How Lead-Lag Correlations Affect the Intraday Pattern of Collective Stock Dynamics [1]

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

[a] *Center for Polymer Studies and Department of Physics, Boston University*
[b] *Office of the Chief Economist, Financial Industry Regulatory Authority (FINRA)*
[c] *Dipartimento di Fisica e Chimica, Università degli Studi di Palermo*
[d] *Department of Computer Science, University College London*
[e] *Complexity Science Hub Vienna*
[f] *Dipartimento di Scienze Economiche, Aziendali e Statistiche, University of Palermo*

Abstract

Properly estimating correlations and understanding how they change under different economic conditions plays a key role in asset pricing models, risk management, and many econometric models. In this paper we introduce a robust framework to identify a meaningful correlation relationship, address different types of correlations and their interplay, and address correlations across different time scales. First, we present a methodological framework to estimate synchronous, lagged, and autocorrelations for stock price return time series, and validate their statistical significance across different time horizons. Second, we explore the interplay between these different comovement relationships, using a model to uncouple the factors contributing to the intraday pattern of contemporaneous correlations, including volatility,

[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, the U.S. Department of the Treasury, or the Financial Industry Regulatory Authority (FINRA). This paper was partly produced while Dror Kenett was employed by the OFR. This paper benefited from helpful comments by Daniel Barth, Robert Engle, Mark Flood, Shlomo Havlin, Benjamin Kay, Bruce Mizrach, Jonathan Sokobin, and participants of the NYU Volatility Institute Quantitative Financial Econometrics Seminar series. CC, and HES are thankful to the Office of Naval Research (ONR Grant N000141410738) for financial support. Corresponding author email: dror.kenett@finra.org (Dror Y. Kenett), michele.tumminello@unipa.it (Michele Tumminello)

autocorrelations and lagged cross-correlations. Third, we use the methodological framework to investigate correlations between stocks traded on the New York Stock Exchange in the periods 2001-03 and 2011-13, and provide insights on how correlations and their dynamics have changed over time.

Keywords: Financial markets, Market structure, Correlation analysis, Epps effect, Lead-lag
JEL: G21, D85, N26, G18

---

1. Introduction

While the concept of correlations is a widely used and highly studied one, some key aspects still require additional insights. For instance, researchers have long understood that non-synchronous trading (trading on different time scales) can lead to misleading inferences about the price and return cross-correlations among stocks. Moreover, as technologies advance and traders rely upon systems that measure trade and order submission times in micro- and milli-seconds, understanding how to address correlations across different time scales has becomes more important. The way in which different types of meaningful correlation relationships interact synchronously and across different time scales is still not fully understood. This is of importance when one type of correlation can result in an increase in another type of correlation, which ultimately leads to acceleration in market downturns. When constructing an optimal portfolio of assets, one's goal is typically to allocate resources so as to balance the tradeoff 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 market-wide systemic risks (Bisias et al., 2012).

In equity markets, 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 are likely 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. Robust, meaningful information on such a lead-lag relationship

is critical for the understanding of the market dynamics and the underlying mechanisms responsible for it. Recently, Curme et al. (2015) have made use of the Statistically Validated Network (SVN) methodology to estimate and validate lead-lag relationships. In this paper we make use of the SVN methodology (Tumminello et al., 2011) to study the structure and dynamics of of intraday returns correlation of the stocks trading on the New York Stock Exchange (NYSE), throughout the trading day. To this end, we analyze high frequency data of stock traded at the New York Stock Exchange in the periods 2001-03 and 2011-13. To control for potential noise arising from market structure issues at very short time frequencies, we focus on price movements at 15 minute intervals [2]. 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 changes in trading technologies, investor behaviors, and the regulatory environment. Our results present evidence for the existence of lead-lag relationships among price movements. 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. 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. Finally, we propose a model for decomposing contemporaneous cross-correlation at a given time scale to different types of cross-correlations in shorter time scales. Our model provides new insights into the Epps effect (Epps, 1979), providing information on the interplay between different forms of comovement.

It is a well-known empirical fact that intraday trading volume has a U-shaped pattern. Heavy trading occurs at the beginning and the end of the trading day, while light trading occurs in the middle of the day. Admati and

[2]Microstructure noise, in equity markets, is typically considered to be significant for trading at time scales smaller than 5 minutes. This noise is attributed to such factors as bid-ask bounce, execution quality, discreteness of price change, and latency issues. These factors, at short time scales, can introduce spurious, or synthetic correlation that is not informative or statistically significant. See for example Madhavan (2000).

Pfleiderer (1988) show theoretically that informed traders will act strategically by timing their trades for high trading volume periods, or during the first and the last half-hours. Gerety and Mulherin (1992) analyze the extent to which the opening and close of financial markets effect trading volume. They find that the clustering of volume around the open and close of trading results from the desire of investors to reduce the risk of overnight exposures of their positions. This paper provides a new approach to studying the collective dynamics of stock pricing throughout the trading day, and thus contributes to this stream of literature.

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 coefficient 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 model (Laloux et al., 1999). Clustering algorithms can also be applied to the correlation matrix to elicit information about emergent structures in the system (Mantegna, 1999). The aim of using such methods is to provide the means to filter the pairwise correlation coefficients to find some meaningful structures, usually thought of in terms of groups, or clusters, of the underlying assets. Some commonly used procedures of filtering correlation matrices have been proposed in the context of correlation-based network methodologies. For example, such structures can 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 the similarities between the different variables (Mantegna, 1999). 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, 2015) and by investigating Granger-causality networks (Billio et al., 2012).

In the context of financial markets, the correlation matrix among asset returns is an object of central importance for different applications, such

as for the measurement, management, and mitigation of 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 lagged cross-correlation relationships among the 80 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 the beginning of the current decade. The resulting 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 two decades, 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 an increase of statistically validated lagged cross-correlation and autocorrelations 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. 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.

The remainder of this paper is organized as follows: In Section 2, we provide a literature review of the main issues related to this paper. In Section 3 and Section 4, we develop our methodological framework, including the statistical validation methodology, the correlation decomposition and recon-

struction models, and some measures of testing the outcomes of the models. In Section 5 we present an empirical application of the methodological framework. We discuss the data, the results of the validation process, the intraday dynamics of synchronous, auto and lagged correlations, and a comparison of the findings between the two investigated datasets. We further present a focused case study, by studying only stocks belonging to the banking sector. Finally, in Section 6 we present a summary of the main findings and their applications.

2. Literature review

There exists a large body of literature on the topics of correlation and comovement within and between financial markets. For example, Forbes and Rigobon (2002) compare cross-country correlations of asset returns in tranquil and crisis periods, testing for contagion and dependence. The authors claim that cross country relationships are that of dependence, which they define as having a high level of market co-movement at both tranquil and crisis period. The authors reject the hypothesis of having contagion relationships, which they define as significant increase in market comovement following a crisis event in one country. This was later revisited by Corsetti et al. (2005), who argue that under different model setups, that consider differently the features of the country specific-crisis event, it is possible to observe both contagion and dependence relationships across countries. Pollet and Wilson (2010) study the average correlation between daily stock returns in the U.S. stock market, and find that it can predict subsequent quarterly stock market excess returns. Furthermore, they show that changes in stock market risk, holding average correlation constant, can be interpreted as changes in the average variance of individual stocks, and have a negative relation with future stock market excess returns. Moreover, correlation based measures, and their derivatives, have been recently introduced as measures of systemic risk in financial markets (see Kritzman et al., 2011; Brownlees and Engle, 2016)

The main body of work on intra-day patterns has focused on returns, volume, and volatility. For example, Hasbrouck and Seppi (2001) investigate how important are cross-stock common factors in the price discovery/liquidity provision process in equity markets. Using principal components and canonical correlation analyses, they show that both returns and order flows, for a sample of the 30 Dow Jones Index components, are characterized by common factors. Furthermore, they show that commonality in

the order flows explains roughly two-thirds of the commonality in returns. Allez and Bouchaud (2011) study several stylized facts concerning the intraday seasonalities of stock dynamics. Using 5-minute returns for 126 large market cap stocks traded in the New York Stock Exchange (NYSE), for the period January 2000 - December 2009, they find that the average correlation between stocks increases throughout the day, leading to a smaller relative dispersion between stocks. However, they find that the kurtosis reaches a minimum at the open of the market, when the volatility is at its peak. This can be interpreted as that during large market swings, the idiosyncratic component of the stock dynamics deceases significantly. Thus, the authors claim that early hours of trading are dominated by idiosyncratic or sector specific effects, whereas the influence of the market factor increases throughout the day.

When investigating intraday to high frequency data, a major issue that arises is the choice of sampling frequency and time horizon. Andersen et al. (2003) use a vector autoregressive (VAR) framework for the daily realized variances and covariance of two exchange rates (DEM/USD and YEN/USD) for the period of December 1986 to June 1999, for all exchange rate quotes. The authors discuss the information content of the raw quotes, and discuss the various market microstructure "frictions" they are exposed to, including strategic quote positioning and standardization of the size of the quoted bid/ask spread. The authors claim that such features are generally immaterial when analyzing longer horizon returns, but that such longer time horizons may distort the statistical properties of the underlying high-frequency intraday returns. Moreover, the authors claim that the sampling frequency at which such considerations become a concern is intimately related to market activity. The authors claim that for the investigated exchange rate data, the use of equally-spaced thirty-minute returns strikes a satisfactory balance between these two considerations.

Fleming et al. (2003) use five-minute returns on three actively traded futures contracts (S&P 500 index, Treasury bonds, and gold) to show that a mean-variance efficient investor would be willing to pay 50 to 200 basis points per annum for being able to use daily covariance matrix forecasts based on high-frequency intraday returns instead of daily returns. They make use of the 5-minute time horizon following the discussion of the effects of microstructure issues in shorter time horizons in Andersen et al. (2003) and Andersen et al. (2001). Similarly, Liu (2009) constructs and assesses the performance of the minimum variance portfolio and the minimum tracking

error portfolio (tracking the S&P 500 index) using five-minute returns for the 30 Dow Jones index constituents. Pooter et al. (2008) find that using daily conditional covariance matrix forecasts based on high-frequency intraday returns instead of daily returns considerably improves portfolio performance. For the global minimum risk portfolios, the authors claim that the optimal sampling frequency for the S&P 100 constituents ranges between 30 and 65 minutes (see also Bandi and Russell (2008).

Researchers have proposed a variety of methods to estimate and study intraday patterns of stock dynamics. Engle and Granger (1987) proposed a representation theorem that connects the moving average, autoregressive, and error correction representations for cointegrated systems. McInish et al. (1995) investigate synchronous transactions data for IBM from the New York, Pacific, and Midwest Stock Exchanges, to estimate an error correction model to investigate whether each of the exchanges is contributing to price discovery. They find that all three exchanges maintain a long-run cointegration equilibrium; that is, IBM prices on the NYSE adjust toward IBM prices on the Midwest and Pacific Exchanges, just as Midwest and Pacific prices adjust to the NYSE. Lin et al. (1994) propose a GARCH process to investigate how returns and volatilities of stock indices are correlated between the Tokyo and New York markets. Using intra-day data for both daytime and overnight returns for both markets, the authors find that Tokyo (New York) daytime returns are correlated with New York (Tokyo) overnight returns. De Jong and Nijman (1997) propose an estimator that avoids imputation and uses all available transactions to calculate (cross) covariances. They propose this measure as a possible way to analyze lead-lag relationships at arbitrarily high frequencies without additional imputation bias. The authors apply this measure empirically to test lead-lag relationship between the S&P 500 index and futures written on it. Hayashi et al. (2005) consider the case when two continuous diffusion processes are observed only at discrete times in a non-synchronous manner, and propose an estimator that is free of synchronization biases. In the case of diffusion-type processes with independent random observation times, they showed that their estimator is consistent for the underlying covariation as the size of observation intervals goes to zero. Hansen and Lunde (2005) derive an empirical measure of daily integrated variance (IV), for cases where high-frequency price data are unavailable for part of the day. They propose that the optimal combination of the realized variance and squared overnight return can be determined, and relate their results, both theoretical and empirical, to the problem of combining forecasts.

Christensen et al. (2010) propose a pre-averaged realized covariance estimator, and discuss the problem of measuring the ex-post covariance of financial asset returns under microstructure noise and non-synchronous trading.

At short time scales, synchronous correlations among stock returns tend to be lower in magnitude, which is known as the Epps effect (Epps, 1979). In his seminal paper, Epps reported results showing that stock return correlations decrease as the sampling frequency of data increases. Two main factors have been proposed as the underlying processes that result in the Epps effect. The first one is a possible lead-lag effect between stock returns. In this case, the maximum of the time-dependent cross-correlation function can be observed at non zero time lag, resulting in increasing cross-correlations as the sampling time horizon approaches the same order of magnitude as the characteristic lag. Tóth and Kertész (2006) have shown that throughout the past few years this effect has become less important as the characteristic time lag has shrunk, which could possibly be associated with an increase in market efficiency. The second factor is the asynchronicity of ticks in case of different stocks. Empirical results reported by Renò (2003) show that taking into account only the synchronous ticks decreases the Epps effect to a great extent, i.e., the observed correlation coefficients for short sampling time horizons actually increase. One could reasonably expect that for a given sampling frequency, an increase in trading activity would result in a decrease of the the asynchronicity, leading to a weaker Epps effect. Renò (2003) investigated this hypothesis using Monte Carlo experiments, and found an inverse relation between trading activity and the correlation drop. Tóth and Kertész (2009) study the Trade and Quote (TAQ) Database of the New York Stock Exchange (NYSE) for the period of 1/4/1993 to 12/31/2003, and propose a new description of the Epps effect by studying a decomposition of cross-correlations estimators into the different elements of their analytical expression. The authors find that the origin of the independence of the characteristic time scale of the Epps effect on the trading frequency is the presence of a human time scale in the time lagged autocorrelation functions.

3. Statistically validated correlation analysis

At short time scales, lagged correlations among assets may become non-negligible (Toth and Kertesz, 2009; Curme et al., 2015). 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 thresholding 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., 2015). 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

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

where $P_i(t)$ is the price of stock i at time t, and Δt is the sampling time resolution.

We perform a lagged-correlation analysis between all possible stock pairs. Lagged-correlation is a standard method of estimating the degree to which two series are correlated in a nonsynchronous way (see for example (Muchnik et al., 2009; Arianos and Carbone, 2009)). In this paper, we propose a lagged correlation estimator to account for intraday patterns in return time series. Specifically, we focus on the returns matrix at the $\Delta t = 15$ minute time horizon, which avoids the potential noise resulting from mechanical market structure issues, and divide each trading day into non-overlapping Δt parts $(\Delta t_1, \Delta t_2, ..., \Delta t_{26})$[3,4]. 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 and lag q, 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 -

[3]A trading day at NYSE is 390 minutes long, which implies 26 time windows of 15 minutes each.

[4]The statistically validated correlation analysis is applicable at additional time horizons, see for example Curme et al. (2015).

10 a.m.), assuming that lag $q = 1$, 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 coefficient, $\rho_{X,Y}(\Delta t_i, \Delta t_{i+q})$ between return of stock X at Δt_i and stock Y at Δt_{i+q}, as the standard Pearson coefficient between the corresponding time series of returns:

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

where all the sums are over the T trading days in the considered dataset. Such a lagged correlation estimator has the advantage that it allows one to eventually describe the intraday pattern of correlation. The disadvantage is that the length of time series, T, which is the number of trading days, is much shorter than the one used to calculate the usual lagged correlation coefficient reported in Eq. (2). For the sake of simplicity, when lag 1 is considered, we indicate the empirical lagged correlation matrix as $C(\Delta t_i)$, by neglecting the indication of lag 1 in the notation.

For each chosen Δt_i, the matrix $C(\Delta t_i) \equiv C$ can be considered a weighted adjacency matrix for a fully connected, directed graph including self-links (representing auto-correlations). 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 matrix 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 (for positive lagged correlations) and D (for negative lagged correlations), where an entry $U_{m,n}$ of U is the number of realizations for which entry $\widetilde{C}_{m,n}$ of C is larger or equal to entry $C_{m,n}$ of C, and entry $D_{m,n}$ of D is the number of realizations in which $\widetilde{C}_{m,n} \leq C_{m,n}$.

Matrix U allows one to 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 (positive) correlation. Similarly, from D we can associate a one-tailed p-value with all negative correlations. In this analysis, we choose as univariate threshold of statistical significance

$p = 0.01$. However, we must adjust our statistical threshold to account for multiple comparisons. We use the False Discovery Rate (FDR) protocol for N stocks, so that adjusted threshold takes into account the number of tests we run for each part of day Δt_i. This correction is constructed as follows. For a sample of $N = 100$ stocks, we construct 10^6 independently shuffled surrogate time series, and, an approximation of the p-value associated with an actual positive (negative) lagged correlation coefficient between two stocks, say, from stock m to stock n, is obtained as the number of shuffled realizations in which the resulting correlation coefficient is larger (smaller) or equal to the actual one, that is $U_{m,n}$, divided by the total number of independently shuffled surrogate time series. 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 $p_k < k\ 0.01/N^2$, if the univariate threshold of statistical significance is set to 0.01. Therefore, for the FDR network, our threshold for the matrices U (or D) 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).

4. Modeling the impact of lead-lag relationships on synchronous correlations

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.

4.1. The contemporaneous correlation reconstruction model

Here, we propose a correlation reconstruction model, which describes how the synchronous correlation between two stock returns, as evaluated at a certain intraday time horizon, can be described by auto- or lagged-correlations estimated at shorter time horizons. For such a reconstruction model, it is useful to compare a relatively long time horizon with a relatively short time horizon. For example, one can consider dividing the trading day in three equal parts of morning, noon, and afternoon. This results in three 130 minute periods of the trading day. For each of these 130 minute periods of the trading day, the estimated correlation coefficients can be decomposed

in order to make apparent the individual contribution of auto-correlations and lagged cross-correlations evaluated at smaller time windows, such as $\Delta t = 5$ minutes. Our approach is similar to the one presented in Toth and Kertesz (2009). 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. Whereas In Appendix B we derive the most general model, here we present and discuss a simple example. 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:

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

This equation clearly shows how the interplay between short-term lagged cross-correlations and auto-correlations 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.

4.2. Testing the impact of autocorrelations

It is useful to validate that the identified statistically significant lagged cross-correlations are not spuriously the result of autocorrelations. To this end, we employ a methodology that uses the concept of partial correlations (Kenett et al., 2010), that results in subtracting off the influence of autocorrelations. This is achieved by recalculating the lagged correlation coefficients, when controlling for a potential mediating effect of the autorcorrelations of either stock X or Y, as follows:

$$\rho(x(t), y(t+\tau)|y(t)) = \frac{\rho(x(t), y(t+\tau)) - \rho(y(t), y(t+\tau))\rho(x(t), y(t))}{\sqrt{[1-\rho(y(t), y(t+\tau))^2][1-\rho(x(t), y(t))^2]}}. \quad (4)$$

$$\rho(x(t), y(t+\tau)|x(t+\tau)) = \frac{\rho(x(t), y(t+\tau)) - \rho(x(t+\tau), y(t+\tau))\rho(x(t), x(t+\tau))}{\sqrt{[1-\rho(x(t+\tau), y(t+\tau))^2][1-\rho(x(t), x(t+\tau))^2]}}, \quad (5)$$

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 (4) and (5), again choosing our statistical threshold to be $p = 0.01$. We then test the extent of overlap between the original networks, and the partial correlation networks. The probability of randomly sampling these intersections x from the $L = N \times (N-1)$ total possible lagged cross-correlation links in n "draws" (links in the original network) is given by the hypergeometric distribution:

$$P(x|n, k, L) = \frac{\binom{k}{x}\binom{L-k}{n-x}}{\binom{L}{n}}, \quad (6)$$

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. In section 5.5 we conclude that statistically significant lagged cross-correlations are not a mere consequence of the presence of statistically validated autocorrelations.

5. Empirical analysis

Making use of the Trades and Quotes (TAQ) database, we consider intraday pricing data for stocks traded on the New York Stock Exchange (NYSE) during two periods: 2001-2003 and 2011-2013. In each period we consider all trading days, with the exception of certain days surrounding U.S. national holidays in which there was little trading activity and early market closures. We remove eleven days in the 2001-2003, dataset for a total of 737 trading days, and seven days in the 2011-2013 dataset, for a total of 747 trading days.

In each period we consider 100 companies with the largest market capitalization as of the final trading day of the period. To control for variations in liquidity, which may induce lagged correlations between returns data, we retain in this set only stocks that exhibit at least 15 transactions in at least 95% of 15 minute intervals considered in the analysis. Of this subset we consider the top 80 stocks in terms of market capitalization (see Appendix A), in order to construct equal-sized sets in the two periods. In Appendix A we present the list of stocks used for the two time periods. In Table 1 we present summary statistics of the number of trades and their distributions across the cross section of securities and time windows of the day.

5.1. Intraday periodicities

Constructing 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.

Using the previous-tick standard, we sample prices for each stock every 15 minutes throughout each trading day and calculate the corresponding logarithmic returns over each interval. These returns form the basis of our analysis. In Table 2 we provide summary statistics of each intraday period in the 2001-2003 and 2011-2013 datasets, including the mean $\langle \rho \rangle$ and standard deviation σ_ρ of synchronous Pearson correlation coefficients between distinct returns series. We also show the mean $\langle C \rangle$ and standard deviation σ_C of entries in the lagged correlation matrix C. For summaries of lagged correlations, the intraday period represents the ending points of two adjacent 15

Table 1: Summary statistics of intraday periods in 2001-2003 and 2011-2013 datasets., For each period, separately, for a given interval $\Delta t = 15$ minutes of the day, we report the maximum number of trades (max), median number of trades (median), and mean number of trades (mean), across all stocks and all days.

Intraday period	2001-2003			2011-2013		
	max	median	mean	max	median	mean
9:00 - 9:30	7539	115	171.33	165366	1032	1971.22
9:45 - 10:00	7967	109	143.29	84884	972	1682.03
10:00 - 10:15	6444	111	143.95	57275	977	1664.13
10:15 - 10:30	4440	103	132.84	55128	827	1382.66
10:30 - 10:45	3369	98	125.83	49252	781	1290.35
10:45 - 11:00	3488	93	118.57	41084	724	1189.45
11:00 - 11:15	3078	90	114.58	66788	705	1134.10
11:15 - 11:30	4115	85	107.94	43440	660	1034.14
11:30 - 11:45	3341	83	104.82	49573	607	972.23
11:45 - 12:00	2487	78	99.29	38060	515	860.38
12:00 - 12:15	4467	77	97.86	55427	487	824.99
12:15 - 12:30	3340	72	90.69	39799	457	765.60
12:30 - 12:45	4474	71	89.21	32877	445	749.10
12:45 - 1:00	4257	69	86.79	47980	425	727.31
1:00 - 1:15	2901	70	87.76	33194	425	725.73
1:15 - 1:30	3612	71	88.11	23008	425	710.80
1:30 - 1:45	2893	73	90.74	28776	444	751.61
1:45 - 2:00	3337	75	93.30	44094	452	761.75
2:00 - 2:15	3324	82	101.13	36955	520	873.10
2:15 - 2:30	2807	84	104.91	44418	514	865.98
2:30 - 2:45	4005	89	110.34	32828	545	895.18
2:45 - 3:00	6125	91	113.18	26517	564	919.65
3:00 - 3:15	4908	99	121.86	51962	671	1080.01
3:15 - 3:30	2847	103	126.10	55906	757	1209.42
3:30 - 3:45	4257	109	135.20	40091	1091	1653.77
3:45 - 4:00	4379	124	152.41	75071	1991	2971.27

minute windows. For example, the 9:45 - 10:00 intraday period represents lagged correlations between returns from 9:30 - 9:45 and those from 9:45 to 10:00.

Figure 1 displays the intraday pattern of the average synchronous correlation between returns of all stock pairs for the sample of top 80 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 as 2011-13. We observe a difference over the past decade in the magnitude of the measured correlations. Both periods exhibit a similar profile in the first parts of the trading day, in terms of the intraday pattern of synchronous correlations, with an abrupt growth in the first thirty minutes of the trading day, that then levels off 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). However, a striking difference is observed between the two periods at the end of the trading day. While in the 2001-03 data the average synchronous correlation drops at the end of the trading day, almost in a reverse "U"-like convex manner, in the 2011-13 we observe a jump in the average correlation. Thus, while at the 2001-03 period the average correlation dampens at the end of the day, it is found to be strongest at the end of the day for the 2011-13 period.

We use the statistical methodology introduced above to construct an analogous profile for lagged correlations. In Figure 2 we plot the average lagged correlation between the same stock pairs from Fig. 1. 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 [5]. In the two upper panels of Figure 2 we plot the intraday pattern of lagged correlations for the stock pairs belonging to the FDR network in blue, for all validated links, only positive validated links, and only negative validated links. 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. We also provide the probability density function of all lagged correlation coefficients for two intraday periods in 2001-03 (bottom left panel) and 2011-13 (bottom

[5] These tails are with respect to the top 1% of the probability distribution, corrected for multiple hypothesis testing, which for example is calculated as $0.01/80^2 = 1.5 \cdot 10^{-6}$.

Table 2: Summary statistics of intraday periods in 2001-2003 and 2011-2013 datasets.

Intraday period	2001-2003				2011-2013			
	$\langle\rho\rangle$	σ_ρ	$\langle C\rangle$	σ_C	$\langle\rho\rangle$	σ_ρ	$\langle C\rangle$	σ_C
9:00 - 9:30	0.12	0.08	-	-	0.20	0.13	-	-
9:45 - 10:00	0.20	0.09	-0.012	0.048	0.33	0.11	0.001	0.045
10:00 - 10:15	0.30	0.12	-0.020	0.049	0.39	0.12	0.002	0.042
10:15 - 10:30	0.25	0.10	0.016	0.052	0.38	0.11	-0.002	0.041
10:30 - 10:45	0.29	0.11	0.038	0.057	0.38	0.11	0.014	0.042
10:45 - 11:00	0.25	0.10	-0.008	0.057	0.36	0.10	0.012	0.039
11:00 - 11:15	0.24	0.09	0.008	0.057	0.38	0.10	-0.005	0.045
11:15 - 11:30	0.25	0.09	-0.008	0.053	0.32	0.09	-0.026	0.042
11:30 - 11:45	0.21	0.09	-0.006	0.055	0.38	0.09	-0.004	0.041
11:45 - 12:00	0.24	0.09	0.032	0.052	0.36	0.10	0.010	0.043
12:00 - 12:15	0.24	0.09	-0.015	0.053	0.41	0.10	-0.030	0.052
12:15 - 12:30	0.23	0.09	0.004	0.051	0.35	0.10	0.003	0.043
12:30 - 12:45	0.20	0.08	0.014	0.052	0.44	0.10	-0.004	0.045
12:45 - 1:00	0.23	0.09	-0.021	0.048	0.43	0.10	-0.018	0.047
1:00 - 1:15	0.22	0.09	-0.019	0.046	0.37	0.10	0.001	0.047
1:15 - 1:30	0.29	0.12	0.004	0.052	0.41	0.10	-0.009	0.041
1:30 - 1:45	0.27	0.10	-0.015	0.089	0.46	0.10	0.070	0.046
1:45 - 2:00	0.28	0.09	-0.005	0.053	0.43	0.10	-0.019	0.045
2:00 - 2:15	0.28	0.09	-0.028	0.053	0.51	0.10	-0.008	0.040
2:15 - 2:30	0.29	0.10	0.008	0.046	0.44	0.10	0.021	0.041
2:30 - 2:45	0.31	0.10	0.003	0.053	0.58	0.10	-0.018	0.056
2:45 - 3:00	0.32	0.10	-0.000	0.048	0.52	0.11	-0.166	0.067
3:00 - 3:15	0.31	0.09	0.023	0.052	0.58	0.11	0.064	0.049
3:15 - 3:30	0.33	0.10	0.034	0.051	0.52	0.11	-0.001	0.056
3:30 - 3:45	0.33	0.09	0.052	0.052	0.55	0.10	0.080	0.047
3:45 - 4:00	0.22	0.08	0.041	0.058	0.62	0.11	0.126	0.049

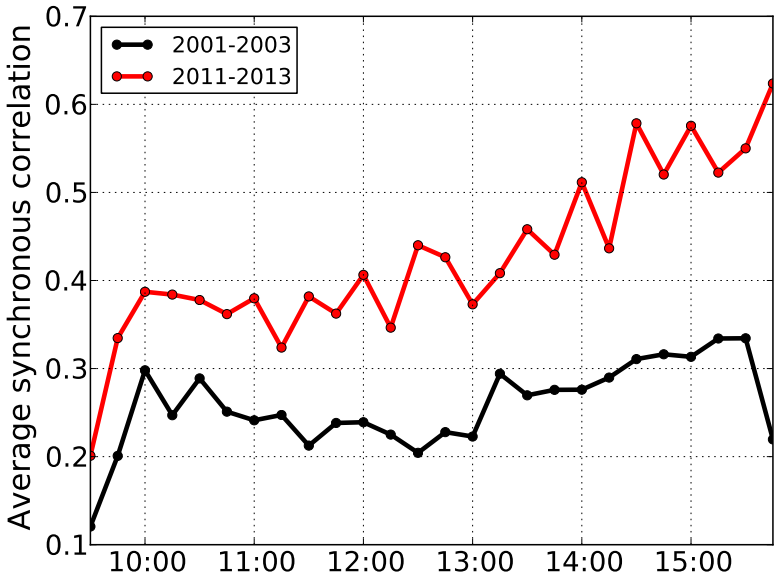


Figure 1: Intraday pattern of the average synchronous correlation between fifteen minute stock returns of 80 highly capitalized stocks traded at NYSE in the period 2001-03 (black continuous line) and 2011-13 (red dashed line). One can observe that at the 2001-03 period the average correlation dampens at the end of the day, while it is found to be strongest at the end of the day for the 2011-13 period.

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. While this is observed in the 2001-03 data, it is significantly more pronounced in the 2011-13 data, which is also consistent with the observation regarding the difference in patterns of the average synchronous correlation at the end of the trading day, presented in Figure 1.

The results presented in Figure 2, and values presented in Table 2 describe the differences between the two time periods, in terms of the magnitude of the validated lagged correlation coefficients. One may also consider the difference in frequency of such statistically significant relationships. In Figure 3 we plot the number of positive (left) and negative links (right), for the 2001-2003 (top) and 2011-2013 (bottom) datasets, for the FDR validation procedure. Studying the evolution of number of validated links provides additional insights into the differences between the two time periods. First, the 2011-13 data exhibits a growing number of both positive and negative validated links towards the end of the trading day. However, in the last 30 minutes of the day, the number of validated negative links becomes minimal, whereas the number of positive links becomes maximal. Second, the 2001-03 data exhibits presence of both positive and negative validated links throughout the day. However, in the 2011-13 data, we observe very few validated positive links before the second half of the trading day. Third, the maximum number of validated links in the 2011-13 data is approximately one order of magnitude larger than that observed for the 2001-03 data.

5.2. Persistence of intraday lead-lag relationships

To what extent do the statistically validated lead-lag relationships between individual stock pairs persist across different parts of 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 for both time periods. However, we find that such persistence

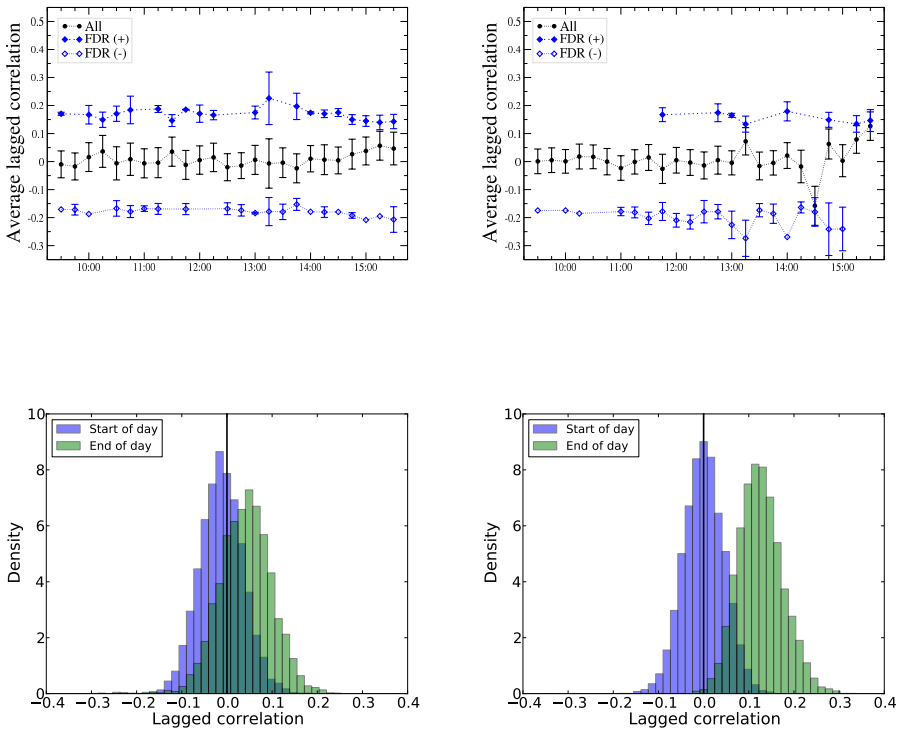
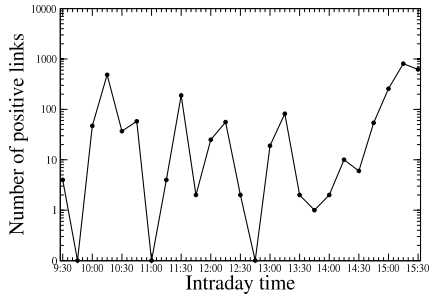
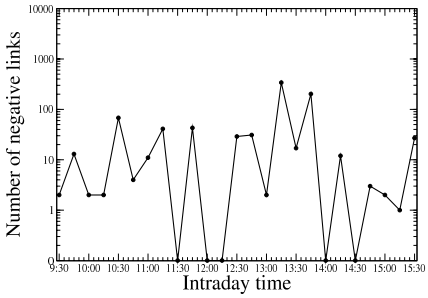


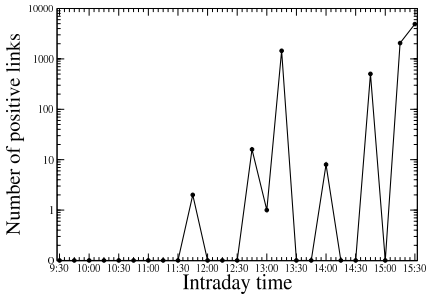
Figure 2: Intraday pattern of the average lagged correlation, evaluated at one lag, between $\Delta t = 15$ minute stock returns of 80 highly capitalized stocks traded at NYSE in the period 2001-03 (top left panel) and 2011-13 (top right panel). In each panel, we also report the pattern of lagged correlation with average taken over all the links that belong to the FDR network (blue diamonds), by distinguishing between positive (+) and negative (-) statistically validated correlations. We also provide the probability density function of all 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.



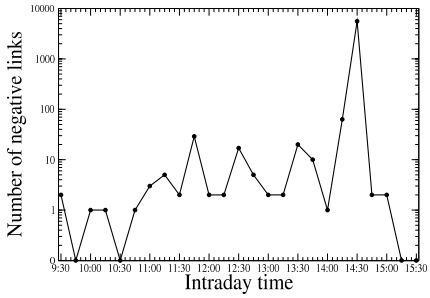
(a) Links of positive correlation, 2001-2003



(b) Links of negative correlation, 2001-2003

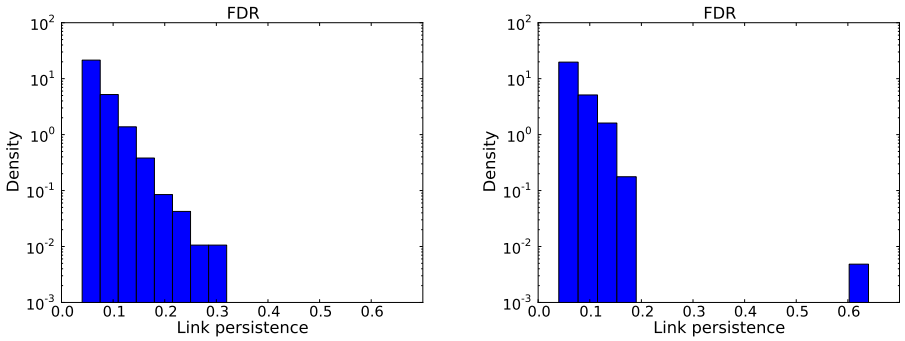


(c) Links of positive correlation, 2011-2013



(d) Links of negative correlation, 2011-2013

Figure 3: Plots of the number of positive and negative validated links for FDR lagged correlation networks. The decrease in number of validated links for increasing time horizon is apparent in both the 2001-2003 and 2011-2013 datasets. The vertical axis is presented on a logarithmic scale that is linearized near zero.



(a) FDR, 2001-2003 (b) FDR, 2011-2013

Figure 4: Distributions of link persistence for all links in networks at a time horizon $\Delta t = 15$ min. Left panel shows results using data from 2001-03; right panel shows results using data from 2011-13. We find that the validated links are generally more persistent in the 2011-13 data throughout the trading day.

is more strongly dependent on the particular intraday period in the 2001-03 data. We support this finding with two analyses.

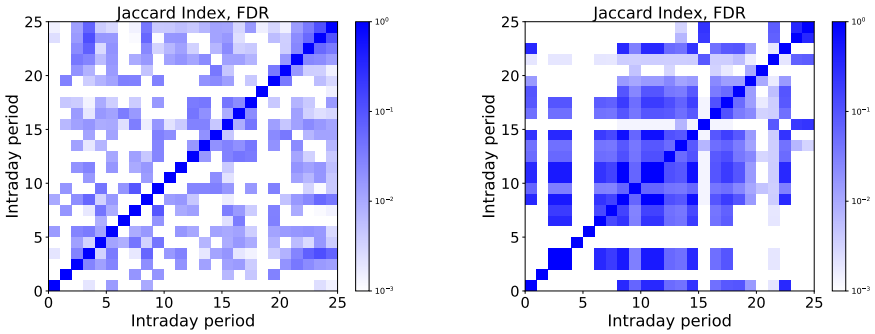
First, we examine this similarity at the level of individual links, by quantifying the persistence of links. This persistence is defined as the fraction of intraday periods (of which there are 26 for $\Delta t = 15$ min.) in which a given link appears. We plot the distributions of link persistence for all intraday periods in Figure 4, where it is possible to observe that individual links seem to be more persistent in the 2011-13 data; however, this analysis does not convey information regarding the number or strength of the validated links.

Second, we quantify the extent to which two intraday periods (defined by the time horizon) exhibit similarity of statistically significant validated links, using the Jaccard Index (Chou, 1975):

$$J(i,j) = \frac{|L_i \cap L_j|}{|L_i \cup L_j|}, \tag{7}$$

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 intraday periods share a large proportion of their total statistically validated links.

In Figure 5 we display matrices of Jaccard Indices $J(i,j)$ between sets of links corresponding to all intraday periods at a time horizon $\Delta t = 15$ min.



(a) FDR, 2001-2003 (b) FDR, 2011-2013

Figure 5: Matrices of Jaccard Indices between sets of links corresponding to networks for all intraday periods at a time horizon $\Delta t = 15$ min. Left panel shows results using data from 2001-03; right panel shows results using data from 2011-13. We find that the validated links are generally more persistent in the 2011-13 data throughout the trading day.

We find that the Jaccard Indices are generally high, suggesting that the links we validate are indeed persistent across many of the intraday 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 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. For the two datasets, it is possible to observe that in the FDR network, there exists persistence of the validated links in the last two intraday periods of the day. Moreover, it is possible to observe that while in the 2001-03 data the persistence is longer (four last intraday periods in the FDR validation), the persistence is stronger in the 2011-13 data. We also observe some persistence in the FDR validation regarding the links observed for the first 15 minutes of the trading day. In both the 2001-03 data. and more significantly in the 2011-13 data, we find time periods later in the trading day that exhibit an overlap of validated links with the first 15 minute period of the day.

5.3. Reconstructing correlations

In the previous section we made use of the statistical validation framework, the SVN approach, to validate different types of empirical intraday correlations, and track their evolution throughout the trading day for the

two investigated periods. To further explore the statistically significant empirical cross-correlation relationships between the different stocks, at different time scales, we make use of the correlation reconstruction model presented above.

In Fig. 6, we show some results of the reconstruction analysis of the 80 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:

$$F(C_{or}, C_{rec}) = \sqrt{\mathrm{tr}\,[(C_{or} - C_{rec})(C_{or} - C_{rec})^T]}, \tag{8}$$

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 of Fig. 6) 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

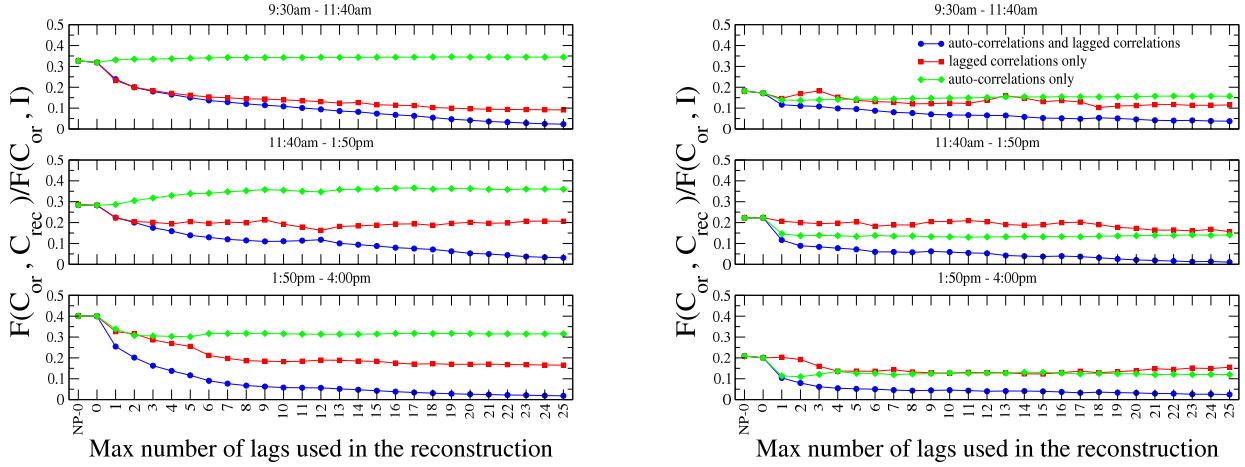
autocorrelations or lagged cross-correlations. In the 2011-13 time period the situation is different. If one uses both autocorrelations and lagged cross-correlations to reconstruct the correlation matrix, then all the reconstructed matrices are positive definite for any lags considered in the reconstruction. 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.

5.4. Intraday correlation patterns and the Epps effect

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. The Epps effect itself can be seen as an average effect. Indeed, as figure 7 shows, there are time periods of the day, especially at the beginning of the trading activity, where a *reverse* Epps effect is observed. This so called reverse Epps effect is associated to the observation that the average correlation calculated using 5-minute returns is *larger* than the one observed for average correlation calculated using 65-minute returns. However, the usual Epps effect is magnified at the end of the trading activity, where many significant lagged-cross correlations are observed. The bottom line is that the Epps effect also shows an intraday dynamics that depends on the dynamics of factors, mostly lagged cross-correlations and autocorrelations, which compete to form synchronous correlations of returns in a given intraday time window.

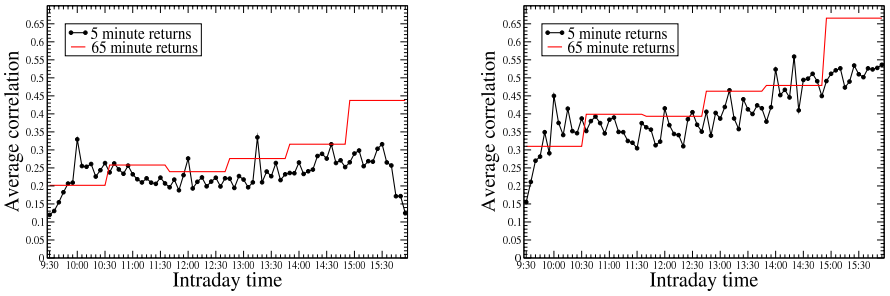
5.5. Partial lagged correlation validation.

The reconstruction analysis presented above 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.



(a) 2001-2003 (b) 2011-2013

Figure 6: Normalized Frobenius distance between the 130 minute return correlation matrix for 80 of the 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.



(a) 2001-2003 (b) 2011-2013

Figure 7: Intraday dynamics of average synchronous correlation at 5 minute and 65 minute time horizons, for the 2001-2003 dataset (a) and the 2011-2013 dataset (b). At the beginning of the trading activity, it is possible to observe a *reverse* Epps effect. This so called reverse Epps effect is associated to the observation that the average correlation calculated using 5-minute returns is *larger* than the one observed for average correlation calculated using 65-minute returns. However, the usual Epps effect is magnified at the end of the trading activity, where many significant lagged-cross correlations are observed.

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 $\rho(x(t), y(t))$.

Following the methods described in section 4.2, 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 FDR correction for multiple com-

parisons, we validate 2,498 positive links using the partial correlation matrix, as calculated using equation 4, and 1,688 positive links, as calculated using equation 5. We validate no links of negative correlation. Using the original lagged correlation matrix, we validate 619 positive links and no 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 (4) and (5) share an intersection of 614 and 617 links, respectively, with the original network. The probability of randomly sampling these intersections x from the $L = 80 \times 79 = 6320$ total possible lagged cross-correlation links in $n = 619$ "draws" (links in the original network) is given by the hypergeometric distribution:

$$P(x|n,k,L) = \frac{\binom{k}{x}\binom{L-k}{n-x}}{\binom{L}{n}}, \tag{9}$$

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:

$$p = P(j > x|n,k,L) = 1 - P(j < x|n,k,L) = 1 - \sum_{j=0}^{x} P(j|n,k,L). \tag{10}$$

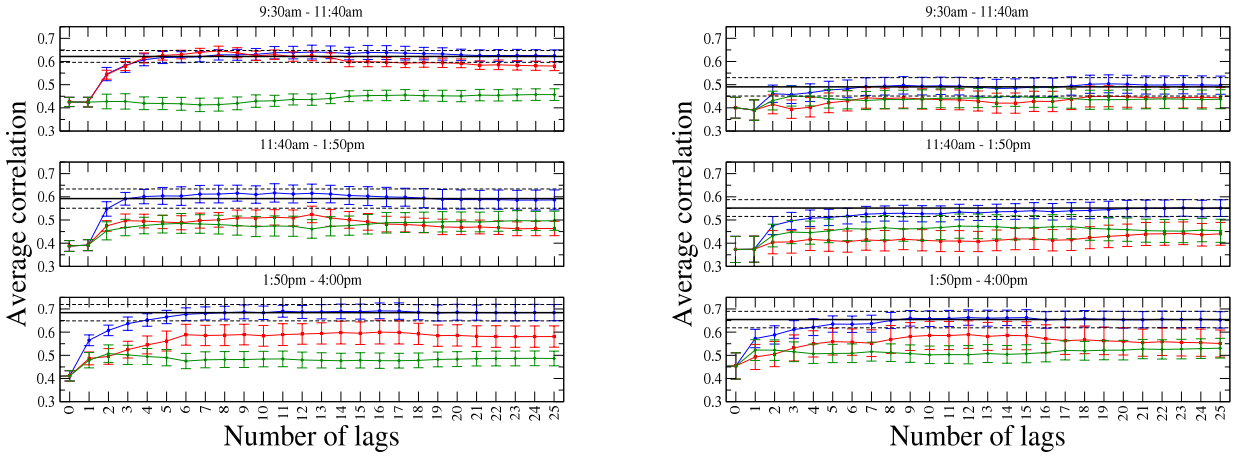
This number is vanishingly small for the numbers of links k validated in each partial correlation network, and the intersections x between the directed links obtaained in each one of the FDR networks from partial correlations and the directed links validated in the original lagged correlation network. 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 3,196 positive links using the partial correlation matrix (4), and 3,257 positive links using the matrix (5). We validate no links of negative correlation. Using the original lagged correlation matrix, we validate 4,875 positive links and no negative links. We note that the partial lagged correlation networks using the matrices (4) and (5) share an intersection of 3,080 and 3,055 links, respectively, with the original network. Again, we may associate a p-value to these intersections using the hypergeometric distribution, which is vanishingly small for both networks.

5.6. Case study: the banking sector

To better understand the specific contribution of autocorrelations and lagged cross-correlation to the Epps effect, we consider a subset of companies that belong to the same sector of activity — the banking sector. We chose this subset of stocks, since they exhibit a strong average pairwise correlation, and share several similar characteristics. These include Bank of America (BAC), Bank of NY Mellon (BK), National City Corp (NCC), Suntrust Bank Inc. (STI), PNC Financial Services Group (PNC), and Mellon Financial Corp (MEL) for the 2001-03 data; and Wells Fargo Corp. (WFC), JPMorgan Chase & Co. (JPM), Bank of America (BAC), Citigroup Inc. (C), Royal Bank of Canada (RY), Lloyds Banking Group PLC-ADR (LYG), MITSUBISHI UFJ FINL-SPON ADR (MTU), Toronto-Dominion Bank (TD), Goldman Sachs (GS), Bank of Nova Scotia (BNS), US Bankcorp (USB), UBS Group (UBS), and Banco Bilbao VIZCAYA-SP ADR (BBVA) for the 2011-13 data. Fig. 8 shows that, in order to properly reconstruct the average correlation coefficient between stock returns for the Banking sector, at 130 minute time horizon starting from returns calculated at 5 minute time horizons, both lagged cross-correlation and autocorrelations should be used. However, the figure also shows that the interplay between lagged cross-correlation and autocorrelations is different in the two time periods. Specifically, lagged cross correlations are more relevant to the Epps effect than autocorrelations in the 2001-2003 time period, while the opposite is observed in the 2011-2013 time period.

According to Eq. (3) and Eq. (B.10), the Epps effect, that is, the empirical evidence that synchronous correlations between the returns of stocks belonging to the same economic sector of activity increase as the time horizon at which returns are calculated increases, is favored by the presence of positive lagged cross-correlations (numerator of Eq. (3) and Eq. (B.10) and negative autocorrelations (denominator of Eq. (3) and Eq. (B.10) at shorter time horizons. Such a difference can be further understood with the results presented in Table 3, which presents the summary statistics of average autocorrelation and cross correlation in the baking sector, for the 2001-03 and 2011-13 data. The trading day is divided into three equal 130 minute parts, representing morning, mid-day, and afternoon. For each time interval, we present the average auto-correlation, and average lagged-correlation, for a lag of one, or two, $\Delta t = 15$ minutes, as well as the absolute value of the ratio between the two. We observe that the average autocorrelation is always negative in the 2011-2013 dataset for the first two lags (the most influential



(a) 2001-2003

(b) 2011-2013

Figure 8: Epps effect in the banking sector. Left panels: 2001-2003 (Company tickers included in the sector Banks: BAC, BK, NCC, STI, PNC, MEL) Right panels: 2011-2013 (Company tickers included in the sector Banks: WFC, JPM, BAC, C, RY, LYG, MTU, TD, GS, BNS, USB, UBS, BBVA)

ones) and its absolute value is larger than the one observed in the 2001-2003 dataset. At the same time, average lagged cross correlations in the 2011-2013 time period are smaller (in one case even negative) than average lagged cross correlations in the 2001-2003 time period.

6. Summary

The methodological framework presented here provides a validation of lead-lag relationships in financial markets, a picture of their intra-day dynamics. Moreover, this framework quantifies the impact of short term lead-lag relationships on longer term synchronous correlations among equity returns throughout different parts of a trading day. First, we validate the existence of such relationships using empirical data from two different time periods. Second, the validated lead-lag relationships provide new insights into the dynamics of equity markets, and provide new understandings into such phenomena as the Epps effect. Third, the proposed methodology provides the means to understand the interplay between different types of correlations, on different time scales. As such, this provides the means to both deconstruct cross sectional correlations into lagged and auto correlations, and also predict the correlation coefficients for different time scales.

Comparing the time periods 2001–2003 and 2011–2013, the synchronous correlations among these large market capitalization stocks have grown considerably, whereas the number of validated lagged-correlation relationships has decreased. We relate these two behaviors to an increase in the risks of financial spillovers and an increase in the informational efficiency of the market, respectively. Furthermore, our different analyses show a change in the role of auto-correlation in market dynamics, which is increasing. This is possibly related to the growing use of algorithmic and high frequency trading (HFT), in the U.S. market, between the two investigated time periods. The market share of HFT in terms of trading volume is estimated to have grown from under 10% in the early 2000's to approximately 60% in the early 2010's. Moreover, the observed change in correlation patterns between the two periods could be explained by the evolution of the Exchange Traded Fund (ETF) market. Trading in ETFs has been significantly growing, from their inception at the early 2000's, to approximately 20% at the early 2010's. The interplay between the ETF and its underlying assets could lead to the observed end-of-day increase in the synchronous and lagged correlations, in the 2011-2013 period. Another possibility, not alternative to the previous one, is that lagged

Table 3: Summary statistics of average autocorrelation and cross correlation in the baking sector, for the 2001-03 and 2011-13 data. The trading day is divided into three equal 130 minute parts, representing morning, mid-day, and afternoon. For each time interval, we present the average auto-correlation, and average lagged-correlation, for a lag of one, or two, $\Delta t = 15$ minutes, as well as the absolute value of the ratio between the two.

Time window		9:30am - 11:40am		11:40am - 1:50pm		1:50pm - 4:00pm	
Time period		2001-2003	2011-2013	2001-2003	2011-2013	2001-2003	2011-2013
$\langle \rho_{i,i} \rangle$	Lag 1	-0.0184	-0.0593	-0.0690	-0.0835	-0.0709	-0.0790
	Lag 2	0.0024	-0.0178	-0.0159	-0.0143	-0.0258	-0.0063
$\langle \rho_{i,j} \rangle$	Lag 1	0.0515	0.0113	0.0411	0.0155	0.0384	0.0178
	Lag 2	0.0221	-0.0079	0.0111	0.0004	0.0044	0.0004
$\left\lvert \frac{\langle \rho_{i,i} \rangle}{\langle \rho_{i,j} \rangle} \right\rvert$	Lag 1	0.4	5.3	1.7	5.4	1.8	4.4
	Lag 2	0.1	2.3	1.4	35.8	5.9	15.8

correlations might depend on company exposures to other stocks, so that, a change on the value of the latter might indirectly influence the market value of the former.

The decomposition and reconstruction procedures of synchronous correlations applied to both datasets, demonstrates that (1) the Epps effect is due to the presence of autocorrelation and lagged cross-correlation of stock returns, and (2) the interplay between autocorrelations and lagged cross-correlations has changed over time, potentially due to an increased presence of algorithmic trading and pair-trading strategies in the 2011-2013 time period. The proposed methodological framework thus provides the means to test the effect of market developments, such as HFT or ETF trading.

In summary, we introduce the statistically validated methodological framework for validating lead-lag relationships, and are able to empirically identify and validate such relationships. This sheds new light into the underlying dynamics of the U.S. equity 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 and price formation, risk management, and monitoring of financial spillovers.

Refrences

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

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 financial markets. Journal of empirical finance 4 (2), 115–158.

Andersen, T. G., Bollerslev, T., Diebold, F. X., Ebens, H., 2001. The distribution of realized stock return volatility. Journal of financial economics 61 (1), 43–76.

Andersen, T. G., Bollerslev, T., Diebold, F. X., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71 (2), 579–625.

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

Bandi, F. M., Russell, J. R., 2008. Microstructure noise, realized variance, and optimal sampling. The Review of Economic Studies 75 (2), 339–369.

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.

Bisias, D., Flood, M., Lo, A. W., Valavanis, S., 2012. A survey of systemic risk analytics. Annu. Rev. Financ. Econ. 4 (1), 255–296.

Brownlees, C., Engle, R. F., 2016. Srisk: A conditional capital shortfall measure of systemic risk. Review of Financial Studies, hhw060.

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

Christensen, K., Kinnebrock, S., Podolskij, M., 2010. Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. Journal of Econometrics 159 (1), 116–133.

Corsetti, G., Pericoli, M., Sbracia, M., 2005. 'some contagion, some interdependence': More pitfalls in tests of financial contagion. Journal of International Money and Finance 24 (8), 1177–1199.

Curme, C., Tumminello, M., Mantegna, R. N., Stanley, H. E., Kenett, D. Y., 2015. Emergence of statistically validated financial intraday lead-lag relationships. Quantitative Finance 15 (8), 1375–1386.

De Jong, F., Nijman, T., 1997. High frequency analysis of lead-lag relationships between financial markets. Journal of Empirical Finance 4 (2-3), 259–277.

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

Engle, R. F., Granger, C. W., 1987. Co-integration and error correction: representation, estimation, and testing. Econometrica: journal of the Econometric Society, 251–276.

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

Fleming, J., Kirby, C., Ostdiek, B., 2003. The economic value of volatility timing using "realized" volatility. Journal of Financial Economics 67 (3), 473–509.

Forbes, K. J., Rigobon, R., 2002. No contagion, only interdependence: measuring stock market comovements. The journal of Finance 57 (5), 2223–2261.

Gerety, M. S., Mulherin, J. H., 1992. Trading halts and market activity: An analysis of volume at the open and the close. The Journal of Finance 47 (5), 1765–1784.

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.

Hansen, P. R., Lunde, A., 2005. A realized variance for the whole day based on intermittent high-frequency data. Journal of Financial Econometrics 3 (4), 525–554.

Hasbrouck, J., Seppi, D. J., 2001. Common factors in prices, order flows, and liquidity. Journal of financial Economics 59 (3), 383–411.

Hayashi, T., Yoshida, N., et al., 2005. On covariance estimation of non-synchronously observed diffusion processes. Bernoulli 11 (2), 359–379.

Kenett, D. Y., Huang, X., Vodenska, I., Havlin, S., Stanley, H. E., 2015. Partial correlation analysis: Applications for financial markets. Quantitative Finance 15 (4), 569–578.

Kenett, D. Y., Tumminello, M., Madi, A., Gur-Gershgoren, G., Mantegna, R. N., 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., Page, S., Rigobon, R., 2011. Principal components as a measure of systemic risk. The Journal of Portfolio Management 37 (4), 112–126.

Laloux, L., Cizeau, P., Bouchaud, J., Potters, M., 1999. Noise dressing of financial correlation matrices. Physical Review Letters 83 (7), 1467–1470.

Lin, W.-L., Engle, R. F., Ito, T., 1994. Do bulls and bears move across borders? international transmission of stock returns and volatility. Review of financial studies 7 (3), 507–538.

Liu, Q., 2009. On portfolio optimization: How and when do we benefit from high-frequency data? Journal of Applied Econometrics 24 (4), 560–582.

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

Madhavan, A., 2000. Market microstructure: A survey. Journal of financial markets 3 (3), 205–258.

Mantegna, R. N., 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.

McInish, T. H., Shoesmith, G. L., Wood, R. A., et al., 1995. Cointegration, error correction, and price discovery on informationally linked security markets. Journal of financial and quantitative analysis 30 (04), 563–579.

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.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L., Stanley, H., 1999. Universal and nonuniversal properties of cross correlations in financial time series. Physical Review Letters 83 (7), 1471–1474.

Pollet, J. M., Wilson, M., 2010. Average correlation and stock market returns. Journal of Financial Economics 96 (3), 364–380.

Pooter, M. d., Martens, M., Dijk, D. v., 2008. Predicting the daily covariance matrix for s&p 100 stocks using intraday data?but which frequency to use? Econometric Reviews 27 (1-3), 199–229.

Renò, R., 2003. A closer look at the epps effect. International Journal of theoretical and applied finance 6 (01), 87–102.

Tóth, B., Kertész, J., 2006. Increasing market efficiency: Evolution of cross-correlations of stock returns. Physica A: Statistical Mechanics and its Applications 360 (2), 505–515.

Toth, B., Kertesz, J., 2009. Accurate estimator of correlations between asynchronous signals. Physica A: Statistical Mechanics and its Applications 388 (8), 1696–1705.

Tóth, B., Kertész, J., 2009. The epps effect revisited. Quantitative Finance 9 (7), 793–802.

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

Appendix A. List of securities

Table A1: List of securities, 2001-2003: company ticker, company name, and company industry

TICKER	NAME	INDUSTRY
GE	GENERAL ELECTRIC CO	Miscellaneous Manufacturing
PFE	PFIZER INC	Pharmaceuticals
WMT	WAL-MART STORES INC	Retail
AIG	AMERICAN INTERNATIONAL GROUP	Insurance
IBM	INTL BUSINESS MACHINES CORP	Computers
KO	COCA-COLA CO/THE	Beverages
JNJ	JOHNSON & JOHNSON	Pharmaceuticals
PG	PROCTER & GAMBLE CO/THE	Cosmetics/Personal Care
MRK	MERCK & CO. INC.	Pharmaceuticals
BAC	BANK OF AMERICA CORP	Banks
WFC	WELLS FARGO & CO	Banks
SBC	SBC COMMUNICATIONS INC	Telecommunications
FNM	FANNIE MAE	Diversified Finan Serv
HD	HOME DEPOT INC	Retail
PEP	PEPSICO INC	Beverages
LLY	ELI LILLY & CO	Pharmaceuticals
BUD	ANHEUSER-BUSCH INBEV-SPN ADR	Beverages
ABT	ABBOTT LABORATORIES	Healthcare-Products
BMY	BRISTOL-MYERS SQUIBB CO	Pharmaceuticals
AXP	AMERICAN EXPRESS CO	Diversified Finan Serv
MER	MERRILL LYNCH & CO INC	Diversified Finan Serv
MDT	MEDTRONIC PLC	Healthcare-Products
UTX	UNITED TECHNOLOGIES CORP	Aerospace/Defense
BLS	BELLSOUTH LLC	Telecommunications
ONE	HIGHER ONE HOLDINGS INC	Diversified Finan Serv
TYC	TYCO INTERNATIONAL PLC	Building Materials
TXN	TEXAS INSTRUMENTS INC	Semiconductors
G	GENPACT LTD	Computers
DD	DU PONT (E.I.) DE NEMOURS	Chemicals
DIS	WALT DISNEY CO/THE	Media
LOW	LOWE'S COS INC	Retail
BA	BOEING CO/THE	Aerospace/Defense
FRE	FREDDIE MAC	Diversified Finan Serv
MCD	MCDONALD'S CORP	Retail
DOW	DOW CHEMICAL CO/THE	Chemicals
GM	GENERAL MOTORS CO	Auto Manufacturers
ALL	ALLSTATE CORP	Insurance
WAG	WALGREEN CO	Retail
FDC	FIRST DATA CORP- CLASS A	Software
CL	COLGATE-PALMOLIVE CO	Cosmetics/Personal Care
SLB	SCHLUMBERGER LTD	Oil&Gas Services
SGP	SCHERING-PLOUGH CORP/PRE-MER	Pharmaceuticals
BK	BANK OF NEW YORK MELLON CORP	Banks
CAT	CATERPILLAR INC	Machinery-Constr&Mining
KMB	KIMBERLY-CLARK CORP	Household Products/Wares
MOT	MOTOROLA INC	Telecommunications
KRB	MBNA CORP	Diversified Finan Serv
EMR	EMERSON ELECTRIC CO	Electrical Compo&Equip
BSX	BOSTON SCIENTIFIC CORP	Healthcare-Products
EMC	EMC CORP/MA	Computers
GCI	GANNETT CO INC	Media
CCU	CIA CERVECERIAS UNI-SPON ADR	Beverages
CAH	CARDINAL HEALTH INC	Pharmaceuticals
SYY	SYSCO CORP	Food
MMC	MARSH & MCLENNAN COS	Insurance
RD	ROYAL DUTCH PETRO-NY SHARES	Oil&Gas
ITW	ILLINOIS TOOL WORKS	Miscellaneous Manufacturing
AVP	AVON PRODUCTS INC	Cosmetics/Personal Care
AFL	AFLAC INC	Insurance
GIS	GENERAL MILLS INC	Food
GPS	GAP INC/THE	Retail
NCC	NATIONAL CITY CORP	Banks
SO	SOUTHERN CO/THE	Electric
GD	GENERAL DYNAMICS CORP	Aerospace/Defense
STI	SUNTRUST BANKS INC	Banks
IP	INTERNATIONAL PAPER CO	Forest Products&Paper
LEH	LEHMAN BROTHERS HOLDINGS INC	Diversified Finan Serv
BAX	BAXTER INTERNATIONAL INC	Healthcare-Products
S	SPRINT CORP	Telecommunications
PNC	PNC FINANCIAL SERVICES GROUP	Banks
UNP	UNION PACIFIC CORP	Transportation
MEL	MELLON FINANCIAL CORP	Banks
GDT	GUIDANT LLC	Healthcare-Products
PPG	PPG INDUSTRIES INC	Chemicals
DUK	DUKE ENERGY CORP	Electric
NEM	NEWMONT MINING CORP	Mining
DE	DEERE & CO	Machinery-Diversified
OMC	OMNICOM GROUP	Advertising
CA	CA INC	Software
SLE	HILLSHIRE BRANDS CO/THE	Food

Table A2: List of securities, 2011-2013: company ticker, company name, and company industry

TICKER	NAME	INDUSTRY
XOM	EXXON MOBIL CORP	Oil&Gas
GE	GENERAL ELECTRIC CO	Miscellaneous Manufacturing
BRK-B	BERKSHIRE HATHAWAY INC-CL B	Insurance
JNJ	JOHNSON & JOHNSON	Pharmaceuticals
WMT	WAL-MART STORES INC	Retail
CVX	CHEVRON CORP	Oil&Gas
WFC	WELLS FARGO & CO	Banks
RDS-A	ROYAL DUTCH SHELL-SPON ADR-A	Oil&Gas
PG	PROCTER & GAMBLE CO/THE	Cosmetics/Personal Care
JPM	JPMORGAN CHASE & CO	Banks
TM	TOYOTA MOTOR CORP -SPON ADR	Auto Manufacturers
CHL	CHINA MOBILE LTD-SPON ADR	Telecommunications
IBM	INTL BUSINESS MACHINES CORP	Computers
PFE	PFIZER INC	Pharmaceuticals
NVS	NOVARTIS AG-SPONSORED ADR	Pharmaceuticals
T	AT&T INC	Telecommunications
BHP	BHP BILLITON LTD-SPON ADR	Mining
KO	COCA-COLA CO/THE	Beverages
BUD	ANHEUSER-BUSCH INBEV-SPN ADR	Beverages
BAC	BANK OF AMERICA CORP	Banks
C	CITIGROUP INC	Banks
BP	BP PLC-SPONS ADR	Oil&Gas
MRK	MERCK & CO. INC.	Pharmaceuticals
VZ	VERIZON COMMUNICATIONS INC	Telecommunications
SNY	SANOFI-ADR	Pharmaceuticals
V	VISA INC-CLASS A SHARES	Diversified Finan Serv
TOT	TOTAL SA-SPON ADR	Oil&Gas
PM	PHILIP MORRIS INTERNATIONAL	Agriculture
DIS	WALT DISNEY CO/THE	Media
GSK	GLAXOSMITHKLINE PLC-SPON ADR	Pharmaceuticals
PEP	PEPSICO INC	Beverages
UL	UNILEVER PLC-SPONSORED ADR	Cosmetics/Personal Care
SLB	SCHLUMBERGER LTD	Oil&Gas Services
SI	SIEMENS AG-SPONS ADR	Miscellaneous Manufactur
HD	HOME DEPOT INC	Retail
UTX	UNITED TECHNOLOGIES CORP	Aerospace/Defense
RIO	RIO TINTO PLC-SPON ADR	Mining
SAP	SAP SE-SPONSORED ADR	Software
BA	BOEING CO/THE	Aerospace/Defense
MA	MASTERCARD INC - A	Diversified Finan Serv
UPS	UNITED PARCEL SERVICE-CL B	Transportation
RY	ROYAL BANK OF CANADA	Banks
AXP	AMERICAN EXPRESS CO	Diversified Finan Serv
MCD	MCDONALD'S CORP	Retail
LYG	LLOYDS BANKING GROUP PLC-ADR	Banks
MTU	MITSUBISHI UFJ FINL-SPON ADR	Banks
MMM	3M CO	Miscellaneous Manufactur
DEO	DIAGEO PLC-SPONSORED ADR	Beverages
TSM	TAIWAN SEMICONDUCTOR-SP ADR	Semiconductors
PBR	PETROLEO BRASILEIRO-SPON ADR	Oil&Gas
E	ENI SPA-SPONSORED ADR	Oil&Gas
BMY	BRISTOL-MYERS SQUIBB CO	Pharmaceuticals
TD	TORONTO-DOMINION BANK	Banks
COP	CONOCOPHILLIPS	Oil&Gas
OXY	OCCIDENTAL PETROLEUM CORP	Oil&Gas
CVS	CVS HEALTH CORP	Retail
AMX	AMERICA MOVIL-SPN ADR CL L	Telecommunications
GS	GOLDMAN SACHS GROUP INC	Banks
EC	ECOPETROL SA-SPONSORED ADR	Oil&Gas
VALE	VALE SA-SP ADR	Iron/Steel
STO	STATOIL ASA-SPON ADR	Oil&Gas
UNP	UNION PACIFIC CORP	Transportation
MO	ALTRIA GROUP INC	Agriculture
BNS	BANK OF NOVA SCOTIA	Banks
AIG	AMERICAN INTERNATIONAL GROUP	Insurance
HMC	HONDA MOTOR CO LTD-SPONS ADR	Auto Manufacturers
UNH	UNITEDHEALTH GROUP INC	Healthcare-Services
AZN	ASTRAZENECA PLC-SPONS ADR	Pharmaceuticals
TEF	TELEFONICA SA-SPON ADR	Telecommunications
USB	US BANCORP	Banks
HON	HONEYWELL INTERNATIONAL INC	Electronics
UBS	UBS GROUP AG-REG	Banks
BBVA	BANCO BILBAO VIZCAYA-SP ADR	Banks
NKE	NIKE INC -CL B	Apparel
LVS	LAS VEGAS SANDS CORP	Lodging
TWX	TIME WARNER INC	Media
F	FORD MOTOR CO	Auto Manufacturers
GM	GENERAL MOTORS CO	Auto Manufacturers
MON	MONSANTO CO	Chemicals
ABB	ABB LTD-SPON ADR	Machinery-Constr&Mining

Appendix B. 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 $p\Delta t = 195$ min of a trading day. We are interested in the correlation coefficient between the time series

$$\begin{aligned} \{x\} &= \{x_1, x_2, ..., x_T\} \text{ and} \\ \{y\} &= \{y_1, y_2, ..., y_T\}, \end{aligned} \tag{B.1}$$

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 Δt min, e.g., if $p\Delta t = 195$ min one can set $p = 13$ and $\Delta t = 15$ min:

$$\begin{aligned} \{x\} &= \left\{ \sum_{j=1}^{p} x_1(j), \sum_{j=1}^{p} x_2(j), ..., \sum_{j=1}^{p} x_T(j) \right\}; \\ \{y\} &= \left\{ \sum_{j=1}^{p} y_1(j), \sum_{j=1}^{p} y_2(j), ..., \sum_{j=1}^{p} y_T(j) \right\}; \end{aligned} \tag{B.2}$$

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

$$< x(j) >= \frac{1}{T} \sum_{i=1}^{T} x_i(j) =< y(j) >= \frac{1}{T} \sum_{i=1}^{T} y_i(j) = 0, \quad \forall j = 1, ..., p. \tag{B.3}$$

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

$$< x >= 0 \text{ and } \quad < y >= 0 \tag{B.4}$$

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:

$$\begin{aligned}\sigma_x^2 =& < x^2 >= \frac{1}{T}\sum_{i=1}^{T}\left[\sum_{j=1}^{p} x_i(j)\right]^2 = \\ &= \frac{1}{T}\sum_{i=1}^{T}\left[\sum_{j=1}^{p} x_i(j)^2 + 2\sum_{j=1}^{p-1} x_i(j)\, x_i(j+1) + 2\sum_{j=1}^{p-2} x_i(j)\, x_i(j+2) + ... + 2\, x_i(1)\, x_i(p)\right] = \\ &= \sum_{j=1}^{p}\sigma_x(j)^2 + 2\sum_{j=1}^{p-1}\sigma_x(j)\,\sigma_x(j+1)\rho_{x_j,x_{j+1}} + 2\sum_{j=1}^{p-2}\sigma_x(j)\,\sigma_x(j+2)\rho_{x_j,x_{j+2}} \\ &+ ... + 2\,\sigma_x(1)\,\sigma_x(p)\rho_{x_1,x_p},\end{aligned} \tag{B.5}$$

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:

$$\sigma_x(j) = k_x \cdot f(j);\;\; \sigma_y(j) = k_y \cdot f(j),\;\; \forall j = 1, ..., p \tag{B.6}$$

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:

$$\begin{aligned}\sigma_x^2 =& k_x^2[\sum_{j=1}^{p} f(j)^2 + 2\sum_{j=1}^{p-1} f(j)\, f(j+1)\rho_{x_j,x_{j+1}} + 2\sum_{j=1}^{p-2} f(j)\, f(j+2)\rho_{x_j,x_{j+2}} + \\ &... + 2\, f(1)\, f(p)\rho_{x_1,x_p}],\end{aligned} \tag{B.7}$$

where Eq. B.6 has been used to describe the intraday pattern of volatility. Similarly, we obtain the variance of y:

$$\sigma_y^2 = k_y^2\left[\sum_{j=1}^{p} f(j)^2 + 2\sum_{j=1}^{p-1} f(j)\, f(j+1)\rho_{y_j,y_{j+1}} + 2\sum_{j=1}^{p-2} f(j)\, f(j+2)\rho_{y_j,y_{j+2}} + ... + 2\, f(1)\, f(p)\rho_{y_1,y_p}\right]. \tag{B.8}$$

The covariance of x and y is then:

$$\begin{aligned} cov(x,y) &=< x\, y > \\ &= \frac{1}{T}\sum_{i=1}^{T}\left[\left(\sum_{j=1}^{p} x_i(j)\right)\cdot\left(\sum_{l=1}^{p} y_i(l)\right)\right] = \\ &= k_x\, k_y\left[\sum_{j=1}^{p} f(j)^2\rho_{x_j,y_j}\right] + \\ &+ k_x\, k_y\left[\sum_{j=1}^{p-1} f(j)f(j+1)(\rho_{x_j,y_{j+1}}+\rho_{x_{j+1},y_j})\right] \quad + \\ &+ ... + k_x\, k_y[f(1)f(p)(\rho_{x_1,y_p}+\rho_{x_p,y_1})]. \end{aligned} \tag{B.9}$$

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

$$\rho_{x,y} = \frac{\left[\sum_{j=1}^{p} f(j)^2\rho_{x_j,y_j}\right] + \left[\sum_{j=1}^{p-1} f(j)f(j+1)(\rho_{x_j,y_{j+1}}+\rho_{x_{j+1},y_j})\right] + ... + f(1)f(p)(\rho_{x_1,y_p}+\rho_{x_p,y_1})}{\sqrt{\left(\sum_{j=1}^{p} f(j)^2 + 2\sum_{j=1}^{p-1} f(j)\,f(j+1)\rho_{x_j,x_{j+1}} + ...\right)\left(\sum_{j=1}^{p} f(j)^2 + 2\sum_{j=1}^{p-1} f(j)\,f(j+1)\rho_{y_j,y_{j+1}} + ...\right)}} \tag{B.10}$$

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:

$$\rho_{x,y} = \frac{\sum_{j=1}^{p} f(j)^2\rho_{x_j,y_j}}{\sum_{j=1}^{p} f(j)^2} + \frac{\sum_{j=1}^{p-1} f(j)f(j+1)(\rho_{x_j,y_{j+1}}+\rho_{x_{j+1},y_j})}{\sum_{j=1}^{p} f(j)^2}. \tag{B.11}$$

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.