Supplementary Material: Quantifying the diversity of news around stock market moves

Chester Curme,[1,2*] Ying Daisy Zhuo,[2,3] Helen Susannah Moat,[2,4] Tobias Preis[2,4]

[1]Center for Polymer Studies and Department of Physics, Boston University,
590 Commonwealth Avenue, Boston, MA 02215, USA
[2]Data Science Lab, Warwick Business School,
Scarman Road, CV4 7AL Coventry, United Kingdom
[3]Operations Research Center, Massachusetts Institute of Technology,
77 Massachusetts Avenue, Cambridge, MA 02139, USA
[4]The Alan Turing Institute, British Library,
96 Euston Road, NW1 2DB London, United Kingdom

[*]To whom correspondence should be addressed; E-mail: ccurme@bu.edu.

1 Financial data

Closing price figures and daily trade volumes for the FTSE 100 index were obtained from Yahoo Finance (`https://uk.finance.yahoo.com/`).

2 *Financial Times* text preprocessing

We analyze a corpus of daily issues of the *Financial Times* from January 2, 2007 to December 31, 2012. Issues were retrieved from `http://www.ft.com/` in Portable Document Format (PDF). All issues were retrieved for this period, with the exception of five dates due to technical problems. These dates were February 22, 2007, March 8, 2007, May 12, 2007, January 28, 2009, and November 8, 2012. Each PDF was converted to text format (.txt) using the open source software *pdftotext*, which is freely available and included in most Linux distributions.

Documents for input to the Latent Dirichlet Allocation (LDA) were defined as blocks of text that were separated by isolated newline sequences "\n" and contained greater than 30 words. All characters were processed to unicode, forced to lowercase, and hyphens were replaced with whitespace. All characters other than the letters "a" to "z" were removed. The remaining text was then stemmed using the Porter stemming algorithm (Porter 1980), cleaned of single-letter words, and cleaned of stemmed stopwords. We used the MySQL stopword list (`http://dev.mysql.com/doc/refman/5.6/en/fulltext-stopwords.html`), supplemented with the words "ft", "financial" "times", "xd", "gbp", "usd", "euro", "acc", "eur", "page", "per", "cent", and "mr". Processed documents containing fewer than 30 words were removed.

3 Latent Dirichlet Allocation

We configure a weighted LDA (Hazen 2010, Řehůřek & Sojka 2010, Wilson & Chew 2010, Xu et al. 2013) to model each document as a mixture of $K = 50$ topics. In order to reduce the influence of common words when identifying topics, we weight word counts inversely to their frequency in the entire corpus, using the TF-IDF weighting scheme for individual words (Salton & McGill 1986). We find that this scheme helps to control for certain words that were abundant in the financial literature, but absent from conventional stopword lists.

The *gensim* Python package (Řehůřek & Sojka 2010) was used for the LDA on the full set of processed documents. We configured a batch LDA, with ten passes over the entire corpus. The top ten (stemmed) words for each of the 50 topics are provided in Table S1.

Once the LDA is trained, each document d in the corpus is represented by the K-dimensional topic vector $\theta_d = (\theta_{d,1}, \theta_{d,2}, ..., \theta_{d,K})$. The terms in this vector may be interpreted as probabilities, and therefore sum to one. In order to quantify the distribution of topics in the financial news on a given day, we computed a normalized sum of the distribution of topics over each document (paragraph) in the corresponding issue of the *Financial Times*. That is, from the set of documents $\mathcal{D}_t$ in the *Financial Times* issue on day t, we construct the vector

$$\rho_t \equiv \frac{1}{|\mathcal{D}_t|} \sum_{d \in \mathcal{D}_t} \theta_d, \tag{1}$$

where $|\mathcal{D}_t|$ denotes the number of documents in the set $\mathcal{D}_t$. This vector also sums to one, and quantifies the distribution of topics represented in the *Financial Times* on day t.

The collection of all ρ_t form the rows of a matrix ρ, which provides rich information regarding both the detailed and large-scale structure of news to which investors, traders, and the public are exposed. The columns of ρ, for example, represent time series of weights for individual topics in the *Financial Times*. Analyses of these individual time series can provide insight into the ebbs and flows of stories into and out of public attention. Figure S1 depicts the autocorrelation functions (ACF) for two topic time series $\rho^T_{k,t}$. In Fig. S1A we show the ACF for a topic regarding events in Egypt (“mubarak”, “egypt”, “protest”,...), while in Fig. S1B we show the same for a topic regarding events in Korea (“korea”, “seoul”, “kim”,...). These two represent topics with slow and fast decays in their autocorrelation functions, respectively. We quantify the lifetime of a topic as the first lag (in weekdays) at which the ACF falls within the 95% confidence bands for an uncorrelated signal. In Fig. S1C we show the distribution of all 50 topic lifetimes. Some lifetimes are on the order of years, but these tend to constitute topics which occur regularly in issues of the *Financial Times* (e.g., topics relating to weather reports, or market performance). 50% of topics have lifetimes shorter than 13 weekdays. Note that these calculations exclude weekend issues of the *Financial Times*. Such analyses, while simple, give valuable insight into “meta characteristics” that may be common to distinct topics in the news.

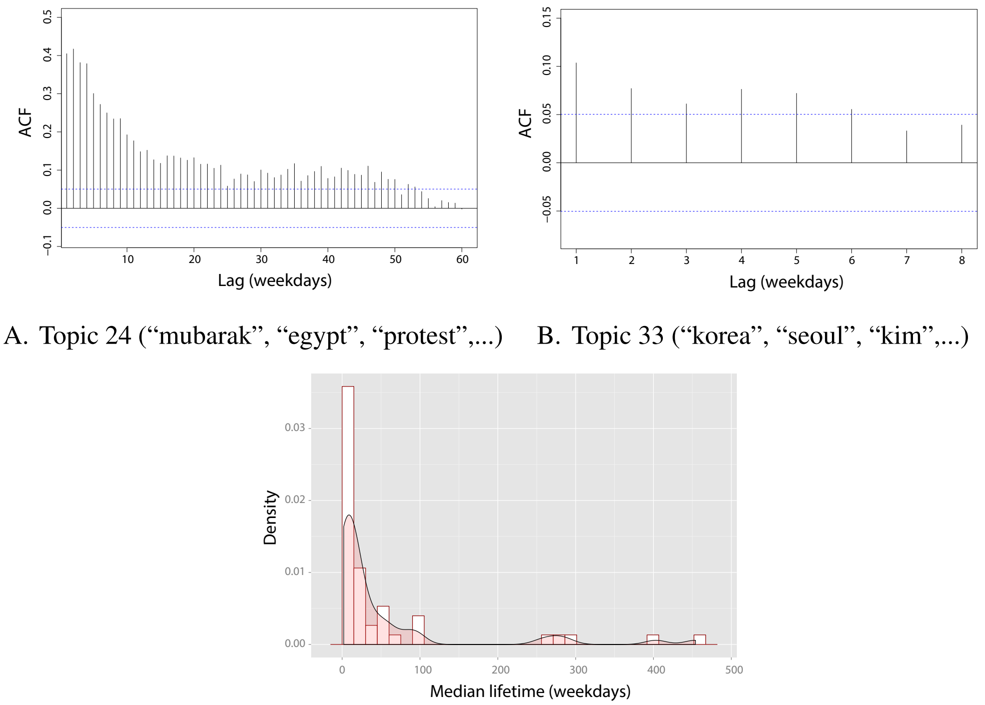
We also examine the presence of weekly and monthly seasonalities in the diversity H_t in Fig. S2. In Fig. S2A we observe characteristically low values of the diversity in weekend issues of the *Financial Times*, as noted in the main text. In Fig. S2B we show the seasonal variation in the diversity H_t, excluding weekend issues. We observe little seasonal variation in H_t, although the diversity appears somewhat higher during the “silly season” in the summer months.

4 Relation of diversity H_t to financial market movements

As in the main text, the entropy, which we refer to as the diversity, is computed as

$$H_t \equiv - \sum_{k=1}^{K} \rho_{t,k} \log(\rho_{t,k}) \tag{2}$$

where $\rho_{t,k}$ is entry k of the vector ρ_t, and represents the relative weight of topic k in the *Financial Times* on day t. We used the natural logarithm in this analysis, although alternative choices, such as the logarithm base 2, will simply scale measurements of H_t.



A. Topic 24 (“mubarak”, “egypt”, “protest”,...) B. Topic 33 (“korea”, “seoul”, “kim”,...)

C. Distribution of lifetimes

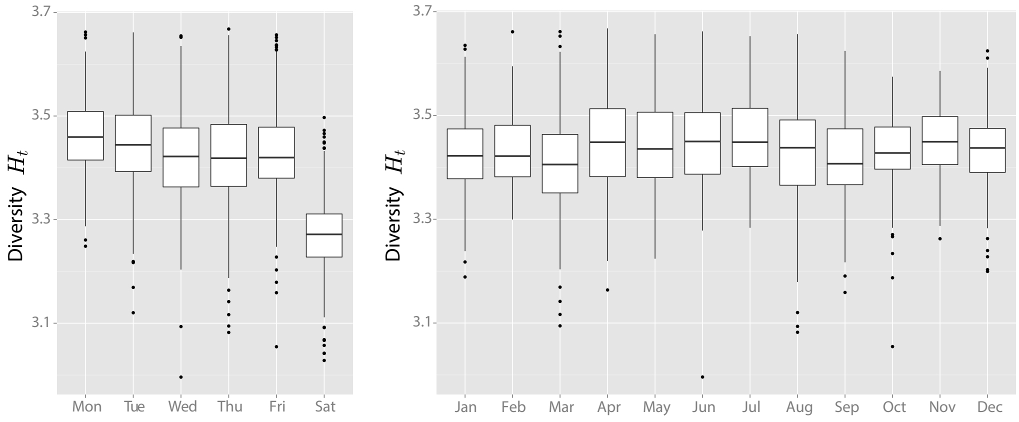
Figure S1: **Variation in topic lifetimes in the *Financial Times*.** (A) ACF for the topic time series relating to events in Egypt, $\rho^T_{24,t}$. (B) The same for a topic relating to events in Korea, $\rho^T_{33,t}$. (C) The distribution of lifetimes for all 50 topics, defined as the first lag at which the corresponding ACF falls at or below the 95% confidence band for an uncorrelated signal.

4.1 Diversity relates to same-day trading volume

We quantify daily trade volume by differencing the total daily trade volume in the FTSE 100 after a log transformation:

$$v_t \equiv \log(V_t) - \log(V_{t-1}) \tag{3}$$

where V_t represents the total trade volume on day t. One order of differencing, as above, is sufficient to render the series $\log(V_t)$ stationary. Specifically, according to a KPSS test (Hyndman & Khandakar 2008) for $\log(V_t)$, testing a null hypothesis of a stationary root against a unit-root alternative, we reject the null hypothesis: $\text{KPSS} = 10.6, N = 1516, p < 0.05$. We accept the



A. Weekly seasonalities

B. Monthly seasonalities

Figure S2: **Boxplots of the diversity H_t, aggregated by weekday and by month.** (A) Weekly seasonalities in the diversity. Weekend issues of the *Financial Times* exhibit characteristically low values of H_t, as a large portion of these issues are devoted to a small number of topics that appear infrequently in weekday issues, such as the topic containing the words "book", "music", and "film". (B) Yearly seasonalities in the diversity. In this figure we exclude weekend issues from our measurements of the diversity H_t. We see visually that there is little seasonal variation in H_t, although the diversity appears somewhat higher during the "silly season", or "slow news season" in the summer months.

null hypothesis for v_t: $\text{KPSS} = 0.04, N = 1515, p > 0.05$.

To isolate the predictive power of the differenced diversity ΔH_t with respect to changes in daily trade volume v_t, we first examine the extent to which v_t may be modeled using only its past values $\{v_{t-1}, v_{t-2}, ...\}$. A scan of ARMA models reveals the presence of both significant autoregressive and moving average terms. For this purpose we model v_t as an ARMA(1,1) process. To account for finite-size effects, we also add the logarithm of the number of paragraphs in each issue as an external regressor:

$$v_t = v_0 + \alpha_1 v_{t-1} + \beta_1 \epsilon_{t-1} + \theta_1 \log(N) + \epsilon_t \qquad (4)$$

Using maximum-likelihood estimation (Hyndman & Khandakar 2008), we find $v_0 = 0.12 \pm 0.09$, $\alpha_1 = 0.33 \pm 0.05$, $\beta_1 = -0.82 \pm 0.03$, and $\theta_1 = -0.02 \pm 0.01$.

We find that a significant portion of the variance of the residuals in model (4) can be explained using changes in the diversity ΔH_t. We find that in the model

$$\epsilon_t = \alpha_0 + \alpha_1 \Delta H_t + \eta_t, \tag{5}$$

the coefficient $\alpha_1 = -0.30 \pm 0.07$ is significant according to a standard t-test ($t = -4.0$, $N = 1459$, $p < 0.0001$). This motivates us to include the change in diversity ΔH_t, measured in the *Financial Times* on the morning of day t, in our model of the volume signal v_t for the same trading day. We therefore fit

$$v_t = v_0 + \gamma_1 v_{t-1} + \gamma_2 \epsilon_{t-1} + \gamma_3 \Delta H_t + \gamma_4 \log(N) + \epsilon_t \tag{6}$$

finding $v_0 = -0.04 \pm 0.07$, $\gamma_1 = 0.29 \pm 0.04$, $\gamma_2 = -0.82 \pm 0.03$, $\gamma_3 = -0.41 \pm 0.09$, and $\gamma_4 = 0.006 \pm 0.01$. The coefficient of ΔH_t is again significant according to a standard t-test ($t = -4.6$, $N = 1459$, $p < 0.0001$). The negative coefficients α_1 and γ_3 in models (5) and (6) indicate that falls in the diversity H_t tend to precede increased transaction volumes in the FTSE 100, and that increases in diversity tend to precede trading days in which transaction volumes are relatively diminished.

We supplement our in-sample tests through a comparison of errors from out-of-sample one-step forecasts between the models (4) and (6). We fit both models using only the first 70% of the dataset, and evaluate one-step forecasts on the remaining 30% of the dataset. Using the Diebold-Mariano test for predictive accuracy (Diebold & Mariano 1995, Hyndman & Khandakar 2008) with a quadratic loss function, we reject the hypothesis that inclusion of the diversity signal ΔH_t in model (6) fails to provide an increased out-of-sample accuracy ($\mathrm{DM} = 2.2$, $N = 431$, $p = 0.015$).

We provide an additional check on the robustness of our results by bootstrapping on empirical residuals. In particular, we compute the N differences in squared residuals in the out-of-sample forecasts, and re-sample with replacement N of these points. We then calculate the mean M of the differences, where a positive mean indicates that the model including the diversity fluctuations results in smaller squared residuals, on average, than the model that does not include the diversity. We repeat this procedure 10,000 times, finding that the model incorporating ΔH_t consistently outperforms the model without it, as M is positive in over 99% of re-samplings, with a 95% confidence interval of $[4.18 \times 10^{-4}, 5.39 \times 10^{-3}]$. By re-centering the distribution of M to zero, we can test the null hypothesis that including the volume signal does not change the out-of-sample accuracy. We record the mean difference in squared residuals from the test set, and compare the absolute value of this mean to this re-centered distribution. We find that the absolute deviation of the re-centered distribution is greater than the absolute value of this mean only 2.98% of the time, and so we reject the null hypothesis ($p = 0.0298$).

A cursory analysis reveals no evidence for a reciprocal relationship in which the volume signal v_t anticipates changes in the diversity H_t, as indicated by correlations between v_{t-1} and next-day changes in diversity ΔH_t (Pearson $r = -0.03, N = 1450, p > 0.05$). For a more thorough investigation, we include the volume signal v_{t-1} in our MA(1) model of ΔH_t. Here, we find that although previous-day changes in trade volume are significant when modeling the diversity ΔH_t in-sample ($\gamma_2 = -0.025 \pm 0.007$, $t = -3.5$, $p < 0.001$), they fail to offer any advantage in out-of-sample predictions upon repetition of the Diebold-Mariano test ($\mathrm{DM} = 0.87$, $N = 428$, $p > 0.1$).

4.2 Price changes of the FTSE drive changes in diversity

To isolate the influence of the FTSE 100 returns r_t on the diversity H_t, we first determine the extent to which H_t may be modeled using only its past values $\{H_{t-1}, H_{t-2}, ...\}$. There exist general methods to model a time series using only its past values – autoregressive (AR) terms – as well as the model's own residuals – moving average (MA) terms. A popular, classical approach to modeling stationary time series in this way is to train an ARMA model (Chan &

Cryer 2010).

To this end we model H_t as an ARMA process, finding that one order of differencing is sufficient to achieve stationarity in H_t. We therefore model the differenced diversity $\Delta H_t \equiv H_t - H_{t-1}$. Specifically, according to a KPSS test (Hyndman & Khandakar 2008) for H_t, testing the null hypothesis of a stationary root against a unit-root alternative, we reject the null hypothesis: $\mathrm{KPSS} = 6.8, N = 1520, p < 0.05$. We accept the null hypothesis for ΔH_t: $\mathrm{KPSS} = 0.07, N = 1479, p > 0.05$. To determine how many elements of the lagged time series we must include in our model, we scan over several ARMA(p, q) models (p = 1,...,5; q = 1,...,5) and find that the Akaike information criterion (AIC) is minimized with a simple MA(1) process. This is corroborated by the autocorrelation function of ΔH_t (Chan & Cryer 2010), which exhibits an isolated negative spike at lag 1 and is otherwise featureless. We therefore fit

$$\Delta H_t = \beta_0 + \epsilon_t + \beta_1 \epsilon_{t-1}, \tag{7}$$

finding $\beta_0 = -0.0001 \pm 0.0003$ and $\beta_1 = -0.88 \pm 0.02$ using maximum-likelihood estimation (Hyndman & Khandakar 2008). We find no significant dependence of ΔH_t on the day of the week, as would be indicated by the presence of significant five-day seasonality. A plot of the signal ΔH_t, as well as its autocorrelation function (ACF) and partial autocorrelation function (PACF; Chan & Cryer 2010), is provided in Fig. S3.

A simple least-squares linear regression of the residuals of model (7) against the returns of the FTSE 100 on the previous day suggests that these residuals are at least in part related to financial market movements. We find that in the model

$$\epsilon_t = \alpha_0 + \alpha_1 r_{t-1} + \eta_t,$$

with η_t an error term, the coefficient $\alpha_1 = 0.5 \pm 0.1$ is significant according to a standard t-test (t = 3.8, N = 1450, $p <$ 0.001). This motivates us to include the previous-day returns of the

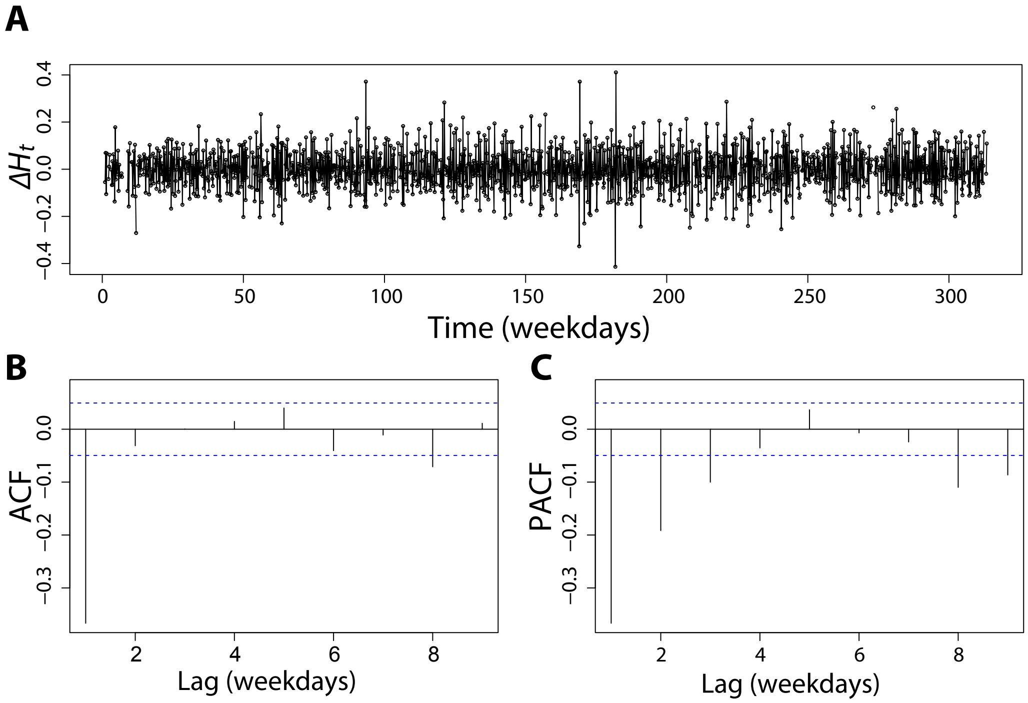


Figure S3: **Time-series features of the differenced diversity ΔH_t.** (A) Plot of the differenced diversity ΔH_t. (B) The ACF of ΔH_t. (C) The PACF of ΔH_t. The time series exhibits characteristics of a MA(1) process. The *forecast* package for R was used in creating this plot (Hyndman & Khandakar 2008).

FTSE 100 to our model of diversity fluctuations. We therefore fit

$$\Delta H_t = \gamma_0 + \gamma_1 \epsilon_{t-1} + \gamma_2 r_{t-1} + \epsilon_t \tag{8}$$

finding $\gamma_0 = 0.0000 \pm 0.0003$, $\gamma_1 = -0.87 \pm 0.02$ and $\gamma_2 = 0.30 \pm 0.07$. The coefficient γ_2 of the previous day's returns r_{t-1} is again significant according to a standard t-test ($t = 4.3$, $N = 1450, p < 0.0001$). The positive coefficients in models (4.2) and (8) indicate that decreases in diversity H_t follow stock market falls, while increases in diversity follow stock market rises.

Ultimately, the utility of the returns r_{t-1} in predicting changes in the diversity ΔH_t can be decided in a comparison of errors from out-of-sample one-step forecasts between the models

(7) and (8). For this purpose, we fit both models using only the first 70% of the dataset – from January 4, 2007 to March 16, 2011. We then compare one-step forecasts on the remainder of the data, from March 17, 2011 to December 31, 2012. A scan of ARMA models again finds that the MA(1) model best fits the training data, according to the AIC statistic.

We compare errors from the out-of-sample forecasts using the Diebold-Mariano test for predictive accuracy (Diebold & Mariano 1995, Hyndman & Khandakar 2008) with a quadratic loss function. To interpret the results of this test we need not assume that the forecast errors are Gaussian, of zero-mean, or serially or contemporaneously uncorrelated (Diebold & Mariano 1995). We find marginal support for the hypothesis that including the previous-day returns of the FTSE 100, as in model (8), results in an increased out-of-sample accuracy ($\mathrm{DM} = 1.4$, $N = 428, p = 0.078$).

We again check the robustness of our results by bootstrapping on empirical residuals. We compute the N differences in squared residuals in the out-of-sample forecasts, and re-sample with replacement N of these points. We then calculate the mean M of the differences, where a positive mean indicates that the model including the returns r_{t-1} results in smaller squared residuals, on average, than the model that does not include them. We repeat this procedure 10,000 times, finding that the model incorporating r_{t-1} consistently outperforms the model without it, as M is positive in over 93% of re-samplings, with a 95% confidence interval of $[-4.96 \times 10^{-5}, 4.07 \times 10^{-4}]$. By re-centering the distribution of M to zero, we can test the null hypothesis that including the volume signal does not change the out-of-sample accuracy. We record the mean difference in squared residuals from the test set, and compare the absolute value of this mean to this re-centered distribution. We find that the absolute deviation of the re-centered distribution is greater than the absolute value of this mean 14.98% of the time, and so we fail to reject the null hypothesis ($p = 0.1498$).

4.3 Influence of individual topics

To search for individual topics that may dominate the observed relationship between ΔH_t and r_{t-1}, we isolate the 50 columns ρ_k^T of ρ corresponding to individual topics. For each topic, we compute the Pearson correlation between the differences $\Delta\rho_{k,t}^T \equiv \rho_{k,t}^T - \rho_{k,t-1}^T$ and the previous-day returns of the FTSE 100. We associate a p-value to each correlation in the usual way, using the Fisher transformation (Fisher 1915). As explained in the main text, only one topic relating to the recent financial crisis of 2008 ("mortgage", "loan", "credit", "debt",...) was found to be significantly impacted by previous-day returns of the FTSE 100 ($p < 0.05$ after FDR correction for multiple comparisons, Benjamini & Hochberg 1995). We find that the sign of this relationship is negative, implying a greater interest in this topic following falls in the FTSE 100, and vice-versa.

We check the influence of this topic on our previous results by removing it from the analysis. That is, we remove the entry corresponding to this topic from each topic vector θ_d, re-compute ρ_t and H_t, and repeat the comparison with the returns r_t of the FTSE 100. Exclusion of this topic leaves the results qualitatively unchanged, as is evident in Table S2. We again find that the differenced diversity ΔH_t is best modeled as an MA(1) process, according to the AIC statistic. Moreover, upon repetition of the Diebold-Mariano test on the errors of one-step out-of-sample forecasts, we find that inclusion of the previous-day returns of the FTSE 100 results in significantly greater accuracy in predicting ΔH_t (DM $= 1.8$, $N = 428$, $p = 0.03$).

Table S1: LDA Topics

Topic	Top 10 words
1	“ge”, “dhabi”, “abu”, “nt”, “goldman”, “sach”, “en”, “capit”, “verizon”, “codelco”
2	“libor”, “cd”, “share”, “profit”, “month”, “sale”, “compani”, “year”, “revenu”, “bn”
3	“rio”, “stock”, “group”, “xstrata”, “gilt”, “list”, “yield”, “price”, “mine”, “bhp”
4	“market”, “rate”, “bank”, “price”, “dollar”, “economi”, “inflat”, “growth”, “year”, “bond”
5	“iceland”, “group”, “suez”, “french”, “compani”, “carrefour”, “sale”, “brazil”, “share”, “bn”
6	“busi”, “compani”, “googl”, “work”, “peopl”, “facebook”, “school”, “skill”, “job”, “social”
7	“investig”, “case”, “fraud”, “compani”, “alleg”, “court”, “bank”, “ivco”, “vw”, “porsch”
8	“cf”, “airlin”, “aircraft”, “trail”, “carrier”, “airbu”, “jet”, “passeng”, “boe”, “air”
9	“car”, “carmak”, “gm”, “compani”, “sale”, “vehicl”, “year”, “bn”, “market”, “plant”
10	“fund”, “manag”, “hedg”, “equiti”, “invest”, “asset”, “investor”, “global”, “incom”, “market”
11	“properti”, “etf”, “fund”, “market”, “investor”, “bank”, “invest”, “uk”, “year”, “compani”
12	“appl”, “phone”, “shown”, “mobil”, “limit”, “hlc”, “yr”, “trade”, “free”, “content”
13	“parti”, “labour”, “minist”, “elect”, “tori”, “brown”, “govern”, “cameron”, “polit”, “prime”
14	“art”, “design”, “work”, “artist”, “galleri”, “london”, “museum”, “build”, “citi”, “hous”
15	“bbc”, “film”, “weather”, “itv”, “show”, “seri”, “live”, “hollyoak”, “channel”, “region”
16	“murdoch”, “broadband”, “farmer”, “food”, “agricultur”, “bt”, “crop”, “bskyb”, “compani”, “corp”
17	“pe”, “chile”, “denmark”, “rep”, “hungari”, “colombia”, “group”, “indonesia”, “malaysia”, “argentina”
18	“cadburi”, “kraft”, “drug”, “dubai”, “lm”, “compani”, “ship”, “ord”, “shipp”, “gsk”
19	“carbon”, “emiss”, “ser”, “energi”, “climat”, “prog”, “fund”, “rbsg”, “environment”, “invest”
20	“aig”, “islam”, “pru”, “bn”, “compani”, “aia”, “bank”, “insur”, “busi”, “execut”
21	“nh”, “health”, “patient”, “hospit”, “care”, “healthcar”, “servic”, “privat”, “drug”, “compani”
22	“china”, “school”, “chines”, “busi”, “peopl”, “music”, “year”, “beij”, “work”, “univers”
23	“stock”, “call”, “request”, “fund”, “mail”, “minut”, “charg”, “price”, “thaksin”, “servic”
24	“mubarak”, “egypt”, “elect”, “egyptian”, “brotherhood”, “presid”, “protest”, “ahmadi”, “nejad”, “polit”
25	“equip”, “servic”, “leisur”, “ga”, “industri”, “telecommun”, “good”, “oil”, “materi”, “food”
26	“abn”, “emi”, “amro”, “terra”, “firma”, “bank”, “bn”, “forti”, “group”, “compani”
27	“und”, “fd”, “ssga”, “bd”, “om”, “ho”, “bs”, “class”, “govt”, “editor”
28	“index”, “fell”, “stock”, “cl”, “bank”, “rose”, “share”, “market”, “gain”, “data”
29	“properti”, “fd”, “brand”, “hotel”, “luxuri”, “hous”, “watch”, “yacht”, “sundai”, “residenti”
30	“peso”, “fund”, “equiti”, “dinar”, “privat”, “invest”, “bank”, “egypt”, “bn”, “compani”
31	“oil”, “iran”, “ga”, “bp”, “iraq”, “nuclear”, “militari”, “countri”, “govern”, “energi”
32	“price”, “dec”, “yield”, “south”, “turkei”, “pe”, “nav”, “sep”, “poland”, “venezuela”
33	“korea”, “korean”, “clear”, “lg”, “otc”, “deriv”, “south”, “trade”, “seoul”, “kim”
34	“pension”, “tax”, “scheme”, “annuiti”, “incom”, “retir”, “list”, “pai”, “benefit”, “rate”
35	“coal”, “ivco”, “aim”, “compani”, “share”, “mine”, “group”, “price”, “enrc”, “china”
36	“eu”, “european”, “eurozon”, “govern”, “bank”, “countri”, “greec”, “union”, “minist”, “debt”
37	“wine”, “russia”, “china”, “russian”, “putin”, “kairo”, “chines”, “moscow”, “restaur”, “georgia”
38	“melchior”, “opp”, “tesco”, “calculat”, “share”, “class”, “date”, “uk”, “shower”, “store”
39	“rate”, “convent”, “ng”, “market”, “appli”, “bond”, “currenc”, “il”, “meril”, “par”
40	“sun”, “fair”, “cloudi”, “shower”, “rain”, “xr”, “priceslast”, “shown”, “thunder”, “microsoft”
41	“palestinian”, “israel”, “isra”, “gaza”, “hama”, “flu”, “netanyahu”, “peac”, “dress”, “minist”
42	“compani”, “govern”, “account”, “school”, “busi”, “manag”, “rail”, “audit”, “fund”, “regul”
43	“jpm”, “siemen”, “vodafon”, “compani”, “sale”, “bn”, “group”, “year”, “deut”, “eq”
44	“gam”, “polic”, “pakistan”, “sky”, “kill”, “attack”, “sport”, “bbb”, “war”, “footbal”
45	“bank”, “fund”, “fin”, “market”, “manag”, “invest”, “investor”, “bn”, “int”, “compani”
46	“bank”, “mortgag”, “loan”, “bn”, “credit”, “capit”, “fund”, “market”, “debt”, “asset”
47	“ftse”, “cap”, “republican”, “obama”, “msci”, “global”, “romnei”, “democrat”, “dj”, “world”
48	“work”, “plai”, “book”, “music”, “life”, “peopl”, “love”, “live”, “film”, “make”
49	“cp”, “sempra”, “roch”, “prologi”, “rockwel”, “safewai”, “sherwil”, “rockwlcol”, “questdg”, “repsrv”
50	“quot”, “euriborlibor”, “libor”, “basi”, “annual”, “month”, “rate”, “icap”, “euroswiss”, “semi”

Table S2: In-sample model results with and without Topic 46 (“mortgage”, “loan”, “credit”, “debt”,...)

Model	All topics	Topic 46 removed
$\Delta H_t = \beta_0 + \beta_1 \epsilon_{t-1} + \epsilon_t$	$\beta_1 = -0.88 \pm 0.02^{***}$	$\beta_1 = -0.91 \pm 0.02^{***}$
$\epsilon_t = \alpha_0 + \alpha_1 r_{t-1} + \eta_t$	$\alpha_1 = 0.5 \pm 0.1^{***}$	$\alpha_1 = 0.3 \pm 0.1^{*}$
$\Delta H_t =$ $\gamma_0 + \gamma_1 \epsilon_{t-1} + \gamma_2 r_{t-1} + \epsilon_t$	$\gamma_2 = 0.30 \pm 0.07^{***}$	$\gamma_2 = 0.20 \pm 0.07^{**}$

Note: Signif. codes: *** 0.001 ** 0.01 * 0.05

References

Benjamini, Y. & Hochberg, Y. (1995), ‘Controlling the false discovery rate: a practical and powerful approach to multiple testing’, *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.

Chan, K. S. & Cryer, J. D. (2010), *Time Series Analysis with Applications in R*, 2nd edn, Springer, New York.

Diebold, F. X. & Mariano, R. S. (1995), ‘Comparing predictive accuracy’, *Journal of Business and Economic Statistics* **13**, 253–263.

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

Hazen, T. J. (2010), Direct and latent modeling techniques for computing spoken document similarity, *in* ‘Spoken Language Technology Workshop (SLT), 2010 IEEE’, IEEE, pp. 366–371.

Hyndman, R. J. & Khandakar, Y. (2008), ‘Automatic time series forecasting: the forecast package for R’, *Journal of Statistical Software* **27**, 1–22.

Porter, M. F. (1980), ‘An algorithm for suffix stripping’, *Program* **14**(3), 130–137.

Řehůřek, R. & Sojka, P. (2010), Software Framework for Topic Modelling with Large Corpora, *in* ‘Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks’, ELRA, Valletta, Malta, pp. 45–50.

Salton, G. & McGill, M. J. (1986), *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA.

Wilson, A. T. & Chew, P. A. (2010), Term weighting schemes for latent Dirichlet allocation, *in* 'Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, pp. 465–473.

Xu, R., Ye, L. & Xu, J. (2013), 'Reader's emotion prediction based on weighted latent Dirichlet allocation and multi-label k-nearest neighbor model', *Journal of Computational Information Systems* **9**(6), 2209–2216.