Coupling news sentiment with web browsing data predicts intra-day stock prices

Gabriele Ranco \textsuperscript{*} *, Ilaria Bordino \textsuperscript{1} , Giacomo Bormetti \textsuperscript{‡} §, Guido Caldarelli \textsuperscript{*} § †, Fabrizio Lillo \textsuperscript{‡} † † † || Michele Treccani \textsuperscript{‡‡}

\textsuperscript{*}IMT Institute for Advanced Studies, Piazza San Francesco 19, 56100 Lucca, Italy, \textsuperscript{1} Yahoo Labs, Barcelona, Spain, \textsuperscript{2} Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy, \textsuperscript{3} QUANTLab, Via Pietrasantina 123, 56122 Pisa, Italy, \textsuperscript{4} ISC-CNR, Via dei Taurini 19, 00185 Roma, Italy, \textsuperscript{5} London Institute for Mathematical Science, South St. 35 Mayfair, London W1K 2XF, UK, \textsuperscript{6} Dipartimento di Fisica e Chimica, Università degli Studi di Palermo, Viale delle Scienze Ed. 18, 90128 Palermo, Italy, \textsuperscript{7} Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA, and \textsuperscript{8} MedioBanca S.p.A, Piazzetta E. Cuccia 1, 20121 Milan, Italy

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The new digital revolution of big data is deeply changing our capability of understanding society and forecasting the outcome of many social and economic systems. Unfortunately, information can be very heterogeneous in the importance, relevance, and surprise it conveys, affecting severely the predictive power of semantic and statistical methods. Here we show that the aggregation of web users’ behavior can be elicited to overcome this problem in a hard to predict complex system, namely the financial market. Specifically, we show that the combined use of sentiment analysis of news and browsing activity of users of Yahoo! Finance allows to forecast intra-day and daily price changes of a set of 100 highly capitalized US stocks traded in the period 2012-2013. Sentiment analysis or browsing activity when taken alone have very small or no predictive power. Conversely, when considering a news signal where in a given time interval we compute the average sentiment of the clicked news, weighted by the number of clicks, we show that for more the 50% of the companies such signal Granger-causes price returns. Our result indicates a “wisdom-of-the-crowd” effect that allows to exploit users’ activity to identify and weigh properly the relevant and surprising news, enhancing considerably the forecasting power of the news sentiment.

The recent technological revolution with widespread presence of computers, users and media connected by Internet has created an unprecedented situation of data deluge, changing dramatically the way in which we look at social and economic sciences. As people increasingly use the Internet for information such as business or political news, online activity has become a mirror of the collective consciousness, reflecting the interests, concerns, and intentions of the global population with respect to various economic, political, and cultural phenomena. Humans’ interactions with technological systems are generating massive datasets documenting collective behaviour in a previously unimaginable fashion \cite{1,2}. By properly dealing with such data collections, for instance representing them by means of network structures \cite{3,4}, it is possible to extract relevant information about the evolution of the systems considered (i.e. trading \cite{5}, disease spreading \cite{6,7}, political elections \cite{8}).

A particularly interesting case of study is that of the financial markets. Markets can be seen as collective decision making systems, where exogenous (news) as well as endogenous (price movements) signals convey valuable information on the value of a company. Investors continuously monitor these signals in the attempt of forecasting future price movements. Because of their trading based on these signals, the information is incorporated into prices, as postulated by the Efficient Market Hypothesis \cite{9}. Therefore the flow of news and data on the activity of investors can be used to forecast price movements. The literature on the relation between news and price movement is quite old and vast. In order to correlate news and price returns one needs to assess whether the former is conveying positive or negative information about a company, a particular sector or on the whole market. This is typically done with the sentiment analysis, often performed with dedicated semantic algorithms as described and reviewed in the Methods Section.

In this paper, we combine the information coming from the sentiment conveyed by public news with the browsing activity of the users of a finance specialized portal to forecast price returns at daily and intra-day time scale. To this aim we leverage a unique dataset consisting of a fragment of the log of Yahoo! Finance, containing the news articles displayed on the web site and the respective number of “clicks”, i.e. the visualizations made by the users. Our analysis considers 100 highly capitalized US stocks in a one-year period between 2012 and 2013.

For each of these companies we build a signed time series of the sentiment expressed in the related news. The sentiment expressed in each article mentioning a company is weighted by the number of visualizations of the article. In our dataset each click action is associated with a timestamp recording the exact point in time when such action took place. Thus we are able to construct time series at the time resolution of the minute. To the best of our knowledge, this is the first time that an analysis like the one described in this paper is conducted at such intra-day granularity. The main idea behind this approach is that the sentiment analysis gives information
on the news, while the browsing volume enable us to properly weigh news according to the attention received from the users.

We find that news on the same company are extremely heterogeneous in the number of clicks they receive, an indication of the huge difference in their importance and the interest these news generate on users. For 70% of the companies examined, there is a significant correlation between the browsing volumes of financial news related to the company, and its traded volumes or absolute price returns. More important, we show that for more than 50% of the companies (at hourly time scale), and for almost 40% (at daily time scale), the click weighted average sentiment time series Granger-cause price returns, indicating a rather large degree of predictability.

**Data**

**Stocks considered.** Our analysis is conducted on 100 highly capitalized stocks traded in the US equity markets, which we monitor during a period of one year between 2012 and 2013. The ticker list of the investigated stocks with a distinctive numerical company identifier follows: 1 KBH, 2 LEN, 3 CRZ, 4 DTV, 5 AMGN, 6 XOM, 7 UPS, 8 V, 9 VZ, 10 GRPN, 11 ZNGA, 12 ABT, 13 LUV, 14 RTN, 15 HAL, 16 ATVI, 17 MRK, 18 GPS, 19 GILD, 20 LCC, 21 NKE, 22 MCD, 23 UNH, 24 DOW, 25 M, 26 CBS, 27 COP, 28 CHK, 29 CAT, 30 HON, 31 TWX, 32 AIG, 33 UAL, 34 TXN, 35 BIIB, 36 WAG, 37 PEP, 38 VMW, 39 KO, 40 QCOM, 41 ACN, 42 NOC, 43 DISH, 44 BBY, 45 HD, 46 FG, 47 JNJ, 48 AXP, 49 MAR, 50 TWC, 51 UTX, 52 MA, 53 BLK, 54 EBAY, 55 DAL, 56 NWSA, 57 MSCI, 58 LNKD, 59 TSLA, 60 CVX, 61 AA, 62 NYX, 63 JCP, 64 CMCSA, 65 NDAQ, 66 IT, 67 YHOO, 68 DIS, 69 SBUX, 70 PFE, 71 ORCL, 72 HPQ, 73 S, 74 LMT, 75 XOM, 76 IBM, 77 NFLX, 78 INTC, 79 CSCO, 80 GE, 81 WFC, 82 WMT, 83 AMZN, 84 VOD, 85 DELL, 86 F, 87 TRI, 88 GM, 89 FRT, 90 VZ, 91 FB, 92 BAC, 93 MS, 94 JPM, 95 C, 96 BA, 97 GS, 98 MSFT, 99 GOOG, 100 AAPL. The numerical identifiers are assigned according to the increasing order of the total number of published news in Yahoo! Finance. We considered three main sources of data for these stocks:

**Market data.** The first source contains information on price returns and trading volume of the stock at the resolution of the minute. We consider different time scales of investigation, corresponding to 1, 10, 30, 65, and 130 minutes. The above values are chosen because they are sub-multiple of the trading minute. We consider different time scales of investigation, corresponding to 1, 10, 30, 65, and 130 minutes. The above values are chosen because they are sub-multiple of the trading minute.

Finally, in our analysis we use the information on the browsing volume, that is, the time series of “clicks” that the web users made on each article displayed on Yahoo! Finance to visualize its content. Given that the users’ activity on this domain-specific portal proved to provide a clean signal of interest in financial stocks [16], we exploit it in this work to weight the sentiment of each article on a given financial company. Specifically, we use the number of clicks of an article as a proxy for the level of attention that users gave to that news. By aggregating over a time window the clicks on all the articles, even published earlier, that mention a particular company, it is possible to derive an estimation of the attention around that company.

In summary, for each time scale and for each stock, the variables we extract from the database are (see Support Information):

- $C$, the time series of the total number of clicks in a time window,
- $S$, the sum of the sentiment of all news related to each company,
- $W/S$, the sum of the sentiment of all news weighted by the number of clicks.

The first quantity $C$ is non negative and measures the level of attention in a given time interval for news about a specific company. The $S$ quantity is the usual (in absolute terms) measure employed in numerous studies and provides the aggregated sentiment of the company specific news published in a given time interval. The most important and novel quantity is $W/S$, which combines the two previous ones by assigning a sign to each click depending on the sentiment of the clicked news. As for the market variables, we remove the intra-day pattern from the click time series. In fact, both the publication of news [15] and the clicking activity of users [16] show a strong intra-day seasonality. These patterns are probably related to

Yahoo! Finance is the most popular web portal for news and data related to financial companies, offering news and information around stock quotes, stock exchange rates, corporate press releases, financial reports, and message boards for discussion. We analyze a portion of the web-site log, containing news about the portal. The articles are manually tagged by a team of trained domain experts, who associate each article with the specific companies (e.g., Google, Yahoo!, Apple, Microsoft) or financial entities (e.g., market indexes, commodities, derivatives) that are mentioned in its text.
the way humans carry out their activities during the day (e.g., small activity during the lunch time, more hectic activity at the beginning or at the end of the business day).

Methods

Sentiment Analysis. Regarding the analysis of search-engine queries, some recent works [17, 18, 19, 20, 21] have studied the relation between the daily number of queries related to a particular stock and the trading volume over the same stock.

An incomplete list of contributions on the role of news includes studies investigating (i) the relation between exogenous news and price movements [12, 22, 23, 24], (ii) the correlation between high or low pessimism of media and high market trading volume [13]; (iii) the relation between the sentiment of news, earnings and return predictability [25, 26], (iv) the role of news in the trading action [15, 27, 28, 29]; (v) the role of macroeconomic news in the performance of stock returns [30], and (vi) the high-frequency market reaction to news [31].

For example, in a recent paper [18], related to ours, authors show that daily trading volumes of stocks traded at NASDAQ can be forecasted with the daily volumes of queries related to the same stocks. In another paper a similar analysis shows that an increase in queries predicts higher stock prices in the next two weeks [32]. As for social networks and micro-blogging platforms [33], Twitter data is becoming an increasingly popular choice for financial forecasting. For example some have investigated whether the daily number of tweets predicts SP 500 stock prices [34, 35]. A textual analysis approach to Twitter data can be found in other works [14, 36, 37, 38] where the authors find clear relations between mood indicators and Dow Jones Industrial Average. Some other authors have used news, Wikipedia data or search trends to predict market movements [22, 39, 40, 41].

There are two main critical aspects in the kind of analyses described above. First the unfiltered news or social network users is probably too large and the fraction of users truly interested in finance is likely quite low, with the consequence that signals based on generic Internet users are very noisy. Second, as we will empirically show below, the universe of news considered is very heterogeneous in terms of their relevance as a signal of future price movement. For example, one can forecast financial variables, and more specifically price returns, by using the information on the browsing activity of the users. Namely if, by weighting the sentiment of the clicked news by the number of clicks each news receives, one can improve significantly the predictability of returns.

Synchronous correlation. In order to understand how the relation between financial and news variables depends on the time scale, we perform a synchronous correlation analysis. We remark that some variables (C, σ, V) are non negative, while others (R, S, WS) are signed. Not surprisingly thus, we find that the linear correlation between any one of the former and anyone of the latter is always very close to zero. For each of the 100 companies, we compute the Pearson’s correlation coefficient ρ between the three sensible pairs made by one “news” time series and one “financial” time series. Figure 2 summarizes the results for the 65 min time series. The x axis lists the companies, uniquely identified by a number that provides the rank of the company in the order from the least to the most cited one (as measured by the absolute number of associated news). Thus, 1 corresponds to the company KBH with the least number of news, while 100 to the most cited AAPL. We label the y axis with the pairs (C, V), (C, σ), and (WS, R), while the color scale indicates the level of correlation. We compute the correlation sampling the original time series every 65 minutes, equalizing to zero those values whose significance does not reject the null hypothesis of zero correlation with a significance level of 5%.

### Footline Author

PNAS | Issue Date | Volume | Issue Number

[3]
Time scale. In order to investigate how the correlation changes with the time scale, in Table 1 we also show the percentage of rejection for the 1, 10, 30, and 130 minutes time scales. As a general comment, we observe that the number of companies with a significant correlation becomes higher at finer time resolution. This is a known fact for market variables (e.g. volume and volatility), while we document it for the first time at intraday scale also for sentiment and browsing variables. As expected the correlations are stronger for unsigned rather than for signed variables. Nevertheless, we can state the presence of a significant relation between the attention given to news articles (signed on the basis of the sentiment expressed in them) related to a given stock and the evolution of the stock price. Remarkably, in the case of the sentiment time series a progressive increase of the number of companies can be observed, until we reach a hourly granularity, suggesting that this could be a characteristic time scale of the process.

Dynamics of attention. The time scale might in principle depend on the relevance of the news. As we have seen, not all news are equal in terms of the attention they receive from the users. To investigate this dependence, we study the dynamics of the number of clicks an article received after its publication. We compute the cumulative number of clicks received by a given news until a given minute after the publication. We perform this for all minutes in a week after the publication.

We then normalize this cumulative time series by dividing it by the total amount of clicks received by the news. We construct ten groups of news based on the deciles of the total number of news they eventually receive, and we compute for each group the average cumulative sum of clicks. The result is shown in Figure 3. The inset reports for each decile the typical time scale of attention obtained by an exponential fit of the curves. Remarkably, the time scale of attention is an increasing function of the importance of the news (as measured by the total number of clicks). Irrelevant news are immediately recognized as such, while important news continue to receive attention well after their publication. In general, the time scale of the users’ attention ranges between one and two hours after the publication, suggesting that this intraday time scale is probably the most appropriate to detect dependencies among financial variables and browsing activity.

Causality. The synchronous correlation is an important measure of dependence, but not necessarily a sign of causality. Thus we perform a causality analysis by applying Granger’s test. We present the results of this analysis, for the 65-minute time horizon, in Figure 4. The x axis lists the companies as done in Figure 2, while the y axis labels the eight tests that we perform. Black cells correspond to rejection of the null hypothesis of no Granger causality, and the opposite for the white cells. When considering the non-negative variables (V, C, and σ) we observe strong causal relations. Specifically, in 65% of the cases the clicking activity causes the trading volume and in 69% of the cases it causes price volatility. The causality is very strong also in the opposite direction, i.e. volume and volatility cause click volume. This is probably due in part to a reaction of users to anomalously high activity in the market (in terms of volume and/or volatility), while in part it might be a statistical effect due to the fact that all the three variables are very autocorrelated in time, creating strong Granger causal relations in both directions.

We obtain the most interesting and unexpected results when considering the signed variables (R, S, and WS). All these variables are weakly serially autocorrelated. When we consider the sentiment of the news (S, without the clicks), we find that only in 4% of the cases S causes returns, and in 13% of the cases price return causes S. Especially the first value is expected under the null, since at 5% confidence level we expect 5% of false positive. This means that the simple sentiment of the news does not allow to forecast price returns at intraday (hourly) time scale. On the contrary, when we consider the clicks weighted by the sentiment of the news (WS), we find that in 53% of the cases it allows predicting returns and only in 19% of the cases the opposite occurs. In general, companies with more news have higher causality.

Weighting news by users’ browsing behavior. These results show that, on a hourly time scale, the simple news sentiment time series alone (i.e. the one without browsing activity) is not able to predict the price returns; instead, if we add the information provided by the browsing activity, we are then able to properly weigh the news (and its sentiment) by the importance the users give to it by clicking the page. Thus, we find the interesting result that the browsing activity combined with the sentiment analysis of the news increases significantly the level of prediction of price returns for the stocks.

Comparison with existing literature. Most of the existing studies on sentiment and predictability of returns focus on daily or longer time scale. In order to compare properly our result with the existing literature, we present in Table 2 the results of the above Granger tests on a daily time scale. Table 2 shows that, without the browsing activity, S causes R for 18% of the companies and R Granger-causes S in 9% of the cases. Thus there is now some predictability of sentiment, even if the number of companies is quite limited. This is consistent with the existing literature, which reports a weak daily predictability of returns by using sentiment. It is important to note that by adding the browsing activity we can double the number of companies for which there is predictability. In fact, WS Granger-causes R for 37% of the companies and 11% in the opposite direction.

Discussion. The semantic analysis of the news on a specific company is known to have a small predictive power on the future price movements. At the light of our findings, we argue that this effect could be related to the distribution in the attention that the news receive, as clearly emerge in Figure 1: its scale-free behaviour reflects the extreme heterogeneity in the information they convey and the surprise they generate in the readers.

We show that by adding the clicking activity of the web users, we can increase dramatically the predictive power of the news for the price returns. This occurs because the time series built with only the sentiment of the news gives the same weight to all the news. In this way even irrelevant news are considered, adding noise to the sentiment time series and reducing the predictive power of the signal. Adding the browsing activity means giving a meaningful weight to each news according to its importance, as measured by the attention it receives by the users, and this largely improves the predictability of returns.

The approach to collective evaluation that we proposed in this paper can be useful in many other non financial contexts, since the overflow of information is a common aspect in our lives. In the financial domain, a natural extension of the present work concerns market instabilities and crashes. The analysis presented here is in fact unconditional, i.e. it does not take into account large price movements or, more generally, abnormal returns. From our societal perspective it would be extremely valuable to have a collective evaluation system, like
the one presented here, capable of sifting the relevant information from the pool of data, news, blogs, etc. and to provide early warning indicators of large price movements. Since we have shown that predictability exists also at intraday time scale, this approach could be also used for real-time indicators, as well as for high-frequency instabilities and systemic price cojumps [44], which are becoming increasingly more frequent in our highly automated financial world.

ACKNOWLEDGMENTS. The authors warmly thank Lucio Calcagnile for the valuable technical support during the final stage of this work. GR and GC acknowledge support from the FP7/2007-2013 under grant agreement CRISIS-ICT-CNR “CrisisLab” and partial support from the U.S. Department of the Defense, Defense Threat Reduction Agency, grant HDTRA1-11-1-0048. FL acknowledges support from the FP7/2007-2013 under grant agreement CRISS-ICT-2011-288001. All the authors developed the concept of the study. GR conducted the analysis, and the statistical methods needed for data interpretation. Data have been made available by Yahoo! Inc and Quantlab groups.

The opinions expressed here are solely those of the authors and do not represent in any way those of their institutions or employers.

10. Loughran, T. & McDonald, B. When is a liability not a liability? Textual analysis, and the statistical methods needed for data interpretation. Data have been made available by Yahoo! Inc and Quantlab groups. The opinions expressed here are solely those of the authors and do not represent in any way those of their institutions or employers.

Footline Author

PNAS | Issue Date | Volume | Issue Number | 5
**Fig. 1.** Complementary of the cumulative distribution function of the number of clicks a news receives for the ten assets with the largest number of news and the aggregate portfolio of 100 stocks. Both coordinates have been rescaled by a common factor preserving the power law scaling of the right tail and normalizing the maximum number of clicks to the value $10^{10}$. The dotted line corresponds to a power law with tail exponent fitted from the portfolio time series. We provide details about the standard error and the complete list of tail exponents for all the companies in Support Information.

**Fig. 2.** Pearson’s correlation coefficients for the de-seasonalized time series of all the 100 companies at hourly scale. The x axis reports the list of companies identified by a unique number, as detailed in the main text. Among the several possibilities, we consider only three couples and the color scale corresponds to the level of correlation. We plot those values for which we reject the null of zero correlation at 5% significance level and equalize non significant values to zero (light green color).

**Fig. 3.** Time evolution of the cumulative number of clicks per news in a time interval of five hours after the publication. We normalize the cumulated amount by a constant which corresponds to the total number of clicks received by a single news during the first week after publication. The news are grouped in deciles according to the total number of clicks they have received until October 2013 and the curves represent average values. Inset: estimated values and standard errors of the attention time scale obtained by an exponential fit of the decile curves.
Fig. 4. Granger Causality tests at hourly scale between de-seasonalized time series (x axis as in figure 2). The white cells correspond to tests for which we do not reject the null hypothesis of no Granger causality at 5% significance level. A black cell corresponds to a statistically significant Granger causality.

Table 1. Percentage of companies for which we reject the null hypothesis of zero correlation at 5% confidence level.

<table>
<thead>
<tr>
<th>Time interval (minutes)</th>
<th>$\rho(WS,R)$</th>
<th>$\rho(C,\sigma)$</th>
<th>$\rho(C,V)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>77</td>
<td>78</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>48</td>
<td>69</td>
</tr>
<tr>
<td>30</td>
<td>10</td>
<td>32</td>
<td>62</td>
</tr>
<tr>
<td>65</td>
<td>13</td>
<td>30</td>
<td>57</td>
</tr>
<tr>
<td>130</td>
<td>8</td>
<td>29</td>
<td>52</td>
</tr>
</tbody>
</table>

Table 2. Number of companies for which we reject the null hypothesis of no Granger causality at 5% confidence level.

<table>
<thead>
<tr>
<th>Causality relation</th>
<th>Hourly scale</th>
<th>Daily scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow R$</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>$R \rightarrow S$</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>$R \rightarrow WS$</td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>$WS \rightarrow R$</td>
<td>53</td>
<td>37</td>
</tr>
<tr>
<td>$V \rightarrow C$</td>
<td>100</td>
<td>97</td>
</tr>
<tr>
<td>$C \rightarrow V$</td>
<td>65</td>
<td>52</td>
</tr>
<tr>
<td>$C \rightarrow \sigma$</td>
<td>69</td>
<td>52</td>
</tr>
<tr>
<td>$\sigma \rightarrow C$</td>
<td>96</td>
<td>16</td>
</tr>
</tbody>
</table>