Boston University

OpenBU **http://open.bu.edu**

Theses & Dissertations Boston University Theses & Dissertations

2015

Lagged correlation networks

https://hdl.handle.net/2144/15992

Boston University

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

LAGGED CORRELATION NETWORKS

by

CHESTER CURME
B.A., Middlebury College, 2011

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
2015

Approved by

First Reader

H. Eugene Stanley, Ph.D.
William Fairfield Warren Distinguished Professor
Professor of Physics

Second Reader

Irena Vodenska, Ph.D., C.F.A.
Professor of Administrative Sciences

For Ying.

Acknowledgments

First and foremost I'd like to thank my advisor, Gene Stanley, for his countless words of wisdom, for welcoming me into his academic family, for his friendship, and of course for his financial support. I am extraordinarily lucky to have been Gene's student, and to have worked with him to push the boundaries of interdisciplinary science.

I'd also like to thank my collaborators around the world: Prof. Tobias Preis, Prof. Suzy Moat, Prof. Michele Tumminello, Dr. Dror Kenett, Prof. Irena Vodenska, Prof. Rosario Mantegna, Prof. Boris Podobnik, Prof. Viktoria Dalko, Prof. Andrea Gabrielli, Prof. Doug Rosene, Prof. Luciano Costa, Antonio Majdandzic, Adam Avakian, Will Morrison, João Santos, Dario Corradini, and César Comin.

I'd like to thank my thesis committee, for their valuable feedback: Prof. Gene Stanley, Prof. Irena Vodenska, Prof. Robert Carey, Prof. Shyam Erramilli, and Prof. Mark Kon.

I'd like to thank my friends and intellectual compatriots: Antonio Majdandzic, Xin Yuan, Shuai Shao, Nagendra Panduranga, Adam Avakian, Nima Dehmami, Asher Mullokandov, Will Morrison, João Santos, and everyone else in Gene's group. Special thanks to Erik Lascaris for his friendship and technical assistance.

Finally I'd like to thank my parents, Ollie and Cynthia, my brothers, Eli and Harry, and my dear wife Ying.

LAGGED CORRELATION NETWORKS

(Order No.)

CHESTER CURME

Boston University Graduate School of Arts and Sciences, 2015

Major Professor: H. Eugene Stanley, William Fairfield Warren Distinguished Professor,

Professor of Physics

ABSTRACT

Technological advances have provided scientists with large high-dimensional datasets that describe the behaviors of complex systems: from the statistics of energy levels in complex quantum systems, to the time-dependent transcription of genes, to price fluctuations among assets in a financial market. In this environment, where it may be difficult to infer the joint distribution of the data, network science has flourished as a way to gain insight into the structure and organization of such systems by focusing on pairwise interactions.

This work focuses on a particular setting, in which a system is described by multivariate time series data. We consider time-lagged correlations among elements in this system, in such a way that the measured interactions among elements are asymmetric. Finally, we allow these interactions to be characteristically weak, so that statistical uncertainties may be important to consider when inferring the structure of the system. We introduce a methodology for constructing statistically validated networks to describe such a system, extend the methodology to accommodate interactions with a periodic component, and show how consideration of bipartite community structures in these networks can aid in the construction of robust statistical models.

An example of such a system is a financial market, in which high frequency returns data may be used to describe contagion, or the spreading of shocks in price among assets. These data provide the experimental testing ground for our methodology. We study NYSE data from both the present day and one decade ago, examine the time scales over which the validated lagged correlation networks exist, and relate differences in the topological properties of the networks to an increasing economic efficiency. We uncover daily periodicities in the validated interactions, and relate our findings to explanations of the Epps Effect, an empirical phenomenon of financial time series. We also study bipartite community structures in networks composed of market returns and news sentiment signals for 40 countries. We compare the degrees to which markets anticipate news, and news anticipate markets, and use the community structures to construct a recommender system for inputs to prediction models. Finally, we complement this work with novel investigations of the exogenous news items that may drive the financial system using topic models. This includes an analysis of how investors and the general public may interact with these news items using Internet search data, and how the diversity of stories in the news both responds to and influences market movements.

Contents

1 Introduction 1

2 Statistically-validated network methodology 6

2.1 Motivation 6

2.2 Methodology 8

2.2.1 Probability of a false positive link 10

2.3 Relevance to financial data 12

2.3.1 Statistically validated lagged correlation networks in financial markets 15

2.3.2 Synchronous correlation networks 23

2.4 Discussion 25

2.5 Comparison of the bootstrap method and an analytical one to calculate correlation p-values 26

2.6 Effect of lag and time series length on validated links for a fixed time horizon 28

3 Seasonalities and the Epps Effect **31**

3.1 Incorporating seasonalities 32

3.2 Application to equity returns 34

3.3 Methodology 36

3.4 Intraday seasonalities 39

3.5 Reconstructing the Epps Effect 42

3.6 Regression model 48

3.7 Discussion 50

3.8 Robustness Checks 51
  3.8.1 Visualization of 2011-2013 networks 51
  3.8.2 Contribution of high volatility period to lagged correlations 52
  3.8.3 Stability of reconstructed correlation matrices to noise 54
  3.8.4 Persistence of links 60
  3.8.5 Influence of autocorrelations on linear models 62
  3.8.6 Partial lagged correlation networks 63

4 Community structures in lagged correlation networks and their relevance to feature selection 66

4.1 Bipartite communities in directed networks 67
4.2 Relevance to financial data 69
4.3 Data and summary statistics 71
4.4 Synchronous Correlations 73
  4.4.1 Methodology 73
  4.4.2 Results 76
4.5 Lagged Correlations 77
  4.5.1 Methodology 77
  4.5.2 Results 79
4.6 Relation between the structure of the statistically-validated network and prediction model performance 82
4.7 Conclusions 89
4.8 Lagging procedure 91
4.9 Statistical validation of directed links 92
4.10 Tests with synthetic data 94
  4.10.1 Effect of signal-to-noise ratio 95
  4.10.2 Effect of network bipartivity 97

5 Using topic models to explain market movements **101**
5.1 Application to financial news data 102
5.1.1 Quantifying the diversity of financial news with LDA 104
5.1.2 Price changes of the FTSE drive changes in news diversity 111
5.1.3 Influence of individual topics 117
5.1.4 News diversity relates to same-day trading volume 119
5.1.5 Discussion 123
5.2 Application to Internet search data 125
5.2.1 Method 127
5.2.2 Results 129
5.2.3 Discussion 133

Appendices **139**

A Stemmed word distributions from LDA of *The Financial Times* **140**

Bibliography **144**

Curriculum Vitae **160**

List of Tables

2.1 Summary statistics of 2002-2003 and 2011-2012 datasets. 17

2.2 Threshold-correlation values and validated links according to a normal distribution of returns 27

2.3 Comparison between number of positive validated links according to the bootstrap method and a normal distribution of returns 28

4.1 Summary statistics for returns $\tilde{r}_{i,t}$ and de-trended news sentiment signals $\tilde{s}_{i,t}$ for the period January 8, 2002 to December 31, 2012. 74

4.1 *Continued:* Summary statistics for returns $\tilde{r}_{i,t}$ and de-trended news sentiment signals $\tilde{s}_{i,t}$ for the period January 8, 2002 to December 31, 2012. 75

4.2 Largest five components of the first three left- and right-singular vector pairs. Entries refer to market indices, unless otherwise specified as news. 83

5.1 In-sample model results with and without Topic 46 (“mortgage”, “loan”, “credit”, “debt”,...) 119

5.2 Regression results using search volume signals $x_{\text{Politics I}}$, x_{Business}, and $x_{\text{Politics II}}$. 132

5.3 Quantiles of test-statistics $|t|$ and R^2 using randomized search volume data. 132

5.4 Comparison of observed test statistics with those obtained from bootstrapping procedure. 133

A.1 LDA Topics from Chapter 5.1 141

A.1 *Continued:* LDA Topics from Chapter 5.1 142

A.1 *Continued:* LDA Topics from Chapter 5.1 143

List of Figures

1.1 Hierarchical organization of correlations among synchronous equity returns. 4

2.1 Cumulative distribution functions for returns of the top 100 largest market capitalization stocks in the NYSE from 2011-2012. 7

2.2 Schematic of lagged correlation calculation for a time horizon $h = 130$ minutes. 16

2.3 Distribution of lagged correlation coefficients for all $N = 100$ stocks at a time horizon $h = 15$ minutes. 18

2.4 Illustrations of Bonferroni networks constructed from statistically-validated lagged correlations for various time horizons h. 19

2.5 Plots of the number of positive and negative validated links for both Bonferroni and FDR lagged correlation networks. 20

2.6 Average magnitude (absolute value) of lagged correlation coefficients filtered in Bonferroni and FDR networks. 21

2.7 In- and out-degree distributions for FDR networks 22

2.8 Percentage occurrence of all 14 possible directed three-node motifs with more than one link in Bonferroni networks. 23

2.9 (a) Plot of mean synchronous correlation coefficients in both 2002-2003 and 2011-2012 data. (b) Histograms of correlation coefficients for returns sampled at a 15 minute time horizon. 24

2.10 Numbers of positive and negative validated links for both Bonferroni and FDR correlation networks for varying lag l. 29

2.11 Numbers of positive and negative validated links for both Bonferroni and FDR lagged correlation networks for time series segments of length $T = 1004$ at $h = 15$ minutes. 30

3.1 Power spectrum of $\langle C \rangle_t$. 34

3.2 Intra-day pattern of the average synchronous correlation between fifteen minute stock returns. 40

3.3 Intra-day pattern of the average lagged correlation between fifteen minute stock returns. 41

3.4 Visualization of the Bonferroni networks from periods in the beginning, middle, and end of the trading day in the period 2001-2003. 42

3.5 Frobenius distance between the 130 minute return correlation matrix and the corresponding reconstructed correlation matrix. 49

3.6 Difference in BICs between the models in equation (3.6) and the randomized models described in the text. 50

3.7 Visualization of the Bonferroni networks from periods in the beginning, middle, and end of the trading day in the period 2011-2013. 52

3.8 Contributions $\langle C \rangle_t$ of each day t in the 2011-2013 data to the mean lagged correlation measured for each intraday period. 53

3.9 Top row: number of validated positive links in the 2011-2013 data. Bottom row: number of validated negative links in the 2011-2013 data. 55

3.10 Normalized histograms of all $N^2 = 100^2 = 10,000$ lagged correlation coefficients for two intraday periods in 2011-2013. 56

3.11 Frobenius distance between the 130 minute return correlation matrix and the corresponding reconstructed correlation matrix. 57

3.12 Probability of observing at least one negative eigenvalue in each 130 min. correlation matrix after perturbing correlation matrices with a given level of noise. 59

3.13 Matrices of Jaccard Indices between sets of links corresponding to networks for all intraday periods at a time horizon $\Delta t = 15$ min. 61

3.14 Distributions of link persistence for all links in networks at a time horizon $\Delta t = 15$ min. 62

3.15 Difference in BICs between the linear models with inputs prescribed by the validated network and the randomized models described in the text. 63

4.1 (a) Plot of the Minimum Spanning Tree of the synchronous correlations. (b) Histogram of the betweenness centrality of financial markets and news sentiment separately. 77

4.2 Histograms of lagged correlation coefficients. 80

4.3 Plots of each subgraph of the statistically validated network. 81

4.4 (a) Display of the complete directed network. In (b), (c), and (d) we plot the sorted components of the first three pairs of left- and right-singular vectors. 84

4.5 ROC curves for the performance of the logistic regression models. 86

4.6 Pairwise differences in AUCs between models with inputs defined by the validated network and those using all possible inputs. 87

4.7 AUC for all markets against the magnitudes of the entry of the corresponding right singular vector. 88

4.8 AUC, as averaged among the top five nodes in $|V^1|$ (markets), $|V^2|$ (markets), and $|V^3|$ (news), for each additional model input. 89

4.9 Diagram of lagging procedure for measuring lagged correlations. 93

4.10 Pairwise differences in AUCs between logistic regression models with inputs given by case (ii), and inputs given by case (i), for varying α. 97

4.11 (a) Bipartite adjacency matrix used to simulate time series, with $p = 0$. (b) Adjacency matrix of the resulting statistically-validated network. 99

4.12 (a) Pairwise differences in AUCs between logistic regression models with inputs given by case (ii), and inputs given by case (i), for varying p. 100

5.1 The prominence of topics in the *Financial Times*. 107
5.2 Variation in topic lifetimes in the *Financial Times*. 109
5.3 Boxplots of the diversity H_t, aggregated by month. 110
5.4 Changes in topic diversity across time. 112
5.5 Time-series features of the differenced diversity ΔH_t. 114
5.6 Improvement in out-of-sample model errors in forecasts of diversity fluctuations ΔH_t using returns from the FTSE 100. 116
5.7 Identification of topics that correlate individually with previous-day market movements. 118
5.8 Improvement in out-of-sample model errors in forecasts of trade volumes v_t using changes in the diversity of news ΔH_t. 122
5.9 (Caption next page.) 136
5.9 (Previous page.) *Google Trends* based trading strategies for 55 different semantic topics. 137
5.10 Effect of changing time window on returns. 138

Chapter 1

Introduction

Recent decades have witnessed a large body of work dedicated to uncovering the organization of complex systems in the physical, biological, and social sciences. These systems are composed of a large number of components, the interactions among which typically induce large-scale collective structures or behavior. Technological developments have allowed researchers to produce multivariate datasets that quantify some aspects of the behaviors of individual members of these complex systems. A major challenge has been to extract, from these multivariate datasets, insights into regularities in the large-scale organization of the system.

A common approach is to study measures of similarity among pairs of elements, often quantified using the pair cross-correlation. Examples range from the study of energy correlations in quantum spectra, to the time-dependent transcription of genes, to price fluctuations among assets in a financial market. These correlation measurements are assembled into a matrix, and the task is then to investigate collective structures represented in this matrix. Many tools have been evaluated for this purpose, such as principal components analysis, random matrix theory, hierarchical clustering, factor analysis, and graph theory. Each tool may uncover a different aspect of the system's organization, and the success of each tool rests on the form of the experimental data, such as the degree of statistical uncertainty in the measured correlations, and the underlying organization of the system.

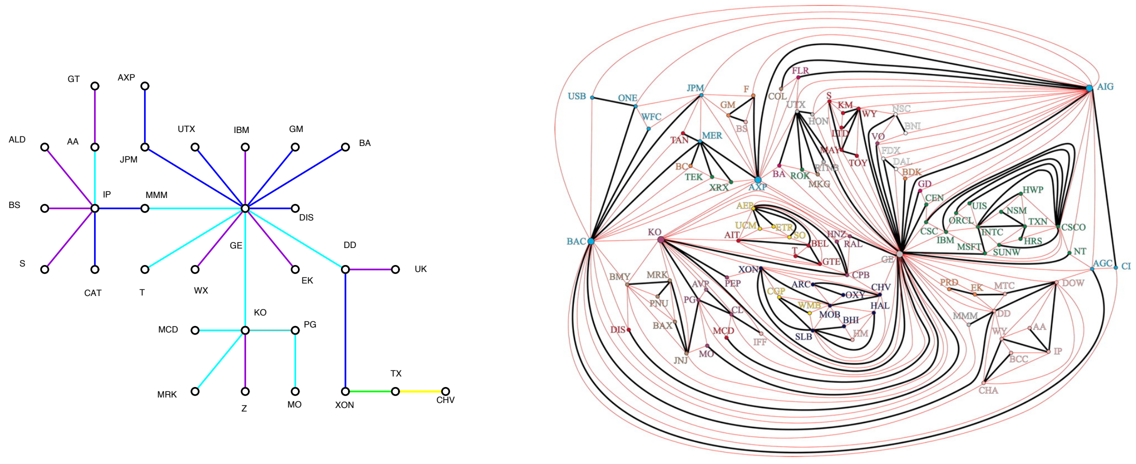
This thesis focuses on a particular experimental setting, in which the system is described by multivariate time-series data. We consider time-lagged correlations among elements in this system, in such a way that the resulting measure of pair-similarity is asymmetric. Finally, we allow for characteristically low values of the correlation coefficient, so that statistical uncertainties may be important to consider when deducing the structure of the system.

An example of such a system is the stock market, in which high frequency returns data may be used to describe financial contagion, or the spreading of shocks in price among assets. Indeed, such data will provide the experimental testing ground for much of the work in this thesis. As demonstrated by the events of recent past, the functioning of this system, for better or worse, is tied to the everyday life of the world's population. A large portion of financial market activity is reflected in stock market movements, which are driven by the trading decisions of many investors. The motivating forces behind these decisions, whether they are exogenous news items or the endogenous influences of other traders, therefore warrant scientific attention.

Networks formed from synchronous correlations among financial assets have been studied for some time. The earliest studies of correlation-based networks of equity returns involve hierarchical clustering procedures. In 1999 Mantegna [52] projected equities to a common T-dimensional space using time series of length T: each time series i was treated as a T-vector $\vec{r}_i$, which can be normalized to $\tilde{r}_i \equiv (\vec{r}_i - \langle\vec{r}_i\rangle)/(\sqrt{T-1}\sigma_i)$, so that $\tilde{r}_i$ is a unit vector (here, $\langle\vec{r}_i\rangle$ indicates the vector's mean and σ_i indicates its sample standard deviation). The correlation between two such time series i and j is then the dot-product $\rho_{i,j} = \vec{r}_i \cdot \vec{r}_j$, and the Euclidean distance between two vectors is $d_{i,j} = \sqrt{2(1-\rho_{i,j})}$. The hierarchical organization of the system can then be studied using the Minimal Spanning Tree, which is the graph connecting all N nodes with $N-1$ edges such that there are no loops and such that the sum of the distances $d_{i,j}$ between all vertices joined by an edge is minimized. This graph is commonly constructed using Kruskal's algorithm [86], which relies only on a ranking of the measured correlation coefficients. Mantegna found that this structure revealed "a meaningful economic taxonomy" [52], in which equities of similar economic sectors cluster in the same branches of the tree. Such a structure both offers a data-driven description of financial markets and highlights common factors influencing price movements in groups of stocks. Indeed, hierarchical clustering procedures have been applied to construct robust factor models of equity returns [2]. We provide a plot of an MST studied in [52] in Fig. 4.1(a).

In 2005 Tumminello [43] provided an extension of the MST through the Planar Maximally Filtered Graph (PMFG), which is the graph that maximizes the sum of the correlations between all vertices joined by an edge such that the resulting graph is planar, i.e., it can be embedded on a sphere. This construction is especially useful for the study of correlation-based networks, as it preserves the hierarchical organization of the MST but conveys a larger amount of information, allowing triangular loops and four-element cliques. Indeed, the MST is a subgraph of the PMFG. In Fig. 1.1(b) we provide a plot of the PMFG studied in [43] connecting the 100 most capitalized stocks in the U.S. equity markets from 1995 to 1998.

These hierarchical clustering procedures have since been widely applied to the study of financial markets– see [46] for a review of the subject. Crucially, both the MST and the PMFG rely on a ranking of the estimated correlation coefficients, and are therefore sensitive to uncertainties in the estimation process. This property renders hierarchical filtering procedures undesirable for the study of lagged correlation-based networks, as these correlations are typically low in magnitude (see Chapter 2). Moreover, such topological procedures do not readily accommodate the asymmetric nature of lead-lag relationships. An alternative approach that avoids these pitfalls is to simply threshold the correlation matrix, filtering into the network only pairwise correlations beyond a chosen magnitude. In 2010 Tse et al. [1] explored the networks that result from a range of thresholds, finding degree distributions resembling a power law at sufficiently high thresholds. Of course, the properties of the correlation network are highly dependent on this threshold, which should be justified beyond a post-hoc description of the resulting network. A natural idea is to assume some form for the univariate distributions of each signal (such as a normal distribution),



(a) Minimal Spanning Tree (from [52])

(b) Planar Maximally Filtered Graph (from [43])

Figure 1.1: Hierarchical organization of correlations among synchronous equity returns. In Fig. 4.1(a) we display the Minimal Spanning Tree (MST) constructed in [52] that connects the 30 stocks used to compute the Dow Jones Industrial Average from 1989 to 1995. Colors are proportional to the distance represented in the corresponding edge. In Fig. 1.1(b) we show the Planar Maximally Filtered Graph (PMFG) connecting the 100 most capitalized stocks in the U.S. equity markets from 1995 to 1998. The thicker, black lines belong to the associated MST, which is a subgraph of the PMFG.

and to select a threshold that has a sufficiently small probability of being generated by uncorrelated signals of the same length. The resulting threshold would be the same for all edges, however. This is a problem in the case of lagged correlations among equities, as the statistical significance of a lagged correlation is sensitive to the return distributions of the corresponding pair of equities, and such distributions might vary across equities– a consequence, for example, of varying levels of liquidity.

We introduce a methodology for filtering the lagged correlation matrix of such a system into a statistically-validated network, in a way that preserves information about the individual distributions of each signal. We show how consideration of these statistically-robust features of the system allow one to address scientific questions, such as how the nature of risk and contagion is changing in financial markets. Further, we examine the presence of mesoscopic community structures in these networks, and show that consideration of such community structures can provide a new method for building more robust statistical models. We complement this work with novel investigations of the exogenous news items that may drive the financial system, including an analysis of their role in the lead-lag network; how investors and the general public may interact with these news items using Internet search data; and how certain “large-scale” properties of financial news, such as the diversity of stories in the news, may respond to and influence market movements.

Chapter 2

Statistically-validated network methodology

2.1 Motivation

We consider the empirical scenario in which there are many "nodes", each of which is represented by a single time-series. We are also interested in the case in which there is a low signal-to-noise ratio (SNR) among the lagged signals. This is characteristic of returns time series in financial markets, as consistent, high SNR lead-lag relationships tend to be arbitraged away. The entries in the measured lagged correlation matrix will therefore typically be low in magnitude. There are a variety of methods available to filter this matrix, retaining only its robust features. A large class of such methods, such as the minimal spanning tree (MST) or planar maximally-filtered graph (PMFG)– i.e., hierarchical filtering procedures– rely on a ranking of the measured correlations. This is an undesirable property in this scenario, as the results will be sensitive to uncertainties in the measured correlations. Moreover, they often rely on a measure of "distance" and therefore do not readily accommodate the asymmetric nature of lead-lag relationships.

For these reasons we elect to filter the matrix according to a threshold of statistical significance. We must take special care with our statistics, in this scenario, as equity return signals have well-documented heavy tails and thus may violate the normality assumptions of many statistical tests. In Fig. 2.1 we display the cumulative distribution functions for empirical intraday price returns of the top 100 largest market capitalization stocks in the NYSE from 2011-2012. These return signals are defined in section 2.3.1. The figure is represented on a double-logarithmic scale, illustrating the extreme deviation of these signals from the normal CDF curve to a straight line, resembling a power law. In particular, this deviation appears to grow with increasing price sampling frequency, highlighting the importance of this observation when considering intra-day equity returns.

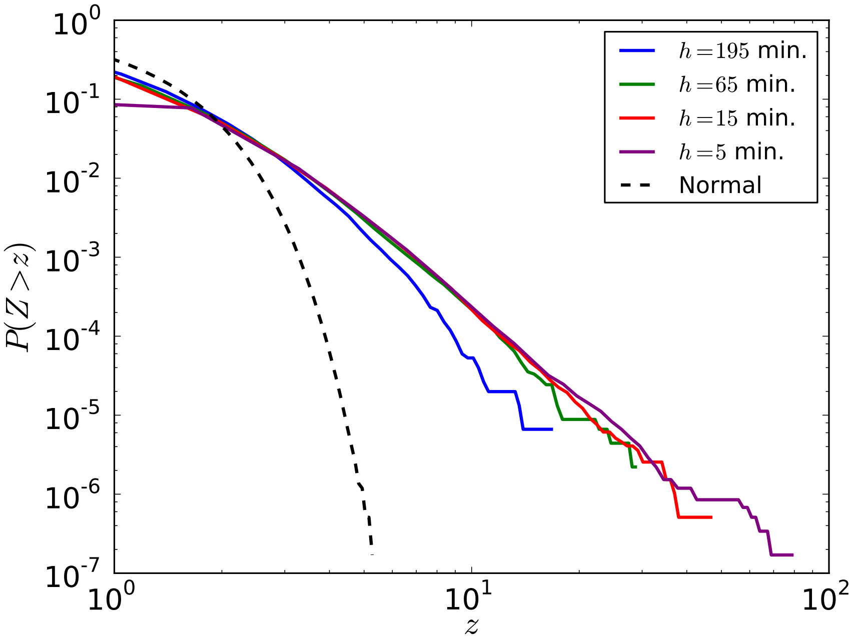


Figure 2.1: Cumulative distribution functions for returns of the top 100 largest market capitalization stocks in the NYSE from 2011-2012. Returns are standardized to z-scores and are aggregated among all stocks for different sampling horizons h, where the parameter h is defined in section 2.3.1. Positive and negative tails of the distribution are aggregated together. We also display the CDF for the standard normal distribution, illustrating the departure of the return signals from a Gaussian as the sampling frequency h^{-1} increases.

We therefore base our statistical framework on permutation tests, taking the distributions of each individual signal from the data in order to validate each directed link.

Crucially, the procedure accounts for multiple hypothesis testing over all pairs of nodes.

Our procedure therefore allows for the filtering of multivariate time-series data into a network of lead-lag relationships in a way that retains information about the distribution of each individual signal. We may use this methodology to learn about real-world systems. We consider the equity market as a case study. In this chapter we introduce the methodology, and apply it to the price fluctuations of stocks in the NYSE in two periods: from 2002-2003 and 2011-2012. For each period, we vary the sampling frequency over which we evaluate price returns. We find that lead-lag relationships are strongly dependent on the sampling frequency, becoming virtually non-existent at sampling frequencies longer than one hour. We also compare the numbers of validated links, network in- and out-degree distributions, and distributions of 3-node directed motifs in order to characterize what is an increasing market efficiency from 2002-2003 to 2011-2012.

We compare our results to those obtained assuming Gaussian distributions for each signal, confirming that the restrictions we impose on our methodology are crucial. In addition, we analytically compute the false positive rate for the methodology. This provides for the quantification of uncertainty in our method, and is the inspiration for work in subsequent chapters that attempts to mitigate the influence of false positives.

2.2 Methodology

Consider N time series, organized as columns in a matrix R. In order to infer directed pairwise relationships among the time series, special care must be taken to render them in an appropriately stationary form. This often involves throwing away information, such as in differencing the time series, so that a minimal amount of pre-processing is desirable. The amount of processing that is necessary varies with the nature of the time series under consideration. In this thesis I consider several different examples.

We then filter R into two matrices, A and B, where B is a version of A that has been time-lagged by some lag l. From these data we construct an empirical lagged correlation matrix C using the Pearson correlation coefficient of columns of A and B,

$$C_{m,n} = \frac{1}{T-1} \sum_{i=1}^{T} \frac{(A_{m,i} - \langle A_m \rangle)(B_{n,i} - \langle B_n \rangle)}{\sigma_m \sigma_n}, \tag{2.1}$$

where $\langle A_m \rangle$ and σ_m are the mean and sample standard deviation, respectively, of column m of A, and T is the number of rows in A (and B).

The matrix C can be considered a weighted adjacency matrix for a fully connected, directed graph. To filter the links in this graph according to a threshold of statistical significance, we apply a shuffling technique [16]. The rows of A are shuffled repeatedly without replacement in order to create a large number of surrogated time series of returns. After each shuffling we re-calculate the lagged correlation matrix (3.8) and compare this shuffled lagged correlation matrix $\widetilde{C}$ to the empirical matrix C. For each shuffling we thus have an independent realization of $\widetilde{C}$. We then construct the matrices U and D, where $U_{m,n}$ is the number of realizations for which $\widetilde{C}_{m,n} \geq C_{m,n}$, and $D_{m,n}$ is the number of realizations for which $\widetilde{C}_{m,n} \leq C_{m,n}$.

From matrix U we associate a one-tailed p-value with all positive correlations as the probability of observing a correlation that is equal to or higher than the empirically-measured correlation. Similarly, from D we associate a one-tailed p-value with all negative correlations. In this analysis we set the threshold at $p = 0.01$. We must adjust our statistical threshold, however, to account for multiple comparisons. We use the conservative Bonferroni correction for a given sample size of N stocks. For example, for $N = 100$ stocks the corrected threshold will be $0.01/N^2 = 10^{-6}$. We thus construct 10^6 independently shuffled surrogate time series. If $U_{m,n} = 0$ we can associate a statistically-validated positive link from stock m to stock n ($p = 0.01$, Bonferroni correction). Likewise, if $D_{m,n} = 0$ we can associate a statistically-validated negative link from stock m to stock n. In this way we construct the Bonferroni network [47]. In section 2.2.1 we discuss the probability that using our approximated method we will wrongly identify a link as statistically significant (i.e., have a false positive).

For the sake of comparison, for each time horizon h we also construct the network using

p-values corrected according to the false discovery rate (FDR) protocol [5]. This correction is less conservative than the Bonferroni correction and is constructed as follows. The p-values from each individual test are arranged in increasing order ($p_1 < p_2 < \cdots < p_{N^2}$), and the threshold is defined as the largest k such that $p_k < k\ 0.01/N^2$. In the FDR network our threshold for the matrices U or D is thus not zero but the largest integer k such that U or D has exactly k entries fewer than or equal to k. From this threshold we can filter the links in C to construct the FDR network [47]. We note that the Bonferroni network is a subgraph of the FDR network.

Because we make no assumptions about the return distributions, this randomization approach is especially useful in high-dimensional systems in which it can be difficult to infer the joint probability distribution from the data [44]. We also impose no topological constraints on the Bonferroni or FDR networks. This method serves to identify the significant positive and negative lagged correlation coefficients in a way that accounts for heterogeneities in relationships between signals under consideration. An alternative, but closely related approach would be to construct a theoretical distribution for correlation coefficients under the null hypothesis of uncorrelated returns sampled from a given joint distribution [7]. For a desired confidence level, one could then construct a threshold correlation, beyond which empirical correlations are validated. Such an approach typically assumes equal marginal distributions for returns, and must fix a uniform correlation threshold for all relationships. At the expense of computational time, our method is flexible in that it permits heterogeneities in marginal distributions. We compare the results of the two approaches in section 2.5.

2.2.1 Probability of a false positive link

The one-tailed p-value associated with positive correlations represents the probability of observing a correlation between two elements, i and j, that is greater than or equal to the observed correlation ρ_{obs} under the null hypothesis that i and j are uncorrelated:

$$p\text{-value}(\rho_{obs}) = P(\rho > \rho_{obs}). \tag{2.2}$$

Our objective is to select all the correlations with a p-value smaller than a given univariate statistical threshold q_0, corrected for multiple hypothesis testing through the Bonferroni correction (i.e., divided by the total number of tests, N^2 in our case, $N = 100$ is the number of signals). Here we compute the probability that a correlation with a p-value p greater than or equal to $p_0 = q_0/N^2$ is (falsely) validated as statistically significant according to the shuffling method. That is, we ask: what is the probability that, over the $Q = k\,N^2$ independent replicates of the data, a correlation between i and j larger than the observed one is never observed?

If we set the p-value, p, of ρ_{obs} equal to $\frac{q}{N^2}$ (where q is a quantity that ranges between 0 and N^2) the question is: what is the probability that, over $Q = k\,N^2$ independent draws ($Q = 100 \cdot N^2 = 10^6$ bootstrap replicates with our method) a value of correlation larger than ρ_{obs} is never obtained? This probability is

$$P(\text{null}|p) = (1-p)^Q, \tag{2.3}$$

where "null" indicates the event that a value of correlation larger than ρ_{obs} has never been obtained over $Q = k\,N^2$ random replicates of data. This probability can be used to calculate the probability that $p = q/N^2$ is larger than or equal to $p_0 = q_0/N^2$, conditioned to the event that a value of correlation larger than ρ_{obs} has never been obtained over $Q = k\,N^2$ draws. This is done using Bayes' rule, under the assumption that the marginal distribution of p-value p is uniform in $[0,1]$, i.e., the density function is $f(p) = 1$. Then, integrating over p,

$$P(p \geq p_0|\text{null}) = \int_{p_0}^{1} \frac{P(\text{null}|p) f(p)}{P(\text{null})} dp = \int_{p_0}^{1} (Q+1)(1-p)^Q dp = (1-p_0)^{Q+1}, \tag{2.4}$$

where we used the fact that $P(\text{null}) = \int_0^1 P(\text{null}|p) f(p) dp = \frac{1}{Q+1}$. In our method, $k = 100$, and the sample size is $N = 100$. Therefore

$$P(p \geq p_0|\text{null}) = \left(1 - \frac{q_0}{N^2}\right)^{k\,N^2+1} \cong \left(1 - \frac{q_0}{N^2}\right)^{k\,N^2} \cong e^{-k\,q_0}. \tag{2.5}$$

It is interesting to note that, as soon as the level of statistical significance is corrected through the Bonferroni correction ($p_0 = q_0/N^2$, where q_0 is the univariate level of statistical significance, and the number, Q, of independent replicates is a multiple of the number of tests, $Q = k\,N^2$), then the probability $P(p \geq p_0|\text{null})$ is approximately independent of the sample size (N).

With our method to estimate correlation p-values, the probability that we select a positive correlation as statistically significant at the confidence level $p_0 = q_0/N^2 = 0.01/100^2 = 10^{-6}$, while it is actually not significant at that level of statistical confidence, is $P(q \geq 0.01|\text{null}) = \frac{1}{e} \cong 0.368$. However, the probability that a significant correlation according to our method has a p-value larger then $0.05/N^2 = 0.05/100^2 = 5 \cdot 10^{-6}$ is already quite small: $P(q \geq 0.05|\text{null}) = \frac{1}{e^5} \cong 0.0067$. In other words, if we obtain a validated network with 1,000 links, i.e., 1,000 validated positive correlations according to our approximated method, we expect that, on average, only 7 correlations will have a one-tailed p-value larger than $0.05/100^2 = 5 \cdot 10^{-6}$.

2.3 Relevance to financial data

Modern financial markets have developed lives of their own. This fact makes it necessary that we not only monitor financial markets as an "auxiliary system" of the economy, but that we develop a methodology for evaluating them, their feedback on the real economy, and their effect on society as a whole [13, 21]. The events of the recent past have clearly demonstrated that the everyday life of the majority of the world's population is tied to the well-being of the financial system. Individuals are invested in stock markets either directly or indirectly, and shocks to the system (be they endogenous or exogenous) have an immense and immediate impact. Thus the need for a robust and efficient financial system is becoming stronger and stronger. These two critical concepts have been discussed and heatedly debated for the past century, with the efficient market hypothesis (EMH) in the center of the debate.

The EMH [32, 41] stipulates that all available information (or only past prices in the weak variant of the hypothesis) is already reflected in the current price and it is therefore not possible to predict future values in any statistical method based on past records [31].

The EMH has been questioned by applying statistical tests to NYSE returns [30, 39] in which the authors formulated the problem equivalent to the EMH, and showed by contrast that an efficient compression algorithm they proposed was able to utilize structure in the data—which would not be possible if the hypothesis were in fact true. The possibility for such compression suggests the data must be somehow structured. This encourages us to explore methods of modeling and exploring this structure in ways that can be applied to real-world markets.

Many efforts have thus been devoted to uncovering the true nature of the underlying structure of financial markets. Much attention has been given to understanding correlations in financial markets and their dynamics, for both daily [4, 11, 12, 14, 18, 19, 25, 26, 37, 38, 52] and intra-day time scales [9, 10, 35, 45]. More recently, other measures of similarity have been introduced, such as Granger-causality analysis [6] and partial correlation analysis [27], both of which aim to quantify how the behavior of one financial asset provides information about the behavior of a second asset. For these different measures of co-movement in financial markets, however, the main question that remains is how to uncover underlying meaningful information.

An analysis of synchronous correlations of equity returns has shown that a financial market usually displays a nested structure in which all the stock returns are driven by a common factor, e.g., a market index, and are then organized in groups of like economic activity—such as technology, services, utilities, or energy—that exhibit higher values of average pair correlation. Within each group, stocks belonging to the same sub-sector of economic activity, e.g., “insurance” and “regional banks” within the financial sector, show an even higher correlation degree. Such a structure has been recognized using very different methods of analysis, ranging from random matrix theory [20, 29], to hierarchical clustering [52], to correlation based networks [8, 36, 52]. The several methods devised to construct correlation based networks can be grouped into two main categories: threshold methods and topological/hierarchical methods. Both approaches start from a sample correlation matrix or, more generally, a sample similarity measure. Using the threshold method we set a correlation threshold and construct a network in which any two nodes are linked if their correlation is larger than the threshold. As we lower the threshold value we see the formation of groups of stocks (economic sub-sectors) that progressively merge to form larger groups (economic sectors) and finally merge into a single group (the market). The advantage of this approach is that, due to the finite length of data series, threshold networks are very robust to correlation uncertainty. The disadvantage of threshold based networks is that it is difficult to find a single threshold value to display, in a single network, the nested structure of the correlation matrix of stock returns (see [27]). Topological methods to construct correlation based networks, such as the minimal spanning tree (MST) [8, 9, 36, 52] or the planar maximally-filtered graph (PMFG) [43], are based solely on the ranking of empirical correlations. The advantage of this approach is that these methods are intrinsically hierarchical and are able to display the nested structure of stock-return correlations in a financial market. The disadvantage of this approach is that these methods are less stable than threshold methods with respect to the statistical uncertainty of data series, and it is difficult to include information about the statistical significance of correlations and their ranking [44]. Thus it is a challenge of modern network science to uncover the significant relationships (links) between the components (nodes) of the investigated system [22].

Although much attention has been devoted to the study of synchronous correlation networks of equity returns (see [46] for a review of the topic), comparatively few results have been obtained for networks of lagged correlations [24]. Neither method of constructing correlation based networks is readily extendable to the study of directed lagged correlations in a financial market. The lagged correlations in stock returns are small, even at time horizons as short as five minutes, and are thus strongly influenced by the statistical uncertainty of the estimation process. The use of topological methods to construct a lagged-correlation based network of stock returns is difficult because they only take into consideration the ranking of correlations and not their actual values. The result could be a network in which many links are simply caused by statistical fluctuations. On the other hand, standard threshold methods are also difficult to apply because it is difficult to find an appropriate threshold level and, more importantly, the threshold selected in these methods is usually the same for all stock pairs. This is a problem if we want to study lagged correlations because the statistical significance of a lagged-correlation may depend on the return distribution of the corresponding pair of stocks, and such distributions might vary across stocks—a consequence, for example, of the different liquidity of stocks.

Here we apply the method of section 2.2 to describe the structure of lagged relationships between intraday equity returns sampled at high frequencies in financial markets. In particular, we investigate how the structure of the network changes with increasing return sampling frequency, and compare the results using data from both the periods 2002–2003 and 2011–2012. It should be noted that the two investigated time periods are quite different if we consider that the fraction of volume exchanged by algorithmic trading in the US equity markets has increased from approximately 20% in 2003 to more than 50% in 2011. In both periods we find a large growth in the connectedness of the networks as we increase the sampling frequency.

2.3.1 Statistically validated lagged correlation networks in financial markets

We begin the analysis by calculating the matrix of logarithmic returns over given intraday time-horizons. We denote by $p_n(t)$ the most recent transaction price for stock n occurring on or before time t during the trading day. We define the opening price of the stock to be the price of its first transaction of the trading day. Let h be the time horizon. Then for each stock we sample the logarithmic returns,

$$r_{n,t} \equiv \log(p_n(t)) - \log(p_n(t-h)), \tag{2.6}$$

every h minutes throughout the trading day, and assemble these time series as columns in a matrix R. We then filter R into two matrices, A and B, in which we exclude returns during the last period h of each trading day from A and returns during the first period h of each trading day from B. Here we set the lag to be one time horizon h. A schematic of this sum is diagrammed in Fig. 2.2. This forms the matrices A and B from section 2.2,

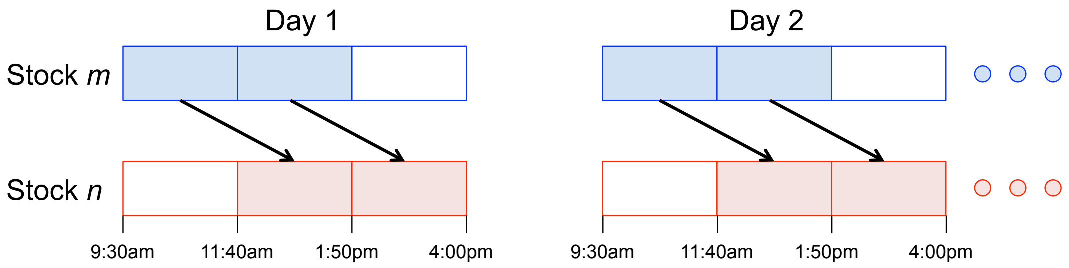


Figure 2.2: Schematic of lagged correlation calculation for a time horizon $h = 130$ minutes. The sum $C_{m,n}$ is generated using products of returns from stocks m and n that are linked by an arrow. We consider only time horizons h that divide evenly into the 390 minute trading day.

and we proceed to construct the statistically-validated lagged correlation networks exactly as described.

We study and compare two different datasets. The first dataset comprises returns of 100 companies with the largest market capitalization on the New York Stock Exchange (NYSE) during the period 2002–2003 (501 trading days), which was investigated in [45]. For the second dataset we consider returns during the period 2011–2012 (502 trading days) of 100 companies with the largest market capitalization on the NYSE as of December 31, 2012 (retrieved from the Trades and Quotes database, Wharton Research Data Services, http://wrds-web.wharton.upenn.edu/wrds/). Market capitalization figures were obtained from Yahoo Finance web service (http://finance.yahoo.com). For each company we obtain intraday transaction records. These records provide transaction price data at a time resolution of one second. The stocks under consideration are quite liquid, helping to control for the problem of asynchronous transactions and artificial lead-lag relationships due to different transaction frequencies [15]. Transaction data were cleaned the for canceled trades and trades reported out of sequence. We then sample returns at time horizons of 5, 15, 30, 65, and 130 minutes.

We report summary statistics in Table 2.1, including the lengths of time series T from

Table 2.1: Summary statistics of 2002-2003 and 2011-2012 datasets.

Period	T	h	$\langle\rho\rangle$	σ_ρ	$\langle C_{m,n}\rangle$	σ_C
2002-2003	38,577	5 min.	0.267	0.077	0.008	0.024
	12,525	15 min.	0.290	0.092	0.007	0.025
	6,012	30 min.	0.307	0.102	0.005	0.025
	2,505	65 min.	0.317	0.110	0.015	0.029
	1002	130 min.	0.327	0.115	0.022	0.038
2011-2012	38,654	5 min.	0.380	0.121	0.006	0.024
	12,550	15 min.	0.411	0.115	0.006	0.022
	6,024	30 min.	0.422	0.115	0.017	0.024
	2,510	65 min.	0.448	0.119	-0.003	0.027
	1004	130 min.	0.452	0.126	-0.019	0.033

equation (3.8), as well as the mean $\langle\rho\rangle$ and standard deviation σ_ρ of synchronous Pearson correlation coefficients between distinct columns of the returns matrix R for each time horizon h. We also show the mean $\langle C_{m,n}\rangle$ and standard deviation σ_C of entries in the lagged correlation matrix C.

Figure 2.3 displays bounds on the positive and negative coefficients selected by this method for both Bonferroni and FDR networks at a time horizon of $h = 15$ minutes.

In Fig. 2.4 we display plots of each statistically validated lagged correlation network obtained from the 2011–2012 data (Bonferroni correction). At time horizons of $h = 130$ minutes and $h = 65$ minutes we validate one and three links, respectively. It is somewhat remarkable that we uncover any persistent relationships at such long time horizons.

We see a striking increase in the number of validated links at small intraday time horizons, below $h = 30$ minutes in particular. This is likely due to a confluence of two effects: (i) with decreasing h we increase the length T of our time series, gaining statistical power and therefore the ability to reject the null hypothesis; (ii) at small h we approach the timescales over which information and returns spill over across different equities. In section 2.6 we provide evidence that diminishing the time horizon h reveals more information about the

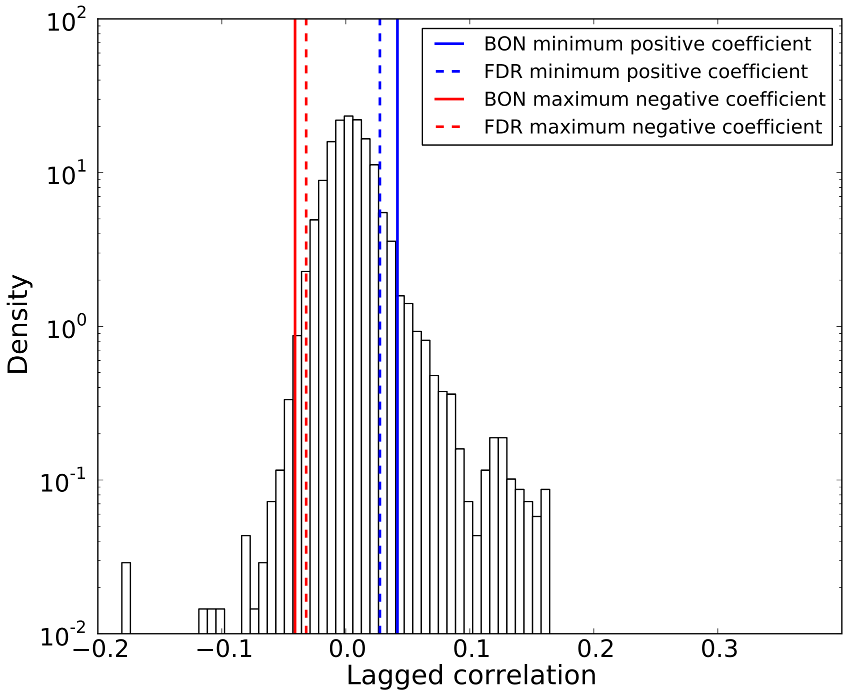
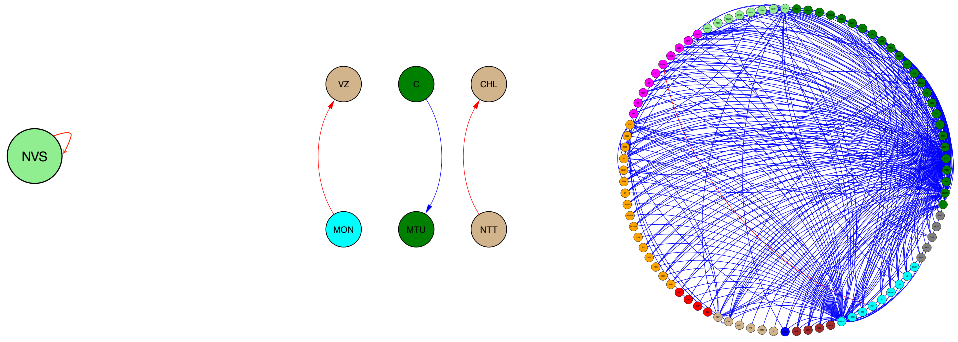


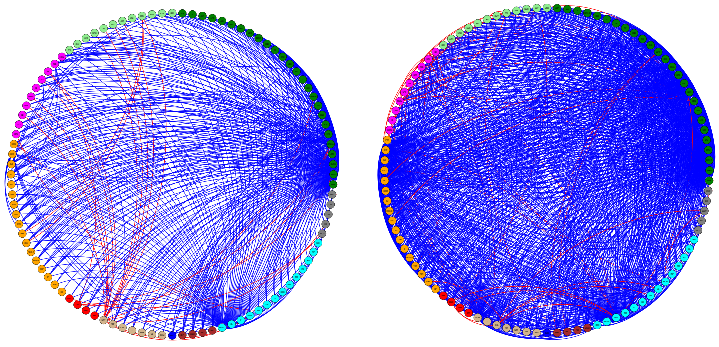
Figure 2.3: Distribution of lagged correlation coefficients for all $N = 100$ stocks at a time horizon $h = 15$ minutes. The minimum positive coefficients and maximum negative coefficients selected using both Bonferroni and FDR filtering procedures are shown. We note that these methods select coefficients from the tails of the distribution, without fixing a uniform threshold for all pairs of stocks.

system than is obtained by increasing the time series length T alone.

It is clear visually that the validated links of positive correlation vastly outnumber the validated links of negative correlation. We plot the number of validated links in both the Bonferroni and FDR networks for the 2002–2003 and 2011–2012 datasets in Fig. 2.5, where the decrease in number of all validated links for increasing time horizon is apparent. Note that for a given time horizon we usually validate more links in the 2002–2003 dataset than in the 2011–2012 dataset. This suggests that there has been an increase in market efficiency over the past decade. We revisit this idea in subsequent portions of this chapter, where we



(a) $h = 130$ minutes (b) $h = 65$ minutes (c) $h = 30$ minutes



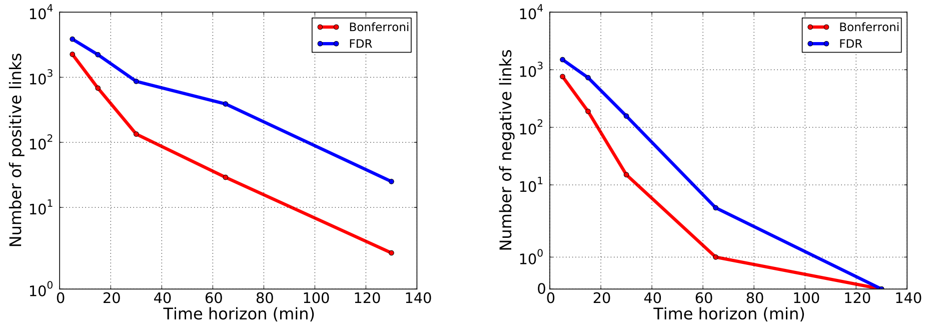
(d) $h = 15$ minutes (e) $h = 5$ minutes

Figure 2.4: Illustrations of Bonferroni networks constructed from statistically-validated lagged correlations for various time horizons h. Data were obtained from returns of large market-capitalization companies on the NYSE in 2011-2012. Nodes are colored by industry. Blue links represent positive correlations; red links represent negative correlations.

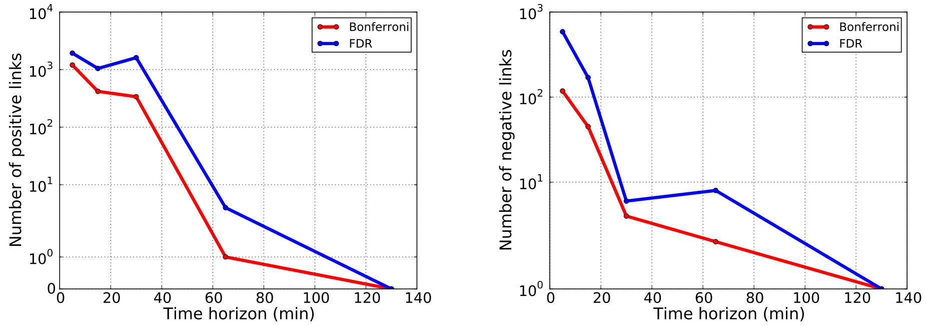
study the properties of the network in- and out-degree distributions and the characterization of three-node motifs.

We also explore how the number of validated links decreases for a fixed time horizon h but a changing time lag. We build a lag l into the lagged correlation matrix (3.8) by excluding the last l returns of each trading day from matrix A and the first l returns of each trading day from matrix B. Thus the present analysis uses $l = 1$. In section 2.6 we plot the decreasing number of validated links with increasing l for $h = 15$ minutes.

We must also measure the extent to which the number of validated lead-lag relationships



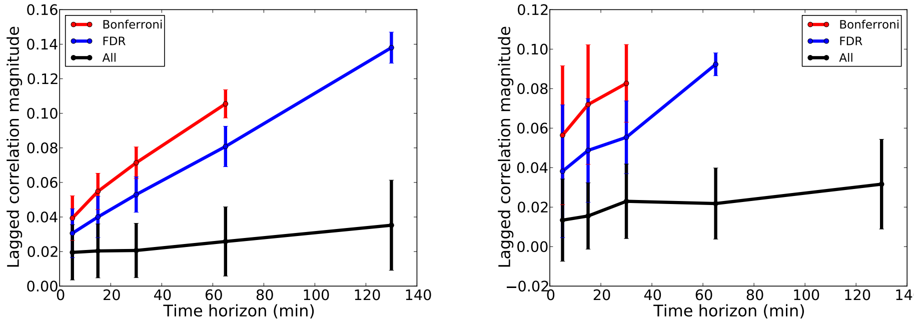
(a) Links of positive correlation, 2002-2003 (b) Links of negative correlation, 2002-2003



(c) Links of positive correlation, 2011-2012 (d) Links of negative correlation, 2011-2012

Figure 2.5: Plots of the number of positive and negative validated links for both Bonferroni and FDR lagged correlation networks. The decrease in number of validated links for increasing time horizon is apparent in both the 2002-2003 and 2011-2012 datasets. The vertical axis is presented on a logarithmic scale that is linearized near zero.

can be disentangled from the strength of those relationships. Figure 2.6 thus shows plots of the average magnitude of lagged correlation coefficients selected by the Bonferroni and FDR networks. Although we validate more links at small time horizons, we note that the average magnitude of the selected coefficients tends to decrease. At short time horizons h we correlate time series of comparatively large length T, narrowing the distribution of entries in the shuffled lagged correlation matrix $\widetilde{C}$ and gaining statistical power. We are thus able to reject the null hypothesis even for lead-lag relationships with a modest correlation coefficient.

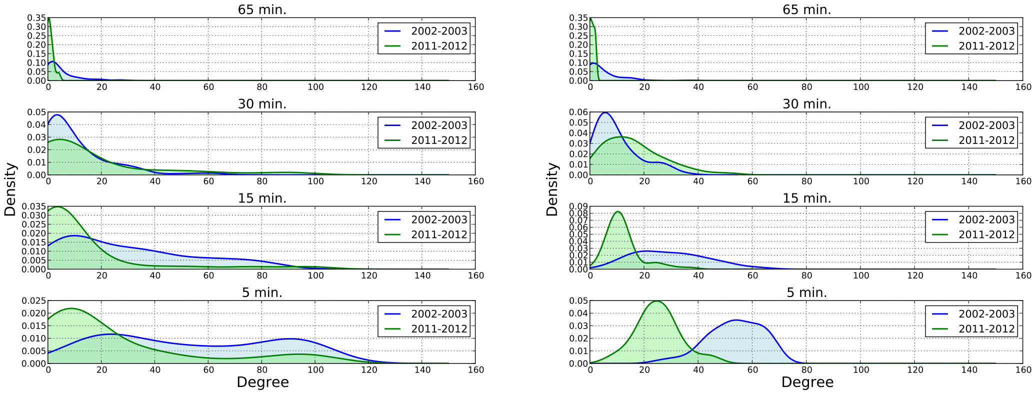


(a) 2002-2003 (b) 2011-2012

Figure 2.6: Average magnitude (absolute value) of lagged correlation coefficients filtered in Bonferroni and FDR networks. Magnitudes appear to grow with increasing time horizon of return sampling. Error bars represent plus-or-minus one standard deviation. Results are displayed only for networks containing at least five links.

Finally, in Fig. 2.7 we characterize the topologies of the statistically-validated networks by studying the properties of their in-degree and out-degree distributions. We make two observations. First, we note that both the in-degree and out-degree distributions appear more homogeneous in the 2002–2003 period than the 2011–2012 period, i.e., the 2011–2012 data exhibit large heterogeneities, particularly in the in-degree distributions, in which many nodes have small degrees but few nodes have very large degrees, as can be seen in the extended tails of the distributions. Second, we observe that in both the 2002–2003 and 2011–2012 data there are more nodes with large in-degrees than out-degrees. Although few individual stocks have a strong influence on the larger financial market, it appears that the larger financial market has a strong influence on many individual stocks, especially at short time horizons.

We further investigate this point by studying the relative occurrence of three-node network motifs in the Bonferroni networks [34]. We find that, of all motifs featuring more than one link, the "021U" motif (two nodes influencing a common third node) occurs frequently in the recent data, and in fact occurs in over 80% of node triplets having more than one link



(a) In-degree distributions of FDR networks (b) Out-degree distributions of FDR networks

Figure 2.7: In- and out-degree distributions for FDR networks from 2002–2003 (blue) and 2011–2012 (green). Smoothed distributions are obtained using a kernel density estimate with a Gaussian kernel. With the exception of the $h = 30$ minute in-degree distributions, at each of the 30 min., 15 min., and 5 min. time horizons the distributions from 2002–2003 and 2011–2012 are statistically distinct ($p < 0.05$, all test statistics $W > 130$, two-tailed two-sample Wilcoxon rank-sum tests, Bonferroni correction applied).

between them for time horizons greater than $h = 65$ minutes. In the 2002–2003 data this motif is also the most common at every time horizon except $h = 65$ minutes. Figure 2.8 plots the occurrence frequencies of these motifs. These features can be related to the information efficiency of the market. In the 2011–2012 dataset we find a dominant motif in which a large number of stocks influence only a few other stocks. Predictive information regarding a given stock, therefore, tends to be encoded in the price movements of many other stocks and so is difficult to extract and exploit. In contrast, the distributions of degrees and motifs in the 2002–2003 data are more homogeneous. Although there are more nodes with large in-degrees, there are also more nodes with large out-degrees. If a stock has a large out-degree, its price movements influence the price movements of many other stocks. These sources of exploitable information have all but disappeared over the past decade.

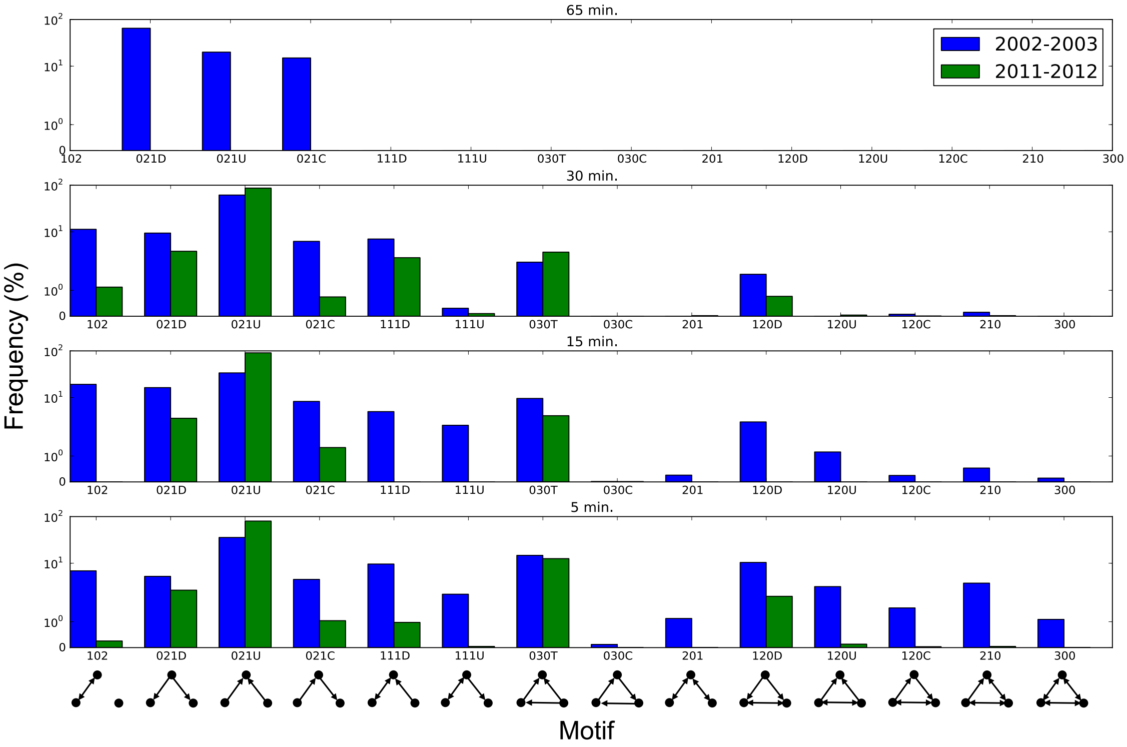
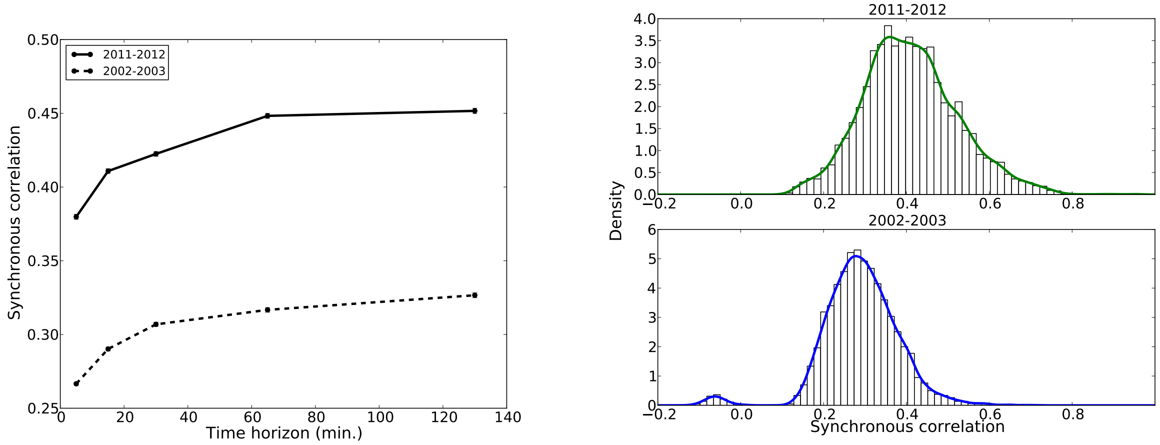


Figure 2.8: Percentage occurrence of all 14 possible directed three-node motifs with more than one link in Bonferroni networks. The total number of such motifs in 2002-2003 are 40 ($h = 65$ min.), 1,296 ($h = 30$ min.), 17,545 ($h = 15$ min.), and 92,673 ($h = 5$ min.). In 2011-2012 these counts are 0 ($h = 65$ min.), 8,710 ($h = 30$ min.), 13,850 ($h = 15$ min.), and 46,687 ($h = 5$ min.).

2.3.2 Synchronous correlation networks

To construct synchronous correlation networks using the methodology described in Sec. 2.3.1, we use the unfiltered columns of R as our time series such that each entry $C_{m,n}$ of the empirical correlation matrix is the Pearson correlation between columns m and n of R. We then independently shuffle the columns of R, without replacement, when constructing the surrogated time series. We find that with the same significance threshold of $p = 0.01$, in 2011-2012 both the Bonferroni and FDR networks are almost fully connected, with well over 4500 of the $N(N-1)/2 = 4950$ possible links validated in all networks over all time horizons. Our method is thus quite sensitive to the presence or absence of correlations between time series.

Figure 2.9(a) plots the empirical synchronous correlations against time horizon for all stocks considered in both datasets. We see a clear increase in the magnitude of these coefficients as the time horizon grows, a phenomenon known as the Epps Effect [17, 45]. It is known that lagged correlations may in part contribute to this effect [42]. The extent of this contribution is an active area of investigation [48]. The synchronous correlations are also significantly higher in the recent data, suggesting that, despite the increased efficiencies shown in Fig. 2.5, there is also an increase in co-movements in financial markets since 2003, heightening the risk of financial contagion (see for example [26, 40]).



(a) Epps curves for 2002-2003 and 2011-2012 data.

(b) Distributions of correlations at a 15 minute time horizon.

Figure 2.9: (a) Plot of mean synchronous correlation coefficients in both 2002-2003 and 2011-2012 data. Error bars represent plus-or-minus one standard deviation of the mean. (b) Histograms of correlation coefficients for returns sampled at a 15 minute time horizon. Solid curves show kernel density estimates using a Gaussian kernel. Depicted distributions are statistically distinct ($p < 0.001$, test statistic $W = 19415612$, two-tailed two-sample Wilcoxon rank-sum test).

Figure 2.9(b) shows the distribution of correlation coefficients at $h = 15$ minutes for both 2002–2003 and 2011–2012 datasets. We observe a slightly bi-modal distribution of synchronous correlation coefficients in the 2002–2003 data across all time horizons h. Most coefficients are positive, but there is also a small number of negative coefficients among these high market capitalization stocks. This quality disappears in the 2011–2012 data, and all correlation coefficients are positive.

2.4 Discussion

In this chapter, we propose a method for the construction of statistically validated correlation networks. The method is applicable to the construction of both lagged (directed) and synchronous (undirected) networks, and imposes no topological constraints on the networks. The sensitivity of the method to small deviations from the null hypothesis of uncorrelated returns makes it less useful for studying the synchronous correlations of stocks, as these equities tend to display a considerable degree of correlation and we validate almost all possible links in the network. The method is apt, however, for the study of lagged correlation networks. We are able to adjust the sensitivity of the method with our choice of p-value and protocol for multiple comparisons. Here we show that, with the conservative Bonferroni correction and p-value=0.01, we are able to compare changes in network connectivity with increasing return sampling frequency between old and new datasets. The primary drawback to our method is its computational burden, which grows as $\mathcal{O}(N^4)$ for N time series.

We find that for timescales longer than one hour, significant lead-lag relationships that capture return and information spill-over virtually disappear. For timescales smaller than 30 minutes, however, we are able to validate hundreds of relationships. According to the efficient market hypothesis there can be no arbitrage opportunities in informationally-efficient financial markets. However, lagged correlations may not be easily exploitable due to the presence of market frictions, including transaction costs, the costs of information processing, and borrowing constraints.

Between the time periods 2002–2003 and 2011–2012, the synchronous correlations among these high market capitalization stocks grow considerably, but the number of validated lagged-correlation relationships diminish. We relate these two behaviors to an increase in the risks of financial contagion and an increase in the informational efficiency of the market, respectively. We find that networks from both periods exhibit asymmetries between their in-degree and out-degree distributions. In both there are more nodes with large in-degrees than large out-degrees, but in the 2011–2012 data, nodes with large in-degrees are represented by the extended tails of the degree distribution and, in contrast, the 2002–2003 distribution exhibits a greater uniformity. A comparison between in-degree and out-degree distributions shows that nodes with high in-degree are much more likely than nodes with high out-degree, especially for the 2011–2012 data. This evidence is also interpreted in terms of informational efficiency of the market. Indeed a large out-degree of a stock implies that knowledge of its return, at a given time, may provide information about the future return of a large number of other stocks. On the other hand, a large in-degree of a stock indicates that information about its return at a given time can be accessed through the knowledge of past returns of many stocks. There are also many more nodes with large out-degrees in the 2002–2003 data than in the 2011–2012 data. We relate these observations to an increased information efficiency in the market. Such an interpretation is also supported by the analysis of three-node motifs, which shows an apparent dominance of motif 021U with respect to all the others.

2.5 Comparison of the bootstrap method and an analytical one to calculate correlation p-values

Here we compare (for a sub-set of our data) the number of significant correlations obtained according to the presented bootstrap approach and the number of significant correlations that we may have obtained relying upon the analytical distribution of sample pair correlations of normally distributed data.

If x and y are uncorrelated variables that follow a normal distribution, then the proba-

Table 2.2: Threshold-correlation values and validated links according to a normal distribution of returns

T	h	ρ_t	# pos. valid.	# neg. valid
38,577	5 min	0.0242	2,398	793
12,525	15 min	0.0425	754	212
6,012	30 min	0.0613	158	19
2,505	65 min	0.0948	43	3
1002	130 min	0.1496	3	0

bility density function of the sample correlation coefficient, r, between x and y is [28]

$$f(r,T) = \frac{(1-r^2)^{\frac{T-1}{2}-2}}{B(\frac{1}{2}, \frac{T-1}{2}-1)}, \tag{2.7}$$

where T is the length of the sample and $B(q,p)$ is the Euler beta function of parameters q and p. Given a level of statistical significance, q_0/N^2 (already corrected for multiple hypothesis testing), $f(r,T)$ can be used to set a threshold for the correlation value ρ_t such that the probability $P(\rho > \rho_t) = \frac{q_0}{N^2}$ is

$$P(\rho > \rho_t) = \int_{\rho_t}^{1} f(r,T)dr = \frac{q_0}{N^2}. \tag{2.8}$$

According to this analysis, for a data sample of N time series, each one of length T, we can say that an observed correlation, ρ_{obs}, is statistically significant if $\rho_{obs} > \rho_t$, where ρ_t is obtained by (numerically) solving the previous non linear equation.

Table B1 shows the 2002–2003 dataset and reports the length of data series used to calculate lagged correlations (column 1) at a given time horizon (column 2), the quantity ρ_t such that $P(\rho > \rho_t) = 0.01/N^2 = 10^{-6}$ (column 3), the number of validated positive correlations (column 4), and the number of validated negative correlations (column 5).

Table B2 shows the number of validated positive correlations (i) according to the shuffling method (column 3), (ii) according to the analytical method discussed above (column 4), and (iii) common to both methods (column 5). The results reported in the table show that the bootstrap method we used is more conservative than the analytical method based on the assumption that return time series follow a normal distribution. Indeed the number

Table 2.3: Comparison between number of positive validated links according to the bootstrap method and a normal distribution of returns

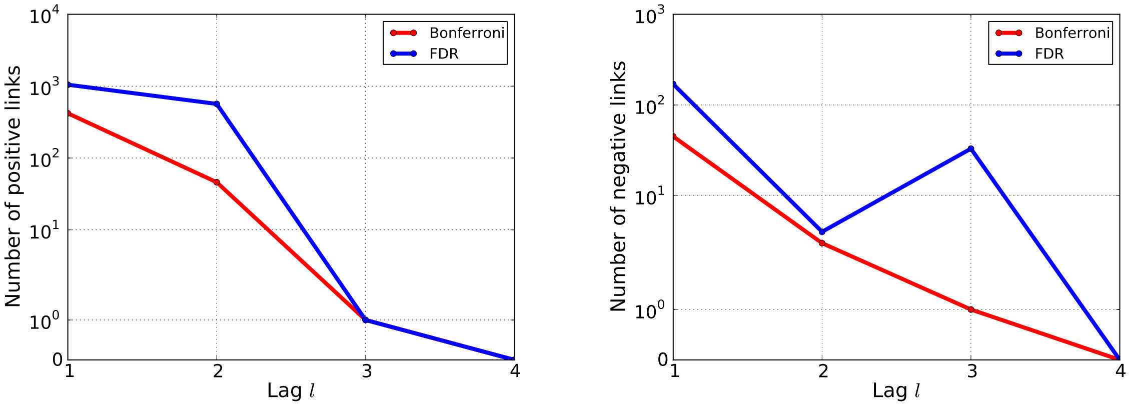
T	h	bootstrap	normal	both
38,577	5 min	2,252	2,398	2,230
12,525	15 min	681	754	666
6,012	30 min	134	158	131
2,505	65 min	29	43	26
1002	130 min	2	3	2

of validated positive correlations according to the bootstrap method is always smaller than the one obtained using the theoretical approach. Furthermore, most of the correlations validated according to the bootstrap method are also validated according to the theoretical method.

A similar discussion can be held about the validation of negative correlations.

2.6 Effect of lag and time series length on validated links for a fixed time horizon

We explore how the number of validated links decreases when the time horizon h is fixed and the time lag variable l increases. A lag l is built into the lagged correlation matrix (3.8) by excluding the last l returns of each trading day from matrix A and the first l returns of each trading day from matrix B. Thus the results presented in the main text are restricted to $l = 1$. Figure 2.10 plots the number of positive links and negative links validated in the 2011–2012 data for $h = 15$ minutes as l increases. Although for this h the length T of the time series in A and B decrease by only $\approx 4\%$ for each additional lag l (as each 390 minute trading day includes $390/15 - l = 26 - l$ returns), we observe a sharp decrease in the number of validated links as l increases. The number of validated negative links is an order of magnitude smaller than the number of positive links, so the small peak in negative links at $l = 3$ for the FDR network is likely an artifact of noise.

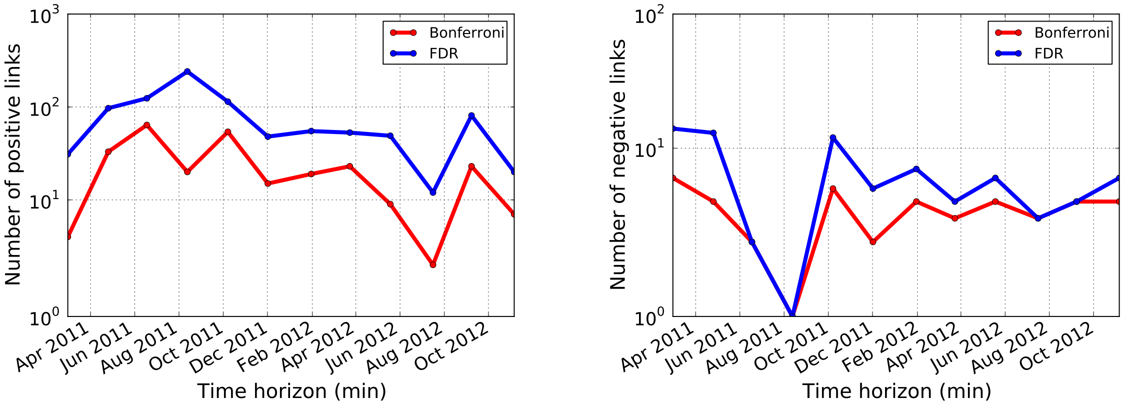


(a) Links of positive correlation

(b) Links of negative correlation

Figure 2.10: Numbers of positive and negative validated links for both Bonferroni and FDR correlation networks for varying lag l. Returns are sampled every $h = 15$ minutes from the 2011-2012 data.

We also investigate the effect of the time series length T on the numbers of validated links. For $h = 15$ minutes, we partition the entire 2011-2012 time series into segments of length $T = 1004$, as this is the length of the time series for the longest time horizon considered ($h = 130$ minutes). For each segment we generate the lagged correlation network using 10^6 surrogate time series, as before. We find that the union of all such Bonferroni networks consists of 125 distinct links, 106 of which are positive and 19 of which are negative. Although this number is 30% of the number of links validated in the $h = 15$ minute network that was not partitioned ($T = 12,550$), it stands in contrast to the single link that was validated in the $h = 130$ minute Bonferroni network using the entire time period. The number validated in each partition is shown in Figure 2.11. We can thus safely conclude that decreasing the time horizon h provides information independent of the increased time series length T.



(a) Links of positive correlation

(b) Links of negative correlation

Figure 2.11: Numbers of positive and negative validated links for both Bonferroni and FDR lagged correlation networks for time series segments of length $T = 1004$ at $h = 15$ minutes. Horizontal axis gives date of the final return in each network.

Chapter 3

Seasonalities and the Epps Effect

In this chapter we extend the methodology of Chapter 2 to accommodate seasonalities in the multivariate data under consideration. The approach is to identify periodicities in the terms composing the Pearson product-moment sums (as averaged across all node pairs, for example), using either Fourier analysis or the autocorrelation function. Once this periodicity is established, separate networks can be constructed using only terms that are spaced according to the desired periodicity.

This approach allows us to investigate seasonal effects in multivariate systems. As an example, we return to price fluctuation data for equities in the NYSE. We identify strong daily periodicities in the measured synchronous correlations, motivating us to explore the intraday profile of synchronous and lagged correlations in a characteristic trading day. We construct separate networks for each intraday period, providing a picture of the dynamics of lagged correlations among equities in a single trading day. We report several novel phenomena. Most notably, while the network is sparse and clustered by economic sectors in the early portion of the trading day, toward the end of the trading day we observe an explosion in network connectivity that is largely agnostic of economic sector. We suggest several explanations for this observation, which is consistent between datasets from 2001-2003 and 2011-2013.

The Epps Effect is an empirical phenomenon in financial markets whereby measured synchronous correlations grow as the sampling frequency over which one computes returns decreases. There are several explanations which compete in the literature, including variations in human reaction time to news; liquidity effects; and the influence of lagged correlations. We quantify the contribution of lagged correlations to the Epps Effect by analytically decomposing the synchronous correlation coefficient at long time horizons into terms corresponding to the synchronous correlation, lagged cross-correlation, and autocorrelation at shorter time-horizons. In this way we may "reconstruct" the synchronous correlation matrix, under a minimal set of assumptions, using combinations of lagged cross-correlations and autocorrelations at various time lags. Our finding of persistent intraday seasonalities motivates us to trace how the contributions of autocorrelations and cross-correlations evolve during the trading day. We also compare results using data from 2001-2003 and 2011-2013. We find structural problems in the reconstructed correlation matrix in the 2011-2013 data, as it is not positive definite, indicating the "tangling" of autocorrelations and lagged cross-correlations. We suggest several explanations of this phenomenon.

3.1 Incorporating seasonalities

In Chapter 2 we introduced a methodology for associating a statistically-validated network of directed (time-lagged) relationships to multivariate datasets. Given a procedure for accounting for multiple hypothesis testing (e.g., Bonferroni or FDR), we describe a system with a single, static network.

A natural extension of this methodology is to accommodate networks that are not static, but dynamic. A relationship between two nodes may exist in only a fraction of the data, or may appear at regular intervals. This latter phenomenon may indicate the presence of seasonal effects, or periodicities in the underlying interactions. Here we show how to uncover such seasonal effects in correlation-based networks, and how to extend the methodology of Chapter 2 to accommodate these effects.

Using the notation of Chapter 2, we may write the mean synchronous correlation as averaged over all L validated links as the sum:

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

with $\langle C\rangle_t$ the defined as the term in brackets in the second line. The term $a_{m,n}$ takes value one if there is a link from m to n, and is zero otherwise. The sum of these terms is then the mean lagged correlation as averaged over all validated links.

Periodicities in the time series $\langle C\rangle_t$ can be uncovered using standard tools, such as the autocorrelation function or Fourier transform. In Figure 3.1 we show the power spectrum of the discrete Fourier transform $f(\omega)$ of $\langle C\rangle_t$, where the matrices A and B consist of five minute returns using the NYSE data of Chapter 2. We identify peaks in this function at frequencies ω corresponding to one trading day: $\omega = 390\ \text{min}^{-1}$, suggesting the presence of daily periodicities in the terms that contribute to the average validated correlation. This suggests the existence of intraday patterns in the collective dynamics of the system under investigation, which may be important to take into account when modeling interactions among elements of the system. In addition, if we don't restrict our averaging to validated links, i.e., if $a_{m,n} = 1\forall m, n$, then the corresponding power spectrum for these data is featureless. The statistically validated network of Chapter 2 exposed this periodicity in the system's lagged correlations.

In particular, a single, static network associated to such a multivariate system may aggregate together relationships that "come and go" periodically in the data. We suggest a simple resolution to this problem: if we identify seasonal effects at some discrete periodicity τ, then we partition our data into τ regularly-spaced buckets and construct τ distinct networks. This approach can highlight the dynamics of the multivariate system during one characteristic period of its evolution. In this chapter we illustrate the approach in the context of the same returns data from Chapter 2, evaluated at a time horizon of $\Delta t =$

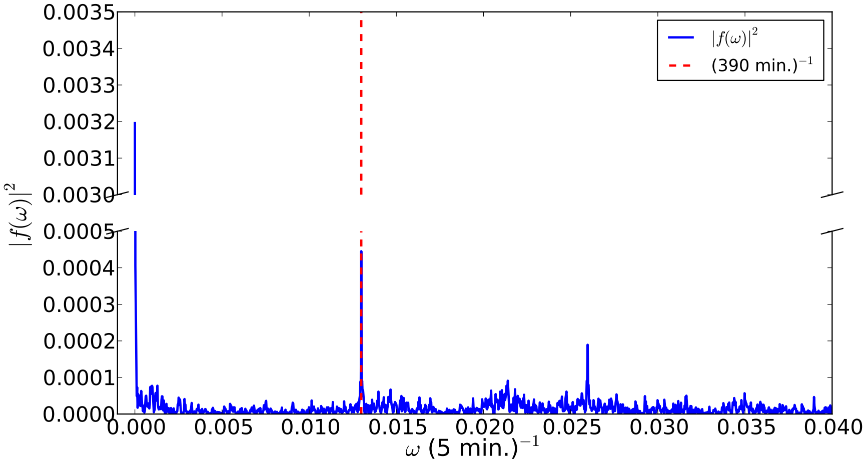


Figure 3.1: Power spectrum of $\langle C \rangle_t$. The terms $\langle C \rangle_t$ are calculated using the data of Chapter 2, where returns are evaluated at five minute intervals. We identify peaks in this function at frequencies ω corresponding to one trading day: $\omega = 390 \text{ min}^{-1}$. Secondary peaks are visible at twice this frequency, as well.

15 minutes. Our approach reveals several novel phenomena. In addition, we then use this approach to explain how the factors contributing to the Epps effect, an empirical phenomenon in financial markets, themselves evolve during a characteristic trading day.

3.2 Application to equity returns

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

Financial markets are a primary example of a setting in which this approach has value. When constructing an optimal portfolio of assets, for example, the goal is typically to allocate resources so as to balance the tradeoff between return and risk. As has been understood at least since the work of Markowitz [49], 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.

Time series obtained by monitoring the evolution of a multivariate complex system, such as time series of price returns in a financial market, can be used to extract information about the structural organization of such a system. This is generally accomplished by using the correlation between pairs of elements as a similarity measure, and analyzing the resulting correlation matrix. A spectral analysis of the sample correlation matrix can indicate deviations from a purely random matrix [50, 51] or more structured models, such as the single index [50]. Clustering algorithms can also be applied to elicit information about emergent structures in the system from a sample correlation matrix [52]. Such structures can also be investigated by associating a (correlation-based) network with the correlation matrix. One popular approach has been to extract the minimum spanning tree (MST), which is the tree connecting all the elements in a system in such a way to maximize the sum of node similarities [53, 54]. Different correlation based networks can be associated with the same hierarchical tree, putting emphasis on different aspects of the sample correlation matrix. For instance, while the MST reflects the ranking of correlation coefficients, other methods, such as threshold methods, emphasize more the absolute value of each correlation coefficient. Researchers have also aimed to quantify the extent to which the behavior of one market, institution or asset can provide information about another through econometric studies [55], and by investigating Granger-causality networks [6] and partial-correlation networks [27].

In the context of financial markets, the correlation matrix among asset returns is an object of central importance in measuring risk. The filtering procedures described above may reveal statistically reliable features of the correlation matrix [46, 50], 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 [56, 57], especially those dynamics described by intraday patterns in volume, price and volatility [3, 58–60].

What is largely missing is an understanding of the drivers of these synchronous correlations, using the properties of the collective stock dynamics at shorter time-scales. Here, we apply a statistical methodology, detailed below, in order to study directed networks of lagged correlations among the 100 largest market capitalization stocks in the New York Stock Exchange (NYSE). In particular, we consider data from both the beginning of the previous decade and today. The resulting network representations of the systems provide insights into their underlying structure and dynamics. Our analysis reveals how the interplay of price movements at short time-scales evolves during a trading day, how it has changed over the past decade, and quantifies how it contributes to structural properties of the synchronous correlation matrix at longer time scales. We find striking periodicities in the validated lagged correlations, characterized by surges in network connectivity at the end of the trading day, which are crucial to account for when modeling equity price fluctuations. We show how these periodicities can refine our understanding of empirical phenomena, such as the Epps effect, and how they may be incorporated into regression models. We subject our analysis to a variety of robustness checks, which are detailed in section 3.8. Our analysis provides a deeper understanding of market risk by focusing on the short-term drivers of collective stock dynamics.

3.3 Methodology

At short time scales, measured synchronous correlations among stock returns tend to be lower in magnitude [17], and lagged correlations among assets may become non-negligible [61, 63]. 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 threshing 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 [63]. The SVN methodology provides the means to choose a statistically significant threshold for each variable independently, retaining information about the distribution of each individual time-series. We apply this methodology at different points in the trading day in order to explore the intraday pattern of collective stock dynamics.

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

$$r_i(t) = \log(P_i(t + \Delta t)) - \log(P_i(t)). \tag{3.1}$$

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 in signal processing of estimating the degree to which two series are correlated (see for example [12, 64–66]). The discrete lagged-correlation function between two time series X and Y is given by [67]

$$\rho_{X,Y}(d) = \frac{\sum_{i=1}^{N-d} \left[(X(i) - \langle X \rangle) \cdot (Y(i-d) - \langle Y \rangle) \right]}{\sqrt{\sum_{i=1}^{N-d} (X(i) - \langle X \rangle)^2} \cdot \sqrt{\sum_{i=1}^{N-d} (Y(i-d) - \langle Y \rangle)^2}} \tag{3.2}$$

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

In this work we focus on the returns matrix at the $\Delta t = 15$ minute time horizon, and divide each trading day into non-overlapping Δt parts $(\Delta t_1, \Delta t_2, ..., \Delta t_{26})$. We partition the contributions to each lagged correlation based on the period Δt_i, in order to explore seasonal effects in the data. For each time of day, we construct two matrices, A and B. For example, starting with the first 15 minutes of the day represented by Δt_1, then row m, column n of A is the return of stock n during the first 15 minutes (9:30 - 9:45am) of day m of the data. Row m, column n of B is the return of stock n during the second 15 minutes (9:45 - 10:00am) of day m of the data. So the number of rows of A or B is the number of days in the investigated dataset. We then calculate the lagged correlation matrix, where each entry (m, n) is the Pearson correlation coefficient of column m of matrix A with column n of matrix B. This process results in the empirical lagged correlation matrix, $C_{\Delta t_i}(m, n)$.

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

From the construction U we will associate a one-tailed p-value with all positive correlations as the probability to observe, by chance, a correlation which is equal to or higher than the empirically-measured correlation. Similarly, from D we will associate a one-tailed p-value with all negative correlations. In this analysis we choose our threshold to be $p = 0.01$. We must adjust our statistical threshold, however, to account for multiple comparisons. We use the conservative Bonferroni correction for N stocks, so that our new threshold is $0.01/N^2$. Thus, for a sample of $N = 100$ stocks, we construct 10^6 independently shuffled surrogate time series; if $U_{m,n} = 0$ we may associate a statistically-validated positive link from stock m to stock n ($p = 0.01$, Bonferroni correction). Likewise, if $D_{m,n} = 0$, we may associate a statistically-validated negative link from stock m to stock n. In this way we construct the Bonferroni network [47].

For comparison, for each part of day Δt_i we also construct the network using p-values that are corrected according to the False Discovery Rate (FDR) protocol. This correction is less conservative than the Bonferroni correction, and is constructed as follows. The p-values from each individual test are arranged in increasing order ($p_1 < p_2 < \cdots < p_{N^2}$), and the threshold is defined as the largest k such that $p_k < k\ 0.01/N^2$. Therefore, for the FDR network, our threshold for the matrices U (or D) is not zero but instead is the largest integer k such that U (or D) has exactly k entries less than or equal to k. From this threshold we may filter the links in C to construct the FDR network [47].

3.4 Intraday seasonalities

This approach, in which we construct a distinct network for each interval of Δt minutes between 9:30am and 4:00pm, 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 seasonal patterns at the daily level. These seasonalities can be important features to account for when modeling stock price movements.

Figure 1 displays the intra-day pattern of the average synchronous correlation between returns of all stock pairs in the top 100 most capitalized stocks traded on the NYSE. Prices are sampled at a time resolution of $\Delta t = 15$ min. We include results for data from the time period 2001-2003, as well as 2011-2013, where we observe striking changes over the past decade in the magnitude of the measured correlations. Both periods exhibit a similar profile in the intra-day pattern of synchronous correlations, with an explosive growth in the first hour of the trading day that levels in the late morning, followed by a steady increase in the afternoon. A similar profile has been observed in other studies [3].

We use the statistical methodology introduced above to construct an analogous profile for lagged correlations. In Figure 3.3 we plot the average lagged correlation between the same stock pairs from Figure 1. Prices are again sampled at a time resolution of $\Delta t = 15$ min., with correlations evaluated at one sampling time horizon. We find that, although the distributions of lagged correlation coefficients are on average quite small, there exist pairs of stocks in the tails of these distributions that represent a statistically-significant lagged correlation, in the sense of the methodology described above. These stock pairs form the links in a series of statistically-validated networks. We plot the intra-day pattern of lagged correlations for the stock pairs belonging to the Bonferroni network in red, and the FDR

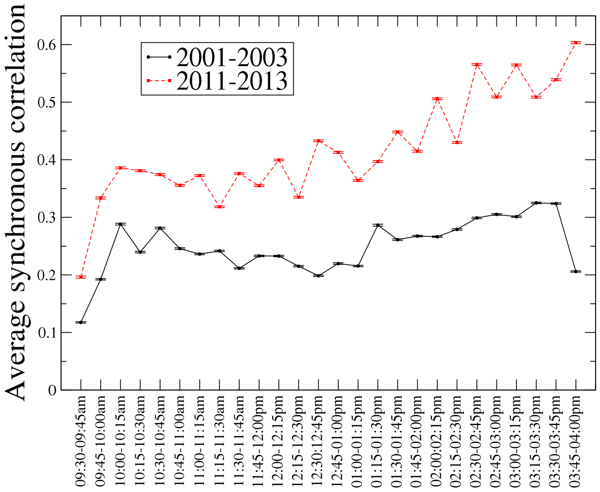


Figure 3.2: Intra-day pattern of the average synchronous correlation between fifteen minute stock returns of the 100 most capitalized stocks traded at NYSE in the period 2001-2003 (black continuous line) and 2011-2013 (red dashed line).

network in blue. In both the data from 2001-2003 and 2011-2013 we find that the bulk of the lagged correlations tends to shift to the positive regime during the final minutes of the trading day.

The positive shift in the bulk of the lagged correlation coefficients manifests as an increase in network connectivity. In Figure 2 we display visualizations of the Bonferroni networks for both the beginning and end of the trading day for the period 2001-2003. We include the corresponding visualizations for the 2011-2013 data in section 3.8. In both periods we observe an explosive growth in the significance of positive lagged correlations during the final minutes of the trading day, underscoring dramatic seasonal effects in the co-movements of asset prices. Despite these effects, we find that the validated links are largely persistent throughout the trading day, as detailed in section 3.8.

The positive shift in the bulk of the lagged correlation coefficients manifests as an increase in network connectivity. In Figure 2 we display visualizations of the Bonferroni networks for both the beginning, middle and end of the trading day for the period 2001-2003. We include the corresponding visualizations for the 2011-2013 data in section 3.8. In

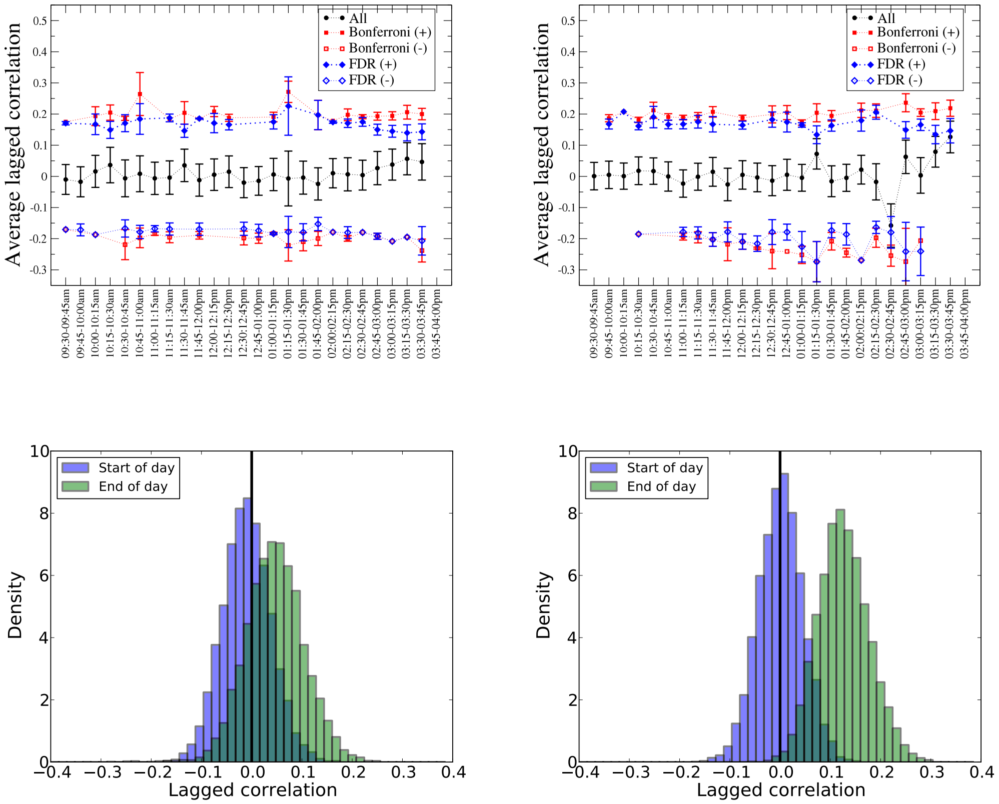


Figure 3.3: Intra-day pattern of the average lagged correlation, evaluated at one lag, between fifteen minute stock returns of the 100 most capitalized stocks traded at NYSE in the period 2001-2003 (top left panel) and 2011-2013 (top right panel). In each panel, we also report the pattern of lagged correlation with average taken over all the links that belong to the Bonferroni network (red squares) and the FDR network (blue diamonds), by distinguishing between positive (+) and negative (-) statistically validated correlations. We also provide normalized histograms of all $N^2 = 10,000$ lagged correlation coefficients for two intraday periods in 2001-2003 (bottom left panel) and 2011-2013 (bottom right panel). The blue shaded histogram corresponds to correlations between returns in the first 15 minutes of the trading day (9:30am to 9:45am) and those in the second 15 minutes (9:45am to 10:00am). The green shaded histogram corresponds to correlations between returns in the second-to-last 15 minutes of the trading day (3:30pm to 3:45pm) and those in the last 15 minutes (3:45pm to 4:00pm). We observe a characteristic positive shift in the lagged correlations in the final minutes of the trading day.

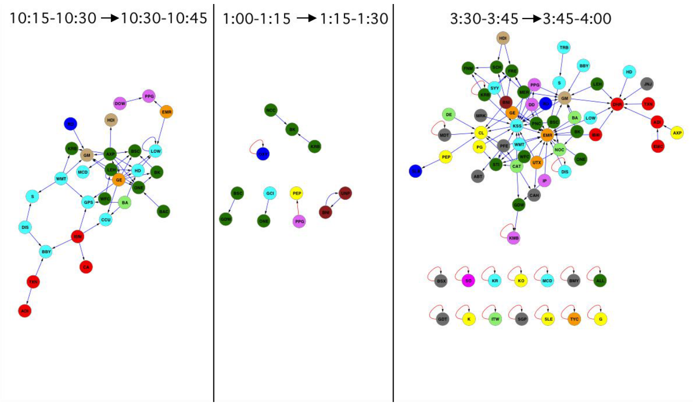


Figure 3.4: Visualization of the Bonferroni networks from periods in the beginning, middle, and end of the trading day in the period 2001-2003. The corresponding visualizations for the 2011-2013 data are included in section 3.8. Stocks are colored by their economic sector. Links of positive correlation are colored blue, while links of negative correlation are colored red.

both periods we observe a decrease in connectivity during the middle of the trading day, followed by an explosive growth in the significance of positive lagged correlations during the final minutes of the trading day, reminiscent of the well-known U-shaped pattern in intraday transaction volume and volatility [58, 60]. Our analysis underscores dramatic seasonal effects in the co-movements of asset prices. Despite these effects, we find that the validated links are largely persistent throughout the trading day, as detailed in section 3.8.

3.5 Reconstructing the Epps Effect

These seasonal effects are crucial to take into account when modeling collective stock dynamics. Here we investigate the impact of high-frequency lagged cross correlations and autocorrelations of returns on synchronous correlations between stock returns evaluated at a larger time horizon. In particular, we retain information on the intraday period when measuring how these lead-lag relationships at short timescales may influence synchronous co-movements among equities at longer timescales. We derive an equation, obtained by taking an approach similar to the one presented in ref. [61], in which we show how the synchronous correlation between two stock returns time series, as evaluated at a certain intraday time window, e.g., the first 130 minutes of the trading day, can be decomposed in order to make apparent the individual contribution of auto-correlations and lagged cross-correlations evaluated at smaller time windows, such as $\Delta t = 5$ minutes. The only assumption we make to obtain that equation is that the intraday volatility pattern $\sigma_i^2(q, \Delta t)$ of a stock i, where q indicates the intraday-time and Δt the time horizon, can be written as an idiosyncratic constant k_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: $\sigma_i^2(q, \Delta t) = k_i \cdot f_q(\Delta t)$.

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

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

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

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

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

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

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

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

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

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

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

It is well known that there is an intraday pattern of volatility, which is common to all the stocks [62]. This means that, without introducing a large error, we can set:

$$\sigma_x(j) = k_x \cdot f(j);\;\; \sigma_y(j) = k_y \cdot f(j),\;\; \forall j = 1, ..., p \tag{3.3}$$

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:

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

where Eq. (3.3) has been used to describe the intra-day pattern of volatility. Similarly, we obtain the variance of y:

$$\begin{aligned}
\sigma_y^2 = k_y^2\Bigg[&\sum_{j=1}^{p} f(j)^2 + 2\sum_{j=1}^{p-1} f(j)\, f(j+1)\rho_{y_j,y_{j+1}} + 2\sum_{j=1}^{p-2} f(j)\, f(j+2)\rho_{y_j,y_{j+2}} + ... \\
&+ 2\, f(1)\, f(p)\rho_{y_1,y_p}\Bigg].
\end{aligned}$$

The covariance of x and y is then:

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

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

$$\rho_{x,y} = \frac{\Big[\sum_{j=1}^{p} f(j)^2 \rho_{x_j,y_j}\Big] + \Big[\sum_{i=1}^{p-1}\sum_{j=1}^{p-i} f(j) f(j+i)(\rho_{x_j,y_{j+i}} + \rho_{x_{j+i},y_j})\Big]}{\sqrt{\Big(\sum_{j=1}^{p} f(j)^2 + 2\sum_{i=1}^{p-1}\sum_{j=1}^{p-i} f(j)\, f(j+i)\rho_{x_j,x_{j+i}}\Big)}\cdot \sqrt{\Big(\sum_{j=1}^{p} f(j)^2 + 2\sum_{i=1}^{p-1}\sum_{j=1}^{p-i} f(j)\, f(j+i)\rho_{y_j,y_{j+i}}\Big)}}$$

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

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

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

For illustration, consider the first 30 minutes of the trading day, and suppose we are interested in the synchronous correlation coefficient $\rho_{x,y}$ between the time series x and y, such that $\{x\} = \{x(1), x(2), ..., x(T)\}$ and $\{y\} = \{y(1), y(2), ..., y_(T)\}$, where T is the number of trading days in the dataset, and $x(i)$ and $y(i)$ represent the return of stock i and stock j, respectively, in the first 30 minutes of day i. Each one of these time series of log-returns can be decomposed in the sum of $p = 2$ time series of log-returns, specifically, the time series of returns in the first $p = 2$ intraday time intervals of $\Delta t = 15$ minutes:

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

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

$$\rho_{x,y} = \frac{f_1^2\, \rho_{x_1,y_1} + f_2^2\, \rho_{x_2,y_2} + f_1\, f_2\, (\rho_{x_1,y_2} + \rho_{x_2,y_1})}{\sqrt{[f_1^2 + f_2^2 + 2 f_1\, f_2\, \rho_{x_1,x_2}]\, [f_1^2 + f_2^2 + 2 f_1\, f_2\, \rho_{y_1,y_2}]}}. \tag{3.4}$$

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 [17]: $\rho_{x,y} > \max(\rho_{x_1,y_1}, \rho_{x_2,y_2})$. It is also worthwhile to point out that the correlation coefficient $\rho_{x,y}$ does not depend on quantities related to other stocks in the system. Therefore, structural properties of the correlation matrix, such as the fact that it should be positive semi-definite, are not forced by our reconstruction equation. In Fig. 3.5, we show some results of the reconstruction analysis of the 100 stock correlation matrix for the two time periods under investigation, 2001-2003 (left panel) and 2011-2013 (right panel). We have divided the trading day in three time windows of 130 minutes each, from 9:30am to 11:40am (top panels), from 11:40am to 1:50pm (mid panels), and from 1:50pm to 4:00pm (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 autocorrelations 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 autocorrelations and lagged cross-correlation in the reconstruction formula, we also neglect the intraday volatility pattern. The curves are obtained by comparing the reconstructed correlation matrix C_{rec} and the original correlation matrix C_{or} through the standard Frobenius norm:

$$F(C_{or}, C_{rec}) = \sqrt{\mathrm{tr}\left[(C_{or} - C_{rec})(C_{or} - C_{rec})^T\right]}, \tag{3.5}$$

where tr[·] is the trace operator, and apex T indicates the transpose operator. The results obtained for the 2001-2003 time period (left panels) indicate that lagged cross-correlations contribute more to synchronous correlations than autocorrelations in all the three time windows, although such a contribution tends to decrease during the day. On the other hand, in the 2011-2013 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 2001-2003 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 2011-2013 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 2011-2013 correlation matrices in the presence of noise, and explore this interpretation in section 3.8.

The presented analysis shows that, in the period 2001-2003 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 2011-2013, 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.

3.6 Regression model

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

We show that, at each intraday period, the relevant inputs to a model of the returns of stock i can be reliably taken as the set of direct predecessors $\{\nu_j\}$ of the corresponding node in the validated network for that period. That is, we need only consider a node j as an input to the model if there is a link from j to i. To demonstrate this, for each intraday period we attempt to model the returns of stocks with an in-degree of at least one with a simple linear model. If we represent column i of matrices A or B from the methodology section with A_i or B_i, then we fit

$$B_i = \beta_0 + \beta_1 A_{\nu_1} + \beta_2 A_{\nu_2} + \cdots + \beta_{k_i} A_{\nu_{k_i}} + \epsilon \tag{3.6}$$

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

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

$$\mathrm{BIC}_i = (k_i + 1)\ln(T) - 2\ln(L_i) \tag{3.7}$$

where T is the number of rows in A and B, equal to the number of days in the analysis, and L_i is the maximized likelihood for the model in equation (3.6). The BIC is a criterion for model selection, and can be interpreted as an anticipation of a model's out-of-sample performance using only in-sample training data.

We compare the measured BICs to a randomised model, in which for each node i we randomly select k_i of the $N = 100$ available nodes as regressors in equation (3.6). This procedure is repeated 100 times for each model. In Figure 3.6, for both the 2001-2003 and

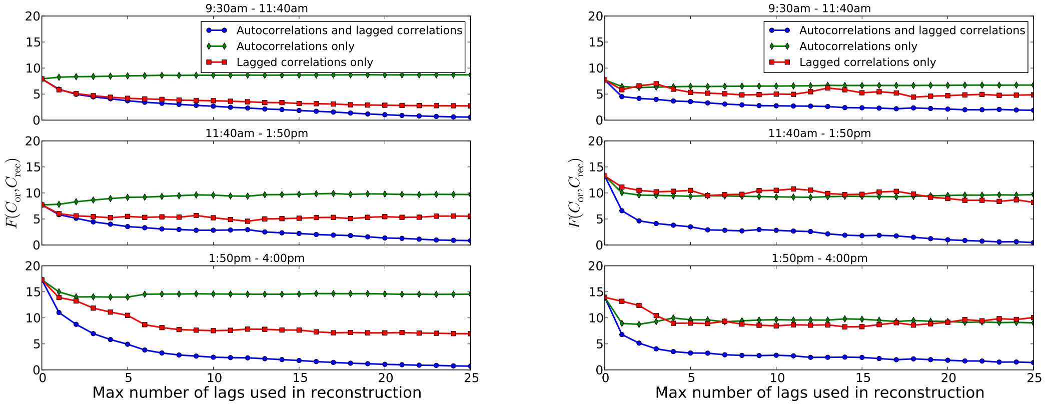


Figure 3.5: Frobenius distance between the 130 minute return correlation matrix of the 100 most capitalized stocks traded at NYSE, C_{or}, and the corresponding correlation matrix, C_{rec}, reconstructed according to the method described in the text in the time period 2001-2003 (left panels) and 2011-2013 (right panels), in the three 130 minute segments of the trading day: from 9:30am to 11:40am (top panels), from 11:40am to 1:50pm (middle panels), and from 1:50pm to 4:00pm (bottom panels). 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.

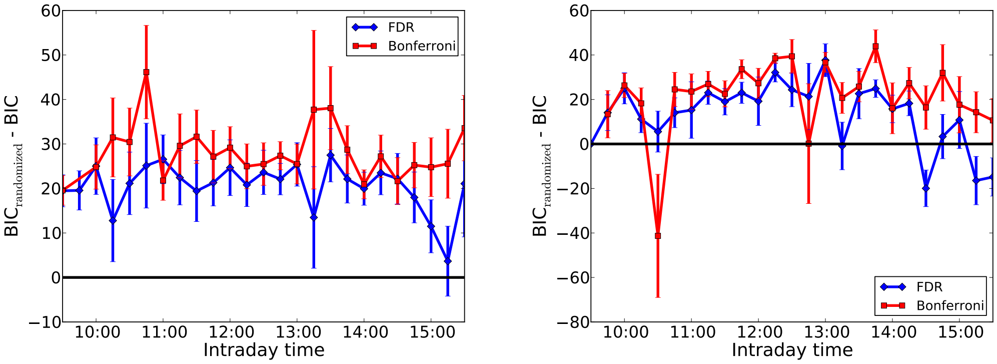


Figure 3.6: Difference in BICs between the models in equation (3.6) and the randomized models described in the text, for both the periods 2001-2003 (left panel) and 2011-2013 (right panel). We generate 100 realisations of the random model for each stock. Points show the mean BIC deviation of all stocks from the mean BIC of the corresponding randomised models. Error bars show the uncertainties in this deviation for all models, added in quadrature.

2011-2013 datasets, we plot the mean difference in BICs for all models. With the exception of one period in the 2011-2013 dataset, the specification of model inputs using the Bonferroni network always outperforms the randomised specification. The specification using the FDR network fares similarly, although it fails to outperform the randomised specification in one period in the 2001-2003 dataset and five periods in the 2011-2013 dataset. These periods fall at the end of the trading day, when, due to the large numbers of of validated links, the relative advantage of the validated networks in feature selection diminishes against a random selection of inputs.

3.7 Discussion

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

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

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

3.8 Robustness Checks

3.8.1 Visualization of 2011-2013 networks

In Figure 3.7 we display visualizations of validated networks using the 2011-2013 data, where we observe qualitatively the same effect as in the 2001-2003 data.

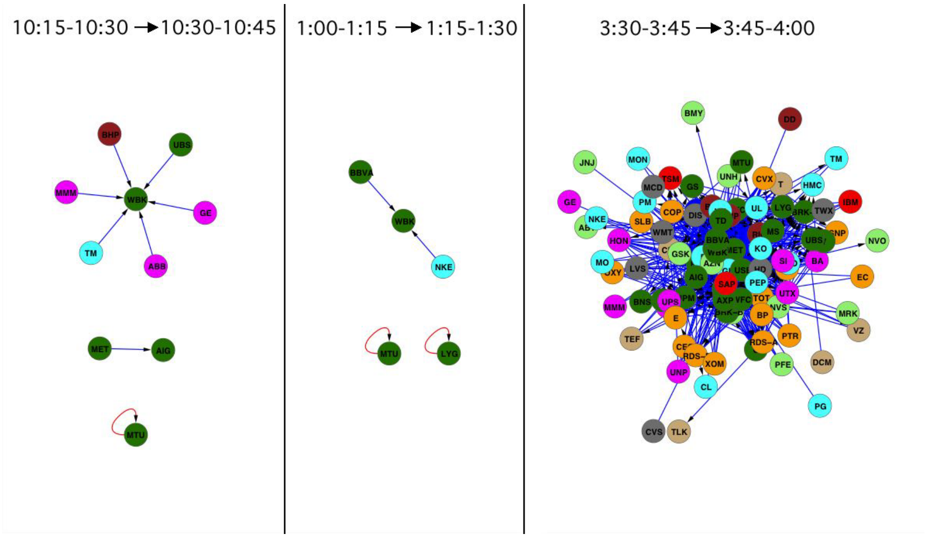


Figure 3.7: Visualization of the Bonferroni networks from periods in the beginning, middle, and end of the trading day in the period 2011-2013. Stocks are colored by their economic sector. Links of positive correlation are colored blue, while links of negative correlation are colored red.

3.8.2 Contribution of high volatility period to lagged correlations

The months of August to October 2011 witnessed a volatile period in U.S. stock exchanges. Here we examine the influence of this period on the results presented in the text. We may quantify the contribution of each day in the data to the average lagged correlation in each intraday period as follows. In analogy with equation (3.1), we may write the mean lagged correlation as averaged overall all N^2 stock pairs as the sum:

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

with $\langle C \rangle_t$ the defined as the term in brackets in the second line. The sum of these terms is then the average lagged correlation associated with each intraday period. We plot the time-series of these terms for each intraday period in Figure 3.8.

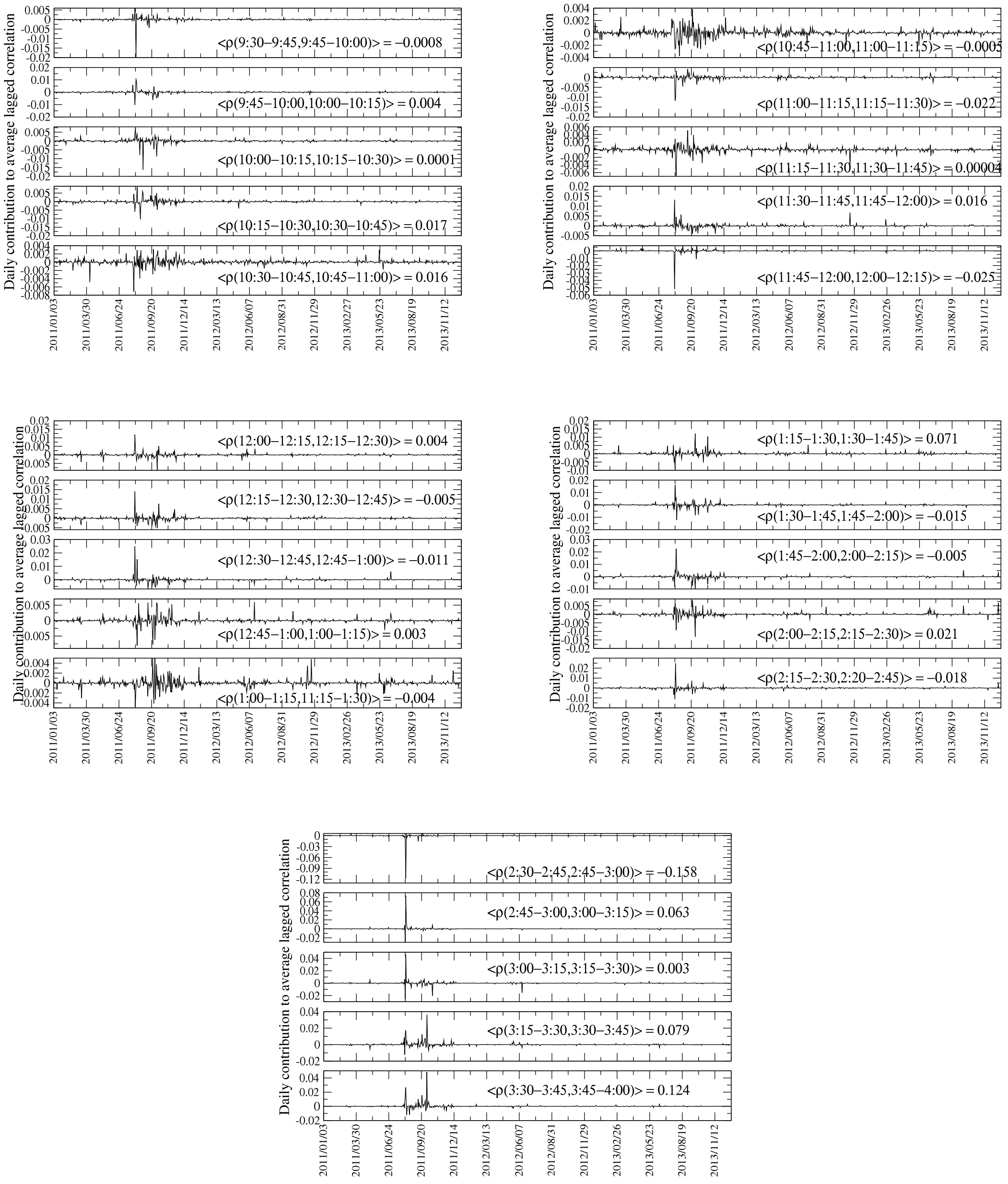


Figure 3.8: Contributions $\langle C \rangle_t$ of each day t in the 2011-2013 data to the mean lagged correlation measured for each intraday period. Each row of each subfigure corresponds to a lagged correlation between two consecutive intraday periods. Inset provides the mean lagged correlation as averaged over all stock pairs.

The period of August through October 2011 appears as a volatile portion of the time series for each intraday period. The contribution of this period is particularly pronounced toward the end of the trading day, where a small number of days seem to contribute disproportionately to the average lagged correlation. We therefore remove all days in August, September, and October 2011 to test the robustness of our results when excluding periods of financial crisis. In Figure 3.9 we compare the numbers of validated positive and negative links using all available days in the data with those excluding the period August-October 2011. We find that the influence of this volatile period on the statistically-validated networks is largest at the end of the trading day, and that the lagged relationships uncovered by the analysis are otherwise robust. This is corroborated by Figure 3.10, where we see that the characteristic positive shift in the distribution of lagged correlations at the end of the trading day is weakened upon excluding the months of August through October 2011.

We additionally examine the effect of this period on the reconstruction analysis presented in the text. In Figure 3.11 we display the results of the reconstruction analysis for the 2011-2013 data both including and excluding the months of August through October 2011. We again find that the effect of these months is most pronounced at the end of the trading day, from 1:50pm to 4:00pm. We also see that, while this period contributed disproportionately to the measured lagged cross-correlations, it had little effect on the measured autocorrelations, which continue to contribute to the reconstructed 195min. synchronous correlation.

3.8.3 Stability of reconstructed correlation matrices to noise

Here we provide a brief explanation of the structural problems uncovered in the reconstructed correlation matrices in 2011-2013. If we constrain ourselves to use only autocorrelations or lagged cross-correlations in the reconstruction analysis, then most matrices in this period are not positive definite as they have some number of negative eigenvalues. On the other hand, all reconstructed correlation matrices in the period 2001-2003 have positive eigenvalues.

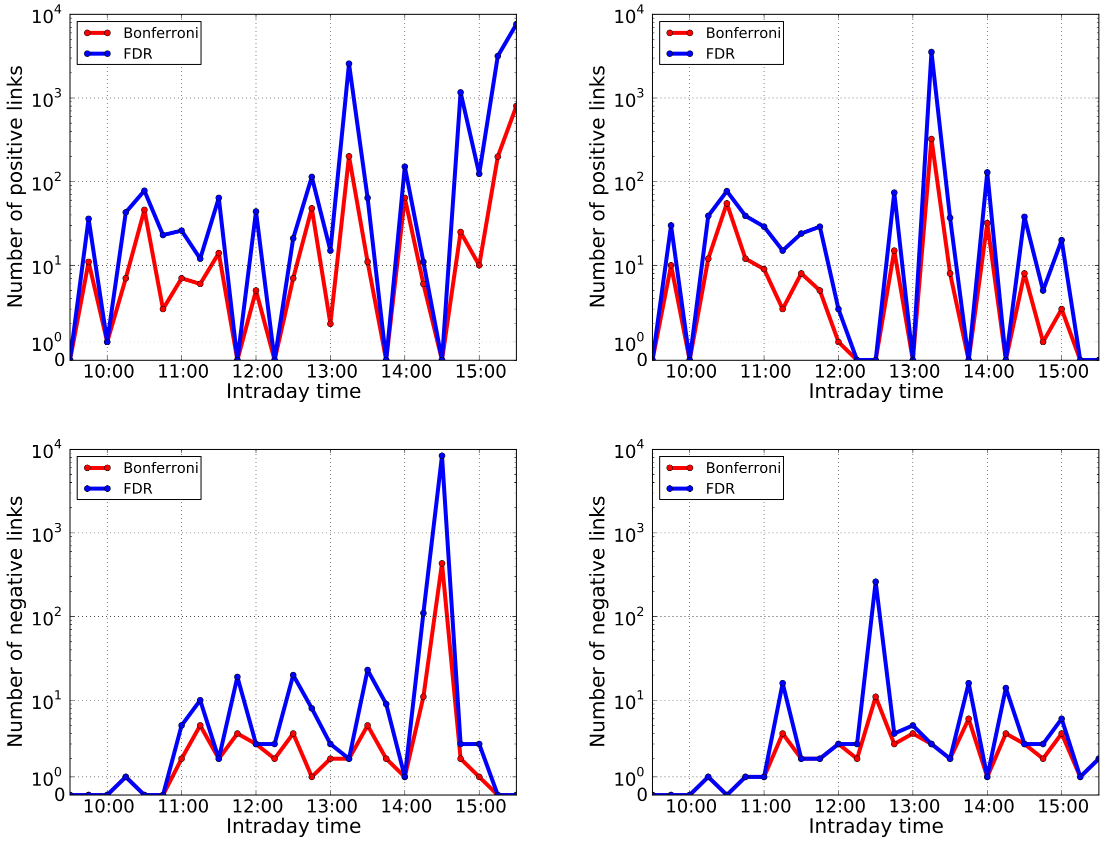


Figure 3.9: Top row: number of validated positive links in the 2011-2013 data for all days (left) and after removal of August-October 2011 (right). Bottom row: number of validated negative links in the 2011-2013 data for all days (left) and after removal of August-October 2011 (right).

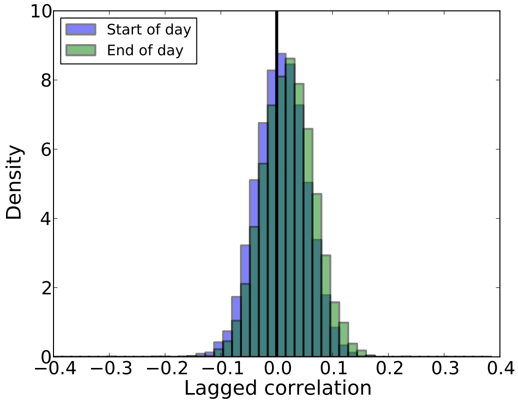


Figure 3.10: Normalized histograms of all $N^2 = 100^2 = 10,000$ lagged correlation coefficients for two intraday periods in 2011-2013, excluding the months of August through October 2011. The blue shaded histogram corresponds to correlations between returns in the first 15 minutes of the trading day (9:30am to 9:45am) and those in the second 15 minutes (9:45am to 10:00am). The green shaded histogram corresponds to correlations between returns in the second-to-last 15 minutes of the trading day (3:30pm to 3:45pm) and those in the last 15 minutes (3:45pm to 4:00pm). The characteristic positive shift in the lagged correlations in the final minutes of the trading day has weakened upon excluding the months of August through October 2011.

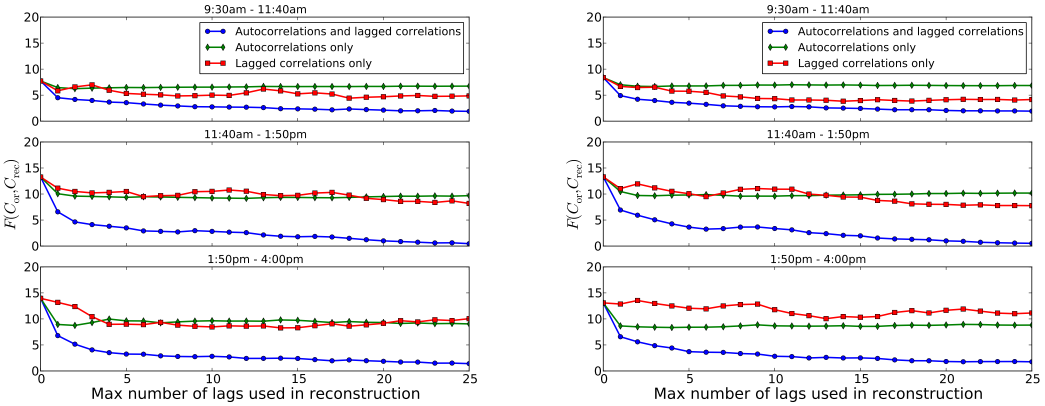


Figure 3.11: Frobenius distance between the 130 minute return correlation matrix of the 100 most capitalized stocks traded at NYSE, C_{or}, and the corresponding correlation matrix, C_{rec}, reconstructed according to the method described in the text in the three 130 minute segments of the trading day: from 9:30am to 11:40am (top panels), from 11:40am to 1:50pm (middle panels), and from 1:50pm to 4:00pm (bottom panels). All data are from 2011-2013. In the left panel we show results using the entire period, while in the middle panel we exclude the months of August through October 2011. 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.

We illustrate this increased "fragility" of the 2011-2013 correlation matrices in Figure 3.12. In this analysis we perturb the 130 min. correlation matrices from each portion of the trading day with a given level of noise. For a noise level x, each symmetric pair of off-diagonal elements (i, j) and (j, i) are perturbed by a number from a uniform distribution on the interval $[-x, x]$. We then measure the probability of observing at least one negative eigenvalue in each matrix through 1000 independent perturbations. In Figure 3.12 we compare results from 2001-2003 with those from 2011-2013, and also show the contribution of the months of August, September and October 2011 by removing it from the analysis (right panel).

We find that the structural properties of the correlation matrices obtained in the period 2001-2003 are significantly more robust than those obtained in 2011-2013. This analysis complements the observation presented in the main text, that the 130 min. correlation matrices reconstructed without contributions from 5 min. lagged cross-correlations or autocorrelations are not always positive definite. Owing in part to an increased level of synchronous correlation, there are tighter bounds constraining each element of the 2011-2013 correlation matrices. Given a noise level, these bounds are more easily violated than in the 2001-2003 data.

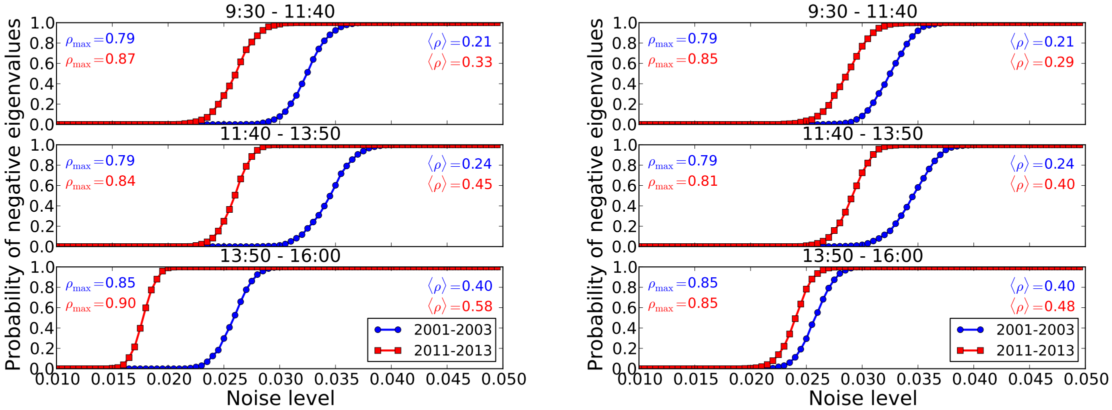


Figure 3.12: Probability of observing at least one negative eigenvalue in each 130 min. correlation matrix after perturbing correlation matrices with a given level of noise. In the right panel we exclude the months of August, September and October 2011 from the analysis. For a noise level x, each symmetric pair of off-diagonal elements (i, j) and (j, i) are perturbed by a number from a uniform distribution on the interval $[-x, x]$. Data from 2001-2003 are shown in blue, while data from 2011-2013 are shown in red. We also show the mean and maximum off-diagonal correlation values from each matrix. We observe that the 2011-2013 data exhibits negative eigenvalues at a consistently lower noise level than the 2001-2003 data. Each probability is evaluated through 1000 independent perturbations of the matrix. In addition, the 2011-2013 data has four pairs of stocks that represent the same firm: BRK-A and BRK-B, RDS-A and RDS-B, BHP and BBL, UN and UL. These stocks have very high synchronous correlations, so we exclude BRK-B, RDS-B, BBL and UL from the analysis. Including them does not qualitatively change the results, but exaggerates the observed pattern.

3.8.4 Persistence of links

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

We find that the validated links are indeed largely persistent throughout the trading day, although they are more strongly dependent on the particular intraday period in the 2001-2003 data. We support this finding with two analyses. First, we may quantify the extent to which two networks share links using the Jaccard Index:

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

where L_i is the set of links in network i. We distinguish edges by both direction and sign when constructing these sets. A high value of the Jaccard Index, in this context, indicates that two networks share a large proportion of their total links. In Figure 3.13 we display matrices of Jaccard Indices $J(i,j)$ between sets of links corresponding to networks for all intraday periods at a time horizon $\Delta t = 15$ min. We find that the Jaccard Indices are generally high, suggesting that the links we validate are indeed persistent across many time periods, although this effect is weaker in the 2001-2003 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 2011-2013 data. Finally, we have verified that these plots are only very weakly affected by the turmoil of August - October 2011, as the corresponding diagrams for the networks that were constructed with this period removed are similar.

The analysis in Figure 3.13 quantifies a degree of similarity among intraday periods. We can also examine this similarity at the level of individual links, by quantifying the persistence of links. This persistence is defined as the fraction of intraday networks (of which there are 25 for $\Delta t = 15$ min.) in which a given link appears. We plot the distributions of link persistence for all networks in Figure 3.14, where we observe again from this perspective

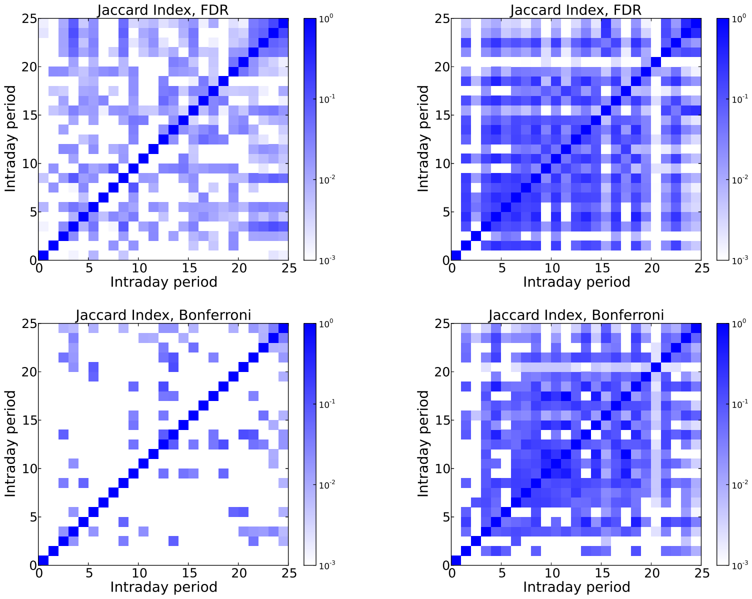


Figure 3.13: Matrices of Jaccard Indices between sets of links corresponding to networks for all intraday periods at a time horizon $\Delta t = 15$ min. Left column shows results using data from 2001-2003 (FDR and Bonferroni networks); right column shows results using data from 2011-2013 (FDR and Bonferroni networks). We find that the validated links are generally more persistent in the 2011-2013 data throughout the trading day.

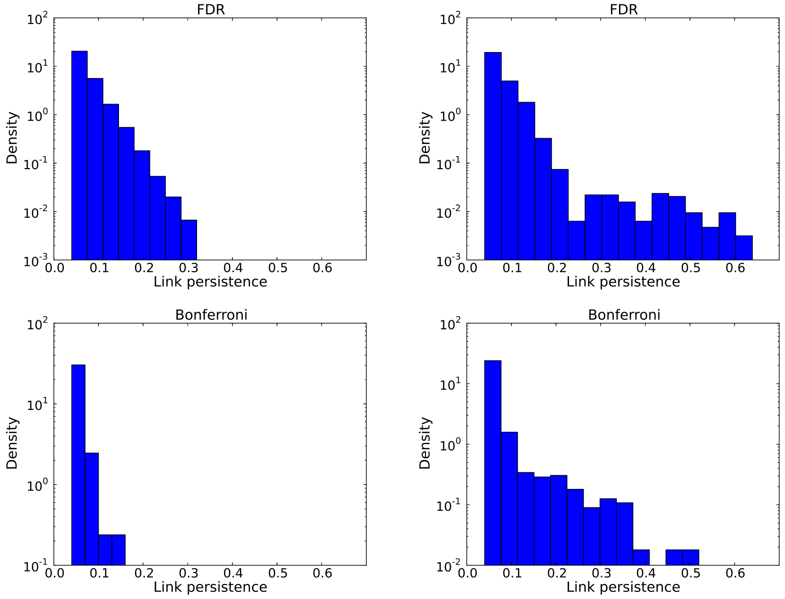


Figure 3.14: Distributions of link persistence for all links in networks at a time horizon $\Delta t = 15$ min. Left column shows results using data from 2001-2003 (FDR and Bonferroni networks); right column shows results using data from 2011-2013 (FDR and Bonferroni networks). We find that the validated links are generally more persistent in the 2011-2013 data throughout the trading day.

that individual links seem to be more persistent in the 2011-2013 data (although, again, this analysis does not convey information regarding the number or strength of the validated links).

3.8.5 Influence of autocorrelations on linear models

To examine the influence of autocorrelations on the performance of the linear models described in the text, we repeat the analysis with validated autocorrelation links removed. That is, the model for each node i has k_i inputs, with k_i the in-degree of node i, disregarding autocorrelation links. As in the text, we compare the BICs of these models with those obtained from randomly selecting k_i of the $N = 100$ possible input nodes as regressors in the model. In Figure 3.15, for both the 2001-2003 and 2011-2013 datasets, we plot the mean difference in BICs for all models. The results highlight the elevated influence

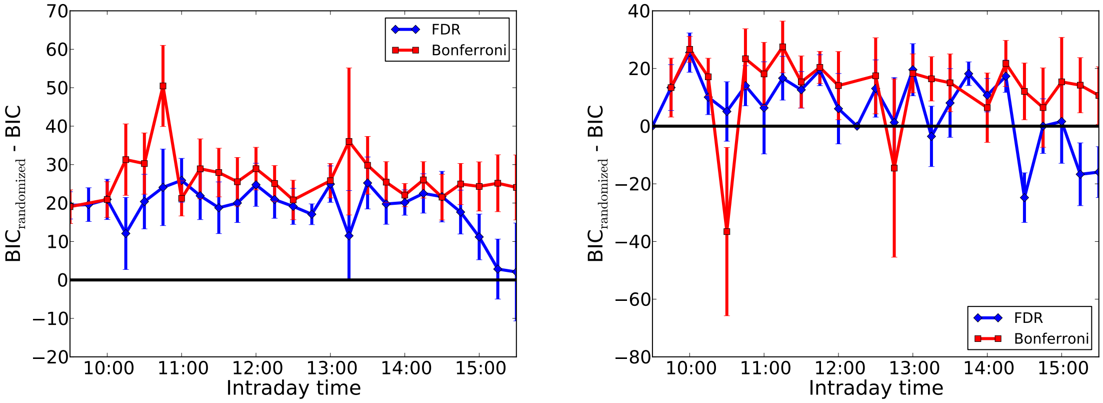


Figure 3.15: Difference in BICs between the linear models with inputs prescribed by the validated network and the randomized models described in the text, for both the periods 2001-2003 (left panel) and 2011-2013 (right panel), upon removal of autocorrelation links. We generate 100 realizations of the random model for each stock. Points show the mean BIC deviation of all stocks from the mean BIC of the corresponding randomized models. Error bars show the uncertainties in this deviation for all models, added in quadrature.

of autocorrelations in the recent data: whereas the models in 2001-2003 continue to outperform the randomized models, in 2011-2013 the model performance is markedly worse if autocorrelations are ignored.

3.8.6 Partial lagged correlation networks

The reconstruction analysis presented in the text reveals how both autocorrelations and lagged correlations at a given time horizon compete to form synchronous correlations among stock returns evaluated at a larger time horizon. In the 2011-2013 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 2001-2003 data, and (ii) many statistically significant autocorrelations are observed in the 2011-2013 data, while fewer are observed in the 2001-2003 data. These two observations have the potential to explain the presence of a large number of statistically validated lagged correlations in the 2011-2013 dataset, and could also explain the tight connection between autocorrelations and lagged cross-correlations mentioned above. That is, a lagged cross-correlation between two stock returns $\rho(x(t), y(t+\tau))$ may just reflect the presence of autocorrelation of stock return x, $\rho(x(t), x(t+\tau))$ and the synchronous correlation between stock returns x and y, $\rho(x(t+\tau), y(t+\tau))$. Similarly, we could consider the autocorrelation of returns in stock y, $\rho(y(t), y(t+\tau))$ and the synchronous correlation $\rho(x(t), y(t))$.

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

$$\rho(x(t), y(t+\tau)|y(t)) = \frac{\rho(x(t), y(t+\tau)) - \rho(y(t), y(t+\tau))\rho(x(t), y(t))}{\sqrt{[1-\rho(y(t), y(t+\tau))^2][1-\rho(x(t), y(t))^2]}}, \text{ and} \quad (3.8)$$

$$\rho(x(t), y(t+\tau)|x(t+\tau)) = \frac{\rho(x(t), y(t+\tau)) - \rho(x(t+\tau), y(t+\tau))\rho(x(t), x(t+\tau))}{\sqrt{[1-\rho(x(t+\tau), y(t+\tau))^2][1-\rho(x(t), x(t+\tau))^2]}}, \quad (3.9)$$

subtracting off the influence of autocorrelations.

We thus repeat the statistical validation procedure, using the same shuffling procedure described in the text, with a matrix of lagged partial correlations in place of the lagged correlation matrix (considering only the off-diagonal elements, as the diagonal elements of this partial correlation matrix are undefined). We build separate networks for partial correlations given by (3.8) and (3.9), again choosing our statistical threshold to be $p = 0.01$.

We report results for $\Delta t = 15$ min. in the last time horizon of the trading day, when we find the strongest autocorrelations. Using the Bonferroni correction for multiple comparisons, we validate 448 positive links using the partial correlation matrix (3.8), and 313 positive links using the matrix (3.9). We validate no links of negative correlation. Using the original lagged correlation matrix, we validate 91 positive links and 18 negative links. Because the autocorrelations are negative, we validate many more links in the partial lagged correlation networks; that is, the original lagged correlation networks contain many positive links in spite of the negative correlations, and not because of them. We note that the partial lagged correlation networks using the matrices (3.8) and (3.9) share an intersection of 77 and 83 links, respectively, with the original network. The probability of randomly sampling these intersections x from the $L = 100 \times 99 = 9900$ total possible lagged cross-correlation links in $n = 91$ “draws” (links in the original network) is given by the hypergeometric distribution:

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

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

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

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

We repeat the same procedure on the 2011-2013 data, validating 629 positive links using the partial correlation matrix (3.8), and 831 positive links using the matrix (3.9). We validate no links of negative correlation. Using the original lagged correlation matrix, we validate 801 positive links and no negative links. We note that the partial lagged correlation networks using the matrices (3.8) and (3.9) share an intersection of 295 and 374 links, respectively, with the original network. Again, we may associate a p-value to these intersections using the hypergeometric distribution, which is vanishingly small both networks.

Chapter 4

Community structures in lagged correlation networks and their relevance to feature selection

In this chapter we review a method for identifying communities of nodes that cluster, or share many neighbors. We focus on the complexities that arise when considering directed networks. We then consider a particular form of community that appears in much of the financial data under consideration, and show how to expose these communities using a spectral clustering method. The particular method we employ involves a singular value decomposition of the adjacency matrix. We then provide an argument as to why, in low signal-to-noise environments, these communities provide relevant information when constructing statistical models of the multivariate time series.

We test these ideas in the context of a particular problem: studying the interplay of news and market movements. Our data set consists of returns data from major stock indices in 40 countries, in conjunction with news sentiment time series for the same markets, provided by Thompson Reuters. After rendering the investigated time series stationary, we study the structure of the synchronous correlation matrix, finding that the markets form the "backbone" of the network. We then apply the statistically validated network methodology in order to investigate the extent to which news lead markets, and the extent to which markets anticipate news. We find that the latter effect is much more pronounced in the data.

Finally, using the community-detection procedures described earlier in the chapter, we investigate large-scale flows of information among geographic regions. We find several pronounced large-scale structures, many of which supplement studies from the econometrics literature. We then show how, in such an experimental setting, the identified communities can aid in the construction of more robust statistical models by forming the basis of a recommender-system for model inputs. That is, false positive links, for example, can be highlighted and removed using a simple methodology that we introduce. We confirm this notion using out-of-sample test results from several classes of predictive models, and using both empirical and synthetic data.

4.1 Bipartite communities in directed networks

Clustering in networks is commonly studied using a spectral decomposition of the underlying adjacency matrix. In the case of a symmetric matrix with undirected links, as in a network defined by synchronous correlations, an eigendecomposition of the matrix A or its Laplacian can reveal groups of nodes that cluster together, in the sense of sharing many links (Chung, 1997). The interpretation of the eigenvectors and eigenvalues is less straightforward in directed networks, as the adjacency matrix A is asymmetric and we will generally obtain complex eigenvalues and eigenvectors. The Singular Value Decomposition (SVD), however, has been shown to be a simple method to reveal clustering in even directed graphs (Drineas et al., 2004). The SVD is a matrix factorization of the form

$$A = U\Sigma V^{\dagger}$$

where, in the special case of an $N \times N$ matrix A, U is an $N \times N$ unitary matrix composed of the eigenvectors of AA^T, and $V^{\dagger}$ is the conjugate transpose of an $N \times N$ unitary matrix V, whose columns are composed of the eigenvectors of $A^T A$. Σ is an $N \times N$ diagonal matrix with entries σ_n that are the real square roots of the eigenvalues of U or V. The columns of U and V are known as the left- and right-singular vectors of A, respectively, and the diagonal entries σ_n of Σ are known as the singular values of A.

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

Taylor et al. (2011) prove, in idealized cases of networks composed entirely of fully-connected non-overlapping bipartite structures, that each pair of left- and right-singular vectors corresponds to a bipartite subgraph: the nonzero entries of the left-singular vector are nodes in one layer of the bipartite structure; the nonzero entries of the right-singular vector are nodes in the second layer of the structure, and edges are drawn from the nodes in the left-singular vector to the nodes in the right-singular vector. Furthermore, each singular value gives the geometric mean of the number of nodes represented in the corresponding left- and right-singular vectors. This holds exactly for the highly-idealized situation described above, but is fairly robust in the presence of noise, such as missing edges or overlapping bipartite structures (Taylor et al., 2011).

The robust nature of this spectral clustering method affords it much popularity in the study of recommender systems. Consider, for example, a set of consumers and a set of goods. Online marketplaces, such as Amazon, or media providers, such as Netflix, often collect extensive information on which consumers view or purchase which goods. These data can be interpreted as a bipartite network, in which one layer of the network represents the consumers, and the other layer represents the goods that they are interested in. A singular value decomposition of the corresponding adjacency matrix will reveal groups of people who are interested in the same goods, and is robust to certain “missing links” in these substructures. The principle of collaborative filtering suggests that these missing links can be used to recommend goods to consumers: if Alice is has expressed interest in physics books, and other people who express interest in physics books also tend to express interest in certain math books as well, then it stands to reason that Alice might also be interested in those math books.

In this chapter we will engage with such bipartite substructures of directed networks in two ways. First, given an arbitrary directed network, we will show how elucidation of the bipartite substructures can reduce the dimensionality of the system in a way that allows one to better understand patterns in the directed flows. That is, we will use tools of community detection to describe an existing directed network. Second, we will show how, in the case of lead-lag correlation-based networks, identification of the "missing links" in these substructures can improve the performance of statistical models in out-of-sample tests.

4.2 Relevance to financial data

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

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

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

The information encoded in such sentiment analyses both reflects and influences the decisions of investors, which collectively may shape the gains and losses of financial markets worldwide. To disentangle the directionality of these relationships, here we investigate the interactions among financial markets and news sentiment data for 40 countries for the period from 2002 through 2012. Through the consideration of both synchronous and lagged correlation-based networks, we explore the extent to which news leads financial market movements, and to which markets lead news. Using tools from linear algebra, we abstract away from the level of individual countries in order to identify large-scale flows of information among geographic regions. We find that, at a time resolution of one day, and both at the level of individual nodes and when considering the network's larger-scale structure, financial markets anticipate news much more substantially than news items anticipate market movements. Finally, we use logistic regression models to show that the structures in the lagged networks are indicative of some degree of predictability; some of these structures have been uncovered previously in studies of international return spillover effects (Diebold and Yilmaz, 2009; Ghosh et al., 1999; Hamao et al., 1990).

In section 4.3 we introduce the data sources, provide summary statistics, and explain our procedure for de-trending the data to guard against spurious results due to serial correlation. In section 4.4 we examine the topological structure of the matrix of synchronous correlations among news sentiment signals and market returns. In section 4.5 we describe our methodology for constructing networks of lagged correlations among news sentiment signals and market returns, interpret the results of our method, and summarize the community structures embedded in the directed network. In section 4.6 we show how consideration of these community structures can be useful in building more robust predictive models. We offer concluding remarks and propose extensions of the work in section 4.7.

4.3 Data and summary statistics

We obtain daily news signals for each country from the Thomson Reuters *MarketPsych* "Sentiment" index, which measures "overall positive references, net of negative references"

(MarketPsych, 2013) for a given country and takes a value in the range $[-1, 1]$. The *MarketPsych* signals are computed using textual news from Reuters as well as various third-party sources. Text is also sourced from blogs, microblogs, and other social media.

First, we clean the data for missing values, replacing them by the sentiment value at one day prior. We discard any country with more than 1% missing values from the analysis. We then difference the sentiment data in order to construct stationary time series $s_{i,t}$, for 40 countries indexed $i = 1, ..., 40$. The full list of countries studied here is provided in Table 4.1.

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

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

as is common in mathematical finance— if prices follow geometric brownian motion, as is commonly assumed, then the returns $r_{i,t}$ are i.i.d. normally distributed in time.

We aim to measure both synchronous and one-day lagged relationships among the signals $s_{i,t}$ and $r_{i,t}$. Many of the news sentiment signals, in addition to the return signals from the markets of certain developing countries, exhibit a non-negligible degree of autocorrelation at a lag of one day. To isolate the influences of external signals from the endogenous structure of each time series, we de-trend all signals for one-day autocorrelation. Specifically, we subtract the influences of these autocorrelation features from our signals using one-step rolling forecasts. For each point $s_{i,t}$ in each news sentiment time series, for example, we fit a local regression (Shumway and Stoffer, 2011)

$$s_{i,t} = \beta_0 + \beta_1 s_{i,t-1}, \tag{4.1}$$

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

$$\tilde{s}_{i,t} \equiv s_{i,t} - (\beta_0 + \beta_1 s_{i,t-1}), \tag{4.2}$$

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

We implement the exact same procedure on the returns $r_{i,t}$ in order to construct the de-trended time series $\tilde{r}_{i,t}$. The signals $\tilde{s}_{i,t}$ and $\tilde{r}_{i,t}$ were obtained for a total of 40 countries over a period ranging from January 8, 2002 to December 31, 2012. Summary statistics, including the first two moments of $\tilde{r}_{i,t}$ and $\tilde{s}_{i,t}$ for each country and index considered, are provided in Table 4.1.

4.4 Synchronous Correlations

4.4.1 Methodology

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

$$C_{i,j} = \frac{1}{T-1} \sum_{t=1}^{T} \frac{(X_{i,t} - \langle X_i \rangle)(X_{j,t} - \langle X_j \rangle)}{\sigma_i \sigma_j}, \tag{4.3}$$

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

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

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

Table 4.1: Summary statistics for returns $\tilde{r}_{i,t}$ and de-trended news sentiment signals $\tilde{s}_{i,t}$ for the period January 8, 2002 to December 31, 2012.

Country	Index	$\langle \tilde{r}_{i,t} \rangle$	$\sigma_{\tilde{r}}$	$\langle \tilde{s}_{i,t} \rangle$	$\sigma_{\tilde{s}}$
Argentina	MERVAL	7.33×10^{-5}	1.91×10^{-2}	-3.10×10^{-5}	9.43×10^{-2}
Australia	AS51	2.01×10^{-5}	1.09×10^{-2}	2.10×10^{-5}	6.55×10^{-2}
Austria	ATX	1.43×10^{-5}	1.61×10^{-2}	-7.95×10^{-5}	1.61×10^{-1}
Belgium	BEL20	3.16×10^{-5}	1.40×10^{-2}	-1.20×10^{-4}	1.17×10^{-1}
Brazil	IBOV	4.44×10^{-5}	1.85×10^{-2}	-5.34×10^{-5}	8.18×10^{-2}
Chile	IPSA	8.21×10^{-5}	1.06×10^{-2}	9.34×10^{-5}	1.72×10^{-1}
China	SHSZ300	4.99×10^{-5}	1.72×10^{-2}	1.42×10^{-5}	4.85×10^{-2}
Colombia	IGBC	5.94×10^{-5}	1.37×10^{-2}	-1.24×10^{-4}	1.26×10^{-1}
Denmark	KFX	3.15×10^{-5}	1.36×10^{-2}	1.24×10^{-4}	1.75×10^{-1}
Finland	HEX25	4.83×10^{-5}	1.51×10^{-2}	-1.70×10^{-4}	2.06×10^{-1}
France	CAC	3.99×10^{-5}	1.59×10^{-2}	-2.37×10^{-5}	5.66×10^{-2}
Germany	DAX	5.12×10^{-5}	1.62×10^{-2}	-2.23×10^{-5}	6.34×10^{-2}
Greece	ASE	7.44×10^{-5}	1.76×10^{-2}	-1.09×10^{-4}	1.20×10^{-1}
Hong Kong	HSI	4.87×10^{-5}	1.57×10^{-2}	4.58×10^{-5}	1.41×10^{-1}
Hungary	BUX	-1.47×10^{-5}	1.67×10^{-2}	-1.81×10^{-4}	1.84×10^{-1}
Indonesia	JCI	1.24×10^{-5}	1.45×10^{-2}	-6.33×10^{-5}	1.03×10^{-1}
Ireland	ISEQ	6.21×10^{-6}	1.54×10^{-2}	-3.62×10^{-5}	9.14×10^{-2}
Israel	TA-25	2.58×10^{-5}	1.26×10^{-2}	-3.17×10^{-5}	4.47×10^{-2}
Italy	FTSEMIB	4.34×10^{-5}	1.58×10^{-2}	-8.96×10^{-5}	6.99×10^{-2}
Japan	NKY	4.43×10^{-5}	1.54×10^{-2}	1.32×10^{-5}	7.28×10^{-2}
Malaysia	FBMKLCI	1.89×10^{-5}	7.81×10^{-3}	-2.73×10^{-5}	1.38×10^{-1}
Mexico	MEXBOL	1.41×10^{-5}	1.33×10^{-2}	-9.46×10^{-6}	8.03×10^{-2}
Netherlands	AEX	3.95×10^{-5}	1.61×10^{-2}	-1.08×10^{-4}	1.61×10^{-1}
New Zealand	NZSE50FG	1.81×10^{-5}	7.13×10^{-3}	1.05×10^{-5}	1.26×10^{-1}
Norway	OBX	3.21×10^{-6}	1.75×10^{-2}	-6.08×10^{-5}	1.36×10^{-1}
Pakistan	KSE100	1.22×10^{-5}	1.38×10^{-2}	-3.32×10^{-5}	6.70×10^{-2}
Peru	IGBVL	3.58×10^{-5}	1.57×10^{-2}	-3.54×10^{-5}	1.74×10^{-1}
Philippines	PCOMP	2.12×10^{-5}	1.29×10^{-2}	-1.01×10^{-4}	1.18×10^{-1}

(next page)

Table 4.1: *Continued:* Summary statistics for returns $\tilde{r}_{i,t}$ and de-trended news sentiment signals $\tilde{s}_{i,t}$ for the period January 8, 2002 to December 31, 2012.

Country	Index	$\langle \tilde{r}_{i,t} \rangle$	$\sigma_{\tilde{r}}$	$\langle \tilde{s}_{i,t} \rangle$	$\sigma_{\tilde{s}}$
Poland	WIG	1.90×10^{-6}	1.31×10^{-2}	-9.02×10^{-5}	1.47×10^{-1}
Portugal	PSI20	1.74×10^{-5}	1.18×10^{-2}	-2.22×10^{-4}	1.72×10^{-1}
Russia	INDEXCF	-5.58×10^{-5}	2.28×10^{-2}	-2.37×10^{-5}	5.88×10^{-2}
Saudi Arabia	SASEIDX	6.59×10^{-5}	1.71×10^{-2}	-4.32×10^{-5}	9.95×10^{-2}
South Africa	TOP40	2.02×10^{-5}	1.41×10^{-2}	-1.06×10^{-4}	8.09×10^{-2}
Spain	IBEX	5.01×10^{-5}	1.57×10^{-2}	-6.63×10^{-5}	8.27×10^{-2}
Sweden	OMX	8.51×10^{-5}	1.56×10^{-2}	-1.27×10^{-4}	1.42×10^{-1}
Switzerland	SMI	-9.55×10^{-6}	1.27×10^{-2}	-1.16×10^{-4}	1.14×10^{-1}
Thailand	SET	5.56×10^{-5}	1.39×10^{-2}	2.55×10^{-5}	1.17×10^{-1}
United Kingdom	UKX	2.06×10^{-5}	1.31×10^{-2}	-1.14×10^{-5}	4.14×10^{-2}
United States	SPX	2.96×10^{-5}	1.33×10^{-2}	-1.71×10^{-5}	3.05×10^{-2}
Venezuela	IBVC	7.62×10^{-5}	1.38×10^{-2}	-1.90×10^{-4}	1.24×10^{-1}

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

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

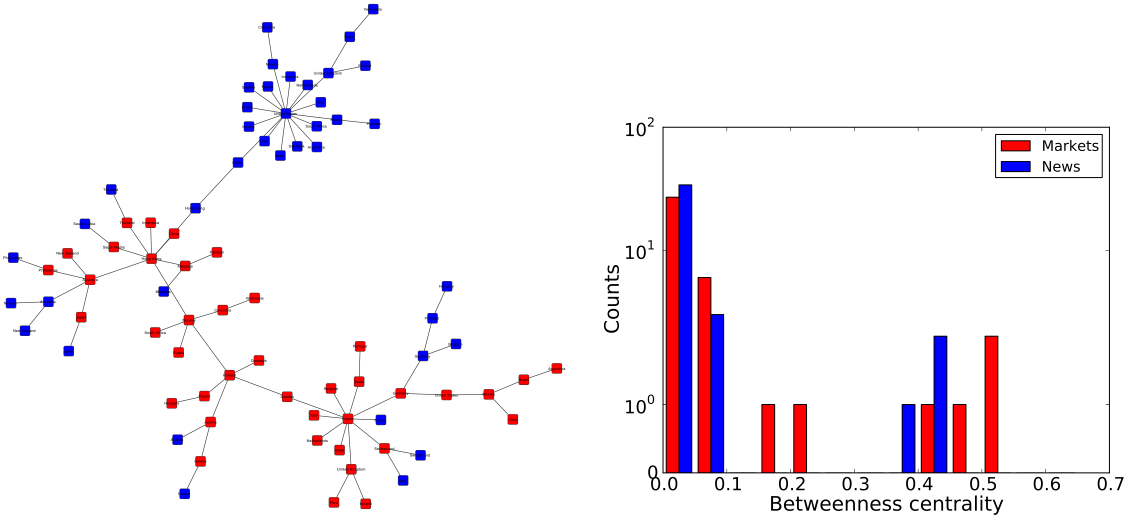
4.4.2 Results

We plot the MST of the data X in Figure 4.1(a), and observe a structure in which the "backbone," or highest-level organization is defined by the financial markets. The lowest-level of the hierarchy, or "leafs" of the tree, are commonly the news sentiment signals. This is corroborated by Figure 4.1(b), which displays histograms of the betweenness-centrality for the financial market nodes and news sentiment nodes separately. The betweenness centrality of a node n is given by (Freeman, 1977)

$$g(n) = \sum_{m \neq n \neq p} \frac{\sigma_{mp}(n)}{\sigma_{mp}}$$

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

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



(a) Minimum Spanning Tree

(b) Associated betweenness centrality

Figure 4.1: (a) Plot of the Minimum Spanning Tree of the synchronous correlations. Financial markets are colored red; news sentiment signals are colored blue. (b) Histogram of the betweenness centrality of financial markets and news sentiment separately. We find that the strongest correlations in the system are among financial markets, and between the news sentiment signals of a country and the same country's market returns. The notable exception is the node corresponding to news sentiment signals from the United States, which is strongly correlated with news from a host of other countries and so represents a hub in the network.

4.5 Lagged Correlations

4.5.1 Methodology

We next study the Pearson correlations at one-day lag. Although the market return data only exists at most between Monday and Friday of each week, the news sentiment data is available seven days per week. We adopt a lagging scheme that maintains a constant time series length T for all relationships studied, but ensures that each term in the Pearson product-moment sum includes signals that are separated by the minimum possible non-zero time lag at a resolution of one day. Our procedure is given in detail in section 4.8.

For each of the four possible categories of relationships– market-market, news-news, news-market, and market-news– we assemble the time series as columns in a matrix $X^{(t)}$. We then shift the time series by one day, as detailed in section 4.8, and assemble them as columns in a matrix $X^{(t+1)}$. We construct the lagged correlation matrix

$$L_{i,j} = \frac{1}{T-1} \sum_{k=1}^{T} \frac{(X_{i,k}^{(t)} - \langle X_i^{(t)} \rangle)(X_{j,k}^{(t+1)} - \langle X_j^{(t+1)} \rangle)}{\sigma_i \sigma_j} \tag{4.4}$$

as in equation (4.3).

To study the structure of this matrix, we aim to filter its elements into a network of directed relationships. The Minimal Spanning Tree relies on a symmetric distance $D_{i,j}$ between any two nodes. It therefore does not readily extend to the study of lagged correlation networks, in which the correlations are asymmetric: in general, $L_{i,j} \neq L_{j,i}$. More generally, such topological methods of filtering a correlation matrix into a network, which rely only on a ranking of the measured correlation coefficients, are less robust to statistical uncertainty than simpler methods, such as applying a threshold to the matrix (Curme et al., 2014). This is especially important when studying lagged correlations, which tend to be much lower in magnitude than synchronous correlations.

We could apply a simple thresholding procedure, choosing a static threshold based on statistical confidence— i.e., a correlation coefficient that has a probability less than p of being generated by uncorrelated variables. But this threshold will vary with the distribution of the signals under consideration, many of which are known to be non-normal (Mantegna and Stanley, 2000). To this end we apply a bootstrapping procedure (Curme et al., 2014) in which the rows of the matrix $X^{(t)}$ are shuffled repeatedly in order to construct a distribution for the sample correlation coefficient as measured using uncorrelated signals of the same distribution as the data. We then apply a uniform statistical threshold of $p = 0.01$, with FDR correction for multiple comparisons (Benjamini and Hochberg, 1995), to obtain thresholds of measured correlation coefficients that vary for each time series pair. Thus, we construct the four different $X^{(t)}$ and $X^{(t+1)}$ matrices described above, perform $100 \times N^2 = 100 \times (80)^2 = 640{,}000$ independent shufflings of the data, construct the distribution for the measured correlation coefficient under the null hypothesis of uncorrelated variables, and accept into our directed network any pair that has a probability $p < 0.01$ of being generated by uncorrelated variables after FDR correction. Further details of this procedure, including the implementation of the FDR correction, are given in section 4.9.

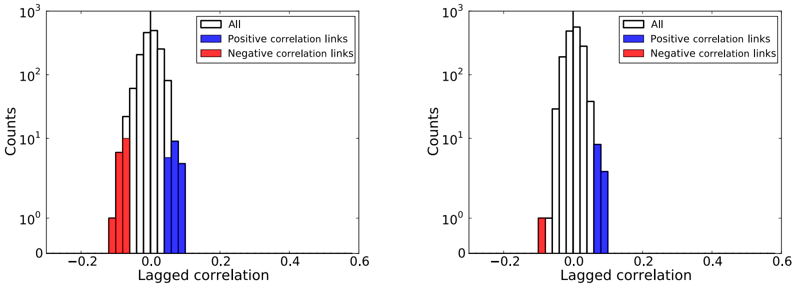
This procedure yields four networks of statistically-validated directed links. In the subsequent portions of the chapter, we will both analyze the structure of these networks, and explore their utility as a feature-selection tool in developing prediction models.

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

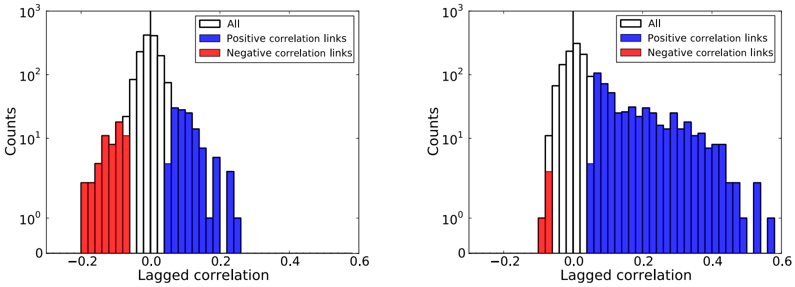
4.5.2 Results

In Figure 4.2 we display histograms of measured lagged correlation coefficients separately for relationships among news sentiment signals, among market returns, and between news and markets. The histograms are shaded according to the numbers of links that are validated according to the statistical validation procedure described above. The corresponding subgraphs of the validated lead-lag relationships are displayed in Figure 4.3, where we preserve the geographical location of each node. We distinguish positive and negative correlations by the colors of the links.

We find that the greatest number of validated links are between financial markets, with 534 links of positive correlation and 4 links of negative correlation. There is also a substantial number of links leading from markets to news sentiments, as we validate 118 links of positive correlation and 56 links of negative correlation. By contrast, we find far fewer entities, among both news sentiments and market returns, that are lead by news.



(a) News → News (b) News → Markets

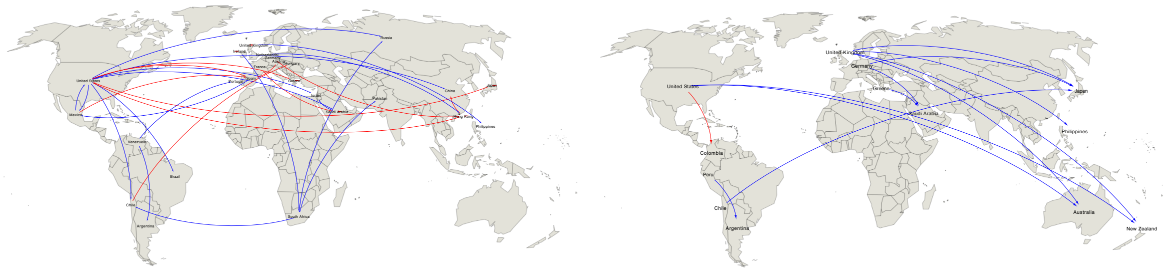


(c) Markets → News (d) Markets → Markets

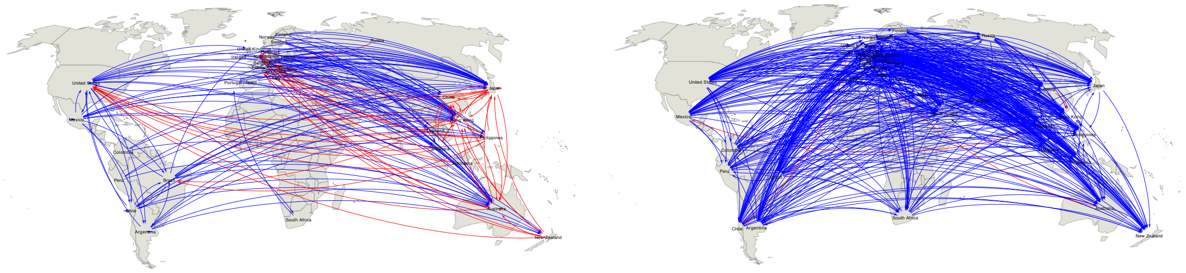
Figure 4.2: Histograms of lagged correlation coefficients (a) among news sentiment signals, (b) in which news anticipate market movements, (c) in which market movements anticipate news, and (d) among market movements. Shading indicates positive (blue) and negative (red) coefficients of pairs that are filtered into the statistically validated network.

In this sense, we find that the system is primarily driven by market movements, which complements our study of the synchronous correlations in which the markets composed the base of the Minimal Spanning Tree. A comparison of the distributions of correlation coefficients in which news leads markets to those in which markets lead news, as displayed in Figure 4.2, again suggests that the stronger relationships are those in which the markets anticipate news sentiment.

At the level of individual lead-lag relationships, then, we find that the strongest correlations are those that are driven by market movements. To analyze the higher-level structure of the networks, we make use of a well-known clustering algorithm involving a spectral



(a) News → News (b) News → Markets



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

Figure 4.3: Plots of each subgraph of the statistically validated network. (a) shows lagged relationships among news sentiment signals; (b) shows lagged relationships from news signals to market returns; (c) shows lagged relationships from market returns to news signals, and (d) shows lagged relationships among market returns. Blue color indicates validated links of positive correlation; red color indicates validated links of negative correlation. Network visualizations are prepared with the Cytoscape software framework (Shannon et al, 2003).

decomposition of the adjacency matrix A, where $A_{i,j} = 1$ if a link exists from i to j, and 0 otherwise. Here we consider the full $N \times N = 80 \times 80$ adjacency matrix that is the union of the graphs displayed in Figure 4.3.

To describe the large-scale flows in the statistically-validated lagged correlation network, we study the Singular Value Decomposition of the full adjacency matrix A, as described in section 4.1. In Table 4.2 we display the largest five components in magnitude of selected left- and right-singular vectors U^n and V^n of A. Included are the top three singular vector pairs in terms of their corresponding singular value σ_n. Plots of all entries of the first three pairs of left- and right-singular vectors are included in Figure 4.4. In Figure 4.4 we also plot the full directed network, arranging the positions of nodes according to their entries in the first three singular vector pairs.

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

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

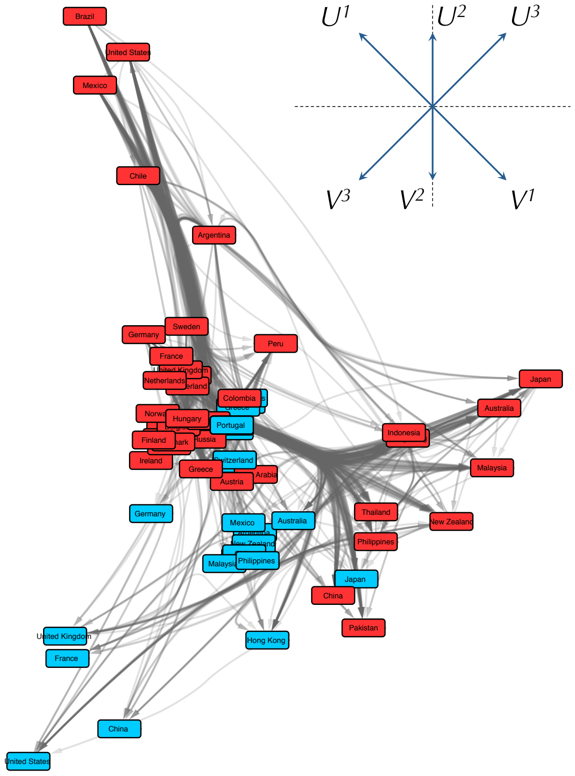
4.6 Relation between the structure of the statistically-validated network and prediction model performance

We further investigate the predictability of node signals within the statistically-validated lead-lag network. We first divide our data into a training set from 2002 to the end of 2010, and a testing set from 2011 to the end of 2012. We construct the statistically-validated network, using the methodology described above, with only the training subset of the data.

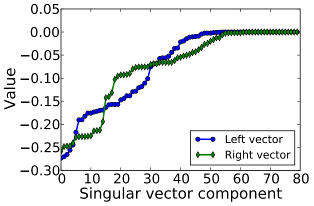
We then employ the networks as a feature-selection step in the training of a classifier. We aim to predict the sign (+1 or -1) of the signals $\tilde{r}_{i,t}$ and $\tilde{s}_{i,t}$, using both the most recent previous index returns and news sentiment data. For each node, we exclude days of sign zero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

Table 4.2: Largest five components of the first three left- and right-singular vector pairs. Entries refer to market indices, unless otherwise specified as news.

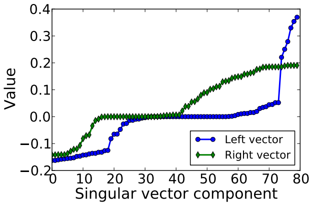
σ_1	U^1	V^1
21.9	United States Mexico Brazil Chile Argentina	New Zealand Philippines Australia Japan Malaysia
σ_2	U^2	V^2
9.74	United States Mexico Brazil Chile Saudi Arabia	France United Kingdom Sweden Finland Belgium
σ_3	U^3	V^3
6.86	Japan Australia Philippines Hong Kong Malaysia	China News United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



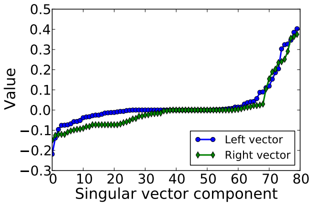
(a) Complete directed network



(b) U^1 and V^1



(c) U^2 and V^2



(d) U^3 and V^3

Figure 4.4: (a) Display of the complete directed network, showing all links among markets (red) and news sentiment signals (blue). Nodes are arranged according to their entries in the first three left and right singular vectors. Specifically, we associate a vector in the plane $\mathbb{R}^2$ to each of the singular vectors U^1, U^2, U^3, V^1, V^2, and V^3 (inset). Each node position in the plane is then a weighted sum of these six 2-vectors, where the weight of each 2-vector is equal to the magnitude of the node's entry in the corresponding singular vector. Edges are bundled according to the algorithm in Holten and van Wijk (2009) to highlight the larger scale flows among groups of nodes. In (b), (c), and (d) we plot the sorted components of the first three pairs of left- and right-singular vectors. For each vector, the largest entries in magnitude tend to be of the same sign. Network visualizations are prepared with the Cytoscape software framework (Shannon et al, 2003).

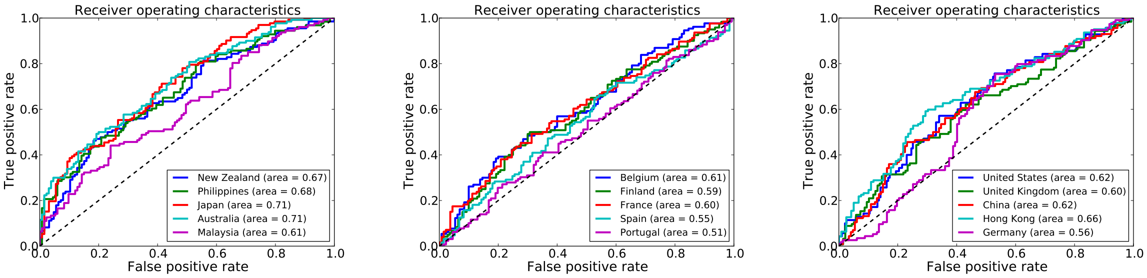
When modeling a given node i, we use as inputs all nodes j for which there is an edge from j to i in the statistically-validated network constructed from the training data. The number of inputs to each logistic regression, therefore, is equal to the in-degree of the desired node. For each node, we assemble the lagged input signals $\tilde{r}_{i,t}$ and $\tilde{s}_{i,t}$ as columns in a matrix X. Signals are lagged as in section 4.5.1, and standardized to Z-scores by subtracting the mean and scaling by the standard deviation of the training set of each column. We then fit a logistic regression using the training data from 2002 through 2010, and test on data from 2011-2012. For a row vector $\vec{x}$ of X, the logistic regression models the probability for an upward movement in $\tilde{r}_{t+1}$ for a desired market as

$$\Pr(\tilde{r}_{t+1} > 0|\vec{x}) = \frac{e^{\beta_0+\vec{\beta}\cdot\vec{x}}}{1+e^{\beta_0+\vec{\beta}\cdot\vec{x}}}, \tag{4.5}$$

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

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

In Figure 4.5 we plot some sample ROC curves for 15 of the logistic regression models. In particular, we repeat the singular value decomposition on the adjacency matrix for the network constructed from the training data, and plot the ROC curves for the largest five entries of the right singular vectors $|V^1|$, $|V^2|$, and $|V^3|$ (note the large overlap of these entries with those from Table 4.2, which was constructed from the full data set). The notation $|V^i|$ indicates the vector of absolute values of the entries of V^i. We find that these models perform reasonably well on the test data.



(a) Largest five entries in V^1 (b) Largest five entries in V^2 (c) Largest five entries in V^3

Figure 4.5: ROC curves for the performance of the logistic regression model in predicting (a) daily returns $\tilde{r}_{i,t}$ of the stock indices in the top five entries of $|V^1|$ and (b) $|V^2|$, and (c) sentiment scores $\tilde{s}_{i,t}$ for the news signals in the top five entries of $|V^3|$. Each ROC curve is generated by varying the threshold probability for the prediction of a positive return. The area under each curve is provided in the legend.

We compare the performance of all logistic regressions, using only the inputs as defined by the validated network, with the performance of models that use all 80 nodes as inputs in the vector $\vec{x}$. In Figure 4.6 we show the distributions of differences in AUCs between these two sets of models, finding that in nearly all cases the feature selection step represented by constraining inputs according to the validated network provides for significant gains in accuracy in the test data. The network is thus highlighting persistent relationships among nodes and excluding noisy inputs that may confound predictive models.

Finally, we explore the extent to which information on the predictive relationships among nodes is encoded in the wiring diagram of the validated network's adjacency matrix. In Figure 4.7 we plot the AUC for all markets against the magnitude of the entry of each market in the right singular vectors V^1 and V^2. Similarly, we plot the AUC for all news sentiments against the magnitude of the entry of each node in the right singular vector

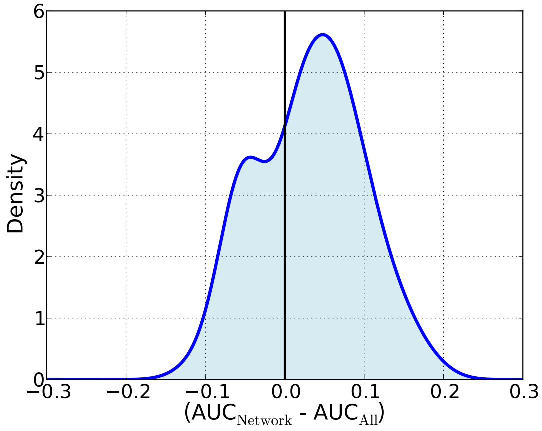
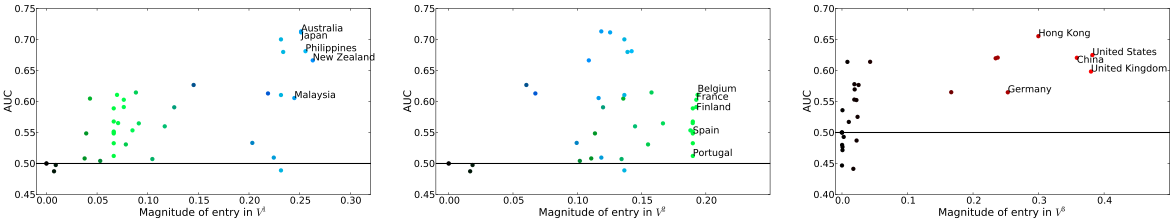


Figure 4.6: Pairwise differences in AUCs between models with inputs defined by the validated network, AUC$_{\text{Network}}$, and those using all possible inputs, AUC$_{\text{All}}$. The distribution of AUC differences is shown for all news sentiment and market return signals, and is represented using a Gaussian kernel density estimate, with a bandwidth calculated using Silverman's rule of thumb. The median of this distribution differs significantly from zero according to a non-parametric Wilcox test ($V = 2407$, $p < 0.001$), suggesting that the networks constructed using the training data uncover persistent lead-lag relationships, and that restricting model inputs to the nodes defined by these networks offer improved model performance.

V^3. We find that the majority of market indices cannot be reliably predicted using data at a time horizon of one day, in accordance with the efficient market hypothesis (Fama, 1970). However, there does exist a group of nodes that exhibits a considerable degree of predictability, and these are precisely the nodes identified in the first right-singular vector V^1 of the adjacency matrix of the full network. Similarly, the most predictable signals among news sentiments are those with the highest entries (in magnitude) in the right singular vector V^3, as shown in Figure 4.7(c). These numerical demonstrations suggest that the SVD of the lagged correlation network's adjacency matrix may be a plausible method for identifying predictable subsets of nodes in a complex network.

We also investigate the extent of the information encoded in the left singular vectors.



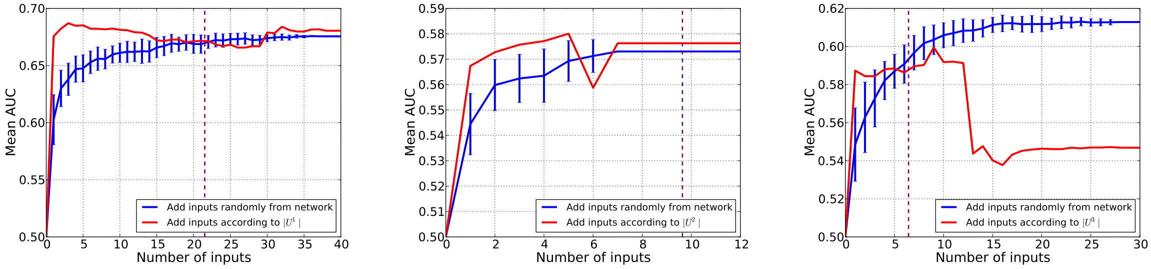
(a) Markets

(b) Markets

(c) News

Figure 4.7: AUC for all markets against (a) the magnitude of the entry of the corresponding element of V^1, and (b) the magnitude of the entry of the corresponding element of V^2. Points are shaded blue according to the magnitude of the entry in V^1, and green according to the magnitude of the entry in V^2. In (c) we plot the AUC for all news against the magnitude of the corresponding entry in V^3 (additionally shaded in red). We observe that the right singular vectors of the adjacency matrix identify subsets of predictable nodes.

To this end, for the top 5 entries in each right singular vector, we add inputs sequentially to each model, and compute the out-of-sample AUC. We compare the effect of two schemes: in the first scheme, when modeling node i, we choose each additional input at random from all nodes j for which there is an edge from j to i in the validated network. In the second scheme, we choose each additional input in the order of their ranking in the corresponding left singular vector. In Figure 4.8 we plot the mean AUC for the top five entries of each right singular vector against the number of inputs in each model. We find that, when modeling the signals of nodes highlighted in the right singular vectors, the corresponding nodes highlighted in the left singular vectors tend to represent the most important inputs to the model. In the case of the nodes in V^1 and V^2, for a node of in-degree k_{in}, using as inputs the largest k_{in} components of U^1 or U^2 will on average result in better model performance than using the inputs selected by the network. The effect is weaker for the nodes in V^3, although choosing nodes from the largest components of the left singular vector U^3 still yields comparable model performance to choosing them from the underlying network, up to the singular value corresponding to this singular vector pair (which, as in



(a) Elements of V^1

(b) Elements of V^2

(c) Elements of V^3

Figure 4.8: AUC, as averaged among the top five nodes in $|V^1|$ (markets), $|V^2|$ (markets), and $|V^3|$ (news), for each additional model input. When the number of model inputs exceeds the in-degree of a node, we cease adding inputs. In blue, we plot the mean AUC when randomly adding input nodes from the validated network, as averaged over 50 iterations. In red, we plot the mean AUC when input nodes are added in order of their magnitudes in the corresponding left singular vectors, regardless of the presence or absence of a link in the validated network. Dashed vertical lines mark the singular value associated with each singular vector pair, approximating the number of nodes involved in the large-scale flow. We find that the most important inputs to nodes with large weight in the first three right singular vectors are nodes with large weights in the corresponding left singular vectors.

Taylor et al. (2011) approximates the geometric mean of the number of nodes involved in the large-scale flow). Whereas the right singular vectors identify subsets of predictable nodes, then, the corresponding left singular vectors seem to identify the most important inputs to these nodes, with respect to the performance of our prediction models. We therefore find that the network's adjacency matrix alone can offer nontrivial insights into global flows of information.

4.7 Conclusions

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

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

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

A possible application of similar analyses in the context of lagged-correlation networks would be a “recommender system” for exogenous inputs in time series models. A preliminary feature-selection, such as the construction of a statistically-validated network, is always subject to false negatives or false positives. A simple singular value decomposition allows one to "recommend" inputs for a model according to the inputs of other nodes— other time series— that otherwise share similar inputs. As demonstrated here, incorporating such inputs can potentially improve performance, though the limitations of this approach are evident in Figure 4.8(c). This approach could also be refined with more sophisticated recommender systems, although we make no claims about the statistical basis for the functioning of these systems.

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

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

4.8 Lagging procedure

In this work we study Pearson correlations among news sentiment signals and market returns at one-day lag. While the news sentiment data is available seven days per week, the market return data only exists at most between Monday and Friday of each week. To account for this difference, we adopt the following scheme, which we diagram in Figure 4.9.

- For correlations between financial markets, we include products between returns on Friday and those on the following Monday in the Pearson product-moment sum, using all available data.

- For correlations between news data and subsequent market movements, we relate news sentiment data between Sunday and Thursday of each week with market data from Monday to Friday.

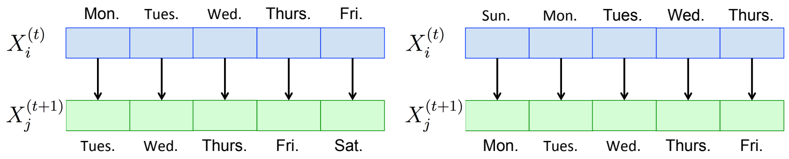
- For correlations between market movements and subsequent news data, we relate market data from Monday to Friday with news sentiment data between Tuesday and Saturday of each week.

- For correlations between news sentiment data, we relate news sentiments between Monday and Friday of each week with those from Tuesday to Saturday of each week. This method allows for a comparison between the effects of market returns and news sentiment signals on subsequent news sentiments.

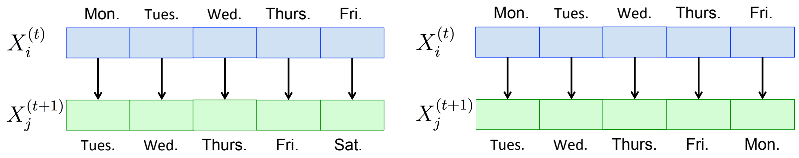
This scheme maintains a five day week, and therefore a constant time series length T, for all relationships studied. We also use all available market data. An alternative scheme is to simply synchronize all time series, removing data from Saturdays and Sundays, as is done in Section 4.4. We would then simply correlate each time series against time series that have been shifted by one day. We have checked to confirm that this change only weakly impacts the results.

4.9 Statistical validation of directed links

We aim to filter the lagged correlation coefficients in L according to a threshold of statistical significance. In this high-dimensional setting, composed of signals that are by no means normally distributed, it can be difficult to infer the joint probability distribution of the data (Tumminello et al., 2007). We will thus apply a bootstrapping procedure (Efron and Tibshirani, 1993) in order to determine the statistical significance of each entry of L sepa-



(a) News → News (b) News → Markets



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

Figure 4.9: Diagram of lagging procedure for measuring lagged correlations. We maintain a five day week, and therefore a constant time series length T, for all four classes of relationships. This scheme uses all available market data, but only includes terms that are spaced exactly one day apart when possible.

rately, and filter L according to a statistical threshold. Although this threshold is uniform among all measured lagged correlations, the lagged correlation coefficient corresponding to this threshold will vary with the distributions of each pair of signals under consideration. See Curme et al. (2014) for an analysis of this method when applied to intraday stock returns.

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

From matrix U we associate a one-tailed p-value with all positive correlations as the probability of observing a correlation that is equal to or higher than the empirically-measured correlation, under the null hypothesis of uncorrelated signals. From D we may similarly associate a one-tailed p-value for all negative correlations. We choose our statistical threshold to be $p = 0.01$. Because we are performing many statistical inferences simultaneously, however, we must correct our p-values to account for multiple comparisons. We use the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) protocol to correct all N^2 p-values. According to this correction, the p-values from each individual test are arranged in increasing order ($p_1 < p_2 < \cdots < p_{N^2}$), and the threshold is defined as the largest k such that $p_k < k\ 0.01/N^2$. In this case, for $N = 80$ nodes, we must construct $100N^2 = 640,000$ independently shuffled surrogate time series. We may then interpret $U_{m,n}/(100N^2)$ as the p-value for the positive one-tailed test, and $D_{m,n}/(100N^2)$ as the p-value for the negative one-tailed test. Directly from the matrices U, and D, then, our threshold is the largest integer k such that U or D has exactly k entries fewer than or equal to k. From this threshold we can filter the links in L to construct the FDR network (Tumminello et al., 2011).

4.10 Tests with synthetic data

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

To this end we engage in two experiments. In the first experiment we generate many simulated time series with the same underlying correlation network as the real-world data. By varying the strength of the correlations, we examine the range over which our method– selecting model inputs according to their ranking in the corresponding left singular vector of the adjacency matrix– outperforms the “null” model of simply choosing inputs according to the adjacency matrix alone. In the second experiment we generate many time series with an underlying bipartite lagged correlation network. We fix the distribution of lagged correlation coefficients to match what we find in empirical data, but vary the bipartivity of the underlying network in order to test the sensitivity of our results to changes in network topology.

4.10.1 Effect of signal-to-noise ratio

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

$$\vec{x}_t = \mathbf{B}\vec{x}_{t-1} + \vec{\epsilon}_t. \tag{4.6}$$

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

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

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

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

$$\mathbf{B} = \alpha \Sigma^{-1} L. \tag{4.7}$$

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

In this way we may construct N time series of length T, and find its associated lagged correlation network as before, using FDR correction for multiple comparisons. Because our signals are homogeneously and normally distributed, we filter our network according to a Gaussian threshold corresponding to $p < 0.01$ using Eq. (2.7). This simplification allows us to generate large numbers of these systems in a reasonable amount of time. We find that, for $\alpha = 1$, the properties of the resulting system– namely, its synchronous correlation matrix, lagged correlation matrix, and validated adjacency matrix– match closely our empirical results.

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

We then continue our time series for another $T = 100$ time-steps, and measure the AUC of each model in each of cases (i) and (ii) on this held-out data set. In Figure 4.10 we show differences in the measured AUCs for varying values of α. We find that for low values of α, there is no difference between cases (i) and (ii). That is, the lagged correlations in the system are so weak that both methods perform equally poorly. For values of α in the range

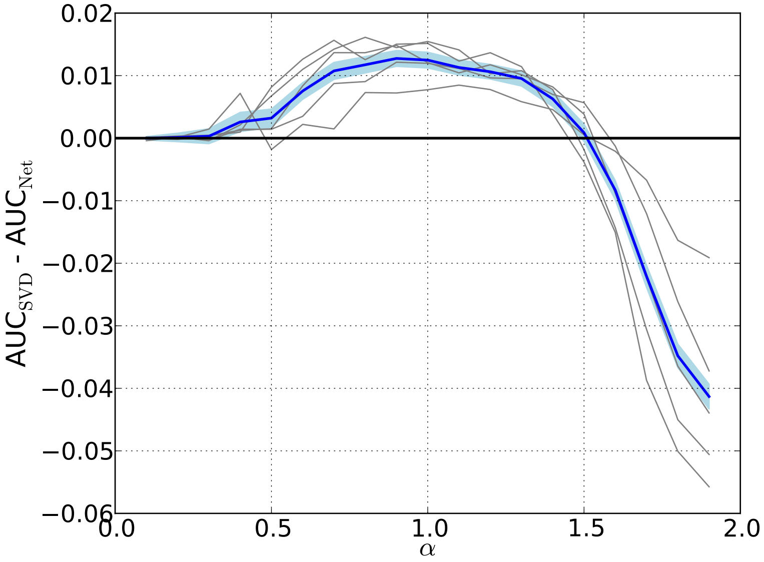


Figure 4.10: Pairwise differences in AUCs between logistic regression models with inputs given by case (ii), $\mathrm{AUC}_{\mathrm{SVD}}$, and inputs given by case (i), $\mathrm{AUC}_{\mathrm{Net}}$, for varying α. In grey we show characteristic trajectories of each of the five largest entries of the first right singular vector. In blue we show results as averaged over each of these entries.

from roughly 0.5 to 1.5, case (ii) outperforms case (i) by 1% - 2%. In this regime, we find that consideration of the network's bipartite community structure can increase the accuracy of our predictions. Note that, if we characterize the strengths of the lagged correlations by what we find in the empirical data (corresponding to $\alpha = 1$), we achieve a near-optimum gain in accuracy. For α larger than 1.5, however, one is much better-off choosing model inputs from the adjacency matrix alone. In this regime the signal-to-noise ratio is sufficiently strong that we find a low rate of false positives and negatives in the validated links, so that our recommender system has little to offer.

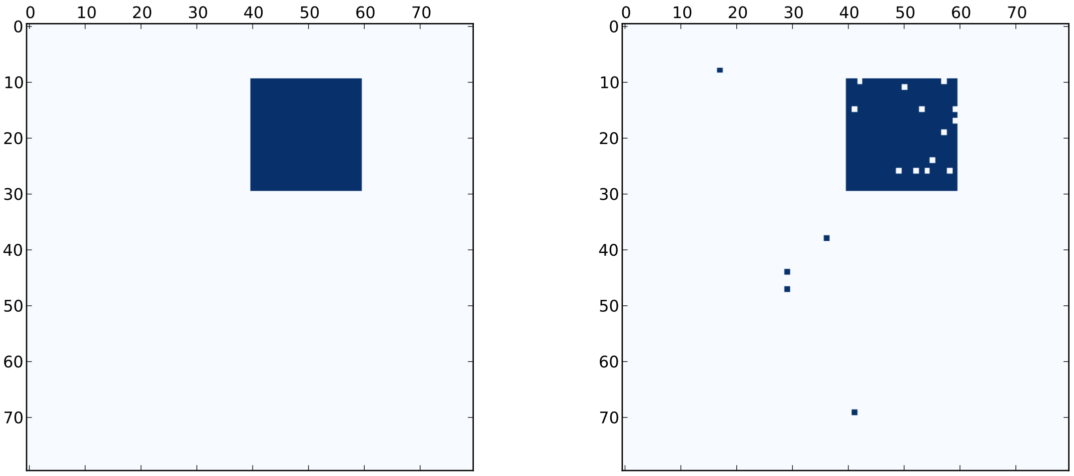
4.10.2 Effect of network bipartivity

In this experiment we fix the distribution of lagged correlation coefficients to match what we find in empirical data, but vary the bipartivity of the underlying network in order to test the sensitivity of our results to changes in network topology. We again construct N time series of length T according to the iterative procedure in Eq. (4.6). We use Eq. (4.7) to construct the matrix $\mathbf{B}$, fixing $\alpha = 1$.

To construct the matrix Σ, we first create a matrix U, the columns of which are an orthonormal basis (we simply construct random vectors in the range (-1,1), and then apply the Gram-Schmidt process). We then construct a diagonal matrix Λ of positive random numbers, and take $\Sigma = U\Lambda U^T$ to be our positive-definite matrix. To construct L, we first construct a perfectly bipartite network. The adjacency matrix for this network is shown in Fig. 4.11(b). We then randomly re-wire the network as follows: each entry in the adjacency matrix is switched (from 0 to 1 or from 1 to 0) with probability p. The parameter p describes the extent to which the underlying network is bipartite. The links are then weighted by lagged correlation values sampled from the same range as our empirical data. We use the resulting weighted adjacency matrix as the L in Eq. (4.7).

For each value of p, we simulate N simulated time series of length T, and construct the corresponding validated network. The adjacency matrix for one such network, corresponding to $p = 0$, is shown in Fig. 4.11(b). Here one can see the influence of false positives and false negatives in the statistical validation process.

We demonstrate how consideration of the network's bipartite community structure can mitigate the effect of these false positives and false negatives. For each value of p, we generate 500 systems, each of which is composed of $N = 80$ simulated time series of length $T = 400$. As before, we compute the singular value decomposition of the resulting adjacency matrix, and train logistic regression models in which we attempt to classify the sign (+1 or -1) of the signal a given node at each time t. We again consider both cases (i) and (ii). We then continue our time series for another $T = 100$ time-steps, and measure the AUC of each model in each of cases (i) and (ii) on this held-out data set. Results are shown in Fig. 4.12. We find that for small p, the methodology of case (ii), in which we use the community structures to recommend features to the models, provides an increased accuracy. For modestly large values of p above 1.5%, however, case (i) provides a larger out of sample

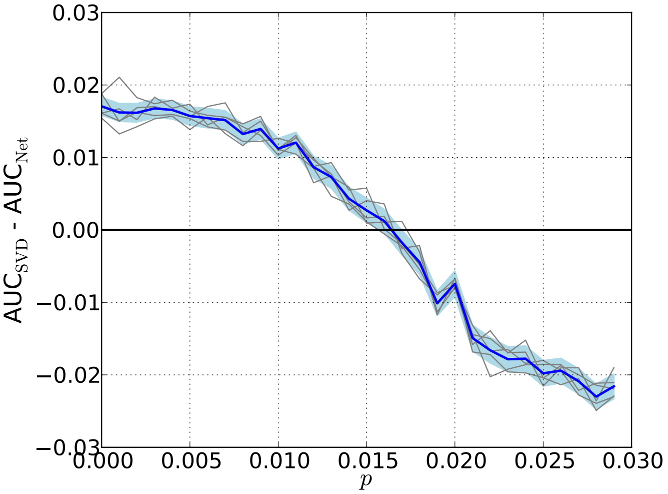


(a) Underlying adjacency matrix

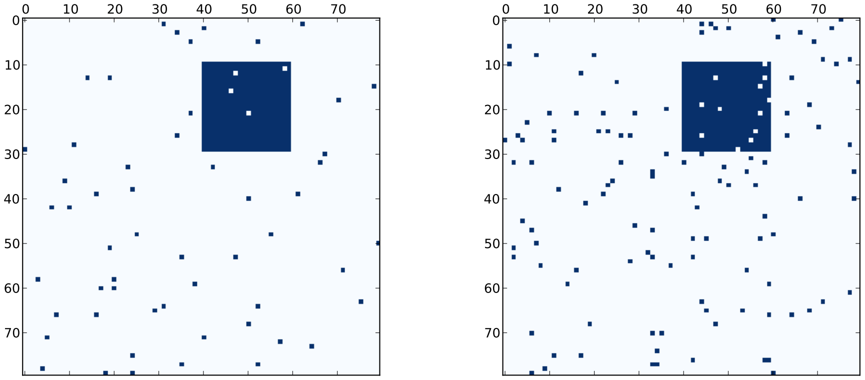
(b) Reconstructed adjacency matrix

Figure 4.11: (a) Bipartite adjacency matrix used to simulate time series, with $p = 0$. (b) Adjacency matrix of the resulting statistically-validated network, showing the influence of false positives and false negatives in the statistical validation process.

accuracy. We therefore conclude that our findings rely on a large degree of bipartivity in the underlying lagged correlation network.



(a) Difference in AUC



(b) Underlying adjacency matrix, $p = 0.01$ (c) Underlying adjacency matrix, $p = 0.02$

Figure 4.12: (a) Pairwise differences in AUCs between logistic regression models with inputs given by case (ii), $\mathrm{AUC}_{\mathrm{SVD}}$, and inputs given by case (i), $\mathrm{AUC}_{\mathrm{Net}}$, for varying p. In grey we show characteristic trajectories of each of the five largest entries of the first right singular vector. In blue we show results as averaged over each of these entries. In (b) and (c) we show sample underlying adjacency matrices for two values of p.

Chapter 5

Using topic models to explain market movements

In this chapter we complement our studies of financial news data with investigations of "large scale" properties of financial news and Internet search data, and their relationships with market movements. That is, we abstract away from the incidence of individual keywords, which form the basis of the sentiment analysis of Chapter 4, to study the dynamics of semantic topics. Our tool for this task is a hierarchical Bayesian model for text known as Latent Dirichlet Allocation (LDA). We review LDA and apply it to (i) Internet search data from *Google Trends* and (ii) financial news from *The Financial Times*. Using common techniques from time series analysis, we study how the dynamics of the topics in these domains relate to market movements. We report that (i) only changes in *Google* searches for words related to finance and politics, out of a large universe of potential topics, tend to precede stock market movements; in particular increases in these searches tend to precede falls in the market. Using the *Financial Times* data, we find that attention in the news condenses to a small number of high-interest topics immediately following falls in the stock market, and immediately preceding jumps in trading volume. We bolster our conclusions with a number of statistical robustness checks.

5.1 Application to financial news data

The well-being of individuals and entire economies is increasingly tied to activity in the financial sector, a point emphasized by the 2008 financial crisis. A large portion of this activity is reflected in stock market movements, which are driven by the trading decisions of many investors. The motivating forces behind these decisions, whether they are exogenous news items, or the endogenous influences of other traders, have therefore naturally received much scientific attention [90, 98, 126, 132–134, 139].

An understanding of systematic relationships between financial news and the actions of traders and investors has largely remained elusive. This is in part because the information embedded in textual documents is difficult to quantify. Nonetheless, one observes certain regularities in the ebbs and flows of stories into and out of the news, at least qualitatively. During the “silly season” or “slow news season” in the summer months, for example, the media may pay increased attention to seemingly frivolous topics. By contrast, attention in the news may be sharply focused on a small number of issues during a war, or following natural or economic disasters. The variety of news story lifetimes is also familiar: whereas some topics remain in the news for great lengths of time, others are forgotten soon after they are first reported. In order to understand the interplay between the actions of investors and issues in financial news, it may be first necessary to grapple with these common “meta-characteristics” of news items.

Recent advances in natural language processing and text analysis have assisted in the quantification of certain features in financial news, and the study of how these features individually relate to market activity. Indeed, automated approaches to forecasting financial market movements through the text-mining of news and social media has driven the development of entire industries [99, 100]. Academic interest has also focused on the reciprocity of the relationship between news and market movements [101]. Recently, much attention has been devoted to the information embedded in novel online sources, such as social media [102, 103] and Internet search records [104, 105, 149, 158].

When one sets out to relate textual information to some real-world activity, one is immediately confronted with a vast universe of words, each of which may or may not be individually relevant to the question at hand. That is, text data is naturally high-dimensional. A first step toward rendering these data tractable for analysis, then, is often to reduce their dimensionality by clustering words together into groups. A common tool for this task is topic modelling. Under this approach, a text corpus is partitioned into documents, each of which is usually treated as an unordered collection of words, or "bag of words". One can then use the co-occurrence of words in documents in order to infer semantic similarities among words and documents. For example, the words "rain", "wind", and "clouds" may naturally occur together frequently in documents, allowing one to associate them as members of a single topic, in this case related to the weather.

Topic modelling algorithms treat each document as a mixture of topics, allowing one to both group words into topics and to measure similarities between the mixtures of topics in two separate documents. One of the simplest and most popular topic modelling algorithms is Latent Dirichlet Allocation (LDA) [159]. LDA and similar methods, such as probabilistic latent semantic analysis (PLSA) [106] represent documents in a low-dimensional "semantic space", allowing one to abstract away from individual keywords in order to describe the distribution of topics– each of which is a distribution of keywords– in a document. A document discussing a hurricane in a certain country, for example, might be represented as 30% in a topic about weather, 30% in a topic about that particular country or region of the world, and 40% in topics about politics or economics, discussing the ramifications of the event. See [107] for a review of the subject.

LDA has been applied to financial news corpora and Internet search data in efforts to understand what groups of keywords may be related to large trading volumes or market returns when searched online [105], or when appearing in the news [108]. Most approaches, however, focus on characterizing the importance of individual topics, such as groups of "bearish" or "bullish" keywords. Just as the collective actions of individual traders are relevant to stock market movements, we hypothesize that larger-scale descriptions of the news, such as the tendency of news to focus on large or small numbers of topics, may also bear relevance to understanding trading decisions.

Here, we investigate the relationship between the diversity of topics appearing in financial news— represented by daily issues of the *Financial Times*— and trading activity in financial markets. Specifically, we apply a topic modelling approach in order to distill to a single number the extent to which a given issue of the *Financial Times* is focusing on a large number of topics, or a small number of topics. We consider the time series of this news diversity, as constructed from a corpus of financial news from 2007 to 2012. Our analysis suggests that large drops in diversity– occurring when attention is focused on a small number of topics in the news– follow falls in the stock market, and that increases in diversity follow upward market movements. Moreover, we present evidence that the time series of diversity can be applied to assist forecasts of daily trading volume, finding that increases in trading volume tend to coincide with falls in the diversity of the *Financial Times* that morning. Our analysis suggests that the breadth of news to which traders are exposed may be important in understanding the information flows that are at play during large stock market movements.

5.1.1 Quantifying the diversity of financial news with LDA

To understand the diversity of news in an issue of the *Financial Times*, a natural first step is to measure what topics are represented in the news, as well as the space devoted to each topic. LDA presents an ideal framework for these measurements, as it is a standard tool for decomposing a text into a mixture of topics, each of which is assigned a “weight” that represents the fraction of content that is devoted to that topic.

We analyze a corpus of daily issues of the *Financial Times* from January 2, 2007 to December 31, 2012. Issues were retrieved from `http://www.ft.com/` in Portable Document Format (PDF). All issues were retrieved for this period, with the exception of five dates due to technical problems. These dates were February 22, 2007, March 8, 2007, May 12, 2007, January 28, 2009, and November 8, 2012. Each PDF was converted to text format

(.txt) using the open source software *pdftotext*, which is freely-available and included in most Linux distributions.

Documents for input to the Latent Dirichlet Allocation (LDA) were defined as blocks of text that were separated by isolated newline sequences "\n" and contained greater than 30 words. All characters were processed to unicode, forced to lowercase, and hyphens were replaced with whitespace. All characters other than the letters "a" through "z" were removed. The remaining text was then stemmed using the Porter stemming algorithm [109], cleaned of single-letter words, and cleaned of (stemmed) stopwords. We used the MySQL stopword list [110], supplemented with the words "ft", "financial" "times", "xd", "gbp", "usd", "euro", "acc", "eur", "page", "per", "cent", and "mr". Processed documents containing fewer than 30 words were removed.

In the framework of LDA, a topic is a distribution over a finite number of words. Each topic is then a list of words, each of which is associated with a numeric weight, such that the weights sum to one. The LDA algorithm models each document in a corpus as a mixture of K topics. We choose to treat each paragraph of the *Financial Times* as a separate document, in order to obtain $937,649$ total documents of roughly equal lengths. Each issue on average contains approximately 515 documents.

We configure a weighted LDA [111]- [114] to model each document as a mixture of $K = 50$ topics. In order to reduce the influence of common words when identifying topics, we weight word counts inversely to their frequency in the entire corpus, using the TF-IDF weighting scheme for individual words [115]. We find that this scheme helps to control for certain words that were abundant in the financial literature, but absent from conventional stopword lists. The selection of $K = 50$ results in a reasonable identification of topics upon post-hoc inspection. Moreover, we can check how well our model fits the text, and we find that changes from this value of K do not considerably augment the model's likelihood, as measured by low model perplexities when testing on held-out corpora. We uncover a range of topics, involving politics ("labour", "elect", "party",...), energy and the environment ("carbon", "energy", "environment",...), technology ("google", "facebook", "social",...) and the economy ("market", "rate", "bank",...). The top ten (stemmed) words for each of the 50 topics are provided in Appendix A.

The *gensim* Python package [111] was used for the LDA on the full set of processed documents. We configured a batch LDA, with ten passes over the entire corpus.

Once the LDA is trained, each document d in the corpus is represented by the K-dimensional topic vector $\theta_d = (\theta_{d,1}, \theta_{d,2}, ..., \theta_{d,K})$. The terms in this vector may be interpreted as probabilities, and therefore sum to one. In order to quantify the distribution of topics in the financial news on a given day, we computed a normalised sum of the distribution of topics over each document (paragraph) in the corresponding issue of the *Financial Times*. That is, from the set of documents $\mathcal{D}_t$ in the *Financial Times* issue on day t, we construct the vector

$$\rho_t \equiv \frac{1}{|\mathcal{D}_t|} \sum_{d \in \mathcal{D}_t} \theta_d, \tag{5.1}$$

where $|\mathcal{D}_t|$ denotes the number of documents in the set $\mathcal{D}_t$. This vector also sums to one, and quantifies the distribution of topics represented in the *Financial Times* on day t. This yields a K-dimensional vector ρ_t, which also sums to one, and quantifies the distribution of topics represented in the *Financial Times* on day t. The collection of all ρ_t form the rows of a matrix ρ. We display the first 100 rows of ρ in Figure 5.1.

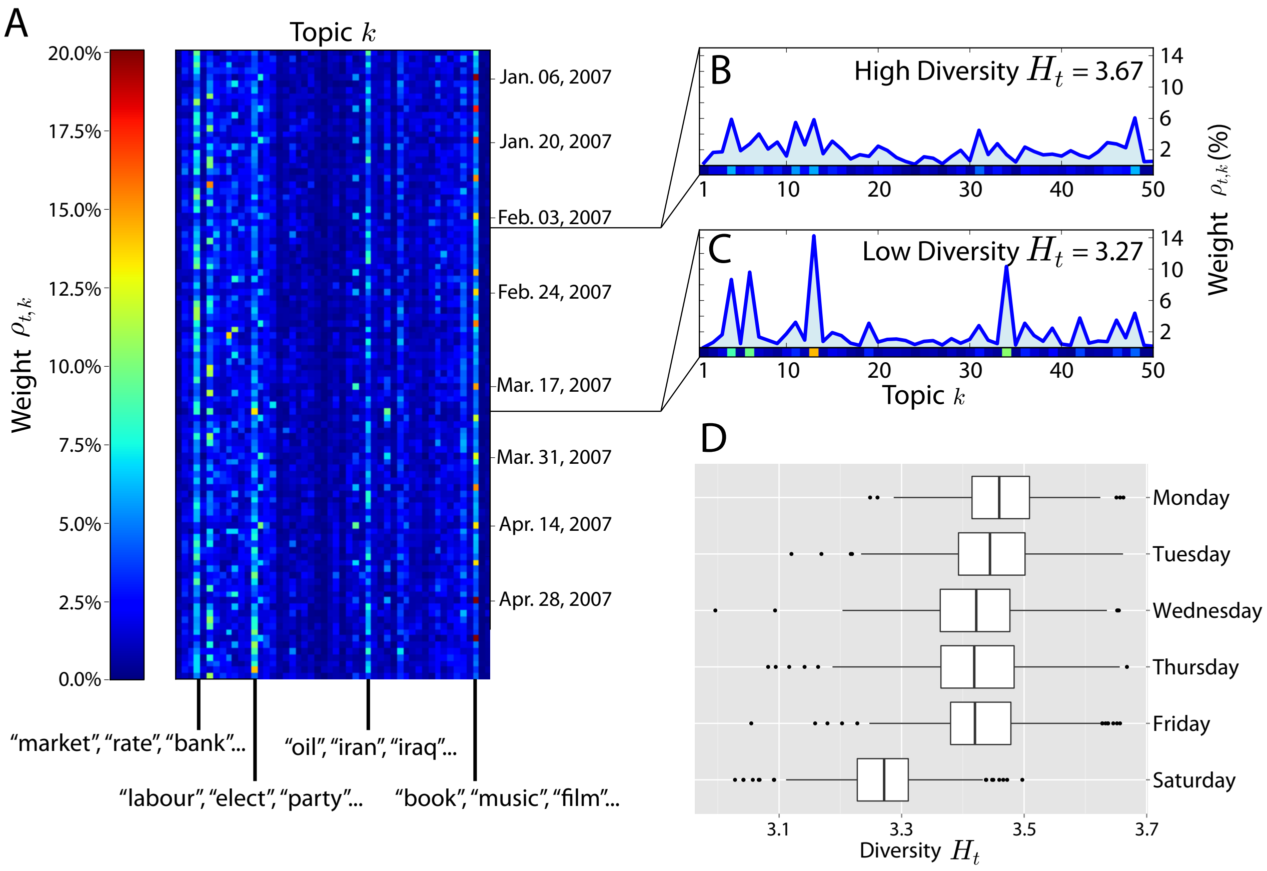
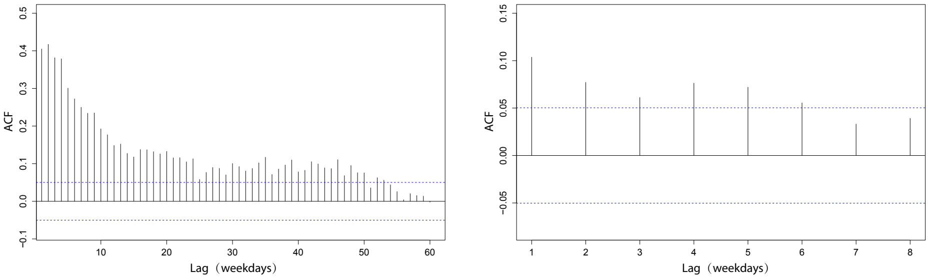


Figure 5.1: The prominence of topics in the *Financial Times*. (A) The weights $\rho_{t,k}$ of each topic k for each day t in the first 100 days of our dataset. We label the date of every other Saturday in the dataset, where the effect of weekend issues is visible. Sample topics are annotated with three of their top ten words by weight, showing the variety of topics in each daily issue of the news. (B and C) display the distributions of topics for two days exhibiting high and low news diversities H_t. (D) Boxplots of the news diversity H_t, aggregated by weekday. Weekend issues of the *Financial Times* exhibit characteristically low values of H_t, as a large portion of these issues are devoted to a small number of topics that appear infrequently in weekday issues of the news, such as the topic containing the words "book", "music", and "film".

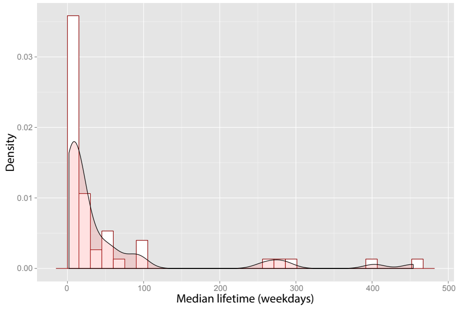
The matrix ρ provides rich information regarding both the detailed and large-scale structure of news to which investors, traders, and the public are exposed. The columns of ρ, for example, represent time series of weights for individual topics in the *Financial Times*. Analyses of these individual time series can provide insight into commonalities among ebbs and flows of stories into and out of public attention. Figure 5.2 depicts the autocorrelation functions (ACF) for two topic time series $\rho^T_{k,t}$. In Figure 5.2(a) we show the ACF for a topic regarding events in Egypt ("mubarak", "egypt", "protest",...), while in Figure 5.2(b) we show the same for a topic regarding events in Korea ("korea", "seoul", "kim",...). These two represent topics with slow and fast decays in their autocorrelation functions, respectively. We quantify the lifetime of a topic as the first lag (in weekdays) at which the ACF falls within the 95% confidence bands for an uncorrelated signal. In Figure 5.2(c) we show the distribution of all 50 topic lifetimes. Some lifetimes are on the order of years, but these tend to constitute topics which occur regularly in issues of the *Financial Times* (e.g., topics relating to weather reports, or market performance). 50% of topics have lifetimes shorter than 13 weekdays. Note that these calculations exclude weekend issues of the *Financial Times*. Such analyses, while simple, give valuable insight into "meta characteristics" that may be common to distinct topics in the news.

The question of interest here is how, if at all, the diversity of topics represented in a single issue of news interacts with financial market movements. To quantify this diversity, we seek to assign a single number to the topic distribution that measures the extent to which discussion is concentrated in few topics, or dispersed in many topics. A natural choice for this quantity is the Shannon entropy [116] of the distribution ρ_t. This quantity can be thought of as a measure of the uncertainty in ρ_t: for small values of the entropy, discussion in the news is focused in a narrow range of topics, lending a certain coherence to the text and resulting in low measured "uncertainties." For large values of the entropy, the topic distribution is relatively uniform, so that there is a comparatively wide diversity of topics represented in the text. The entropy of topic distributions derived from LDA has been applied in other contexts, such as the detection of "false" or semantically incoherent



(a) Topic 24 (“mubarak”, “egypt”, “protest”,...)

(b) Topic 33 (“korea”, “seoul”, “kim”,...)



(c) Distribution of lifetimes

Figure 5.2: Variation in topic lifetimes in the *Financial Times*. (A) ACF for the topic time series relating to events in Egypt, $\rho^T_{24,t}$. (B) The same for a topic relating to events in Korea, $\rho^T_{33,t}$. (C) The distribution of lifetimes for all 50 topics, defined as the first lag at which the corresponding ACF falls at or below the 95% confidence band for an uncorrelated signal.

documents that are constructed to deceive search engines [117]. In our case, the entropy, which we will refer to as the diversity, is computed as

$$H_t \equiv -\sum_{k=1}^{K} \rho_{t,k} \log(\rho_{t,k}) \tag{5.2}$$

where $\rho_{t,k}$ is entry k of the vector ρ_t, and represents the relative weight of topic k in the *Financial Times* on day t. We use the natural logarithm in this analysis, although alternative choices, such as the logarithm base 2, will simply scale measurements of H_t. In Figure 5.1 we plot the topic distributions ρ_t for two issues of the *Financial Times* exhibiting high and low diversities H_t. In Figure 5.1 we also examine the presence of weekly seasonalities in the news diversity H_t. We observe characteristically low values of the diversity in weekend issues of the *Financial Times*, as a large portion of these issues are devoted to a small number of topics that appear infrequently in weekday issues, such as the topic containing the words "book", "music", and "film". The weekday issues otherwise display only marginal seasonal effects.

We also examine the presence of monthly seasonalities in the diversity H_t in Figure 5.3, where we show the seasonal variation in the diversity H_t, excluding weekend issues. We observe little seasonal variation in H_t, although the diversity appears somewhat higher during the "silly season" in the summer months.

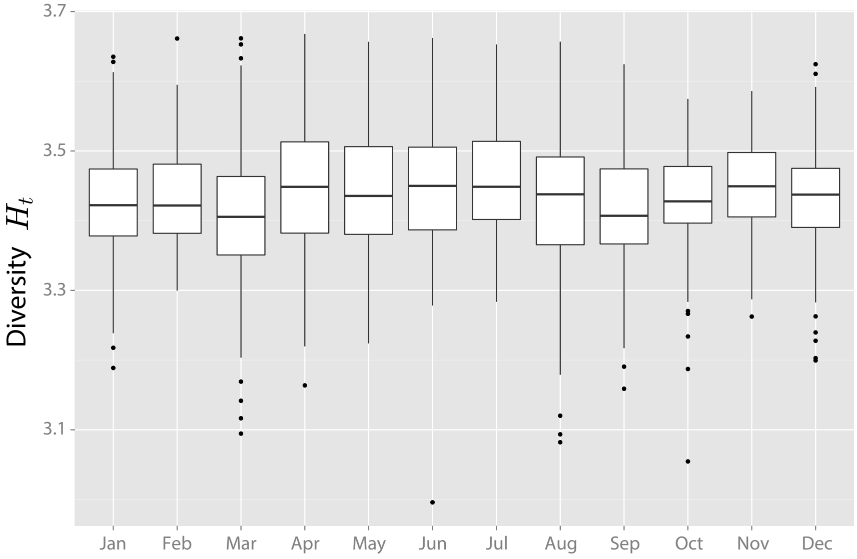


Figure 5.3: Boxplots of the diversity H_t, aggregated by month. In this figure we exclude weekend issues from our measurements of the diversity H_t. We see visually that there is little seasonal variation in H_t, although the diversity appears somewhat higher during the "silly season", or "slow news season" in the summer months.

5.1.2 Price changes of the FTSE drive changes in news diversity

To ease comparison with financial market movements and to exclude the influence of special "weekend issues" of the *Financial Times*, for the remainder of the analysis we exclude weekends from our analysis. In Figure 5.4 we display the topic vectors ρ_t for a subset of time, alongside the univariate time series of diversity, H_t. To give a picture of the interaction between market movements and H_t, we also plot the logarithmic returns r_t of the FTSE 100 index. These returns are defined as

$$r_t \equiv \log(P_t) - \log(P_{t-1}), \tag{5.3}$$

where P_t is the closing price of the FTSE 100 index on day t. Closing price figures and daily trade volumes for the FTSE 100 index were obtained from Yahoo Finance (`https://uk.finance.yahoo.com/`). From Figure 5.4, it seems visually clear that, especially during the economic crisis in 2008, there was a sharp decrease in the diversity H_t of the financial news. If a consistent relationship between the diversity H_t and the returns r_t exists, however, it is unclear whether fluctuations in the returns lead changes in H_t, or changes in H_t lead the returns.

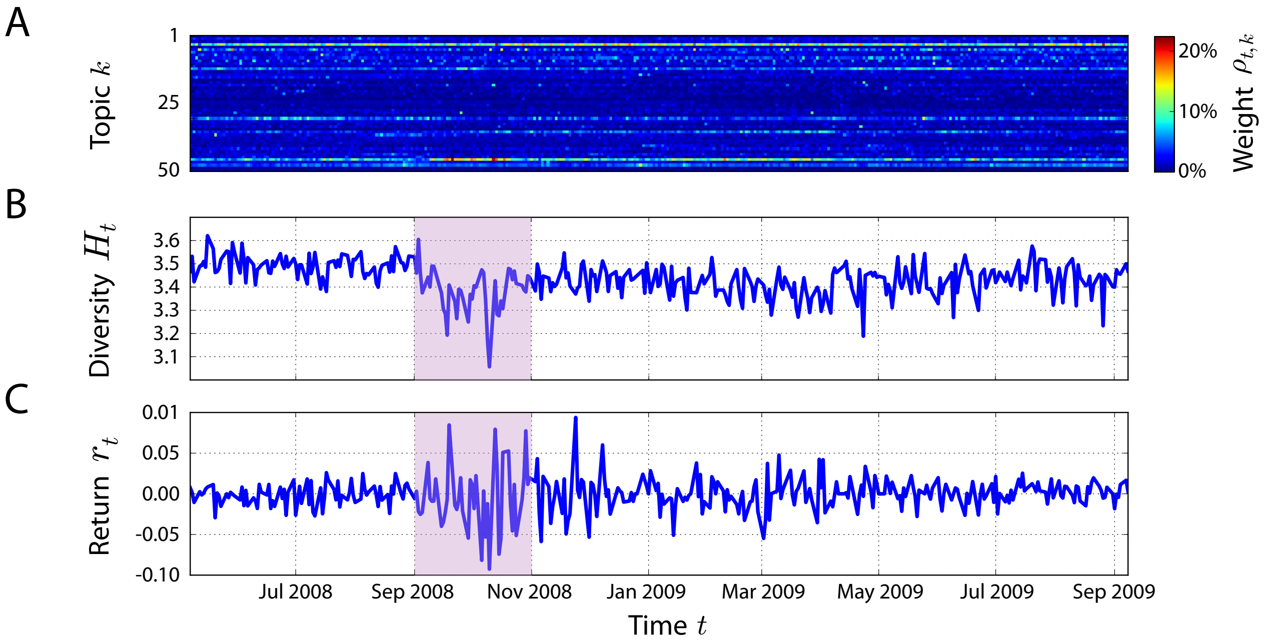


Figure 5.4: Changes in topic diversity across time. (A) The prominence of topics across a subset of the period under consideration, represented by the topic weights $\rho_{t,k}$ as in Fig. 5.1A. (B) News diversity, H_t, across time. In the shaded region we depict a period during the economic turmoil of 2008, in which the news diversity H_t exhibits a sharp downward trend. (C) The returns r_t of the FTSE 100 during the same period.

We first test one potential direction of the relationship: whether price changes in the FTSE 100, represented by the time series r_t, may drive fluctuations in the news diversity H_t. To isolate the influence of the returns r_t, it is necessary to determine the extent to which H_t may be modeled endogenously, i.e., using only its past values $\{H_{t-1}, H_{t-2}, ...\}$ in absence of any external inputs. An improvement on such a model using the returns r_t would suggest that a relationship exists between price changes in the FTSE 100 and the diversity of financial news. There exist general methods to model a time series using only its past values– autoregressive (AR) terms– as well as the model's own residuals– moving average (MA) terms. A popular, classical approach to modeling stationary time series in this way is to train an ARMA model [118, 119], which treats the time series as a linear combination of both AR and MA terms. We find that the changes in the diversity from day $t-1$ to day t, or $H_t - H_{t-1}$ form a stationary time series (KPSS test [119], α=0.05, testing null hypothesis of a stationary root against a unit-root alternative). We therefore model the differenced diversity $\Delta H_t \equiv H_t - H_{t-1}$. To determine how many elements of the lagged time series we must include in our model, we scan over several ARMA(p, q) models ($p =$ 1,...,5; $q =$ 1,...,5) and find that the Akaike information criterion (AIC) is minimised with a simple MA(1) process. This is corroborated by the autocorrelation function of ΔH_t [118], which exhibits an isolated negative spike at lag 1 and is otherwise featureless. We therefore fit

$$\Delta H_t = \epsilon_t + \beta_1 \epsilon_{t-1}, \tag{5.4}$$

finding $\beta_1 = -0.88 \pm 0.02$ using maximum-likelihood estimation [119]. We find no significant dependence of ΔH_t on the day of the week, as would be indicated by the presence of significant five-day seasonality. A plot of the signal ΔH_t, as well as its autocorrelation function (ACF) and partial autocorrelation function (PACF) [118], is provided in Figure 5.5.

The moving average process models the response of ΔH_t to random shocks, as quantified by the model residuals. Moreover, a simple least-squares linear regression of the residuals of the MA(1) model against the returns of the FTSE 100 on the previous day suggests that

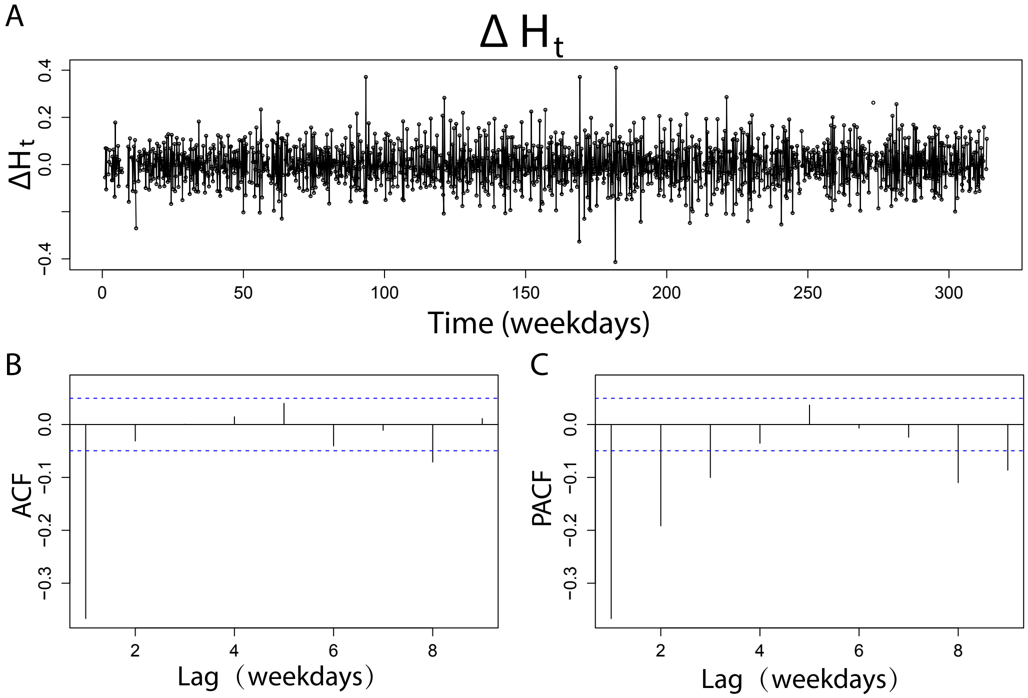


Figure 5.5: Time-series features of the differenced diversity ΔH_t. (A) Plot of the differenced diversity ΔH_t. (B) The ACF of ΔH_t. (C) The PACF of ΔH_t. The time series exhibits characteristics of a MA(1) process. The *Forecast* package for *R* was used in creating this plot [119].

these shocks are at least in part related to financial market movements. We find that in the model

$$\epsilon_t = \alpha_0 + \alpha_1 r_{t-1} + \eta_t,$$

with η_t an error term, the coefficient $\alpha_1 = 0.5 \pm 0.1$ is significant according to a standard t-test ($t = 3.8$, $N = 1450$, $p < 0.001$). This motivates us to include the previous-day returns of the FTSE 100 to our model of diversity fluctuations. We therefore fit

$$\Delta H_t = \epsilon_t + \gamma_1 \epsilon_{t-1} + \gamma_2 r_{t-1}, \tag{5.5}$$

finding $\gamma_1 = -0.87 \pm 0.02$ and $\gamma_2 = 0.30 \pm 0.07$. The coefficient γ_2 of the previous day's returns r_{t-1} is again significant according to a standard t-test ($t = 4.3$, $N = 1450$, $p <$

0.0001). The positive coefficients in models (5.1.2) and (5.5) indicate that decreases in diversity H_t follow stock market falls, while increases in diversity follow stock market rises, and bolsters quantitatively what we see qualitatively in Figure 5.4.

Ultimately, the utility of the FTSE 100 returns in predicting changes in the diversity ΔH_t can be decided in a comparison of errors from out-of-sample one-step forecasts between a purely endogenous model, and a model that includes the returns r_{t-1}. For this purpose, we fit both models (5.4) and (5.5) using only the first 70% of the dataset– from January 4, 2007 to March 16, 2011. We then compare one-step forecasts on the remainder of the data, from March 17, 2011 to December 31, 2012. A scan of ARMA models again finds that the MA(1) model best fits the training data, according to the AIC statistic.

We compare errors from the out-of-sample forecasts using the Diebold-Mariano test for predictive accuracy [119, 120] with a quadratic loss function. To interpret the results of this test we need not assume that the forecast errors are Gaussian, of zero-mean, or serially or contemporaneously uncorrelated [120]. We find marginal support for the hypothesis that including the previous-day returns of the FTSE 100, as in model (5.5), results in an increased out-of-sample accuracy (DM $= 1.4$, $N = 428$, $p = 0.078$). In Figure 5.6 we display the time dependence of the differences in squared out-of-sample errors between the purely endogenous model and the comparable model that includes the FTSE 100 returns. We find that the latter is relatively robust in its outperformance of the former, although there are several periods that contribute disproportionately to the effect.

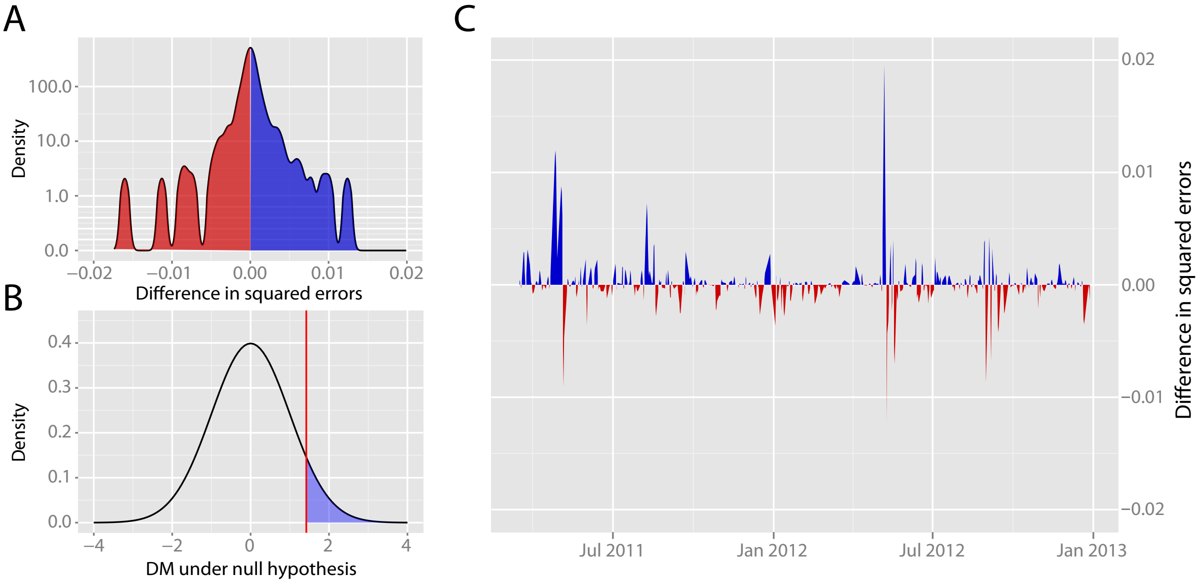


Figure 5.6: Improvement in out-of-sample model errors in forecasts of diversity fluctuations ΔH_t using returns from the FTSE 100. For each day in our out-of-sample test, we compare the errors of the the endogenous model of ΔH_t, which uses only its past values, to the errors of the same model that additionally incorporates the returns of the FTSE 100. (A) Distribution of squared errors when including returns from the FTSE 100, as subtracted from squared errors using the purely endogenous model on the same day. The extended positive tail of the distribution suggests that the FTSE 100 returns are important in explaining changes in news diversity. Distribution is represented using a Gaussian kernel density estimate. (B) Diebold-Mariano (DM) test statistic under the null hypothesis that the exogenous model fails to outperform the endogenous model, displaying visually the statistical significance of the result. (C) Time series of the squared errors that are aggregated in (A). Blue shaded regions indicate periods in the test data during which the model that incorporates the FTSE 100 returns outperforms the endogenous model, and red shaded regions indicate periods during which it failed to outperform the endogenous model. Incorporating the returns r_t of the FTSE 100 appears to consistently improve forecasts of diversity fluctuations ΔH_t, although there are several periods that contribute disproportionately to the effect.

The *Financial Times* is released daily at 5:00am London time, whereas the FTSE opens at 8:00am Monday through Friday. We find no evidence that changes in the news diversity H_t are related to subsequent price movements, as would be indicated by correlations between the returns r_t and same-day (Pearson $R = 0.05$) or previous day (Pearson $R = 0.002$) movements in the news diversity H_t. To bolster this conclusion, we repeat the above analysis, fitting an ARMA model to the returns r_t and testing the effect of the differenced diversity ΔH_t as an external regressor. We find that the coefficient of ΔH_t is insignificant in the ARMA model ($t = 1.73$, $N = 1459$, $p > 0.05$), and that the news diversity signal ΔH_t offers no improvements to out-of-sample predictions upon repetition of the Diebold-Mariano test (DM=-0.04, $N = 431$, $p > 0.1$).

5.1.3 Influence of individual topics

We therefore find evidence of a positive relationship between financial market movements and increases or decreases in the diversity of the next-day financial news. Our analysis suggests that market downturns are followed by a decrease in diversity, as discussion in the financial news is concentrated in a small number of topics; likewise, market upturns are followed by an increased diversity of topics in the financial news.

An alternative hypothesis that may also explain the results is that a small number of topics individually have a strong negative correlation with previous-day financial market movements. A hypothetical topic discussing market downturns, for instance, could naturally arise more often following negative returns of the FTSE 100. Spikes in activity for this topic could then decrease the measured diversity H_t, resulting in the observed pattern.

To search for such a topic, we consider separately the 50 columns ρ_k^T of ρ. Each of these columns corresponds to a time series of weights of a given topic in each issue of the *Financial Times*. For each topic, we compute the Pearson correlation between the differences $\Delta\rho_{k,t}^T \equiv \rho_{k,t}^T - \rho_{k,t-1}^T$ and the previous-day returns of the FTSE 100. A plot of the correlation coefficients measured for all topics is shown in Figure 5.7. Only one topic relating to the recent financial crisis of 2008 (“mortgage”, “loan”, “credit”, “debt”,...)

was found to be significantly impacted by previous-day returns of the FTSE 100 after FDR correction for multiple comparisons [5]. We find that the sign of this relationship is negative, implying a greater interest in this topic following falls in the FTSE 100, and vice-versa.

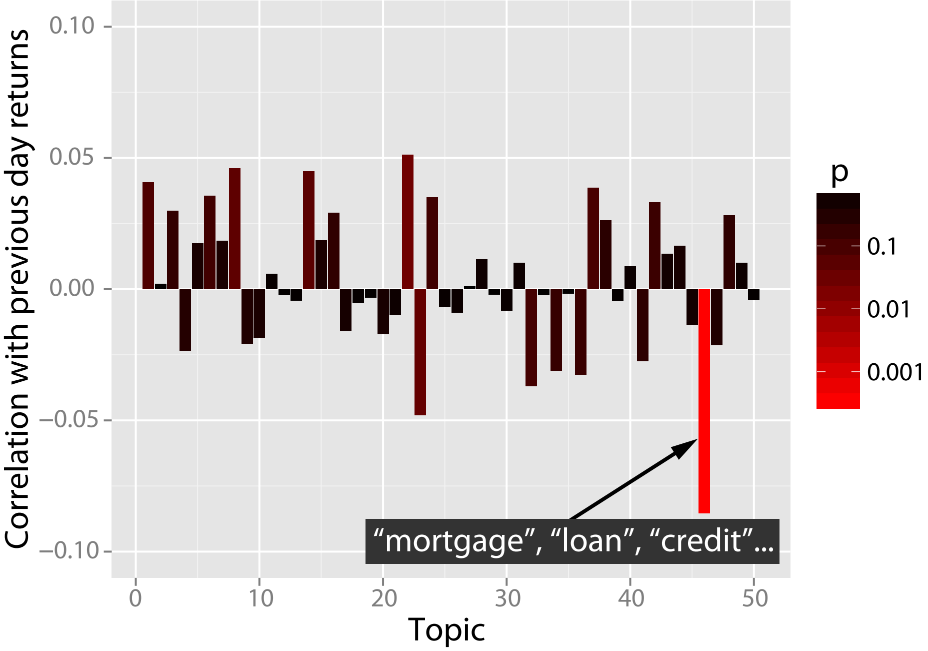


Figure 5.7: Identification of topics that correlate individually with previous-day market movements. We measure the correlations between the changes in topic weights, $\Delta\rho^T_{k,t}$, for each topic and the previous-day returns r_{t-1} of the FTSE 100. Bars are shaded by the corresponding p-value, computed using the Fisher transformation [121]. Only one topic (“mortgage”, “loan”, “credit”,...) was found to have a significant lagged relationship with previous returns of the FTSE 100 ($p < 0.05$ after FDR correction for multiple comparisons [5]). Removing this topic and repeating the analysis leaves the observed relationships between the diversity H_t and financial market movements qualitatively unchanged, providing support for the idea that topic diversity follows market movements in a way that is not captured by individual topics.

Table 5.1: In-sample model results with and without Topic 46 (“mortgage”, “loan”, “credit”, “debt”,...)

Model	All topics	Topic 46 removed
$\Delta H_t = \epsilon_t + \beta_1 \epsilon_{t-1}$	$\beta_1 = -0.88 \pm 0.02^{***}$	$\beta_1 = -0.91 \pm 0.02^{***}$
$\epsilon_t = \alpha_0 + \alpha_1 r_{t-1} + \eta_t$	$\alpha_0 = -(0.7 \pm 1.8) \times 10^{-3}$ $\alpha_1 = 0.5 \pm 0.1^{***}$	$\alpha_0 = -(0.009 \pm 1.8) \times 10^{-3}$ $\alpha_1 = 0.3 \pm 0.1^{*}$
$\Delta H_t =$ $\epsilon_t + \gamma_1 \epsilon_{t-1} + \gamma_2 r_{t-1}$	$\gamma_1 = -0.87 \pm 0.02^{***}$ $\gamma_2 = 0.30 \pm 0.07^{***}$	$\gamma_1 = -0.91 \pm 0.02^{***}$ $\gamma_2 = 0.20 \pm 0.07^{**}$

Note: Signif. codes: *** 0.001 ** 0.01 * 0.05

We check the influence of this topic on our previous results by removing it from the analysis. That is, we remove the entry corresponding to this topic from each topic vector θ_d, re-compute the matrix ρ and the diversity H_t, and repeat the comparison with the returns r_t of the FTSE 100. Exclusion of this topic leaves the results qualitatively unchanged, as is evident in Table 5.1. We again find that the differenced diversity ΔH_t is best modelled as an MA(1) process, according to the AIC statistic. Moreover, upon repetition of the Diebold-Mariano test on the errors of one-step out-of-sample forecasts, we find that inclusion of the previous-day returns of the FTSE 100 results in significantly greater accuracy in predicting changes in news diversity ΔH_t (DM $= 1.8$, $N = 428$, $p = 0.03$). We therefore find that changes in the diversity of topics in the news is influenced by previous-day stock market movements, independent of the reaction of individual topics.

5.1.4 News diversity relates to same-day trading volume

It is perhaps of greater interest to link the diversity of financial news, as quantified by the diversity H_t, to subsequent events in financial markets. Here, we present evidence that the diversity H_t can improve the accuracies of forecasts of daily trade volume in the FTSE 100.

We quantify daily trade volume again by differencing the total daily trade volume in the FTSE 100 after a log-transform:

$$v_t \equiv \log(V_t) - \log(V_{t-1}) \tag{5.6}$$

where V_t represents the total trade volume on day t. One order of differencing, as above, is sufficient to render the series $\log(V_t)$ stationary (KPSS test [119], α=0.05, testing null hypothesis of a stationary root against a unit-root alternative). The quantity v_t captures fluctuations in trading activity, irrespective of the directionality of price changes, and measures the extent to which investors elect to trade on day t.

As before, to isolate the predictive power of the differenced news diversity ΔH_t with respect to changes in daily trade volume v_t, we first examine the extent to which v_t may be modeled using only its past values $\{v_{t-1}, v_{t-2}, ...\}$. We find that the fluctuations in trading volume, v_t, forms a stationary series. A scan of ARMA models reveals the presence of both significant autoregressive and moving average terms; for this purpose we model v_t as an ARMA(1,1) process:

$$v_t = \epsilon_t + \alpha_1 v_{t-1} + \beta_1 \epsilon_{t-1}. \tag{5.7}$$

Using maximum-likelihood estimation [119], we find $\alpha_1 = 0.29 \pm 0.04$, and $\beta_1 = -0.83 \pm 0.03$.

The ARMA(1,1) model captures the degree to which we may model fluctuations in trading volume v_t endogenously, using only its past values. Following the analysis in section 5.1.2, we find that a significant portion of the variance of the residuals in model (5.7) can be explained using changes in the diversity ΔH_t. We find that in the model

$$\epsilon_t = \alpha_0 + \alpha_1 \Delta H_t + \eta_t, \tag{5.8}$$

the coefficient $\alpha_1 = -0.30 \pm 0.07$ is significant according to a standard t-test ($t = -4.0$,$N = 1459$, $p < 0.0001$). This motivates us to include the change in diversity ΔH_t, measured in the *Financial Times* on the morning of day t, in our model of the volume signal v_t for the same trading day. We therefore fit

$$v_t = \epsilon_t + \gamma_1 v_{t-1} + \gamma_2 \epsilon_{t-1} + \gamma_3 \Delta H_t, \tag{5.9}$$

finding $\gamma_1 = 0.29 \pm 0.04$, $\gamma_2 = -0.83 \pm 0.03$, and $\gamma_3 = -0.41 \pm 0.09$. The coefficient of ΔH_t is again significant according to a standard t-test ($t = -4.6$, $N = 1459$, $p < 0.0001$). The negative coefficients α_1 and γ_3 in models (5.8) and (5.9) indicate that falls in the diversity H_t tend to precede increased transaction volumes in the FTSE 100, and that increases in diversity tend to precede trading days in which transaction volumes are relatively diminished.

We supplement our in-sample tests through a comparison of errors from out-of-sample one-step forecasts between the purely endogenous model of v_t, and the model that incorporates fluctuations in the diversity ΔH_t. As in section 5.1.2, we fit both models (5.7) and (5.9) using only the first 70% of the dataset, and evaluate one-step forecasts on the remaining 30% of the dataset. Using the Diebold-Mariano test for predictive accuracy [119, 120] with a quadratic loss function, as before, we reject the hypothesis that inclusion of the diversity signal ΔH_t in model (5.9) fails to provide an increased out-of-sample accuracy (DM $= 2.2$, $N = 431$, $p = 0.013$). In Figure 5.8 we display how the difference in squared out-of-sample errors between these two models depends on time. We find that the model incorporating ΔH_t consistently outperforms the endogenous model throughout the test set.

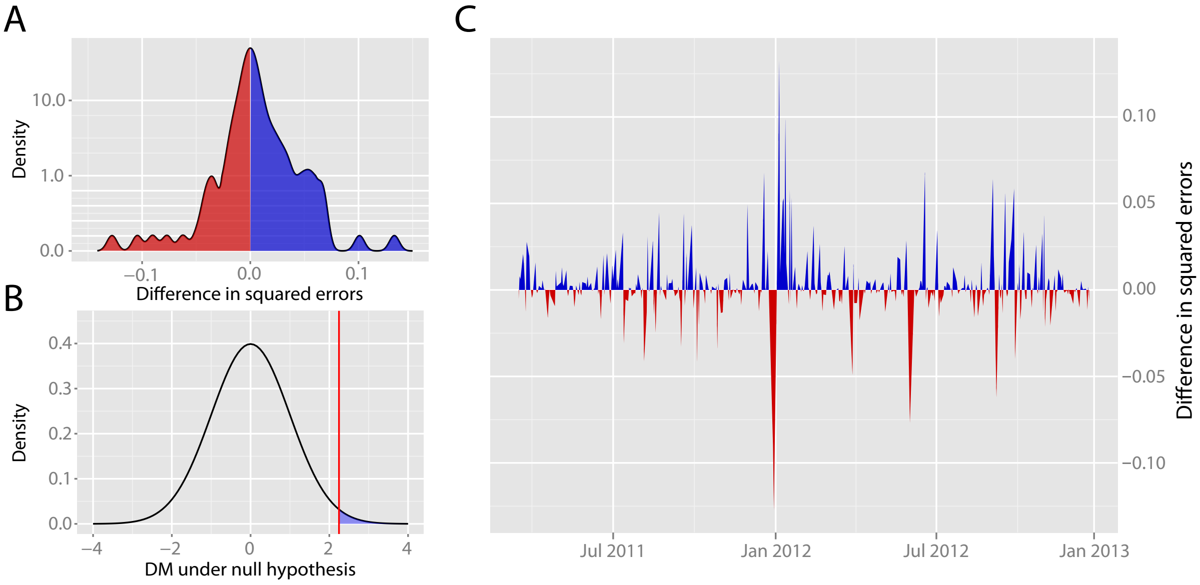


Figure 5.8: Improvement in out-of-sample model errors in forecasts of trade volumes v_t using changes in the diversity of news ΔH_t. For each day in our out-of-sample test, we compare the errors of the the endogenous model of v_t, which uses only its past values, to the errors of the same model that additionally incorporates the changes in news diversity ΔH_t. (A) Distribution of squared errors when including ΔH_t, as subtracted from squared errors using the purely endogenous model on the same day. The extended positive tail of the distribution suggests that fluctuations in news diversity are important in explaining changes daily trading volume. Distribution is represented using a Gaussian kernel density estimate. (B) Diebold-Mariano (DM) test statistic under the null hypothesis that the exogenous model fails to outperform the endogenous model, displaying visually the statistical significance of the result. (C) Time series of the squared errors that are aggregated in (A). Blue shaded regions indicate periods in the test data during which the model that incorporates the news diversity fluctuations ΔH_t outperforms the endogenous model, and red shaded regions indicate periods during which it failed to outperform the endogenous model. Incorporating the news diversity appears to consistently improve forecasts of daily trading volume in the FTSE 100.

A cursory analysis reveals no evidence for a reciprocal relationship in which the volume signal v_t anticipates changes in the diversity H_t. In particular, the correlation between v_{t-1} and next-day changes in diversity ΔH_t is low (Pearson $R = -0.02$). For a more thorough investigation, we include the volume signal v_{t-1} in our MA(1) model of ΔH_t and repeat the analysis in section 5.1.2. Here, we find that although previous-day changes in trade volume are significant when modeling the news diversity ΔH_t in-sample ($\gamma_2 = -0.025 \pm 0.007$, $t = -3.5$, $p < 0.001$), they fail to offer any advantage in out-of-sample predictions upon repetition of the Diebold-Mariano test (DM $= 0.87$, $N = 428$, $p > 0.1$).

5.1.5 Discussion

We find that using topic modelling to quantify the diversity of subjects in the financial news yields fruitful insights into the relationship between investors and the media. Indeed, we find a consistent reaction of the news diversity to falls in the stock market, as discussion concentrates in a small number of topics following drops in the price of the FTSE 100. Moreover, we find that the diversity of topics in the news has utility as a leading indicator of fluctuations in trading volume.

Of interest is the asymmetry in the result that market downturns, and not market upturns, tend to lead to falls in the diversity of financial news. Our finding may provide insight into the psychological and commercial forces that shape the dissemination of information to investors: while much space may be dedicated to the discussion of “bad news” in the market and its perceived causes, on days of comparative “good news” attention is liable to shift to a diversity of topics of interest.

Although we restrict our focus to financial news, we make no efforts to filter topics based on their semantic content. Our approach weights all topics equally, regardless of whether they refer to politics, war, or the economy. Our analysis suggests that the news that drives the actions of investors may not always have obvious semantic connections with finance or the economy. Abstracting away from individual topics, we find that cohesion of financial news in particular can be related to recent market downfalls and same-day rises in trading volume.

The *Financial Times* is one publication in a sea of sources for financial news, with its own biases and dispositions. Nonetheless, boasting an average daily readership of 2.2 million people worldwide [122], it offers a reliable sampling of the information to which investors and the public are exposed.

Our analysis is by no means exhaustive, in the sense that there are many ways to measure activity in financial markets that we did not consider. Changes in price and daily transaction volume are among the simplest measures, and it is for that reason that they were pursued in this work. We suggest that extensions of these analyses could incorporate more nuanced measures of financial activity, such as the prices of various futures contracts. The robustness of these results in other forms of news, such as discussion on social media, could also be studied. Information-gathering processes, as reflected in online search activity, could additionally offer insight into "herding effects" in public sentiment and its relationship to events in the real world.

Going beyond the study of individual keywords or even groups of keywords, the results of our approach suggest that the exploration of "meta-characteristics" of news, of which the diversity is one example, may prove a fruitful avenue for research. We suggest that studies of additional features, such as the lifetime of news stories, may shed light on public engagement with different forms of media surrounding a range of real-world events.

5.2 Application to Internet search data

Financial crises arise from the complex interplay of decisions made by many individuals. Stock market data provide extremely detailed records of such decisions, and as such, both these data and the complex networks which underlie them have generated considerable scientific attention [123]– [142]. However, despite their gargantuan size, such datasets capture only the final action taken at the end of a decision making process. No insight is provided into earlier stages of this process, where traders may gather information to determine what the consequences of various actions may be [143].

Nowadays, the Internet is a core information resource for humans worldwide, and much information gathering takes place online. For many, search engines such as *Google* act as a gateway to information on the Internet. *Google*, like other search engines, collects extensive data on the behavior of its users [144]– [147], and some of these data are made publicly available via its service *Google Trends*. These datasets catalog important aspects of human information gathering activities on a global scale, and thereby open up new opportunities to investigate early stages of collective decision making.

In line with this suggestion, previous studies have shown that the volume of search engine queries for specific keywords can be linked to a range of real world events [148], such as the popularity of films, games and music on their release [149], unemployment rates [150], reports of flu infections [151], and trading volumes in US stock markets [152, 153]. A recent study showed that Internet users from countries with a higher per capita gross domestic product (GDP), in comparison with Internet users from countries with a lower per capita GDP, search for proportionally more information about the future than information about the past [154].

Here, we investigate whether we can identify topics for which changes in online information gathering behavior can be linked to the sign of subsequent stock market moves. A number of recent results suggests that online search behavior may measure the attention of investors to stocks before investing [155]– [157]. We build on a recently-introduced method [155] that uses trading strategies based on search volume data to identify online precursors for stock market moves. This previous analysis of search volume for 98 terms of varying financial relevance suggests that, at least in historic data, increases in search volume for financially relevant search terms tend to precede significant losses in financial markets [155]. Similarly, Moat et al. [158] demonstrated a link between changes in the number of views of *Wikipedia* articles relating to financial topics and subsequent large stock market moves. The importance of the semantic content of these *Wikipedia* articles is emphasized by a parallel analysis, which finds no such link for data from *Wikipedia* pages relating to actors and filmmakers.

Financial market systems are complex however, and trading decisions are usually based on information about a huge variety of socio-economic topics and societal events. The initial examples above [155, 158] focus on a narrow range of pre-identified financially related topics. Instead of choosing topics for which search data should be retrieved and investigating whether links exist between the search data and financial market moves, here we present a method which allows us to identify topics for which levels of online interest change before large movements of the Standard & Poor's 500 index (S&P 500). Though we restrict ourselves to stock market moves in this study, our methodology can be readily extended to determine topics which Internet users search for before the emergence of other large scale real-world events.

Our approach is as follows. Firstly, we take a large online corpus, *Wikipedia*, and use a well-known technique from computational linguistics [159] to identify lists of words constituting semantic topics within this corpus. Secondly, to give each of these automatically identified topics a name, we engage users of the online service *Amazon Mechanical Turk*. Thirdly, we take lists of the most representative words of each of these topics and retrieve data on how frequently *Google* users searched for the terms over the past nine years. Finally, we use the method introduced in [155] to examine whether the search volume for each of these terms contains precursors of large stock market moves. We find that our method is capable of automatically identifying topics of interest before stock market moves, and provide evidence that for complex events such as financial market movements, valuable information may be contained in search engine data for keywords with less obvious semantic connections.

5.2.1 Method

To extract semantic categories from the online encyclopedia *Wikipedia*, we build on a well-known observation [159] that words which frequently appear together in newspaper articles, encyclopedia entries, or other kinds of documents tend to bear semantic relationships to each other. For example, a document containing the word "debt" may be more likely to also contain other words relating to finance than other words relating to, say, fruit. For such an analysis of semantic relationships to produce meaningful results, the overall frequency of terms must also be taken into account. To incorporate these insights, we analyze the semantic characteristics of all the articles and words in the English version of *Wikipedia* using a modeling approach called Latent Dirichlet Allocation (LDA) [159]. We configure the LDA to extract 100 different semantic topics from *Wikipedia*. We note that individual words can occur in multiple semantic topics.

Using the publicly available service *Google Trends*, we obtain data on the frequency with which *Google* users in the United States search for each of these terms. We analyze data generated between 4 January 2004, the earliest date for which *Google Trends* data are available, and 16 December 2012. We consider data at a weekly granularity, the finest granularity at which *Google Trends* provides data for the majority of search terms.

Google Trends provides data on search volume using a finite integer scale from 0 to 100, where 100 represents the highest recorded search volume for all terms in a given *Google Trends* request. If search volume time series for low frequency keywords are downloaded in isolation from other keywords, noisy data can result, as only a small number of searches is required for a unit change in search volume to be registered. To avoid this problem, we download search volume data for the high frequency term "google" alongside search volume data for each of our terms. In this way, we ensure that the value 100 represents the maximum search volume for this high frequency term. However, we also find that the mean search volume for terms in 45 of our extracted topics is too low to register on this "google" based scale, having a value less than one. Below, we describe analyses based on the remaining 55 topics.

To generate labels for the topics, we make use of the online service *Amazon Mechanical Turk*. This service allows small tasks to be taken on by anonymous human workers, who receive a small payment for each task. Through this service, 39 unique human workers provided topic names for the 55 sets of words identified above.

To compare changes in search volume to subsequent stock market moves, we implement for each of these terms the trading strategy introduced in [155]. We use for our analyses the U.S. equities market index S&P 500 which includes 500 leading companies in leading industries of the U.S. economy. We hypothetically trade the S&P 500 Total Return index (SPXT) which also accounts for the reinvestment of dividends. In this strategy, we first use *Google Trends* to measure how many searches $n(t)$ occurred for a chosen term in week t. To quantify changes in information-gathering behavior, we compute the relative change in search volume $\Delta n(t, \Delta t) = n(t) - N(t-1, \Delta t)$ with $N(t-1, \Delta t) = (n(t-1) + n(t-2) + \cdots + n(t - \Delta t))/\Delta t$. We sell the SPXT at the closing price $p(t)$ on the first trading day of week t, if $\Delta n(t-1, \Delta t) > 0$, and buy the index at price $p(t+1)$ at the end of the first trading day of the following week. If instead $\Delta n(t-1, \Delta t) < 0$, then we buy the index at the closing price $p(t)$ on the first trading day of week t and sell the index at price $p(t+1)$ at the end of the first trading day of the coming week. If we sell at the closing price $p(t)$ and buy at price $p(t+1)$, then the arithmetic cumulative return R changes by a factor of $p(t)/p(t+1)$. If we buy at the closing price $p(t)$ and sell at price $p(t+1)$, then the arithmetic cumulative return R changes by a factor of $p(t+1)/p(t)$. The maximum number of transactions per year when using our strategy is only 104, allowing a closing and an opening transaction per week; hence, for the purposes of this analysis of the relationship between search volume and stock market moves, we neglect transaction fees.

We compare the cumulative returns from such strategies with the cumulative returns from 1,000 realizations of an uncorrelated random strategy. In the random strategy, a decision is made each week to buy or sell the SPXT. The probability that the index will be bought rather than sold is 50%, and the decision is unaffected by decisions in previous weeks.

For each of the 55 topics, we calculate R for each of the 30 trading strategies, each based on search volume data for one term belonging to the topic. Strategies trade weekly on the SPXT from January 2004 to December 2012, using $\Delta t = 3$ weeks. We report the arithmetic cumulative returns, $R-1$, in percent. We also report the mean arithmetic cumulative return $\bar{R}$ for each topic.

5.2.2 Results

Figure 5.1 depicts the distributions of R for each of the 55 topics. We compare the arithmetic cumulative returns for search volume based strategies to the distribution of arithmetic cumulative returns from the random strategy using two-sample Wilcoxon rank sum tests, with FDR correction for multiple comparisons, as described in detail in [5], among a range of topics and values of the parameter Δt. We find that strategies based on keywords in the categories Politics I (e.g., Republican, Wisconsin, Senate,...; mean return = 56.4%; $W = 20713$, $p = 0.01$) and Business (e.g., business, management, bank,...; mean return = 38.6%; $W = 19919$, $p = 0.04$) lead to significantly higher arithmetic cumulative returns than those from the random strategy, suggesting that changes in search volume for keywords belonging to these topics may have contained precursors of subsequent stock market moves. These two distributions are colored by their $\bar{R}$.

We examine the effect of changing the value of Δt. In Fig. 5.1B, we depict the results of varying Δt between 1 and 15 weeks for all 55 topics. We color cells according to $\bar{R}$ for a given topic, using a given value of Δt. Where no color is shown, no significant difference is found between the distribution of arithmetic cumulative returns from a random strategy and the distribution of arithmetic cumulative returns for the topic's strategies with the given value of Δt ($p \geq 0.05$). We find that terms within the Business category result in significant values of R for values of Δt of 2 to 15 weeks (all Ws ≥ 19278, all ps < 0.05), with the exceptions of $\Delta t = 4$ weeks and $\Delta t = 12$ weeks. Terms within the category Politics I result in significant returns for $\Delta t = 2$ to 15 weeks (all Ws ≥ 20422, all ps < 0.05), with the exceptions of $\Delta t =$4, 5, and 7 weeks. The relationship between changes in search volume for these topics and movements in the SPXT is therefore reasonably robust to changes in Δt. We also find that terms within the category Politics II (e.g., party, law, government,...) result in significant values of R for $\Delta t = 6$ weeks and $\Delta t = 8$ to 15 weeks (all Ws ≥ 20144, all ps < 0.05). For some values of Δt, we find significant values of R for terms belonging to the categories Medicine, Education I and Education II. The significance of these values of R is however highly dependent on the value of Δt.

As a check of our procedure for multiple hypothesis testing, we repeat the above analysis using randomly-generated search volumes. We construct $55 \cdot 30 = 1650$ time series of search volume data by independently shuffling the time series of search volume for each word in each topic. We then re-create Fig. 5.1A and 5.1B using these 55 "topics" in Fig. 5.1C and 5.1D, respectively. We find that, after FDR correction, no such topic deviates significantly from the cumulative returns from an uncorrelated random strategy.

We next investigate the Politics I, Politics II, and Business categories more carefully. In particular, we examine the effect of changing the period of time during which we analyze this relationship. In Fig. 5.2, we depict the results of using a range of moving four year windows between 2004 and 2012 for the Business, Politics I and Politics II topics with Δt held at 3 weeks. We include an additional time window, from January 2010 to December 2013, to check the present-day performance of the strategies. We depict distributions of R for these periods using a kernel density estimate. As in Fig. 5.1, we compare the distributions of R from each topic with the distribution of R from random strategies. Terms in the Politics I category result in significant values of R (all Ws ≥ 18839, all ps < 0.05 after FDR correction) for all time windows, with the exception of 2009-2012 and 2010-2013. Terms relating to Business result in significant values of R for the periods 2004-2007, 2006-2009, 2007-2010, and 2008-2011 (all Ws ≥ 18511, all ps < 0.05, FDR correction applied). Lastly, terms in the

Politics II category result in significant values of R for the periods 2005-2008, 2006-2009, 2007-2010, and 2008-2011 (all Ws ≥ 19196, all ps < 0.05, FDR correction applied). Our results provide evidence of a historical relationship between the search behavior of *Google* users and financial market movements. However, our analyses suggest that the strength of this relationship has diminished in recent years, perhaps reflecting increasing incorporation of Internet data into automated trading strategies.

We additionally calculate regressions to control for other effects and to check the robustness of our results on a weekly scale. This approach also permits us to explore relationships between the magnitude of the change in search volume and the magnitude of the subsequent return, in addition to its sign. At each week t we monitor the mean relative change in search volume, $x_t \equiv \Delta n(t, \Delta t)/N(t-1, \Delta t)$, for the Politics I, Politics II, and Business topics. We regress the percentage return of the SPXT in the subsequent week, $r_{t+1} \equiv [(p(t+1) - p(t))/p(t)] \cdot 100\%$, against this signal. We also include the S&P 500 Volatility Index (VIX) as a regressor:

$$r_{t+1} = \beta_0 + \beta_1 x_t + \beta_2 \mathrm{VIX}_t + \epsilon_t$$

where ϵ_t is an error term.

Using the mean relative change in search volume for the Politics I category as our signal $x_{\text{Politics I}}$, we report a significantly negative coefficient of -2.80 ($t = -2.65$, $p = 0.024$, Bonferroni correction applied). Using instead the Business category for our signal x_{Business}, we report a significantly negative coefficient of -5.34 ($t = -2.61$, $p = 0.027$, Bonferroni correction applied). We find that the signal generated by the Politics II category $x_{\text{Politics II}}$, however, is not significantly related to subsequent stock market moves, according to this analysis (t = -2.02, p = 0.13, Bonferroni correction applied). The coefficient β_2 of the volatility index VIX was insignificant in all regressions ($p > 0.35$). We detail the results of the regressions in Table 1. Table 2 provides the median, 5% and 95% quantiles for the absolute value of the test-statistics $|t|$ as well as R^2 for all 55 regressions carried out using the same shuffled search volume data that is represented in Fig. 1C. We find that the

Table 5.2: Regression results using search volume signals $x_{\text{Politics I}}$, x_{Business}, and $x_{\text{Politics II}}$.

Regressor	Estimate	Std. Error	t-statistic	Pr($> \|t\|$)	R^2
$x_{\text{Politics I}}$	-2.80	1.06	-2.65	0.024*	0.0169
x_{Business}	-5.34	2.05	-2.61	0.027*	0.0164
$x_{\text{Politics II}}$	-1.65	0.816	-2.02	0.13	0.0107

Note: Signif. codes: ** 0.01 * 0.05 . 0.10

statistics $|t|$ and R^2 for the Politics I and Business topics fall within the top 5% of values obtained using the shuffled search volumes.

Table 5.3: Quantiles of test-statistics $|t|$ and R^2 using randomized search volume data.

Quantile	$\|t\|$	R^2
5%	0.0608	0.00190
Median	0.796	0.00326
95%	2.56	0.0159

To examine the distributions of the test statistics for the Politics I, Business, and Politics II topics, we implement a block bootstrap procedure [160] in which we construct surrogate time series by circularly shifting our signals x_t (i.e., at each shift, the final entry is moved to the first position). We examine the distributions of t-statistics and coefficients of determination R^2 under all such shifts, providing a safeguard against spurious results due to auto-correlative structure in the data. The median, 5%, and 95% quantiles are reported in Table 3, where we find that all observed test-statistics fall within the top 5% of bootstrapped results.

As a final check of our results, we apply the Hansen test for superior predictive ability [160]. For this test we construct 1,000 re-samplings of the data, with replacement, using a stationary bootstrap technique [161, 162]. The continuous block length of the pseudo time-series is chosen to be geometrically distributed with parameter $q = 0.001$, of the order of the inverse length of the time series, in order to preserve effects due to autocorrelation. For each of the topics Politics I, Business, and Politics II, we test the universe of trading strategies

Table 5.4: Comparison of observed test statistics with those obtained from bootstrapping procedure.

Statistic	$x_{\text{Politics I}}$	x_{Business}	$x_{\text{Politics II}}$
Observed $\lvert t\rvert$	2.65	2.61	2.02
5% $\lvert t\rvert$	0.0746	0.0716	0.0577
Median $\lvert t\rvert$	0.627	0.655	0.623
95% $\lvert t\rvert$	1.95	2.13	1.94
Observed R^2	0.0169	0.0164	0.0107
5% R^2	0.00191	0.00191	0.00190
Median R^2	0.00275	0.00282	0.00273
95% R^2	0.0101	0.0115	0.0100

generated by all 30 words in the topic against both a random strategy and a buy and hold strategy. We find that a random strategy is significantly outperformed by strategies generated by words in the Politics I ($T^{\text{SPA}} = 9.06$, $p < 0.001$), Business ($T^{\text{SPA}} = 9.53$, $p < 0.001$), and Politics II ($T^{\text{SPA}} = 6.47$, $p < 0.001$) topics. However, we only find marginal support for these strategies significantly outperforming a buy-and-hold strategy (Politics I: $T^{\text{SPA}} = 2.34$, $p = 0.085$; Business: $T^{\text{SPA}} = 2.62$, $p = 0.071$; Politics II: $T^{\text{SPA}} = 1.23$, $p = 0.143$).

5.2.3 Discussion

In summary, we introduce a method to mine the vast data Internet users create when searching for information online in order to identify topics in which levels of online interest change before stock market moves. We draw on data from *Google* and *Wikipedia*, as well as *Amazon Mechanical Turk*. Our results are in line with the intriguing possibility that changes in online information gathering behavior relating to both politics and business or finance were historically linked to subsequent stock market moves. Crucially, we find no robust link between stock market moves and search engine queries for a wide range of further semantic topics, all drawn from the English version of *Wikipedia*.

We note that the overlap between words in the topics Politics I (e.g., Republican, Wisconsin, Senate,. . .) and Politics II (e.g., party, law, government,. . .) is small, as the two topics, containing thirty words each, share only four words: “president,” “law,” “election,” and “democratic.” Despite this, our method identifies relationships between both politics-related topics and stock market moves, providing further evidence of the importance of underlying semantic factors in keyword search data. We note that a third topic related to politics, Politics III, was not flagged by our method. A close inspection reveals that this topic in fact bears more relevance to politics in the United Kingdom, containing the keywords “parliament,” “british,” “labour,” “london,” etc. This finding is in line with the suggestion that changes in online information gathering specifically relating to politics in Britain may not bear a strong relationship to subsequent financial market moves in the U.S.

Our results provide evidence that for complex events such as large financial market moves, valuable information may be contained in search engine data for keywords with less obvious semantic connections to the event in question. Overall, we find that increases in searches for information about political issues and business tended to be followed by stock market falls. One possible explanation for our results is that increases in searches around these topics may constitute early signs of concern about the state of the economy - either of the investors themselves, or as society as a whole. Increased concern of investors about the state of the economy, or investors’ perception of increased concern on a society wide basis, may lead to decreased confidence in the value of stocks, resulting in transactions at lower prices. However, our analyses provide evidence that the strength of this relationship has diminished in recent years, perhaps reflecting increasing incorporation of Internet data into automated trading strategies.

The method we present here facilitates in a number of ways the interpretation of the relationship between search data and complex events such as financial market moves. Firstly, the frequency of searches for a given keyword can grow and decline for various reasons, some of which may or may not be related to a real world event of interest. This method allows us to abstract away from potentially noisy data for individual keywords, and identify underlying semantic factors of importance. Secondly, our method allows us to extract subsets of search data of relevance to real world events, without privileged access to full data on all search queries made by *Google* users. By identifying representative keywords for a range of semantic topics, such analyses can be carried out despite limitations on the number of keywords for which search data can be retrieved via the *Google Trends* interface. Thirdly, our semantic analysis is based on simple statistics on how often words occur in documents alongside other words. As a result, the analysis presented could be carried out in languages other than English—for example, using other editions of *Wikipedia*—with no extra modifications to the approach required. We suggest that extensions of these analyses could offer insight into large scale information flow before a range of real-world events.

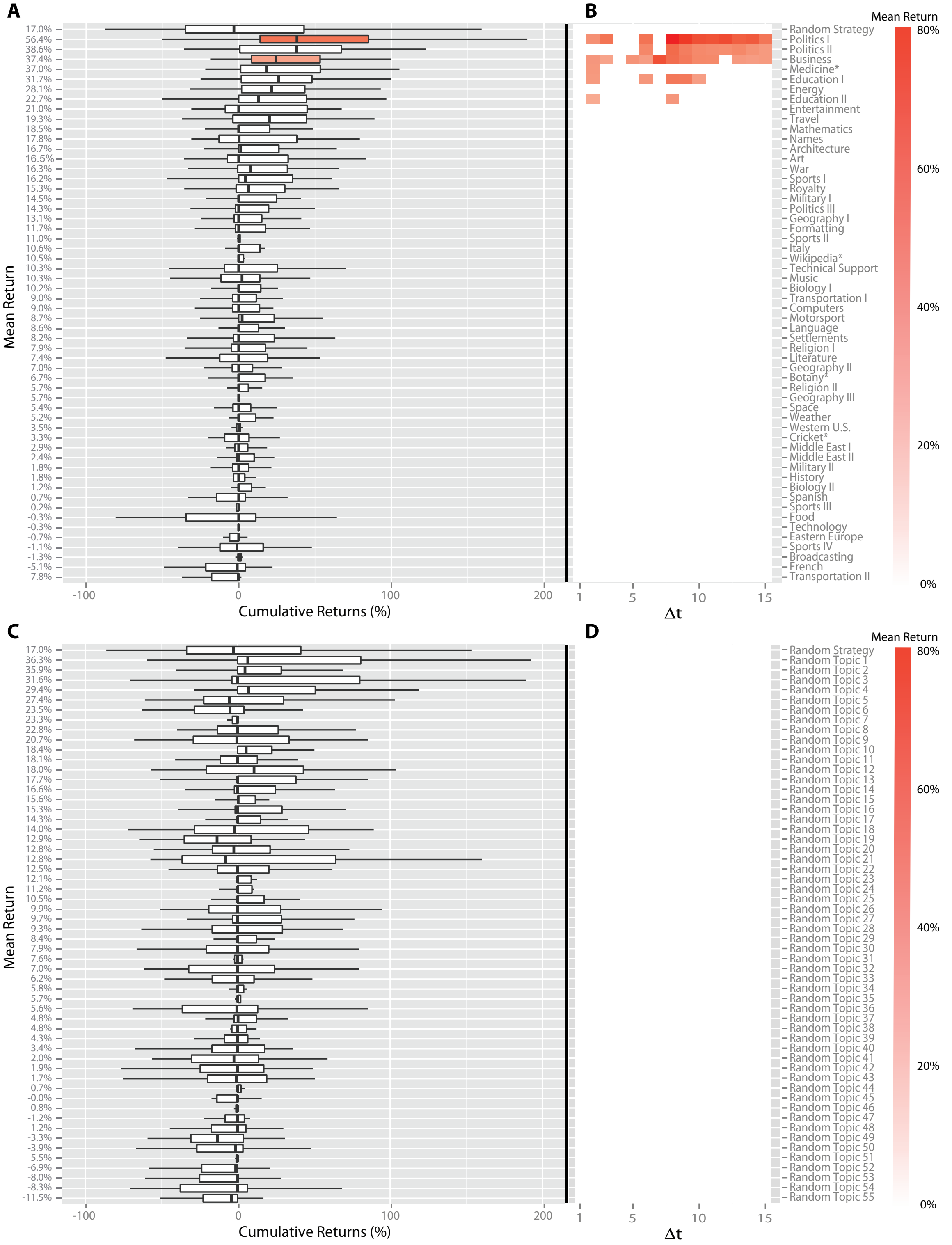


Figure 5.9: (Caption next page.)

Figure 5.9: (Previous page.) *Google Trends* based trading strategies for 55 different semantic topics. (A) For each topic, we depict the distribution of cumulative returns from 30 trading strategies, each based on search volume data for one term belonging to the topic. Strategies trade weekly on the SPXT from 2004 to 2012, using $\Delta t = 3$. We show in the top row the distribution of cumulative returns for a random strategy. The mean percentage returns for each topic appear on the left column. We compare the cumulative returns for search volume based strategies to the distribution of cumulative returns from the random strategy using two-sample Wilcoxon rank sum tests, with FDR correction for multiple comparisons among a range of topics and values of the parameter Δt. We find that strategies based on keywords in the categories Politics I ($W = 20713$, $p = 0.01$) and Business ($W = 19919$, $p = 0.04$), shown in red, lead to higher cumulative returns than the random strategy. (B) Colored cells denote values of Δt for which the cumulative returns for a semantic topic are significantly higher than those of a random strategy ($p < 0.05$). Terms within the categories Business, Politics I and Politics II result in significant returns across a range of values of Δt. (C) and (D) same as (A) and (B), but using shuffled search volumes and finding no significant "topics."

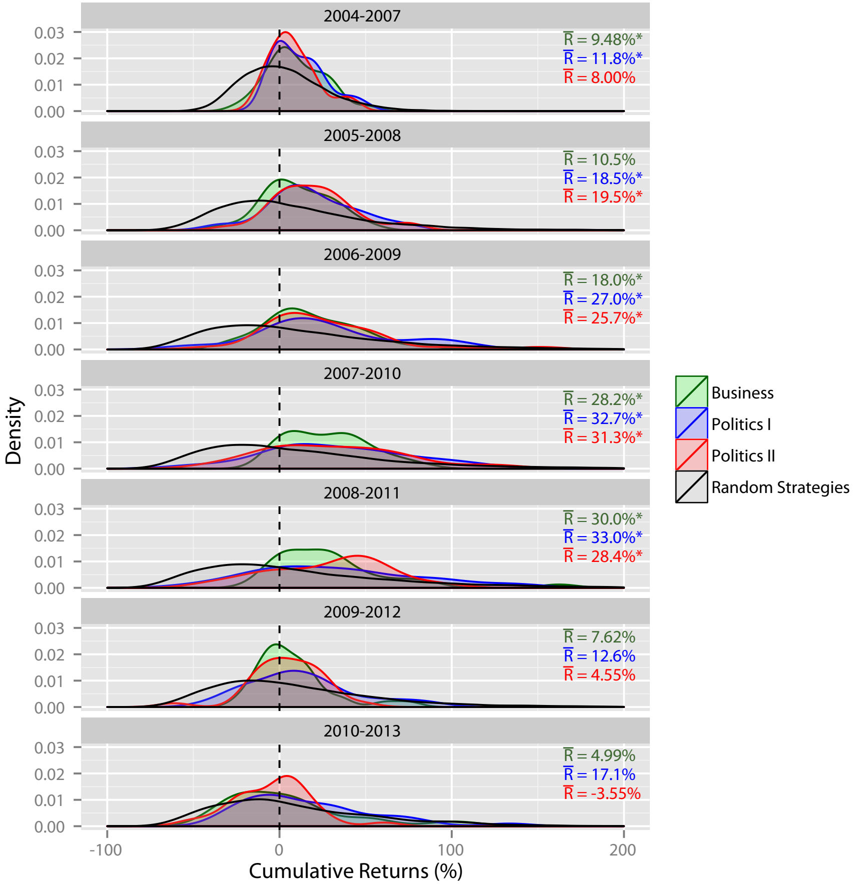


Figure 5.10: Effect of changing time window on returns. For the Business, Politics I, and Politics II topics, we depict the distribution of cumulative returns from the corresponding trading strategies in six overlapping four-year time windows. Distributions are plotted using a kernel density estimate, with a Gaussian kernel and bandwidth calculated with Silverman's rule of thumb [164]. Strategies trade weekly on the SPXT, using $\Delta t = 3$. The distribution of cumulative returns for a random strategy is also shown in each time window. The mean percentage return $\bar{R}$ for each topic is provided on the right of the figure. We compare the cumulative returns for search volume based strategies to the distribution of cumulative returns from the random strategy using two-sample Wilcoxon rank sum tests, with FDR correction for multiple comparisons. Terms in the Politics I category result in significant returns (all Ws ≥ 18839, all ps < 0.05 after FDR correction) for all time windows, with the exception of 2009-2012 and 2010-2013. Terms relating to Business result in significant returns for the periods 2004-2007, 2006-2009, 2007-2010, and 2008-2011 (all Ws ≥ 18511, all ps < 0.05 after FDR correction). Lastly, terms in the Politics II category result in significant returns for the periods 2005-2008, 2006-2009, 2007-2010, and 2008-2011 (all Ws ≥ 19196, all ps < 0.05 after FDR correction).

Appendices

Appendix A

Stemmed word distributions from LDA of *The Financial Times*

Below we provide the top ten (stemmed) words for each of the 50 topics extracted from the Latent Dirichlet Allocation of *The Financial Times* in Chapter 5.1.

Table A.1: LDA Topics from Chapter 5.1

Topic	Top 10 words
1	“ge”,“dhabi”,“abu”,“nt”,“goldman”,“sach”,“en”,“capit”,“verizon”,“codelco”
2	“libor”,“cd”,“share”,“profit”,“month”,“sale”,“compani”,“year”,“revenu”,“bn”
3	“rio”,“stock”,“group”,“xstrata”,“gilt”,“list”,“yield”,“price”,“mine”,“bhp”
4	“market”,“rate”,“bank”,“price”,“dollar”,“economi”,“inflat”,“growth”,“year”,“bond”
5	“iceland”,“group”,“suez”,“french”,“compani”,“carrefour”,“sale”,“brazil”,“share”, “bn”
6	“busi”,“compani”,“googl”,“work”,“peopl”,“facebook”,“school”,“skill”,“job”,“social”
7	“investig”,“case”,“fraud”,“compani”,“alleg”,“court”,“bank”,“ivco”,“vw”,“porsch”
8	“cf”,“airlin”,“aircraft”,“trail”,“carrier”,“airbu”,“jet”,“passeng”,“boe”,“air”
9	“car”,“carmak”,“gm”,“compani”,“sale”,“vehicl”,“year”,“bn”,“market”,“plant”
10	“fund”,“manag”,“hedg”,“equiti”,“invest”,“asset”,“investor”,“global”,“incom”, “market”
11	“properti”,“etf”,“fund”,“market”,“investor”,“bank”,“invest”,“uk”,“year”,“compani”
12	“appl”,“phone”,“shown”,“mobil”,“limit”,“hlc”,“yr”,“trade”,“free”,“content”
13	“parti”,“labour”,“minist”,“elect”,“tori”,“brown”,“govern”,“cameron”,“polit”, “prime”
14	“art”,“design”,“work”,“artist”,“galleri”,“london”,“museum”,“build”,“citi”,“hous”
15	“bbc”,“film”,“weather”,“itv”,“show”,“seri”,“live”,“hollyoak”,“channel”,“region”
16	“murdoch”,“broadband”,“farmer”,“food”,“agricultur”,“bt”,“crop”,“bskyb”, “compani”,“corp”
17	“pe”,“chile”,“denmark”,“rep”,“hungari”,“colombia”,“group”,“indonesia”, “malaysia”,“argentina”
18	“cadburi”,“kraft”,“drug”,“dubai”,“lm”,“compani”,“ship”,“ord”,“shipp”,“gsk”
19	“carbon”,“emiss”,“ser”,“energi”,“climat”,“prog”,“fund”,“rbsg”,“environment”, “invest”
20	“aig”,“islam”,“pru”,“bn”,“compani”,“aia”,“bank”,“insur”,“busi”,“execut”

(next page)

Table A.1: *Continued:* LDA Topics from Chapter 5.1

Topic	Top 10 words
21	“nh”,“health”,“patient”,“hospit”,“care”,“healthcar”,“servic”,“privat”,“drug”, “compani”
22	“china”,“school”,“chines”,“busi”,“peopl”,“music”,“year”,“beij”,“work”,“univers”
23	“stock”,“call”,“request”,“fund”,“mail”,“minut”,“charg”,“price”,“thaksin”,“servic”
24	“mubarak”,“egypt”,“elect”,“egyptian”,“brotherhood”,“presid”,“protest”,“ahmadi”, “nejad”,“polit”
25	“equip”,“servic”,“leisur”,“ga”,“industri”,“telecommun”,“good”,“oil”,“materi”, “food”
26	“abn”,“emi”,“amro”,“terra”,“firma”,“bank”,“bn”,“forti”,“group”,“compani”
27	“und”,“fd”,“ssga”,“bd”,“om”,“ho”,“bs”,“class”,“govt”,“editor”
28	“index”,“fell”,“stock”,“cl”,“bank”,“rose”,“share”,“market”,“gain”,“data”
29	“properti”,“fd”,“brand”,“hotel”,“luxuri”,“hous”,“watch”,“yacht”,“sundai”, “residenti”
30	“peso”,“fund”,“equiti”,“dinar”,“privat”,“invest”,“bank”,“egypt”,“bn”,“compani”
31	“oil”,“iran”,“ga”,“bp”,“iraq”,“nuclear”,“militari”,“countri”,“govern”,“energi”
32	“price”,“dec”,“yield”,“south”,“turkei”,“pe”,“nav”,“sep”,“poland”,“venezuela”
33	“korea”,“korean”,“clear”,“lg”,“otc”,“deriv”,“south”,“trade”,“seoul”,“kim”
34	“pension”,“tax”,“scheme”,“annuiti”,“incom”,“retir”,“list”,“pai”,“benefit”,“rate”
35	“coal”,“ivco”,“aim”,“compani”,“share”,“mine”,“group”,“price”,“enrc”,“china”
36	“eu”,“european”,“eurozon”,“govern”,“bank”,“countri”,“greec”,“union”,“minist”, “debt”
37	“wine”,“russia”,“china”,“russian”,“putin”,“kairo”,“chines”,“moscow”,“restaur”, “georgia”
38	“melchior”,“opp”,“tesco”,“calculat”,“share”,“class”,“date”,“uk”,“shower”,“store”
39	“rate”,“convent”,“ng”,“market”,“appli”,“bond”,“currenc”,“il”,“meril”,“par”
40	“sun”,“fair”,“cloudi”,“shower”,“rain”,“xr”,“priceslast”,“shown”,“thunder”, “microsoft”

(next page)

Table A.1: *Continued:* LDA Topics from Chapter 5.1

Topic	Top 10 words
41	“palestinian”,“israel”,“isra”,“gaza”,“hama”,“flu”,“netanyahu”,“peac”,“dress”, “minist”
42	“compani”,“govern”,“account”,“school”,“busi”,“manag”,“rail”,“audit”,“fund”, “regul”
43	“jpm”,“siemen”,“vodafon”,“compani”,“sale”,“bn”,“group”,“year”,“deut”,“eq”
44	“gam”,“polic”,“pakistan”,“sky”,“kill”,“attack”,“sport”,“bbb”,“war”,“footbal”
45	“bank”,“fund”,“fin”,“market”,“manag”,“invest”,“investor”,“bn”,“int”,“compani”
46	“bank”,“mortgag”,“loan”,“bn”,“credit”,“capit”,“fund”,“market”,“debt”,“asset”
47	“ftse”,“cap”,“republican”,“obama”,“msci”,“global”,“romnei”,“democrat”,“dj”, “world”
48	“work”,“plai”,“book”,“music”,“life”,“peopl”,“love”,“live”,“film”,“make”
49	“cp”,“sempra”,“roch”,“prologi”,“rockwel”,“safewai”,“sherwil”,“rockwlcol”, “questdg”,“repsrv”
50	“quot”,“euriborlibor”,“libor”,“basi”,“annual”,“month”,“rate”,“icap”,“euroswiss”, “semi”

Bibliography

[1] Tse, C.K., Liu, J. and Lau, F.C.M., A network perspective of the stock market. *Journal of Empirical Finance*, 2010, **17**, 659–667.

[2] Tumminello, M., Lillo, F. and Mantegna, R.N., Hierarchically nested factor model from multivariate data. *Europhysics Letters*, 2007, **78**, 30006.

[3] Allez, R. and Bouchaud, J.P., Individual and collective stock dynamics: intra-day seasonalities. *New Journal of Physics*, 2011, **13**, 025010.

[4] Aste, T., Shaw, W. and Di Matteo, T., Correlation structure and dynamics in volatile markets. *New Journal of Physics*, 2010, **12**, 085009.

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

[6] Billio, M., Getmansky, M., Lo, A. and Pelizzon, L., Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 2012, **104**, 535–559.

[7] Biroli, G., Bouchaud, J.P. and Potters, M., The Student ensemble of correlation matrices: eigenvalue spectrum and Kullback-Leibler entropy. *arXiv preprint arXiv:0710.0802*, 2007.

[8] Bonanno, G., Caldarelli, G., Lillo, F. and Mantegna, R.N., Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E*, 2003, **68**, 046130.

[9] Bonanno, G., Lillo, F. and Mantegna, R.N., High-frequency cross-correlation in a set of stocks. *Quantitative Finance*, 2001, **1**, 96–104.

[10] Borghesi, C., Marsili, M. and Miccichè, S., Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode. *Physical Review E*, 2007, **76**, 026104.

[11] Campbell, R., Forbes, C., Koedijk, K. and Kofman, P., Increasing correlations or just fat tails?. *Journal of Empirical Finance*, 2008, **15**, 287–309.

[12] Carbone, A., Detrending Moving Average algorithm: a brief review. In *Proceedings of the Science and Technology for Humanity (TIC-STH), 2009 IEEE Toronto International Conference*, pp. 691–696, 2009.

[13] Cecchetti, S. and Kharroubi, E., Reassessing the impact of finance on growth. *BIS working paper*, 2012, Available at SSRN: http://ssrn.com/abstract=2117753.

[14] Cizeau, P., Potters, M. and Bouchaud, J., Correlation structure of extreme stock returns. *Quantitative Finance*, 2001, **1**, 217–222.

[15] De Jong, F., Nijman, T. and Röell, A., Price effects of trading and components of the bid-ask spread on the Paris Bourse. *Journal of Empirical Finance*, 1996, **3**, 193–213.

[16] Efron, B. and Tibshirani, R., *An introduction to the bootstrap*, Vol. 57, , 1993, CRC press.

[17] Epps, T., Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 1979, pp. 291–298.

[18] Forbes, K. and Rigobon, R., No contagion, only interdependence: measuring stock market comovements. *The Journal of Finance*, 2002, **57**, 2223–2261.

[19] Gopikrishnan, P., Plerou, V., Liu, Y., Amaral, L., Gabaix, X. and Stanley, H., Scaling and correlation in financial time series. *Physica A: Statistical Mechanics and its Applications*, 2000, **287**, 362–373.

[20] Gopikrishnan, P., Rosenow, B., Plerou, V. and Stanley, H., Quantifying and interpreting collective behavior in financial markets. *Physical Review E*, 2001, **64**, 035106.

[21] Hall, R.E., Why does the economy fall to pieces after a financial crisis?. *The Journal of Economic Perspectives*, 2010, **24**, 3–20.

[22] Havlin, S., Kenett, D.Y., Ben-Jacob, E., Bunde, A., Cohen, R., Hermann, H., Kantelhardt, J., Kertész, J., Kirkpatrick, S., Kurths, J. *et al.*, Challenges in network science: Applications to infrastructures, climate, social systems and economics. *European Physical Journal-Special Topics*, 2012, **214**, 273.

[23] Hayashi, T. and Yoshida, N., On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 2005, **11**, 359–379.

[24] Huth, N. and Abergel, F., High Frequency Lead/lag Relationships-Empirical facts. *arXiv preprint arXiv:1111.7103*, 2011.

[25] Kenett, D.Y., Preis, T., Gur-Gershgoren, G. and Ben-Jacob, E., Quantifying Meta-Correlations in Financial Markets. *Europhysics Letters*, 2012a, **99**, 38001.

[26] Kenett, D.Y., Raddant, M., Lux, T. and Ben-Jacob, E., Evolvement of uniformity and volatility in the stressed global financial village. *PloS one*, 2012b, **7**, e31144.

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

[28] Kenney, J.F. and Keeping, E.S., *Mathematics of Statistics, part 2*, 2nd Edition , 1962, D. Van Nostrand Company Inc.

[29] Laloux, L., Cizeau, P., Potters, M. and Bouchaud, J., Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 2000, **3**, 391–398.

[30] Lo, A.W. and MacKinlay, A.C., Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of financial studies*, 1988, **1**, 41–66.

[31] Malkiel, B.G., The efficient market hypothesis and its critics. *The Journal of Economic Perspectives*, 2003, **17**, 59–82.

[32] Malkiel, B.G. and Fama, E.F., Efficient Capital Markets: A Review Of Theory And Empirical Work*. *The journal of Finance*, 1970, **25**, 383–417.

[33] Mantegna, R.N., Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 1999, **11**, 193–197.

[34] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U., Network motifs: simple building blocks of complex networks. *Science*, 2002, **298**, 824–827.

[35] Munnix, M., Schafer, R. and Guhr, T., Impact of the tick-size on financial returns and correlations. *Physica A: Statistical Mechanics and its Applications*, 2010, **389**, 4828–4843.

[36] Onnela, J., Chakraborti, A., Kaski, K. and Kertesz, J., Dynamic asset trees and Black Monday. *Physica A: Statistical Mechanics and its Applications*, 2003, **324**, 247–252.

[37] Podobnik, B. and Stanley, H.E., Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Nonstationary Time Series. *Physical review letters*, 2008, **100**.

[38] Pollet, J. and Wilson, M., Average correlation and stock market returns. *Journal of Financial Economics*, 2010, **96**, 364–380.

[39] Shmilovici, A., Alon-Brimer, Y. and Hauser, S., Using a stochastic complexity measure to check the efficient market hypothesis. *Computational Economics*, 2003, **22**, 273–284.

[40] Song, D.M., Tumminello, M., Zhou, W.X. and Mantegna, R.N., Evolution of worldwide stock markets, correlation structure, and correlation-based graphs. *Physical Review E*, 2011, **84**, 026108.

[41] Tobin, J., A general equilibrium approach to monetary theory. *Journal of money, credit and banking*, 1969, **1**, 15–29.

[42] Toth, B. and Kertesz, J., The Epps effect revisited. *Quantitative Finance*, 2009, **9**, 793–802.

[43] Tumminello, M., Aste, T., Di Matteo, T. and Mantegna, R.N., A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**, 10421-10426.

[44] Tumminello, M., Coronnello, C., Lillo, F., Micciche, S. and Mantegna, R.N., Spanning trees and bootstrap reliability estimation in correlation based networks. *Int. J. Bifurcat. Chaos*, 2007a, **17**, 2319–2329.

[45] Tumminello, M., Di Matteo, T., Aste, T. and Mantegna, R.N., Correlation based networks of equity returns sampled at different time horizons. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2007b, **55**, 209–217.

[46] Tumminello, M., Lillo, F. and Mantegna, R.N., Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior & Organization*, 2010, **75**, 40–58.

[47] Tumminello, M., Miccichè, S., Lillo, F., Piilo, J. and Mantegna, R.N., Statistically validated networks in bipartite complex systems. *PloS one*, 2011, **6**, e17994.

[48] Tumminello, M., Curme, C., Mantegna, R.N., Stanley, H.E. and Kenett, D.Y., How lead-lag correlations affect the intra-day pattern of collective stock dynamics. *Manuscript in preparation.*

[49] Markowitz, H. Portfolio selection. *The Journal of Finance*, 7:77-91, 1952.

[50] L Laloux, P Cizeau, JP Bouchaud, and M Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7):1467–1470, AUG 16 1999.

[51] V Plerou, P Gopikrishnan, B Rosenow, LAN Amaral, and HE Stanley. Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters*, 83(7):1471–1474, AUG 16 1999.

[52] RN Mantegna. Hierarchical structure in financial markets. *European Physical Journal B*, 11(1):193–197, SEP 1999.

[53] G Bonanno, G Caldarelli, F Lillo, and RN Mantegna. Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E*, 68(4, 2), OCT 2003.

[54] JP Onnela, A Chakraborti, K Kaski, J Kertesz, and A Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 68(5, 2), NOV 2003.

[55] Y Hamao, RW Masulis, and V Ng. Correlations in price changes and volatility across international stock markets. *Review of Financial Studies* 3(2):281–307, 1990.

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

[57] MC Lundin, MM Dacorogna, and UA Müller. Correlation of high frequency financial time series Available at SSRN 79848, 1998.

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

[59] LH Ederington and JH Lee. How markets process information: News releases and volatility. *The Journal of Finance* 48(4):1161–1191, 1993.

[60] TG Andersen and T Bollerslev. Intraday periodicity and volatility persistence in financial markets *Journal of Empirical Finance* 4: 115–158.

[61] Bence Toth and Janos Kertesz. Accurate estimator of correlations between asynchronous signals. *Physica A– Statistical Mechanics and its Applications*, 388(8):1696–1705, APR 15 2009.

[62] R. Allez and J.P. Bouchaud. "Individual and collective stock dynamics: intra-day seasonalities". *New Journal of Physics* **13**: 025010, 2011.

[63] Chester Curme, Michele Tumminello, Rosario N Mantegna, H Eugene Stanley, and Dror Y Kenett. Emergence of statistically validated financial intraday lead-lag relationships. Quantitative Finance, 2014.

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

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

[66] A. Carbone and G. Castelli. Scaling properties of long-range correlated noisy signals: application to financial markets. In *Proc. of SPIE Vol*, volume 5114, page 407, 2003.

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

[68] Anderson, T.G., Bollerslev, T., Diebold, F.X. and Vega, C., 2003. "Micro effects on macro announcements: real-time price discovery in foreign exchange" *American Economic Review* **93**: 36–82.

[69] Anderson, T.G., Bollerslev, T., Diebold, F.X. and Vega, C., 2007. "Real-time price discovery in global stock, bond, and foreign exchange markets." *Journal of International Economics* **73** (2): 251–277.

[70] Balduzzi, P., Elton, E.J. and Green, T.C., 2001. “Economic news and bond prices: evidence from the U.S. treasury market”, *Journal of Financial and Quantitative Analysis* **36** (4): 523–543.

[71] Ball, R., and Brown, P., 1968. “An empirical evaluation of accounting income numbers. *Journal of Accounting Research* **6** (2):159–178.

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

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

[74] Bradley, A., 1997. “The use of the area under the ROC curve in the evaluation of machine learning algorithms.” *Pattern Recognition* **30** (7): 1145–1159.

[75] Chordia, T., Goyal, A., Sadka, G., Sadka, R. and Shivakumar, L., 2009. “Liquidity and the post-earnings-announcement drift.” *Financial Analysts Journal* **65** (4): 18–32.

[76] Chung, F. *Spectral Graph Theory.* USA: American Mathematical Society, 1997.

[77] Curme C., Tumminello, M., Mantegna, R.N., Stanley, H.E. and Kenett, D.Y., 2014. “Emergence of statistically validated financial intraday lead-lag relationships”, recently accepted at *Quantitative Finance.* http://arxiv.org/pdf/1401.0462.pdf

[78] Diebold, F.X. and Yilmaz, K., 2009. “Measuring financial asset return and volatility spillovers, with application to global equity markets.” *The Economic Journal* **119** (534): 158–171.

[79] Drineas, P., Frieze, A., Kannan, R., Vempala, S. and Vinay, V., 2004. “Clustering large graphs via the Singular Value Decomposition.” *Machine Learning* **56**: 9-33.

[80] Freeman, L., 1977. “A set of measures of centrality based on betweenness.” *Sociometry* **40**: 35-41.

[81] Ghosh, A., Saidi, R. and Johnson, K.H., 1999. "Who moves the Asia-Pacific stock markets–US or Japan? Empirical evidence based on the theory of cointegration." *The Financial Review* **34** (1): 159–169.

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

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

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

[85] Holten, D. and van Wijk, J. J., 2009. "Force-directed edge bundling for graph visualisation." *Eurographics/ IEEE-VGTC Symposium on Visualization* **28** (3).

[86] Kruskal, J.B., 1956. "On the shortest spanning subtree of a graph and the traveling salesman problem." *Proceedings of the American Mathematical Society* **7**: 48-50.

[87] Mantegna, R.N. and Stanley, H.E., 2000. *An Introduction to Econophysics: Correlations and Complexity in FInance*. Cambridge: Cambridge University Press.

[88] "Thomson Reuters MarketPsych Indices (TRMI)." MarketPsych, 2013. 25 March 2014. <https://www.marketpsych.com/data/>.

[89] Morrison, J.L., Breitling, R., Higham, D.J. and Gilbert, D.R., 2006. "A lock-and-key model for protein-protein interactions." *Bioinformatics* **2**: 2012-2019.

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

[91] Sandoval, L., 2014. “To lag or not to lag? How to compare indices of stock markets that operate at different times.” *Physica A* **403**: 227-243.

[92] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski B. and Ideker, T., 2003. “Cytoscape: a software environment for integrated models of biomolecular interaction networks.” *Genome Research* **13** (11): 2498–504.

[93] Shumway, R.H. and Stoffer, D.S., 2011. *Time Series Analysis and its Applications with R Examples.* New York: Springer.

[94] Taylor, A., Vass, J.K. and Higham, D.J., 2011. “Discovering bipartite substructure in directed networks.” *J. Comput. Math.* **14**: 72-86.

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

[96] Vandewalle, N., Ph. Boveroux, and F. Brisbois. “Domino effect for world market fluctuations.” *The European Physics Journal B*, **15** (2000), 547–549.

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

[98] R. Cont, J.P. Bouchaud, Herd behavior and aggregate fluctuations in financial markets. *Macroecon. Dyn.* **4**, 170–196 (2000).

[99] “Thomson Reuters MarketPsych Indices (TRMI). *MarketPsych.* https://www.marketpsych.com/data/

[100] R. Feldman, B. Rosenfield, R. Bar-Haim, M. Fresko, The Stock Sonar– Sentiment analysis of stocks based on a hybrid approach. *Innovative Applications of Artificial Intelligence*, 1642–1647 (2011).

[101] M. Alanyali, H.S. Moat, T. Preis, Quantifying the relationship between financial news and the stock market. *Scientific Reports* **3**, 3578 (2013).

[102] J. Bollen, H. Mao, X.J. Zeng, Twitter mood predicts the stock market. *Journal of Computational Science* **2**(1), 1–8 (2011).

[103] N. Oliveira, P. Cortez, N. Areal, in *Progress in Artificial Intelligence*, L.M. Correia, L. Reis, J.M. Cascalho, Eds. (Springer, Berlin Heidelberg, 2013) pp. 355–365.

[104] T. Preis, H.S. Moat, H.E. Stanley, Quantifying trading behavior in financial markets using *Google Trends*. *Scientific Reports* **3**, 1684 (2013).

[105] C. Curme, T. Preis, H.E. Stanley, H.S. Moat, Quantifying the semantics of search behavior before stock market moves. *Proc. Natl. Acad. Sci. USA* **111**(32), 11600–11605 (2014).

[106] T. Hofmann, Probabilistic latent semantic analysis. Paper presented at the 15th Conference on Uncertainty in Artificial Intelligence, pp. 289–296, Stockholm, Sweden, 1999. http://www.csail.mit.edu/ jrennie/trg/papers/hofmann-uai99.ps.gz

[107] D.M. Blei, Probabilistic topic models. *Communications of the ACM* **55**(4), 77–84 (2012).

[108] R. Hisano, D. Sornette, T. Mizuno, T. Ohnishi, T. Watanabe, High quality topic extraction from business news explains abnormal financial market volatility. *PLoS ONE* **8**(6), e64846 (2013).

[109] M.F. Porter, An algorithm for suffix stripping. *Program* **14**(3), 130–137 (1980).

[110] Full-Text Stopwords. MySQL Reference Manual. Retrieved from `http://dev.mysql.com/doc/refman/5.6/en/fulltext-stopwords.html`.

[111] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora. Paper presented at the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50, Valletta, Malta, 2010, `http://is.muni.cz/publication/884893/en`.

[112] A.T. Wilson, P.A. Chew, Term weighting schemes for Latent Dirichlet Allocation. Paper presented at Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pp. 465–473, Los Angeles, CA, 2010, `http://www.aclweb.org/anthology/N10-1070`.

[113] T.J. Hazen, Direct and latent modeling techniques for computing spoken document similarity. Paper presented at the IEEE Workshop on Spoken Language Technology (SLT), pp. 366–371, Berkeley, CA, 2010, `http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5700880`.

[114] R. Xu, L. Ye, J. Xu, Reader's emotion prediction based on weighted latent dirichlet allocation and multi-label k-nearest neighbor model. *Journal of Computational Information Systems* **9**(6), 2209–2216 (2013).

[115] G. Salton, M. McGill, *Introduction to Modern Information Retrieval* (McGraw Hill, New York, 1983).

[116] C.E. Shannon, A mathematical theory of communication. *Bell System Technical Journal* **27** (3), 379–423 (1948).

[117] H. Misra, O. Cappé, F. Yvon, Using LDA to detect semantically incoherent documents. *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 41–48 (2008).

[118] K.S. Chan, J.D. Cryer, *Time Series Analysis with Applications in R* (Springer, New York ed. 2, 2010).

[119] R.J. Hyndman, Y. Khandakar, Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software* **27**(3) (2008).

[120] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**, 253–263 (1995).

[121] R.A. Fisher, Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika* **10**(4), 507–521 (1915).

[122] PricewaterhouseCoopers, LLP "Average Daily Global Audience (ADGA)" audited figures, 2010.

[123] Shleifer A (2000) *Inefficient Markets: An Introduction to Behavioral Finance* (Oxford University Press, Oxford).

[124] Lillo F, Farmer JD, Mantegna RN (2003) Econophysics: Master curve for price-impact function. *Nature* **421**, 129–130 (2003).

[125] Gabaix X (2009) Power Laws in Economics and Finance. *Annu. Rev. Econ.* **1**, 255–93 (2009).

[126] Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2003) A theory of power-law distributions in financial market fluctuations. *Nature* **423**, 267–270 (2003).

[127] Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2006) Institutional Investors and Stock Market Volatility. *Quarterly Journal of Economics* **121**, 461-504 (2006).

[128] Preis T, Schneider JJ, Stanley HE (2011) Switching processes in financial markets. *Proc Natl Acad Sci USA* **108**, 7674–7678.

[129] Takayasu H, editor (2006) *Practical Fruits of Econophysics* (Springer, Berlin).

[130] Coval J, Jurek JW, Stafford E (2009) Economic catastrophe bonds. *American Economic Review* **99**, 628–666.

[131] Bouchaud JP, Matacz A, Potters M (2001) Leverage effect in financial markets: The retarded volatility model. *Phys Rev Lett* **87**, 228701.

[132] Hommes CH (2002) Modeling the stylized facts in finance through simple nonlinear adaptive systems. *Proc Natl Acad Sci USA* **99**, 7221–7228.

[133] Haldane AG, May RM (2011) Systemic risk in banking ecosystems. *Nature* **469**, 351–355 (2011).

[134] Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* **397**, 498–500.

[135] Krugman P (1996) *The Self-Organizing Economy* (Blackwell, Cambridge MA).

[136] Sornette D, von der Becke S (2011) Complexity clouds finance-risk models. *Nature* **471**, 166.

[137] Garlaschelli D, Caldarelli G, Pietronero L (2003) Universal scaling relations in food webs. *Nature* **423**, 165–168.

[138] Onnela J-P, Arbesman S, Gonzalez MC, Barabasi A-L, Christakis NA (2011) Geographic constraints on social network groups. *PLoS One* **6**, e16939.

[139] Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, Vespignani A, White DR (2009) Economic networks: The new challenges. *Science* **325**, 422–425.

[140] Gabaix X (2011) The granular origins of aggregate fluctuations, *Econometrica* **79**, 733-772

[141] Gabaix X, Krishnamurthy A, Vigneron O (2007) Limits of Arbitrage: Theory and Evidence from the Mortgage-Backed Securities Market, *Journal of Finance* **62**, 557-595

[142] Lux T (1999) The socio-economic dynamics of speculative markets: interacting agents, chaos, and the fat tails of return distributions, *Journal of Economic Behavior & Organization* **33**, 143-165

[143] Simon HA (1955) A behavioral model of rational choice. *Quarterly Journal of Economics* **69**, 99–118.

[144] King G (2011) Ensuring the data-rich future of the social sciences. *Science* **331**, 719–721.

[145] Vespignani A (2009) Predicting the behavior of techno-social systems. *Science* **325**, 425–428.

[146] Lazer D, Pentland A, Adamic L, Aral S, Barabasi A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Computational social science. *Science* **323**, 721–723.

[147] Conte R, et al. (2012) Manifesto of computational social science. *Eur. Phy. J. Spec. Top.* **214**, 325–346.

[148] Moat HS, Preis T, Olivola CY, Liu C, Chater N (2014) Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences* **37**, 92–93.

[149] Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with Web search. *Proc Natl Acad Sci USA* **107**, 17486–17490.

[150] Askitas N, Zimmermann KF (2009) Google econometrics and unemployment forecasting. *Applied Economics Quarterly* **55**, 107–120.

[151] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014.

[152] Preis T, Reith D, Stanley HE (2010) Complex dynamics of our economic life on different scales: Insights from search engine query data. *Phil Trans R Soc A* **368**, 5707–5719.

[153] Bordino I, Battiston S, Caldarelli G, Cristelli M, Ukkonen A, Weber I (2012) Web search queries can predict stock market volumes. *PLoS One* **7**, e40014.

[154] Preis T, Moat HS, Stanley HE, Bishop SR (2012) Quantifying the advantage of looking forward. *Scientific Reports* **2**, 350.

[155] Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports* **3**, 1684.

[156] Da Z, Engelberg J, Gao P (2011) In search of attention. *Journal of Finance* **66**, 1461–1499.

[157] Bank M, Larch M, Peter G (2011) Google search volume and its influence on liquidity and returns of German stocks. *Financial Markets and Portfolio Management* **25**, 239–264.

[158] Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T (2013) Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports* **3**, 1801.

[159] Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022.

[160] Politis D, Romano J (1992) in *Exploring the Limits of Bootstrap*, eds LePage R, Billard L (John Wiley, New York), pp 263–270.

[161] Hansen PR (2005) A test for superior predictive ability. *Journal of Business & Economic Statistics* **23**, 365–380.

[162] Lux T (2011) Sentiment Dynamics and Stock Returns: The Case of the German Stock Market. *Empirical Economics* **41**, 663–679.

[163] Tversky A, Kahneman D (1991) Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics* **106**, 1039–1061.

[164] Silverman BW (1986) Density Estimation (Chapman and Hall, London).

Curriculum Vitae

Chester Curme

Boston University, Physics Department Telephone: 781-760-3645
590 Commonwealth Avenue
Boston, Massachusetts 02215 USA E-mail: chester.curme@gmail.com

EDUCATION

- 2015, Ph.D. Physics, Boston University, Boston, MA, USA

  Advisor: H. Eugene Stanley

  Thesis: *Lagged correlation networks*

  GPA: 4.00
- 2011, B.A. Physics and Mathematics, *summa cum laude*

  Middlebury College, Middlebury, VT, USA

  GPA: 3.96

  GRE: 168 Verbal (98%), 167 Quantitative (95%), 5.5 Writing (96%)

AWARDS

- Junior Phi Beta Kappa, 2012 (top 2% of graduating class)
- Middlebury College Physics Department Highest Honors
- NSF Graduate Research Fellowship Program Honorable Mention in 2012 and 2014

ACADEMIC APPOINTMENTS

- Graduate Research Assistant, Center for Polymer Studies, Boston University (Summer 2012 – 2015)
- Research Fellow in Data Science, Warwick Business School, Coventry, UK (Summer 2014)

PROGRAMMING

Python R Fortran Mathematica Matlab C++ Java SQL Excel LaTeX Bash

RESEARCH INTERESTS

complex networks, multivariate statistics, statistical inference, interdisciplinary applications of physics

PUBLICATIONS

1. **C. Curme**, I. Vodenska, and H.E. Stanley (submitted). *Coupled network approach to uncovering predictive relationships among financial market returns and news sentiments.*
2. **C. Curme**, M. Tumminello, R.N. Mantegna, H.E. Stanley and D.Y. Kenett (in preparation). *How lead-lag correlations affect the intra-day pattern of collective stock dynamics.*
3. **C. Curme**, Y.D. Zhuo, H.S. Moat, and T. Preis (submitted). *The role of diversity in financial news.*
4. **C. Curme**, T. Preis, H.E. Stanley and H.S. Moat (2014). *Quantifying the semantics of search behavior before stock market moves*, PNAS **111** (32): 11600–11605.

   Coverage in: *The Wall Street Journal*, *Scientific American*, and *The Times*, among others.

5. **C. Curme**, M. Tumminello, R.N. Mantegna, H.E. Stanley and D.Y. Kenett (2014). *Emergence of statistically validated financial intraday lead-lag relationships*, Quantitative Finance. http://arxiv.org/pdf/1401.0462.pdf

6. B. Podobnik, A. Majdandzic, **C. Curme**, Z. Qiao, W.X. Zhou, H.E. Stanley, and B. Li (2014). *Network risk and forecasting power in phase-flipping dynamical networks.* Physical Review E **89**, 042807.

7. H.S. Moat, **C. Curme**, H.E. Stanley and T. Preis (2014). "Anticipating stock market movements with Google and Wikipedia". Book chapter in *Nonlinear Phenomena in Complex Systems: From Nano to Macro Scale*, eds D. Matrasulov, H.E. Stanley.

8. V. Dalko, L.R. Klein, **C. Curme**, D.Y. Kenett, H.E. Stanley, and M.H. Wang (2014). *Financial market dominance - an agent-based model.* Sixth Annual Meeting of the Academy of Behavioral Finance and Economics.

9. H.S. Moat, **C. Curme**, A. Avakian, D.Y. Kenett, H.E. Stanley and T. Preis (2013). *Quantifying Wikipedia usage patterns before stock market moves.* Scientific Reports 3, 1801.
http://www.nature.com/srep/2013/130508/srep01801/pdf/srep01801.pdf

10. C.H. Comin, J.R. Santos, D. Corradini, W. Morrison, **C. Curme**, D.L. Rosene, A. Gabrielli, L.da F. Costa and H.E. Stanley (2014). *Statistical physics approach to quantifying differences in myelinated nerve fibers.* Scientific Reports 4, 4511.
http://www.nature.com/srep/2014/140328/srep04511/pdf/srep04511.pdf

11. H.D. Winter, **C. Curme**, K.K. Reeves and P. Martens (2013). *Simulating emission of coronal loops of non-constant cross-section.* Proceedings of AAS/Solar Physics Division Meeting.

INVITED TALKS

- **Chester Curme**. *Quantifying the semantics of search behavior before stock market moves.* Computational Social Science Conference. June 2014, Coventry, United Kingdom.
- **Chester Curme**. *Statistically-validated networks of lagged correlations in financial markets.* FuturICT International School on Network Science. May 2014, Balatonfüred, Hungary.
- Adam Avakian, **Chester Curme**, Helen Susannah Moat, Tobias Preis and H. Eugene Stanley. *Can we anticipate economic behavior using open source indicators?* Open Source Indicators– Population Behavior Understanding for Large Scale Events. July 2013, HRL Laboratories, Malibu, CA.

REVIEWING ACTIVITIES

- PLOS ONE
- Physica A

UNDERGRADUATE RESEARCH

- Cornell Center for Nanoscale Systems, Ithaca, NY (Summer 2010)

  NSF-Sponsored REU Student

  - Refined a novel lithographic technique known as Magnetophoretic Lithography.
  - Built a numerical simulation of the process to explore its resolution limits.

- Harvard-Smithsonian Center for Astrophysics, Cambridge, MA (Summer 2009)

  NSF-Sponsored REU Student

  - Improved numerical simulation software to better model solar coronal loop structures.
  - Pursued work to poster presentation in 2009 American Geophysical Union Fall Meeting and one published article.

LEADERSHIP

- Solar Decathlon, Middlebury, VT (2009 – 2011)

  Student Engineering Lead

  - Led the systems engineering group of Middlebury College's Solar Decathlon team. Middlebury was one of 19 finalists that competed in the biennial DOE contest to build a solar-powered home.
  - Designed and built the house's heating, ventilation, water and electrical systems with the aid of student teammates and local professionals.
  - Middlebury placed fourth overall in the 2011 competition, winning first place in three sub-contests.

- Physics Department Student Advisory Council, Middlebury, VT (January 2010)

  Junior Representative

  - Elected to a three-person committee by peers within physics department to aid in selection of new tenure-track professor.
  - Responsible for interviewing candidates and acting as liaison between students and faculty.

ADDITIONAL

- Boston University Physics tutor
- Middlebury College Physics Teaching Assistant
- Violinist in the Middlebury College Orchestra
- Weston High School Student Co-President (2006-2007)