Quantifying the semantics of search behavior before stock market moves

Chester Curme[a,b,1], Tobias Preis[b], H. Eugene Stanley[a,1], and Helen Susannah Moat[b]

[a]Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215; and [b]Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom

Contributed by H. Eugene Stanley, December 31, 2013 (sent for review August 6, 2013)

Technology is becoming deeply interwoven into the fabric of society. The Internet has become a central source of information for many people when making day-to-day decisions. Here, we present a method to mine the vast data Internet users create when searching for information online, to identify topics of interest before stock market moves. In an analysis of historic data from 2004 until 2012, we draw on records from the search engine Google and online encyclopedia Wikipedia as well as judgments from the service Amazon Mechanical Turk. We find evidence of links between Internet searches relating to politics or business and subsequent stock market moves. In particular, we find that an increase in search volume for these topics tends to precede stock market falls. We suggest that extensions of these analyses could offer insight into large-scale information flow before a range of real-world events.

complex systems | computational social science | data science | online data | financial markets

Financial crises arise from the complex interplay of decisions made by many individuals. Stock market data provide extremely detailed records of such decisions, and as such both these data and the complex networks that underlie them have generated considerable scientific attention (1–20). However, despite their gargantuan size, such datasets capture only the final action taken at the end of a decision-making process. No insight is provided into earlier stages of this process, where traders may gather information to determine what the consequences of various actions may be (21).

Nowadays, the Internet is a core information resource for humans worldwide, and much information gathering takes place online. For many, search engines such as Google act as a gateway to information on the Internet. Google, like other search engines, collects extensive data on the behavior of its users (22–25), and some of these data are made publicly available via its service Google Trends. These datasets catalog important aspects of human information gathering activities on a global scale and thereby open up new opportunities to investigate early stages of collective decision making.

In line with this suggestion, previous studies have shown that the volume of search engine queries for specific keywords can be linked to a range of real-world events (26), such as the popularity of films, games, and music on their release (27); unemployment rates (28); reports of flu infections (29); and trading volumes in US stock markets (30, 31). A recent study showed that Internet users from countries with a higher per capita gross domestic product (GDP), in comparison with Internet users from countries with a lower per capita GDP, search for proportionally more information about the future than information about the past (32).

Here, we investigate whether we can identify topics for which changes in online information-gathering behavior can be linked to the sign of subsequent stock market moves. A number of recent results suggest that online search behavior may measure the attention of investors to stocks before investing (33–35). We build on a recently introduced method (33) that uses trading strategies based on search volume data to identify online precursors for stock market moves. This previous analysis of search volume for 98 terms of varying financial relevance suggests that, at least in historic data, increases in search volume for financially relevant search terms tend to precede significant losses in financial markets (33). Similarly, Moat et al. (36) demonstrated a link between changes in the number of views of Wikipedia articles relating to financial topics and subsequent large stock market moves. The importance of the semantic content of these Wikipedia articles is emphasized by a parallel analysis that finds no such link for data from Wikipedia pages relating to actors and filmmakers.

Financial market systems are complex, however, and trading decisions are usually based on information about a huge variety of socioeconomic topics and societal events. The initial examples above (33, 36) focus on a narrow range of preidentified financially related topics. Instead of choosing topics for which search data should be retrieved and investigating whether links exist between the search data and financial market moves, here we present a method that allows us to identify topics for which levels of online interest change before large movements of the Standard & Poor's 500 index (S&P 500). Although we restrict ourselves to stock market moves in this study, our methodology can be readily extended to determine topics that Internet users search for before the emergence of other large-scale real-world events.

Our approach is as follows. First, we take a large online corpus, Wikipedia, and use a well-known technique from computational linguistics (37) to identify lists of words constituting semantic topics within this corpus. Second, to give each of these automatically identified topics a name, we engage users of the

Significance

Internet search data may offer new possibilities to improve forecasts of collective behavior, if we can identify which parts of these gigantic search datasets are relevant. We introduce an automated method that uses data from Google and Wikipedia to identify relevant topics in search data before large events. Using stock market moves as a case study, our method successfully identifies historical links between searches related to business and politics and subsequent stock market moves. We find that the predictive value of these search terms has recently diminished, potentially reflecting increasing incorporation of Internet data into automated trading strategies. We suggest that extensions of these analyses could help draw links between search data and a range of other collective actions.

Author contributions: C.C., T.P., H.E.S., and H.S.M. designed research; C.C., T.P., H.E.S., and H.S.M. performed research; C.C., T.P., H.E.S., and H.S.M. analyzed data; and C.C., T.P., H.E.S., and H.S.M. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: ccurme@bu.edu or hes@bu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1324054111/-/DCSupplemental.

online service Amazon Mechanical Turk. Third, we take lists of the most representative words of each of these topics and retrieve data on how frequently Google users searched for the terms over the past 9 y. Finally, we use the method introduced in ref. 33 to examine whether the search volume for each of these terms contains precursors of large stock market moves. We find that our method is capable of automatically identifying topics of interest before stock market moves and provide evidence that for complex events such as financial market movements valuable information may be contained in search engine data for keywords with less-obvious semantic connections.

Method

To extract semantic categories from the online encyclopedia Wikipedia, we build on a well-known observation (37) that words that frequently appear together in newspaper articles, encyclopedia entries, or other kinds of documents tend to bear semantic relationships to each other. For example, a document containing the word "debt" may be more likely to also contain other words relating to finance than other words relating to, say, fruit. For such an analysis of semantic relationships to produce meaningful results, the overall frequency of terms must also be taken into account. To incorporate these insights, we analyze the semantic characteristics of all of the articles and words in the English version of Wikipedia using a modeling approach called latent Dirichlet allocation (LDA) (37). We configure the LDA to extract 100 different semantic topics from Wikipedia and provide lists of the 30 most representative words for each topic. Lists of the topics and their constituent words are provided in Dataset S1. We note that individual words can occur in multiple semantic topics.

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

Google Trends provides data on search volume using a finite integer scale from 0 to 100, where 100 represents the highest recorded search volume for all terms in a given Google Trends request. If search volume time series for low-frequency keywords are downloaded in isolation from other keywords, noisy data can result, because only a small number of searches is required for a unit change in search volume to be registered. To avoid this problem, we download search volume data for the high-frequency term "google" alongside search volume data for each of our terms. In this way, we ensure that the value 100 represents the maximum search volume for this high-frequency term. However, we also find that the mean search volume for terms in 45 of our extracted topics is too low to register on this "google"-based scale, having a value less than 1. Below, we describe analyses based on the remaining 55 topics.

To generate labels for the topics, we make use of the online service Amazon Mechanical Turk. This service allows small tasks to be taken on by anonymous human workers, who receive a small payment for each task. Through this service, 39 unique human workers provided topic names for the 55 sets of words identified above. Both the full list of topic names obtained from Amazon Mechanical Turk and more details on this procedure are provided in *Supporting Information* and Dataset S1.

To compare changes in search volume to subsequent stock market moves, we implement for each of these terms the trading strategy introduced in ref. 33. We use for our analyses the US equities market index S&P 500, which includes 500 leading companies in leading industries of the US economy. We hypothetically trade the S&P 500 Total Return index (SPXT), which also accounts for the reinvestment of dividends. In this strategy, we first use Google Trends to measure how many searches $n(t)$ occurred for a chosen term in week t. To quantify changes in information-gathering behavior, we compute the relative change in search volume $\Delta n(t, \Delta t) = n(t) - N(t-1, \Delta t)$ with $N(t-1, \Delta t) = (n(t-1) + n(t-2) + \ldots + n(t-\Delta t))/\Delta t$. We sell the SPXT at the closing price $p(t)$ on the first trading day of week t, if $\Delta n(t-1, \Delta t) > 0$ and buy the index at price $p(t+1)$ at the end of the first trading day of the following week. If instead $\Delta n(t-1, \Delta t) < 0$, then we buy the index at the closing price $p(t)$ on the first trading day of week t and sell the index at price $p(t+1)$ at the end of the first trading day of the coming week. If we sell at the closing price $p(t)$ and buy at price $p(t+1)$, then the arithmetic cumulative return R changes by a factor of $p(t)/p(t+1)$. If we buy at the closing price $p(t)$ and sell at price $p(t+1)$, then the arithmetic cumulative return R changes by a factor of $p(t+1)/p(t)$. The maximum number of transactions per year when using our strategy is only 104, allowing a closing and an opening transaction per week; hence, for the purposes of this analysis of the relationship between search volume and stock market moves we neglect transaction fees.

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

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

Results

Fig. 1*A* depicts the distributions of R for each of the 55 topics. We compare the arithmetic cumulative returns for search volume-based strategies to the distribution of arithmetic cumulative returns from the random strategy using two-sample Wilcoxon rank sum tests, with false discovery rate (FDR) correction for multiple comparisons, as described in detail in ref. 38, among a range of topics and values of the parameter Δt. Further details are provided in *Supporting Information*. We find that strategies based on keywords in the categories Politics I (e.g., Republican, Wisconsin, Senate,. . .; mean return = 56.4%; $W = 20{,}713$, $P = 0.01$) and Business (e.g., business, management, bank,. . .; mean return = 38.6%; $W = 19{,}919$, $P = 0.04$) lead to significantly higher arithmetic cumulative returns than those from the random strategy, suggesting that changes in search volume for keywords belonging to these topics may have contained precursors of subsequent stock market moves. These two distributions are colored by their $\overline{R}$.

We examine the effect of changing the value of Δt. In Fig. 1*B*, we depict the results of varying Δt between 1 and 15 wk for all 55 topics. We color cells according to $\overline{R}$ for a given topic, using a given value of Δt. Where no color is shown, no significant difference is found between the distribution of arithmetic cumulative returns from a random strategy and the distribution of arithmetic cumulative returns for the topic's strategies with the given value of Δt ($P \geq 0.05$). We find that terms within the Business category result in significant values of R for values of Δt of 2–15 wk (all Ws $\geq 19{,}278$, all Ps < 0.05), with the exceptions of $\Delta t = 4$ wk and $\Delta t = 12$ wk. Terms within the category Politics I result in significant returns for $\Delta t = 2$–15 wk (all Ws $\geq 20{,}422$, all Ps < 0.05), with the exceptions of $\Delta t = 4$, 5, and 7 wk. The relationship between changes in search volume for these topics and movements in the SPXT is therefore reasonably robust to changes in Δt. We also find that terms within the category Politics II (e.g., party, law, government,. . .) result in significant values of R for $\Delta t = 6$ wk and $\Delta t = 8$–15 wk (all Ws $\geq 20{,}144$, all Ps < 0.05). For some values of Δt, we find significant values of R for terms belonging to the categories Medicine, Education I, and Education II. The significance of these values of R is, however, highly dependent on the value of Δt.

As a check of our procedure for multiple hypothesis testing, we repeat the above analysis using randomly generated search volumes. We construct $55 \cdot 30 = 1{,}650$ time series of search volume data by independently shuffling the time series of search volume for each word in each topic. We then recreate Fig. 1 *A* and *B* using these 55 "topics" in Fig. 1 *C* and *D*, respectively. We find that, after FDR correction, no such topic deviates significantly from the cumulative returns from an uncorrelated random strategy.

We next investigate the Politics I, Politics II, and Business categories more carefully. In particular, we examine the effect of changing the period during which we analyze this relationship. In Fig. 2, we depict the results of using a range of moving 4-yr windows between 2004 and 2012 for the Business, Politics I, and

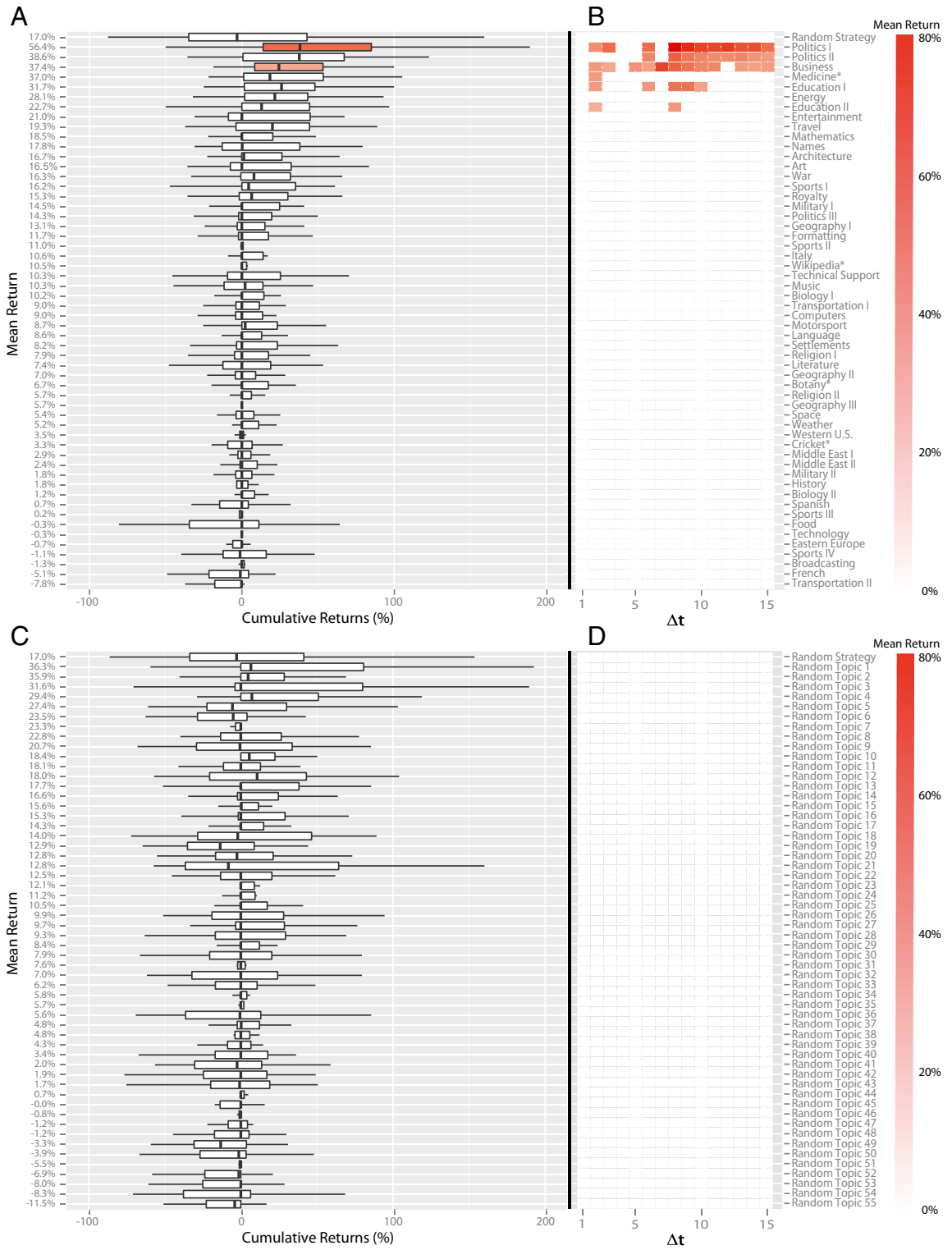

Fig. 1. Google Trends based trading strategies for 55 different semantic topics. (*A*) For each topic, we depict the distribution of cumulative returns from 30 trading strategies, each based on search volume data for one term belonging to the topic. Strategies trade weekly on the SPXT from 2004 to 2012, using $\Delta t = 3$ wk. We show in the top row the distribution of cumulative returns for a random strategy. The mean percentage returns for each topic appear on the left column. We compare the cumulative returns for search volume-based strategies to the distribution of cumulative returns from the random strategy using two-sample Wilcoxon rank sum tests, with FDR correction for multiple comparisons among a range of topics and values of the parameter Δt. We find that strategies based on keywords in the categories Politics I ($W = 20{,}713$, $P = 0.01$) and Business ($W = 19{,}919$, $P = 0.04$), shown in red, lead to higher cumulative returns than the random strategy. (*B*) Colored cells denote values of Δt for which the cumulative returns for a semantic topic are significantly higher than those of a random strategy ($P < 0.05$). Terms within the categories Business, Politics I, and Politics II result in significant returns across a range of values of Δt. (*C* and *D*) same as *A* and *B*, but using shuffled search volumes and finding no significant "topics."

Politics II topics with Δt held at 3 wk. We include an additional time window, from January 2010 to December 2013, to check the present-day performance of the strategies. We depict distributions of R for these periods using a kernel density estimate. As in Fig. 1, we compare the distributions of R from each topic with the distribution of R from random strategies. Terms in the Politics I category result in significant values of R (all Ws $\geq$ 18,839, all Ps < 0.05 after FDR correction) for all time windows, with the exception of 2009–2012 and 2010–2013. Terms relating to Business result in significant values of R for the periods 2004–2007, 2006–2009, 2007–2010, and 2008–2011 (all Ws $\geq$ 18,511, all Ps < 0.05, FDR correction applied). Finally, terms in the Politics II category result in significant values of R for the periods 2005–2008, 2006–2009, 2007–2010, and 2008–2011 (all Ws $\geq$ 19,196, all Ps < 0.05, FDR correction applied). Our results provide evidence of a historical relationship between the search behavior of Google users and financial market movements. However, our analyses suggest that the strength of this relationship has diminished in recent years, perhaps reflecting increasing incorporation of Internet data into automated trading strategies.

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

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

where ϵ_t is an error term.

Using the mean relative change in search volume for the Politics I category as our signal $x_{\text{Politics I}}$, we report a significantly negative coefficient of -2.80 ($t = -2.65$, $P = 0.024$, Bonferroni correction applied). Using instead the Business category for our signal x_{Business}, we report a significantly negative coefficient of -5.34 ($t = -2.61$, $P = 0.027$, Bonferroni correction applied). We find that the signal generated by the Politics II category $x_{\text{Politics II}}$, however, is not significantly related to subsequent stock market moves, according to this analysis ($t = -2.02$, $P = 0.13$, Bonferroni

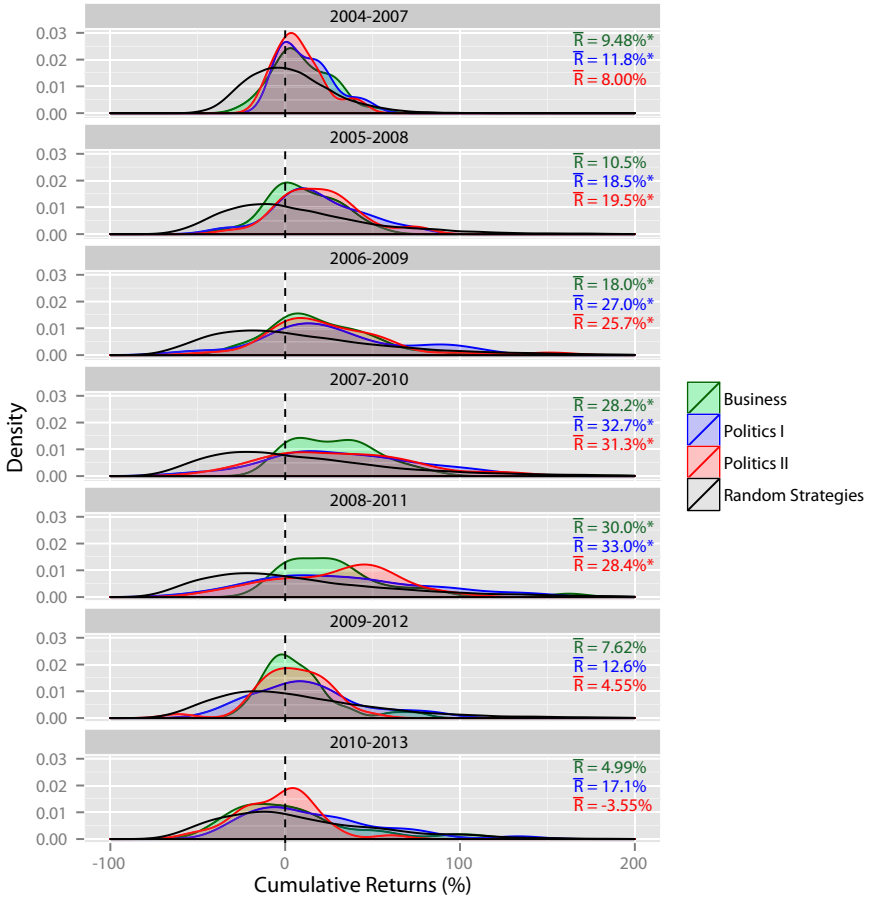

Fig. 2. Effect of changing time window on returns. For the Business, Politics I, and Politics II topics, we depict the distribution of cumulative returns from the corresponding trading strategies in six overlapping 4-yr time windows. Distributions are plotted using a kernel density estimate, with a Gaussian kernel and bandwidth calculated with Silverman's rule of thumb (42). Strategies trade weekly on the SPXT, using $\Delta t = 3$. The distribution of cumulative returns for a random strategy is also shown in each time window. The mean percentage return $\overline{R}$ for each topic is provided on the right of the figure. We compare the cumulative returns for search volume-based strategies to the distribution of cumulative returns from the random strategy using two-sample Wilcoxon rank sum tests, with FDR correction for multiple comparisons. Terms in the Politics I category result in significant returns (all Ws $\geq$ 18,839, all Ps < 0.05 after FDR correction) for all time windows, with the exception of 2009–2012 and 2010–2013. Terms relating to Business result in significant returns for the periods 2004–2007, 2006–2009, 2007–2010, and 2008–2011 (all Ws $\geq$ 18,511, all Ps < 0.05 after FDR correction). Finally, terms in the Politics II category result in significant returns for the periods 2005–2008, 2006–2009, 2007–2010, and 2008–2011 (all Ws $\geq$ 19,196, all Ps < 0.05 after FDR correction).

Table 1. Regression results using search volume signals $x_{\text{Politics I}}$, x_{Business}, and $x_{\text{Politics II}}$

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

*$P < 0.05$.

correction applied). Details of our corrections for multiple comparisons are provided in *Supporting Information*. The coefficient β_2 of the volatility index VIX was insignificant in all regressions ($P > 0.35$). We provide scatter plots of our signals against the subsequent week's return in Fig. S1 and detail the results of the regressions in Table 1. Table 2 provides the median, 5%, and 95% quantiles for the absolute value of the test statistics $|t|$ as well as R^2 for all 55 regressions carried out using the same shuffled search volume data that is represented in Fig. 1*C*. We find that the statistics $|t|$ and R^2 for the Politics I and Business topics fall within the top 5% of values obtained using the shuffled search volumes.

To examine the distributions of the test statistics for the Politics I, Business, and Politics II topics, we implement a block bootstrap procedure (39) in which we construct surrogate time series by circularly shifting our signals x_t (i.e., at each shift, the final entry is moved to the first position). We examine the distributions of t statistics and coefficients of determination R^2 under all such shifts, providing a safeguard against spurious results due to auto-correlative structure in the data. The median, 5%, and 95% quantiles are reported in Table 3, where we find that all observed test statistics fall within the top 5% of bootstrapped results.

As a final check of our results, we apply the Hansen test for superior predictive ability (39). For this test we construct 1,000 resamplings of the data, with replacement, using a stationary bootstrap technique (40, 41). The continuous block length of the pseudo time series is chosen to be geometrically distributed with parameter $q = 0.001$, of the order of the inverse length of the time series, to preserve effects due to autocorrelation. For each of the topics Politics I, Business, and Politics II, we test the universe of trading strategies generated by all 30 words in the topic against both a random strategy and a buy-and-hold strategy. We find that a random strategy is significantly outperformed by strategies generated by words in the Politics I ($T^{\text{SPA}} = 9.06$, $P < 0.001$), Business ($T^{\text{SPA}} = 9.53$, $P < 0.001$), and Politics II ($T^{\text{SPA}} = 6.47$, $P < 0.001$) topics. However, we only find marginal support for these strategies significantly outperforming a buy-and-hold strategy (Politics I: $T^{\text{SPA}} = 2.34$, $P = 0.085$; Business: $T^{\text{SPA}} = 2.62$, $P = 0.071$; Politics II: $T^{\text{SPA}} = 1.23$, $P = 0.143$).

Discussion

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

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

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

Table 3. Comparison of observed test statistics with those obtained from bootstrapping procedure

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

We note that the overlap between words in the topics Politics I (e.g., Republican, Wisconsin, Senate,...) and Politics II (e.g., party, law, government,...) is small; the two topics, containing 30 words each, share only four words: "president," "law," "election," and "democratic." Despite this, our method identifies relationships between both politics-related topics and stock market moves, providing further evidence of the importance of underlying semantic factors in keyword search data. We note that a third topic related to politics, Politics III, was not flagged by our method. A close inspection reveals that this topic in fact bears more relevance to politics in the United Kingdom, containing the keywords "parliament," "british," "labour," "london," etc. This finding is in line with the suggestion that changes in online information gathering specifically relating to politics in Britain may not bear a strong relationship to subsequent financial market moves in the United States.

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

The method we present here facilitates in a number of ways the interpretation of the relationship between search data and complex events such as financial market moves. First, the frequency of searches for a given keyword can grow and decline for various reasons, some of which may or may not be related to a real-world event of interest. This method allows us to abstract away from potentially noisy data for individual keywords and identify underlying semantic factors of importance. Second, our method allows us to extract subsets of search data of relevance to real-world events, without privileged access to full data on all search queries made by Google users. By identifying representative keywords for a range of semantic topics, such analyses can be carried out despite limitations on the number of keywords for which search data can be retrieved via the Google Trends interface. Third, our semantic analysis is based on simple statistics on how often words occur in documents alongside other words. As a result,

the analysis presented could be carried out in languages other than English—for example, using other editions of Wikipedia—with no extra modifications to the approach required. We suggest that extensions of these analyses could offer insight into large-scale information flow before a range of real-world events.

ACKNOWLEDGMENTS. We thank Adam Avakian and Dror Y. Kenett for comments. This work was supported by the Intelligence Advanced Research Projects Activity via Department of Interior National Business Center Contract D12PC00285. C.C., T.P., and H.S.M. acknowledge support from Research Councils UK Grant EP/K039830/1. C.C. and H.E.S. also wish to thank the Office of Naval Research (Grants N00014-09-1-0380 and N00014-12-1-0548), the Defense Threat Reduction Agency (Grants HDTRA-1-10-1-0014 and HDTRA-1-09-1-0035), and the National Science Foundation (Grant CMMI 1125290). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

1. Shleifer A (2000) *Inefficient Markets: An Introduction to Behavioral Finance* (Oxford Univ Press, Oxford).
2. Lillo F, Farmer JD, Mantegna RN (2003) Econophysics: Master curve for price-impact function. *Nature* 421(6919):129–130.
3. Gabaix X (2009) Power laws in economics and finance. *Annu Rev Econ* 1:255–293.
4. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2003) A theory of power-law distributions in financial market fluctuations. *Nature* 423(6937):267–270.
5. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2006) Institutional investors and stock market volatility. *Q J Econ* 121(2):461–504.
6. Preis T, Schneider JJ, Stanley HE (2011) Switching processes in financial markets. *Proc Natl Acad Sci USA* 108(19):7674–7678.
7. Takayasu H, ed (2006) *Practical Fruits of Econophysics* (Springer, Berlin).
8. Coval J, Jurek JW, Stafford E (2009) Economic catastrophe bonds. *Am Econ Rev* 99(3): 628–666.
9. Bouchaud JP, Matacz A, Potters M (2001) Leverage effect in financial markets: The retarded volatility model. *Phys Rev Lett* 87(22):228701.
10. Hommes CH (2002) Modeling the stylized facts in finance through simple nonlinear adaptive systems. *Proc Natl Acad Sci USA* 99(Suppl 3):7221–7228.
11. Haldane AG, May RM (2011) Systemic risk in banking ecosystems. *Nature* 469(7330): 351–355.
12. Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397:498–500.
13. Krugman P (1996) *The Self-Organizing Economy* (Blackwell, Cambridge, MA).
14. Sornette D, von der Becke S (2011) Complexity clouds finance-risk models. *Nature* 471(7337):166.
15. Garlaschelli D, Caldarelli G, Pietronero L (2003) Universal scaling relations in food webs. *Nature* 423(6936):165–168.
16. Onnela J-P, Arbesman S, González MC, Barabási A-L, Christakis NA (2011) Geographic constraints on social network groups. *PLoS ONE* 6(4):e16939.
17. Schweitzer F, et al. (2009) Economic networks: The new challenges. *Science* 325(5939): 422–425.
18. Gabaix X (2011) The granular origins of aggregate fluctuations. *Econometrica* 79: 733–772.
19. Gabaix X, Krishnamurthy A, Vigneron O (2007) Limits of arbitrage: Theory and evidence from the mortgage-backed securities market. *J Finance* 62(2):557–595.
20. Lux T (1999) The socio-economic dynamics of speculative markets: Interacting agents, chaos, and the fat tails of return distributions. *J Econ Behav Organ* 33:143–165.
21. Simon HA (1955) A behavioral model of rational choice. *Q J Econ* 69:99–118.
22. King G (2011) Ensuring the data-rich future of the social sciences. *Science* 331(6018): 719–721.
23. Vespignani A (2009) Predicting the behavior of techno-social systems. *Science* 325(5939):425–428.
24. Lazer D, et al. (2009) Computational social science. *Science* 323(5915):721–723.
25. Conte R, et al. (2012) Manifesto of computational social science. *Eur Phys J Spec Top* 214:325–346.
26. Moat HS, Preis T, Olivola CY, Liu C, Chater N (2014) Using big data to predict collective behavior in the real world. *Behav Brain Sci* 37(1):92–93.
27. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with Web search. *Proc Natl Acad Sci USA* 107(41):17486–17490.
28. Askitas N, Zimmermann KF (2009) Google econometrics and unemployment forecasting. *Appl Econ Q* 55(2):107–120.
29. Ginsberg J, et al. (2009) Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.
30. Preis T, Reith D, Stanley HE (2010) Complex dynamics of our economic life on different scales: Insights from search engine query data. *Philos Trans A Math Phys Eng Sci* 368(1933):5707–5719.
31. Bordino I, et al. (2012) Web search queries can predict stock market volumes. *PLoS ONE* 7(7):e40014.
32. Preis T, Moat HS, Stanley HE, Bishop SR (2012) Quantifying the advantage of looking forward. *Sci Rep* 2:350.
33. Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using Google Trends. *Sci Rep* 3:1684.
34. Da Z, Engelberg J, Gao P (2011) In search of attention. *J Finance* 66(5):1461–1499.
35. Bank M, Larch M, Peter G (2011) Google search volume and its influence on liquidity and returns of German stocks. *Financ Mark Portfolio Manage* 25(3):239–264.
36. Moat HS, et al. (2013) Quantifying Wikipedia usage patterns before stock market moves. *Scientific Rep* 3:1801.
37. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022.
38. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc, B* 57:289–300.
39. Politis D, Romano J (1992) *Exploring the Limits of Bootstrap*, eds LePage R, Billard L (Wiley, New York), pp 263–270.
40. Hansen PR (2005) A test for superior predictive ability. *J Bus Econ Stat* 23:365–380.
41. Lux T (2011) Sentiment dynamics and stock returns: The case of the German stock market. *Empir Econ* 41:663–679.
42. Silverman BW (1986) *Density Estimation* (Chapman and Hall, London).