of the collective stock dynamics at shorter time scales. Here, we apply a statistical methodology, detailed below, in order to study directed networks of lagged correlations among the 100 largest market capitalization stocks in the New York Stock Exchange (NYSE). In particular, we consider data from both the beginning of the previous decade and today. The resulting network representations of the system provide insights into its underlying structure and dynamics. Our analysis reveals how the interplay of price movements at short time scales evolves during a trading day, how it has changed over the past decade, and quantifies how it contributes to structural properties of the synchronous correlation

matrix at longer time scales. For example, we find that unlike in the 2001-03 period, correlations increase throughout the day in the 2011-13 period. Furthermore, auto and lagged correlations play a much more prominent role, compared to what is observed in the 2001-03 data. We find striking periodicities in the validated lagged correlations, characterized by surges in network connectivity at the end of the trading day, which are crucial to account for when modeling equity price fluctuations. We show how these periodicities can refine our understanding of empirical phenomena, such as the Epps effect, and how they may be incorporated into regression models. We subject our analysis to a variety of robustness checks, which are detailed in the Supplementary Information. Our analysis provides a deeper understanding of market risk by focusing on the short-term drivers of collective stock dynamics resulting from lagged and auto-correlations.

In finance, for example, a statistically significant correlation between the price time series of stocks of two companies provides information on their comovement, and provides important information on how they react in times of risk. However, this does not provide the information on how the price movements of one company will influence price movements of a second company. Such a lead-lag relationship Curme et al. (2014) is critical for the understanding of the market dynamics and the underlying mechanisms responsible for it. Thus, in this paper we make use of the SVN methodology to study the structure and dynamics of the U.S. stock market. Our results present evidence for the existence of such lead-lag relationships. By comparing data from from the beginning of the first decade of the century to that of the second, we shed new light on the changes in the market structure, which are potentially related to the current volatile financial reality.

2. Statistically Validated Network methodology

At short time scales, measured synchronous correlations among stock returns tend to be lower in magnitude (Epps, 1979), and lagged correlations among assets may become non-negligible (Toth and Kertesz, 2009; Curme et al., 2014). Hierarchical clustering methods, which rely on a ranking of estimated correlations, will be strongly influenced by statistical uncertainties in this regime. An alternative approach is the use of a thresholding process, admitting all pairwise correlations beyond a threshold as edges in a correlation-based network. The threshholding approach requires fewer assumptions and is less restrictive; however, it requires making an ad hoc choice of the threshold, which is then used for all the variables. Recently, a solution to this issue has been presented through the use of statistically validated networks (Curme et al., 2014). The Statistically Validated Network (SVN) methodology (Tumminello et al., 2011) provides the means to choose a statistically significant threshold for each variable independently, retaining information about the distribution of each individual time-series. We apply this methodology at different points in the trading day in order to explore the intraday pattern of collective stock dynamics.

First, we transform the processed data from price to additive return, using the commonly used transformation

$$r_i(t) = \log(P_i(t + \Delta t)) - \log(P_i(t)). \tag{1}$$

where $P_i(t)$ is the price of stock i at time t, and Δt is the sampling time resolution. We obtain intraday price data from the NYSE Trades and Quotes (TAQ) database (Brownlees and Gallo, 2006).

We perform a lagged-correlation analysis between all possible stock pairs. Lagged-correlation is a standard method in signal processing of estimating the degree to which two series are correlated (see for example (Muchnik et al., 2009; Arianos and Carbone, 2009; Carbone and Castelli, 2003; Carbone, 2009)). The discrete lagged-correlation function between two time series X and Y is given by (Chou, 1975)

$$\rho_{X,Y}(d) = \frac{\sum_{i=1}^{N-d} \left[(X(i) - \langle X \rangle) \cdot (Y(i-d) - \langle Y \rangle) \right]}{\sqrt{\sum_{i=1}^{N-d} (X(i) - \langle X \rangle)^2} \cdot \sqrt{\sum_{i=1}^{N-d} (Y(i-d) - \langle Y \rangle)^2}} \tag{2}$$

where d is the lag used. In this work we use values of $d = \pm 1$. When we consider the case of $d = 0$, then we end up with the standard synchronous Pearson correlation coefficient.

In this work we focus on the returns matrix at the $\Delta t = 15$ minute time horizon, and divide each trading day into non-overlapping Δt parts $(\Delta t_1, \Delta t_2, ..., \Delta t_{26})$. We partition the contributions to each lagged correlation based on the period Δt_i, in order to explore the effects of intraday periodicities in the data. For each time of day, we construct two matrices, A and B. For example, starting with the first 15 minutes of the day represented by Δt_1, then row m, column n of A is the return of stock n during the first 15 minutes (9:30 - 9:45 a.m.) of day m of the data. Row m, column n of B is the return of stock n during the second 15 minutes (9:45 - 10 a.m.) of day m of the data. So the number of rows of A or B is the number of days in the investigated dataset. We then calculate the lagged correlation matrix, where each entry (m, n) is the Pearson correlation coefficient of column m of matrix A with column n of matrix B. This process results in the empirical lagged correlation matrix, $C_{m,n}(\Delta t_i)$.

For each chosen Δt_i, the matrix $C_{m,n}(\Delta t_i) \equiv C$ can be considered a weighted adjacency matrix for a fully connected, directed graph. We aim to filter the links in this graph according to a threshold of statistical significance. To this end we apply a shuffling technique as follows: the rows of A are shuffled repeatedly, without replacement, so as to create a large number of surrogated time series of returns. After each shuffling we recalculate the lagged correlation matrix, and compare this shuffled lagged correlation matrix $\widetilde{C}$ to the empirical matrix C. For each shuffling we thus have an independent realization of $\widetilde{C}$. We then construct the matrices U and D, where $U_{m,n}$ is the number of realizations for which $\widetilde{C}_{m,n} \geq C_{m,n}$, and $D_{m,n}$ is the number of realizations for which $\widetilde{C}_{m,n} \leq C_{m,n}$.

From the construction U we will associate a one-tailed p-value with all positive correlations as the probability to observe, by chance, a correlation which is equal to or higher than the empirically-measured correlation. Similarly, from D we will associate a one-tailed p-value with all negative correlations. In this analysis we choose our threshold to be $p = 0.01$. We must adjust our statistical threshold, however, to account for multiple comparisons. We use the conservative Bonferroni correction for N stocks, so that our new threshold is $0.01/N^2$. Thus, for a sample of $N = 100$ stocks, we construct 10^6 independently shuffled surrogate time series; if $U_{m,n} = 0$ we may associate a statistically-validated positive link from stock m to stock n ($p = 0.01$, Bonferroni correction). Likewise, if $D_{m,n} = 0$, we may associate a statistically-validated negative link from stock m to stock n. In this way we construct the Bonferroni network (Tumminello et al., 2011).

For comparison, for each part of day Δt_i we also construct the network using p-values that are corrected according to the False Discovery Rate (FDR) protocol. This correction is less conservative than the Bonferroni correction, and is constructed as follows. 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 $p_k < k\ 0.01/N^2$. Therefore, for the FDR network, our threshold for the matrices U (or D) is not zero but instead is the largest integer k such that U (or D) has exactly k entries less than or equal to k. From this threshold we may filter the links in C to construct the FDR network (Tumminello et al., 2011).

3. Intraday periodicities

This approach, in which we construct a distinct network for each interval of Δt minutes between 9:30 a.m. and 4:00 p.m., provides a picture of the dynamics of lagged correlations among equities during a characteristic trading day. We uncover consistent, dramatic changes in network connectivity during the trading day, suggesting that collective stock dynamics exhibit diurnal patterns at the daily level. These diurnalities can be important features to account for when modeling stock price movements.

Figure 1 displays the intraday pattern of the average synchronous correlation between returns of all stock pairs in the top 100 most capitalized stocks traded on the NYSE. Prices are sampled at a time resolution of $\Delta t = 15$ minutes. We include results for data from the time period 2001-03, as well

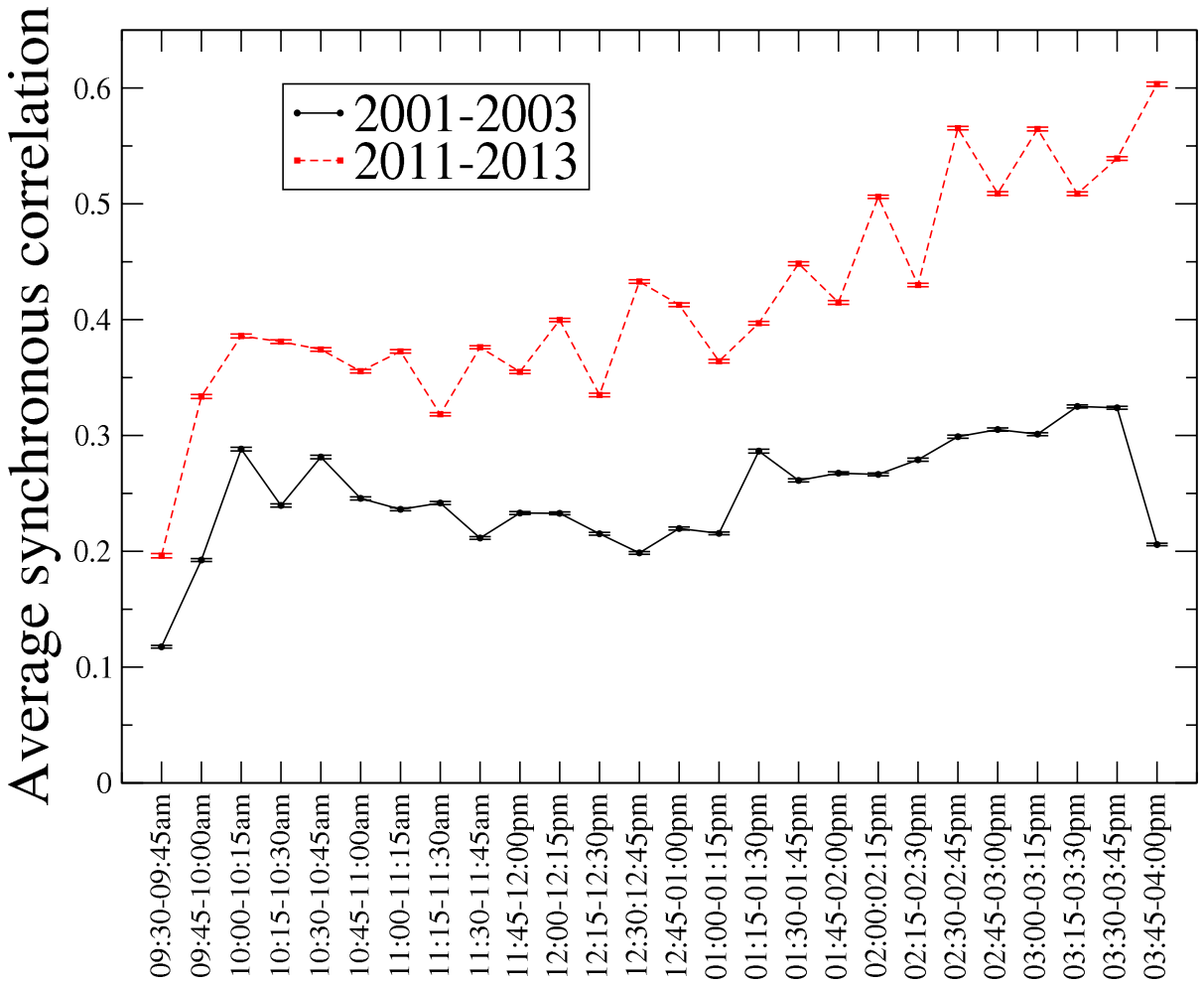

Figure 1: intraday pattern of the average synchronous correlation between fifteen minute stock returns of the 100 most capitalized stocks traded at NYSE in the period 2001-03 (black continuous line) and 2011-13 (red dashed line).

as 2011-13, where we observe striking changes over the past decade in the magnitude of the measured correlations. Both periods exhibit a similar profile in the intraday pattern of synchronous correlations, with an explosive growth in the first hour of the trading day that levels in the late morning, followed by a steady increase in the afternoon. A similar profile has been observed in other studies (Allez and Bouchaud, 2011).

We use the statistical methodology introduced above to construct an analogous profile for lagged correlations. In Fig. 2 we plot the average lagged correlation between the same stock pairs from Fig. 1. Prices are again sampled at a time resolution of $\Delta t = 15$ min., with correlations evaluated at one sampling time horizon. We find that, although the distributions of lagged correlation coefficients are on average quite small, there exist pairs of stocks in the tails of these distributions that represent a statistically-significant lagged correlation, in the sense of the methodology described above. These stock pairs form the links in a series of statistically-validated networks. We plot the intraday pattern of lagged correlations for the stock pairs belonging to the Bonferroni network in red, and the FDR network in blue. In both the data from 2001-03 and 2011-13 we find that the bulk of the lagged correlations tends to shift to the positive regime during the final minutes of the trading day.

The positive shift in the bulk of the lagged correlation coefficients manifests as an increase in network connectivity. In Fig. 4 we display visualizations of the Bonferroni networks for both the beginning,

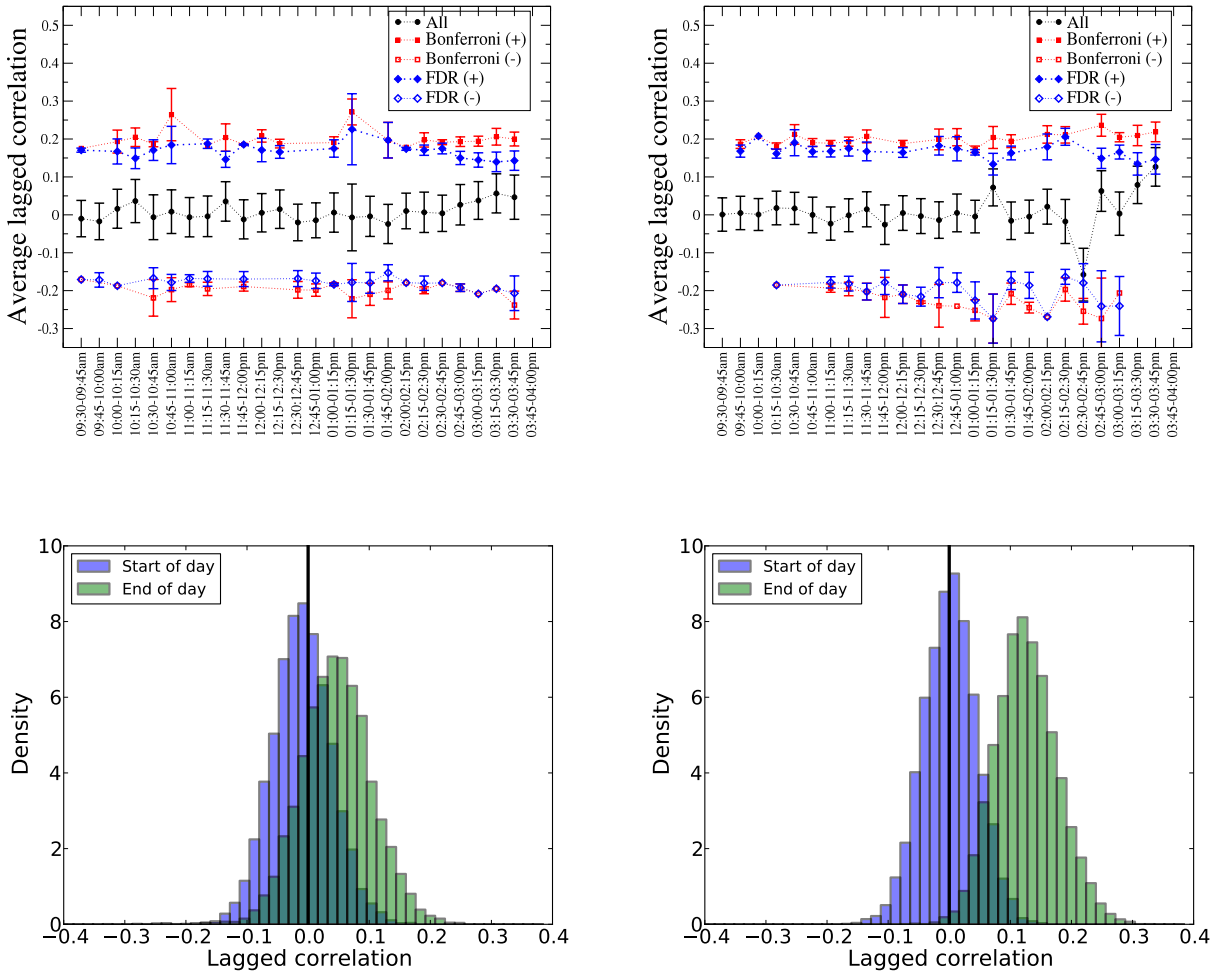

Figure 2: intraday pattern of the average lagged correlation, evaluated at one lag, between fifteen minute stock returns of the 100 most capitalized stocks traded at NYSE in the period 2001-03 (top left panel) and 2011-13 (top right panel). In each panel, we also report the pattern of lagged correlation with average taken over all the links that belong to the Bonferroni network (red squares) and the FDR network (blue diamonds), by distinguishing between positive (+) and negative (-) statistically validated correlations. We also provide the probability density function of all $N^2 = 10,000$ lagged correlation coefficients for two intraday periods in 2001-03 (bottom left panel) and 2011-13 (bottom right panel). The blue shaded histogram corresponds to correlations between returns in the first 15 minutes of the trading day (9:30 a.m. to 9:45 a.m.) and those in the second 15 minutes (9:45 a.m. to 10:00 a.m.). The green shaded histogram corresponds to correlations between returns in the second-to-last 15 minutes of the trading day (3:30 p.m. to 3:45 p.m.) and those in the last 15 minutes (3:45 p.m. to 4:00 p.m.). We observe a characteristic positive shift in the lagged correlations in the final minutes of the trading day.

middle and end of the trading day for the period 2001-03, and the corresponding visualizations for the 2011-13 data in Fig. **??**. In both periods we observe a decrease in connectivity during the middle of the trading day, followed by an explosive growth in the significance of positive lagged correlations during the final minutes of the trading day, reminiscent of the well-known U-shaped pattern in intraday transaction volume and volatility (Admati and Pfleiderer, 1988; Andersen and Bollerslev, 1997). Our analysis underscores dramatic intraday periodicities in the co-movements of asset prices. Despite these effects, we find that the validated links are largely persistent throughout the trading day, as detailed in the Supplementary Information.

4. Reconstructing correlations

These periodic effects are crucial to take into account when modeling collective stock dynamics. Here we investigate the impact of high-frequency lagged cross correlations and autocorrelations of returns on synchronous correlations between stock returns evaluated at a larger time horizon. In particular, we retain information on the intraday period when measuring how these lead-lag relationships at short timescales may influence synchronous co-movements among equities at longer timescales. In the Supplementary Information we derive an equation, obtained by taking an approach similar to the one presented in ref. (Toth and Kertesz, 2009), in which we show how the synchronous correlation between two stock returns time series, as evaluated at a certain intraday time window, e.g., the first 130 minutes of the trading day, can be decomposed in order to make apparent the individual contribution of autocorrelations and lagged cross-correlations evaluated at smaller time windows, such as $\Delta t = 5$ minutes. The only assumption we make to obtain that equation is that the intraday volatility pattern $\sigma_i^2(q, \Delta t)$ of a stock i, where q indicates the intraday-time and Δt the time horizon, can be written as an idiosyncratic constant c_i, associated with each stock, times a function $f_q(\Delta t)$ that describes the intraday variations of volatility, and which is common to all the stocks: $\sigma_i^2(q, \Delta t) = c_i \cdot f_q(\Delta t)$. Here we show the equation in a simple case, in order to make apparent the contributions of each term. Consider the first 30 minutes of the trading day, and suppose we are interested in the synchronous correlation coefficient $\rho_{x,y}$ between the time series x and y, such that $\{x\} = \{x(1), x(2), ..., x(T)\}$ and $\{y\} = \{y(1), y(2), ..., y_(T)\}$, where T is the number of trading days in the dataset, and $x(i)$ and $y(i)$ represent the return of stock i and stock j, respectively, in the first 30 minutes of day i. Each one of these time series of log-returns can be decomposed in the sum of $p = 2$ time series of log-returns, specifically, the time series of returns in the first $p = 2$ intraday time intervals of $\Delta t = 15$ minutes:

$$\begin{aligned}\{x\} =& \{x_1(1) + x_2(1), x_1(2) + x_2(2), ..., x_1(T) + x_2(T)\};\\ \{y\} =& \{y_1(1) + y_2(1), y_1(2) + y_2(2), ..., y_1(T) + y_2(T)\};\end{aligned}$$

where $x_1(i)$ and $y_1(i)$ ($x_2(i)$ and $y_2(i)$) are the returns of the two stocks observed in the first (second) 15 minutes of day i. In this way we obtain that:

$$\rho_{x,y} = \frac{f_1^2\,\rho_{x_1,y_1} + f_2^2\,\rho_{x_2,y_2} + f_1\,f_2\,(\rho_{x_1,y_2} + \rho_{x_2,y_1})}{\sqrt{[f_1^2 + f_2^2 + 2 f_1\,f_2\,\rho_{x_1,x_2}]\,[f_1^2 + f_2^2 + 2 f_1\,f_2\,\rho_{y_1,y_2}]}}. \tag{3}$$

This equation clearly shows how the interplay between short-term lagged cross-correlations and autocorrelations contributes to the value of the longer-term synchronous correlation $\rho_{x,y}$. For instance, the equation above shows how negative values of autocorrelations, ρ_{x_1,x_2} and ρ_{y_1,y_2}, and/or positive values of lagged cross correlations, ρ_{x_1,y_2} and ρ_{x_2,y_1} may be responsible for the well known Epps effect (Epps, 1979): $\rho_{x,y} > \max(\rho_{x_1,y_1}, \rho_{x_2,y_2})$. It is also worthwhile to point out that the correlation coefficient $\rho_{x,y}$ does not depend on quantities related to other stocks in the system. Therefore, structural properties of the correlation matrix, such as the fact that it should be positive semi-definite, are not forced by our reconstruction equation. In Fig. 5, we show some results of the reconstruction analysis of the 100

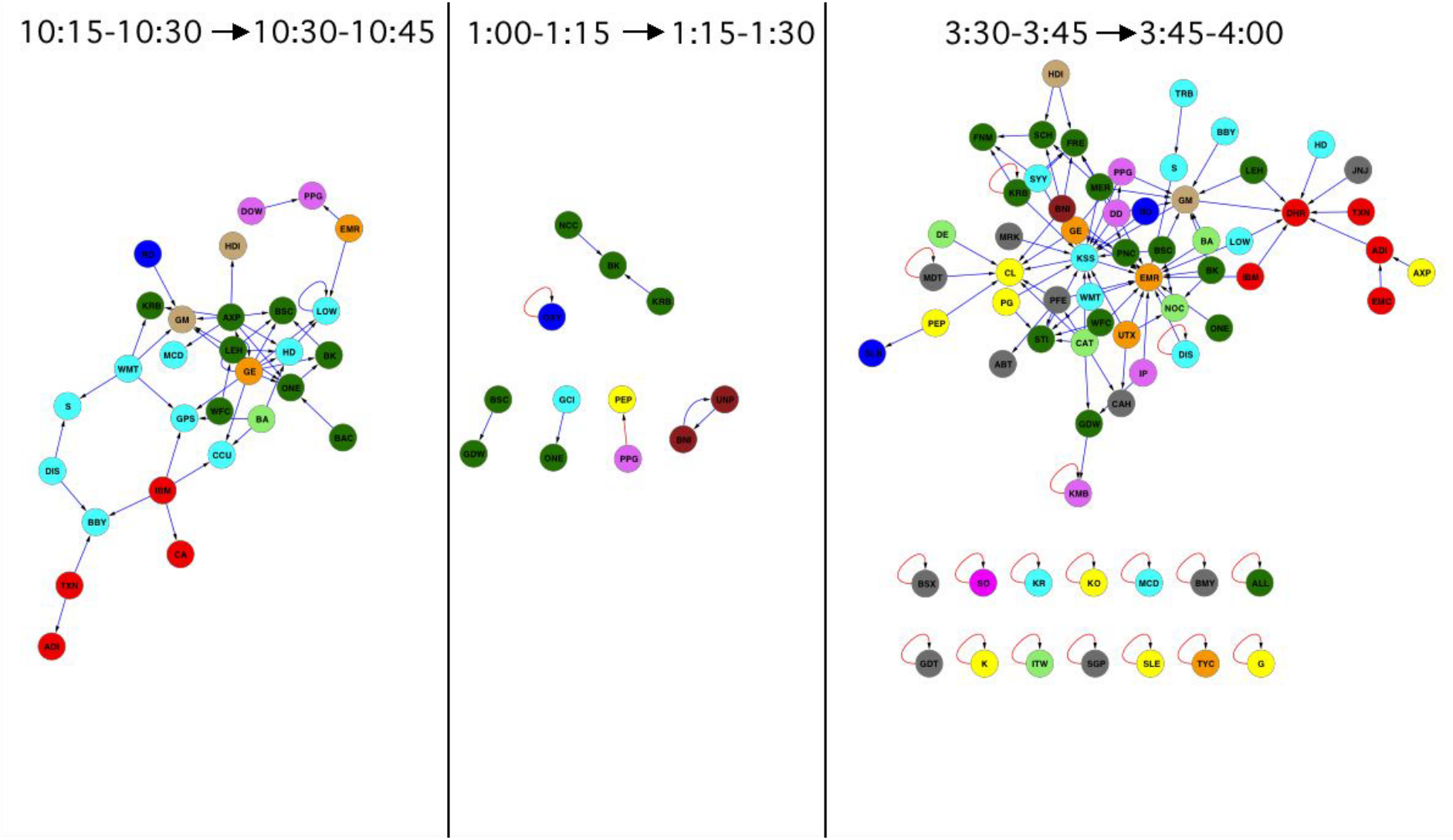

Figure 3: Visualization of the Bonferroni networks from periods in the beginning, middle, and end of the trading day in the period 2001-03. Stocks are colored by their economic sector. Links of positive correlation are colored blue, while links of negative correlation are colored red.

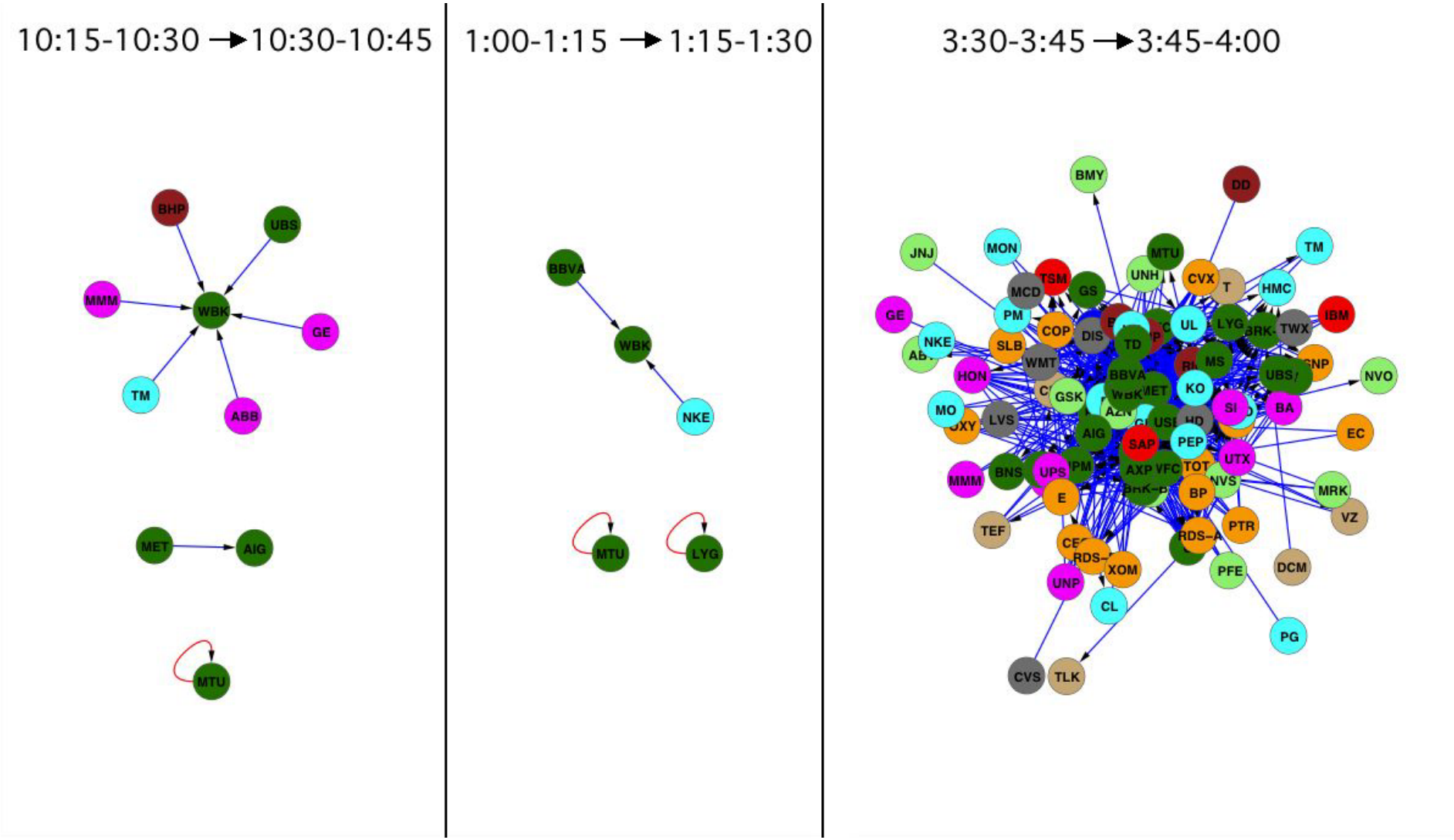

Figure 4: Visualization of the Bonferroni networks from periods in the beginning, middle, and end of the trading day in the period 2011-13. Stocks are colored by their economic sector. Links of positive correlation are colored blue, while links of negative correlation are colored red.

stock correlation matrix for the two time periods under investigation, 2001-03 (left panel) and 2011-13 (right panel). We have divided the trading day in three time windows of 130 minutes each, from 9:30 a.m. to 11:40 a.m. (top panels), from 11:40 a.m. to 1:50 p.m. (mid panels), and from 1:50 p.m. to 4:00 p.m. (bottom panels), and reconstructed synchronous correlations in each time window by considering a subdivision of it in 26 time windows of 5 minutes. In each panel we show three curves, one obtained by considering the contribution of both auto-correlations and lagged cross-correlations up to a given lag, as reported on the x-axis, one obtained by only retaining the contribution of lagged cross-correlations, and one obtained by only considering the contribution of autocorrelations. The first point from the left on the x-axis, labeled NP-0, corresponds to the case in which, besides neglecting all the auto-correlations and lagged cross-correlation in the reconstruction formula, we also neglect the intraday volatility pattern. The curves are obtained by comparing the reconstructed correlation matrix C_{rec} and the original correlation matrix C_{or} through the standard Frobenius norm:

$$F(C_{or}, C_{rec}) = \sqrt{\mathrm{tr}\left[(C_{or} - C_{rec})(C_{or} - C_{rec})^T\right]}, \tag{4}$$

where tr$[\cdot]$ is the trace operator, and apex T indicates the transpose operator. We normalize each distance by the Frobenius distance between C_{or} and the identity matrix, representing the distance that would be obtained under maximal ignorance of the system's correlations. The results obtained for the 2001-03 time period (left panels) indicate that lagged cross-correlations contribute more to synchronous correlations than autocorrelations in all the three time windows, although such a contribution tends to decrease during the day. On the other hand, in the 2011-13 time period, the relative impact of lagged cross-correlations decreases, and the interplay between auto-correlations and lagged cross-correlations becomes stronger. This evidence is also confirmed by an analysis of the spectrum of correlation matrices: indeed, all the correlation matrices reconstructed in the period 2001-03 turn out to be positive definite, regardless of the number of lags considered in the reconstruction, or if we ignore autocorrelations or lagged cross-correlations. In the 2011-13 time period the situation is different. If one uses both autocorrelations and lagged cross-correlations to reconstruct the correlation matrix, then all the reconstructed matrices are positive definite for any lags considered in the reconstruction. However, if we constrain ourselves to use either autocorrelations or cross-correlations in the reconstruction equation, then most of the reconstructed matrices display some negative eigenvalues. We may interpret this result as an increased fragility of the structural properties of the 2011-13 correlation matrices in the presence of noise, and explore this interpretation in the Supplementary Information.

The presented analysis shows that, in the period 2001-03 1) the effect of lagged cross correlations on determining synchronous correlations at larger time horizons is stronger than the effect of autocorrelations and 2) the interplay between these two effects is moderate. At the contrary, in the period 2011-13, we observe that 1) the effect of lagged cross correlations on determining synchronous correlations at larger time horizons is comparable with the effect of autocorrelations and 2) the interplay between these two effects is much stronger in this period. We find that the magnitudes of the lagged cross-correlation, autocorrelation, and volatility terms vary throughout the trading day. Thus, the roles of the factors contributing to the Epps effect are dynamic, both during a single trading day and over the span of years.

5. Regression model

The intraday signals we uncover are of potential use as a feature-selection stage in modeling stock price dynamics. If one aims to model the returns of a given asset using only previous returns of other assets as inputs, the careful selection of these inputs is of critical importance to prevent overfitting and to aid in a model's interpretation.

We show that, at each intraday period, the relevant inputs to a model of the returns of stock i can be reliably taken as the set of direct predecessors $\{\nu_j\}$ of the corresponding node in the validated network

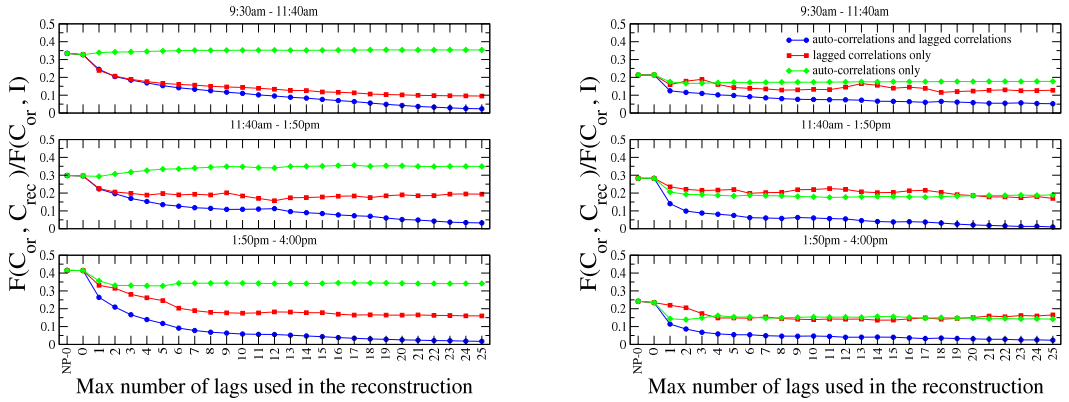

Figure 5: Normalized Frobenius distance between the 130 minute return correlation matrix of the 100 most capitalized stocks traded at NYSE, C_{or}, and the corresponding correlation matrix, C_{rec}, reconstructed according to the method described in the text in the time period 2001-03 (left panels) and 2011-13 (right panels), in the three 130 minute segments of the trading day: from 9:30 a.m. to 11:40 a.m. (top panels), from 11:40 a.m. to 1:50 p.m. (middle panels), and from 1:50 p.m. to 4:00 p.m. (bottom panels). Distances are normalized by the Frobenius distance between C_{or} and the identity matrix. Each value reported in the horizontal axis indicates the number of lags used to reconstruct 130 minute return correlations from from 5 minute return (lagged and synchronous) correlations. The first point from the left in each panel, labeled "NP-0", is obtained by disregarding the intraday pattern of volatility, which is considered in all the other reconstructed matrices. Three curves are shown in each panel: the green (red) curve describes the results obtained by only including autocorrelation (lagged cross-correlation) terms in the equation used to reconstruct synchronous correlations, while the blue curve shows results in the case in which both autocorrelation and lagged cross-correlation terms are included in the reconstruction equation.

for that period. That is, we need only consider a node j as an input to the model if there is a link from j to i. To demonstrate this, for each intraday period we attempt to model the returns of stocks with an in-degree of at least one with a simple linear model. If we represent column i of matrices A or B from the methodology section with A_i or B_i, then we fit

$$B_i = \beta_0 + \beta_1 A_{\nu_1} + \beta_2 A_{\nu_2} + \cdots + \beta_{k_i} A_{\nu_{k_i}} + \epsilon \tag{5}$$

where k_i is the in-degree of node i and there is a directed edge to i from each node $j \in \{\nu_j\}$.

For each model we compute the Bayesian information criterion, or BIC, where for each node i

$$\mathrm{BIC}_i = (k_i + 1)\ln(T) - 2\ln(L_i) \tag{6}$$

where T is the number of rows in A and B, equal to the number of days in the analysis, and L_i is the maximized likelihood for the model in equation (5). The BIC is a criterion for model selection, and can be interpreted as an anticipation of a model's out-of-sample performance using only in-sample training data.

We compare the measured BICs to a randomised model, in which for each node i we randomly select k_i of the $N = 100$ available nodes as regressors in equation (5). This procedure is repeated 100 times for each model. In Fig. 5, for both the 2001-03 and 2011-13 datasets, we plot the mean difference in BICs for all models. With the exception of one period in the 2011-13 dataset, the specification of model inputs using the Bonferroni network always outperforms the randomised specification. The specification using the FDR network fares similarly, although it fails to outperform the randomised specification in one period in the 2001-03 dataset and five periods in the 2011-13 dataset. These periods fall at the end of the trading day, when, due to the large numbers of validated links, the relative advantage of the validated networks in feature selection diminishes against a random selection of inputs.

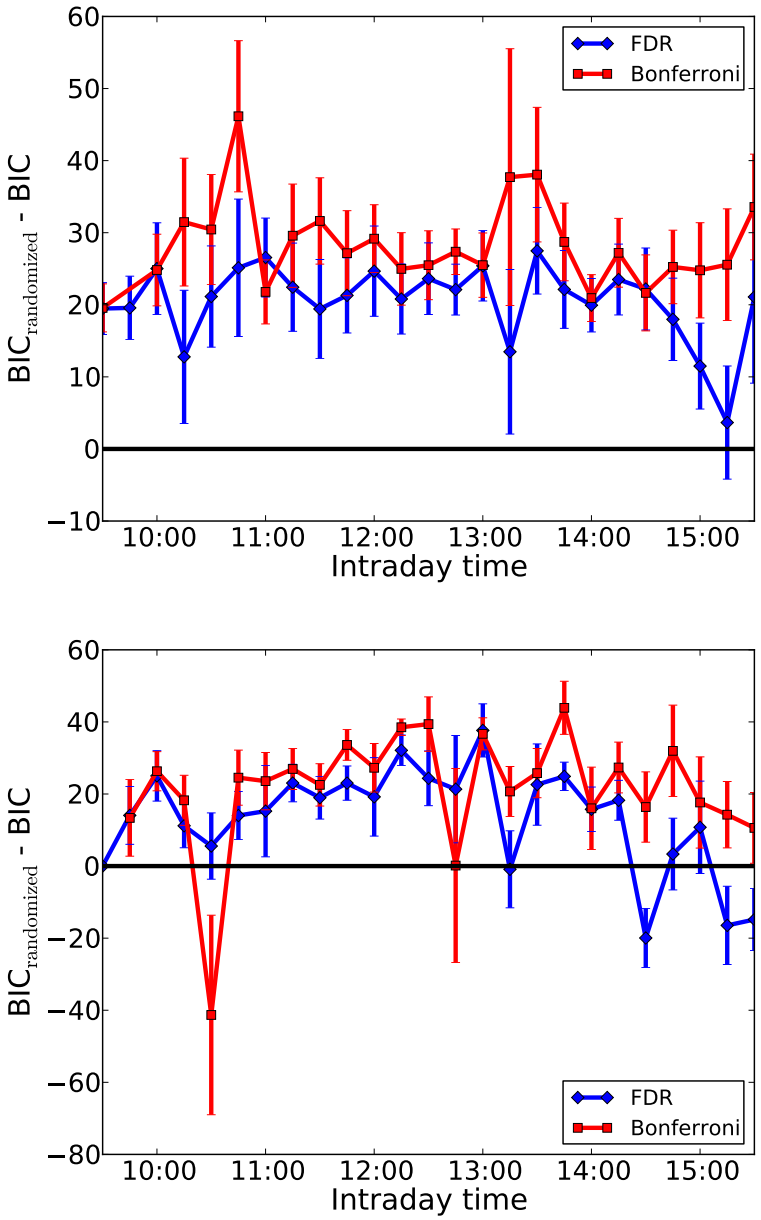

Figure 6: Difference in BICs between the models in equation (5) and the randomized models described in the text, for both the periods 2001-03 (top panel) and 2011-13 (bottom panel). We generate 100 realisations of the random model for each stock. Points show the mean BIC deviation of all stocks from the mean Bayesian Information Criterion (BIC) of the corresponding randomised models. Error bars show the uncertainties in this deviation for all models, added in quadrature.

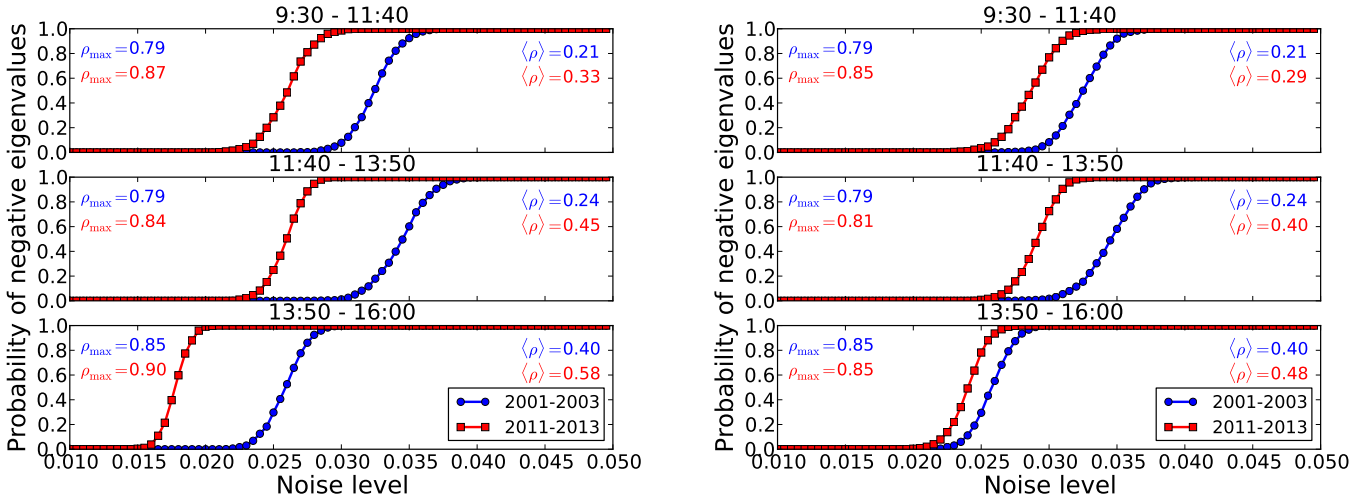

Figure 7: Probability of observing at least one negative eigenvalue in each 130 min. correlation matrix after perturbing correlation matrices with a given level of noise. In the right panel we exclude the months of August, September and October 2011 from the analysis. For a noise level x, each symmetric pair of off-diagonal elements (i, j) and (j, i) are perturbed by a number from a uniform distribution on the interval $[-x, x]$. Data from 2001-03 are shown in blue, while data from 2011-13 are shown in red. We also show the mean and maximum off-diagonal correlation values from each matrix. We observe that the 2011-13 data exhibits negative eigenvalues at a consistently lower noise level than the 2001-03 data. Each probability is evaluated through 1000 independent perturbations of the matrix. In addition, the 2011-13 data has four pairs of stocks that represent the same firm: BRK-A and BRK-B, RDS-A and RDS-B, BHP and BBL, UN and UL. These stocks have very high synchronous correlations, so we exclude BRK-B, RDS-B, BBL and UL from the analysis. Including them does not qualitatively change the results, but exaggerates the observed pattern.

6. Stability of reconstructed correlation matrices to noise

Here we provide a brief explanation of the structural problems uncovered in the reconstructed correlation matrices in 2011-13. If we constrain ourselves to use only autocorrelations or lagged cross-correlations in the reconstruction analysis, then most matrices in this period are not positive definite as they have some number of negative eigenvalues. On the other hand, all reconstructed correlation matrices in the period 2001-03 have positive eigenvalues.

We illustrate this increased "fragility" of the 2011-13 correlation matrices in Figure 7.

In this analysis we perturb the 130 minute correlation matrices from each portion of the trading day with a given level of noise. For a noise level x, each symmetric pair of off-diagonal elements (i, j) and (j, i) are perturbed by a number from a uniform distribution on the interval $[-x, x]$. We then measure the probability of observing at least one negative eigenvalue in each matrix through 1000 independent perturbations. In Figure 7 we compare results from 2001-03 with those from 2011-13, and also show the contribution of the months of August, September and October 2011 by removing it from the analysis (right panel).

We find that the structural properties of the correlation matrices obtained in the period 2001-03 are significantly more robust than those obtained in 2011-13. This analysis complements the observation presented in the main text, that the 130 minute correlation matrices reconstructed without contributions from 5 minutes lagged cross-correlations or autocorrelations are not always positive definite. Owing in part to an increased level of synchronous correlation, there are tighter bounds constraining each element of the 2011-13 correlation matrices. Given a noise level, these bounds are more easily violated than in the 2001-03 data.

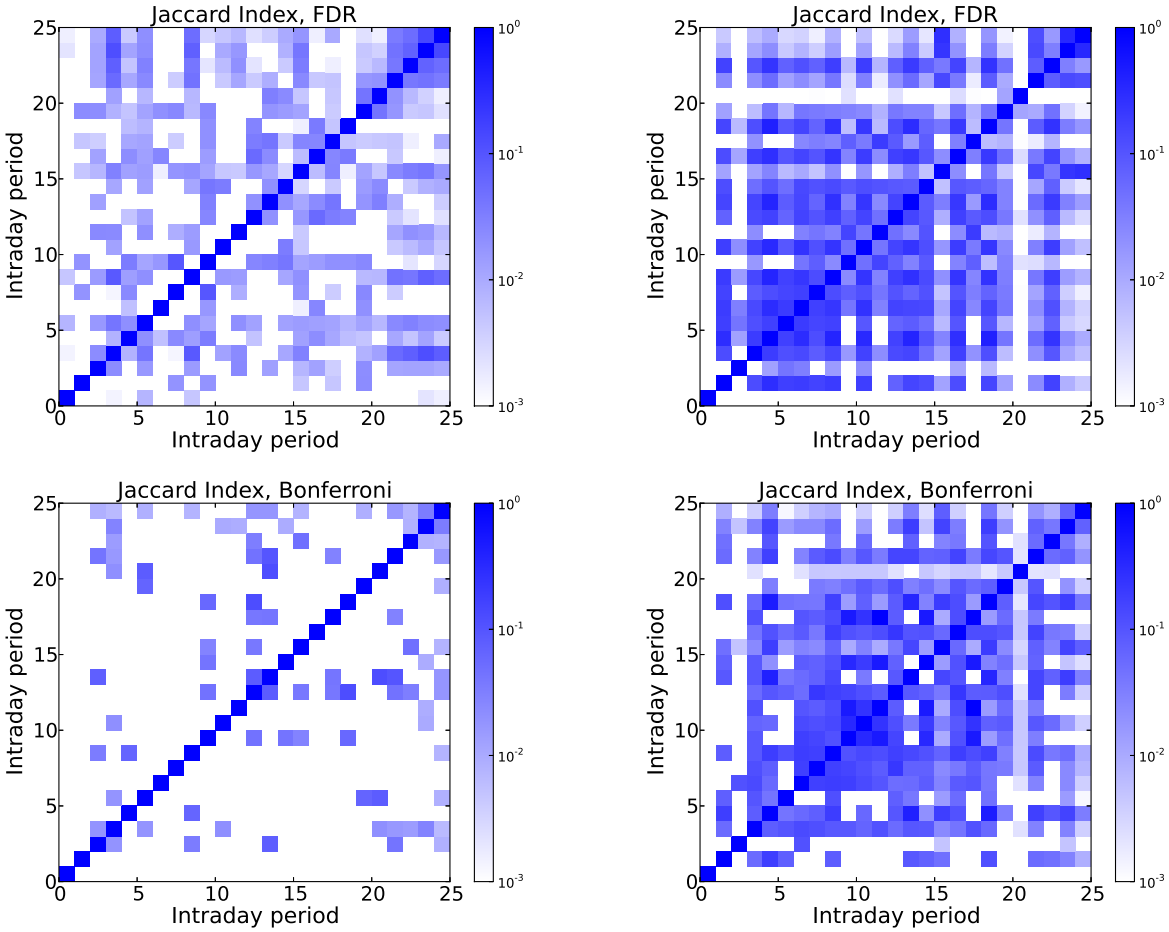

Figure 8: Matrices of Jaccard Indices between sets of links corresponding to networks for all intraday periods at a time horizon $\Delta t = 15$ min. Left column shows results using data from 2001-03 (FDR and Bonferroni networks); right column shows results using data from 2011-13 (FDR and Bonferroni networks). We find that the validated links are generally more persistent in the 2011-13 data throughout the trading day.

7. Persistence of links

To what extent do the lead-lag relationships that we uncover persist during the trading day? Although we find intraday effects that influence the number and strength of the validated lagged correlations, it is a separate question to consider whether a link that is validated in one intraday period will be validated in another.

We find that the validated links are indeed largely persistent throughout the trading day, although they are more strongly dependent on the particular intraday period in the 2001-03 data. We support this finding with two analyses. First, we may quantify the extent to which two networks share links using the Jaccard Index:

$$J(i,j) = \frac{|L_i \cap L_j|}{|L_i \cup L_j|},$$

where L_i is the set of links in network i. We distinguish edges by both direction and sign when constructing these sets. A high value of the Jaccard Index, in this context, indicates that two networks share a large proportion of their total links. In Figure 8 we display matrices of Jaccard Indices $J(i,j)$ between sets of links corresponding to networks for all intraday periods at a time horizon $\Delta t = 15$ min.

We find that the Jaccard Indices are generally high, suggesting that the links we validate are indeed persistent across many time periods, although this effect is weaker in the 2001-03 data. Moreover, we find that the Jaccard Indices are largely homogeneous throughout the trading day; i.e., it does not seem

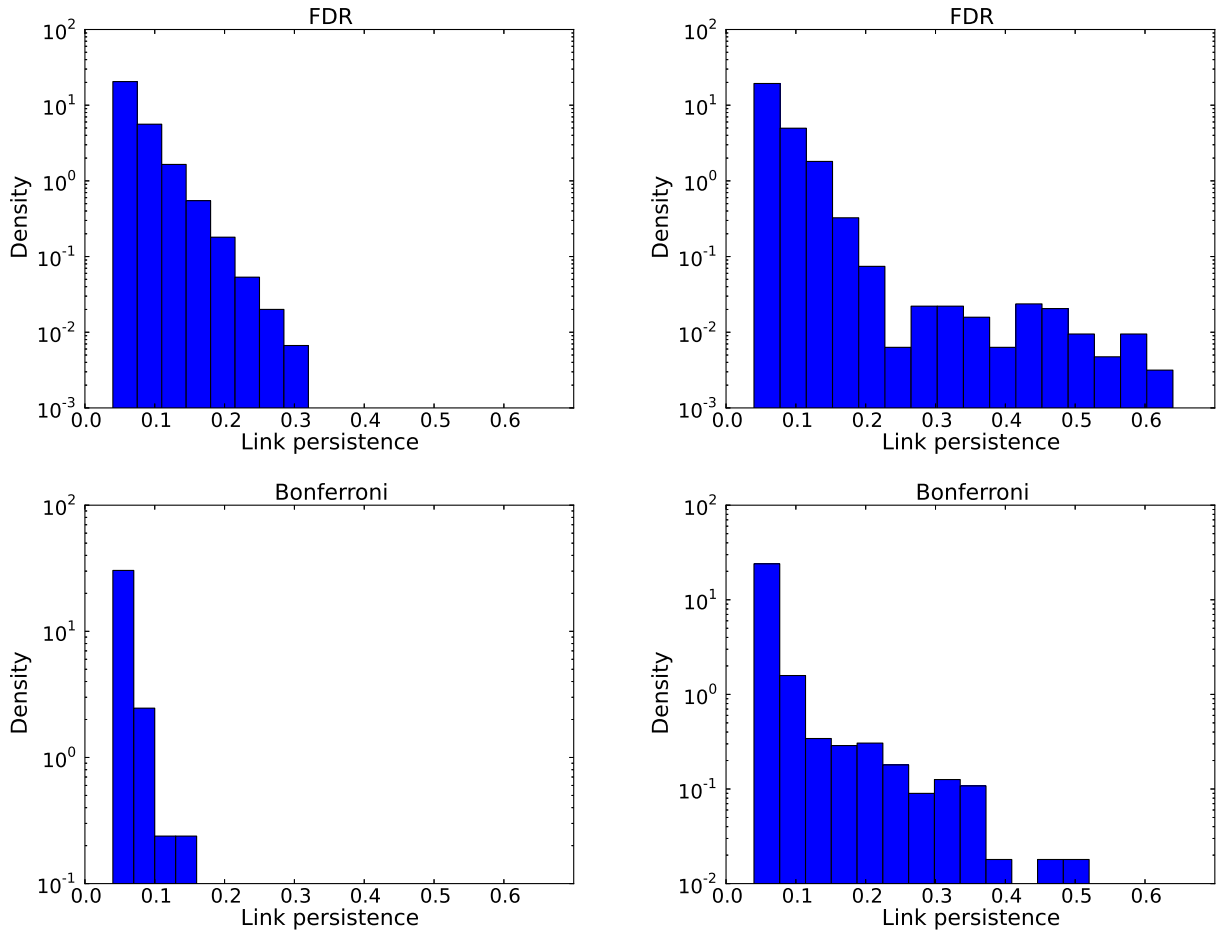

Figure 9: Distributions of link persistence for all links in networks at a time horizon $\Delta t = 15$ min. Left column shows results using data from 2001-03 (FDR and Bonferroni networks); right column shows results using data from 2011-13 (FDR and Bonferroni networks). We find that the validated links are generally more persistent in the 2011-13 data throughout the trading day.

to be the case that links are shared preferentially in neighboring time periods. We find that this effect is stronger in the 2011-13 data. Finally, we have verified that these plots are only very weakly affected by the turmoil of August - October 2011, as the corresponding diagrams for the networks that were constructed with this period removed are similar.

The analysis in Figure 8 quantifies a degree of similarity among intraday periods. We can also examine this similarity at the level of individual links, by quantifying the persistence of links. This persistence is defined as the fraction of intraday networks (of which there are 25 for $\Delta t = 15$ min.) in which a given link appears. We plot the distributions of link persistence for all networks in Figure 9, where we observe again from this perspective that individual links seem to be more persistent in the 2011-13 data (although, again, this analysis does not convey information regarding the number or strength of the validated links).

8. Influence of autocorrelations

8.1. Effect of autocorrelations on linear models

To examine the influence of autocorrelations on the performance of the linear models described in the text, we repeat the analysis with validated autocorrelation links removed. That is, the model for each node i has k_i inputs, with k_i the in-degree of node i, disregarding autocorrelation links. As in the text, we compare the BICs of these models with those obtained from randomly selecting k_i of the

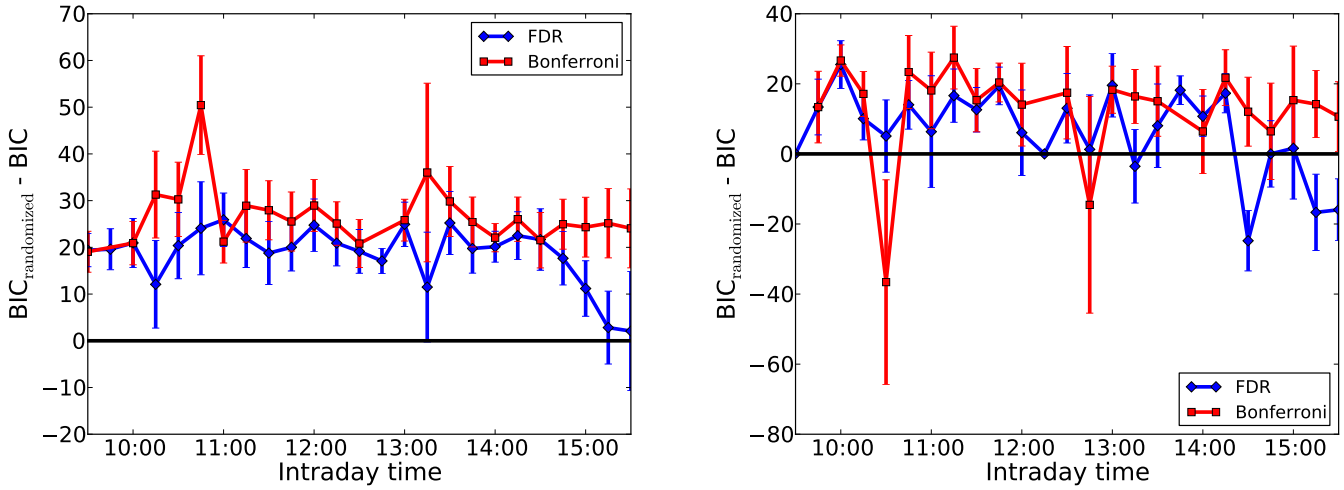

Figure 10: Difference in BICs between the linear models with inputs prescribed by the validated network and the randomized models described in the text, for both the periods 2001-03 (left panel) and 2011-13 (right panel), upon removal of autocorrelation links. We generate 100 realizations of the random model for each stock. Points show the mean BIC deviation of all stocks from the mean BIC of the corresponding randomized models. Error bars show the uncertainties in this deviation for all models, added in quadrature.

$N = 100$ possible input nodes as regressors in the model. In Figure 10, for both the 2001-03 and 2011-13 datasets, we plot the mean difference in BICs for all models. The results highlight the elevated influence of autocorrelations in the recent data: whereas the models in 2001-03 continue to outperform the randomized models, in 2011-13 the model performance is markedly worse if autocorrelations are ignored.

8.2. Partial lagged correlation networks

The reconstruction analysis presented in the text reveals how both autocorrelations and lagged correlations at a given time horizon compete to form synchronous correlations among stock returns evaluated at a larger time horizon. In the 2011-13 dataset, we find that the two contributions are tangled, and when one attempts to uncouple them the result is a reconstructed correlation matrix that exhibits severe structural problems, such as negative eigenvalues. This result might be due to the fact that (i) the average synchronous correlation among stock returns is quite large in this period– significantly larger than in the 2001-03 data, and (ii) many statistically significant autocorrelations are observed in the 2011-13 data, while fewer are observed in the 2001-03 data. These two observations have the potential to explain the presence of a large number of statistically validated lagged correlations in the 2011-13 dataset, and could also explain the tight connection between autocorrelations and lagged cross-correlations mentioned above. That is, a lagged cross-correlation between two stock returns $\rho(x(t), y(t+\tau))$ may just reflect the presence of autocorrelation of stock return x, $\rho(x(t), x(t+\tau))$ and the synchronous correlation between stock returns x and y, $\rho(x(t+\tau), y(t+\tau))$. Similarly, we could consider the autocorrelation of returns in stock y, $\rho(y(t), y(t+\tau))$ and the synchronous correlation $\rho(x(t), y(t))$.

To check that the lagged cross-correlations we validate are not spuriously the result of autocorrelations, we construct networks derived from partial lagged correlations

$$\rho(x(t), y(t+\tau)|y(t)) = \frac{\rho(x(t), y(t+\tau)) - \rho(y(t), y(t+\tau))\rho(x(t), y(t))}{\sqrt{[1-\rho(y(t), y(t+\tau))^2][1-\rho(x(t), y(t))^2]}}, \text{ and} \quad (7)$$

$$\rho(x(t), y(t+\tau)|x(t+\tau)) = \frac{\rho(x(t), y(t+\tau)) - \rho(x(t+\tau), y(t+\tau))\rho(x(t), x(t+\tau))}{\sqrt{[1-\rho(x(t+\tau), y(t+\tau))^2][1-\rho(x(t), x(t+\tau))^2]}}, \quad (8)$$

subtracting off the influence of autocorrelations.

We thus repeat the statistical validation procedure, using the same shuffling procedure described in the text, with a matrix of lagged partial correlations in place of the lagged correlation matrix (considering only the off-diagonal elements, as the diagonal elements of this partial correlation matrix are undefined). We build separate networks for partial correlations given by (7) and (8), again choosing our statistical threshold to be $p = 0.01$.

We report results for $\Delta t = 15$ min. in the last time horizon of the trading day, when we find the strongest autocorrelations. Using the Bonferroni correction for multiple comparisons, we validate 448 positive links using the partial correlation matrix (7), and 313 positive links using the matrix (8). We validate no links of negative correlation. Using the original lagged correlation matrix, we validate 91 positive links and 18 negative links. Because the autocorrelations are negative, we validate many more links in the partial lagged correlation networks; that is, the original lagged correlation networks contain many positive links in spite of the negative correlations, and not because of them. We note that the partial lagged correlation networks using the matrices (7) and (8) share an intersection of 77 and 83 links, respectively, with the original network. The probability of randomly sampling these intersections x from the $L = 100 \times 99 = 9900$ total possible lagged cross-correlation links in $n = 91$ "draws" (links in the original network) is given by the hypergeometric distribution:

$$P(x|n,k,L) = \frac{\binom{k}{x}\binom{L-k}{n-x}}{\binom{L}{n}},$$

where k is the number of validated links in the partial correlation network. We can thus associate a p-value to these intersections as the probability of validating at least x links common to both the original and partial lagged correlation networks under the null hypothesis of random sampling:

$$p = P(j > x|n,k,L) = 1 - P(j < x|n,k,L) = 1 - \sum_{j=0}^{x} P(j|n,k,L).$$

This number is vanishingly small for the numbers of links k validated in each partial correlation network, and the intersections x between the directed links in this network and the directed links validated in the original lagged correlation network. So we may safely conclude that the lagged cross-correlations we validate in the data are not artifacts of autocorrelation effects in the time series.

We repeat the same procedure on the 2011-13 data, validating 629 positive links using the partial correlation matrix (7), and 831 positive links using the matrix (8). We validate no links of negative correlation. Using the original lagged correlation matrix, we validate 801 positive links and no negative links. We note that the partial lagged correlation networks using the matrices (7) and (8) share an intersection of 295 and 374 links, respectively, with the original network. Again, we may associate a p-value to these intersections using the hypergeometric distribution, which is vanishingly small both networks.

9. Discussion

The methodological framework presented here provides a validation of lead-lag relationships in financial markets, and quantifies the impact of underlying networks of short term lead-lag relationships on longer term synchronous correlations among equities throughout different parts of a trading day. First, we validate the existence of such relationships using empirical data from two different periods. The validated lead-lag relationships provide new insights into the dynamics of financial markets, and provide new understandings into such phenomena as the Epps effect. Finally, we present an example of the use of such new information on market dynamics, by performing a regression model which incorporates the information on the validated lead-lag relationships.

Comparing the time periods 2001–2003 and 2011–2013, the synchronous correlations among these high market capitalization stocks have grown considerably, whereas the number of validated lagged-correlation relationships have decreased. We relate these two behaviors to an increase in the risks of financial spillover and an increase in the informational efficiency of the market, respectively. Furthermore, our different analyses all show a change in the role of auto-correlation in market dynamics, which is increasing. This is possibly related to the growing use of automated and high frequency trading, in the U.S. market and elsewhere.

In summary, we introduce the statistically validated network framework for validating lead-lag relationships in the U.S. market, and are able to empirically identify and validate such relationships. This sheds important new light into the underlying dynamics of the U.S. financial market, and provides critical information into future risk management strategies. Furthermore, it provides policy and decision makers new information on the structure and stability of the market, and lays the ground for new models and theories for asset management, risk management, and financial spillover.

Acknowledgments

The authors would like to thank Shlomo Havlin for his comments and suggestions on this work. CC, HES, and DYK are thankful to the Office of Naval Research (ONR Grant N000141410738) for financial support. The views and opinions expressed are those of the individual authors and do not necessarily represent official positions or policy of the Office of Financial Research or the U.S. Treasury.

Appendix A. Reconstruction of synchronous correlations using autocorrelations and lagged cross-correlations

Consider two time series of log-returns, $\{x\}$ and $\{y\}$, associated with a certain intraday window $p\Delta t$, with integer $p > 2$, e.g. the first $p\Delta t = 195min$ of a trading day. We are interested in the correlation coefficient between the time series

$$\begin{aligned}\{x\} =& \{x_1, x_2, ..., x_T\} \text{ and} \\ \{y\} =& \{y_1, y_2, ..., y_T\},\end{aligned}$$

where T is the number of trading days in the dataset. Each one of these time series of log-returns can be decomposed as the sum of p time series of log-returns— specifically, the time series of returns in the first p intraday time intervals of $\Delta tmin$, e.g., if $p\Delta t = 195min$ one can set $p = 13$ and $\Delta t = 15min$:

$$\begin{aligned}\{x\} &= \left\{ \sum_{j=1}^{p} x_1(j), \sum_{j=1}^{p} x_2(j), ..., \sum_{j=1}^{p} x_T(j) \right\}; \\ \{y\} &= \left\{ \sum_{j=1}^{p} y_1(j), \sum_{j=1}^{p} y_2(j), ..., \sum_{j=1}^{p} y_T(j) \right\};\end{aligned}$$

where $x_i(j)$ and $y_i(j)$ are the returns of the two stocks observed in jth 15 minute time window of day i, $j = 1, ..., p$. We further assume that

$$< x(j) >= \frac{1}{T} \sum_{i=1}^{T} x_i(j) =< y(j) >= \frac{1}{T} \sum_{i=1}^{T} y_i(j) = 0, \quad \forall j = 1, ..., p.$$

This is not a very restrictive hypothesis because it's (usually) appropriate to assume that the expected return is 0. Therefore, we obtain that:

$$< x >= 0 \text{ and } \quad < y >= 0$$

as a consequence of the additivity of log-returns and the linearity of the average. Let's now consider the (maximum likelihood estimate of the) the variance of the variable x:

$$\begin{aligned}\sigma_x^2 =& < x^2 >= \frac{1}{T} \sum_{i=1}^{T} \left[\sum_{j=1}^{p} x_i(j) \right]^2 = \\ &= \frac{1}{T} \sum_{i=1}^{T} \left[\sum_{j=1}^{p} x_i(j)^2 + 2 \sum_{j=1}^{p-1} x_i(j)\, x_i(j+1) + 2 \sum_{j=1}^{p-2} x_i(j)\, x_i(j+2) + ... + 2\, x_i(1)\, x_i(p) \right] = \\ &= \sum_{j=1}^{p} \sigma_x(j)^2 + 2 \sum_{j=1}^{p-1} \sigma_x(j)\, \sigma_x(j+1) \rho_{x_j, x_{j+1}} + 2 \sum_{j=1}^{p-2} \sigma_x(j)\, \sigma_x(j+2) \rho_{x_j, x_{j+2}} \\ &+ ... + 2\, \sigma_x(1)\, \sigma_x(p) \rho_{x_1, x_p},\end{aligned}$$

where $\sigma_x(j)^2$ is the variance of $x(j)$, and $\rho_{x_j, x_{j+1}}$ is the autocorrelation of x. We also have an analogous equation for the variance of the variable y.

It is well known that there is an intraday pattern of volatility, which is common to all the stocks (Allez and Bouchaud, 2011). This means that, without introducing a large error, we can set:

$$\sigma_x(j) = k_x \cdot f(j); \;\; \sigma_y(j) = k_y \cdot f(j), \;\; \forall j = 1, ..., p \tag{A.1}$$

where k_x and k_y are parameters specific to the two stocks, and $f(j)$ describes the (common) intraday pattern of volatility. This assumption can be used to simplify the expression for the variance of x:

$$\sigma_x^2 = k_x^2 \left[\sum_{j=1}^{p} f(j)^2 + 2 \sum_{j=1}^{p-1} f(j)\, f(j+1) \rho_{x_j, x_{j+1}} + 2 \sum_{j=1}^{p-2} f(j)\, f(j+2) \rho_{x_j, x_{j+2}} + ... + 2\, f(1)\, f(p) \rho_{x_1, x_p} \right],$$

where Eq.1 has been used to describe the intraday pattern of volatility. Similarly, we obtain the variance of y:

$$\sigma_y^2 = k_y^2 \left[\sum_{j=1}^{p} f(j)^2 + 2 \sum_{j=1}^{p-1} f(j)\, f(j+1) \rho_{y_j, y_{j+1}} + 2 \sum_{j=1}^{p-2} f(j)\, f(j+2) \rho_{y_j, y_{j+2}} + ... + 2\, f(1)\, f(p) \rho_{y_1, y_p} \right].$$

The covariance of x and y is then:

$$cov(x,y) = < x\, y > = \frac{1}{T} \sum_{i=1}^{T} \left[\left(\sum_{j=1}^{p} x_i(j) \right) \cdot \left(\sum_{l=1}^{p} y_i(l) \right) \right] =$$

$$= k_x\, k_y \left\{ \left[\sum_{j=1}^{p} f(j)^2 \rho_{x_j, y_j} \right] + \left[\sum_{j=1}^{p-1} f(j) f(j+1) (\rho_{x_j, y_{j+1}} + \rho_{x_{j+1}, y_j}) \right] + ... + f(1) f(p) (\rho_{x_1, y_p} + \rho_{x_p, y_1}) \right\}.$$

Therefore the synchronous correlation coefficient between x and y is given by:

$$\rho_{x,y} = \frac{\left[\sum_{j=1}^{p} f(j)^2 \rho_{x_j, y_j} \right] + \left[\sum_{j=1}^{p-1} f(j) f(j+1) (\rho_{x_j, y_{j+1}} + \rho_{x_{j+1}, y_j}) \right] + ... + f(1) f(p) (\rho_{x_1, y_p} + \rho_{x_p, y_1})}{\sqrt{\left(\sum_{j=1}^{p} f(j)^2 + 2 \sum_{j=1}^{p-1} f(j)\, f(j+1) \rho_{x_j, x_{j+1}} + ... \right) \left(\sum_{j=1}^{p} f(j)^2 + 2 \sum_{j=1}^{p-1} f(j)\, f(j+1) \rho_{y_j, y_{j+1}} + ... \right)}}$$

If we assume that all lagged cross-correlations evaluated at a lag larger than 1 are equal to 0, and that all the auto-correlations are negligible then:

$$\rho_{x,y} = \frac{\sum_{j=1}^{p} f(j)^2 \rho_{x_j, y_j}}{\sum_{j=1}^{p} f(j)^2} + \frac{\sum_{j=1}^{p-1} f(j) f(j+1) (\rho_{x_j, y_{j+1}} + \rho_{x_{j+1}, y_j})}{\sum_{j=1}^{p} f(j)^2}.$$

This expression for $\rho_{x,y}$ is easy to interpret as the sum of two terms with different meanings. The first term is a weighted average of the synchronous correlations between x and y in the p sub-intervals of Δt minutes, with weights that solely depend on the intraday volatility pattern. This term cannot be larger than $\max(\{\rho_{x_j, y_j}; j = 1, ..., p\})$, so it cannot be used to explain the Epps effect. The second term involves lagged correlations $\rho_{x_j, y_{j+1}}$ and ρ_{x_{j+1}, y_j}. If their sum is positive then this term will be positive, and, therefore, may explain the Epps effect.

Appendix B. Contribution of high volatility period to lagged correlations

The months of August to October 2011 witnessed a volatile period in U.S. stock exchanges. Here we examine the influence of this period on the results presented in the text. We may quantify the contribution of each day in the data to the average lagged correlation in each intraday period as follows. Using equation (2) of the text, we may write the mean lagged correlation as averaged overall all N^2

stock pairs as the sum:

$$\begin{aligned}\langle C\rangle &= \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}\left[\frac{1}{T-1}\sum_{i=1}^{T}\frac{(A_{m,i}-\langle A_m\rangle)(B_{n,i}-\langle B_n\rangle)}{\sigma_m\sigma_n}\right] \\ &= \sum_{i=1}^{T}\left[\frac{1}{N^2(T-1)}\sum_{m=1}^{N}\sum_{n=1}^{N}\frac{(A_{m,i}-\langle A_m\rangle)(B_{n,i}-\langle B_n\rangle)}{\sigma_m\sigma_n}\right] \\ &\equiv \sum_{i=1}^{T}\langle C\rangle_i\end{aligned}$$

with $\langle C\rangle_i$ the defined as the term in brackets in the second line. The sum of these terms is then the average lagged correlation associated with each intraday period. We plot the time-series of these terms for each intraday period in Figure B.11.

The period of August through October 2011 appears as a volatile portion of the time series for each intraday period. The contribution of this period is particularly pronounced toward the end of the trading day, where a small number of days seem to contribute disproportionately to the average lagged correlation. We therefore remove all days in August, September, and October 2011 to test the robustness of our results when excluding periods of financial crisis. In Figure B.12 we compare the numbers of validated positive and negative links using all available days in the data with those excluding the period August-October 2011. We find that the influence of this volatile period on the statistically-validated networks is largest at the end of the trading day, and that the lagged relationships uncovered by the analysis are otherwise robust. This is corroborated by Figure B.13, where we see that the characteristic positive shift in the distribution of lagged correlations at the end of the trading day is weakened upon excluding the months of August through October 2011.

We additionally examine the effect of this period on the reconstruction analysis presented in the text. In Figure B.14 we display the results of the reconstruction analysis for the 2011-13 data both including and excluding the months of August through October 2011. We again find that the effect of these months is most pronounced at the end of the trading day, from 1:50 p.m. to 4:00 p.m.. We also see that, while this period contributed disproportionately to the measured lagged cross-correlations, it had little effect on the measured autocorrelations, which continue to contribute to the reconstructed 195 minutes synchronous correlation.

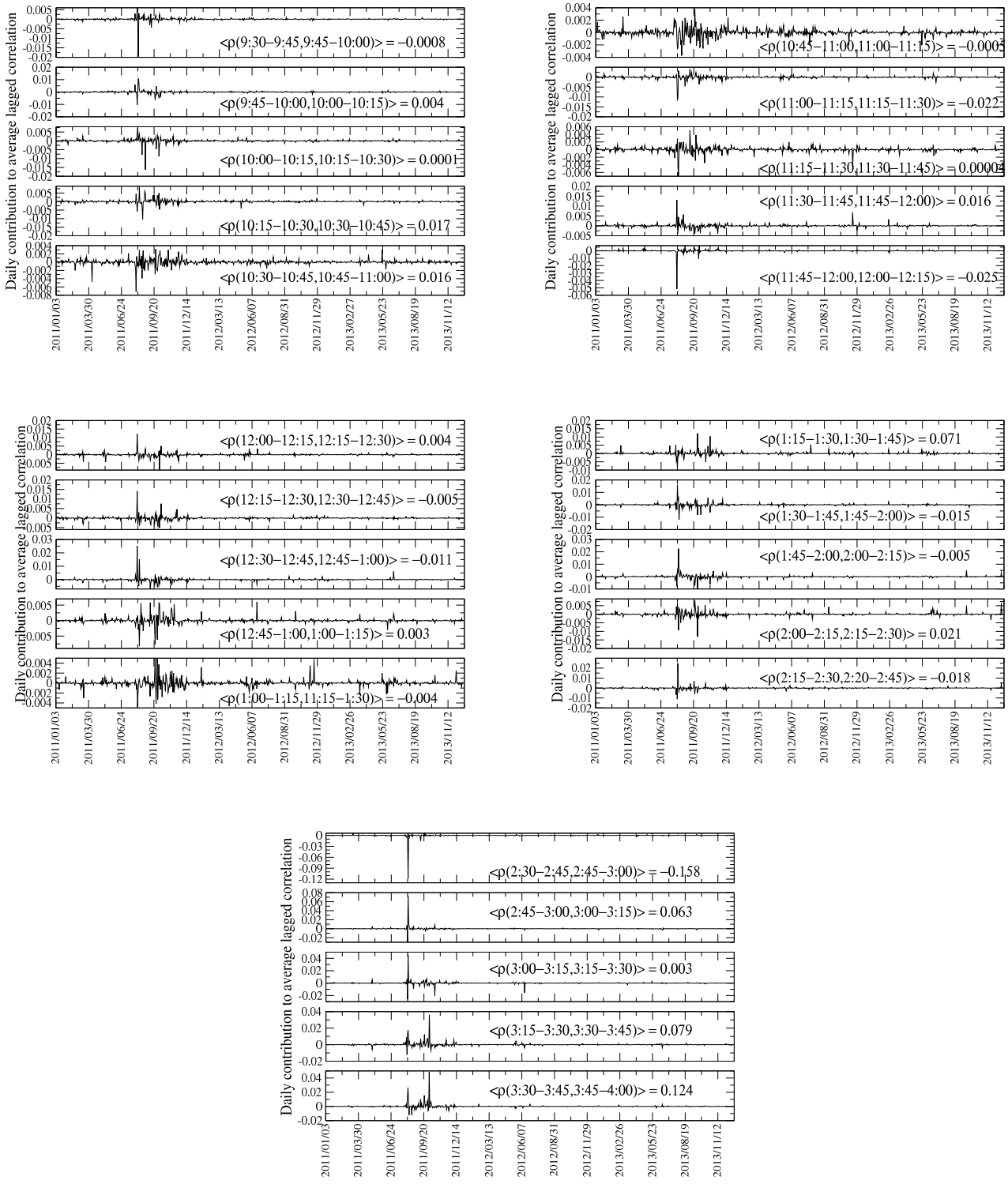

Figure B.11: Contributions $\langle C \rangle_i$ of each day i in the 2011-13 data to the mean lagged correlation measured for each intraday period. Each row of each subfigure corresponds to a lagged correlation between two consecutive intraday periods. Inset provides the mean lagged correlation as averaged over all stock pairs.

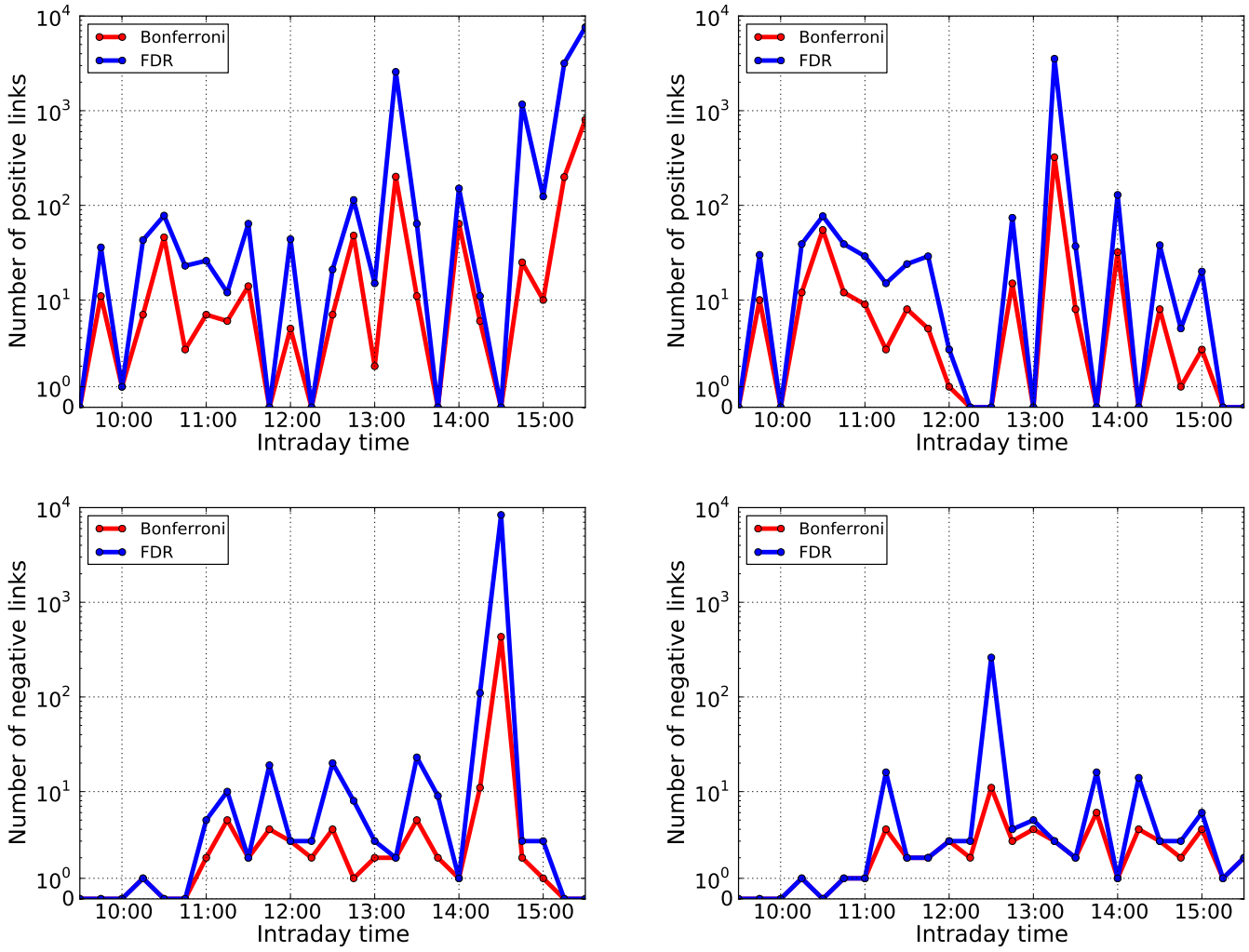

Figure B.12: Top row: number of validated positive links in the 2011-13 data for all days (left) and after removal of August-October 2011 (right). Bottom row: number of validated negative links in the 2011-13 data for all days (left) and after removal of August-October 2011 (right).

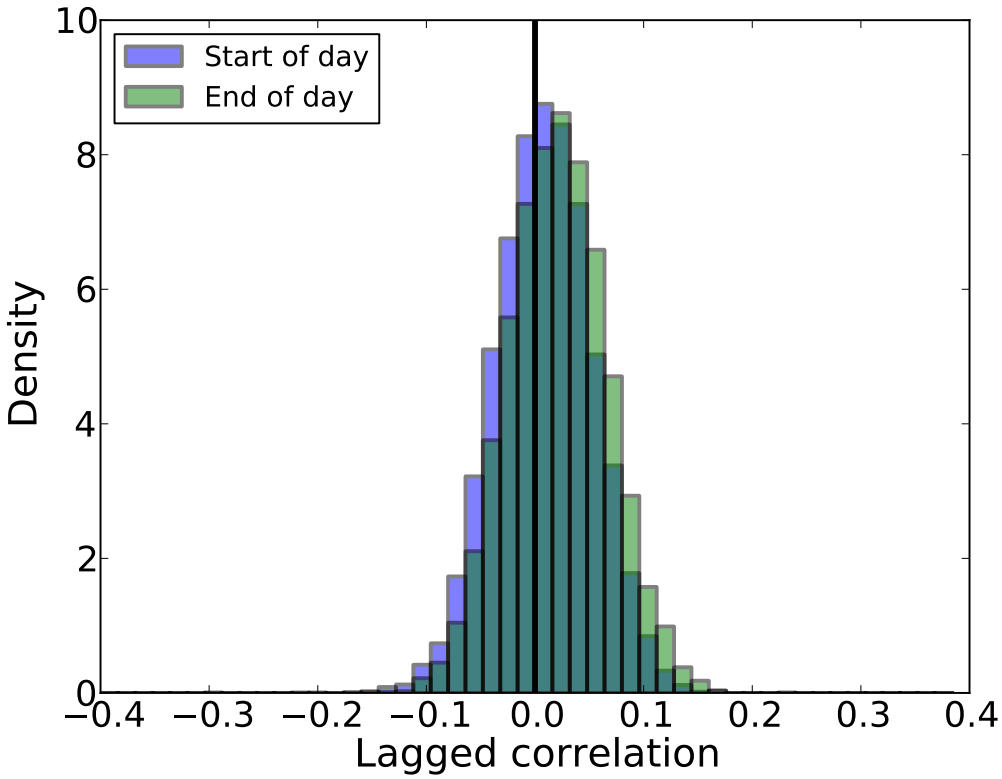

Figure B.13: Normalized histograms of all $N^2 = 100^2 = 10,000$ lagged correlation coefficients for two intraday periods in 2011-13, excluding the months of August through October 2011. The blue shaded histogram corresponds to correlations between returns in the first 15 minutes of the trading day (9:30 a.m. to 9:45 a.m.) and those in the second 15 minutes (9:45 a.m. to 10:00 a.m.). The green shaded histogram corresponds to correlations between returns in the second-to-last 15 minutes of the trading day (3:30 p.m. to 3:45 p.m.) and those in the last 15 minutes (3:45 p.m. to 4:00 p.m.). The characteristic positive shift in the lagged correlations in the final minutes of the trading day has weakened upon excluding the months of August through October 2011.

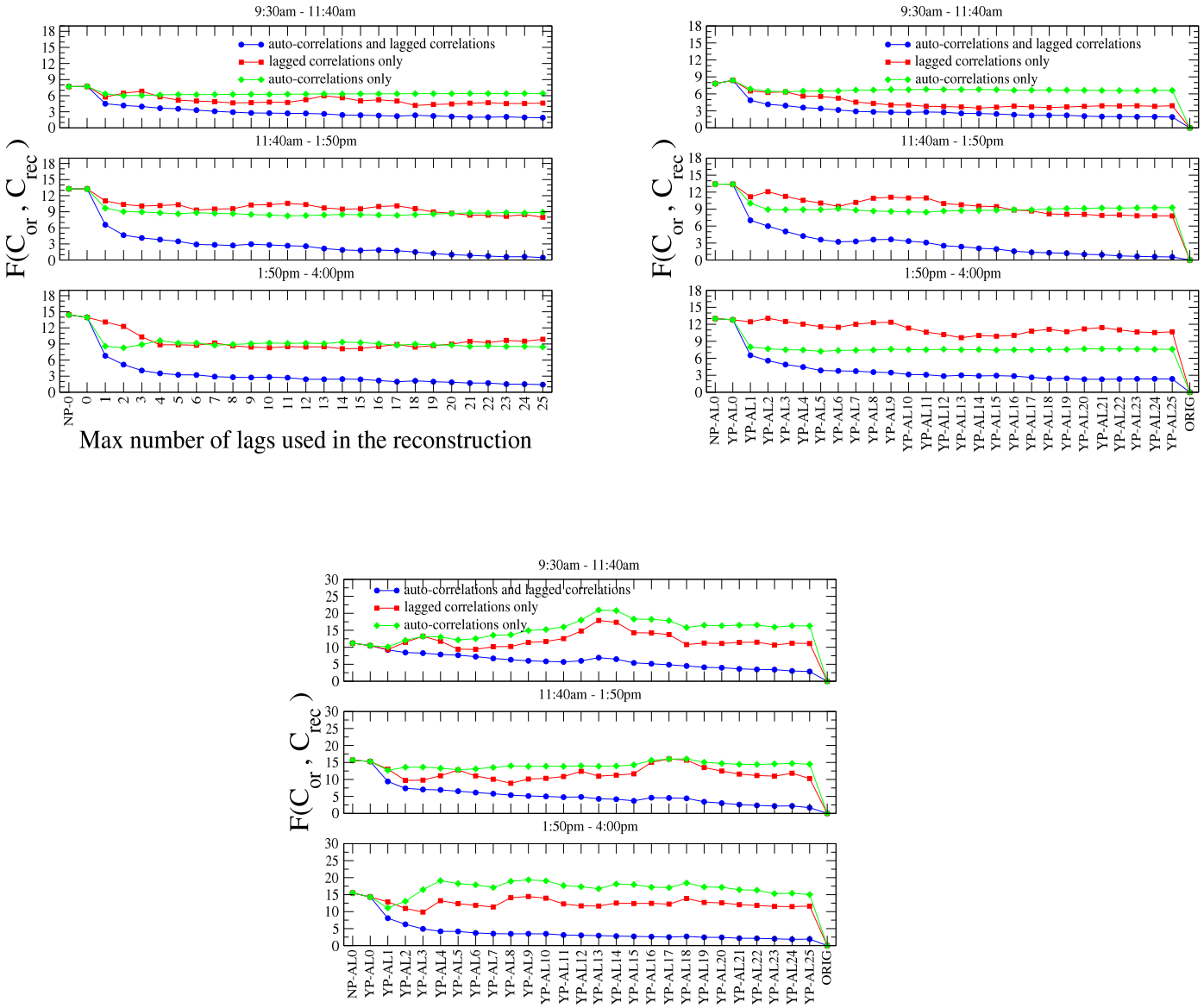

Figure B.14: Frobenius distance between the 130 minute return correlation matrix of the 100 most capitalized stocks traded at NYSE, C_{or}, and the corresponding correlation matrix, C_{rec}, reconstructed according to the method described in the text in the three 130 minute segments of the trading day: from 9:30 a.m. to 11:40 a.m. (top panels), from 11:40 a.m. to 1:50 p.m. (middle panels), and from 1:50 p.m. to 4:00 p.m. (bottom panels). All data are from 2011-13. In the left panel we show results using the entire period, as in the manuscript; in the middle panel we exclude the months of August through October 2011; and in the right panel we show the results using only this period. Each value reported in the horizontal axis indicates the number of lags used to reconstruct 130 minute return correlations from from 5 minute return (lagged and synchronous) correlations. The first point from the left in each panel, labeled "NP-0", is obtained by disregarding the intraday pattern of volatility, which is considered in all the other reconstructed matrices. Three curves are shown in each panel: the green (red) curve describes the results obtained by only including autocorrelation (lagged cross-correlation) terms in the equation used to reconstruct synchronous correlations, while the blue curve shows results in the case in which both autocorrelation and lagged cross-correlation terms are included in the reconstruction equation.

Refrences

Admati, A. R., Pfleiderer, P., 1988. A theory of intraday patterns: Volume and price variability. Review of Financial studies 1 (1), 3–40.

Allez, R., Bouchaud, J.-P., FEB 17 2011. Individual and collective stock dynamics: intra-day seasonalities. NEW JOURNAL OF PHYSICS 13.

Allez, R., Bouchaud, J.-P., 2011. Individual and collective stock dynamics: intra-day seasonalities. New Journal of Physics 13 (2), 025010.

Andersen, T. G., Bollerslev, T., 1997. Intraday periodicity and volatility persistence in.

Arianos, S., Carbone, A., 2009. Cross-correlation of long-range correlated series. Journal of Statistical Mechanics: Theory and Experiment 2009 (03), P03037.

Barndorff-Nielsen, O. E., Shephard, N., 2004. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. Econometrica 72 (3), 885–925.

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.

Bonanno, G., Caldarelli, G., Lillo, F., Mantegna, R., OCT 2003. Topology of correlation-based minimal spanning trees in real and model markets. PHYSICAL REVIEW E 68 (4, 2).

Brownlees, C. T., Gallo, G. M., 2006. Financial econometric analysis at ultra-high frequency: Data handling concerns. Computational Statistics & Data Analysis 51 (4), 2232–2245.

Carbone, A., 2009. Detrending moving average algorithm: a brief review. In: Science and Technology for Humanity (TIC-STH), 2009 IEEE Toronto International Conference. IEEE, pp. 691–696.

Carbone, A., Castelli, G., 2003. Scaling properties of long-range correlated noisy signals: application to financial markets. In: Proc. of SPIE Vol. Vol. 5114. p. 407.

Chou, Y., 1975. Statistical analysis: with business and economic applications. Holt, Rinehart and Winston New York.

Curme, C., Tumminello, M., Mantegna, R. N., Stanley, H. E., Kenett, D. Y., 2014. Emergence of statistically validated financial intraday lead-lag relationships. arXiv preprint arXiv:1401.0462.

Ederington, L. H., Lee, J. H., 1993. How markets process information: News releases and volatility. The Journal of Finance 48 (4), 1161–1191.

Epps, T., 1979. Comovements in stock prices in the very short run. J. Amer. Statist. Assoc. 74 (12), 291–298.

Glasserman, P., Young, H. P., 2015. How likely is contagion in financial networks? Journal of Banking & Finance 50, 383–399.

Hamao, Y., Masulis, R. W., Ng, V., 1990. Correlations in price changes and volatility across international stock markets. Review of Financial studies 3 (2), 281–307.

Kenett, D. Y., Raddant, M., Lux, T., Ben-Jacob, E., 2012. Evolvement of uniformity and volatility in the stressed global financial village. PloS one 7 (2), e31144.

Kenett, D. Y., Tumminello, M., Madi, A., Gur-Gershgoren, G., Mantegna, R., Ben-Jacob, E., 2010. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. PloS one 5 (12), e15032.

Kritzman, M., Li, Y., 2010. Skulls, financial turbulence, and risk management. Financial Analysts Journal 66 (5), 30–41.

Laloux, L., Cizeau, P., Bouchaud, J., Potters, M., AUG 16 1999. Noise dressing of financial correlation matrices. PHYSICAL REVIEW LETTERS 83 (7), 1467–1470.

Lundin, M. C., Dacorogna, M. M., Müller, U. A., 1998. Correlation of high frequency financial time series. Available at SSRN 79848.

Mantegna, R., SEP 1999. Hierarchical structure in financial markets. EUROPEAN PHYSICAL JOURNAL B 11 (1), 193–197.

Markowitz, H., 1952. Portfolio selection*. The journal of finance 7 (1), 77–91.

Muchnik, L., Bunde, A., Havlin, S., 2009. Long term memory in extreme returns of financial time series. Physica A: Statistical Mechanics and its Applications 388 (19), 4145–4150.

Onnela, J., Chakraborti, A., Kaski, K., Kertesz, J., Kanto, A., NOV 2003. Dynamics of market correlations: Taxonomy and portfolio analysis. PHYSICAL REVIEW E 68 (5, 2).

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L., Stanley, H., AUG 16 1999. Universal and nonuniversal properties of cross correlations in financial time series. PHYSICAL REVIEW LETTERS 83 (7), 1471–1474.

Toth, B., Kertesz, J., APR 15 2009. Accurate estimator of correlations between asynchronous signals. PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS 388 (8), 1696–1705.

Tumminello, M., Lillo, F., Mantegna, R., 2010. Correlation, hierarchies, and networks in financial markets. Journal of Economic Behavior & Organization 75 (1), 40–58.

Tumminello, M., Miccichè, S., Lillo, F., Piilo, J., Mantegna, R., 2011. Statistically validated networks in bipartite complex systems. PloS one 6 (3), e17994.