ISBN 978-94-017-8707-9

NATO Science for Peace and Security Series - C:
Environmental Security

Nonlinear Phenomena in Complex Systems: From Nano to Macro Scale

Edited by
Davron Matrasulov
H. Eugene Stanley

Springer

This publication is supported by: The NATO Science for Peace and Security Programme

Chapter 4
Anticipating Stock Market Movements with *Google* and *Wikipedia*

Helen Susannah Moat, Chester Curme, H. Eugene Stanley, and Tobias Preis

Abstract Many of the trading decisions that have led to financial crises are captured by vast, detailed stock market datasets. Here, we summarize two of our recent studies which investigate whether Internet usage data contain traces of attempts to gather information before such trading decisions were taken. By analyzing changes in how often Internet users searched for financially related information on *Google* (Preis et al., Sci Rep 3:1684, 2013) and *Wikipedia* (Moat et al., Sci Rep 3:1801, 2013), patterns are found that may be interpreted as "early warning signs" of stock market moves. Our results suggest that online data may allow us to gain new insight into early information gathering stages of economic decision making.

4.1 Introduction

Stock market data provide extremely detailed records of decisions that traders have made, in an area in which disasters have a widespread impact. As a result, these stock market records have generated considerable scientific attention [7, 8, 11–13, 18–24, 26–28].

H.S. Moat (✉) • T. Preis
Behavioural Science, Warwick Business School, University of Warwick, Scarman Road, Coventry, CV4 7AL, UK
e-mail: Suzy.Moat@wbs.ac.uk; Tobias.Preis@wbs.ac.uk

C. Curme
Center for Polymer Studies, Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, MA 02215, USA
e-mail: ccurme@bu.edu

H.E. Stanley
Boston University, Department of Physics, Boston, MA, USA
e-mail: hes@bu.edu

D. Matrasulov and H.E. Stanley (eds.), *Nonlinear Phenomena in Complex Systems: From Nano to Macro Scale*, NATO Science for Peace and Security Series C: Environmental Security, DOI 10.1007/978-94-017-8704-8_4,

Decisions, such as trading decisions, do not however consist solely of the final execution of a chosen action, such as a trade recorded at the stock exchange. Instead, humans often begin by gathering information to help identify what the consequences of possible actions might be [33].

Nowadays, the Internet has greatly extended human capabilities to distribute and gather information [1, 6, 14, 15, 31]. As a result, online resources have become the first port of call in many quests for new information. Providers of such online resources often collect extensive data on their usage, adding to a range of new large-scale measurements of collective human behavior [5, 17]. These new Internet derived datasets open up new avenues for scientists to investigate the early information gathering stages of decision making processes.

Previous studies have demonstrated that analysis of search data can provide insight into current or even subsequent behavior in the real world. For example, changes in the frequency with which users look for certain terms on search engines such as *Google* and *Yahoo!* have been correlated with changes in the numbers of reports of flu infections across the USA [9], the popularity of films, games and music on their release [10], unemployment rates [2, 4], tourist numbers [4], and trading volumes in the US stock markets [3, 25]. A recent study showed that Internet users from countries with a higher per capita gross domestic product (GDP) search for proportionally more information about the future than information about the past, in comparison with Internet users from countries with a lower per capita GDP [29].

In the two studies summarized here and described in [16] and [30] in full length, we ask whether online searches for information might contain information relevant not only to the current state of the stock market, but also to subsequent trends. Specifically, can we find any evidence that changes in the volume of searches for financial information on *Google* and *Wikipedia* may provide insight into the information gathering process of investors before they make decisions to buy or sell?

4.2 *Google* Searches and Subsequent Stock Market Moves

To investigate whether changes in information gathering behavior as captured by *Google Trends* data were related to later changes in stock price in the period between 2004 and 2011, in [30] we implemented a hypothetical investment strategy for a portfolio using search volume data, called '*Google Trends* strategy' in the following. In this strategy, as described in both [30] and [16], we quantify changes in information gathering behavior by using the relative change in search volume: $\Delta n(t, \Delta t) = n(t) - N(t-1, \Delta t)$ with $N(t-1, \Delta t) = (n(t-1) + n(t-2) + \ldots + n(t - \Delta t))/\Delta t$, where t is measured in units of weeks. We sell the DJIA at the closing price $p(t+1)$ on the first trading day of week $t+1$ if search volume has increased in week t such that $\Delta n(t, \Delta t) > 0$. We then close the position by buying

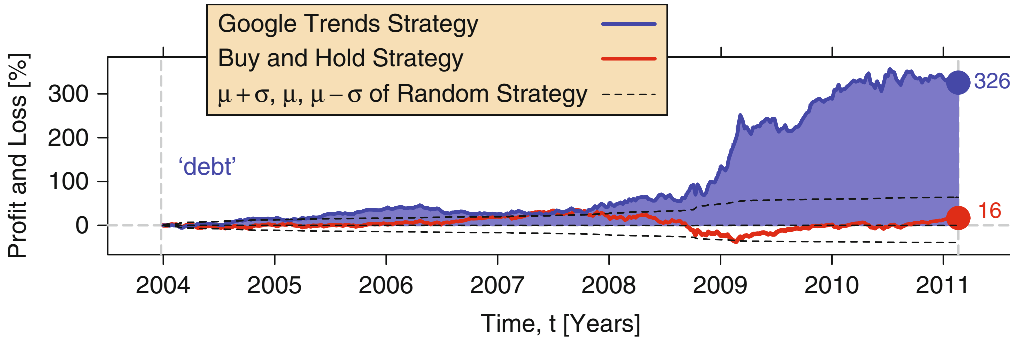


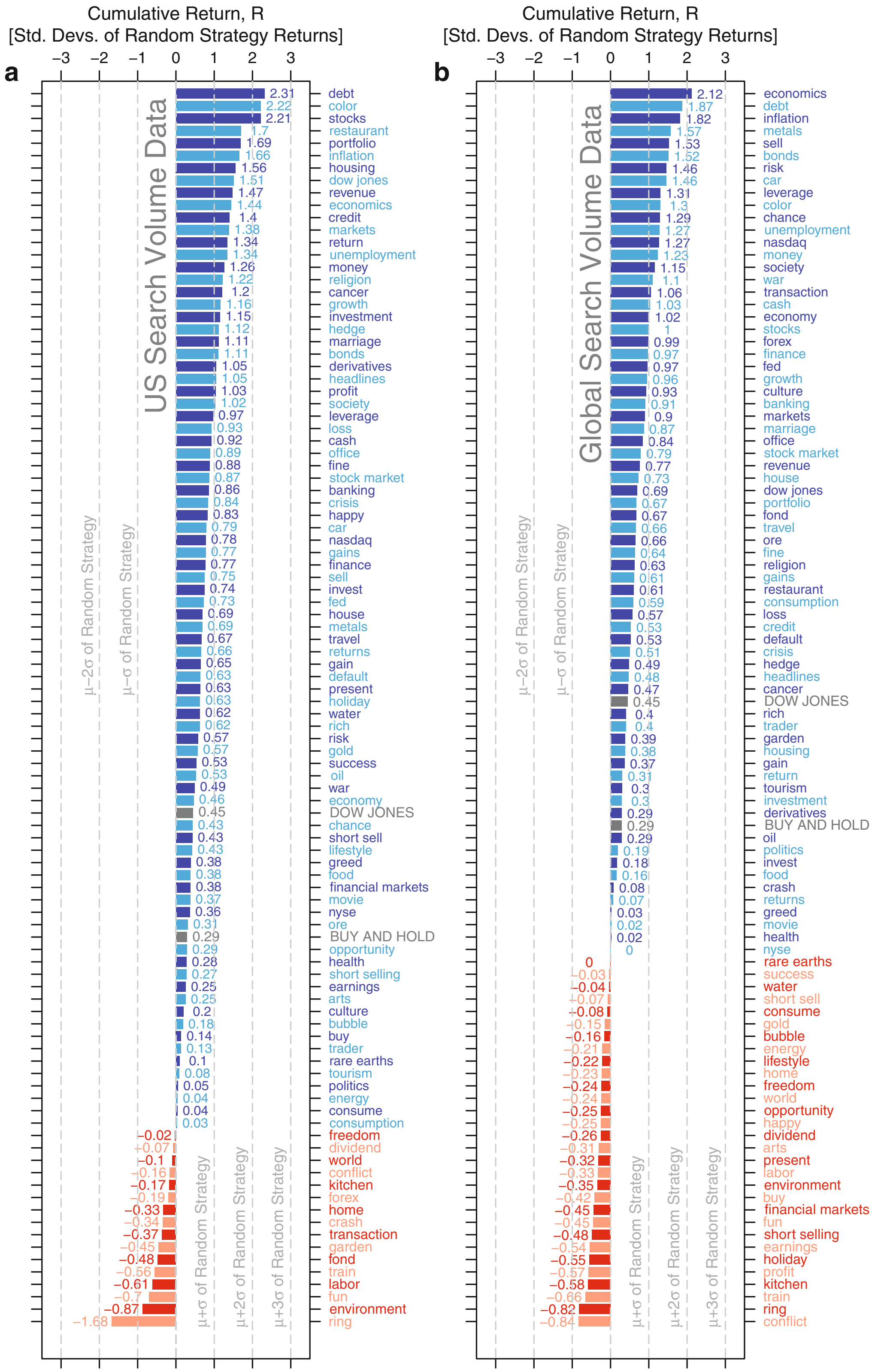
Fig. 4.1 Cumulative performance of an investment strategy based on *Google Trends* data (Reproduced from [30]). Profit and loss for an investment strategy based on the volume of the search term *debt*, the best performing keyword in our analysis, with $\Delta t = 3$ weeks, plotted as a function of time *(blue line)*. This is compared to the "buy and hold" strategy *(red line)* and the standard deviation of 10,000 simulations using a purely random investment strategy *(dashed lines)*. The Google Trends strategy using the search volume of the term *debt* would have yielded a profit of 326 %

the DJIA at price $p(t+2)$ at the end of the first trading day of the following week $t+2$. If instead search volume has decreased or remained the same in week t such that $\Delta n(t, \Delta t) \leq 0$, then we buy the DJIA at the closing price $p(t+1)$ on the first trading day of week $t+1$, and sell the DJIA at price $p(t+2)$ at the end of the first trading day of the coming week $t+2$ to close the position.

In [30], we analyzed the performance of a set of 98 *Google* search terms. We included terms related to the concept of stock markets, with some terms suggested by the *Google Sets* service, a tool which identifies semantically related keywords.

In Fig. 4.1, taken from [30], we depict the performance of our strategy between 2004 and 2011 using the search term *debt*, a keyword with an obvious semantic connection to the most recent financial crisis, and overall the term which performed best in our analyses. The performance of the *Google Trends* strategy based on the search term *debt* is depicted by a blue line, whereas dashed lines indicate the standard deviation of the cumulative return from a strategy in which we buy and sell the market index in an uncorrelated, random manner ('random investment strategy'). The standard deviation is derived from simulations of 10,000 independent realizations of the random investment strategy. Figure 4.1 shows that the use of the *Google Trends* strategy, based on the search term *debt* and $\Delta t = 3$ weeks, would have increased the value of a portfolio by 326 %. The performance of *Google Trends* strategies based on all other search terms that we analyze is depicted in a similar manner in [30].

We rank the full list of the 98 investigated search terms by their trading performance when using search data for U.S. users only (Fig. 4.2a) and when using globally generated search volume (Fig. 4.2b). In order to ensure the robustness of our results, the overall performance of a strategy based on a given search term


a
Cumulative Return, R
[Std. Devs. of Random Strategy Returns]
−3 −2 −1 0 1 2 3
US Search Volume Data
μ−2σ of Random Strategy
μ−σ of Random Strategy
μ+σ of Random Strategy
μ+2σ of Random Strategy
μ+3σ of Random Strategy
2.31 debt
2.22 color
2.21 stocks
1.7 restaurant
1.69 portfolio
1.66 inflation
1.56 housing
1.51 dow jones
1.47 revenue
1.44 economics
1.4 credit
1.38 markets
1.34 return
1.34 unemployment
1.26 money
1.22 religion
1.2 cancer
1.16 growth
1.15 investment
1.12 hedge
1.11 marriage
1.11 bonds
1.05 derivatives
1.05 headlines
1.03 profit
1.02 society
0.97 leverage
0.93 loss
0.92 cash
0.89 office
0.88 fine
0.87 stock market
0.86 banking
0.84 crisis
0.83 happy
0.79 car
0.78 nasdaq
0.77 gains
0.77 finance
0.75 sell
0.74 invest
0.73 fed
0.69 house
0.69 metals
0.67 travel
0.66 returns
0.65 gain
0.63 default
0.63 present
0.63 holiday
0.62 water
0.62 rich
0.57 risk
0.57 gold
0.53 success
0.53 oil
0.49 war
0.46 economy
0.45 DOW JONES
0.43 chance
0.43 short sell
0.43 lifestyle
0.38 greed
0.38 food
0.38 financial markets
0.37 movie
0.36 nyse
0.31 ore
0.29 BUY AND HOLD
0.29 opportunity
0.28 health
0.27 short selling
0.25 earnings
0.25 arts
0.2 culture
0.18 bubble
0.14 buy
0.13 trader
0.1 rare earths
0.08 tourism
0.05 politics
0.04 energy
0.04 consume
0.03 consumption
−0.02 freedom
−0.07 dividend
−0.1 world
−0.16 conflict
−0.17 kitchen
−0.19 forex
−0.33 home
−0.34 crash
−0.37 transaction
−0.45 garden
−0.48 fond
−0.56 train
−0.61 labor
−0.7 fun
−0.87 environment
−1.68 ring
b
Cumulative Return, R
[Std. Devs. of Random Strategy Returns]
−3 −2 −1 0 1 2 3
Global Search Volume Data
μ−2σ of Random Strategy
μ−σ of Random Strategy
μ+σ of Random Strategy
μ+2σ of Random Strategy
μ+3σ of Random Strategy
2.12 economics
1.87 debt
1.82 inflation
1.57 metals
1.53 sell
1.52 bonds
1.46 risk
1.46 car
1.31 leverage
1.3 color
1.29 chance
1.27 unemployment
1.27 nasdaq
1.23 money
1.15 society
1.1 war
1.06 transaction
1.03 cash
1.02 economy
1 stocks
0.99 forex
0.97 finance
0.97 fed
0.96 growth
0.93 culture
0.91 banking
0.9 markets
0.87 marriage
0.84 office
0.79 stock market
0.77 revenue
0.73 house
0.69 dow jones
0.67 portfolio
0.67 fond
0.66 travel
0.66 ore
0.64 fine
0.63 religion
0.61 gains
0.61 restaurant
0.59 consumption
0.57 loss
0.53 credit
0.53 default
0.51 crisis
0.49 hedge
0.48 headlines
0.47 cancer
0.45 DOW JONES
0.4 rich
0.4 trader
0.39 garden
0.38 housing
0.37 gain
0.31 return
0.3 tourism
0.3 investment
0.29 derivatives
0.29 BUY AND HOLD
0.29 oil
0.19 politics
0.18 invest
0.16 food
0.08 crash
0.07 returns
0.03 greed
0.02 movie
0.02 health
0 nyse
0 rare earths
−0.03 success
−0.04 water
−0.07 short sell
−0.08 consume
−0.15 gold
−0.16 bubble
−0.21 energy
−0.22 lifestyle
−0.23 home
−0.24 freedom
−0.24 world
−0.25 opportunity
−0.25 happy
−0.26 dividend
−0.31 arts
−0.32 present
−0.33 labor
−0.35 environment
−0.42 buy
−0.45 financial markets
−0.45 fun
−0.48 short selling
−0.54 earnings
−0.55 holiday
−0.57 profit
−0.58 kitchen
−0.66 train
−0.82 ring
−0.84 conflict

is determined as the mean value over the six returns obtained for $\Delta t = 1 \ldots 6$ weeks. Returns of the strategies are calculated as the logarithm of percentage profit, following the usual definition of returns. Here we report R, the cumulative returns of a strategy, in standard deviations of the cumulative returns of these uncorrelated random investment strategies. In [30], we find that returns from the *Google Trends* strategies we tested are significantly higher overall than returns from the random strategies ($\langle R \rangle_{US} = 0.60$; $t = 8.65$, $df = 97$, $p < 0.001$, one sample t-test).

We compare the performance of these search terms with two benchmark strategies. The 'buy and hold' strategy is implemented by buying the index in the beginning and selling it at the end of the hold period. This strategy yields 16 % profit, equal to the overall increase in value of the DJIA in the time period from January 2004 until February 2011. We further implement a 'Dow Jones strategy' by using changes in $p(t)$ in place of changes in search volume data as the basis of buy and sell decisions. In [30] we find that this strategy also yields only 33 % profit with $\Delta t = 3$ weeks, or when determined as the mean value over the six returns obtained for $\Delta t = 1 \ldots 6$ weeks, 0.45 standard deviations of cumulative returns of uncorrelated random investment strategies (Fig. 4.2a, b).

It is widely recognized that investors prefer to trade on their domestic market, suggesting that search data for U.S. users only, as used in analyses so far, should better capture the information gathering behavior of U.S. stock market participants than data for *Google* users worldwide. Indeed, in [30] we find that strategies based on global search volume data are less successful than strategies based on U.S. search volume data in anticipating movements of the U.S. market ($\langle R \rangle_{US} = 0.60$, $\langle R \rangle_{Global} = 0.43$; $t = 2.69$, $df = 97$, $p < 0.01$, two-sided paired t-test).

Fig. 4.2 Performances of investment strategies based on search volume data (Reproduced from [30]). (**a**) Cumulative returns of 98 investment strategies based on search volumes restricted to search requests of users located in the United States for different search terms, displayed for the entire time period of our study from 5 January 2004 until 22 February 2011 – the time period for which *Google Trends* provides data. We use *two shades of blue* for positive returns and *two shades of red* for negative returns to improve the readability of the search terms. The cumulative performance for the "buy and hold strategy" is also shown, as is a "Dow Jones strategy", which uses weekly closing prices of the Dow Jones Industrial Average (DJIA) rather than *Google Trends* data (see *gray bars*). Figures provided next to the bars indicate the returns of a strategy, R, in standard deviations from the mean return of uncorrelated random investment strategies, $\langle R \rangle_{RandomStrategy} = 0$. *Dashed lines* correspond to −3, −2, −1, 0, +1, +2, and +3 standard deviations of random strategies. We find that returns from the *Google Trends* strategies tested are significantly higher overall than returns from the random strategies ($\langle R \rangle_{US} = 0.60$; $t = 8.65$, $df = 97$, $p < 0.001$, one sample t-test). (**b**) A parallel analysis shows that extending the range of the search volume analysis to global users reduces the overall return achieved by *Google Trends* trading strategies on the U.S. market ($\langle R \rangle_{US} = 0.60$, $\langle R \rangle_{Global} = 0.43$; $t = 2.69$, $df = 97$, $p < 0.01$, two-sided paired t-test). However, returns are still significantly higher than the mean return of random investment strategies ($\langle R \rangle_{Global} = 0.43$; $t = 6.40$, $df = 97$, $p < 0.001$, one sample t-test)

4.3 *Wikipedia* Views and Edits and Subsequent Stock Market Moves

In [16], we investigate whether data from the popular online encyclopedia *Wikipedia* may hold similar insights. We consider data on both how often pages on the English language *Wikipedia* have been viewed, and how often pages on the English language *Wikipedia* have been edited. We calculate our weekly measure of information gathering behavior, $n(t)$, as previously described, but using either view or edit data for *Wikipedia* in place of search volume data from *Google*. Data on *Wikipedia* page views were downloaded from the online service http://stats.grok.se, and data on *Wikipedia* page edits were obtained by parsing the *Wikipedia* "Revision history" page associated to the article. In [16], we then implement the same trading strategy described above using data generated between 10th December 2007, the earliest date for which *Wikipedia* views data are available from http://stats.grok.se, and 30th April 2012.

Figure 4.3, taken from [16], shows the distributions of returns from two portfolios of 30 hypothetical strategies, trading weekly on the DJIA. These trading strategies are based on changes in how often the 30 Wikipedia pages describing the companies in the DJIA were viewed (*blue)*) and edited (*red*) during the period December 2007–April 2012, with $\Delta t = 3$ weeks. The distribution of returns from 10,000 independent realizations of a random strategy is also shown (*gray*).

We find that there are significant differences between these three distributions ($\chi^2 = 10.21$, $df = 2$, $p = 0.006$, Kruskal-Wallis rank sum test). Our analysis shows that the returns of *Wikipedia* page view based strategies for this period are significantly higher than the returns of the random strategies ($\langle R \rangle_{Views} = 0.50$; $W = 199{,}690$, $p = 0.005$, two-tailed two-sample Wilcoxon rank-sum test, Bonferroni correction applied). There is however no statistically significant difference between the returns from the Wikipedia edit based strategies and the random strategies ($\langle R \rangle_{Edits} = -0.09$; $W = 140{,}781$, $p > 0.9$, two-tailed two-sample Wilcoxon rank-sum test, Bonferroni correction applied).

We investigate whether these results extend to *Wikipedia* articles on more general financial topics. To address this question, we make use of the fact that *Wikipedia* contains lists of pages relating to specific topics. In [16], we examine view and edit data for 285 pages relating to general economic concepts, as listed in the subsection "General Economic Concepts" on the English language *Wikipedia* page "Outline of Economics".

Figure 4.4 shows the results of an analysis of the distribution of returns from two portfolios of 285 hypothetical strategies, trading weekly on the DJIA. These strategies are based on changes in how often these 285 financially related *Wikipedia* pages were viewed (*blue*) and edited (*red*) during the same period, again with $\Delta t = 3$ weeks. As before, we find that there is a significant difference between the returns generated by the random strategies, the *Wikipedia* view based strategies and the *Wikipedia* edit based strategies ($\chi^2 = 307.88$, $df = 2$,

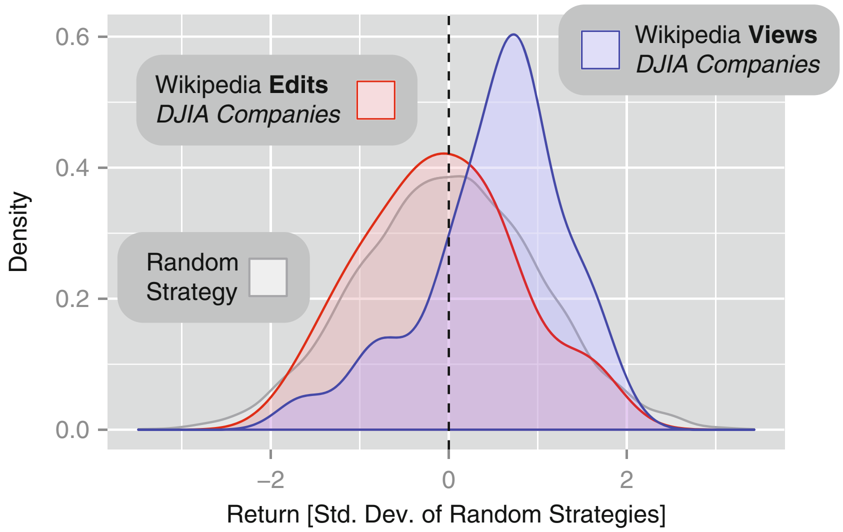


Fig. 4.3 Returns from trading strategies based on Wikipedia view and edit logs for articles relating to the companies forming the Dow Jones Industrial Average (DJIA) (Reproduced from [16]). The distributions of returns from two portfolios of 30 hypothetical strategies, trading weekly on the DJIA, based on changes in how often the 30 *Wikipedia* articles describing the companies listed in the DJIA were viewed (*blue*) and edited (*red*) during the period December 2007–April 2012, with $\Delta t = 3$ weeks. The distribution of returns from 10,000 independent realizations of a random strategy is also shown (*gray*). Data is displayed using a kernel density estimate and the *ggplot2* library [35], with a Gaussian kernel and bandwidth calculated using Silverman's rule of thumb [32]. Whereas we show in the text that random strategies lead to no significant profit or loss, we find that the returns of *Wikipedia* article view based strategies for this period are significantly higher than the returns of the random strategies ($\langle R \rangle_{Views} = 0.50$; $W = 199{,}690$, $p = 0.005$, two-tailed two-sample Wilcoxon rank-sum test, Bonferroni correction applied). There is however no statistically significant difference between the returns from the Wikipedia edit based strategies and the random strategies ($\langle R \rangle_{Edits} = -0.09$; $W = 140{,}781$, $p > 0.9$, two-tailed two-sample Wilcoxon rank-sum test, Bonferroni correction applied)

$p < 0.001$, Kruskal-Wallis rank sum test). Again, the returns of *Wikipedia* page view based strategies are significantly higher than the returns of random strategies for this period ($\langle R \rangle_{Views} = 1.10$; $W = 2{,}286{,}608$, $p < 0.001$, two-tailed two-sample Wilcoxon rank-sum test, Bonferroni correction applied). In contrast, we find no evidence of a statistically significant difference between the returns from the *Wikipedia* edit based strategies, and the random strategies ($\langle R \rangle_{Edits} = 0.12$; $W = 1{,}516{,}626$, $p = 0.19$, two-tailed two-sample Wilcoxon rank-sum test, Bonferroni correction applied).

We note in [16] that the lack of relationship found for the data on *Wikipedia* edits may simply reflect the substantial difference in the volume of data available for views and for edits, despite the much larger number of pages considered in this second analysis, where further relevant statistics on views and edits of *Wikipedia* pages are provided in [16]. For the purposes of these investigations, we therefore do not consider edit data further.

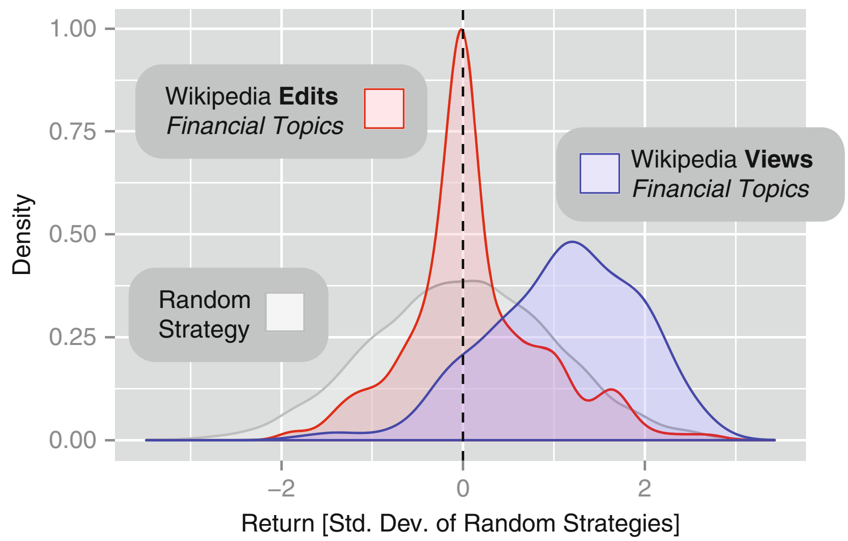


Fig. 4.4 Returns from trading strategies based on *Wikipedia* access and edit logs for pages relating to finance (Reproduced from [16]). Parallel analysis of the distribution of returns from two much larger portfolios of 285 hypothetical strategies, based on changes in how often a set of 285 financially related *Wikipedia* pages were viewed (*blue*) and edited (*red*) during the same period as Fig. 4.3, again with $\Delta t = 3$ weeks. Our analysis shows that the returns of *Wikipedia* page view based strategies are significantly higher than the returns of random strategies for this period ($\langle R \rangle_{Views} = 1.10$; $W = 2{,}286{,}608$, $p < 0.001$, two-tailed two-sample Wilcoxon rank-sum test, Bonferroni correction applied). Once again however, we find no evidence of a statistically significant difference between the returns from the *Wikipedia* edit based strategies, and the random strategies ($\langle R \rangle_{Edits} = 0.12$; $W = 1{,}516{,}626$, $\alpha = 0.05$, two-tailed two-sample Wilcoxon rank-sum test, Bonferroni correction applied)

4.4 Financial Relevance of Information Searched for Before Stock Market Falls

Our assumption so far was that only *Google* and *Wikipedia* usage data relating to financial topics would provide any insight into information gathering processes before trading decisions, and therefore future changes in the DJIA. To verify this assumption, in [16] we carry out a further analysis of view data relating to 233 *Wikipedia* pages describing actors and filmmakers, where further details of these pages are provided in [16]. We suggest that such pages have less obvious financial connotations.

We analyze the distribution of returns for a portfolio of 233 hypothetical trading strategies based on changes in how often these pages were viewed, trading weekly on the DJIA with $\Delta t = 3$ weeks during the period December 2007–April 2012, as in the previous *Wikipedia* analyses. We ensured that this set of pages, of similar size to the set of pages relating to financial topics, had at least equivalent traffic during the period of investigation, to ensure that any failure to find a relationship was not due to power issues caused through lack of data on *Wikipedia* views.

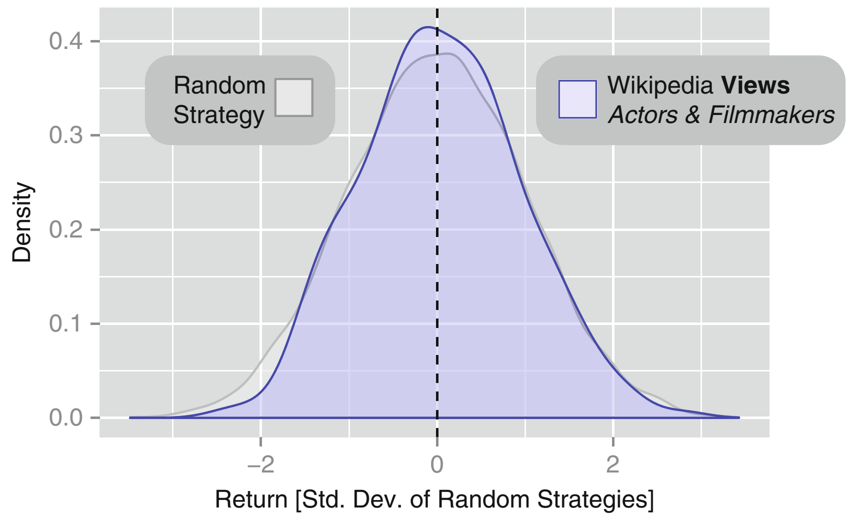


Fig. 4.5 Returns from trading strategies based on *Wikipedia* access logs for pages relating to actors and filmmakers (Reproduced from [16]). Parallel analysis of the distribution of returns for another portfolio of 233 hypothetical strategies based on changes in how often a set of 233 Wikipedia pages relating to actors and filmmakers were viewed (*blue*). Here, we find that there is no significant difference between the returns generated by the random strategies and the *Wikipedia* view based strategies ($\langle R \rangle_{Views} = 0.04$; $W = 1{,}189{,}114$, $p = 0.59$, two-tailed two-sample Wilcoxon rank-sum test)

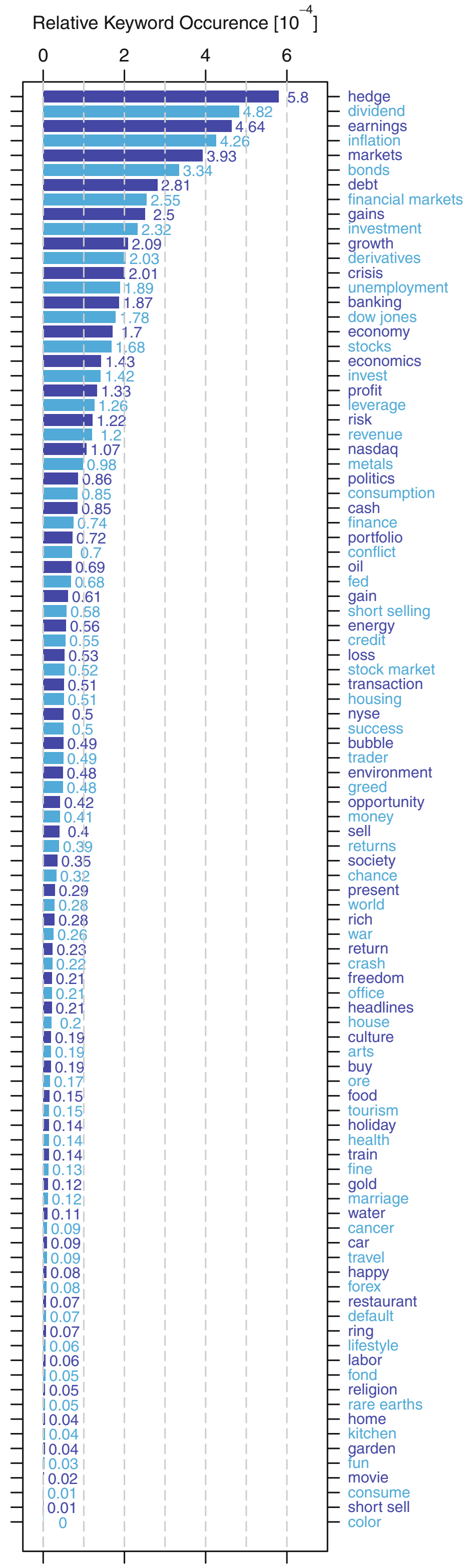
In Fig. 4.5, we show the returns from these 233 strategies based on changes in the number of views of *Wikipedia* articles on actors and filmmakers (*blue*), alongside returns from the random strategies (*gray*). We find that there is no significant difference between the returns generated by the random strategies and the *Wikipedia* view based strategies ($\langle R \rangle_{Views} = 0.04$; $W = 1{,}189{,}114$, $p = 0.59$, two-tailed two-sample Wilcoxon rank-sum test).

Similarly, in [30], we investigate whether differences in performance of the 98 *Google Trends* strategies we tested can be partially explained using an indicator of the extent to which different terms are of financial relevance. We quantify financial relevance by calculating the frequency of each search term in the online edition of the *Financial Times* from August 2004 to June 2011, normalized by the number of *Google* hits for each search term (Fig. 4.6). We find that the return associated with a given search term is correlated with this indicator of financial relevance (Kendall's tau $= 0.275$, $z = 4.01$, $N = 98$, $p < 0.001$) using Kendall's tau rank correlation coefficient.

4.5 Discussion

In the investigations described in [16] and [30], summarized here, we find evidence of increases in searches for financially related information before stock market falls. These results are consistent with the hypothesis that historic usage data from *Google*

Fig. 4.6 How related are search terms to the topic of finance? (Reproduced from [30]). We quantify financial relevance by calculating the frequency of each search term in the online edition of the *Financial Times* (http://www.ft.com) from August 2004 to June 2011, normalized by the number of *Google* hits (http://www.google.com) for each search term



and the online encyclopedia *Wikipedia* may have provided some insight into future trends in the behavior of financial market actors.

In [16], we have proposed one potential explanation in line with these results. We first suggest that *Google* and *Wikipedia* records may provide a proxy measurement of the information gathering process of a subset of investors for the investigated period. We further note that previous studies in behavioral economics have demonstrated that humans are loss averse [34]: that is, they are more concerned about losing $5 than they are about missing an opportunity to gain $5. By this logic, it could be argued that the trading decision of greatest consequence for a trader would be to sell a stock at a lower price than they had previously believed it was worth. If we assume that investors may be willing to invest more efforts in information gathering before making a decision which they view to be of greater consequence, then it would follow that increases in information gathering would precede falls in stock market prices, in line with our results.

Our results suggest that Internet usage data may offer a window into the information gathering processes which precede real world actions captured in large behavioral data sets. By combining these new data sets, we may be able to gain new insight into different stages of collective economic decision making.

Acknowledgements The results summarized here were first reported in Preis, T., Moat, H.S., Stanley. H.E.: Quantifying trading behavior in financial markets using Google Trends. Sci. Rep. 3, 1684 (2013), http://www.nature.com/srep/2013/130425/srep01684/full/srep01684.html, and Moat, H.S. et al.: Quantifying Wikipedia usage patterns before stock market moves. Sci. Rep. 3, 1801 (2013), http://www.nature.com/srep/2013/130508/srep01801/full/srep01801.html. We thank Adam Avakian and Dror Y. Kenett for comments. H.S.M. and T.P. acknowledge the support of the Research Councils UK Grant EP/K039830/1. This work was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC00285. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

1. Alanyali M, Moat HS, Preis T (2013) Quantifying the relationship between financial news and the stock market. Sci Rep 3:3578
2. Askitas N, Zimmermann KF (2009) Google econometrics and unemployment forecasting. Appl Econ Q 55:107–120
3. Bordino I et al (2012) Web search queries can predict stock market volumes. PLOS One 7:e40014
4. Choi H, Varian H (2012) Predicting the present with Google Trends. Econ Rec 88:2–9
5. Conte R et al (2012) Manifesto of computational social science. Eur Phys J Spec Top 214:325–346

6. Corley M, Brocklehurst PH, Moat HS (2010) Error biases in inner and overt speech: evidence from tongue twisters. J Exp Psychol: Learn Mem Cognit 37:162–175
7. Feng L, Li B, Podobnik B, Preis T, Stanley HE (2012) Linking agent-based models and stochastic models of financial markets. Proc Natl Acad Sci USA 109:8388–8393
8. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2003) A theory of power-law distributions in financial market fluctuations. Nature 423:267–270
9. Ginsberg J et al (2009) Detecting influenza epidemics using search engine query data. Nature 457:1012–1014
10. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with web search. Proc Natl Acad Sci USA 107:17486–17490
11. Hommes CH (2002) Modeling the stylized facts in finance through simple nonlinear adaptive systems. Proc Natl Acad Sci USA 99:7221–7228
12. Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of a financial market. Nature 397:498–500
13. Mantegna RN, Stanley HE (2002) Scaling behaviour in the dynamics of an economic index. Nature 376:46–49
14. Moat HS (2010) Modelling subphonemic information flow: an investigation and extension of Dell's (1986) model of word production. Doctoral dissertation, University of Edinburgh
15. Moat HS, Corley M, Hartsuiker RJ (2008) Connecting phonological encoding to articulation: is cascading required? A computational investigation. In: Love BC, McRae K, Sloutsky VM (eds) Proceedings of the 30th annual conference of the Cognitive Science Society, Washington, DC, pp 1320–1325. Cognitive Science Society, Austin
16. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T (2013) Quantifying Wikipedia usage patterns before stock market moves. Sci Rep 3:1801
17. Moat HS, Preis T, Olivola CY, Liu C, Chater N (in press) Using big data to predict collective behavior in the real world. Behav Brain Sci
18. Preis T (2010) Quantifying and modeling financial fluctuations. Doctoral dissertation, University of Mainz
19. Preis T (2011) Econophysics—complex correlations and trend switchings in financial time series. Eur Phys J Spec Top 194:5–86
20. Preis T (2011) GPU-computing in econophysics and statistical physics. Eur Phys J Spec Top 194:87–119
21. Preis T, Stanley HE (2010) Switching phenomena in a system with no switches. J Stat Phys 138:431–446
22. Preis T, Golke S, Paul W, Schneider JJ (2006) Multi-agent-based order book model of financial markets. EPL 75:510
23. Preis T, Golke S, Paul W, Schneider JJ (2007) Statistical analysis of financial returns for a multiagent order book model of asset trading. Phys Rev E 76:016108
24. Preis T, Paul W, Schneider JJ (2008) Fluctuation patterns in high-frequency financial asset returns. EPL 82:68005
25. Preis T, Reith D, Stanley HE (2010) Complex dynamics of our economic life on different scales: insights from search engine query data. Philos Trans R Soc A 368:5707–5719
26. Preis T, Schneider JJ, Stanley HE (2011) Switching processes in financial markets. Proc Natl Acad Sci USA 108:7674–7678
27. Preis T, Virnau P, Paul W, Schneider JJ (2011) Accelerated fluctuation analysis by graphic cards and complex pattern formation in financial markets. New J Phys 11:093024
28. Preis T, Kenett DY, Stanley HE, Helbing D, Ben-Jacob E (2012) Quantifying the behavior of stock correlations under market stress. Sci Rep 2:752
29. Preis T, Moat HS, Stanley HE, Bishop SR (2012) Quantifying the advantage of looking forward. Sci Rep 2:350
30. Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using Google Trends. Sci Rep 3:1684
31. Preis T, Moat HS, Bishop SR, Treleaven P, Stanley HE (2013) Quantifying the digital traces of Hurricane Sandy on Flickr. Sci Rep 3:3141

32. Silverman BW (1986) Density estimation. Chapman and Hall, London
33. Simon HA (1955) A behavioral model of rational choice. Q J Econ 69:99–118
34. Tversky A, Kahneman D (1991) Loss aversion in riskless choice: a reference-dependent model. Q J Econ 106:1039–1061
35. Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer, New York