The Impact of Social Mood on Stock Markets

Thársis Tuani Pinto Souza

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

July 28, 2019

I, Thársis Tuani Pinto Souza, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Assuming social media as a proxy for human activity, behavior and opinion, we aim to test the extent to which financial dynamics can be explained by collective opinion extracted from social media.

First, we present an analysis of Twitter sentiment extracted from U.S.-listed retail brands. We investigate whether there is a significan causal link between Twitter sentiment, and stock returns and volatility. The results suggest that social media is indeed a valuable source in the analysis of financial dynamics, sometimes carrying more prior information than mainstream news such as the Wall Street Journal and Dow Jones Newswires.

Second, we provide empirical evidence that suggests social media and stock markets have a nonlinear causal relationship. By using information-theoretic measures to cope with possible nonlinear causal effects, we point out large differences in the results with respect to linear coupling. Our findings suggest that the significant causal relationship between social media and stock returns is purely nonlinear in most cases. Furthermore, social media dominates directional coupling with the stock market, an effect that is not observable within linear modeling.

Finally, we propose a model that predicts future correlation structure, based on a mechanism of link formation by triadic closure, that combines information from social media and financial data in a multiplex structure. The results demonstrate that the proposed model can achieve up to 40% out-of-sample performance improvement, compared to a benchmark model that assumes that correlation structure is time invariant. Social media information leads to improved models for all settings tested, particularly in the long-term prediction of a financial market structure.

Our findings indicate that social media sentiment dominates directional coupling with the stock market in the prediction of individual asset dynamics as well as the overall market structure.

Impact Statement

Online social networks offer a new way to investigate financial markets' dynamics by enabling the large-scale analysis of investors' collective behavior. This work studies the nature of the relationship between collective opinion extracted from social media and the collective behavior of stock market prices. We expect that the results of this thesis will have an impact in various areas both inside and outside academia. Our work can influence future scholarship and research methods as it serves as empirical guidance on model adequacy, market efficiency, and predictability, in the investigation of causal relationships between social and financial systems. To achieve this impact, our work has been published in a handbook, conference proceedings and it has been published or submitted to peer-reviewed journals, as well as made available at arXiv.org promoting open access research.

We believe that our work will influence how financial market participants are using social media data to inform their decisions. As evidence of interest on our work, the Twitter Blog [1] cited our research as evidence that social media conversations can reliably anticipate market movement in an article that discusses how financial analysts, traders and market professionals are using Twitter to stay abreast of the market and make critical decisions.

Acknowledgements

This work was supported by OptiRisk Systems and PsychSignal. I acknowledge financial support from CNPq - The Brazilian National Council for Scientific and Technological Development.

I am deeply indebted to Professor Tomaso Aste and Dr. Soong Moon Kang, my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to extend my appreciation to my previous academic mentors Prof. Carlile Lavor and Prof. Walter Mascarenhas.

I have had the support and encouragement of brilliant friends Andre Valloto, Olya Kolchyna, Adriano Koshiyama, Rodrigo Mazorra, Humberto Brandao, Cristiano Arbex, Cicero Zandona and Thiago Winkler, whom I thank for providing such a rich source of advice, wisdom and friendship.

I would also like to extend my gratitude to Sandro Manteiga, Clay Susini, Prof. Gautam Mitra and Ruggero Gramatica who taught me invaluable academic and professional lessons.

I wish to express my deepest gratitude to my brother Heli Samuel and my parents Maria de Lourdes Pinto de Lacerda Souza and Isaias de Souza Neto for their love, encouragement, wisdom and for teaching me the value of knowledge, integrity and hard work. Finally, I am deeply grateful to my wife Yoko Furusho for putting up with my idiosyncrasies and for being supportive, intellectually stimulating, constructive and for providing warm encouragement throughout this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background and Context

Investors' decisions are modulated not only by companies' fundamentals but also by personal beliefs, peer influence and informational impact from mainstream news and the Internet [2, 3]. Investors rational and irrational behavior and their relation to the market efficiency hypothesis [4] have been largely debated in economics and financial literature [5]. However, the availability of vast amounts of data from online systems have only recently paved the way for the large-scale investigation into investors collective behavior in financial markets [6, 7].

It is known that news plays a key role in financial markets [8]. Therefore, collecting and understanding the news announcements of private and public companies, macroeconomic news or even rumors have become essential to quickly adapt trading strategies and portfolio optimization in general. News provides information about an event and, as such, may be considered to be an event in itself [9]. The arrival of news influences a market's expectations of future price movements and it has a significant effect on investors sentiment and behavior [10]. There is upcoming and growing literature regarding the influence of news on financial markets, including the analysis of the behavior of price, volatility, liquidity and risk [2, 11, 12, 13, 14, 9, 15, 16, 17, 18].

In addition to mainstream news, the analysis of digital traces of collective human behavior has been recently used as a novel tool to quantify and statistically validate hypotheses about financial dynamics in an ex-ante fashion. Sentiment, emotions, and behavioral analytics can provide valuable information before the market moves [19, 13, 20]. Examples range from the use of Google Trends [21, 22] and Wikipedia [3, 23] to social media [6, 24, 7, 19, 25]. Social media, in particular, has become an increasingly important source to describe financial dynamics, as it provides a fine-grained, real-time information channel that includes not only major news stories but also information on granular events.

While recent literature provides evidence that exogenous information gathered from online social systems may be useful to describe financial dynamics, the literature still presents major gaps including the following:

- There are mixed results on the predictability of stock returns. On the one hand, some evidence is provided in favor of the predictability of price movements using news and social media [2, 11, 6, 26]. On the other hand, other studies report weak results [27, 19] suggesting that social media analytics has low predictive power when used alone.

- There is lack of evidence on the nature of the relationship between social opinion and price movements. The use of ad hoc functional forms and assumptions in different studies makes it difficult to draw general conclusions about the nature of the relationship between collective opinion and stock markets.

- Empirical studies thus far have been limited to the investigation of individual securities, often neglecting joint dependencies and the multi-asset case. There is little evidence on the value of social media data in the prediction of financial correlation structure.

1.2 Objectives

Assuming social media as a proxy for human activity, behavior and collective opinion, the main objective of this work is to test whether and to what extent financial dynamics can be better explained by collective opinion extracted from social media.

The following research questions are addressed:

- Does social media sentiment cause stock prices movements?
- What is the nature of the relationship between social media sentiment and the stock market?
- Can social media predict stock market structure?

In order to address the stated research questions, we test the following hypotheses:

- H1: Social media sentiment has a significant causal relationship with stocks returns and volatility;
- H2: Social media sentiment has a nonlinear impact in stock price returns;
- H3: Social media sentiment dominates directional coupling with the stock market; i.e., information provided by social media contributes to the description of stock market dynamics more than the opposite;
- H4: Social media sentiment structure predicts stock market structure.

1.3 Relevance and Contribution

Financial analysts, traders and market professionals globally are increasingly using Twitter to stay abreast of the market and make critical decisions [1]. From the launch of the first social media-based ETF in 2016 [28] to the release of several alpha-seeking social media data feeds including those launched by the Nasdaq Analytics Hub in 2017 [29] and the event-driven Twitter feed released by Bloomberg in 2018 [30], social media is poised to change how financial practitioners develop financial products and models to gain an edge in the market. Hence, this thesis advances academic research in the area and it also has timely and practical implications in the financial industry. We summarize the relevance and contribution of this thesis as follows:

- We provide evidence that social media sentiment has a significant causal relationship with price movements by analyzing constituents of the Dow Jones Industrial Average index. While analyzing retail brands, in particular, results suggest that social media can be a complementary source in the analysis of the financial dynamics to mainstream news such as the Wall Street Journal and Dow Jones Newswires.
- Taking social media as a proxy for investor's collective attention over the stock market, we provide the first empirical evidence that characterizes social media impact on market prices as nonlinear. This indicates that the impact of social media on stock returns may be higher than those currently reported in related studies. Our results serve as empirical guidance on model adequacy, market efficiency, and predictability, in the investigation of causal relationships between social and financial systems. Testing for nonlinear dependencies is of great importance in financial econometrics due to its implications in model adequacy, market efficiency, and predictability [31].
- Most of the related literature focuses on the investigation of the impact of social opinion on individual stocks often neglecting the multi-asset case. To the best of our knowledge, we provide the first empirical evidence that social opinion structure is relevant to the prediction of future financial correlation structures. This result has important consequences because of the fundamental importance of financial correlation structure in any study of portfolio risk, capital allocation or hedging in trading strategies as well as fundamental literature in Modern Portfolio Theory (MPT) [32], Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) [33].

1.4 Thesis Content

This thesis is organized as follows. In Chapter 2, we provide the methodology background. In Chapter 3, we present a literature review covering related work that leverages news and social media analytics in financial markets applications. In Chapter 4, we investigate whether there is a significant causal relationship between social media sentiment from retails brands and stock returns. The results reveal a dependency structure between sentiment and financial variables for both traditional newswires and social media. The findings suggest that social media sentiment plays an important role in explaining the price dynamics of the retail sector, even when compared to traditional newswires. In Chapter 5, we provide empirical evidence indicating that social media and stock markets have a nonlinear causal relationship. We observe that the significant causal relationship between social media and stock returns is purely nonlinear in most cases. Furthermore, a transfer entropy analysis reveals that more information is transferred from social media to the stock market than the other way around. In Chapter 6, we combine social and financial network information to predict a stock market correlation structure by assuming that financial links are formed through a mechanism of triadic closure, whereby triangles are formed by social and financial links. The results suggest that social media opinion structure can be used to better predict a future stock market structure, particularly in the long term. Finally, in Chapter 7, we conclude the work, describe limitations and suggest avenues for future research.

1.5 Publications and Manuscripts

1. O. Kolchyna and **T. T. P. Souza** and P. C. Trealaven and T. Aste. Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. in Handbook of Sentiment Analysis in Finance, Chapter 5. Mitra, G. and Yu, X. (Eds.) 2016. Also available as e-print: arxiv.org/abs/1507.00955.

2. **T. T. P. Souza** and O. Kolchyna and P. C. Trealaven and T. Aste. Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry. in Handbook of Sentiment Analysis in Finance, Chapter 23. Mitra, G. and Yu, X. (Eds.) 2016. Also available as e-print: arxiv.org/abs/1507.00784.

3. O. Kolchyna and **T. T. P. Souza** and P. C. Trealaven and T. Aste. A Framework for Twitter Events Detection, Differentiation and its Application for Retail Brands. Future Technologies Conference (FTC), 323-331, IEEE, 2016. Also available as e-print: arxiv.org/abs/1508.03981.

4. J. Manfield, D. Lukacsko and **T. T. P. Souza**. Bull Bear Balance: A Cluster Analysis of Socially Informed Financial Volatility. IEEE Proceedings of Computing Conference 2017. London, UK.

5. **T. T. P. Souza** and T. Aste. A nonlinear impact: evidence of causal effects of social media on market prices. *Submitted to PLOS ONE*. Also available as e-print: http://arxiv.org/abs/1601.04535.

6. L. S Rocha, F. S. A. Rocha and **T. T. P. Souza**. Is the public sector of your country a diffusion borrower? Empirical evidence from Brazil. PLOS ONE 12(10): e0185257. `https://doi.org/10.1371/journal.pone.0185257`.

7. **T. T. P. Souza** and T. Aste. Predicting future market structure by combining social and financial network information. *Submitted to Physica A*.

Chapter 2

Methodology

2.1 Data

2.1.1 News Analytics

We consider news meta-data supplied by Ravenpack [34], which consists in 10,949 news stories from Dow Jones Newswires, the Wall Street Journal and Barrons. Each news article receives scores for characteristics such as relevance, novelty and sentiment according to a related individual including the following:

- TIMESTAMP_UTC: The date/time at which the news item was received in Coordinated Universal Time (UTC).

- COMPANY: This field includes a company identifier in the format ISO_CODE/TICKER. The ISO_CODE is based on the company's original country of incorporation and TICKER on a local exchange ticker or symbol. If the company detected is a privately held company, there will be no ISO_CODE/TICKER information, COMPANY_ID.

- ISIN: An International Securities Identification Number (ISIN) to identify the company referenced in a story. The ISINs used are accurate at the time of story publication. Only one ISIN is used to identify a company, regardless of the number of securities traded for any particular company. The ISIN used will be the primary ISIN for the company at the time of the story.

- COMPANY_ID: A unique and permanent company identifier. Every company tracked is assigned an unique identifier comprised of six alphanumeric characters. The RP_COMPANY_ID field consistently identifies companies throughout the historical archive.

- RELEVANCE: A score between 0 and 100 that indicates how strongly related the company is to the underlying news story, with higher values indicating greater relevance. A score of 0 means the company was passively mentioned while a score of 100 means the company was predominant in the news story.

- CATEGORIES: An element or "tag" representing a company-specific news announcement or formal event. Relevant stories about companies are classified in a set of predefined event categories following a pre-defined taxonomy. When applicable, the role played by the company in the story is also detected and tagged.

- ESS - EVENT SENTIMENT SCORE: A granular score between 0 and 100 that represents the news sentiment for a given company by measuring various proxies sampled from the news. The score is determined by systematically matching stories typically categorized by financial experts as having short-term positive or negative share price impact. The strength of the score is derived from training sets where financial experts classified company-specific events and agreed these events generally convey positive or negative sentiment and to what degree. Their ratings are encapsulated in an algorithm that generates a score range between 0 and 100 where higher values indicate more positive sentiment while values below 50 show negative sentiment.

- ENS - EVENT NOVELTY SCORE: A score between 0 and 100 that represents how "new" or novel a news story is within a 24-hour time window. The first story reporting a categorized event about one or more companies is considered to be the most novel and receives a score of 100. Subsequent stories within the 24-hour time window about the same event for the same companies receive lower scores.

Table 2.1 shows a sample of the news-sentiment analytics data provided. The relevance score of a news article ranges between 0 and 100 and indicates how strongly related the company is to the underlying news story, whereby higher values indicate greater relevance. We filter the news stories with a relevance of 100. This increases the likelihood that the story is related to the underlying company. We also consider the event sentiment score (ESS). This measure indicates a short-term positive or negative financial or economic impact of the news in the underlying company; higher values indicate more positive impact.

Table 2.1: News sentiment analytics. Each row represents a news story related to a company. The meta-data considered consists of relevance and sentiment scores and a timestamp.

Story ID	Company	Date	Hour	Relevance	Event Sentiment Score
1	NIKE INC.	20140104	210130	33	64
2	MATTEL INC.	20140105	41357	100	50
3	NIKE INC.	20140105	145917	93	88
4	NIKE INC.	20140105	150523	100	61
5	GAMESTOP CORP.	20140105	193507	44	50
6	GAMESTOP CORP.	20140106	170040	99	44
7	MATTEL INC.	20140106	222532	100	61
8	GAMESTOP CORP.	20140107	32601	100	50
9	MATTEL INC.	20140107	172628	55	40
10	NIKE INC.	20140110	204027	100	67

2.1.2 Social Media Sentiment on Retail Brands

In Chapter 4, we conducted our analysis on a subset of listed retail brands with stocks traded in the U.S. stock market, which we monitored from November 01, 2013 to September 30, 2014. The examined stocks and their Reuters Instrument Codes (RIC) are as follows: ABERCROMBIE & FITCH CO. (ANF.N), NIKE INC. (NKE.N), HOME DEPOT INC. (HD.N), MATTEL INC. (MAT.N) and GAMESTOP CORP. (GME.N).

The choice of companies was given by data availability, which corresponds to the companies analyzed in [35] that were also publicly-listed. The data was extracted from the sentiment dataset that we created in [35], which was previously used to predict sales of retail brands [36].

For the Twitter sentiment data, we consider the number of positive, negative and neutral English messages which are related to the underlying company on a daily basis. We also take the total number of messages, regardless of their language, as a proxy for volume. Table 2.2 shows an example of the Twitter sentiment analytics for MATTEL INC. Table 2.3 offers a summary description of the selected companies and the number of stories considered. The methodology for the Twitter sentiment data is described in [35], where we proposed a new ensemble method that combines a lexicon and machine-learning approaches to best estimate Twitter sentiment analytics.

Table 2.2: Twitter Sentiment Analytics. Sample of meta-data information for MATTEL INC. On a daily basis, we considered the number of positive, negative and neutral English Twitter messages related to the company; we also considered the total number of messages (regardless of the language) as a proxy for volume.

Date	CompanyID	Volume	#Positive	#Negative	#Neutral
01/11/2013	MATTEL INC.	1,980	8	4	485
02/11/2013	MATTEL INC.	1,750	12	2	339
03/11/2013	MATTEL INC.	1,700	8	1	518
04/11/2013	MATTEL INC.	2,720	19	2	429
05/11/2013	MATTEL INC.	1,980	11	8	793
06/11/2013	MATTEL INC.	1,580	11	4	470
07/11/2013	MATTEL INC.	1,770	7	1	498
08/11/2013	MATTEL INC.	1,900	5	4	288
09/11/2013	MATTEL INC.	1,260	16	2	236
10/11/2013	MATTEL INC.	1,700	7	8	313

Table 2.3: Summary table of selected companies. Here we present the five retail brands selected for the analysis along with their market capitalization. We show the number of news and Tweets in the selected period.

Company	Market Cap.*	No. of News**	No. of Tweets
ABERCROMBIE & FITCH CO.	2.86	174	1,352,643
NIKE INC.	67.39	178	38,033,900
HOME DEPOT INC.	111.57	241	1,593,204
MATTEL INC.	15.02	125	613,798
GAMESTOP CORP.	6.41	167	1,209,680

(*) Market Capitalization ($Billions) as of October 31, 2013. Source: Thomson Reuters Eikon.
(**) Number of news analyzed, i.e., filtered with relevance score equals to 100.

2.1.3 Social Media Sentiment on U.S.-listed Companies

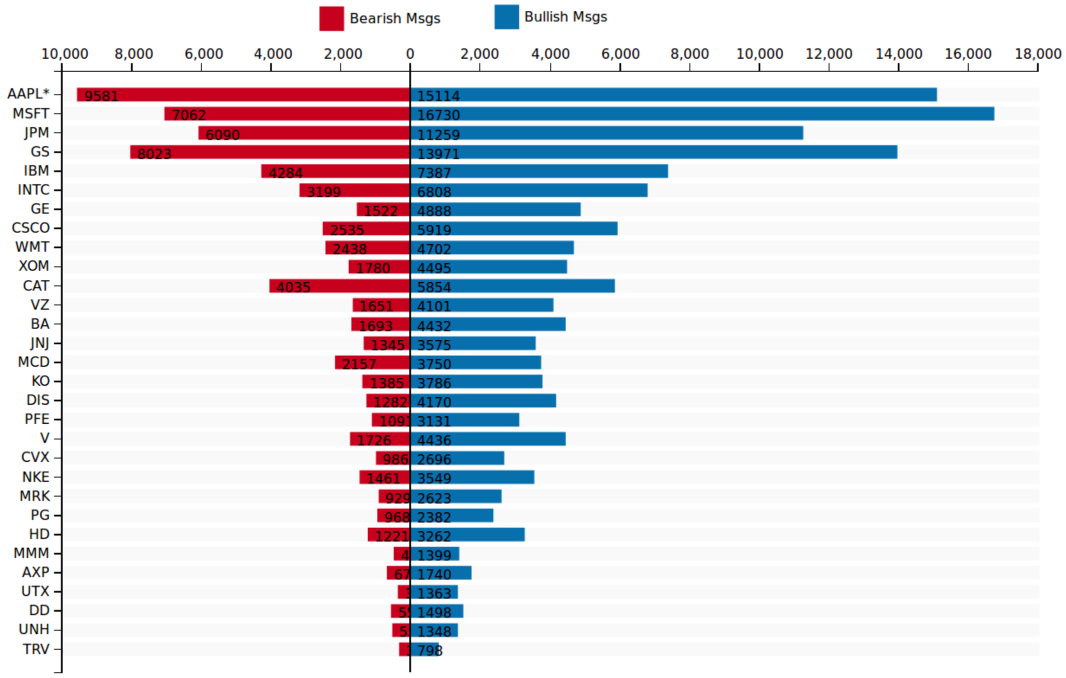
In Chapters 5 and 6, we expanded the scope of social media sentiment to companies that were representative of the entire U.S. market. In Chapter 5, we considered companies constituents of the Dow Jones Industrial Average Index while in Chapter 6, we considered social media sentiment on the top 100 most market capitalized companies constituents of the S& P 500 index.

The data were supplied by PsychSignal.com [37] and they were comprised of volume and sentiment measures. In this dataset, a company is defined to be related to a given message if its ticker-id is mentioned as a *cashtag*, i.e., with its name preceded by a dollar symbol, e.g., $CSCO for the company CISCO SYSTEMS INC. In Twitter, a *cashtag* is a standard way to refer to a listed security. Twitter messages are classified according to their likelihood of bullishness and bearishness toward a company. Fig. 2.1 shows the volume of bearish and bullish messages for the selected companies. The dataset is based on English language content and it ignores the source country. The information is aggregated in a daily fashion within a 24-hour window that ends at 8AM EST and it is composed of the following variables:

- symbol: the stock symbol (ticker) for which the sentiment data refers to,
- timestamp_utc: date and time of the analyzed data in UTC format,
- bull_scored_messages: daily total count of bullish sentiment messages, and
- bear_scored_messages: daily total count of bearish sentiment messages.

Some messages may be classified as “neutral” or at least not having relevant bullish or bearish tones. That type of messages does not affect the bull_scored_messages and bear_scored_messages scores. It is also possible that no messages cite a company in a given day. In that case, the scores are zero.

Table 2.4 shows a summary description of the selected companies with the number of bearish/bullish Twitter messages identified in the period. We have provided further descriptive analytics of the Twitter sentiment dataset used in related literature [38, 39].



(*) A factor of 0.1 was applied in Apple's Twitter volume in the figure to conform all stocks in the same scale.

Figure 2.1: Volume of bearish and bullish Twitter messages mentioning a ticker of a stock component of the Dow Jones Industrial Average (DJIA) index monitored during the two-year period from March 31, 2012 to March 31, 2014.

2.2 Methods

2.2.1 Granger-causality

We quantify causality by using the notion of the causal relation introduced by Granger [40, 41] where a signal X is said to Granger-cause Y if the future realizations of Y can be better explained using the past information from X and Y rather than Y alone.

The most common definitions of Granger-causality (G-causality) rely on the prediction of a future value of the variable Y by using the past values of X and Y itself. In that form, X is said to G-cause Y if the use of X improves the prediction of Y. We follow the notation from [42, 43]. Let X_t be a random variable associated at time t while X^t represents the collection of random variables up to time t. We consider X_t, Y_t and Z_t to be three stochastic processes. Let $\hat{Y}_{t+1}$ be a predictor for the value of the variable Y at time $t+1$. We compare the expected value of a loss function $g(e)$ with the error $e = \hat{Y}_{t+1} - Y_{t+1}$ of two models:

Table 2.4: Summary table of the selected companies. It includes the DJIA index constituents along with their total and daily mean numbers of bearish and bullish tweets during the period from March 31, 2012 to March 31, 2014. The number of total messages processed includes bullish, bearish and neutral messages.

	Bullish messages		Bearish messages		
Ticker	Total	Daily mean	Total	Daily mean	Total Messages
AAPL	151143	279.89	95819	177.443	800638
MSFT	16730	30.98	7062	13.078	139343
JPM	11259	20.85	6090	11.278	82265
GS	13971	25.87	8023	14.857	75578
IBM	7387	13.68	4284	7.933	53547
INTC	6808	12.61	3199	5.924	47653
GE	4888	9.05	1522	2.819	41271
CSCO	5919	10.96	2535	4.694	39665
WMT	4702	8.71	2438	4.515	39607
XOM	4495	8.32	1780	3.296	33194
CAT	5854	10.84	4035	7.472	31911
VZ	4101	7.59	1651	3.057	30936
BA	4432	8.21	1693	3.135	30421
JNJ	3575	6.62	1345	2.491	28392
MCD	3750	6.94	2157	3.994	28059
KO	3786	7.01	1385	2.565	26331
DIS	4170	7.72	1282	2.374	25863
PFE	3131	5.80	1091	2.020	24817
V	4436	8.21	1726	3.196	24118
CVX	2696	4.99	986	1.826	21322
NKE	3549	6.57	1461	2.706	20941
MRK	2623	4.86	929	1.720	20708
PG	2382	4.41	968	1.793	20226
HD	3262	6.04	1221	2.261	17550
MMM	1399	2.59	465	0.861	12382
AXP	1740	3.22	674	1.248	12072
UTX	1363	2.55	369	0.690	11255
DD	1498	2.78	559	1.037	10746
UNH	1348	2.50	532	0.987	9196
TRV	798	1.53	316	0.604	7990
TOTAL	287195	-	157597	-	1767997

1) The expected value of the prediction error given only Y^t

$$\mathscr{R}(Y^{t+1}\,|\,Y^t,Z^t) = \mathbb{E}[g(Y_{t+1} - f_1(X^t,Z^t))] \tag{2.1}$$

2) The expected value of the prediction error given Y^t and X^t

$$\mathscr{R}(Y^{t+1} \,|\, X^t, Y^t, Z^t) = \mathbb{E}[g(Y_{t+1} - f_2(X^t, Y^t, Z^t))]. \tag{2.2}$$

In both models, the functions $f_1(.)$ and $f_2(.)$ are chosen to minimize the expected value of the loss function. In most cases, these functions are retrieved with linear and, possibly, with nonlinear regressions. Typical forms for $g(.)$ are the $l1$- or $l2$-norms.

Definition 1. *X does not Granger-cause Y relative to side information Z if and only if* $\mathscr{R}(Y_{t+1} \mid X^t, Y^t, Z^t) = \mathscr{R}(Y_{t+1} \mid Y^t, Z^t)$.

A more general definition than Def. 1 that does not depend on assuming prediction functions can be formulated by considering the conditional probabilities. A probabilistic definition of G-causality assumes that Y_{t+1} and X^t are independent given the past information (X^t, Y^t) if and only if $p(Y_{t+1} \,|\, X^t, Y^t, Z^t) = p(Y_{t+1} \,|\, Y^t, Z^t)$, where $p(.\,|\,.)$ represents the conditional probability distribution. In other words, omitting past information from X does not change the probability distribution of Y.

Definition 2. *X does not Granger-cause Y relative to side information Z if and only if* $Y_{t+1} \perp\!\!\!\perp X^t \mid Y^t, Z^t$.

Def. 2 does not assume any functional form in the coupling between X and Y. Nevertheless, it requires a method to assess the conditional dependency.

In Section 2.2.2, we define a parametric linear specification of G-causality based on Def. 1; In Section 2.2.4, we define a non-linear specification of G-causality based on Def. 2 using an information-theoretical framework.

2.2.2 Linear G-causality

Standard Granger-causality tests assume a linear relationship among the causes and effects and are implemented by fitting autoregressive models [40, 41].

Consider the linear vector-autoregressive (VAR) equations:

$$R(t) = \alpha + \sum_{\Delta t=1}^{k} \beta_{\Delta t} R(t-\Delta t) + \varepsilon_t, \tag{2.3}$$

$$R(t) = \widehat{\alpha} + \sum_{\Delta t=1}^{k} \widehat{\beta}_{\Delta t} R(t-\Delta t) + \sum_{\Delta t=1}^{k} \widehat{\gamma}_{\Delta t} SM(t-\Delta t) + \widehat{\varepsilon}_t, \tag{2.4}$$

where k is the number of lags considered. From Def 1, SM does not G-cause R if and only if the prediction errors of R in the restricted Eq. (4.4) and unrestricted regression models Eq. (4.5) are equal (i.e., they are statistically indistinguishable). A one-way ANOVA test is utilized to test if the residuals from Eqs. (4.4) and (4.5) differ from each other significantly. When more than one lag k is tested, a Bonferroni correction is applied to control for multiple hypotheses testing. Finally, a significant causal relationship can be reported only if the linear models from Eqs. (4.4) and (4.5) are not misspecified. For that purpose, we utilize the BDS test [44] for the model misspecification (see Section 2.2.3).

2.2.3 BDS Test for Linear Misspecification

The BDS test [44] is used to detect nonlinear dependence in time series. When applied to the residuals of a linear model, the BDS tests the null hypothesis that these residuals are independent and identically distributed. The BDS test is a powerful test to detect linear misspecification and nonlinearity [44, 45]. Let $\varepsilon_t = (\varepsilon_{t=1}, \ldots, \varepsilon_{t=n})$ be the residuals of the linear fitted model and define its m-embedding as $\varepsilon_t^m = (\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-m+1})$. The m-embedding correlation integral is given by

$$C_{m,n}(\Delta\varepsilon) = \frac{2}{k(k-1)} \sum_{s=1}^{t} \sum_{t=s}^{n} \chi(\|\varepsilon_s^m - \varepsilon_t^m\|, \Delta\varepsilon),$$

and

$$C_m(\Delta\varepsilon) = \lim_{n\to\infty} C_{m,n}(\Delta\varepsilon),$$

where χ is an indicator function where $\chi(\|\varepsilon_s^m - \varepsilon_t^m\|, \Delta\varepsilon) = 1$ if $\|\varepsilon_s^m - \varepsilon_t^m\| < \Delta\varepsilon$ and zero, otherwise. The null hypothesis of the BDS test assumes that ε_t is iid. In this case,

$$C_m(\Delta\varepsilon) = C_1(\Delta\varepsilon)^m.$$

The BDS statistic is a measure of the extent that this relation holds in the data. This statistic is given by the following:

$$V_m(\Delta\varepsilon) = \sqrt{n}\frac{C_m(\Delta\varepsilon) - C_1(\Delta\varepsilon)^m}{\sigma_m(\Delta\varepsilon)},$$

where $\sigma_m(\Delta\varepsilon)$ can be estimated as described in [44]. The null hypothesis of the BDS test indicates that the model tested is not misspecified and it is rejected at the 5% significance level if $\|V_m(\Delta\varepsilon)\| > 1.96$.

The parameter $\Delta\varepsilon$ is commonly set as a factor of the variance (σ_ε) of ε. We report results for $\Delta\varepsilon = \sigma_\varepsilon/2$ and the embedding dimension $m = 2$. We also performed tests for $\Delta\varepsilon = \sigma_\varepsilon$ and $m = 3$ with no significant differences in the results.

2.2.4 Nonlinear G-Causality

To compute the nonlinear G-Causality, we use the concept of Transfer Entropy that, since its introduction by Schreiber (2000) [46], has been recognized as an important tool in the analysis of causal relationships in nonlinear systems [47].

Let us first introduce basic information theory concepts. Next, we define Transfer Entropy and its estimation as a causality measure. Let X be a random variable and $P_X(x)$ be its probability density function (pdf). The entropy $H(X)$ is a measure of the uncertainty of X and is defined in the discrete case as follows:

$$H(X) = -\sum_{x \in X} P_X(x) \log P_X(x). \tag{2.5}$$

Given a coupled system (X, Y), where $P_Y(y)$ is the pdf of the random variable Y and $P_{X,Y}$ is the joint pdf between X and Y, the joint entropy between X and Y is given by the following:

$$H(X,Y) = -\sum_{x \in X}\sum_{y \in Y} P_{X,Y}(x,y)\log P_{X,Y}(x,y). \tag{2.6}$$

The conditional entropy is defined by the following:

$$H(Y|X) = H(X,Y) - H(X). \tag{2.7}$$

We can interpret $H(Y|X)$ as the uncertainty of Y given a realization of X.

The Transfer Entropy can be defined as the difference between the conditional entropies:

$$TE(X \rightarrow Y|Z) = H\left(Y^F \middle| Y^P, Z^P\right) - H\left(Y^F \middle| X^P, Y^P, Z^P\right), \tag{2.8}$$

where Y^F is a forward time-shifted version of Y at lag Δt relatively to the past time-series X^P, Y^P and Z^P. Within this framework we say that X does not G-cause Y relative to side information Z if and only if $H\left(Y^F \middle| Y^P, Z^P\right) = H\left(Y^F \middle| X^P, Y^P, Z^P\right)$, i.e., when $TE\left(X \rightarrow Y, Z^P\right) = 0$.

Empirically, we reject this null hypothesis of causality if the Transfer Entropy from X to Y is significantly higher than the shuffled version of the original data. For this we estimate 400 replicates of $TE(X_{Shuffled} \rightarrow Y)$, where $X_{Shuffled}$ is a random permutation of X relatively to Y. We compute the randomized Transfer Entropy at each permutation for each time-shift (Δt) from 1 to 10 days. We then calculated the frequency at which the observed Transfer Entropy was equal or more extreme than the randomized Transfer Entropy. The statistical significance was assessed using p-value < 0.05 after Bonferroni correction.

The estimation of the empirical probability density distribution, which is required for the entropy estimation, was performed using the Kernel Density Estimation (KDE) method, which has several advantages over the frequently used histogram-based methods (see Section 2.2.5 for more details).

2.2.5 Kernel Density Estimation

In the entropy computation, the empirical probability distribution must be estimated. Histogram-based methods and kernel density estimations are the two main methods for that. Histogram-based is the simplest and most used nonparametric density estimator. Nonetheless, it yields density estimates that have discontinuities and vary significantly depending on the bin size choice.

Also known as the Parzen-Rosenblatt window method, the kernel density estimation (KDE) approach approximates the density function at point x using neighboring observations. However, instead of building up the estimate according to the bin edges as in histograms, the KDE method uses each point of estimation x as the center of a bin of width $2h$ and weight it according to a kernel function. Thereby, the kernel estimate of the probability density function $f(x)$ is defined as

$$\hat{f} = \frac{1}{nh} \sum_{x' \in X} K\left(\frac{x - x'}{h}\right). \tag{2.9}$$

A usual choice for the kernel K, which we use here, is the (Gaussian) radial basis function:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}x^2}. \tag{2.10}$$

The problem of selecting the bandwidth h in equation (2.9) is crucial in the density estimation. A large h will oversmooth the estimated density and mask the structure of the data. On the other hand, a small bandwidth will reduce the bias of the density estimate at the expense of a larger variance in the estimates. If we assume that the true distribution is Gaussian and we use a Gaussian kernel, the optimal value of h that minimizes the mean integrated squared error (MISE) is

$$h^* = 1.06\sigma N^{-1/5},$$

where N is the total number of points and σ can be estimated as the sample standard deviation. This bandwidth estimation is often called the Gaussian approximation or Silverman's rule of thumb for kernel density estimation [48]. This is the most commonly-used method and it is here employed. Other common methods are given by Sheather and Jones [49] and Scott [50].

2.2.6 Quantifying Linear Granger-causality using Transfer Entropy

Barnett et al. (2009) [51] showed that linear G-causality and Transfer Entropy are equivalent if all processes (X and Y) are jointly Gaussian. In particular, by assuming the standard measure ($l2$-norm loss function) of linear G-causality for the bivariate case as

$$GC_{X\rightarrow Y} = \log\left(\frac{var(\varepsilon_t)}{var(\widehat{\varepsilon}_t)}\right), \tag{2.11}$$

the following can be proved [51]:

$$TE_{X\rightarrow Y} = GC_{X\rightarrow Y}/2. \tag{2.12}$$

This result provides a direct mapping between the Transfer Entropy and the linear G-causality implemented in the standard VAR framework. Hence, it is possible to estimate the TE both in its general form and with its equivalent form for linear G-causality.

2.2.7 Net Information Flow

Transfer-entropy is an asymmetric measure, i.e., $T_{X\rightarrow Y} \neq T_{Y\rightarrow X}$, and it thus allows the quantification of the directional coupling between systems. The Net Information Flow is defined as

$$\widehat{TE}_{X\rightarrow Y} = TE_{X\rightarrow Y} - TE_{Y\rightarrow X}\ . \tag{2.13}$$

One can interpret this quantity as a measure of the dominant direction of the information flow. In other words, a positive result indicates a dominant information flow from X to Y compared to the other direction or, similarly, it indicates which system provides more predictive information about the other system [52].

For the nonlinear case, the Transfer Entropy was computed as defined in Eq. (2.8). Conversely, to estimate the linear version of the Net Information Flow, we computed the Transfer Entropy using Eq (2.12), i.e., we estimated the linear G-causality (2.11) and multiplied it by a factor of 1/2.

In the next section, we construct simulated systems and test the nonlinear and linear formulations of the net information flow. We show that only the nonlinear formulation of net information flow is able to capture the nonlinear relationships in the simulated systems.

2.2.8 Empirical Experiment: Information Flow on Simulated Systems

In this section, we simulate two systems: the first has purely linear dependencies, and the second introduces nonlinear dependencies. We quantify the information flow among the variables of each system using the linear and nonlinear methods introduced in Section 2.2.6. We show that nonlinear interactions are captured only with the nonlinear formulation of the information flow, i.e., the approach based on the Transfer Entropy. The simulated systems were based on [53], which performed similar analysis while comparing the linear and nonlinear dependencies of artificially created systems.

We first define a linear system composed of 5 variables with the dependencies described in Eqs. 2.14-2.18 as following:

$$x_1(n) = 0.95\sqrt{2}x_1(n-1) - 0.9025x_1(n-1) + w_1 \tag{2.14}$$

$$x_2(n) = 0.5x_1(n-1) + w_2 \tag{2.15}$$

$$x_3(n) = -0.4x_1(n-1) + w_3 \tag{2.16}$$

$$x_4(n) = -0.5x_1(n-1) + 0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_4 \tag{2.17}$$

$$x_5(n) = -0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_5, \tag{2.18}$$

where $w_1, w_2, w_3, w_4, w_5 \sim N(0,1)$. To simulate this system we assume $x_i(0) = 0, i \in (1,2,...,5)$ as initial condition and then iteratively generate x_i for $n \in (1,2,...,N)$ with a total of $N = 200,000$ iterations by randomly sampling $w_i, i \in (1,2,...,5)$ from a normal distribution with zero mean and unit variance.

The Fig. 2.2 A) represents the dependencies of the simulated linear system.

The Fig. 2.2 B) and Fig. 2.2 C) show the linear and nonlinear Information Flows among the system's variables, respectively. A cell (x,y) presents the information flow from variable y to variable x. From Fig. 2.2, we observe that both the linear and nonlinear approaches presented similar results, i.e., both methods captured the system's dependencies similarly. This result is expected as the system is purely linear and the nonlinear information flow is able to capture both the linear and nonlinear interactions. We define a second system by introducing nonlinear interactions in Eqs. 2.15 and 2.17 as defined in Eqs. 2.19-2.23 as following:

$$x_1(n) = 0.95\sqrt{2}x_1(n-1) - 0.9025x_1(n-1) + w_1 \tag{2.19}$$

$$x_2(n) = 0.5x_1^2(n-1) + w_2 \tag{2.20}$$

$$x_3(n) = -0.4x_1(n-1) + w_3 \tag{2.21}$$

$$x_4(n) = -0.5x_1^2(n-1) + 0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_4 \tag{2.22}$$

$$x_5(n) = -0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_5, \tag{2.23}$$

where w_1, w_2, w_3, w_4 and $w_5 \sim N(0,1)$. To simulate this system we assume $x_i(0) = 0, i \in (1,2,...,5)$ as initial condition and then iteratively generate x_i for $n \in (1,2,...,N)$ with a total of $N = 200,000$ iterations by randomly sampling $w_i, i \in (1,2,...,5)$ from a normal distribution with zero mean and unit variance.

The Fig. 2.3 A) represents the dependencies of the simulated nonlinear system. This system has two nonlinear interactions: the first is between variables x_1 and x_2 as defined by Eq. 2.20, and the second is between variables x_1 and x_4 as defined by Eq. 2.22. From Fig. 2.3 B) and Fig. 2.3 C), we observe that the nonlinear interactions introduced were not captured by the linear form of the information flow. While all linear interactions presented similar linear and nonlinear information flows, the two nonlinear interactions introduced in the system presented relatively higher nonlinear information flow compared to the linear formulation.

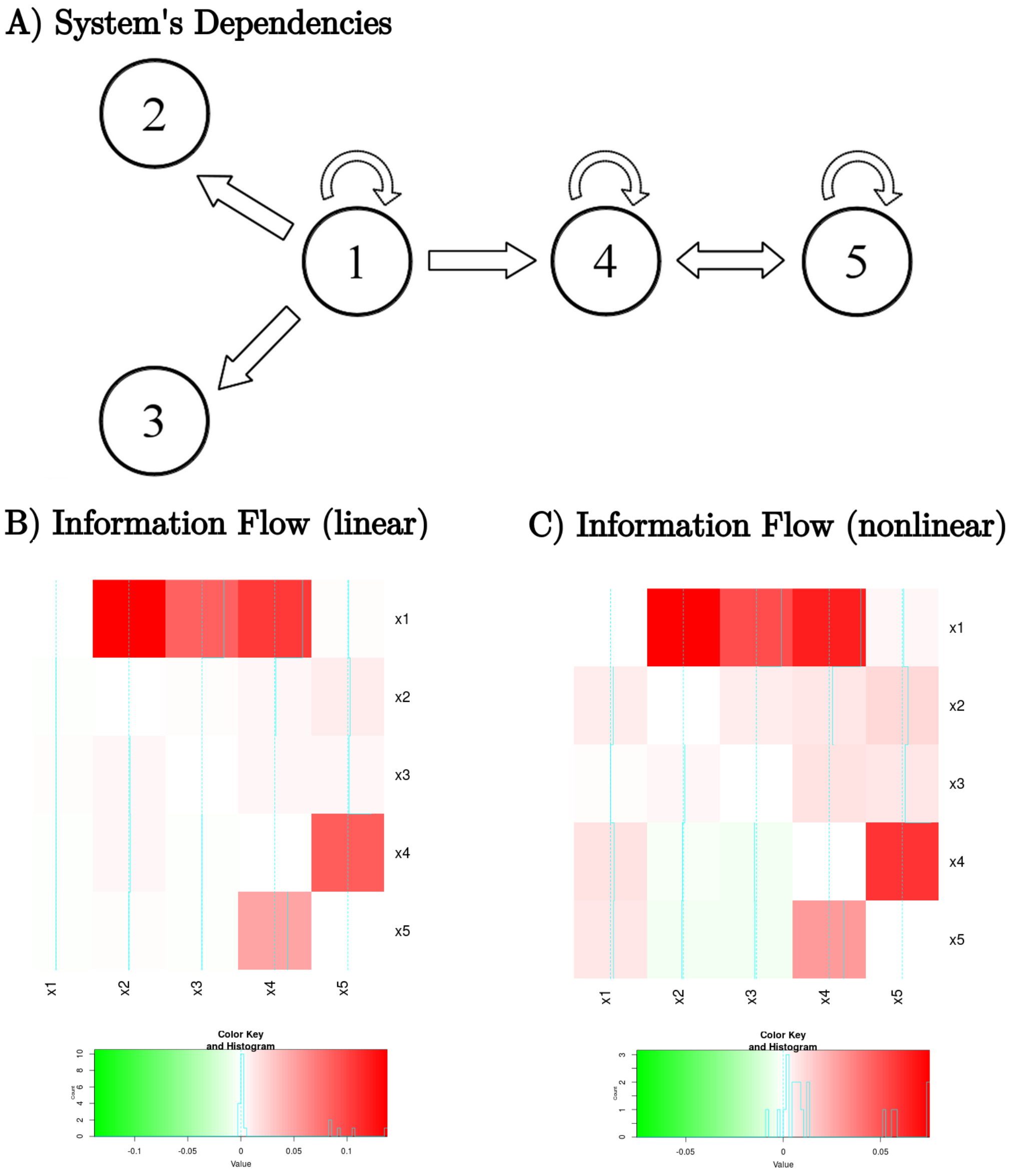


Figure 2.2: Evidence that the linear and nonlinear formulations of the information flow are able to capture the dependencies among the simulated system. The panel A) represents the dependencies of the simulated linear system. The panel B) shows the information flow among the variables using a linear formulation based on the Granger-causality estimated with a linear vector autoregression formulation. The panel C) shows the nonlinear information flow among the system's variables based on the Transfer Entropy. A cell (x, y) represents the information flow from variable y to x.

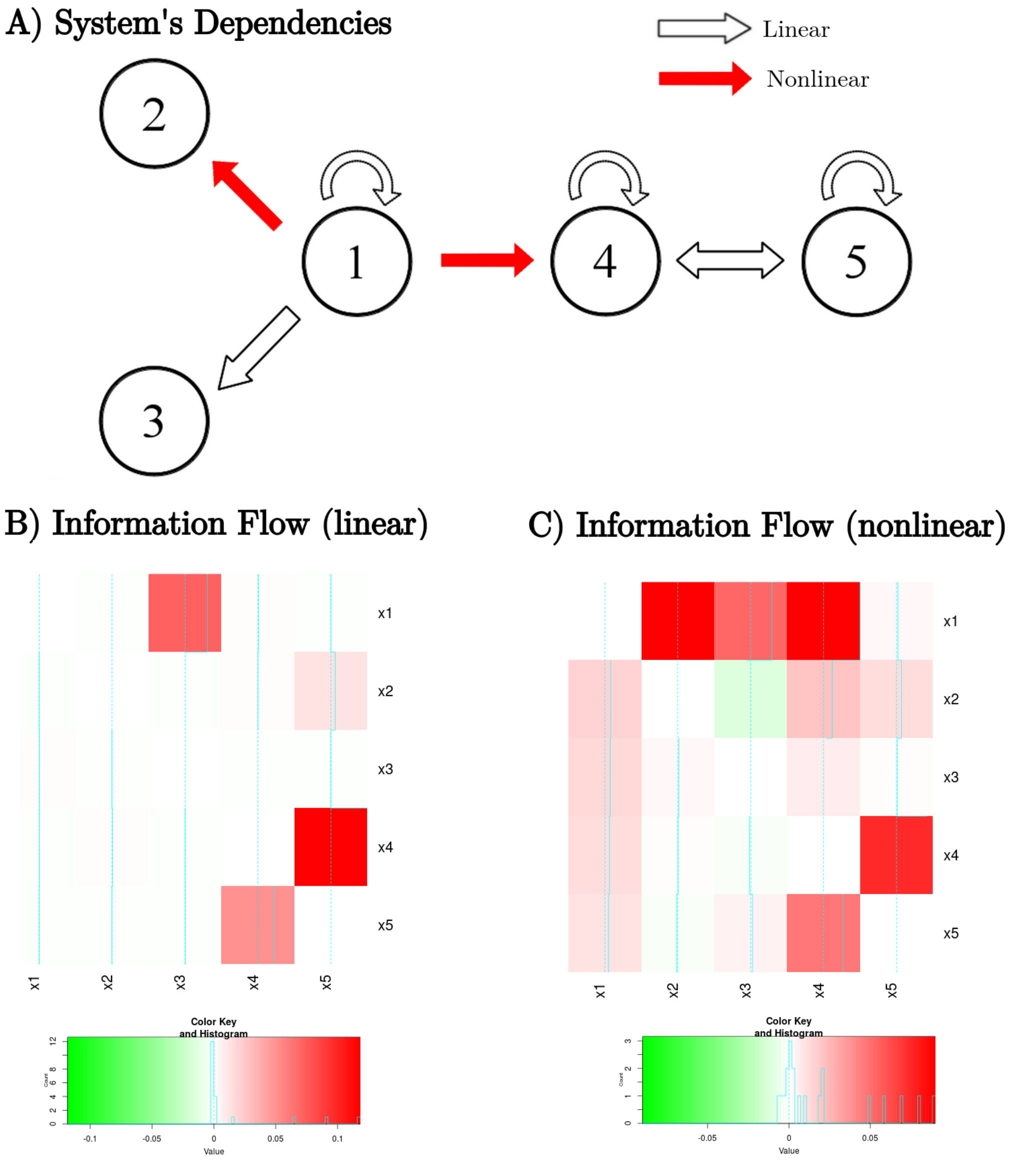


Figure 2.3: Evidence that only the nonlinear formulation of information flow was able to capture the nonlinear dependencies introduced in the simulated system. The panel A) represents the dependencies of the simulated system, which presents nonlinear dependencies between variables x_1 and x_2, and between variables x_1 and x_4. The panel B) shows the information flow among the variables using a linear formulation based on the Granger-causality estimated with a linear vector autoregression formulation. The panel C) shows the nonlinear information flow among the system's variables based on the Transfer Entropy. A cell (x, y) represents the information flow from variable y to x.

Chapter 3

Literature Review

Different sources of exogenous information can impact financial markets [54] including (i) News, (ii) Pre-News, (iii) Rumours and (iv) Social Media. Mitra and Mitra (2011) [9] present the corresponding descriptions as following:

- News: this refers to mainstream media and comprises the news stories produced by reputable sources. These are broadcast via newspapers, radio and television. They are also delivered to traders' desks on newswire services. On-line versions of newspapers may also exist.

- Pre-News: this refers to the source data that reporters research before they write news articles. It comes from primary information sources such as, Securities and Exchange Commission reports and filings, court documents and government agencies. It also includes scheduled announcements such as macro economic news, industry statistics, company earnings reports and other corporate news.

- Rumours: these are blogs and websites that broadcast "news", and are less reputable than news and pre-news sources. The quality of these vary significantly. Some may be blogs associated with highly reputable news providers and reporters. At the other end of the scale some blogs may lack any substance and may be entirely fueled by rumour.

- Social Media: these websites fall at the lowest end of the reputation scale. Barriers to entry are extremely low and the ability to publish "information" is easy. These can be dangerously inaccurate sources of information. However, if carefully applied there may be some value to be gleaned from these.

Financial news can also be divided into two categories [9]: (i) regular synchronous announcements (scheduled or expected news) and (ii) event driven asynchronous announcements (unscheduled or unexpected news). Main-stream news, rumors and social media normally arrive asynchronously in an unstructured textual form. A substantial portion of pre-news arrive at pre-scheduled times and generally in a structured form. Both categories have in common that the analyzed data is textual, non-numeric and qualitative in nature. We may wish to distinguish whether a story's informational content is positive or negative, that is, determine its sentiment. For that we need to transform the data into quantitative information before it can be considered in the financial decision making process. After this "pre-analysis" phase of turning qualitative text into quantified metrics we can develop predictive models in order to update beliefs and provide ex-ante view of the market environment. Figure 3.1 shows a summary of this news analytics information flow [8].

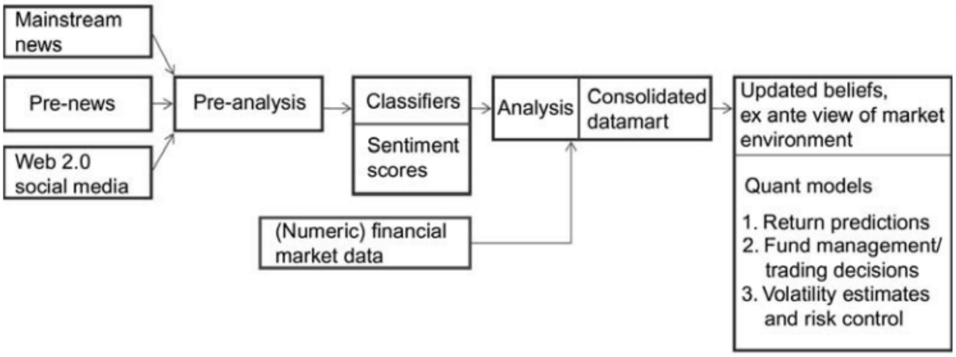

Mainstream news
Pre-news
Web 2.0 social media
Pre-analysis
Classifiers
Sentiment scores
(Numeric) financial market data
Analysis
Consolidated datamart
Updated beliefs, ex ante view of market environment
Quant models
1. Return predictions
2. Fund management/ trading decisions
3. Volatility estimates and risk control

Figure 3.1: News analytics information flow architecture [8]. Stories from News and Social Media are classified and transformed into quantitative scores. These metadata are combined with market data to update ex-ante beliefs of the market.

3.1 Sentiment Analysis

Sentiment analysis is an area of research that investigates people's opinions towards different matters, e.g., products, events or organizations [55]. The role of sentiment analysis has been growing significantly with the rapid spread of social networks, microblogging applications and forums. Mining this volume of opinions provides information for understanding collective human behaviour and it is of valuable commercial interest. For instance, an increasing amount of evidence points out that by analyzing sentiment of social media content it might be possible to predict relevant economic and financial indicators such as sales [36], stock prices [6] and unemployment rates over time [56].

The field of text categorization was initiated long time ago [57], however categorization based on sentiment was introduced more recently in [58, 59, 60]. The main two approaches to implement sentiment analysis are the lexicon-based method (unsupervised approach) and the machine learning based method (supervised approach). Both approaches rely on the bag-of-words method [61, 62], where a document is represented as a vector of words in Euclidean space where each word is independent from others.

In the lexicon-based method, the unigrams which are found in the lexicon are assigned a polarity score, the overall polarity score of the text is then computed as sum of the polarities of the unigrams. In the machine learning supervised method, the classifiers are using the unigrams or their combinations (N-grams) as features. In [35], we described several techniques to implement these approaches and discuss how they can be adopted for sentiment classification of Twitter messages. We presented a comparative study of different lexicon combinations and showed that enhancing sentiment lexicons with emoticons, abbreviations and social media slang expressions increases the accuracy of lexicon-based classification for Twitter. We discussed the importance of feature generation and feature selection processes for machine learning sentiment classification. We presented a new ensemble method that uses a lexicon based sentiment score as input feature for the machine learning approach. The combined method proved to produce more precise sentiment classifications. Later we leveraged this methodology to predict sales of retail brands [36]. In Chapter 4, we leverage the same dataset to investigate the interplay of Twitter sentiment extracted from listed retail brands with stock returns and volatility.

3.2 News Analytics

Major news announcements can have a high impact on financial markets and investors behaviour resulting in rapid changes or abnormal effects in financial portfolios. Although the relevance of news is widely acknowledged, how to incorporate this information channel effectively in quantitative models is a open question [9].

As human responsiveness is limited, automated news analysis has recently been developed as a fundamental component to algorithmic trading. In this way, traders can shorten the time of reaction in response to breaking stories. The basic idea behind news analytics technologies is to quantify human sentiment and automate human behaviour systematically, so traders may be able to anticipate asset movements before making an investment or risk management decision.

Das [18] presents a framework for news analytics techniques used in finance and Banerjee [17] describes a general architecture for news analytics predictive analysis (see Figure 3.2). Typically, a model is composed of endogenous and exogenous variables. The endogenous variables are comprised by market data such as bid and ask prices. The exogenous variables include news metadata information such as sentiment, news novelty and relevance or an event category. A news predictive model can be constructed by combining endogenous and exogenous variables for the prediction of a target outcome variable such as asset prices, returns, volatility or liquidity. The assumption is that news analytics can provide information exogenous to the market that might not have been yet captured in the prices hence providing information that can complement past market data information.

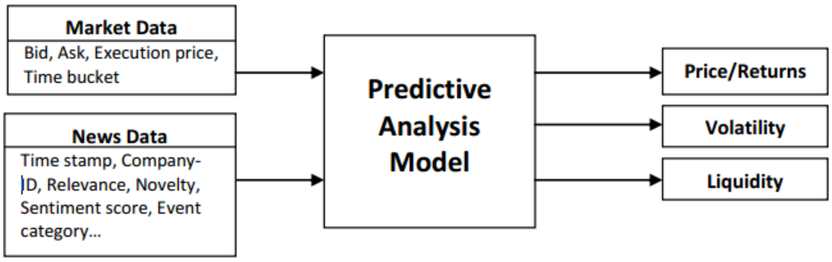

Market Data
Bid, Ask, Execution price, Time bucket
News Data
Time stamp, Company-ID, Relevance, Novelty, Sentiment score, Event category...
Predictive Analysis Model
Price/Returns
Volatility
Liquidity

Figure 3.2: Architecture of predictive analysis model [17]. Market data are combined with News meta-data in order to enhance the prediction of financial variables such as stock price returns, volatility and liquidity.

The investigation of the market impact of News has been long studied since the seminal work of Cutler et al. [63], where the authors estimated for the first time the explanatory power of economic news on stock returns while putting into question the view that large stock price movements are preceded by news. Tetlock [2] provides the first evidence that news media content can predict movements in broad indicators of stock market activity. His findings suggested the presence of a relationship between pessimism of media and high market trading volume. While news volume is skewed to blue chip companies [64, 65], it has been shown that news sentiment is widely linked to overall corporate earnings [11]. Since then, with the availability of machine readable news and the use of sentiment analysis [35], several works have found news as a significant source to explain financial dynamics. For instance in [14] is has been shown that the rate of information arrival impacts return volatility by analyzing both the S&P/ASX 200 Index and SPI 200 Futures. The authors showed that there is a positive correlation between the frequency of the incoming news and the volatility of the stock returns. Also, the systematic risk (or beta) of individual stocks increases by an economically and statistically significant amount on days of firm-specific news announcements [66].

Firm-specific news information flow has been shown to impact the systematic risk of an individual firm, measured by its CAPM beta [15]. Findings indicated that (i) betas increase on announcement days by a statistically and economically significant amount; (ii) covariance of the announcing stock returns with the returns of other stocks in the market index increases significantly on announcement dates. News analytics have also been used to model volatility by extending GARCH models [67, 68, 69]. For instance, the "daily number of press releases on a stock" (news intensity) has been considered as an exogenous variable in the traditional GARCH model [68]. The results of the likelihood ratio test indicated that the GARCH(1,1) model augmented with the news intensity performs better than the original GARCH model.

When new information hits the market, investors may behave differently according to its investment type, risk profile, regulatory constraints etc. Indeed, it has been shown [12] that households and companies are very sensitive to both endogenous (return, volatility) and exogenous (news volume and sentiment) factors. On the other hand, governmental and non-profit organization are weakly affected by those factors. While the presence of news can impact the market in different forms, the absence of news is also an important indicator. For instance, it has been found that stocks with no media coverage earn, on average, higher returns than stocks with high media coverage even after controlling for well-known risk-factors [70].

Positive correlation has been found [13] between the number of mentions of a company in the Financial Times and its stock's trading volume. News sentiment available from Thomson Reuters News Analytics (TRNA) were found to have causality with stock volatility and liquidity [71]. Moreover, it has been shown that news analytics can be a useful data source for commodity [72, 73], fixed-income [74] and FX trading [75, 76] as well as used in high-frequency trading [77], short-selling [78] and event detection [79].

3.3 Social Media Analytics

Twitter data have become an increasingly important source to describe financial dynamics. It provides a fine-grained real-time information channel that includes not only major news stories but also minor events that, if properly modeled, can provide ex-ante information about the market even before the main newswires. Recent developments have reflected this prominent role of social media in the financial markets; for instance, the U.S. Securities and Exchange Commission report allowing companies to use Twitter to announce key information in compliance with Regulation Fair Disclosure [80]. Another example is the so-called *Hash Crash* which happened in 2013 when the Twitter account of American news agency Associated Press was hacked and used to falsely disclose a message about an attack on the White House causing a drop in the Dow Jones Industrial Average of 145 points in minutes [81]. More recently on August 28, 2018, the stock price of the company Tesla, Inc. surged 10% after its CEO, Elon Musk, shocked the marked with a tweet indicating that the company would be considering going private.

Some of research has shown that Twitter data contain relevant information related to financial indicators. As one of the first investigations analyzing Twitter in the context of financial markets, the work in [6] analyzed the text content of daily Twitter feeds to identify two types of mood: (i) polarity (positive vs. negative) and (ii) emotions (calm, alert, sure, vital, kind, and happy). By using a non-linear model based on self-organizing map (SOM), the authors were able to increase the accuracy in the prediction of the DJIA index when using social media analytics. Another research [19] combined Information Theory with sentiment analysis to demonstrate that Twitter sentiment can contain statistically significant ex-ante information on individual securities prices including future prices of the S&P500 index. In Chapter 4 we provide evidence that suggests that social media analytics play an important role in explaining the price dynamics of the retail sector even when compared to mainstream news. J. Manfield, D. Lukacsko and T. T. P. Souza [38] later reported broader results covering a large set of stocks from both the NYSE and NASDAQ exchanges, which confirms a statistically significant coupling between social sentiment and stock prices volatility.

As a contribution to the field of event study research, market reactions to combinations of different types of news events have been studied [7] using Twitter to identify which news are more important from the investor perspective. In a similar way, sentiment analytics has been combined with the identification of Twitter peaks in an event study approach [82] to imply directions of market evolution.

By utilizing a nonparametric formulation of statistical causality, in Chapter 4 we uncover that information flows from social media to stock markets, revealing that tweets are causing markets movements through a nonlinear complex interaction. Our findings thus question some current modeling and analytics that assume linearity. The results also serve as empirical guidance on model adequacy, market efficiency, and predictability in the investigation of causal relationships between social and financial systems.

3.4 News Analytics and Social Media Opinion: The Portfolio Case

The relationship between social and financial systems is commonly modeled as a univariate problem where single assets are isolated from the system when its dynamics are analyzed. In that way, only individual links are investigated and the multi-asset case is often neglected. For instance, social media based polarity signals are utilized to predict stock returns of an individual stock and data on other correlated assets and portfolio dynamics are neglected.

Harry Markowitz formulated the first portfolio theory, entitled of “Modern Portfolio Theory” which was the first systematic financial theory [83]. Modern portfolio theory evaluates the portfolio using a mean-variance pattern and represents a normative pattern for portfolio selection. Portfolio choice is made by solving an optimization problem, in which the portfolio risk is minimized and a desired level of expected return is specified as a constraint. This theory assumes economic equilibrium and was the basis for other financial theories such as the efficient market hypothesis by Fama [84]. Further, the need to penalize different undesirable aspects of the return distribution and the consideration of asymmetric risk led to the proposal of alternative risk measures that penalize the downside return and not its upside. These considerations constitute the basis of the Post Modern Portfolio Theory (PMPT) [85]. Examples of such risk measures are lower partial moments, Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR) [86].

Traditional portfolio selection and risk models have taken historic asset prices as fundamental data in order to predict future behaviors of the financial market. These traditional approaches have the disadvantage that they provide ex-post retrospective strategies. They generally do not take sudden market changes into account and fail to account changes in the investor sentiment. Classical VaR calculation assumes that only the risk of single assets and their correlation (or dependence) matters. As a consequence, this makes VaR inflexible and unresponsive with regard to abnormal market conditions. By incorporating news, social media and investor opinion into the portfolio risk calculation, sudden impactful events can help estimate the probability of emerging abnormal market conditions.

It has been found that updating portfolio risk estimates using news data can provide dynamic (adaptive) measures that account for the market environment [16]. Further these measures may be useful in identifying and giving early ex-ante warning of extreme risk events. Incorporating recent sentiment of the market environment within the estimation of portfolio risk is important, since the market conditions are likely to vary from historic observations. This is particularly important when there are sudden major changes in the market. In these cases, risk measures, calibrated using historic data alone, fail to capture the true level of risk [16, 87].

The evidence that social media is a valuable source of information about the future evolution of the stock market supports the idea that, apart from pure economic and fundamental factors, there is an emotional component that drives these systems. Nonetheless, empirical studies thus far have been limited to the investigation of individual securities, often neglecting joint dependencies and the multi-asset case. Financial markets and, ultimately, human interactions are complex systems that need to be handled as such to explain financial dynamics in a realistic way.

As one of the first studies to model the collective behavior of news sentiment with a network approach, the authors in [88] defined a financial (network) community in order to model multi-variate structure of stock market and its relationship with on-line boards collective opinion. The authors use on-line message boards/forums to uncover the number and structure of these communities and investigate the empirical relationship of these financial communities with return co-variation patterns amongst stocks in the U.S. market. They found that (i) the greater the connectedness in a financial community, the greater the covariance of returns within the community; (ii) highly connected stocks, on average, have lower return variance and higher mean returns; (iii) stocks with high centrality scores tend to have greater average covariance with other stocks than those with low scores.

Further, a model [89, 90] has been proposed to evaluate the impact of news on portfolio return through a *Corporate Network*. The network is created computing the co-occurrence of company names in blog posts. Nodes represent companies whereas edges represent their co-occurrence frequencies. The authors perform Granger-causality tests between centrality measures and financial variables. They found that the average eigenvector centrality of companies in the corporate news network has a two-way Granger-causality on return and volatility with the STOXX 50 index.

In Chapter 5, we model the collective behavior of market prices as a financial network, where nodes represent stocks and edges measure the co-movement of asset prices returns. In an analogous manner, we measure the structure of social opinion as the co-movement of social media opinion on those same assets. We demonstrate that future market correlation structure can be predicted with high out-of-sample accuracy using a multiplex network approach that combines information from social media and financial data. In agreement with the results obtained in Chapter 4, we observe that social media is the dominant information source, indicating that the information provided by social media contributes more to predict stock market structure than stock returns contribute to the prediction of social opinion structure on the same assets.

Chapter 4

Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Sector

Social media and news analytics bring a new possibility to quantify and statistically validate hypotheses in financial dynamics in an ex-ante fashion. In this way, sentiment, emotions, and behavioral analytics can provide valuable information before the market moves [19, 13, 20]. Here, we take advantage of a unique dataset of social media and news analytics to investigate the interplay between market sentiment and stocks returns. We ask whether social media sentiment from retail brands has significant causal links with respect to stock returns. Results reveal a dependency structure between sentiment and financial variables for both traditional newswires and social media. Surprisingly, Twitter's sentiment of selected retail brands exhibited a relatively stronger relationship with stock returns than does that of traditional newswires. Results suggest that social media analytics play an important role in explaining the price dynamics of the retail sector.

4.1 Introduction

In [35] we developed a new methodology to derive sentiment analytics from Twitter messages which we later used to predictive sales of retail brands [36]. In this Chapter, we leverage the data available from this previous work [35] to investigate the interplay of Twitter sentiment extracted from listed retail brands with stock returns and volatility. We also compare results with a corresponding analysis that uses sentiment from traditional newswires. We consider volatility and log-returns as financially endogenous variables, and we take Twitter sentiment and volume as an exogenous explanatory variable. We also consider traditional newswires as datasources for comparative purposes.

The results presented in this Chapter suggest that social media sentiment analytics can be a complementary proxy of market's sentiment compared to news in the analysis of financial dynamics for the retail brands analyzed. Surprisingly, Twitter's sentiment presented a relatively stronger relationship with the stock returns compared to traditional newswires. Results suggest that social media can be a relevant source to explain stock price dynamics in the retail sector.

4.2 Dataset

We conducted our analysis on a subset of listed retail brands with stocks traded in the U.S. stock market, which we monitored from November 01, 2013 to September 30, 2014. The examined stocks and their Reuters Instrument Codes (RIC) are as follows: ABERCROMBIE & FITCH CO. (ANF.N), NIKE INC. (NKE.N), HOME DEPOT INC. (HD.N), MATTEL INC. (MAT.N) and GAMESTOP CORP. (GME.N). As discussed in Section 2.1.2, the choice of companies was given by data availability from previous work [35].

Given the companies selected, we consider three streams of time series data: (i) market data, which is given at the daily stock price; (ii) news meta-data (see Section 2.1.1); and (iii) social media sentiment (see Section 2.1.2).

4.3 Financial and Sentiment Variables

In this section, we define both the financial variables derived from market data and the sentiment variables extracted from news and social media.

Let $P(t)$ be the closing price of an asset at day t and $R(t) = \log P(t) - \log P(t-1)$ its daily log-return. We consider the excess of log-return of the asset over the return of the market benchmark $\widehat{R}$ as follows:

$$ER(t) = R(t) - \widehat{R}(t), \tag{4.1}$$

where we consider the S&P 500 index as the market benchmark. As a proxy of the daily volatility of a stock we define:

$$VOL(t) = 2\frac{P_{high}(t) - P_{low}(t)}{P_{high}(t) + P_{low}(t)} \in [-1,1], \tag{4.2}$$

where $P_{high}(t)$ and $P_{low}(t)$ are the highest and the lowest price of the stock at day t, respectively.

In terms of Twitter analytics, we count the number of positive ($G(t)$) and negative ($B(t)$) English messages at day t which mention a given company, and we define the following variables [12]:

$$S_A(t) = G(t) - B(t), \quad S_R(t) = \frac{G(t) - B(t)}{G(t) + B(t)} \in [-1,1] \tag{4.3}$$

as the absolute and relative sentiments of that company on a given day, respectively. Notice that $S_R(t_0) = +1$, represents a day t_0 with the highest positive sentiment for the company considered; conversely $S_R(t_0) = -1$ indicates the highest negative sentiment, whereas we consider neutrality when $S_R(t_0) = 0$. We also define $V(t)$ as the total number of stories observed at day t regardless of their language.

For the news analytics, we consider the event sentiment score of each news story. This score ranges between 0 and 100. High values indicate more positive sentiment while values below 50 represent negative sentiment. We then normalize this score so that it ranges between -1 and 1, and we consider its daily mean as the relative sentiment for news $S_R(t) \in [-1,1]$. We label a news story as positive if $S_R(t) > 0$ and as negative if $S_R(t) < 0$. We then count the daily number of positive $G(t)$ and negative $B(t)$ stories per company to obtain the relative news sentiment score defined as $S_A(t) = G(t) - B(t)$. The news volume $V(t)$ of a given company is defined as the total number of news stories observed at day t that are related to the company, where a company is assigned to a news story if the news storys relevance score relative to that company is equal to 100. It is important to mention that the news dataset considered here contains only English stories whereas the Twitter data has no such limitation.

Fig. 4.1 shows a sample of the variables calculated from Twitter for Home-Depot Inc. Furthermore, Fig. 4.2 shows the distribution of values of the relative sentiment obtained from Twitter and news. We observe that both Twitter and news present skewed distributions; news has a more neutral-centered distribution than Twitter. It is important to note that the sentiment provided by the Twitter analytics presents a distinct proxy for sentiment compared to news, as each company analyzed depicts different positive/negative sentiment tones: E.g., NIKEs Twitter sentiment is highly positive while the news sentiment exhibits a mean around a neutral point.

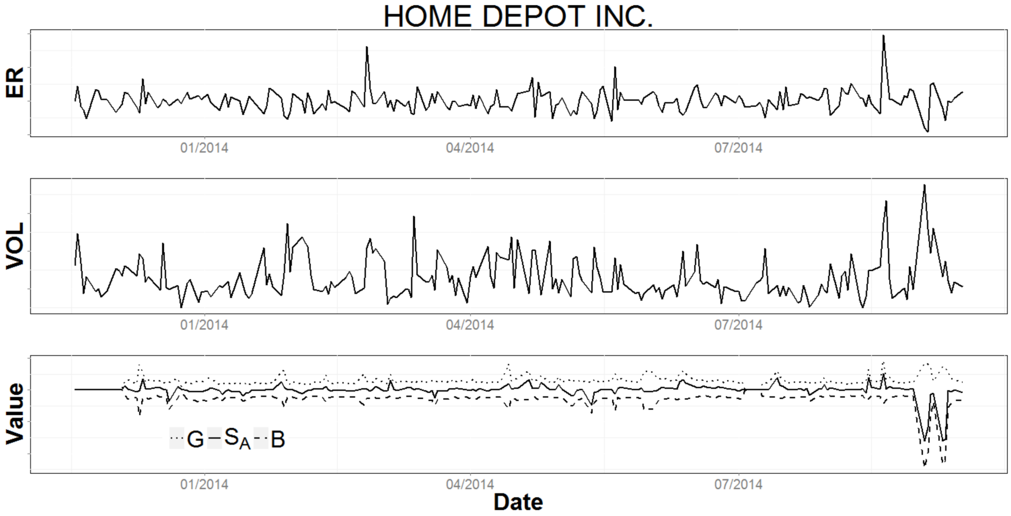


Figure 4.1: Descriptive Analysis for the company Home-Depot Inc. Variables: Excess of log-return ER, Volatility VOL, S_A, G and B.

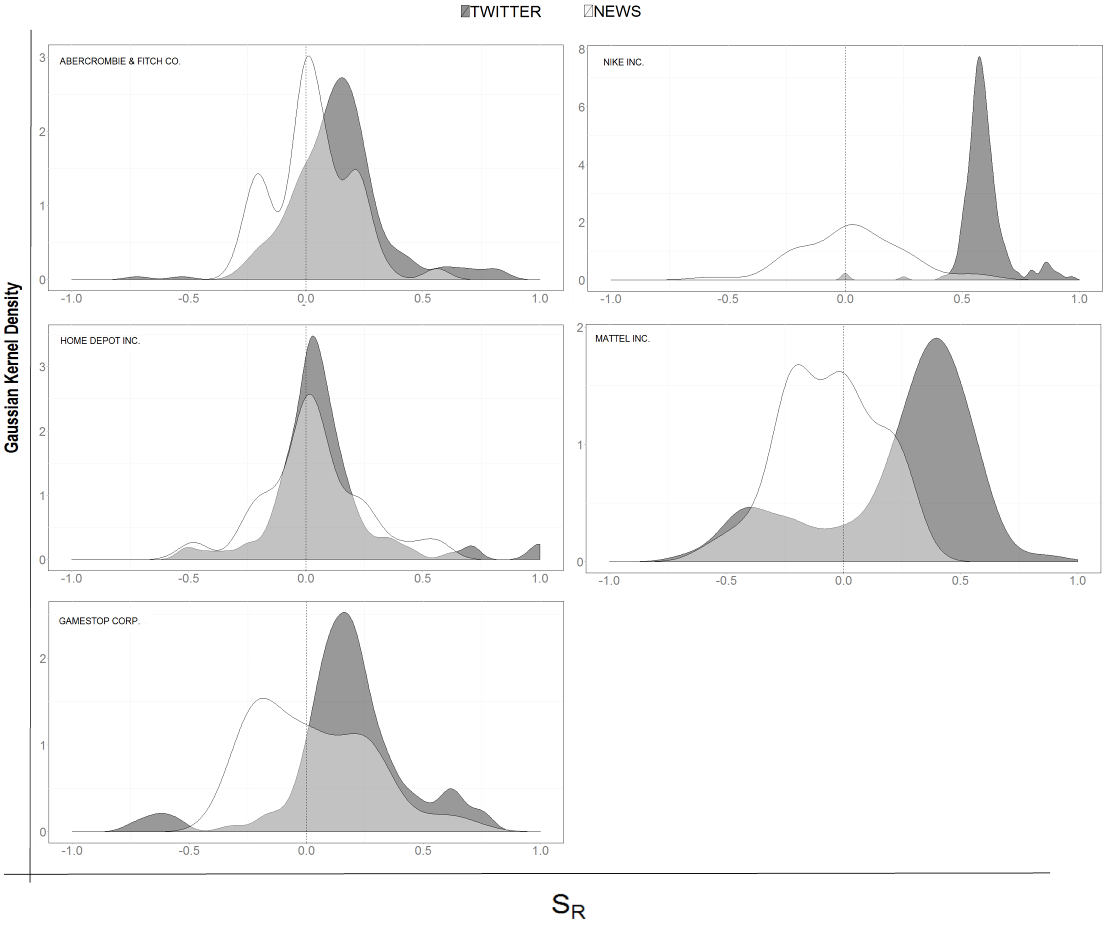


Figure 4.2: Twitter and news sentiment present skewed sentiment distributions. Distribution of relative sentiment $S_R(t)$ from Twitter and news for the following companies: ABERCROMBIE & FITCH CO. (ANF.N), GAMESTOP CORP. (GME.N)., HOME DEPOT INC. (HD.N), MATTEL INC. (MAT.N) and NIKE INC. (NKE.N)

4.4 Methods

4.4.1 Granger Causality

We investigate the statistical causality of social media and news sentiment on the financial variables analyzed, i.e., stock returns and volatility. For this purpose, we utilize Granger-causality as a concept of cause-effect dependence (see Section 2.2.1).

We test the Granger-causality of the excess of log-return ER and the number of positive stories G, the number of negative stories B, the relative sentiment S_R and the absolute sentiment S_A. For the volatility VOL of a given stock, we will also consider the total volume of stories V in addition to previously mentioned variables. Furthermore, we will perform the Granger-causality test over the normally standardized versions of the time series analyzed such as they have zero mean and standard deviation 1.

To visualize the Granger-causality results, we created a Granger-causality graph $G = [V, E]$, where V is a node set and E is an edge set. A node $u \in V$ represents a variable in the causality test and an edge $e = (u, v)$ indicates that u Granger-causes v within a pre-defined significance level. Furthermore, we define $p(e)$ as an attribute of the edge. If C is the set of companies in which we see causality between u and v, then we set $p(e) = C$. Fig. 4.3 shows an example of a Granger-causality graph that indicates that u Granger-causes v for the set of companies C.

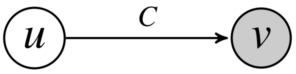


Figure 4.3: Granger-causality graph. The variable u Granger-causes the variable v for the set of companies C.

4.4.2 Predictive Analysis

To evaluate the predictive power of sentiment, we consider two auto-regressive models with and without sentiment. We then conduct a one-step-ahead prediction analysis:

$$\mathscr{M}_0 : X(t) = \alpha + \sum_{\tau=1}^{k} \beta_\tau X(t-\tau) + \varepsilon_t, \tag{4.4}$$

$$\mathscr{M}_1 : X(t) = \alpha + \sum_{\tau=1}^{k} \beta_\tau X(t-\tau) + \sum_{i=1}^{k} \gamma_i Y(t-\tau) + \widehat{\varepsilon}_t \tag{4.5}$$

where,

$$X(t) \in \{ER(t), VOL(t)\}, \tag{4.6}$$

$$Y(t) \in \{G(t), B(t), S_R(t), V(t)\}. \tag{4.7}$$

As $S_A(t)$ is a linear combination of $G(t)$ and $B(t)$, we will not consider it in the linear regression for any dataset. $G(t)$ and $B(t)$ are already considered in the model. Moreover, we consider only one day of lag for the sentiment variables and a lag of two days for the financial variables[1]. Again, we will consider the normally standardized versions of the time series analyzed.

Hence, we will consider the following regression model for the excess of log-return prediction:

$$\mathcal{M}_0 : ER(t) = \alpha + \beta_1 ER(t-1) + \beta_2 ER(t-2) + \varepsilon_t, \tag{4.8}$$

$$\begin{aligned} \mathcal{M}_1 : ER(t) = {} & \alpha + \beta_1 ER(t-1) + \beta_2 ER(t-2) \\ & + \gamma_1 G(t-1) + \gamma_2 B(t-1) + \gamma_3 S_R(t-1) + \widehat{\varepsilon}_t \end{aligned} \tag{4.9}$$

For the volatility prediction using news as a data source, we will not include the volume time series $V(t)$ as an explanatory variable in the regression because of its high correlation with the amount of positive and negative news already considered by the model. Notice that, for the Twitter case, the volume time series considers also non-English messages, which are not considered by the time series given by $G(t)$ and $B(t)$. Therefore, we keep $V(t)$ as an explanatory variable in the Twitter model. As a result, we have the following for news:

$$\mathcal{M}_0 : VOL(t) = \alpha + \beta_1 VOL(t-1) + \beta_2 VOL(t-2) + \varepsilon_t, \tag{4.10}$$

$$\begin{aligned} \mathcal{M}_1 : VOL(t) = {} & \alpha + \beta_1 VOL(t-1) + \beta_2 VOL(t-2) \\ & + \gamma_1 G(t-1) + \gamma_2 B(t-1) + \widehat{\varepsilon}_t \end{aligned} \tag{4.11}$$

and the corresponding model for Twitter:

$$\mathcal{M}_0 : VOL(t) = \alpha + \beta_1 VOL(t-1) + \beta_2 VOL(t-2) + \varepsilon_t, \tag{4.12}$$

$$\begin{aligned} \mathcal{M}_1 : VOL(t) = {} & \alpha + \beta_1 VOL(t-1) + \beta_2 VOL(t-2) \\ & + \gamma_1 G(t-1) + \gamma_2 B(t-1) + \gamma_3 V(t-1) + \widehat{\varepsilon}_t. \end{aligned} \tag{4.13}$$

Forecasting accuracy is measured by comparing the two residuals ε_t and $\widehat{\varepsilon}_t$ in

[1] A model-selection approach can also be used to find an optimal lag for the explanatory variables. Examples of selection criteria include the following: the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and Mallow's Cp. See [91].

terms of residual standard error:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{T} (y_i - \widehat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{T} \hat{\varepsilon}_i^2}{n}} \tag{4.14}$$

where T is the total number of points, n is the number of degrees of freedom of the model, $\widehat{y}_i$ is the predicted value and y_i is the observed one.

4.5 Results and Discussion

We present the results from the Granger-causality tests and the predictive analysis of the financial variables and sentiment data from Twitter and news. The sentiment predictive power and its Granger-causality tests are fulfilled in a one-step-ahead fashion. We investigate the statistical significance of the sentiment variables with respect to movements in returns and volatility, and we compare the Twitter results with news. We provide empirical evidence that Twitter is moving the market in respect to the excess of log-returns for a subset of stocks. Also, Twitter presents a stronger relationship with stock returns than with news for the selected retail companies. On the other hand, Twitter sentiment analytics showed a weaker relationship with volatility compared to news.

4.5.1 Excess of Log-Returns

We analyze the dynamics of the excess of log-returns of the stocks considered in relation to absolute and relative sentiments and with the number of positive and negative stories.

Fig. 4.4 shows a Granger-causality graph that summarizes the significant Granger-causalities (p-value < 0.05) between the excess of log-return and the sentiment variables for both news and Twitter. See Table 4.2 for detailed results. We observe that Twitter's sentiment analytics presents more significant causal relationships than news. Twitter's relative sentiment and its number of positive messages Granger-cause log-returns for GAMESTOP CORP. and MATTEL INC. Twitter's absolute sentiment also Granger-causes log-returns for MATTEL INC. with a two-way significant (p-value < 0.01) Granger-causality for HOME DEPOT INC. The number of negative stories alone has no significant relationship with returns; however, combined with the number of positive stories in the form of relative and absolute sentiment, it constitutes an important measure. The news analytics exhibit only one significant relationship, which is observed in the number of positive news Granger-causing the excess of log-returns for GAMESTOP CORP.

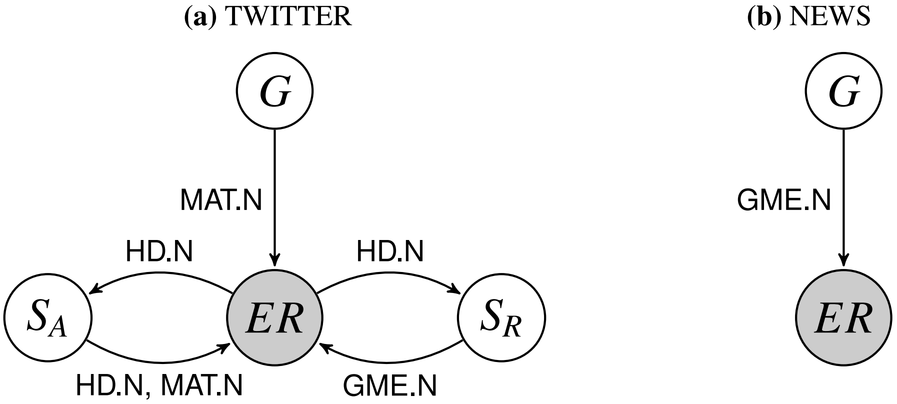


Figure 4.4: Granger-causality graph for (a) Twitter and (b) news. It shows significant causal relationships in the Granger-causality test between excess of log-returns (ER) and the sentiment analytics: number of positive stories (G), number of negative stories (B), absolute sentiment (S_A) and relative sentiment S_R. Sentiment variables that present no significant causal links are not shown in the graph.

The solution of the multiple regression analysis presented in Table 4.3 agrees with the Granger-causality tests, as it shows Twitter to have a larger number of significant sentiment coefficients than news. MATTEL INC. particularly presents all sentiment coefficients with high significance (p-value < 0.01), thus suggesting that the Twitter sentiment analytics is indeed relevant to the prediction of the next-day excess of log-return. The companies HOME DEPOT INC. and GAMESTOP CORP. have also presented significant sentiment coefficients. For the news analytics, the sentiment was significant only for GAMESTOP CORP. Furthermore, analysis of the residual standard error (RES) of the models with and without sentiment variables in Table 4.1 shows that use of the Twitter sentiment variables reduced the error of the model with market data only for MATTEL INC., HOME DEPOT INC. and GAMESTOP CORP. while the news sentiment improved the prediction only for the company GAMESTOP CORP.

Table 4.1: Difference between the residual standard error of the model that considered market data only and the model that considered the sentiment variables $G(t)$, and $B(t)$ in the prediction of excess of log-return $ER(t)$

	Error Reduction (%)	
Company	NEWS	TWITTER
NIKE INC.	-2.41	-0.58
ABERCROMBIE & FITCH CO.	-1.26	-0.60
HOME DEPOT INC.	-0.99	1.23
MATTEL INC.	-0.48	2.82
GAMESTOP CORP.	8.34	1.10

Table 4.2: Statistical significance (p-values) of Granger-causality analysis between excess log-return and $S_A(t), S_R(t), G(t)$ and $B(t)$ for the companies: ABERCROMBIE & FITCH CO. (ANF.N), NIKE INC. (NKE.N), HOME DEPOT INC. (HD.N), MATTEL INC. (MAT.N) and GAMESTOP CORP. (GME.N). Table shows the regression coefficients by fitting the model in Eq. 4.9.

	TWITTER ANALYTICS					NEWS ANALYTICS				
	HD.N	MAT.N	GME.N	NKE.N	ANF.N	HD.N	MAT.N	GME.N	NKE.N	ANF.N
$S_A \to ER$	0.003***	0.046**	0.140	0.888	0.477	0.303	0.411	0.140	0.621	0.707
$ER \to S_A$	0.006***	0.404	0.231	0.354	0.937	0.423	0.451	0.230	0.546	0.281
$S_R \to ER$	0.449	0.497	0.032**	0.680	0.591	0.747	0.977	0.696	0.816	0.814
$ER \to S_R$	0.024**	0.855	0.196	0.995	0.875	0.942	0.314	0.162	0.564	0.213
$G \to ER$	0.182	0.016**	0.885	0.400	0.685	0.203	0.228	0.014**	0.304	0.231
$ER \to G$	0.050*	0.305	0.957	0.380	0.197	0.388	0.382	0.171	0.199	0.518
$B \to ER$	0.327	0.559	0.267	0.344	0.763	0.681	0.976	0.920	0.796	0.398
$ER \to B$	0.219	0.792	0.538	0.166	0.480	0.855	0.646	0.894	0.769	0.863

Significance codes: p-value < 0.01: ***, p-value < 0.05: **, p-value < 0.1: *

Table 4.3: Summary statistics of multiple regression. Prediction of excess of log-return (ER) using sentiment (S_R, G, B) for the companies: ABERCROMBIE & FITCH CO. (ANF.N), NIKE INC. (NKE.N), HOME DEPOT INC. (HD.N), MATTEL INC. (MAT.N) and GAMESTOP CORP. (GME.N).

	TWITTER ANALYTICS					NEWS ANALYTICS				
	HD.N	MAT.N	GME.N	NKE.N	ANF.N	HD.N	MAT.N	GME.N	NKE.N	ANF.N
$ER_{(t-1)}$	4.414018e-04	1.192513e-01*	-1.704137e-01**	-3.001408e-02	1.203011e-02	2.592302e-01	6.129545e-01***	-9.060641e-02	4.018227e-02	-1.457410e-01
$ER_{(t-2)}$	-9.806531e-02	9.192829e-03	1.535305e-02	9.752118e-02	-4.741730e-02	-5.695452e-02	-2.674834e-01*	4.458446e-01***	4.363861e-02	-6.310047e-02
$SR_{(t-1)}$	-2.258381e-02	3.718939e-01***	1.821209e-01**	1.817089e-02	2.558886e-02	-1.092573e-01	-8.211806e-02	-2.852138e-01	4.487921e-02	9.364799e-02
$G_{(t-1)}$	2.536661e-01***	-4.286924e-01***	-1.229142e-02	-3.956882e-02	4.445273e-02	2.600786e-01	2.683852e-01	4.962514e-01***	-2.144345e-01	-2.279505e-01
$B_{(t-1)}$	-2.612111e-01**	4.801024e-01***	-1.159965e-01	-1.459096e-02	-3.347246e-02	-9.858689e-02	3.625886e-03	-5.315150e-02	-6.179474e-02	3.457372e-02

Significance codes: p-value < 0.01: ***, p-value < 0.05: **, p-value < 0.1: *

4.5.2 Volatility

In this section, we analyze the interplay of Twitter, news volumes and sentiment with stock-returns volatility. As volume measures we consider the following: the number of positive G and negative B English stories and the total volume V of stories regardless of its language. As daily sentiment analytics, we consider the absolute sentiment S_A and the relative sentiment S_R.

Fig. 4.5 shows the significant causal relationships (p-value < 0.05) of the Granger-causality test between the volatility and the sentiment variables. See Table 4.5 for detailed results. Overall, there are more significant causal links for the news sentiment analytics than for Twitter in respect to volatility. We observe that the number of positive stories and the total volume both Granger-cause volatility for Twitter and for news but more companies are affected for news. The absolute sentiment Granger-causes volatility only for news, in the case of ABERCROMBIE & FITCH CO. (ANF.N). The relative sentiment and the number of negative stories do not Granger-cause volatility; on the other hand, volatility Granger-causes negative news for the company GAMESTOP CORP. (GME.N).

The solution of the multiple regression analysis in Table 4.6 shows that the number of positive stories is a significant variable for both news and Twitter. It is more significant for the former than for the latter. The number of negative stories exhibits no relevance in either regression. The total volume of Twitter messages is relevant only for NIKE (NKE.N). Moreover, analysis of the residual standard error of the models with and without the sentiment variables in Table 4.4 shows that both Twitter and news are able to reduce the error in prediction for a subset of the companies. In cases where the model was improved with sentiment, news provides a higher reduction of error than Twitter.

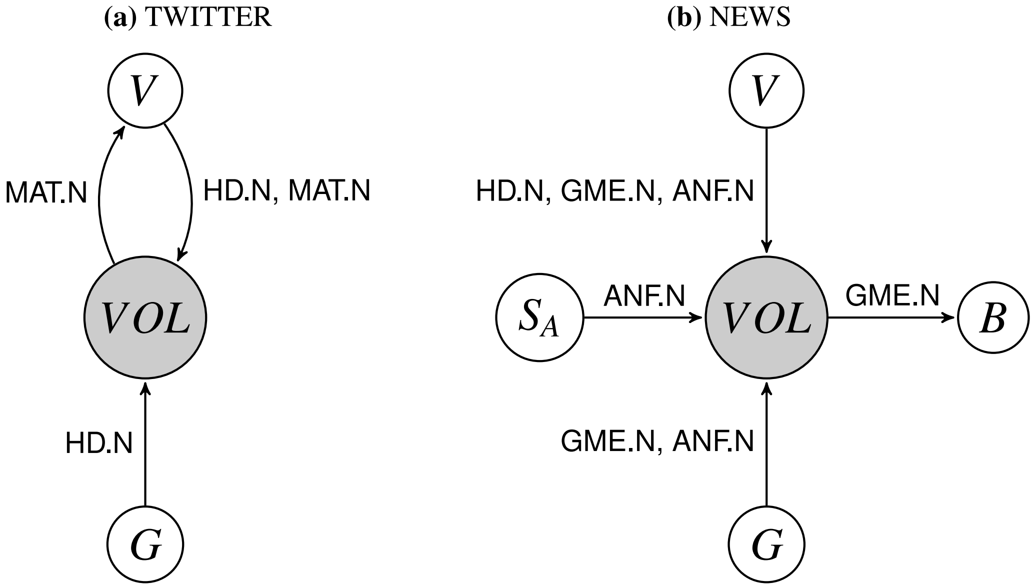


Figure 4.5: Granger-causality graph for (a) Twitter and (b) news. Figure shows the significant causal relationships in the Granger-causality test between volatility (VOL) and sentiment analytics: total number of stories (V), number of positive stories (G), number of positive stories (B), absolute sentiment (S_A) and relative sentiment (S_R). Sentiment variables that present no significant causal links are not shown in the graph.

Table 4.4: Difference between the residual standard error of the model that considered market data only and the model that considered the sentiment variables $G(t)$, and $B(t)$ in the prediction of volatility $VOL(t)$.

	Error Reduction (%)	
Company	NEWS	TWITTER
NIKE INC.	1.36	1.08
ABERCROMBIE & FITCH CO.	4.03	-0.52
HOME DEPOT INC.	2.46	1.10
MATTEL INC.	-2.21	-0.36
GAMESTOP CORP.	14.99	0.20

Table 4.5: Statistical significance (p-values) of Granger-causality analysis between volatility ($VOL(t)$) and $G(t), B(t)$, and $V(t)$ for ABERCROMBIE & FITCH CO. (ANF.N), NIKE INC. (NKE.N), HOME DEPOT INC. (HD.N), MATTEL INC. (MAT.N) and GAMESTOP CORP. (GME.N).

	TWITTER ANALYTICS					NEWS ANALYTICS				
	HD.N	MAT.N	GME.N	NKE.N	ANF.N	HD.N	MAT.N	GME.N	NKE.N	ANF.N
$V \rightarrow VOL$	0.025**	0.004***	0.751	0.976	0.729	0.029**	0.320	0.020**	0.846	0.001***
$VOL \rightarrow V$	0.228	0.016**	0.980	0.611	0.924	0.307	0.453	0.053*	0.462	0.560
$G \rightarrow VOL$	0.004***	0.072*	0.678	0.397	0.228	0.060*	0.341	0.031**	0.961	<0.001***
$VOL \rightarrow G$	0.560	0.273	0.668	0.330	0.690	0.146	0.390	0.211	0.859	0.615
$B \rightarrow VOL$	0.053*	0.118	0.720	0.636	0.884	0.522	0.301	0.204	0.526	0.540
$VOL \rightarrow B$	0.420	0.477	0.363	0.729	0.967	0.484	0.620	0.026**	0.391	0.801
$S_R \rightarrow VOL$	0.539	0.305	0.489	0.786	0.220	0.416	0.925	0.510	0.441	0.056*
$VOL \rightarrow S_R$	0.944	0.867	0.791	0.611	0.623	0.854	0.352	0.590	0.184	0.340
$S_A \rightarrow VOL$	0.274	0.736	0.437	0.142	0.216	0.435	0.194	0.497	0.652	0.010**
$VOL \rightarrow S_A$	0.458	0.900	0.747	0.120	0.950	0.173	0.352	0.187	0.707	0.996

Significance codes: p-value < 0.01: ***, p-value < 0.05: **, p-value < 0.1: *

Table 4.6: Summary statistics of multiple regression. Prediction of volatility (VOL) using sentiment (G,B) and volume (V) for ABERCROMBIE & FITCH CO. (ANF.N), NIKE INC. (NKE.N), HOME DEPOT INC. (HD.N), MATTEL INC. (MAT.N), and GAMESTOP CORP. (GME.N). Table shows the regression coefficients by fitting the models in Eq. 4.13 and Eq. 4.11 for Twitter and news, respectively.

	TWITTER ANALYTICS					NEWS ANALYTICS				
	HD.N	MAT.N	GME.N	NKE.N	ANF.N	HD.N	MAT.N	GME.N	NKE.N	ANF.N
$VOL_{(t-1)}$	2.305851e-01***	2.211374e-01***	1.019004e-01	1.893666e-01***	1.308395e-01*	3.891e-01*	6.265e-02	9.587e-02	3.785e-01**	3.792e-01**
$VOL_{(t-2)}$	1.250733e-01*	1.238580e-01*	-6.173133e-02	4.307688e-02	8.735957e-02	-9.120e-02	-4.451e-03	5.092e-02	3.093e-02	1.054e-01
$G_{(t-1)}$	1.845026e-01*	1.073941e-01	-2.807417e-01*	-1.586209e-03	3.533443e-02	3.018e-01**	2.772e-03	6.046e-01***	1.274e-01	-3.329e-01*
$B_{(t-1)}$	1.420223e-03	-7.759929e-02	4.461084e-02	-2.181354e-03	-8.569212e-02	9.294e-02	-1.105e-01	1.184e-01	-1.400e-01	-1.300e-01
$V_{(t-1)}$	-1.468387e-03	-1.862391e-02	1.920558e-01	1.901394e-01**	5.742508e-02	N/A	N/A	N/A	N/A	N/A

Significance codes: p-value < 0.01: ***, p-value < 0.05: **, p-value < 0.1: *

4.6 Conclusion

We showed that measures of the Twitter sentiment extracted from listed retail brands exhibit a significant causal relatioship with stock returns and volatility. While analyzing the interplay of the excess of log-return and the Twitter sentiment variables, we concluded that, (i) Twitter presents a stronger Granger-causality than news with respect to the stock returns compared, (ii) positive tweets and Twitter's sentiment Granger-cause an excess of stock returns for a subset of companies, and (iii) Twitter's sentiment analytics is indeed relevant to the prediction of the next-day excess of log-return - even when compared to traditional newswires.

Moreover, in the volatility analysis we found that, (i) Twitter's analytics reveals a weaker relationship with volatility than the one observed with returns, (ii) number of positive tweets and total volume both Granger-cause volatility for some companies but present reduced Granger-causality compared to news, and (iii) the number of positive tweets is a significant variable for the one-step-ahead prediction of volatility while the number of negative messages shows no relevance.

It is also important to remark that the asset and content universes of the present work are limited. Social media data are often sparse and of difficult acquisition. The results presented in this Chapter were limited to 5 retail companies that the authors had access to social media sentiment data within a 2-year period. J. Manfield, D. Lukacsko and T. T. P. Souza [38] later reported a broader study, which results suggested that social media can be informative of financial dynamics for a large set of stocks across 500 stocks from both NYSE and NASDAQ exchanges. In the next Chapter, we expand the universe of companies considered and we show that DJIA index constituents have a significant causal relationship with financial returns when nonlinear dynamics are considered.

Chapter 5

A Nonlinear Impact: Evidence of Causal Effects of Social Media On Market Prices

In this Chapter, we expand the asset universe of retail companies studied in the Chapter 4 to a broader set of companies representative of the US stock market by analyzing social media messages related to the DJIA index's constituents. We also leverage a non-parametric framework, instead of assuming linear coupling as previously studied in the Chapter 4 to evaluate to what extent the assumption of linear coupling affects the quantification of causality between social media sentiment and stocks' returns. Two main conclusions are drawn. First, social media's significant causal relationship on stocks' returns is purely nonlinear in most cases. Second, social media dominates the directional coupling with the stock market, which is an effect that is not observable when using linear modeling. The results also serve as empirical guidance on models' adequacy in the investigation of social and financial systems.

5.1 Introduction

Recent research provides evidence that the exogenous information gathered from on-line social systems may be useful to describe financial dynamics [27, 19] . However, to date, there are mixed results on the capability to predict stock returns with social media analytics. On the one hand, some evidence is provided in favor of the predictability of price movements using news and social media [2, 11, 6, 26]. On the other hand, other studies report weak results [27, 19] suggesting that social media analytics have low predictive power when used alone. Moreover, the use of ad hoc functional forms and assumptions in different studies makes it difficult to draw general conclusions regarding the nature of the relationship between collective opinions and stock markets.

In this Chapter, we test Granger-causality of social media sentiment on stock returns using linear and non-linear frameworks. We analyze an extensive dataset comprising both the time series of Twitter sentiment and the stock market returns related to the stocks components of the DJIA index. By comparing the results from the linear and nonlinear frameworks, we detected interactions that are purely nonlinear. We also estimate the information flow between these two systems further providing useful information regarding which system is leading the other or whether a bidirectional coupling is observed.

To the best of our knowledge, our results provide the first empirical evidence that suggests that social media and stock markets not only have a significant causal relationship but also this relationship is dominated by nonlinear interactions. Furthermore, we highlight that common approaches, which assume linear interactions, can hide significant information that is revealed in this study under a nonlinear framework. In particular, we show that the evidence that net information is flowing from social media (Twitter) to markets is revealed only by using the nonlinear approach. Our findings call into question certain current modeling and analytics that assume linearity.

5.2 Data

Our analysis was conducted on the 30 components of the Dow Jones Industrial Average (DJIA) index as of March 31, 2012, which we monitored over the 500 trading days during the two-year period from March 31, 2012 to March 31, 2014. The choice of these stocks was made due to their representativeness for the stock market (see Table 2.4 for the list of stock tickers and Appendix A.1 for the complete list of selected companies). We consider two streams of time series data: (i) daily stock prices, and (ii) social media sentiment analytics based on 1,767,997 Twitter messages.

We considered the closing price $P_i(t)$ of stock $i = 1...30$ on day $t = 1...500$. The financial variables that we considered were the stocks' daily log-returns, which were defined as $R_i(t) = \log P_i(t) - \log P_i(t-1)$. These financial variables were compared with the Twitter data measured during the same period of time.

We considered Twitter sentiment data [35] as a proxy for the collective opinion regarding a stock (see Section 2.1.3). In our analysis, we take the daily total number of bullish tweets related to a company as the social media time series $SM(t)$. This measure is indicative of how positive the messages are toward the mentioned stock.

5.3 Results and Discussion

5.3.1 Social Media and Stocks Returns: Linear and Nonlinear Causality

We test the null hypothesis that social media does not cause stocks returns. First, we test this hypothesis with a standard G-causality test under a linear vector-autoregressive framework. Second, we test this hypothesis with a nonlinear, non-parametric Transfer Entropy approach. We interpret the Transfer Entropy as the information flow between the social media opinion and future outcomes of stocks returns at lag Δt controlled by the current information on stocks returns.

Fig. 5.1 shows the significant causal links between social media and stocks returns considering both cases: nonlinear (TE) and linear G-causality. With the linear analysis, we discover only three stocks with significant causality: INTEL CORP., NIKE INC. and WALT DISNEY CO. Conversely, with the nonlinear approach, in addition to the 3 stocks identified with significant causal linear relationship, we also discover 8 other stocks with purely nonlinear causal relationships.

The low level of causality obtained under linear constraints is in-line with some of the results found in the literature, where it has been shown that stocks returns

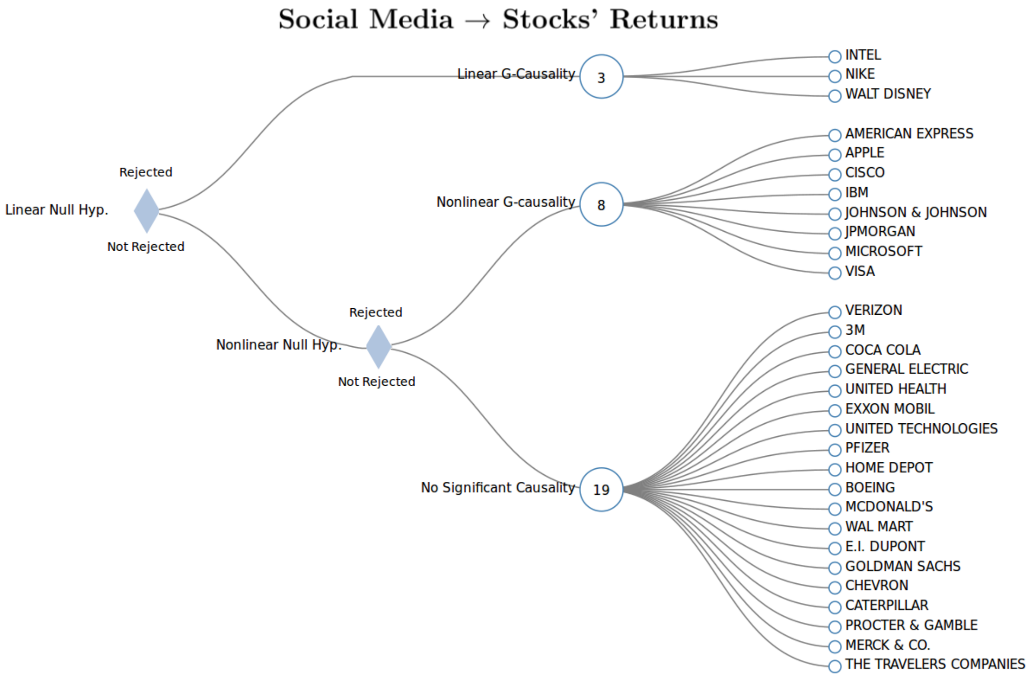


Figure 5.1: Demonstration that the causality between social media and stocks returns is mostly nonlinear. The linear causality test indicated that social media caused stocks returns only for 3 stocks. The nonlinear analysis showed that almost 1/3 of the stocks that were rejected in the linear case have significant nonlinear causality. In the nonlinear case, the Transfer Entropy was used to quantify the causal inference between the systems with randomized permutations test for the statistical significance estimation. In the linear case, a standard linear G-causality test was performed under a linear vector-autoregressive framework. A significant linear G-causality was accepted if its linear specification was not rejected by the BDS test. The p-values were adjusted with the Bonferroni correction. Significance is noted at p-values < 0.05.

have weak causality links [13, 92] with social media sentiment analytics resulting in small or no predictive power [27] and no significant lead-time information about stock's movements for the majority of the stocks [19]. Conversely, the results from the nonlinear analyses unveiled a much higher level of causality, thus indicating that linear constraints may be neglecting the nature of the relationship between social media and stock markets. We also analyze the number of stocks with significant causal relationships aggregated by lag of interaction. Social media causality on stocks return is mostly nonlinear in the next-day period. The causality between social media and next-day stocks returns presents a large difference when the linear and nonlinear cases are compared. From the linear G-causality there is a significant

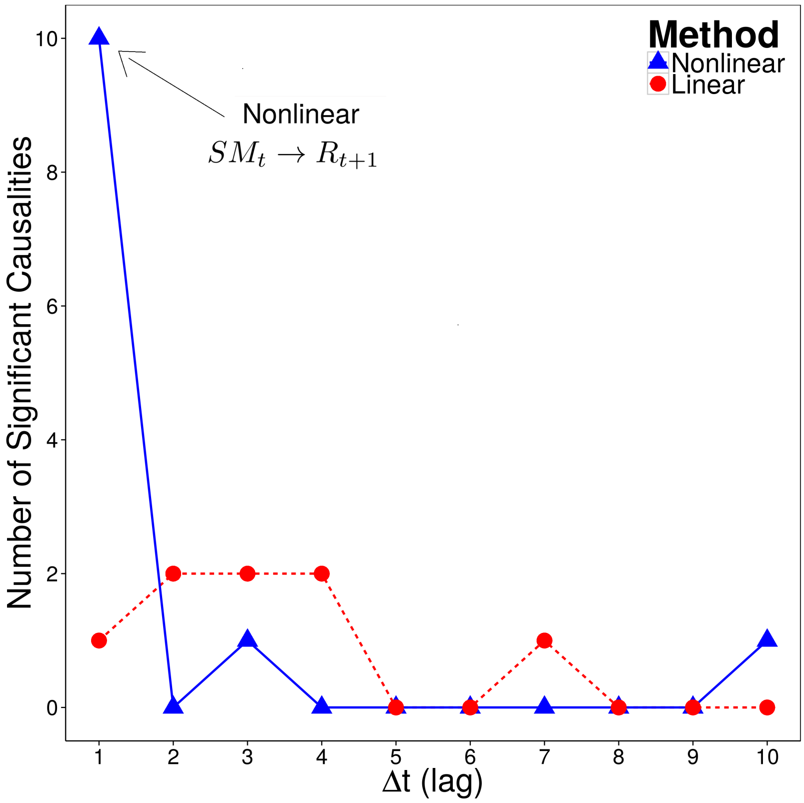


Figure 5.2: The social media causality on stocks return is mostly nonlinear in the next-day period. Figure shows the number of companies with significant causal relationships aggregated by lag. Nonlinear analysis identify the highest number of causal relationships in the first lag. Hence, linear-constraints may be neglecting social media causality over stocks returns, especially in the next-day period. Further lags present a lower number of significant causal relationships in both methods. Statistical significance is noted at p-values < 0.05.

causal relationship between social media and next-day stocks movements for one stock only. Conversely, nonlinear measures indicate that 10 companies have significant causal links in this direction. Higher delays show a drop on this number. These results suggest that linear constraints are neglecting social media causality over stocks returns especially in the short-term.

For the companies identified with nonlinear causality only, we tested whether the common functional forms and transformations that have been used in the literature can explain the observed nonlinearities. We checked the model's adequacy and causality significance for the various functional forms listed in Table 5.1 and reported the results. We observe that the linear functional form is adequate for 5 companies but none presented significant causal relationships. The second-order differencing $\nabla^2 x$ makes a linear functional adequate for the company VISA Inc. (V), but it turns out that Microsoft is mis-specified. GARCH and ARIMA filtering were tentatively applied to separate signal from noise and to linearize the original time series. Nonetheless, no significant causal relationships were observed. Other functional forms performed no better than the original linear specification a part from the absolute value transformation ($|x|$). It is indeed known that social media and news analytics predict absolute changes in market prices [19, 13] better than stocks returns. The absolute value of stock log-returns is a proxy for stock returns' volatility and therefore it has higher predictability than stock returns. However, half of the companies still had an unexplained nonlinear causality.

Table 5.1: Demonstration that the nonlinearities found are nontrivial. For the companies identified with nonlinear causality only, we tested whether common functional forms and transformations can explain the nonlinearities found. The test for mis-specification is performed using the BDS test. x represents the standard linear regression of returns on the social media time series. ∇x and $\nabla^2 x$ are, respectively, the first and second differencing taken in both time series. $f(x, vol)$ represents a linear regression of returns on social media controlled by the stocks returns' daily volatility. In the log-transformation we apply the function $\log(1+x)$ in both time series. The module $|x|$ is applied in the returns time-series, which is then regressed over the original social media data. The GARCH(1,1) and ARIMA(1,1,1) transformations were applied on the returns. Then, we regressed the resulting residuals on the original social media time series. See Appendix A.2 for the formal definition of the functional forms that were used.

Ticker	x (linear)	∇x	$\nabla^2 x$	$f(x, vol)$	$\log(1+x)$	$\|x\|$	GARCH(1,1)	ARIMA(1,1,1)
CSCO	○		○	○	○	●	○	○
MSFT	○			○	○	●	○	○
AXP	○		○	○	○	○	○	○
JPM	○		○	○	○	●	○	○
IBM	○		○	○	○		○	
V			○					
JNJ								
AAPL								

○: Not misspecified; ●: Not misspecified and with significant G-causality.

It is clear from the results that are reported in Table 5.1 that naive transformations, which are often applied to linearize nonlinear interactions, were unable to fully explain the observed causal relations. This indicates that the nonlinear causality observed is nontrivial and that there is a forecastable structure that cannot be explained by commonly used functional forms. Therefore, the impact of social media on market prices may be higher than what is currently reported in related studies since the frequently used functional forms are hiding significant causal relationships that are instead revealed here with a nonlinear analysis.

5.3.2 Quantifying the Direction of Information Flow

We quantified the Net Information Flow from social media to stocks returns using both the nonlinear and linear frameworks (see Section 2.2.7). We investigated which coupling direction is the strongest and to what extent the consideration of nonlinear dynamics affects the results compared to a linearly constrained analysis. Fig. 5.3 A) shows the results for the linear case. We observe an asymmetry of information, i.e., the systems are not coupled with the same amount of information flow in both directions. The stocks are clearly divided in two groups of approximately the same size. One group shows the stocks with a positive Net Information Flow, indicating that social media provides more predictive information about the stock market than the opposite. A second group of stocks indicates the opposite, i.e., information flows more from stocks returns to social media rather than in the other direction. In both cases, the absolute value of the Net Information Flow decreases with the lag.

Surprisingly, the consideration of nonlinear dynamics unveils a considerably different scenario. Fig. 5.3 B) shows the results of the same analysis without linear constraints, i.e., using the Transfer Entropy to estimate the information flow. We observe that the Net Information Flow becomes positive for all stocks that are analyzed. This result suggests that social media is the dominant information source and indicates that the information provided by social media contributes more to the description of the stock markets' dynamics than the opposite.

5.4 Conclusions

The main outcome of the present study is the evidence that social media, namely, Twitter sentiment, has a significant nonlinear causal relationship on stocks returns. This discovery is demonstrated by analyzing an extensive dataset comprising both the time series of Twitter sentiment and the stock market returns related to the stocks

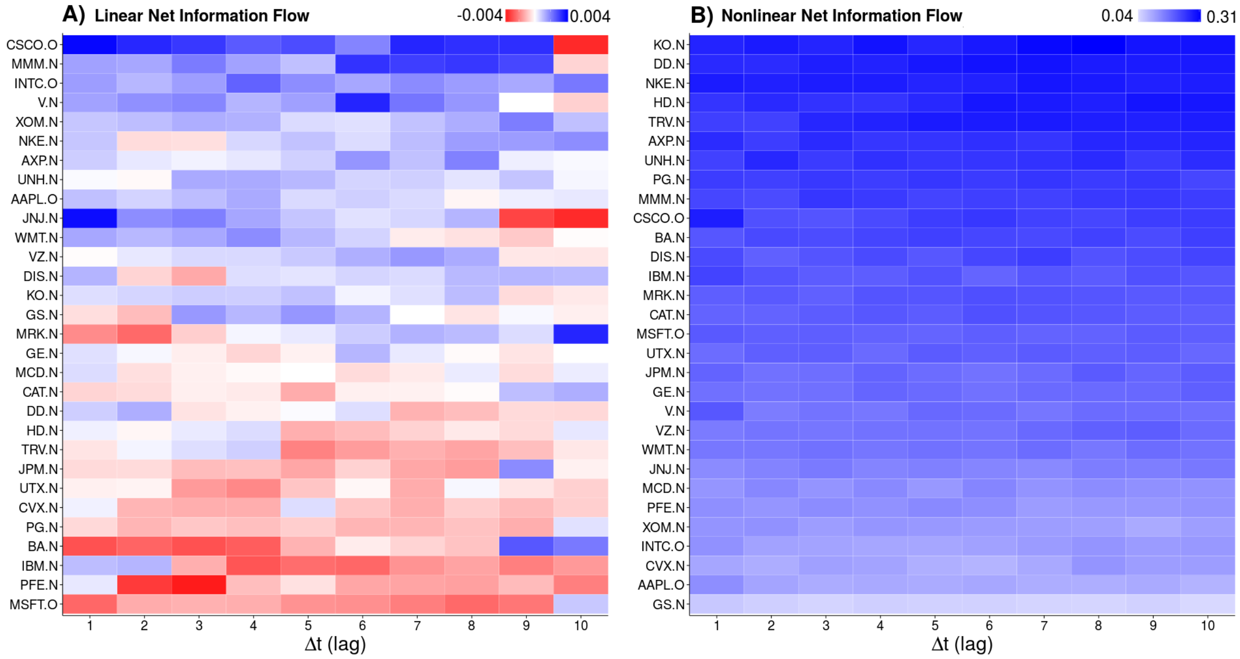


Figure 5.3: Evidence that the linear constraints greatly change the direction of Information Flow between social media and the stock market. The figure shows the Net Information Flow from social media to stocks returns: $\widehat{TE}_{SM\rightarrow R} = TE_{SM\rightarrow R} - TE_{R\rightarrow SM}$. In A), the Net Information Flow is estimated with linear constraints. Positive values (blue) indicate that $TE_{SM\rightarrow R} > TE_{R\rightarrow SM}$, thus providing evidence that information flows from social media to stock returns. Conversely, negative values (red) indicate that the stock market provides more information about social media movements than the opposite. In B), the estimation of the Net Information Flow considers nonlinear dynamics. Differently from the linear case, all companies show a positive information flow from social media to stocks' returns. This indicates that information flows predominantly from social media to the stock market with a nonlinear causality relation. We observe a change of the direction of the information flow in about half of the companies compared to the same analysis with linear constraints. Stocks on the Y-axis are ranked (top to bottom) by their total Net Information Flow considering all lags, i.e., $\sum_{\Delta t=1}^{10} \widehat{TE}_{SM\rightarrow R}$

components of the DJIA index. Linear and nonlinear tests for causality reveal three major empirical findings.

1. Social media and stocks' returns have a significant causal relation that is purely nonlinear in most cases. Specifically, we observed that the consideration of nonlinear causal relations increased the number of stocks with a relevant social media causal effect on stock price from 1/10, in the linear case, to more than 1/3 in the nonlinear case.

2. The functional forms of nonlinear interactions are non-trivial and cannot be explained by common functional forms used in the literature. This indicates that the impact of social media on stocks' returns may be higher than currently reported in related studies.

3. Net Information Flow analysis indicates that social media dominates the directional coupling with stock market, an effect not observable within linear modeling.

From a methodological point of view, the results indicate that a nonlinear approach is highly preferable for the investigation of causal relationships between social and financial systems. A better understanding of the nature of these nonlinear relations and assessing whether these nonlinear relations are common across different systems will be the focus of future research. The fact that social media is a valuable source of information regarding the future evolution of the stock market supports the idea that there is an emotional component that drives these systems beyond pure economic and fundamental factors. This predictive power of social media on future prices has started to be exploited by market players and the effects we have uncovered with this work could fade away in the future when speculation erodes arbitrage opportunities. However, its nonlinear nature might indicate that there are causal effects that might not be trivially incorporated in trading strategies.

Chapter 6

Predicting Future Stock Market Structure by Combining Social and Financial Network Information

In the previous Chapters, we demonstrated that social sentiment can inform the description of financial dynamics. Related research has focused in the analysis of individual stock returns. In this Chapter, instead, we test whether social sentiment can predict the entire market structure. We demonstrate that future market correlation structure can be predicted with high out-of-sample accuracy using a multiplex network approach that combines information from social media and financial data. Market structure is measured by quantifying the co-movement of asset prices returns, while social structure is measured as the co-movement of social media opinion on those same assets. Predictions are obtained with a simple model that uses link persistence and link formation by triadic closure across both financial and social media layers. Results demonstrate that the proposed model can predict future market structure with up to a 40% out-of-sample performance improvement compared to a benchmark model that assumes a time-invariant financial correlation structure. Social media information leads to improved models for all settings tested, particularly in the long-term prediction of financial market structure. Surprisingly, financial market structure exhibited a higher predictability than social opinion structure.

6.1 Introduction

Financial markets can be regarded as a complex network in which nodes represent different financial assets and edges represent one or many types of relationships among those assets. Filtered correlation-based networks have successfully been used in the literature to study financial markets structure particularly from observational data derived from empirical financial time series [93, 94, 95, 96, 97, 98]. The underlying principle is to use correlations from empirical financial time series to construct a sparse network representing the most relevant connections. Analyses on filtered correlation-based networks for information extraction [99, 96] have widely been used to explain market interconnectedness from high-dimensional data. Applications include asset allocation [100, 101], market stability assessments [102], hierarchical structure analyses [95, 96, 97, 103, 104] and the identification of lead-lag relationships [105].

The majority of literature so far has focused on the analysis of financial time series. However, in recent years a large amount of information about financial markets has become available from exogenous sources such as social media. It is reasonable to conceive that changes in social media sentiment [35] and changes in asset prices might be related. Some previous studies have indeed demonstrated the existence of relationships which in some cases indicated that social media can be used to predict changes in asset prices [38, 106, 19, 2, 13, 26]. When new information hits the markets, investors may react either rationally or irrationally [107, 5]. They may express opinions on social media that can later become market actions, thus enabling opportunities to forecast future asset prices. However, as demonstrated in the Chapter 5 not all assets behave in the same way. Some are more influenced by social media sentiment, while others, on the contrary, are more influential on the social media sentiment. Besides each single financial asset, we address in this Chapter whether the entire stock market structure is related to the structure constructed from social media sentiment and whether there exist lead-lag relationships exist that can be used for forecasting one structure in terms of the other.

We use dynamical Kendall correlations computed over rolling windows to take into account possible non-linear coupling in the investigation of the temporal evolution of market structure represented by filtered correlation-based networks constructed from stock market prices and from Twitter sentiment signals. We generate two networks: one from log-returns of stock prices and the other from Twitter sentiment. The two networks are treated as a multilayer problem with two layers of networks that share the same nodes but have different edge sets. We investigate whether financial market structure can be better predicted by combining past financial information with past social media sentiment information. The market structure forecasting problem is formulated as a link prediction problem where we estimate the probability of addition or removal of a link in the future based on information about the structure of the financial and social networks in the past.

6.2 Methods

6.2.1 Financial and Social Networks

We selected $N = 100$ of the most capitalized companies that were part of the S&P500 index from 09/05/2012 to 08/25/2017. The list of these companies' ticker symbols is reported in the Appendix B.1. For each stock i the financial variable was defined as the daily stock's log-return $R_i(\tau)$ at time τ. The social media variable was defined as the the social media opinion O_i of stock i which was estimated as the total number of bullish daily tweets related to the stock i at time τ (see Section 2.1.3).

Stock returns R_i and social media opinion scores O_i each amounted to a time series of length equals to 1251 trading days. These series were divided time-wise into $M = 225$ windows $t = 1, 2, \dots, M$ of width $T = 126$ trading days. A window step length parameter of $\delta T = 5$ trading days defined the displacement of the window, i.e., the number of trading days between two consecutive windows. The choice of window width T and window step δT is arbitrary, and it is a trade-off between having analysis that is either too dynamic or too smooth. The smaller the window width and the larger the window steps, the more dynamic the data are.

To characterize the synchronous time evolution of assets, we used equal time

Kendall's rank coefficients between assets i and j, defined as

$$\rho_{i,j}(t) = \sum_{t'<\tau} sgn(V_i(t') - V_i(\tau))sgn(V_j(t') - V_j(\tau)), \tag{6.1}$$

where t' and τ are time indexes within the window t and $V_i \in \{R_i, O_i\}$.

Kendall's rank coefficients takes into account possible nonlinear (monotonic) relationships. It fulfill the condition $-1 \leq \rho_{i,j} \leq 1$ and form the $N \times N$ correlation matrix $C(t)$ that served as the basis for the networks constructed in this work. To construct the asset-based financial and social networks, we defined a distance between a pair of stocks. This distance was associated with the edge connecting the stocks, and it reflected the level at which they were correlated. We used a simple non-linear transformation $d_{i,j}(t) = \sqrt{2(1-\rho_{i,j}(t))}$ to obtain distances with the property $2 \geq d_{i,j} \geq 0$, forming a $N \times N$ symmetric distance matrix $D(t)$.

We extracted the $N(N-1)/2$ distinct distance elements from the upper triangular part of the distance matrix $D(t)$, which were then sorted in an ascending order to form an ordered sequence $d_1(t), d_2(t), \ldots, d_{N(N-1)/2}(t)$. Since we require the graph to be representative of the market, it is natural to build the network by including only the strongest connections. This is a network filtering procedure that has been successfully applied in the construction of *asset graphs* for the analyses of market structure [108, 109]. The number of edges to include is arbitrary, and we included those from the bottom quartile, which represented the 25% shortest edges in the graph (largest correlations), thus giving $E(t) = \{d_1(t), d_2(t), \ldots, d_{\lfloor N/4 \rfloor}(t)\}$.

We denoted $E^F(t)$ and $E^S(t)$ as the set of edges constructed from the distance matrices derived from stock returns $R(t)$ and social media opinion $O(t)$, respectively. Two networks were considered as two layers of a multiplex structure [110] $\mathcal{G} = \{G^F, G^S\}$ where $G^F = (V, E^F)$, $G^S = (V, E^S)$ and V is the vertex set of stocks which is common to both layers.

6.2.2 Persistence

The state of an edge between vertices u and v in the financial layer at time t was represented with the corresponding adjacency matrix element $E^F_{u,v}(t)$: a binary variable with $E^F_{u,v}(t) = 1$ indicating the existence of the edge and $E^F_{u,v}(t) = 0$ its absence. Analogously, the variable $E^S_{u,v}(t)$ accounted for the presence or absence of edge (u,v) in the social (S) layer. The variable $E_{u,v}(t) = E^F_{u,v}(t) \vee E^S_{u,v}(t) = 1$ indicates instead the presence of at least one edge between u and v in the two layers; $E_{u,v}(t) = 0$ indicates that no edges are present between u and v in any layer.

6.2.3 Triadic Closure

Let $\mathscr{N}_{uv}$ be the set of nodes that are common neighbors to vertices u and v. We defined the triadic closure $T^F_{u,v}(t)$ of an edge (u,v) at layer F and time t as the mean of the clustering coefficients of vertices in $\mathscr{N}_{uv}$:

$$T^F_{u,v}(t) = \frac{1}{|\mathscr{N}_{uv}|} \sum_{i \in \mathscr{N}_{uv}} C^F_i(t), \tag{6.2}$$

where term C^F_i is the clustering coefficient of node i which accounts for the fraction of triads in the neighbors of i that are closed in triangles. This is defined as

$$C^F_i = 2\frac{\text{Number of triangles with a vertex on } i}{k_i(k_i-1)} = \frac{\sum_{j,k \in \mathscr{N}_i} E^F_{j,k}}{k_i(k_i-1)}, \tag{6.3}$$

where k_i is the degree of vertex i and $\mathscr{N}_i$ is the neighborhood of i.

In the multiplex case, we kept the same definition but allowed triangles to form across several layers [110, 111]. For the multiplex case, we used the symbol $T_{u,v}(t)$.

6.2.4 Link Prediction

We aim to predict the probability that an edge is inserted or removed in the financial network, $G^F(t+h)$, at a future time $t+h$ by using the information about the past structures of the financial and social networks at previous times $t' \leq t$. For this purpose we considered two mechanisms:

1) the tendency of an edge present at a previous time to persist in the future (*edge persistence*);

2) the propensity of triangles within or across layers to close (*triadic closure*).

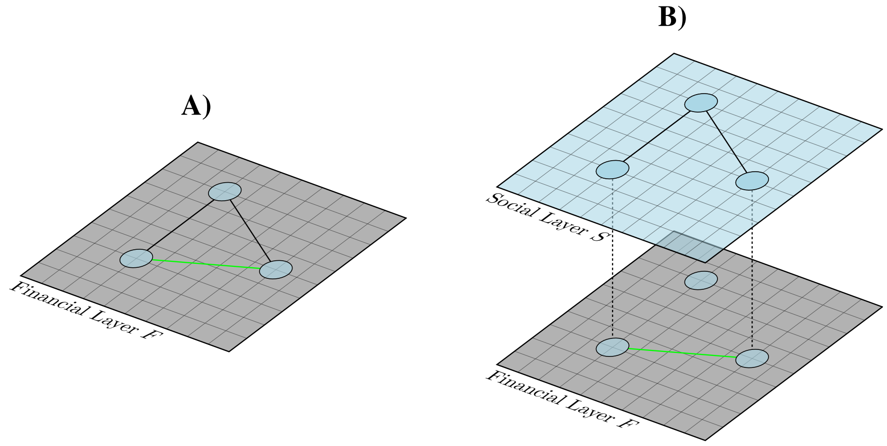


Figure 6.1: Triads on a single layered network (Panel A) and on a multiplex network (Panel B). The clustering coefficient of node i accounts for the fraction of triads in the neighborhood of i that are closed in triangles. The triadic closure of an edge (u,v) at layer F is a function of the clustering coefficients of the common neighbors of the vertices u and v. Triangles can be formed in a single layer or across layers.

The mechanism of growth by triadic closure is based on a principle of transitivity, often observed in real-world networks, where there is a tendency to form triangles [111]. Under this principle, two nodes tend to be connected if they share common neighbors with high transitivity, i.e., propensity to close triangles.

The probability that an edge will be inserted in the future is computed by means of a logistic regression of the edge persistence and the triadic closure coefficients. We estimated regression coefficients by best fitting on a training set which was composed of rolling windows of 126 trading days that initially ranged from 09/05/2012 to 09/10/2014. Predictions concerning the presence of edges in the financial network were made at $h = 1$ to $h = 20$ weeks ahead of the end of the training set. The test set initially ranged from 09/17/2014 to 08/25/2017. The procedure was repeated by moving the training window forward in 1-week steps.

The probability $p_{u,v}(t+h)$ to observe vertices u, v connected by an edge at $t+h$ can be inferred in terms of the set of previous triadic closure coefficients, $T_{u,v}(t)$, and edge persistence scores $E_{u,v}(t)$. We first considered a restricted model that used financial information only, which is given by the following:

$$\log \frac{p^F_{u,v}(t+h)}{1-p^F_{u,v}(t+h)} = \tilde{\beta}_0^h + \tilde{\beta}^h T^F_{u,v}(t) + \tilde{\gamma}^h E^F_{u,v}(t). \tag{6.4}$$

For this restricted model, we performed a 1-step ahead prediction for $h \in (1, 2, \ldots, 19, 20)$ weeks.

To calibrate the parameters in Eq. 6.4, we considered a training window of $W = 126$ days which ends at time t. The log-likelihood function [112] over the training window for the logistic model from Eq. 6.4 is given by

$$\begin{aligned} \mathscr{L}^F(t) = & \sum_{t'=t-W+1}^{t} \sum_{uv \in E^F(t'+h)} -\log\left(1+e^{\tilde{\beta}_0^h+\tilde{\beta}^h T^F_{u,v}(t')+\tilde{\gamma}^h E^F_{u,v}(t')}\right) + \\ & \sum_{t'=t-W+1}^{t} \sum_{uv \in E^F(t'+h)} (1-E^F_{uv}(t'+h))(\tilde{\beta}_0^h + \tilde{\beta}^h T^F_{u,v}(t') + \tilde{\gamma}^h E^F_{u,v}(t')). \end{aligned} \tag{6.5}$$

We differentiated the log-likelihood function given by Eq. 6.5 in order to find maximum log-likelihood estimates for the coefficients of Eq. 6.4.

To test whether the multiplex information is relevant in the prediction of links in the financial network compared to past a financial network alone, we considered a full regression model that takes the set of previous triadic closure coefficients and edge persistence from the financial layer $(T^F_{u,v}(t), E^F_{u,v}(t))$, social layer $(T^S_{u,v}(t), E^S_{u,v}(t))$ and the multiplex network $(T^F_{u,v}(t), E^F_{u,v}(t))$. The full model is

$$\begin{aligned} \log \frac{p_{u,v}(t+h)}{1-p_{u,v}(t+h)} = & \beta_0^h + \beta_1^h T^F_{u,v}(t) + \beta_2^h E^F_{u,v}(t) + \\ & \gamma_1^h T^S_{u,v}(t) + \gamma_2^h E^S_{u,v}(t) + \theta_1^h T_{u,v}(t) + \theta_2^h E_{u,v}(t). \end{aligned} \tag{6.6}$$

The log-likelihood function $\mathscr{L}(t)$ of the full model in Eq. 6.6 and the model fitting can be obtained in an analogous manner to the previously performed procedure for the restricted model from Eq. 6.4.

The likelihood ratio statistic is

$$\lambda(t) = -2(\mathscr{L}_{max}(t) - \mathscr{L}^F_{max}(t)), \tag{6.7}$$

where $\mathscr{L}_{max}(t)$ and $\mathscr{L}^{F}_{max}(t)$ are,respectively, the maxima of the log-likelihood functions of the full and restricted models in the training set window. The likelihood ratio statistic $\lambda(t)$ can be assumed to follow a χ^2 distribution [112] with 4 degrees of freedom where a value of $\lambda > 18.47$ is assumed to be statistically significant at $p = 0.001$. In that case, there is evidence to accept the full model that considers social and financial information over the restricted model that considers financial information only.

The model performance was estimated by counting both the true positives (edges predicted to be there and indeed present in the future network) and the false positives (edges predicted to be there but not present in the future network) and measuring of AUC (area under the receiver operating characteristic curve) in the test set that originally ranged from 09/17/2014 to 08/25/2017. AUC ranges from 0.50 to 1.00, with higher values indicating that the model discriminates better between the two categories of edge-present and edge-absent.

6.3 Results

6.3.1 Market structure dynamics

We first investigated financial network persistence by comparing the financial network $G^F(t)$ at time t with a future financial network, $G^F(t+h)$ at h steps ahead. To quantify the changes in the correlation network structure, we used two measures: A) the fraction of new edges in $G^F(t+h)$ that were not present in $G^F(t)$; B) the Jaccard Distance, defined as

$$Jaccard(G^F(t'), G^F(t)) = \frac{||G^F(t') \cap G^F(t)||}{||G^F(t') \cup G^F(t)||}.$$

Results are reported in Fig. 6.2, panels A) and B), respectively.

Fig. 6.2 panel A) shows the mean percentage of new edges in the financial network at time $t+h$ with respect to the edge set at time t ($1 \leq h \leq 20$ trading weeks). We observe that edges change considerably in the financial network with almost 40% of edges in financial networks changing after a period of $h = 20$ trading

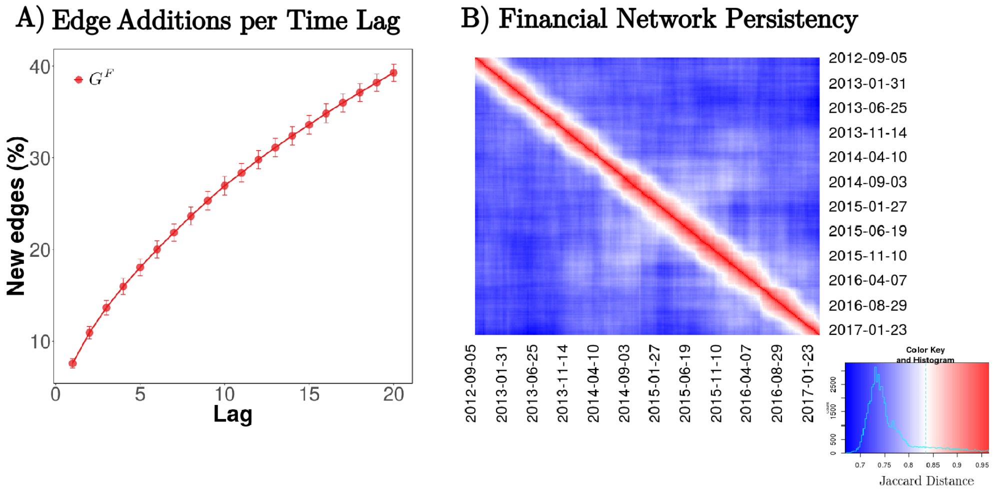


Figure 6.2: Evidence that financial correlation structure changes considerably with time. Panel A) shows the mean percentage of new edges in the financial network at time $t+h$ with respect to the edge set at time t ($1 \leq h \leq 20$ trading weeks). We observe that edges change considerably in the financial network with almost 40% of edges in financial networks changing after a period of $h = 20$ trading weeks. Panel B) shows the cross-similarity among financial networks measured as the Jaccard Distance between $G^F(t')$ and $G^F(t)$ with t and t' ranging from 09/05/2012 to 21/02/2017. We observe that edge changes (persistence) are quite stable overtime, i.e., the number of edges that change is similar throughout the period. Network $G^F(t)$ are constructed at each time t from a correlation structure estimated from a sliding window of 126 trading days starting at time t. The windows move with time step of 1 trading week. Error bars in Panel A) indicate standard error.

weeks. Fig. 6.2 panel B) shows the cross-similarity among financial networks measured as the Jaccard Distance between $G^F(t')$ and $G^F(t)$ with t and t' ranging from 09/05/2012 to 21/02/2017. We observe that edge changes (persistence) are quite stable overtime, i.e., the number of edges that change is similar throughout the period. Hence, results indicate that the constructed financial networks are time-variant across the entire period studied, with a stable rate of edge changes over time.

6.3.2 Prediction of Stock Market Structure

We used Eq. 6.6 to predict a the financial network $G^F(t+h)$ at a future time $t+h$ by using the information about the past structures of the financial and social networks at previous times $t' \leq t$. Fig. 6.3 panel A) shows the performance obtained in the prediction of out-of-sample edges for $h \in (1,5,10,15,20)$ trading steps ahead.

We achieved an overall high out-of-sample performance in financial network link prediction, with performances in the range of 73% to 95% depending on time-lag and time-period. Prediction power improved with a smaller time lag.

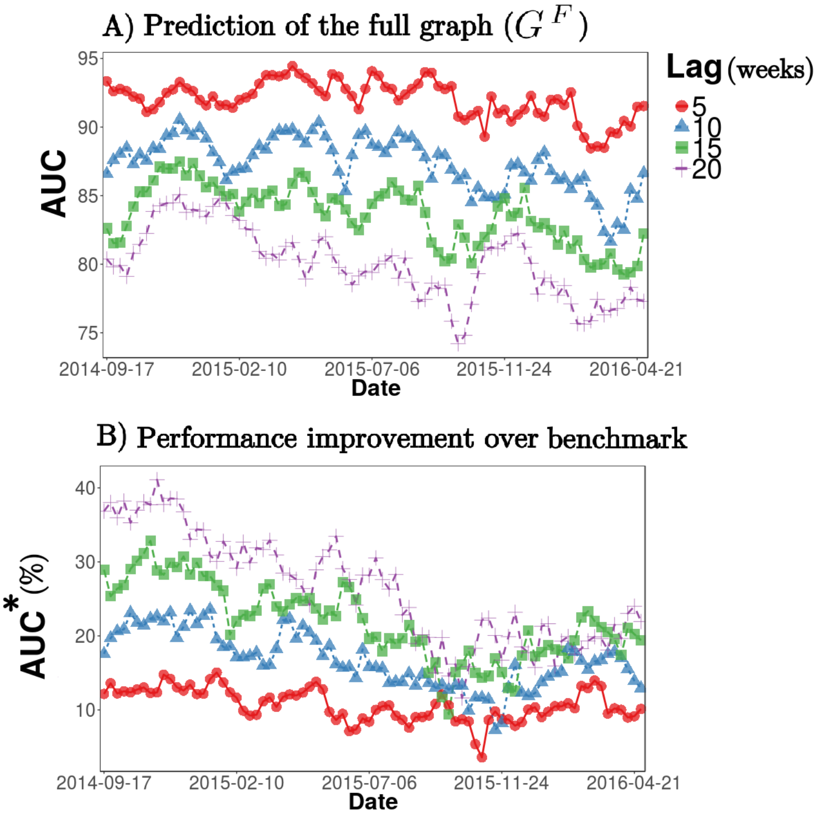


Figure 6.3: Evidence of high out-of-sample performance in financial network link prediction. Models were trained in an expanding window with initial start and end dates 09/05/2012. and 09/10/2014, respectively. Test period ranges from 09/17/2014 and 08/25/2017. Plots display the performance results (AUC) of a model to predict edges in a financial network at time $t+h$ trained with information up to date t. Panel A) shows the performance obtained in the prediction of out-of-sample edges for $h \in (1, 5, 10, 15, 20)$ trading weeks. Panel B) shows the performance improvement (AUC^*) compared to a naive benchmark that assumes that the correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$.

We compared our results to those obtained using a benchmark model that assumes that correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$. The performance improvement against the benchmark is estimated as $AUC^* = (AUC - 0.5)/(\widehat{AUC} - 0.5) - 1$, where AUC represents the performance of the proposed model and $\widehat{AUC}$ is the performance of the benchmark. From Fig. 6.3 panel B), we observe that the higher the time lag, the higher the performance improvement over the benchmark. Let us note that performance improvement over the naive benchmark reached values as high as 40% for a long-term prediction with a lag of 20 trading weeks.

Fig. 6.4 reports an aggregate overview of the previous results for the out-of-sample prediction in terms of the number of weeks ahead. We observe that as the lag increases, the prediction performance declines (panel A). However, the improvement in performance over the naive benchmark improves (panel B).

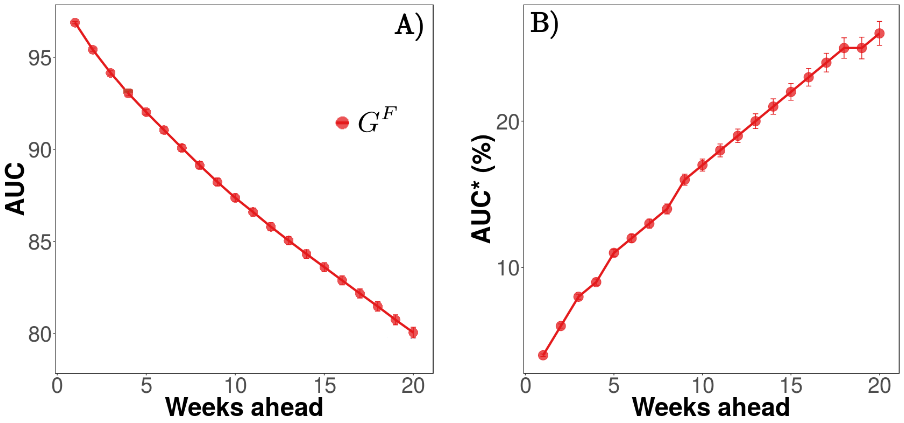


Figure 6.4: The effect of time-lag on out-of-sample predictive performance. Panel A) shows the mean performance (AUC) of the prediction of out-of-sample edges of the full financial network G^F. Panel B) shows the performance improvement (AUC^*) against a naive benchmark that assumes that correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$. Error bars indicate standard error.

In Appendix B.2, we report the results obtained by using an expanding window rather than a rolling window as a training set. We observe that expanding the training set does not necessarily lead to better performance. In fact, the rolling window analysis yielded better performance overall.

To test whether the multiplex network provides additional information to that from the financial network only, we re-computed the same out-of-sample edge prediction by using the financial network only and compared this to the results from the full model that considers both the financial and social information layers. A comparison between the two models was performed by comparing their respective likelihoods. We have also disaggregated the prediction of the insertion of new edges E^+ and the prediction of edge deletions E^-. We report the likelihood values and

AUC performance obtained for the fit of each model in Table 6.1.

We observed that the model that includes both financial and social information better fit the data compared to the model that considers financial data only, particularly for the case of the prediction of insertion of new edges. The likelihood ratio increases with prediction lag indicating that full models (i.e. those that consider both financial and social networks) are particularly important in long-term link prediction. Results confirm that the multiplex network is distinctly better than the single financial layer with all likelihood ratios having p-value < 0.001 for all configurations tested.

6.3.3 Prediction of Social Opinion Structure

We have so far established that social opinion structure can provide statistically significant information about the future financial market structure. In this section, we investigate the opposite relationship of whether financial market structure can also significantly improve the prediction of future social opinion structure, and we determine if this effect is larger or smaller.

The comparison between performance results is summarized in Fig. 6.5, where the prediction of social opinion structure G^S is plotted together with the results for the prediction of financial market structure G^F that was discussed previously. Surprisingly, results suggest that financial market structure has a higher predictability than social opinion structure. We also observe that both the financial network and social opinion network predictions lead to an improvement compared to the naive benchmark that considers time invariance in social network structure. As previously observed, the relative performance improvement increases with time lag. In this case, the relative improvement in prediction is higher for the social opinion structure than for the financial network as observed in Fig. 6.5 panel B).

One of the possible reasons why social opinion structure is less predictable compared to financial network structure is the higher structural variability of the former compared to the latter. Fig. 6.6 provides evidence that social media structure is less stable than financial market structure in terms of the number of edge changes over time. More edges changed in the social opinion network than in the financial

Table 6.1: Financial Link Prediction Performance Results. High out-of-sample AUCs obtained indicate that the model has high performance balancing both false positives and false negatives predictions relative to true positive and negative values. Log-likelihood ratios (λ) increase with prediction lag indicating that social media features are particularly important for long-term prediction. The table reports mean AUC values and log-likelihood ratios λ over the test period with corresponding standard deviations in parentheses. Results are reported for the prediction of new edges E^* and edge deletions E^-. We also report the average performance AUC obtained in the prediction of the full-graph G^F, as well as, the performance improvement AUC^* over the benchmark that assumes that correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$. Models were trained with a rolling window with initial start and end dates of 09/05/2012 and 09/10/2014, respectively. The test period ranged from 09/17/2014 to 08/25/2017.

	E^+		E^-		G^F	
Lag	AUC	λ	AUC	λ	AUC	AUC^* (%)
1	87 (0.33)	21 (0.76)	93 (0.11)	34 (1.2)	97 (0.064)	4 (0.091)
2	87 (0.37)	33 (1.2)	93 (0.1)	45 (1.5)	95 (0.092)	6 (0.14)
3	86 (0.39)	48 (1.5)	93 (0.11)	60 (1.6)	94 (0.11)	8 (0.17)
4	86 (0.39)	65 (2)	93 (0.11)	65 (1.9)	93 (0.13)	10 (0.21)
5	85 (0.41)	85 (2.6)	93 (0.11)	66 (1.9)	92 (0.15)	11 (0.24)
6	85 (0.41)	100 (3.2)	93 (0.1)	74 (2)	91 (0.16)	12 (0.27)
7	84 (0.42)	120 (3.5)	93 (0.1)	70 (2.2)	90 (0.18)	13 (0.3)
8	84 (0.43)	150 (4.3)	93 (0.1)	72 (1.9)	89 (0.19)	15 (0.33)
9	83 (0.44)	180 (5.7)	93 (0.1)	74 (2.2)	88 (0.21)	16 (0.37)
10	83 (0.43)	220 (6.3)	93 (0.096)	79 (1.9)	87 (0.21)	17 (0.4)
11	82 (0.43)	260 (7.2)	93 (0.094)	78 (2)	87 (0.22)	18 (0.43)
12	82 (0.42)	300 (7.9)	93 (0.09)	86 (2.4)	86 (0.22)	19 (0.45)
13	82 (0.43)	330 (7.9)	93 (0.09)	95 (2.1)	85 (0.22)	20 (0.49)
14	81 (0.43)	360 (9.2)	93 (0.084)	100 (2.4)	84 (0.23)	21 (0.51)
15	81 (0.43)	390 (9.9)	93 (0.083)	110 (2.3)	84 (0.24)	22 (0.55)
16	81 (0.43)	410 (10)	93 (0.08)	120 (3)	83 (0.24)	23 (0.58)
17	80 (0.43)	440 (11)	94 (0.079)	130 (2.6)	82 (0.25)	24 (0.62)
18	80 (0.44)	470 (12)	94 (0.076)	150 (3)	82 (0.25)	25 (0.67)
19	80 (0.46)	500 (12)	94 (0.072)	160 (3.6)	81 (0.27)	26 (0.71)
20	80 (0.48)	510 (12)	94 (0.068)	170 (3.7)	80 (0.28)	27 (0.79)

*A likelihood ratio of $\lambda > 18.47$ indicates statistical significance at $p = 0.001$.

network for all lags tested. We observed that more than 50% of the edges in the social media opinion structure changed compared to 40% in the financial network over a time lag of 20 trading weeks.

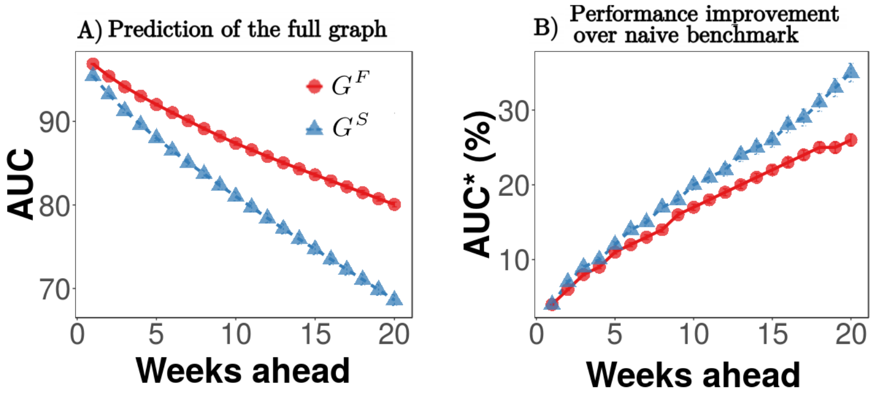


Figure 6.5: Evidence that financial market structure has higher predictability than social media structure. Panel A) shows mean performance (AUC) in the prediction of out-of-sample edges of the full financial network G^F and the social opinion network G^S. Panel B) shows the performance improvement (AUC^*) against a naive benchmark that assumes that the correlation structure is time-invariant. Error bars indicate standard error.

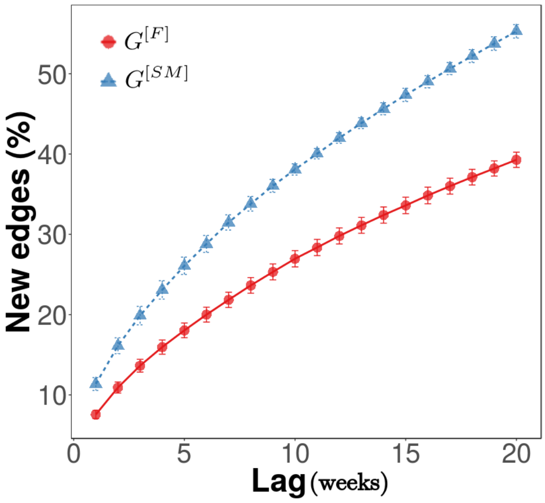


Figure 6.6: Evidence that social media structure is less stable than financial market structure in terms of number of edge changes in time. We observe that almost 40% of edges in Financial Networks changed after a period of 20 trading weeks while the social media structure changed more than 50% of its edges over the same time lag. A network at time t is constructed from a correlation structure estimated from a sliding window of 126 trading days starting at time t that moves with time step of 1 trading week. The financial network measures co-movement of stock returns while the social network measures co-movement of opinion over the same stocks. Error bars indicate standard error.

6.4 Discussion and Conclusions

We investigated whether financial market structure can be better predicted by combining past financial information with past social media sentiment information. We considered the $N = 100$ most capitalized companies that were part of the S&P500 index in the period between May 2012 and August 2017. We generated two networks: A financial network constructed from log-returns of equity prices and a social network constructed from Twitter sentiment analytics. We constructed filtered correlation-based networks by keeping the strongest top quartile correlations only that considered a rolling window of $T = 126$ trading days. The two networks were treated as a multiplex problem with two layers of networks that share the same nodes (stocks) but have different edge sets.

The financial market structure forecasting problem was formulated as a link prediction problem where we estimated the probability of the addition or removal of a link in the future on information about the past structure of financial and social opinion networks.

We proposed that financial network links were formed by a combination of the two mechanisms of triadic closure and edge persistence. The first mechanism assumes that two stocks have a propensity to be correlated if they share common neighbors. The edge persistence mechanism assumes that two connected stocks tend to remain connected in the future. A logistic model was trained over a set of data between 09/05/2012 and 09/10/2014 and then results were reported for the validation set over the following period from 09/17/2014 and 08/25/2017.

Our results indicate that financial market structure is considerably time variant, which invalidates the commonly used assumption of time invariance in the determination of stock correlation structure. The proposed model exhibited high out-of-sample performance in financial network link prediction, particularly in the case of long-term predictions where we observed a performance improvement of up to 40% over a naive benchmark that assumed that the correlation structure of the financial market was time invariant. Likelihood ratio analysis demonstrated that models that considered both financial and social information better fit the data when compared to a restricted model that considers financial information only. This provides evidence that supports the use of social information in the prediction of financial market structure.

Finally, our findings indicate that social opinion structure is less stable than financial market structure. Surprisingly, the prediction of financial market structure using past social and financial information presented higher performance compared to the problem of predicting social opinion structure using past social and financial information.

Let us note that network link formation can occur due to mechanisms beyond the ones we studied here. For instance, networks can form links as a result of a growth process that adds new nodes in the network, e.g., IPOs can generate growth in a financial network. Among other possible mechanisms, link formation can occur due to preferential attachment, a phenomenon widely observed in real networks where new nodes tend to link to the more connected ones [113].

In summary, this study indicates that social opinion structure is relevant to the prediction of future financial correlation structures. This result has important consequences because of the fundamental importance of financial correlation structure in Modern Portfolio Theory (MPT) [32], Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) [33]. Future work should focus on the investigation of further mechanisms of financial link formation and on applications in portfolio allocation strategies.

Chapter 7

Conclusion

The opinions of traders, analysts and other professionals, along with laypeople's opinions, can be widely expressed in the form of news articles, research reports, company transcripts and blogs, among many other sources available on the internet. Social media is an information channel of particular interest due to the high volume and velocity of activity of an ever-evolving network that is constantly providing and creating information. In this work, we provide evidence that supports the use of social media opinion as a relevant signal in the prediction of stock market movements.

Investors may react either rationally or irrationally [107, 5] in the presence of new information. Irrational behavior and investment opinions can influence market actions, thus enabling opportunities to predict asset prices. Therefore, better understanding the nature of the relationship between market opinion and financial dynamics can benefit current financial modeling, which is mostly focused on financial time-series data only.

Assuming social media as a proxy for human activity, behavior and collective opinion, the main objective of this work was to provide evidence on whether and to what extent financial dynamics can be better explained by collective opinion extracted from social media. The following hypotheses were evaluated:

- **H1: Social media sentiment has statistically significant causal relationship with stocks returns and volatility.** In Chapter 5, we provide evidence that social media sentiment has a significant causal relationship with price movements by analyzing constituents of the Dow Jones Industrial Average index. While analyzing retail brands, in particular, results suggested that social media can be a complementary source in the analysis of the financial dynamics to mainstream news such as the Wall Street Journal and Dow Jones Newswires.

- **H2: Social media sentiment has a nonlinear impact in stock price returns.** In Chapter 5, we provided empirical evidence that indicates that social media and stock markets have a nonlinear causal relationship. Our results serve as empirical guidance on model adequacy, market efficiency, and predictability, in the investigation of causal relationships between social and financial systems.

- **H3: Social media sentiment dominates directional coupling with the stock market; i.e., information provided by social media contributes to the description of stock market dynamics more than the opposite.** By analyzing a sentiment dataset composed of social media messages related to DJIA index components, in Chapter 5, we uncovered that information flows predominantly from social media to stock markets. Results from Chapter 6 were aligned with those from Chapter 5, while showing that social media opinion structure dominated the directional coupling with stock market structure; i.e., the prediction of the stock market structure using past social and financial information presented higher performance, compared to the prediction of social opinion structure using past social and financial information.

- **H4: Social media sentiment structure predicts stock market structure.** In Chapter 6, we demonstrated that social media mood can be used to predict not only individual asset prices but also overall market structure. We quantified the collective behavior of asset returns by constructing filtered correlation-based networks, and we showed that social media opinion structure predicts stock market structure. The proposed model exhibited high out-of-sample performance in the prediction of future market correlations among a subset of the most market capitalized S&P500 index constituents. This result has important consequences in any study of portfolio risk, capital allocation or hedging in trading strategies, which typically depend on the estimation of an expected asset correlation structure.

The study of how social mood can impact the stock market is an area of research on its early stages, and while promising, it imposes many challenges. This thesis has many limitations, which include the following:

- **Limitation in the asset universes and data sample.** Social media data are often sparse and of difficult acquisition. Results provided in Chapter 4 were limited to a small number of retail companies within a two-year period; thus, those results might differ if a different analysis were made in a different time period, or if it included a broader selection of companies. While Chapters 5 and 6 expanded the asset universe to a broader segment of the US market, the availability of longer history and expanded number of assets would greatly improve robustness of the results.

- **Limitation in the content universe.** We considered Twitter as main content source that served as a proxy for social opinion. Many other social networks can also be considered such as Facebook and Linkedin.

- **Other mechanisms can take place in financial link formation.** We demonstrated that stock market structure can be better predicted using social media opinion structure by assuming link formation mechanisms based on link persistence and triadic closure. Nonetheless, market dynamics can often be unpredictable; also, links between assets can be formed by a combination of many other reasons beyond those studied in Chapter 6. Other mechanisms can take place in financial link formation such as network growth and preferential attachment, which are two mechanisms among many that are widely observed in real networks [113].

- **Limitations in the quantification of causality.** The quantification of true causality is an open research problem. We approached the problem of measuring causality under a notion of statistical Granger-causality. Nonetheless, it is important to note that many other confounding factors could have contributed to the Granger-causality tests performed, which were limited to the variables here considered (i.e. social sentiment, news sentiment, stock returns and volatility). Moreover, the limitations in the sample data size limited the robustness of the tests performed.

- **Past evidence is not indicative of future performance.** Market players have already begun to exploit the predictive power of social media over future prices, and the effects we have uncovered with this work could fade in the future when speculation erodes arbitrage opportunities.

We consider many future research avenues including the following:

- **Multi-asset class investment.** Future research should investigate the impact of social opinion in multiple asset classes beyond equity. Evidence that supports the use of social opinion in fixed income, FX, indices, commodities, equity among other asset classes can allow for the construction of well-diversified portfolio strategies.

- **Modeling multiple information channels.** Future research should consider multiple information channels that could measure crowd option captured from additional social networks (e.g. Instagram, Facebook and Linkedin), as well as information channels that could capture expert opinion such as Traders chat rooms. Each information channel can be modeled as a network, whereby nodes represent assets and links represent co-movement of opinion. This multiplex framework allows for the investigation of how collective opinion affects financial markets while handling high dimensional data.

- **Portfolio optimization.** In Chapter 6, we provided evidence that market structure can be inferred from social opinion structure. Future research should consider the incorporation of collective opinion into traditional portfolio optimization models, which currently rely in the estimation of financial correlation structure.

In summary, we argue that social opinion is a relevant exogenous signal, compared to traditional market data in the prediction of stock market prices and structure. We observed that social media sentiment can be a complementary signal to mainstream news, when analyzing a subset of US retail brands. We demonstrated that the significant causal relationship between social opinion and stock returns are purely nonlinear for most of the DJIA index constituents. We also provided evidence that suggests that social media predicts stock market structure for S&P500 index constituents. Finally, our findings suggest that social media sentiment dominates the directional coupling with the stock market in the prediction of individual asset dynamics, as well as in the prediction of the overall market structure.

Appendix A

Supporting Information to Chapter 5

A.1 List of Companies Analyzed

The names with their respective Reuters Instrument Codes (RIC) of the stocks investigated in Chapter 5 are the following: INTEL CORP. (INTC.O), VISA INC. (V.N), NIKE INC. (NKE.N), E.I. DUPONT DE NEMOURS & CO. (DD.N), JP-MORGAN CHASE & CO. (JPM.N), BOEING CO. (BA.N), MERCK & CO. INC. (MRK.N), PFIZER INC. (PFE.N), MICROSOFT CORP. (MSFT.O), COCA-COLA CO. (KO.N), GOLDMAN SACHS GROUP INC. (GS.N), MCDONALD'S CORP. (MCD.N), GENERAL ELECTRIC CO. (GE.N), 3M CO. (MMM.N), UNITED TECHNOLOGIES CORP. (UTX.N), VERIZON COMMUNICATIONS INC. (VZ.N), CISCO SYSTEMS INC. (CSCO.O), HOME DEPOT INC. (HD.N), INTERNATIONAL BUSINESS MACHINES CORP. (IBM.N), AMERICAN EXPRESS CO. (AXP.N), PROCTER & GAMBLE CO. (PG.N), APPLE INC. (AAPL.O), UNITEDHEALTH GROUP INC. (UNH.N), CATERPILLAR INC. (CAT.N), EXXON MOBIL CORP. (XOM.N), JOHNSON & JOHNSON (JNJ.N), WAL-MART STORES INC. (WMT.N), WALT DISNEY CO. (DIS.N), CHEVRON CORP. (CVX.N) and THE TRAVELERS COMPANIES INC. (TRV.N).

A.2 Functional Forms Tested in Table 5.1.

Differencing: ∇x**.** The first differencing is taken in both social media and returns time series.

$$\nabla(R(t),1) = \widehat{\alpha} + \sum_{\Delta t=1}^{k} \widehat{\beta}_{\Delta t} \nabla(R(t-\Delta t),1) + \sum_{\Delta t=1}^{k} \widehat{\gamma}_{\Delta t} \nabla(SM(t-\Delta t),1) + \widehat{\varepsilon}_t. \quad (A.1)$$

The second differencing $\nabla^2 x$ was tested in analogous way.

Linear regression: $f(x,vol)$**.** Represents a linear regression of returns on social media controlled by the stocks' returns daily volatility.

$$R(t) = \widehat{\alpha} + \sum_{\Delta t=1}^{k} \widehat{\beta}_{\Delta t} R(t-\Delta t) + \sum_{\Delta t=1}^{k} \widehat{\gamma}_{\Delta t} SM(t-\Delta t) + \sum_{\Delta t=1}^{k} \widehat{\theta}_{\Delta t} vol(t-\Delta t) + \widehat{\varepsilon}_t, \quad (A.2)$$

where we consider

$$vol(t) = 2\frac{P_{high}(t) - P_{low}(t)}{P_{high}(t) + P_{low}(t)} \quad (A.3)$$

as an approximation of the daily returns volatility. P_{high} and P_{low} are the highest and lowest intraday price value, respectively.

Log-transformation: $\log(x+1)$**.**

$$\log(R(t)+1) = \widehat{\alpha} + \sum_{\Delta t=1}^{k} \widehat{\beta}_{\Delta t} \log(R(t-\Delta t)+1) + \sum_{\Delta t=1}^{k} \widehat{\gamma}_{\Delta t} \log(SM(t-\Delta t)+1) + \widehat{\varepsilon}_t. \quad (A.4)$$

Absolute value: $|x|$**.**

$$|R(t)| = \widehat{\alpha} + \sum_{\Delta t=1}^{k} \widehat{\beta}_{\Delta t} |R(t-\Delta t)| + \sum_{\Delta t=1}^{k} \widehat{\gamma}_{\Delta t} SM(t-\Delta t) + \widehat{\varepsilon}_t. \quad (A.5)$$

GARCH(1,1). A GARCH filtering was applied in the original returns time series as follows:

$$R(t) = \alpha + \sum_{\Delta t=1}^{p} \beta_{\Delta t} R(t-\Delta t) + \sum_{\Delta t=1}^{q} \gamma_{\Delta t} \varepsilon_{t-\Delta t}, \tag{A.6}$$

with $p = 1$, $q = 1$ and

$$\varepsilon_t \sim N(0, R(t)). \tag{A.7}$$

The resulting residuals ε_t were then used instead of the original stock returns time series $R(t)$.

ARIMA(1,1,1). ARIMA filtering was applied in the original returns time series as follows:

$$R(t) = R(t-1) + \alpha(SM(t-1) - SM(t-2)) + \beta \varepsilon_{t-1}. \tag{A.8}$$

The resulting residuals ε_t were then used instead of the original stock returns time series $R(t)$.

Appendix B

Supporting Information to Chapter 6

B.1 Ticker Codes of Selected Companies

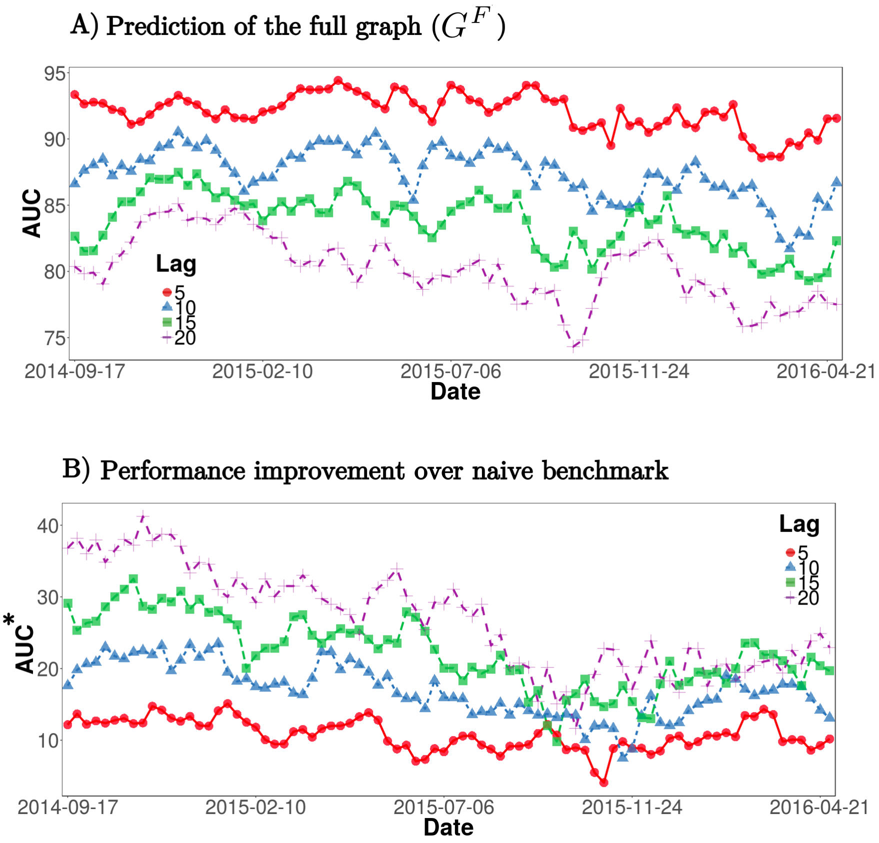
The list of companies used in the experiment performed in Chapter 6 follows: AAPL, AMZN, NFLX, MSFT, GS, GOOGL, BAC, JPM, IBM, DIS, GILD, INTC, YHOO, WMT, GE, XOM, SBUX, CSCO, WFC, NVDA, PCLN, JNJ, MCD, NKE, BA, VZ, ES, PFE, KO, CVX, CAT, MU, MRK, CELG, EBAY, MS, CRM, FCX, QCOM, TGT, HD, CHK, BMY, AMGN, PG, HPQ, ORCL, FSLR, WFM, COST, BIIB, PEP, EA, AXP, WYNN, CMCSA, CL, AIG, DOW, NEM, MA, BBY, COP, LOW, TWX, ADBE, HAL, LLY, UNH, LUV, MMM, CVS, MO, FDX, DD, ED, KR, MON, UTX, ABT, SLB, YUM, MCO, AMAT, EXPE, AET, DE, GPS, UPS, VLO, CBS, HAS, COH, ALL, WDC, JWN, TXN, PM, UNP, EOG.

B.2 Stock Market Structure Prediction Results Using an Expanding Window Training Set

In this section, we report results using models that were trained in an expanding window, instead of a rolling window, using initial start and end dates of 09/05/2012 and 09/10/2014, respectively. The test period ranges from 09/17/2014 to 08/25/2017.

B.3 Model parameters t-statistic

Plot shows in log scale the mean of t-statistic of triadic closure variables for social layer ($triadic^{[SM]}$), financial layer ($triadic^{F}$) and multiplex ($triadic^{[\mathscr{G}]}$) in the link



A) Prediction of the full graph (G^F)

B) Performance improvement over naive benchmark

Figure B.1: Link prediction results using an expanding window training set. Evidence of high out-of-sample performance in financial network link prediction. Models were trained in an expanding window with initial start and end dates 09/05/2012. and 09/10/2014, respectively. Test period ranges from 09/17/2014 and 08/25/2017. Plots display the performance results (AUC) of a model to predict edges in a financial network at time $t+h$ trained with information up to date t. Panel A) shows the performance obtained in the prediction of out-of-sample edges for $h \in (1,5,10,15,20)$ trading weeks. Panel B) shows the performance improvement (AUC^*) compared to a naive benchmark that assumes that the correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$.

prediction of financial networks across the multiple lags tested. Panel A) shows the t-statistics obtained for the problem of predicting new edges in future financial networks. We observe that statistical significance of social media and multiplex triadic closure variables increase with lag. Panel B) shows the t-statistics obtained for the problem of predicting edge removals. We observe a lower statistical significance of social media and multiplex triadic closure variables compared to a corresponding prediction of new edges in the financial network. All the *autocorrelation* variables presented insignificant t-statistic values.

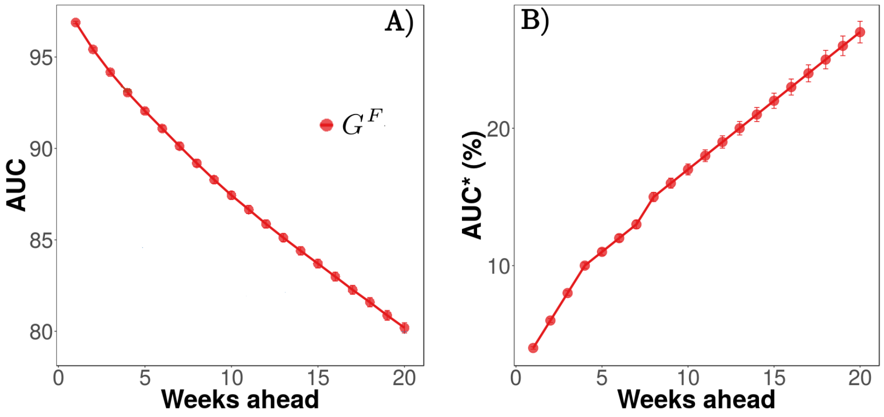


Figure B.2: Link prediction results using an expanding window training set. The effect of time-lag on out-of-sample predictive performance. Panel A) shows the mean performance (AUC) of the prediction of out-of-sample edges of the full financial network G^F. Panel B) shows the performance improvement (AUC^*) against a naive benchmark that assumes that correlation structure is time-invariant, i.e., $G^F(t+h) = G^F(t)$. Error bars indicate standard error.

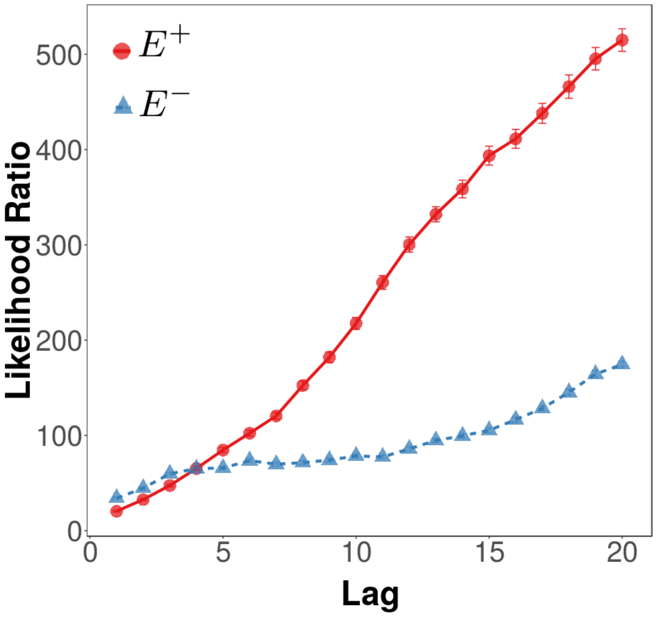


Figure B.3: Expanding Window - Likelihood ratio demonstrates that models that consider both financial and social media features fit the data significantly better than the restricted model that considers financial network features only. Likelihood ratio increases with prediction lag indicating that full models (i.e. those that consider both financial and social networks) are particularly important in long-term link prediction. Likelihood ratios in the prediction of new edges are higher than the likelihood ratios in the prediction of edge deletions indicating that social features are specialily important in the formation of new financial links.

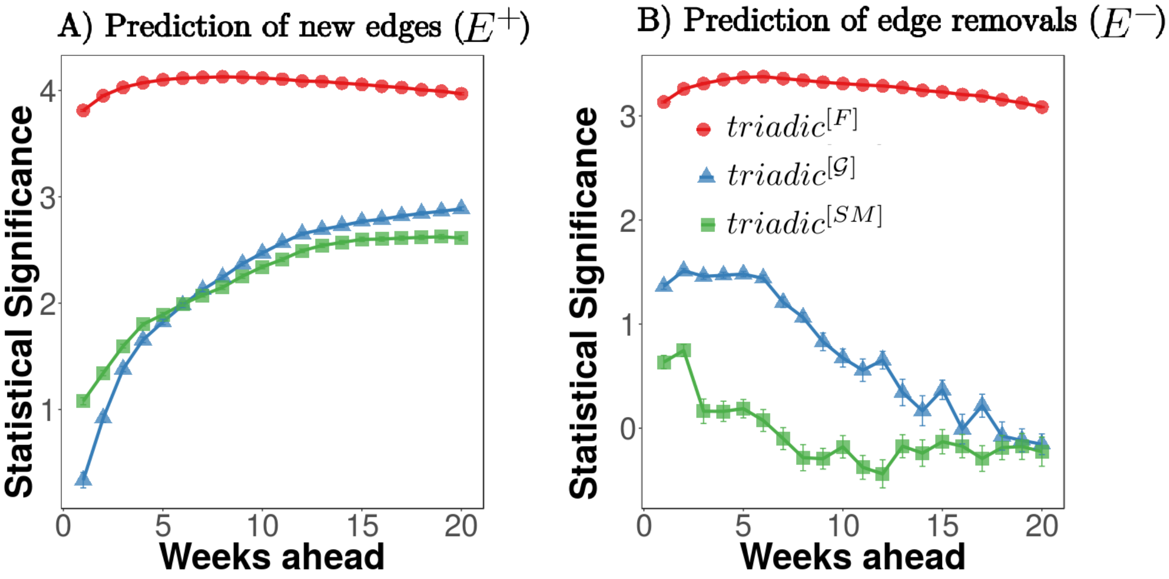


Figure B.4: Expanding Window - Evidence that social and multiplex triadic closure variables are statistically significant in the prediction of links of financial networks, particularly for the case of prediction of newly added edges. Figure shows. in log scale, the mean of t-statistic of triadic closure variables for social layer ($triadic^{[SM]}$), financial layer ($triadic^{F}$) and multiplex ($triadic^{[\mathcal{G}]}$) in the link prediction of financial networks across the multiple lags tested. Panel A) shows the t-statistics obtained for the problem of predicting new edges in future financial networks. We observe that statistical significance of social media and multiplex triadic closure variables increase with lag. Panel B) shows the t-statistics obtained for the problem of predicting edge removals. We observe a lower statistical significance of social media and multiplex triadic closure variables compared to a corresponding prediction of new edges in the financial network.

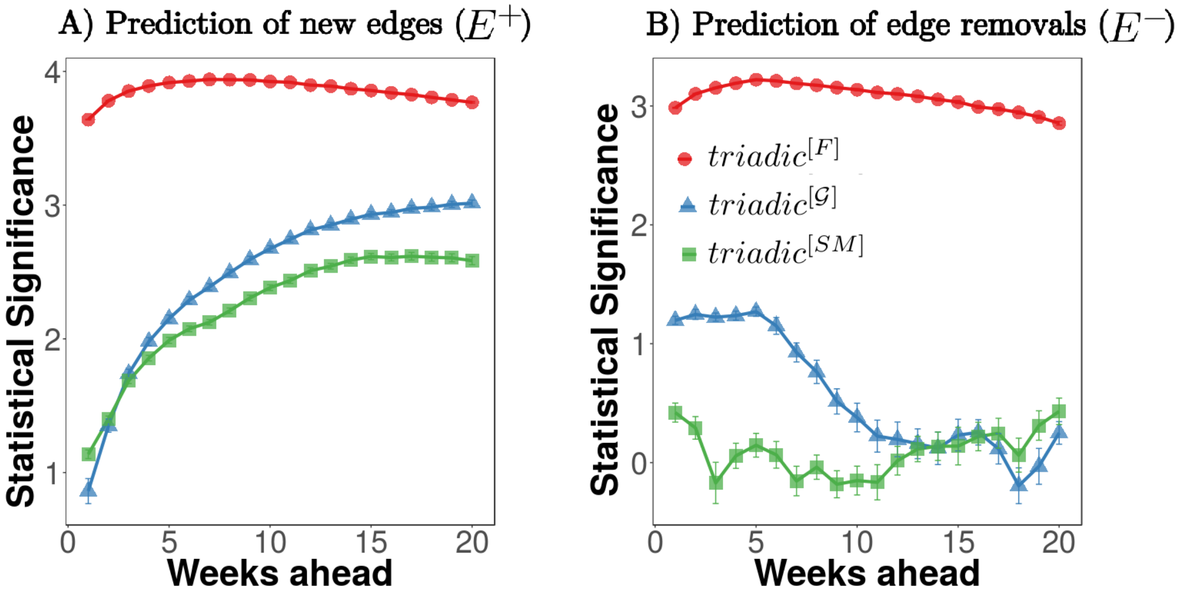


Figure B.5: Evidence that social and multiplex triadic closure variables are statistically significant in the prediction of links of financial networks, particularly for the case of prediction of newly added edges. Plot shows. in log scale, the mean of t-statistic of triadic closure variables for social layer ($triadic^{[SM]}$), financial layer ($triadic^{F}$) and multiplex ($triadic^{[\mathscr{G}]}$) in the link prediction of financial networks across the multiple lags tested. Panel A) shows the t-statistics obtained for the problem of predicting new edges in future financial networks. We observe that statistical significance of social media and multiplex triadic closure variables increase with lag. Panel B) shows the t-statistics obtained for the problem of predicting edge removals. We observe a lower statistical significance of social media and multiplex triadic closure variables compared to a corresponding prediction of new edges in the financial network.

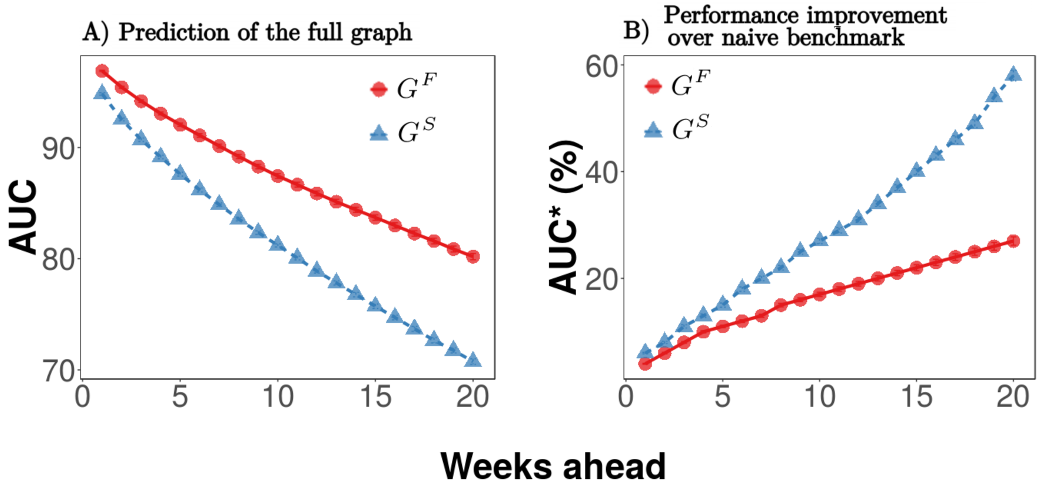


Figure B.6: Link prediction results using an expanding window training set. Evidence that social media structure is less stable than financial market structure in terms of number of edge changes in time. We observe that almost 40% of edges in Financial Networks changed after a period of 20 trading weeks while the social media structure changed more than 50% of its edges over the same time lag. A network at time *t* is constructed from a correlation structure estimated from an expanding window of 126 trading days starting at time *t* that moves with time step of 1 trading week. The financial network measures co-movement of stock returns while the social network measures co-movement of opinion over the same stocks. Error bars indicate standard error.

Bibliography

[1] Twitter. *Twitter Blog.* https://bit.ly/2ygWARh. Last accessed on Oct 14, 2018. 2018.

[2] P. C. Tetlock. "Giving content to investor sentiment: The role of media in the stock market". In: *The Journal of Finance* 62.3 (2007), pp. 1139–1168.

[3] C. Curme, T. Preis, H. E. Stanley, and H. S. Moat. "Quantifying the semantics of search behavior before stock market moves". In: *Proceedings of the National Academy of Sciences* 111.32 (2014), pp. 11600–11605. ISSN: 0027-8424. DOI: 10.1073/pnas.1324054111. eprint: https://www.pnas.org/content/111/32/11600.full.pdf. URL: https://www.pnas.org/content/111/32/11600.

[4] E. F. Fama. "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2 (1970), pp. 383–417. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.1970.tb00518.x. URL: http://dx.doi.org/10.1111/j.1540-6261.1970.tb00518.x.

[5] A. Shleifer. *Inefficient Markets: An Introduction to Behavioral Finance*. Clarendon Lectures in Economics. OUP Oxford, 2000. ISBN: 9780191606892.

[6] J. Bollen, H. Mao, and X. Zeng. "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1 (2011), pp. 1–8. ISSN: 1877-7503.

[7] T. O. Sprenger, P. G. Sandner, A. Tumasjan, and I. M. Welpe. "News or Noise? Using Twitter to Identify and Understand Company-specific News Flow". In: *Journal of Business Finance & Accounting* 41.7-8 (Sept. 2014), pp. 791–830.

[8] L. Mitra and G. Mitra. "Applications of news analytics in finance: A review". In: *The Handbook of News Analytics in Finance*. John Wiley & Sons, Ltd., 2011, pp. 1–39. ISBN: 9781118467411.

[9] G. Mitra and L. Mitra. *The Handbook of News Analytics in Finance*. The Wiley Finance Series. Wiley, 2011. ISBN: 9781119990802.

[10] I. Aldridge. *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*. Wiley Trading. Wiley, 2013. ISBN: 9781118416822.

[11] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy. "More than words: Quantifying language to measure firms' fundamentals". In: *The Journal of Finance* 63.3 (2008), pp. 1437–1467.

[12] F. Lillo, S. Miccichè, M. Tumminello, J. Piilo, and R. N. Mantegna. "How news affects the trading behaviour of different categories of investors in a financial market". In: *Quantitative Finance* 15.2 (2015), pp. 213–229. DOI: 10.1080/14697688.2014.931593. eprint: http://dx.doi.org/10.1080/14697688.2014.931593. URL: http://dx.doi.org/10.1080/14697688.2014.931593.

[13] M. Alanyali, H. S. Moat, and T. Preis. "Quantifying the relationship between financial news and the stock market". In: *Scientific reports* 3 (2013), p. 3578.

[14] P. S. Kalev and H. N. Duong. "Firm-specific news arrival and the volatility of intraday stock index and futures returns". In: *The Handbook of News Analytics in Finance*. John Wiley & Sons, Ltd., 2011, pp. 271–288. ISBN: 9781118467411.

[15] A. J. Patton and M. Verardo. “Does beta move with news? Firm-specific information flows and learning about profitability”. In: *The Review of Financial Studies* 25.9 (2012), pp. 2789–2839.

[16] L. Mitra, G. Mitra, and Dibartolomeo¶. “Equity portfolio risk estimation using market information and sentiment”. In: ().

[17] G. Mitra, D. di Bartolomeo, A. Banerjee, and X. Yu. “Automated Analysis of News to Compute Market Sentiment: Its Impact on Liquidity and Trading”. In: *Available at SSRN 2605049* (2015).

[18] S. R. Das. “News Analytics: Framework, Techniques and Metrics”. In: *The Handbook of News Analytics in Finance*. Wiley Finance, 2010.

[19] I. Zheludev, R. Smith, and T. Aste. “When Can Social Media Lead Financial Markets?” In: *Scientific Reports* 4 (Feb. 2014).

[20] G. Mitra and L. Mitra. *The Handbook of News Analytics in Finance*. The Wiley Finance Series. Wiley, 2011.

[21] T. Preis, H. S. Moat, and H. E. Stanley. “Quantifying trading behavior in financial markets using Google Trends”. In: *Scientific Reports* Volume 3.Article number 1684 (2013), Article number 1684.

[22] T. Preis, D. Reith, and H. E. Stanley. “Complex dynamics of our economic life on different scales : insights from search engine query data”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* Vol.368.No.1933 (2010), pp. 5707–5719.

[23] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis. “Quantifying wikipedia usage patterns before stock market moves”. In: *Scientific Reports* Volume 3 (2013), Article number 1801.

[24] S. Y. Yang, S. Y. K. Mo, and X. Zhu. “An empirical study of the financial community network on Twitter”. In: *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*. IEEE. 2014, pp. 55–62.

[25] X. Zhang, H. Fuehres, and P. A. Gloor. "Predicting Stock Market Indicators Through Twitter 'I hope it is not as bad as I fear'". In: *Procedia - Social and Behavioral Sciences* 26.0 (2011). The 2nd Collaborative Innovation Networks Conference - {COINs2010}, pp. 55 –62. ISSN: 1877-0428. DOI: http://dx.doi.org/10.1016/j.sbspro.2011.10.562. URL: http://www.sciencedirect.com/science/article/pii/S1877042811023895.

[26] H. Mao, S. Counts, and J. Bollen. "Quantifying the effects of online bullishness on international financial markets". In: *European Central Bank Workshop on Using Big Data for Forecasting and Statistics, Frankfurt, Germany* (2014).

[27] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič. "The Effects of Twitter Sentiment on Stock Price Returns". In: *PLoS ONE* 10.9 (Sept. 2015), e0138441. DOI: 10.1371/journal.pone.0138441. URL: http://dx.doi.org/10.1371\%2Fjournal.pone.0138441.

[28] A. Funds. *BUZ ETF.* http://www.alpsfunds.com/overview/BUZ. Last accessed on Oct 14, 2018. 2016.

[29] Nasdaq. *Nasdaq Analytics Hub. Twitter Sentiment.* https://bit.ly/2zhX1LY. Last accessed on Oct 14, 2018. 2017.

[30] Bloomberg. *Bloomberg Launches a Twitter Feed Optimized for Trading.* https://bloom.bg/2LmXW4L. Last accessed on Oct 14, 2018. 2018.

[31] C. Brooks. "Testing for non-linearity in daily sterling exchange rates". In: *Applied Financial Economics* 6.4 (1996), pp. 307–317. DOI: 10.1080/09603109633410 5. eprint: http://dx.doi.org/10.1080/096031096334105. URL: http://dx.doi.org/10.1080/096031096334105.

[32] D. G. Luenberger. *Investment Science*. Oxford University Press, 2014. ISBN: 9780199740086. URL: https://books.google.co.uk/books?id=YMSeDAEACAAJ.

[33] J. Y. Campbell, J. W. Campbell, A. W. C. Lo, A. C. MacKinlay, J. J. Champbell, A. A. LO, A. C.M. K. MacKinlay, P. A. W. Lo, and O. E.P.A.E.J. Y. Campbell. *The Econometrics of Financial Markets*. Princeton University Press, 1997. ISBN: 9780691043012. URL: https://books.google.co.uk/books?id=lkeKhnqUHx8C.

[34] Ravenpack. *Ravenpack official website.* http://www.ravenpack.com/. Last accessed on Oct 14, 2018. 2018.

[35] O. Kolchyna, T. T. P. Souza, P. Treleaven, and T. Aste. "Twitter sentiment analysis: Lexicon method, machine learning method and their combination." In: *Handbook of Sentiment Analysis in Finance. ISBN 1910571571* (2016). eprint: arXiv:1507.00955. URL: http://arxiv.org/abs/1507.00955.

[36] O. Kolchyna, T. T. P. Souza, P. C. Treleaven, and T. Aste. "A framework for Twitter events detection, differentiation and its application for retail brands". In: *2016 Future Technologies Conference (FTC)*. 2016, pp. 323–331. DOI: 10.1109/FTC.2016.7821630.

[37] PsychSignal. *The PsychSignal website*. https://www.psychsignal.com. Last accessed on Oct 14, 2018. 2018.

[38] J. Manfield, D. Lukacsko, and T. T. P. Souza. "Bull bear balance: A cluster analysis of socially informed financial volatility". In: *2017 Computing Conference*. 2017, pp. 421–428. DOI: 10.1109/SAI.2017.8252134.

[39] T. T. P. Souza and T. Aste. "A nonlinear impact: evidences of causal effects of social media on market prices". In: *ArXiv e-prints. http://arxiv.org/abs/1601.04535* (Jan. 2016). eprint: 1601.04535.

[40] N. Wiener. "The theory of prediction". In: *Modern mathematics for engineers*. Ed. by E. F. Beckenbach. McGraw-Hill, New York, 1956. Chap. 8.

[41] C. Granger. “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods”. In: *Econometrica* 37.3 (1969), pp. 424–38.

[42] P.-O. Amblard and O. J. J. Michel. “The Relation between Granger Causality and Directed Information Theory: A Review”. In: *Entropy* 15.1 (2013), pp. 113–143. ISSN: 1099-4300. DOI: 10.3390/e15010113. URL: http://www.mdpi.com/1099-4300/15/1/113.

[43] A. Zaremba and T. Aste. “Measures of Causality in Complex Datasets with Application to Financial Data”. In: *Entropy* 16.4 (2014), pp. 2309–2349. ISSN: 1099-4300. DOI: 10.3390/e16042309. URL: http://www.mdpi.com/1099-4300/16/4/2309.

[44] W. A. Brock, J. A. Scheinkman, W. D. Dechert, and B. LeBaron. “A test for independence based on the correlation dimension”. In: *Econometric Reviews* 15.3 (Jan. 1996), pp. 197–235. DOI: 10.1080/07474939608800353. URL: http://dx.doi.org/10.1080/07474939608800353.

[45] W. A. Barnett, A. R. Gallant, M. J. Hinich, J. A. Jungeilges, D. T. Kaplan, and M. J. Jensen. “A Single-Blind Controlled Competition Among Tests for Nonlinearity and Chaos”. In: *Journal of Econometrics* 82 (1997), pp. 157–192.

[46] T. Schreiber. “Measuring Information Transfer”. In: *Phys. Rev. Lett.* 85 (2 2000), pp. 461–464. DOI: 10.1103/PhysRevLett.85.461. URL: http://link.aps.org/doi/10.1103/PhysRevLett.85.461.

[47] K. Hlavackovaschindler, M. Palus, M. Vejmelka, and J. Bhattacharya. “Causality detection based on information-theoretic approaches in time series analysis”. In: *Physics Reports* 441.1 (Mar. 2007), pp. 1–46. ISSN: 03701573. DOI: 10.1016/j.physrep.2006.12.004. URL: http://dx.doi.org/10.1016/j.physrep.2006.12.004.

[48] B. W. Silverman and P. J. Green. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[49] S. J. Sheather and M. C. Jones. "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 53.3 (1991), pp. 683–690. ISSN: 00359246.

[50] D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. A Wiley-interscience publication. Wiley, 1992. ISBN: 9780471547709.

[51] L. Barnett, A. B. Barrett, and A. K. Seth. "Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables". In: *Phys. Rev. Lett.* 103 (23 2009), p. 238701. DOI: `10.1103/PhysRevLett.103.238701`. URL: `http://link.aps.org/doi/10.1103/PhysRevLett.103.238701`.

[52] J. V. Michalowicz, J. M. Nichols, and F. Bucholtz. *Handbook of Differential Entropy*. Chapman & Hall/CRC, 2013, p. 194. ISBN: 1466583169, 9781466583160.

[53] A. Montalto, L. Faes, and D. Marinazzo. "MuTE: A MATLAB Toolbox to Compare Established and Novel Estimators of the Multivariate Transfer Entropy". In: *PLoS ONE* 9.10 (Oct. 2014), e109462. DOI: `10.1371/journal.pone.0109462`. URL: `http://dx.doi.org/10.1371\%2Fjournal.pone.0109462`.

[54] D. J. Leinweber and T. R. Aronson. *Nerds on Wall Street: Math, Machines and Wired Markets*. Wiley, 2009. ISBN: 9780470500569.

[55] B. Liu. "Sentiment analysis and opinion mining". In: *Synthesis lectures on human language technologies* 5.1 (2012), pp. 1–167.

[56] D. Antenucci, M. Cafarella, M. Levenstein, C. R, and M. D. Shapiro. *Using Social Media to Measure Labor Market Flows*. Working Paper 20010. National Bureau of Economic Research, 2014. DOI: `10.3386/w20010`. URL: `http://www.nber.org/papers/w20010`.

[57] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986. ISBN: 0070544840.

[58] S. R. Das et al. "Yahoo! for amazon: Sentiment extraction from small talk on the web". In: *8th Asia Pacific Finance Association Annual Conference*. 2001.

[59] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. "Mining Product Reputations on the Web". In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '02. Edmonton, Alberta, Canada: ACM, 2002, pp. 341–349. ISBN: 1-58113-567-X. DOI: 10.1145/775047.775098. URL: http://doi.acm.org/10.1145/775047.775098.

[60] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques". In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86. DOI: 10.3115/1118693.1118704. URL: https://doi.org/10.3115/1118693.1118704.

[61] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. "Inductive Learning Algorithms and Representations for Text Categorization". In: *Proceedings of the Seventh International Conference on Information and Knowledge Management*. CIKM '98. Bethesda, Maryland, USA: ACM, 1998, pp. 148–155. ISBN: 1-58113-061-9. DOI: 10.1145/288627.288651. URL: http://doi.acm.org/10.1145/288627.288651.

[62] S. M. Weiss, C. Apte, F. J. Damerau, D. E. Johnson, F. J. Oles, T. Goetz, and T. Hampp. "Maximizing text-mining performance". In: *IEEE Intelligent Systems and their Applications* 14.4 (1999), pp. 63–69. ISSN: 1094-7167. DOI: 10.1109/5254.784086.

[63] D. M. Cutler, J. M. Poterba, and L. H. Summers. "What moves stock prices?" In: *The Journal of Portfolio Management* 15.3 (1989), pp. 4–12.

[64] A. Moniz, G. Brar, and C. Davies. "Have I got news for you." In: *MacQuarie Research Report* (2009).

[65] R. Cahan, J. Jussa, and Y. Luo. "Breaking news: how to use sentiment to pick stocks." In: *MacQuarie Research Report* (2009).

[66] A. Patton and M. Verardo. "Does Beta Move with News? Firm-Specific Information Flows and Learning about Profitability". In: *Review of Financial Studies* 25.9 (2012), pp. 2789–2839.

[67] A. Banerjee, S. Paul, S. Hazra, and R. Dalmia. *Impact of information arrival on volatility of intraday stock returns*. Working paper series : WPS / Indian Institute of Management Calcutta. Calcutta : IIMC, 2011.

[68] P. Date, S. P. Sidorov, and V. Balash. "GARCH Type Volatility Models Augmented with News Intensity Data". In: *Chaos, Complexity and Leadership 2012*. Springer Proceedings in Complexity 2014. Springer Netherlands, 2014, pp. 199–207. ISBN: 978-94-007-7361-5.

[69] Z. A. Sadik, P. M. Date, and G. Mitra. "News augmented GARCH(1,1) model for volatility prediction". In: *IMA Journal of Management Mathematics* (2018), dpy004. DOI: 10.1093/imaman/dpy004. eprint: /oup/backfile/content_public/journal/imaman/pap/10.1093_imaman_dpy004/3/dpy004.pdf. URL: http://dx.doi.org/10.1093/imaman/dpy004.

[70] L. FANG and J. PERESS. "Media Coverage and the Cross-section of Stock Returns". In: *The Journal of Finance* 64.5 (), pp. 2023–2052. DOI: 10.1111/j.1540-6261.2009.01493.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2009.01493.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2009.01493.x.

[71] G. Mitra, K. Yu, and X. Yu. “Impact of news on asset behaviour: return, volatility and liquidity in an intra-day setting”. In: *Available at SSRN: http://ssrn.com/abstract=2296855* (2013).

[72] D. Tsvetanov, J. Coakley, and N. Kellard. “Is news related to GDP growth a risk factor for commodity futures returns?” In: *Quantitative Finance* 16.12 (2016), pp. 1887–1899. DOI: 10.1080/14697688.2016.1211797. eprint: https://doi.org/10.1080/14697688.2016.1211797. URL: https://doi.org/10.1080/14697688.2016.1211797.

[73] J. Shen, M. Najand, F. Dong, and W. He. “News and social media emotions in the commodity market”. In: *Review of Behavioral Finance* 9.2 (2017), pp. 148–168.

[74] N. Gotthelf and M. W. Uhl. “News sentiment: A new yield curve factor”. In: *Journal of Behavioral Finance* 20.1 (2019), pp. 31–41.

[75] M. W. Uhl. “Emotions Matter: Sentiment and Momentum in Foreign Exchange”. In: *Journal of Behavioral Finance* 18.3 (2017), pp. 249–257. DOI: 10.1080/15427560.2017.1332061. eprint: https://doi.org/10.1080/15427560.2017.1332061. URL: https://doi.org/10.1080/15427560.2017.1332061.

[76] K.-Y. Ho, Y. Shi, and Z. Zhang. “Does news matter in Chinas foreign exchange market? Chinese RMB volatility and public information arrivals”. In: *International Review of Economics & Finance* 52 (2017), pp. 302–321.

[77] A. Gro-Klumann and N. Hautsch. “When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions”. In: *Journal of Empirical Finance* 18.2 (2011), pp. 321–340. URL: https://EconPapers.repec.org/RePEc:eee:empfin:v:18:y:2011:i:2:p:321-340.

[78] J. E. Engelberg, A. V. Reed, and M. C. Ringgenberg. "How are shorts informed?: Short sellers, news, and information processing". In: *Journal of Financial Economics* 105.2 (2012), pp. 260 –278. ISSN: 0304-405X. DOI: https://doi.org/10.1016/j.jfineco.2012.03.001. URL: http://www.sciencedirect.com/science/article/pii/S0304405X12000384.

[79] S. P. Sidorov, A. R. Faizliev, M. Levshunov, A. Chekmareva, A. Gudkov, and E. Korobov. "Graph-Based Clustering Approach for Economic and Financial Event Detection Using News Analytics Data". In: *Social Informatics*. Ed. by S. Staab, O. Koltsova, and D. I. Ignatov. Cham: Springer International Publishing, 2018, pp. 271–280. ISBN: 978-3-030-01159-8.

[80] SEC. *U.S. Securities and Exchange Commission. SEC Says Social Media OK for Company Announcements if Investors Are Alerted.* http://1.usa.gov/1zFxUPa. Last accessed on Jan 29, 2015. Apr. 2013.

[81] WSJ. *Wall Street Journal. False AP Twitter Message Sparks Stock-Market Selloff.* http://on.wsj.com/12ms85v. Last accessed on Jan 29, 2015. Apr. 2013.

[82] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grar, and I. Mozeti. "The Effects of Twitter Sentiment on Stock Price Returns". In: *PLOS ONE* 10.9 (Sept. 2015), pp. 1–21. DOI: 10.1371/journal.pone.0138441. URL: https://doi.org/10.1371/journal.pone.0138441.

[83] H. Markowitz. "Portfolio Selection". In: *The Journal of Finance* 7.1 (1952), pp. 77–91. ISSN: 00221082, 15406261. URL: http://www.jstor.org/stable/2975974.

[84] E. F. Fama. "Market efficiency, long-term returns, and behavioral finance1". In: *Journal of financial economics* 49.3 (1998), pp. 283–306.

[85] B. M. Rom and K. W. Ferguson. "Post-Modern Portfolio Theory Comes of Age". In: *The Journal of Investing* 2.4 (1993), pp. 27–33. ISSN: 1068-0896. DOI: 10.3905/joi.2.4.27. eprint: http://joi.

iijournals.com/content/2/4/27.full.pdf. URL: http://joi.iijournals.com/content/2/4/27.

[86] E. Wipplinger. “Philippe Jorion: Value at Risk-The New Benchmark for Managing Financial Risk”. In: *Financial Markets and Portfolio Management* 21.3 (2007), p. 397.

[87] D. diBartolomeo and S. Warrick. “Making covariance based portfolio risk models sensitive to the rate at which markets reflect new information”. In: *Linear Factor models*. Elsevier Finance, 2005.

[88] S. R. Das and J. Sisk. “Financial Communities”. In: *The Journal of Portfolio Management* 31.4 (2005), pp. 112–123.

[89] G. H. Creamer, Y. Ren, and J. Nickerson. “News, Corporate Network and Price Discovery”. In: *Workshop on Information in Networks (WIN)* (2011).

[90] G. G. Creamer, Y. Ren, and J. V. Nickerson. “Impact of dynamic corporate news networks on asset return and volatility”. In: *Social Computing (SocialCom), 2013 International Conference on*. IEEE. 2013, pp. 809–814.

[91] G. Box and G. Jenkins. *Time series analysis: forecasting and control*. Holden-Day series in time series analysis and digital processing. Holden-Day, 1976. ISBN: 9780816211043.

[92] W. Antweiler and M. Z. Frank. “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards”. In: *Journal of Finance* 59.3 (2004), pp. 1259–1294.

[93] M. Bardoscia, S. Battiston, F. Caccioli, and G. Caldarelli. “Pathways towards instability in financial networks”. In: *Nature Communications* 8 (2017), p. 14416.

[94] M. Tumminello, S. Miccich, F. Lillo, J. Piilo, and R. N. Mantegna. “Statistically Validated Networks in Bipartite Complex Systems”. In: *PLoS ONE* 6.3 (Mar. 2011), pp. 1–11. DOI: 10.1371/journal.pone.0017994. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0017994.

[95] R. N. Mantegna. “Hierarchical structure in financial markets”. In: *The European Physical Journal B - Condensed Matter and Complex Systems* 11.1 (1999), pp. 193–197. ISSN: 1434-6036. DOI: 10 . 1007 / s100510050929. URL: http : / / dx . doi . org / 10 . 1007 / s100510050929.

[96] T. Aste, W. Shaw, and T. Di Matteo. “Correlation structure and dynamics in volatile markets”. In: *New Journal of Physics* 12.8 (2010), p. 085009.

[97] M. Tumminello, F. Lillo, and R. N. Mantegna. “Correlation, hierarchies, and networks in financial markets”. In: *Journal of Economic Behavior & Organization* 75.1 (2010). Transdisciplinary Perspectives on Economic Complexity, pp. 40 –58. ISSN: 0167-2681. DOI: http : / / dx . doi . org / 10 . 1016 / j . jebo . 2010 . 01 . 004. URL: http : / / www . sciencedirect . com / science / article / pii / S0167268110000077.

[98] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. “A tool for filtering information in complex systems”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.30 (2005), pp. 10421–10426. DOI: 10.1073/pnas.0500298102. eprint: http://www.pnas.org/content/102/30/10421.full.pdf. URL: http://www.pnas.org/content/102/30/10421.abstract.

[99] W.-M. Song, T. Aste, and T. Di Matteo. “Analysis on filtered correlation graph for information extraction”. In: *Statistical Mechanics of Molecular Biophysics* (2008), p. 88.

[100] Y. Li, X.-F. Jiang, Y. Tian, S.-P. Li, and B. Zheng. “Portfolio optimization based on network topology”. In: *Physica A: Statistical Mechanics and its Applications* (2018). ISSN: 0378-4371. DOI: https : / / doi . org / 10 . 1016 / j . physa . 2018 . 10 . 014. URL: http : / / www . sciencedirect . com / science / article / pii / S0378437118313529.

[101] F. Pozzi, T. Di Matteo, and T. Aste. “Spread of risk across financial markets: better to invest in the peripheries”. In: *Scientific reports* 3 (2013), p. 1665.

[102] R. Morales, T. Di Matteo, R. Gramatica, and T. Aste. “Dynamical generalized Hurst exponent as a tool to monitor unstable periods in financial time series”. In: *Physica A: Statistical Mechanics and its Applications* 391.11 (2012), pp. 3180–3189.

[103] N. Musmeci, T. Aste, and T. di Matteo. “Clustering and hierarchy of financial markets data: advantages of the DBHT.” In: *CoRR* (2014).

[104] W.-M. Song, T. Di Matteo, and T. Aste. “Hierarchical information clustering by means of topologically embedded graphs”. In: *PLoS One* 7.3 (2012), e31929.

[105] C. Curme, H. E. Stanley, and I. Vodenska. “Coupled Network Approach To Predictability Of Financial Market Returns And News Sentiments”. In: *International Journal of Theoretical and Applied Finance* 18.07 (2015), p. 1550043.

[106] T. T. P. Souza, O. Kolchyna, P. Treleaven, and T. Aste. “Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry”. In: *Handbook of Sentiment Analysis in Finance*. Ed. by G. Mitra and X. Yu. 2016. Chap. 23.

[107] W. F.M. D. Bondt and R. Thaler. “Does the Stock Market Overreact?” In: *The Journal of Finance* 40.3 (1985), pp. 793–805. ISSN: 00221082. URL: http://www.jstor.org/stable/2327804.

[108] J. P. Onnela, A. Chakraborti, K. Kaski, J. Kertsz, and A. Kanto. “Asset Trees and Asset Graphs in Financial Markets”. In: *Physica Scripta* 2003.T106 (2003), p. 48. URL: http://stacks.iop.org/1402-4896/2003/i=T106/a=011.

[109] J. P. Onnela, K. Kaski, and J. Kertész. “Clustering and information in correlation based financial networks”. In: *The European Physical Journal B* 38.2 (2004), pp. 353–362. ISSN: 1434-6036. DOI: 10.1140/epjb/e2004-

00128-7. URL: https://doi.org/10.1140/epjb/e2004-00128-7.

[110] F. Battiston, V. Nicosia, and V. Latora. "The new challenges of multiplex networks: Measures and models". In: *The European Physical Journal Special Topics* 226.3 (2017), pp. 401–416. ISSN: 1951-6401. DOI: 10.1140/epjst/e2016-60274-8. URL: http://dx.doi.org/10.1140/epjst/e2016-60274-8.

[111] E. Cozzo, M. Kivel, M. D. Domenico, A. Sol-Ribalta, A. Arenas, S. Gmez, M. A. Porter, and Y. Moreno. "Structure of triadic relations in multiplex networks". In: *New Journal of Physics* 17.7 (2015), p. 073029. URL: http://stacks.iop.org/1367-2630/17/i=7/a=073029.

[112] J. J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press, 2006.

[113] A.-L. Barabási and M. Pósfai. *Network science*. Cambridge University Press, 2016. ISBN: 9781107076266 1107076269. URL: http://barabasi.com/networksciencebook/.