PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2010) **368**, 5707–5719
doi:10.1098/rsta.2010.0284

Complex dynamics of our economic life on different scales: insights from search engine query data

By Tobias Preis[1,2,3,*], Daniel Reith[3] and H. Eugene Stanley[1]

[1]*Center for Polymer Studies, Department of Physics, 590 Commonwealth Avenue, Boston, MA 02215, USA*
[2]*Artemis Capital Asset Management GmbH, Gartenstrasse 14, 65558 Holzheim, Germany*
[3]*Institute of Physics, Johannes Gutenberg University Mainz, Staudingerweg 7, 55128 Mainz, Germany*

Search engine query data deliver insight into the behaviour of individuals who are the smallest possible scale of our economic life. Individuals are submitting several hundred million search engine queries around the world each day. We study weekly search volume data for various search terms from 2004 to 2010 that are offered by the search engine Google for scientific use, providing information about our economic life on an aggregated collective level. We ask the question whether there is a link between search volume data and financial market fluctuations on a weekly time scale. Both collective 'swarm intelligence' of Internet users and the group of financial market participants can be regarded as a complex system of many interacting subunits that react quickly to external changes. We find clear evidence that weekly transaction volumes of S&P 500 companies are correlated with weekly search volume of corresponding company names. Furthermore, we apply a recently introduced method for quantifying complex correlations in time series with which we find a clear tendency that search volume time series and transaction volume time series show recurring patterns.

Keywords: econophysics; cross correlations; autocorrelations; financial markets; search engine queries; pattern recognition

1. Introduction

Econophysics research—*econophysics* forms the interdisciplinary interface between the two disciplines economics[1] and physics[2]—has been addressing a key question of interest in the subfield of financial markets: quantifying and

[1]Ancient Greek: οἰχονομία—management.
[2]Ancient Greek: φυσιχή τέχνη—art of handling nature.

*Author for correspondence (mail@tobiaspreis.de).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsta.2010.0284 or via http://rsta.royalsocietypublishing.org.

One contribution of 14 to a Theme Issue 'Complex dynamics of life at different scales: from genomic to global environmental issues'.

understanding large stock market fluctuations. Previous work was focused on the challenge of quantifying the behaviour of the probability distributions of large fluctuations of relevant variables such as returns, volumes and the number of transactions. Sampling the far tails of such distributions requires a large amount of data. However, there is a truly gargantuan amount of pre-existing precise financial market data already collected, many orders of magnitude more than for other complex systems. Accordingly, financial markets are becoming a paradigm of complex systems, and increasing numbers of scientists are analysing and modelling market data (Stanley *et al.* 1995; Vandewalle & Ausloos 1997; Cont & Bouchaud 2000; Krawiecki *et al.* 2002; Plerou *et al.* 2002*b*; Gabaix *et al.* 2003; Lillo *et al.* 2003; Kiyono *et al.* 2006; Preis *et al.* 2006, 2007; Watanabe *et al.* 2007; Podobnik *et al.* 2009). Empirical analyses have been focused on quantifying and testing the robustness of power-law distributions that characterize large movements in stock market activity. The use of estimators that are designed for serially and cross-sectionally independent data supports the hypothesis that the power-law exponents that characterize fluctuations in stock price, trading volume and the number of trades (Fama 1963; Lux & Marchesi 1999; Plerou *et al.* 2002*a*) are seemingly 'universal' in the sense that they do not change their values significantly for different markets, different time periods or different market conditions.

A reason why the economy is of interest to statistical physicists is that—like an Ising model which is a model of ferromagnetism—it is a system made up of many subunits. The subunits in an Ising model are the interacting spins, and the subunits in the economy are market participants—buyers and sellers. During any time interval, these subunits of the economy may be either positive or negative as regards perceived market opportunities. People interact with each other, and this fact often produces what economists call the herd effect. The orientation of whether they buy or sell is influenced not only by neighbours but also by news usually realized by an external field. If we hear bad news, we may be tempted to sell. So the state of any subunit is a function of the states of all the other subunits and of a field parameter (Preis & Stanley 2010).

One very illustrative example of the herd effect is shown in figure 1. The search engine Google offers the possibility to extract information about how popular are specific search terms'.[3] Thus, one can compare the interest in financial crisis related keywords, such as 'Subprime', 'Lehman Brothers' and 'Financial Crisis', with the fluctuations of the S&P 500 index that has the rank of an international benchmark index. It is easy to understand that peaks in the search volume for the term Subprime coincide with dips in the S&P 500 index time series. At the climax of the crisis, the collapse of Lehman Brothers caused the sell-out of stocks and the public was talking about the Financial Crisis afterwards. Figure 1 documents this course of time and shows that people acted with steadily increasing dynamic. The search volume profiles track the levels of escalation, which can be seen as a prominent example of the herd effect.

This kind of data provides insights into our economic life on different scales. A steadily increasing number of Internet users visit websites of search engines every day. Each query request can be seen as an individual vote: using search engines, we leave information about our interests codified as search terms. Thus, search

[3]More details can be found at http://www.google.com/trends.

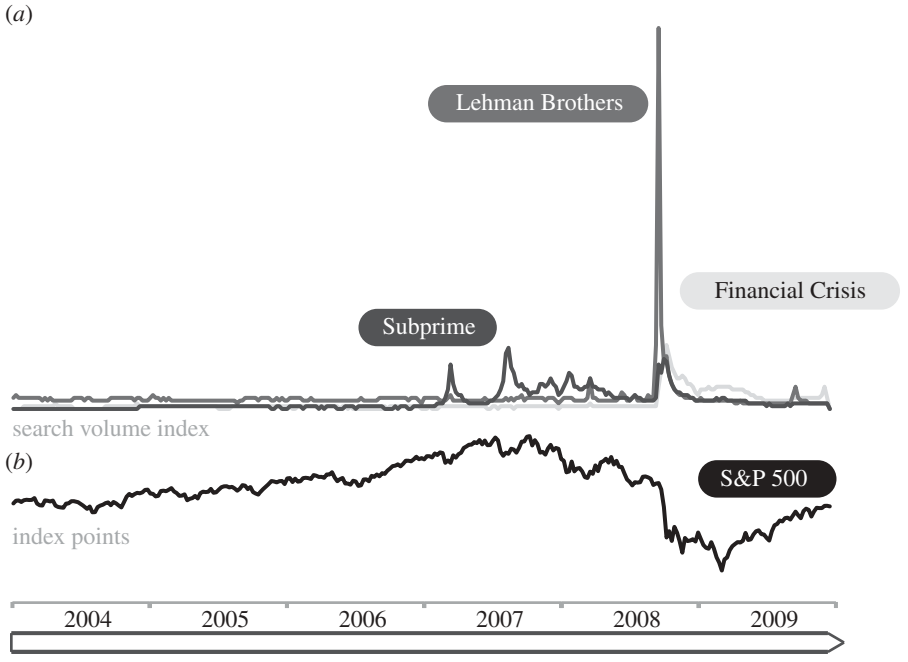

Figure 1. Keyword related search volume illustrates a behaviour similar to a *herd effect*. (*a*) Google Trends analyses a portion of Web searches to compute how many queries have been done for specified keyword terms. The relative number of queries can be plotted over time—here, the keyword terms 'Subprime', 'Lehman Brothers' and 'Financial Crisis' are plotted for the time period from 2004 to 2009. The peak of Lehman Brothers coincides with the bankruptcy of this institution when the investment bank Lehman Brothers filed for chapter 11 bankruptcy protection. (*b*) The US stock index S&P 500 is shown for the same period of time.

engines can collect our interests on the smallest possible scale—the scale of individual requests. On larger time scales, our interest forms trends. Aggregated search volume data can be used for uncovering such trends that affect our economic life on large scales. As seen before, the international financial crisis is one prominent example. However, product trends can be extracted as well—an example for that is the cell phone market. Search volume data provided by Google can also be used to predict spreading of seasonal influenza (Ginsberg *et al.* 2009). In addition, correlations were found linking both the current level of economic activity in given industries and search volume data of industry based query terms (Choi & Varian 2009).

The 'experimental basis' of the interdisciplinary science *econophysics* is given by time series that can be used in their raw form or from which one can derive observables. Such historical price curves can be understood as a macroscopic variable for underlying microscopic processes. The price fluctuations are produced by the superposition of individual actions of market participants, thereby generating cumulative supply and demand for a traded asset—e.g. a stock. The analogue in statistical physics is the emergence of macroscopic properties, which is caused by microscopic interactions among involved subunits.

In this paper, we will ask the question whether there is a link between search volume data and financial market fluctuations. For this task, we study cross correlations between the 'collective intelligence' of Internet users and the change of financial market quantities—weekly stock prices and weekly stock volume. In addition, we apply a method to find complex correlations in search volume data, which was recently introduced by Preis *et al.* (2008). Uncovering mechanisms and dependencies, which are useful to understand the formation of financial crises, is of crucial importance as an effective crises observatory could contribute in protecting the stability of financial systems.

This article is structured as follows. Section 2 describes the datasets that we analyse. In §3, we present correlation analyses between financial market fluctuations and search volume data. In §4, we analyse complex correlations in financial data and search volume data. Finally, §5 summarizes our results.

2. Data analysed

We use weekly closing prices of $N = 500$ US stocks, which were constituents of the S&P 500 index on 31 May 2010. These weekly datasets also contain aggregated transaction volumes covering the time period from the calendar week of 4 January 2004 until the calendar week of 30 May 2010. Thus, $T = 335 \times N = 167\,500$ weekly closing prices and weekly transaction volumes are available for analysis. A detailed list of the S&P 500 index components can be found in the electronic supplementary material. This list contains the exchange trading symbols and the company names.

In order to investigate whether Internet search volume is correlated with financial market fluctuations, we use search volume data provided by the search engine Google, which is available for the same period of time. This service which is called Google Trends analyses a portion of Google Web searches to compute how many searches have been done for specific terms, relative to the total number of searches done on Google over time—here we use all 500 company names of the S&P 500 components. As exact company names—e.g. *Microsoft Corporation*—may result in a weaker search volume quality in comparison to common abbreviations—e.g. *Microsoft*— we optimize the list of company names in order to improve the data quality and availability. The company names that are used for our search volume data requests can be found in the electronic supplementary material.

3. Linear autocorrelations and linear cross correlations

The *Pearson* product-moment correlation coefficient is a measure of the correlation between two variables $X(t)$ and $Y(t)$, giving a value between $+1$ and -1 inclusive (Pearson 1895). This correlation coefficient is widely used as a measure of the strength of linear dependence between two variables. In our case, $X_n(t)$ and $Y_n(t)$ are time series—the change of closing price, $p(t+1) - p(t)$, the change of volume, $v(t+1) - v(t)$, or the change of search volume, $s(t+1) - s(t)$—of stock n with length $T - 1$. As we would like to determine the correlation coefficient in dependence of a time lag parameter Δt,

we use $t \in \{1, 2, \ldots, T - 1 - \Delta t\}$. Thus, the correlation coefficient for stock n ($n \in \{1, 2, \ldots, N\}$) is given by

$$\rho^n_{X,Y}(\Delta t) = \frac{\langle X_n(t)\, Y_n(t + \Delta t)\rangle - \langle X_n(t)\rangle \langle Y_n(t + \Delta t)\rangle}{\sqrt{\langle X_n^2(t)\rangle - \langle X_n(t)\rangle^2}\sqrt{\langle Y_n^2(t + \Delta t)\rangle - \langle Y_n(t + \Delta t)\rangle^2}}, \tag{3.1}$$

with $\langle \ldots \rangle$ denoting the expectation value. Only non-vanishing changes of time series $X_n(t)$ and $Y_n(t)$ are considered as, for example, search volume data are not available for a few search terms at all observation times. Thus, let $T'_n(\Delta t)$ be the number of non-vanishing time series changes of stock n in dependence of Δt. The aggregated correlation coefficient of the set of stocks is calculated by

$$\bar{\rho}_{X,Y}(\Delta t) = \frac{\sum_{n=1}^{N} T'_n(\Delta t) \cdot \rho^n_{X,Y}(\Delta t)}{\sum_{n=1}^{N} T'_n(\Delta t)}. \tag{3.2}$$

For the analysis of cross correlations and autocorrelations ($Y_n(t) = X_n(t)$), we assume that the underlying variables $X_n(t)$ and $Y_n(t)$ have a bivariate normal distribution. Thus, we can use the Fisher transformation (Fisher 1915) for the determination of time lag-dependent confidence intervals. The Fisher transformation of $\bar{\rho}_{X,Y}$ is given by

$$F(\bar{\rho}_{X,Y}) = \frac{1}{2} \ln \frac{1 + \bar{\rho}_{X,Y}}{1 - \bar{\rho}_{X,Y}}. \tag{3.3}$$

For the z-score,

$$z = \sqrt{\left(\sum_{n=1}^{N} T'_n\right) - 3} \cdot F(\bar{\rho}_{X,Y}), \tag{3.4}$$

we obtain the confidence intervals from cumulative distribution function values for the standard normal distribution. An inverted Fisher transformation provides confidence intervals on a correlation scale.

First, we study autocorrelations $\bar{\rho}_{X,X}(\Delta t)$. In figure 2*a*, the autocorrelation coefficients of weekly closing price changes are shown in dependence of Δt. Almost all values are practically negligible and are located close to the 95% confidence interval. Only the negative autocorrelation coefficient at time lag $\Delta t = 1$ week seems to be relevant and reminds us that high-frequency financial market transaction prices exhibit a strong negative autocorrelation at the smallest possible time lag (larger than the trivial case of $\Delta t = 0$) on time scales of individual transactions—Preis *et al.* (2008) report a value of roughly -0.30 for the German DAX Futures contract. On the contrary, the autocorrelation functions of volume changes (figure 2*b*) and search volume changes (figure 2*c*) provide significantly negative values for small time lags ($\Delta t < 4$ weeks).

Figure 3 illustrates cross correlations between weekly closing price changes and search volume changes and between weekly transactions volume changes and search volume changes for one proxy of the S&P 500 index—Apple Incorporated. There are no significant correlations between price changes and search volume changes (figure 3*a*). All values are within the 95% confidence interval. However, increasing/decreasing transaction volumes of this stock

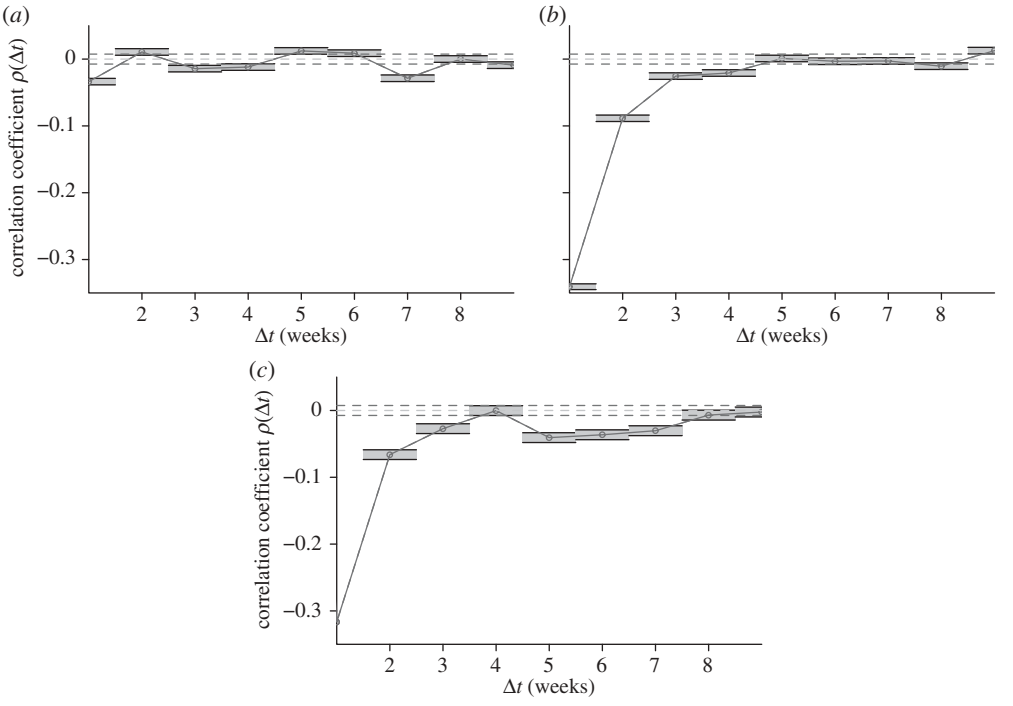

Figure 2. (*a*) Time lag-dependent autocorrelation $\bar{\rho}_{X,X}(\mathbf{\Delta}t)$ of weekly closing price changes. 95% confidence intervals are displayed as rectangles. The dashed lines are 95% confidence interval for autocorrelations of an independent and identically distributed random variables (i.i.d.) process. (*b*) Time lag-dependent autocorrelation $\bar{\rho}_{X,X}(\mathbf{\Delta}t)$ of weekly volume changes. (*c*) Time lag-dependent autocorrelation $\bar{\rho}_{X,X}(\mathbf{\Delta}t)$ of weekly search volume changes.

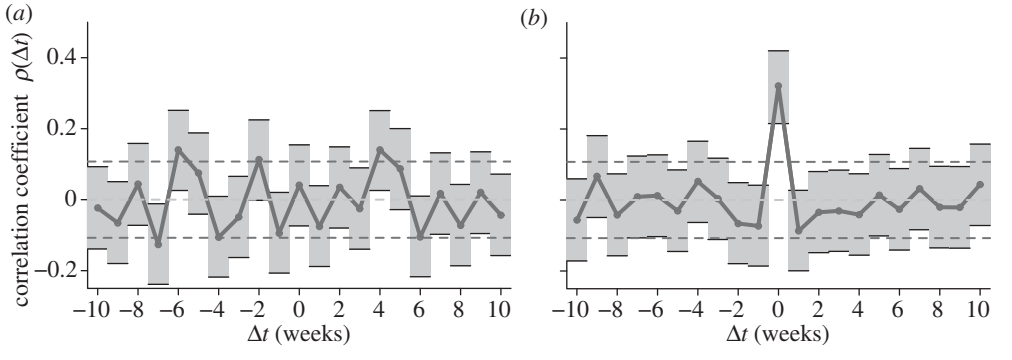

Figure 3. Cross correlations for a single stock—Apple Inc. (*a*) Time lag-dependent cross correlation $\rho_{X,Y}(\mathbf{\Delta}t)$ between weekly closing price changes of Apple stock and weekly search volume changes of the search term 'apple'. Rectangles indicate 95% confidence intervals. Again, the dashed lines are 95% confidence interval for cross correlations of independent and identically distributed random variables (i.i.d.) processes. (*b*) Time lag-dependent cross correlation $\rho_{X,Y}(\mathbf{\Delta}t)$ between weekly transaction volume changes of Apple stock and weekly search volume changes of the search term 'apple'.

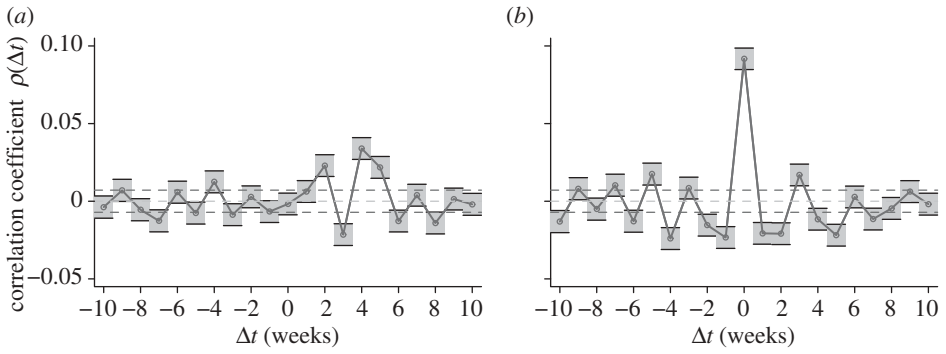

Figure 4. Aggregated cross correlations for all S&P 500 stocks. (*a*) Time lag-dependent cross correlation $\tilde{\rho}_{X,Y}(\Delta t)$ between weekly closing price changes and weekly search volume changes of corresponding company names. (*b*) Time lag-dependent cross correlation $\tilde{\rho}_{X,Y}(\Delta t)$ between weekly transaction volume changes and weekly search volume changes of corresponding company names.

coincide with increasing/decreasing search volumes as one can see at time lag $\Delta t = 0$ weeks in figure 3*b*. Thus, one can conclude that search volume reflects the present attractiveness for trading a stock. But it seems that neither buying transactions nor selling transactions are preferred. This example shows that the commonly accepted reasons for financial market movements—'news moves the market' and 'volume moves the market'—are clearly linked together because news should be the most probably reason for searching company names in Internet search engines. The same effect can be found for aggregated correlation coefficients of all S&P 500 constituents (figure 4), even if the correlation coefficient at time lag $\Delta t = 0$ weeks (figure 4*b*) is smaller than for the single stock, Apple Incorporated. In figure 4*a*, a few correlation coefficients ($\Delta t \approx 4$ weeks) are not in the 95% confidence interval indicating a non-random correlation between weekly closing price changes and search volume changes. In fact, present price movements seem to influence the search volume in the following weeks. However, the correlation coefficients are very small, $|\tilde{\rho}_{X,Y}| < 0.05$. Thus, confirming analyses with more records are necessary. Unfortunately, Google Trends offer search volume data only on a weekly basis.

4. Pattern conformity

These results raise hopes that complex correlations exist on weekly time scales in the data. A sophisticated observable to quantify them was introduced in a recent work (Preis *et al.* 2008).[4] This work was focused on finding complex correlations in high-frequency financial market datasets. In such a context, the existence of complex correlations implies that market participants—human traders and most notably automated trading algorithms—react to a given time series pattern

[4]This approach consumes a huge amount of computing time. However, an accelerated calculation is possible on graphic card architectures (Preis *et al.* 2009*a*,*b*) which can also be used in computational physics (Block *et al.* 2010).

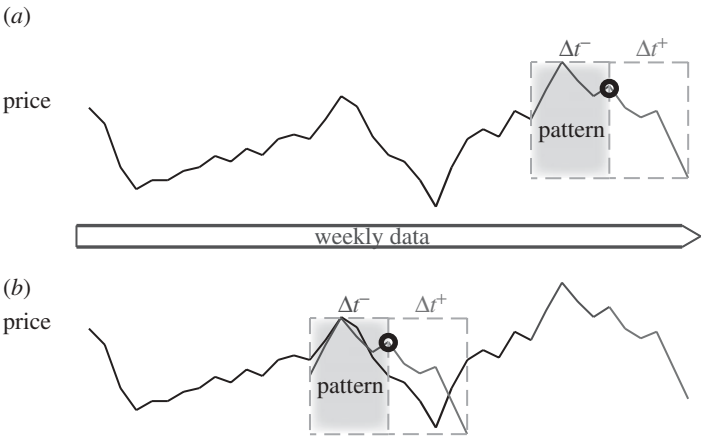

Figure 5. Pattern conformity analysis of financial market fluctuations. The aim is to compare (*a*) a current pattern of a certain time interval length Δt^- with (*b*) all possible previous patterns of the time series.

just like to comparable patterns in the past (figure 5). However, this concept is transferable to medium and large time scales. To quantify additional correlations, we will define a pattern conformity (PC) observable.

The aim is to compare the current reference pattern of time interval length Δt^- with all previous patterns in the time series $p(t)$. The current observation time shall be denoted by $\hat{t}$, then the reference interval is given by $[\hat{t}-\Delta t^-;\hat{t})$. The forward evolution after this current reference interval—the distance to $\hat{t}$ is expressed by Δt^+—is compared with the prediction derived from historical patterns. As the standard deviation is not constant in time, all comparison patterns have to be normalized with respect to the current reference pattern. Thus, we use the true range—the difference between high and low. Let $p_{\mathrm{h}}(\hat{t},\Delta t^-)$ be the maximum value of a pattern of length Δt^- at time $\hat{t}$ and analogously $p_{\mathrm{l}}(\hat{t},\Delta t^-)$ be the minimum value. Note that $p(t)$, $p_{\mathrm{h}}(\hat{t},\Delta t^-)$ and $p_{\mathrm{l}}(\hat{t},\Delta t^-)$ depend also on n, the specific stock. However, we waive the corresponding superscript to improve the readability. We construct a modified time series, which is true range adapted in the appropriate time interval, through

$$\tilde{p}_{\hat{t}}^{\Delta t^-}(t)=\frac{p(t)-p_{\mathrm{l}}(\hat{t},\Delta t^-)}{p_{\mathrm{h}}(\hat{t},\Delta t^-)-p_{\mathrm{l}}(\hat{t},\Delta t^-)} \tag{4.1}$$

with $\tilde{p}_{\hat{t}}^{\Delta t^-}(t)\in[0;1]\ \forall\ t\in[\hat{t}-\Delta t^-;\hat{t})$, as illustrated in figure 6. At this point, the fit quality $Q_{\hat{t}}^{\Delta t^-}(\tau)$ between the current reference sequence $\tilde{p}_{\hat{t}}^{\Delta t^-}(t)$ and a comparison sequence $\tilde{p}_{\hat{t}-\tau}^{\Delta t^-}(t-\tau)$ for $t\in[\hat{t}-\Delta t^-;\hat{t})$ has to be determined by a least mean square fit through

$$Q_{\hat{t}}^{\Delta t^-}(\tau)=\sum_{\theta=1}^{\Delta t^-}\frac{(\tilde{p}_{\hat{t}}^{\Delta t^-}(\hat{t}-\theta)-\tilde{p}_{\hat{t}-\tau}^{\Delta t^-}(\hat{t}-\tau-\theta))^2}{\Delta t^-} \tag{4.2}$$

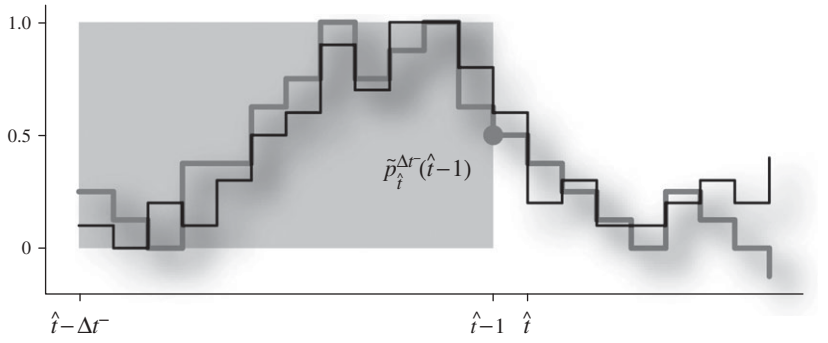

Figure 6. Schematic visualization of the pattern conformity (PC) calculation mechanism. The normalized reference pattern $\tilde{p}_{\hat{t}}^{\Delta t^-}(t)$ and the by τ shifted comparison pattern $\tilde{p}_{\hat{t}-\tau}^{\Delta t^-}(t-\tau)$ have the maximum value 1 and the minimum value 0 in $[\hat{t}-\Delta t^-;\hat{t})$, as illustrated by the filled rectangle. For the PC calculation, it will be checked for each time interval Δt^+ starting at $\hat{t}$ whether reference and comparison pattern are above or below the last value of the reference pattern $\tilde{p}_{\hat{t}}^{\Delta t^-}(\hat{t}-1)$. If both are above or below this level, then $+1$ is added to the non-normalized PC. If one is above and the other below, then -1 is added. Grey line, $\tilde{p}_{\hat{t}}^{\Delta t^-}(t)$; black line, $\tilde{p}_{\hat{t}-\tau}^{\Delta t^-}(t-\tau)$.

with $Q_{\hat{t}}^{\Delta t^-}(\tau)\in[0,1]$ as a result of the true range adaption. With these elements, one can define an observable for the PC, which is not yet normalized by

$$\xi_\chi(\Delta t^+,\Delta t^-)=\sum_{\hat{t}=\Delta t^-}^{T-\Delta t^+}\sum_{\tau=\Delta t^-}^{\hat{t}}\frac{\operatorname{sgn}(\omega_{\hat{t}}^{\Delta t^-}(\tau,\Delta t^+))}{\exp\left(\chi Q_{\hat{t}}^{\Delta t^-}(\tau)\right)}, \tag{4.3}$$

as motivated in figure 6. Furthermore, we use the definition

$$\operatorname{sgn}(x)=\begin{cases}1 & \text{for } x>0\\ 0 & \text{for } x=0\\ -1 & \text{for } x<0.\end{cases} \tag{4.4}$$

The parameter χ weights terms according to their qualities (Preis *et al.* 2008). The larger χ is, the stricter the pattern weighting in order to use only sequences with good agreement to the reference pattern. The expression $\omega_{\hat{t}}^{\Delta t^-}(\tau,\Delta t^+)$ in equation (4.3), which takes into account the value of reference and comparison pattern after $\hat{t}$ for a proposed Δt^+ relative to $\tilde{p}_{\hat{t}}^{\Delta t^-}(\hat{t}-1)$, is given by the following expression:

$$\begin{aligned}\omega_{\hat{t}}^{\Delta t^-}(\tau,\Delta t^+)=&(\tilde{p}_{\hat{t}}^{\Delta t^-}(\hat{t}-1+\Delta t^+)-\tilde{p}_{\hat{t}}^{\Delta t^-}(\hat{t}-1))\\ &\times(\tilde{p}_{\hat{t}-\tau}^{\Delta t^-}(\hat{t}-\tau-1+\Delta t^+)-\tilde{p}_{\hat{t}}^{\Delta t^-}(\hat{t}-1)).\end{aligned} \tag{4.5}$$

We normalize the observable for PC and obtain for stock n

$$\Xi_\chi^n(\Delta t^+,\Delta t^-)=\frac{\xi_\chi(\Delta t^+,\Delta t^-)}{\sum_{\hat{t}=\Delta t^-}^{T-\Delta t^+}\sum_{\tau=\Delta t^-}^{\hat{t}}|\operatorname{sgn}(\omega_{\hat{t}}^{\Delta t^-}(\tau,\Delta t^+))|/\exp\left(\chi Q_{\hat{t}}^{\Delta t^-}(\tau)\right)}, \tag{4.6}$$

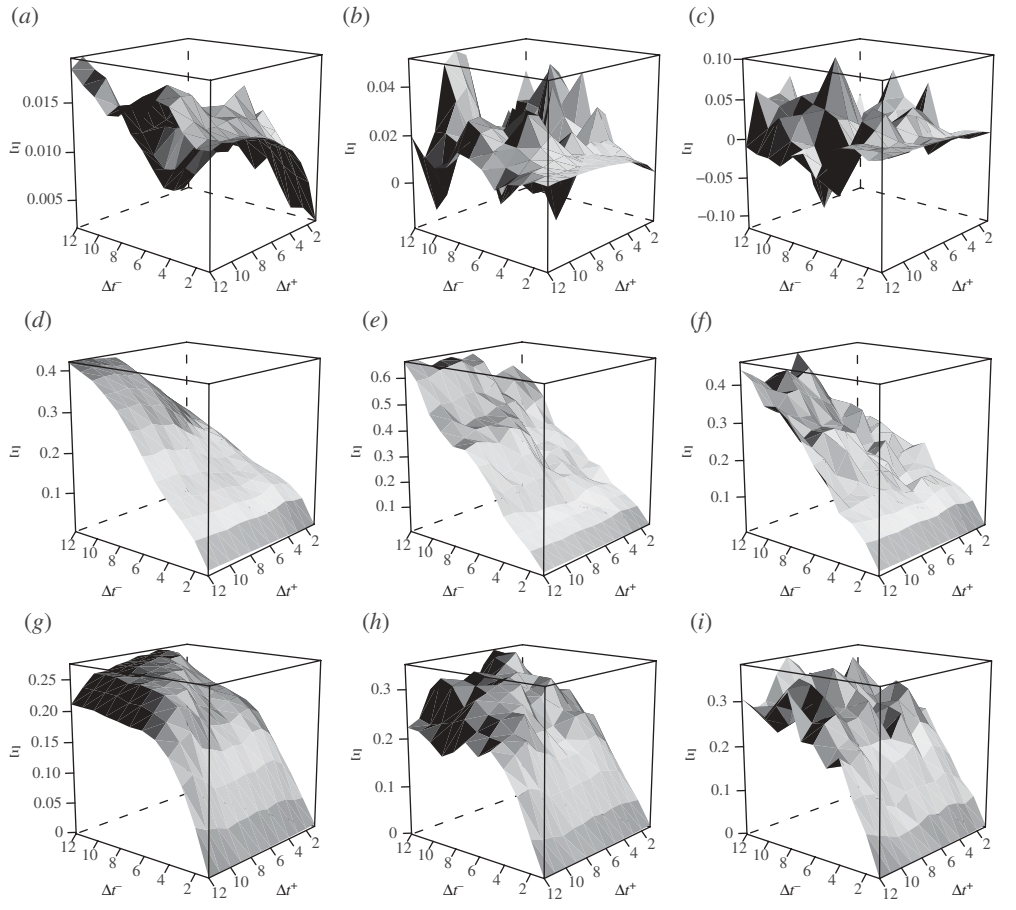

Figure 7. PC—Δt is measured in units of weeks. (a) PC $\Xi^{\mathrm{P}}_{\chi=100}(\Delta t^-, \Delta t^+)$ of weekly closing prices. (b) Identical to (a), but the fit quality of a pattern is not only calculated by prices: appropriate transaction volumes are incorporated, too. (c) Identical to (a), but appropriate search volumes are incorporated, too. (d) PC $\Xi^{\mathrm{S}}_{\chi=100}(\Delta t^-, \Delta t^+)$ of weekly search volumes. (e) Identical to (d), but appropriate transaction volumes are incorporated, too. (f) Identical to (d), but appropriate prices are incorporated, too. (g) PC $\Xi^{\mathrm{V}}_{\chi=100}(\Delta t^-, \Delta t^+)$ of weekly transaction volumes. (h) Identical to (g), but appropriate prices are incorporated, too. (i) Identical to (g), but appropriate search volumes are incorporated, too.

where $\Xi^n_\chi(\Delta t^+, \Delta t^-)$ denotes the PC of a stock n. In order to obtain an aggregated quantity of all S&P 500 stocks, we define

$$\Xi_\chi(\Delta t^+, \Delta t^-) = \frac{1}{N} \sum_{n=1}^{N} \Xi^n_\chi(\Delta t^+, \Delta t^-). \tag{4.7}$$

The PC for a standard random walk time series, which exhibits no correlations by construction, is 0 for all pairs of Δt^+ and Δt^-. The PC for a perfectly correlated time series—a straight line—is 1. With this method, it is possible to search for complex correlations in various time series.

Figure 7 shows the PCs of weekly closing prices (figure 7*a*), weekly search volumes (figure 7*d*) and weekly transaction volumes (figure 7*g*). The tendency to reproduce historic price patterns is very small for weekly closing prices (figure 7*a*). It is difficult to distinguish the given PC from a completely random behaviour (Preis *et al.* 2008). So far, the comparison between reference and historic patterns was only based on the price time series, $Q_{\hat{t}}^{\Delta t^-}(\tau) = Q_{\hat{t}}^{\mathrm{P},\Delta t^-}(\tau)$. Now, we also incorporate the time series of transaction volumes $v(t)$, i.e. $Q_{\hat{t}}^{\Delta t^-}(\tau) = Q_{\hat{t}}^{\mathrm{P},\Delta t^-}(\tau) + Q_{\hat{t}}^{\mathrm{V},\Delta t^-}(\tau)$, to improve the pattern selection. In the same way, it is possible to include the search volume time series $s(t)$ for the pattern selection, i.e. $Q_{\hat{t}}^{\Delta t^-}(\tau) = Q_{\hat{t}}^{\mathrm{P},\Delta t^-}(\tau) + Q_{\hat{t}}^{\mathrm{S},\Delta t^-}(\tau)$. If we include transaction volumes for the selection process (figure 7*b*), then we obtain a noisier PC profile. A still noisier profile can be achieved by using search volume time series as an additional pattern selection criterion (figure 7*c*). Clear recurring tendencies can be found for the search volume time series. Figure 7*d* shows significant non-zero values for the PC. In contrast to results obtained for high-frequency transactions, parameter pairs with large time lags Δt^+ and Δt^- provide the highest level of PC of roughly 0.42—due to the given amount of data points we limit the analyses to the range from one week to three months. The additional incorporation of weekly transaction volumes (figure 7*e*) increases the maximum value of the PC in the range that we analyse. The maximum value is roughly 0.66. This fact supports our finding that there is a clear link between weekly transaction volumes and weekly search volumes. More important, there is not only a linear dependence as found in §3 but also complex dependencies uncovered by the PC approach. Thus, it is evidence that search volume time series and transaction volume time series show recurring patterns. On the contrary, the inclusion of weekly closing prices does not alter the PC significantly (figure 7*f*). Analogously, transaction volume time series are characterized by large PC values (figure 7*g*) that are slightly smaller than in figure 7*d*. If one also incorporates closing price times series (figure 7*h*) or search volume time series (figure 7*i*) for the pattern selection, then a slightly increased PC can be observed.

5. Conclusion

Search engine query data offer insights into our economic life on the smallest possible scale of individual actions. In order to investigate whether Internet search volume is correlated with financial market fluctuations—the largest possible scale of our economic life—we used search volume data provided by the search engine Google. We studied weekly search volume data for various search terms from 2004 to 2010. We asked the question whether there is a link between search volume data and financial market fluctuations on the same, weekly time scale and found clear evidence that weekly transaction volumes of S&P 500 companies are correlated with weekly search volume of the corresponding company names. Increasing transaction volumes of stocks coincide with an increasing search volume and vice versa. Thus, one can conclude that search volume reflects the present attractiveness of trading a stock. But it seems that neither buying transactions nor selling transactions are preferred when one detects an increased search

volume. Thus, the commonly accepted reasons for financial market movements—news and volume—are clearly linked together because news should be the most likely reason for searching company names in Internet search engines. In addition, we have seen that present price movements seem to influence the search volume of the corresponding company name in the following weeks.

Furthermore, we applied a recently introduced method for quantifying complex correlations in time series with which we find the clear tendency that search volume time series and transaction volume time series show recurring patterns. This fact supports our finding that there is a clear link between weekly transaction volumes and weekly search volumes. More important, there is not only a linear dependence but also complex dependencies, which raises hopes that search volume data can contribute to understand financial crises.

The authors are very grateful for helpful discussions with D. Helbing, P. Virnau and K. Yamasaki. In addition, T.P. would like to thank D. Diefenbach for insightful comments.

References

Block, B., Virnau, P. & Preis, T. 2010 Multi-GPU accelerated multi-spin Monte Carlo simulations of the 2D Ising model. *Comp. Phys. Commun.* **181**, 1549–1556. (doi:10.1016/j.cpc.2010.05.005)

Choi, H. & Varian, H. 2009 *Predicting the present with Google Trends*. Working Paper. Mountain View, CA: Google Inc.

Cont, R. & Bouchaud, J.-P. 2000 Herd behavior and aggregate fluctuations in financial markets. *Macroecon. Dyn.* **4**, 170–196. (doi:10.1017/S1365100500015029)

Fama, E. F. 1963 Mandelbrot and the stable paretian hypothesis. *J. Bus.* **36**, 420–429. (doi:10.1086/294633)

Fisher, R. A. 1915 Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507–521.

Gabaix, X., Gopikrishnan, P., Plerou, V. & Stanley, H. E. 2003 A theory of power-law distributions in financial market fluctuations. *Nature* **423**, 267–270. (doi:10.1038/nature01624)

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. 2009 Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014. (doi:10.1038/nature07634)

Kiyono, K., Struzik, Z. R. & Yamamoto, Y. 2006 Criticality and phase transition in stock-price fluctuations. *Phys. Rev. Lett.* **96**, 068701. (doi:10.1103/PhysRevLett.96.068701)

Krawiecki, A., Holyst, J. A. & Helbing, D. 2002 Volatility clustering and scaling for financial time series due to attractor bubbling. *Phys. Rev. Lett.* **89**, 158701. (doi:10.1103/PhysRevLett.89.158701)

Lillo, F., Farmer, J. D. & Mantegna, R. N. 2003 Econophysics: master curve for price-impact function. *Nature* **421**, 129–130. (doi:10.1038/421129a)

Lux, T. & Marchesi, M. 1999 Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* **397**, 498–500. (doi:10.1038/17290)

Pearson, K. 1895 Contributions to the mathematical theory of evolution II: skew variations in homogeneous material. *Phil. Trans. R. Soc. Lond. A* **186**, 343–414. (doi:10.1098/rsta.1895.0010)

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T. & Stanley, H. E. 2002*a* Random matrix approach to cross correlations in financial data. *Phys. Rev. E* **65**, 066126. (doi:10.1103/PhysRevE.65.066126)

Plerou, V., Gopikrishnan, P., Gabaix, X. & Stanley, H. E. 2002*b* Quantifying stock-price response to demand fluctuations. *Phys. Rev. E* **66**, 027104. (doi:10.1103/PhysRevE.66.027104)

Podobnik, B., Horvatic, D., Petersen, A. M. & Stanley, H. E. 2009 Cross-correlations between volume change and price change. *Proc. Natl Acad. Sci. USA* **106**, 22 079–22 084. (doi:10.1073/pnas.0911983106)

Preis, T. & Stanley, H. E. 2010 Switching phenomena in a system with no switches. *J. Stat. Phys.* **138**, 431446. (doi:10.1007/s10955-009-9914-y)

Preis, T., Golke, S., Paul, W. & Schneider, J. J. 2006 Multiagent-based order book model of financial markets. *Europhys. Lett.* **75**, 510–516. (doi:10.1209/epl/i2006-10139-0)

Preis, T., Golke, S., Paul, W. & Schneider, J. J. 2007 Statistical analysis of financial returns for a multiagent order book model of asset trading. *Phys. Rev. E* **76**, 016108. (doi:10.1103/PhysRevE.76.016108)

Preis, T., Paul, W. & Schneider, J. J. 2008 Fluctuation patterns in high-frequency financial asset returns. *Europhys. Lett.* **82**, 68005. (doi:10.1209/0295-5075/82/68005)

Preis, T., Virnau, P., Paul, W. & Schneider, J. J. 2009*a* Accelerated fluctuation analysis by graphic cards and complex pattern formation in financial markets. *New J. Phys.* **11**, 093024. (doi:10.1088/1367-2630/11/9/093024)

Preis, T., Virnau, P., Paul, W. & Schneider, J. J. 2009*b* GPU accelerated Monte Carlo simulation of the 2D and 3D Ising model. *J. Comp. Phys.* **228**, 4468–4477. (doi:10.1016/j.jcp.2009.03.018)

Stanley, M. H. R., Buldyrev, S. V., Havlin, S., Mantegna, R. N., Salinger, M. A. & Stanley, H. E. 1995 Zipf plots and the size distribution of firms. *Econ. Lett.* **49**, 453–457. (doi:10.1016/0165-1765(95)00696-D)

Vandewalle, N. & Ausloos, M. 1997 Coherent and random sequences in financial fluctuations. *Physica A* **246**, 454–459. (doi:10.1016/S0378-4371(97)00366-X)

Watanabe, K., Takayasu, H. & Takayasu, M. 2007 A mathematical definition of the financial bubbles and crashes. *Physica A* **383**, 120–124. (doi:10.1016/j.physa.2007.04.093)