Temporal and Cross Correlations in Business News

Takayuki MIZUNO,[1,2] Kazumasa TAKEI,[3] Takaaki OHNISHI[2,4]
and Tsutomu WATANABE[2,4]

[1] *Division of Information Engineering, Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba 305-8573, Japan*
[2] *Canon Institute for Global Studies, Tokyo 100-6511, Japan*
[3] *The Norinchukin Bank, Tokyo 100-8420, Japan*
[4] *Graduate School of Economics, The University of Tokyo, Tokyo 113-0033, Japan*

We empirically investigate temporal and cross correlations in the frequency of news reports on companies, using a dataset of more than 100 million news articles reported in English by around 500 press agencies worldwide for the period 2003–2009. Our first finding is that the frequency of news reports on a company does not follow a Poisson process, but instead exhibits long memory with a positive autocorrelation for longer than one year. The second finding is that there exist significant correlations in the frequency of news across companies. Specifically, on a daily time scale or longer the frequency of news is governed by external dynamics, while on a time scale of minutes it is governed by internal dynamics. These two findings indicate that the frequency of news reports on companies has statistical properties similar to trading volume or price volatility in stock markets, suggesting that the flow of information through company news plays an important role in price dynamics in stock markets.

§1. Introduction

News is the communication of information about important events. In macroeconomics, quantitative finance, and econophysics, the impact of news on prices and trading volumes in stock markets has previously been studied.[1),2)] Some financial economists have shown that there is only a weak relationship between the number of news reports each day, the trading volume, and the price return in stock markets.[3)] On the other hand, in the area of econophysics, it has been found by using tick-by-tick data that market volatility and volume increases immediately after particular news has been reported.[4)–6)] The influence of exogenous shocks, including news reports, on pricing in financial markets has been examined using numerical models.[7)] Another strand of research has attempted to detect patterns in the flow of information. For instance, it has been suggested that the frequency of use of specific words in blogs on the internet does not follow a Poisson process,[8),9)] while Ref. 10) shows that using latent Dirichlet allocation, news articles appearing in the New York Times can be classified into several topics.

The aim of this paper is to empirically identify certain statistical properties of the frequency of news, with a special focus on the temporal correlation of news frequency for a specific company as well as the cross correlation of news across companies. For this purpose, we use a dataset of news articles reported by around 500 press agencies worldwide. The dataset – “Reuters NewsScope Archive”– was obtained from Thomson Reuters Corporation. The rest of the paper is organized

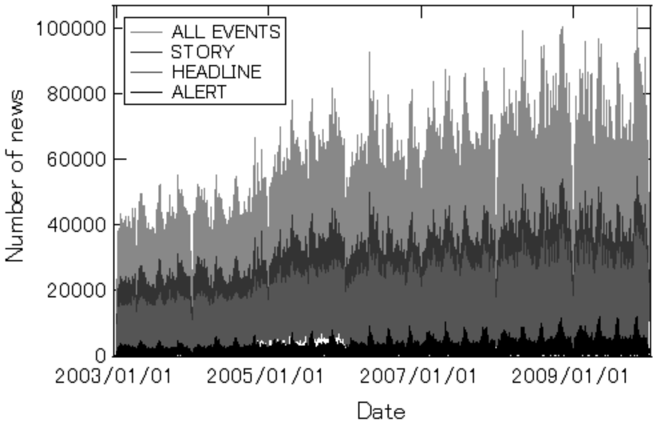

Fig. 1. Time series of the daily number of news reports. The number of news reports in English is counted. The top line is for all news; the second line is for story news; the third line is for headline news; and the bottom line is for alert news.

as follows. Section 2 describes our dataset and shows that there are periodicities in the frequency of news. Section 3 analyzes the autocorrelations of the frequency of news on a particular company and shows that the autocorrelation function follows a power law. Section 4 examines the cross correlations for the frequency of news across companies. We show that the coupling of the average number of news items on a company with its fluctuations obeys a scaling law, and that the frequency of news on a company is not governed solely by internal dynamics (i.e., a Poisson process) but is also affected by external dynamics, such as an increase in the number of news items due to the outbreak of an economic crisis. In §5, we extract common movements across companies using random matrix theory techniques. Section 6 concludes the paper.

§2. Overview of the news data

Thomson Reuters Corporation is a world-famous provider of information for businesses and professionals, providing, among other things, “Reuters 3000 Xtra”, an electronic trading platform typically used by professional traders and investment analysts in trading rooms. “Reuters 3000 Xtra” offers real-time streaming news, comprehensive economic indicators, and financial data, and displays news from not only Thomson Reuters but also around 500 third parties. From 2003 to 2009, approximately 165 million news reports were provided. While these reports were in several languages, about 65 percent of them (107 million) were in English. In this study, we use only the English news reports, all of which are available in the Reuters NewsScope Archive database.

There are three news types in the database. The first type is “alert” news, which covers an urgent newsworthy event and is 80–100 characters long. Alert news is normally followed by another news type. The second type is “headline” news, consisting of the headline of a news report for an event. The third type, finally, is

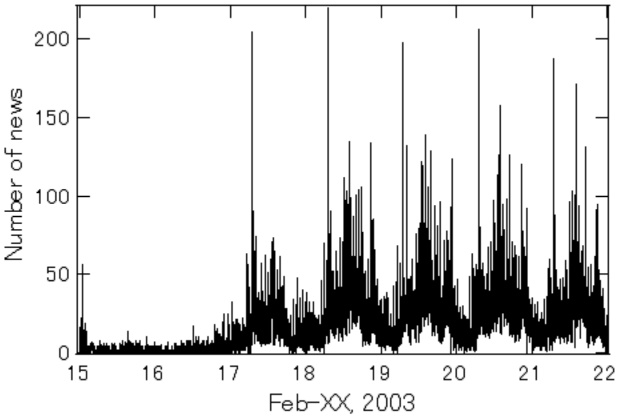

Fig. 2. Number of news reports per minute for a particular week, February 15–22, 2003. The number of news reports in English is counted.

Table I. Mean of the number of news reports on each day of the week. The number of news reports in English is counted.

	Sat.	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.
ALERT	101	148	3010	3807	3960	4366	2666
HEADLINE	2398	3967	20565	23115	23111	23453	20553
STORY	2614	4316	27811	31607	31633	32428	27804

Table II. Example of news.

Date	Time	News type	Text
2010-04-05	00:03:14.307	STORY	...It topped Credit Suisse Group, which jumped from ninth place a year ago to second,...

"story" news, which contains the text that provides further information about the event. If the event is important, story news is often updated.

Figure 1 shows the time series of the number of news reports for each news type. We find that the number of news reports delivered by Reuters increases every year. There were 9.8 million news reports in 2003, but 18.6 million in 2009. Figure 2 shows the time series of the number of news reports per minute for the week starting February 15, 2003. We clearly see that the frequency of news has intraday seasonality, as has been observed previously.[11),12)] Table I presents the mean number of news reports for each news type on each day of the week. There are fewer news reports on the weekend, indicating that before proceeding to a detailed analysis we need to deal with the nonstationarity of the time series. This is discussed in the next section.

To investigate the frequency of company news, we first need to construct time series for the number of news reports for each company. We do so using the following steps. First, we focus on the top 100 companies in the world in terms of market capitalization in 2003 and search the database by company name. For example, we find that the Credit Suisse Group is mentioned in the text of a news report published

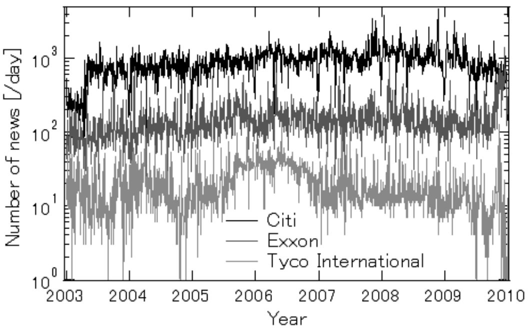

Fig. 3. Time series for the number of company news reports per day excluding the weekend. The top, middle, and bottom lines are for news on Citi, Exxon, and Tyco International, respectively.

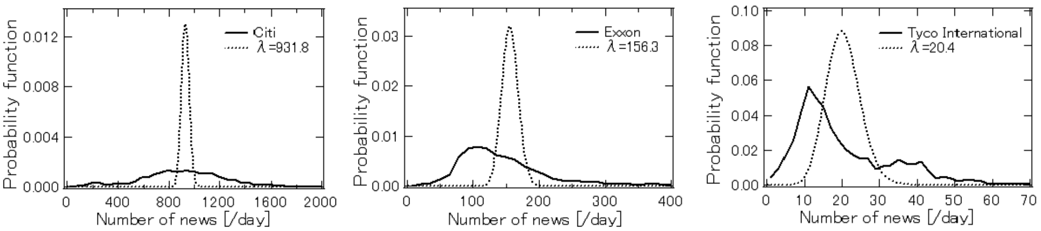

Fig. 4. Probability functions for the frequency of news on Citi, Exxon, and Tyco International. The dashed lines show Poisson distributions with the same mean as in the data. News reports on the weekend are excluded.

at 00:03:14 on April 5 (see Table II). Next, we define company news as news that mentions the name of the company. Finally, by counting the number of news reports for each company, we obtain the relevant time series.

§3. Autocorrelations for the frequency of news reports

We investigate the probability functions and autocorrelations of the frequency of company news for each news type. We focus on the time series for the number of news reports for three companies, Citi, Exxon, and Tyco International, which are shown in Fig. 3. The daily mean number of reports on Citi, Exxon, and Tyco International, excluding the weekend, is 932, 156, and 20, respectively. Figure 4 shows the probability function of the daily number of news reports for each company. Compared to the Poisson distribution that has the same mean as the data, each probability function has a fatter tail, suggesting that the time series for company news do not follow a Poisson process.

As mentioned in the previous section, the news frequency time series are not stationary due to the time trend and daily periodicity. It may be that the fat tails of the probability functions observed in Fig. 4 come from the nonstationarity of the time series. To transform our data into stationary time series, we introduce the concept of "tick time" for news. Tick time refers not to actual time, but is measured in terms of the appearance of news reports, where each news report corresponds to

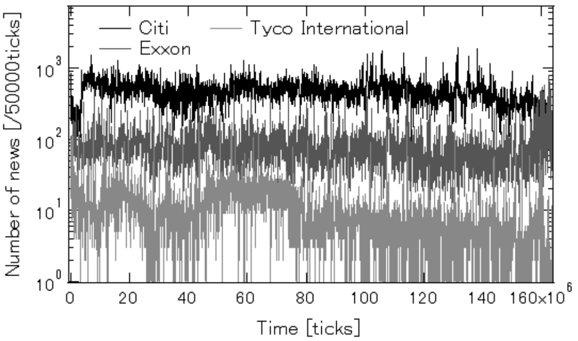

Fig. 5. Time series of the number of company news reports per 50,000 ticks. The top, middle, bottom lines are for Citi, Exxon, and Tyco International, respectively.

Table III. Augmented Dickey-Fuller test for the Citi, Exxon, and Tyco International time series.

		Trend	Drift	None
Citi	t-values	−19.7	−18.6	−1.08
	p-values	0.00	0.00	0.28
Exxon	t-values	−27.5	−26.8	−2.5
	p-values	0.00	0.00	0.01
Tyco International	t-values	−18.4	−18.0	−9.0
	p-values	0.00	0.00	0.00

a unit of "time". That is, the tick time increases by one whenever a fresh news item in any language is reported. Note that because news reports are less frequent, and the interval between "ticks" in actual time therefore longer, on weekends, tick time passes more slowly on Saturdays and Sundays than during the weekdays, when the number of news reports is larger. We set the tick time to zero at the beginning of our sample period (January 1, 2003). Thus, using tick time allows us to eliminate the periodicity and trends observed in the original data. Figure 5 shows the time series of the number of news reports measured by tick time. In this figure, we count the number of news reports per 50,000 ticks, corresponding to about half a day, for each of the three companies. Comparing this with Fig. 3, we see that the upward trend and daily periodicity have been eliminated.

To check the stationarity of the time series with tick time, we use the Augmented Dickey-Fuller (ADF) test, which is a test for a unit root of a time series.[13),14)] We choose the lag order of the ADF test using Akaike's Information Criterion and conduct three types of ADF test ("none", "drift", and "trend") for the time series for Citi, Exxon, and Tyco International. If the type is set to "none", neither an intercept nor a trend is included in the test regression; if it is set to "drift", an intercept is added; and if it is set to "trend", both an intercept and a trend are added. Table III presents the results of the ADF test for each type, showing that the null hypothesis that the time series measured by tick time are not stationary is rejected for eight of the nine cases. In the rest of the paper, we will use tick time unless otherwise indicated.

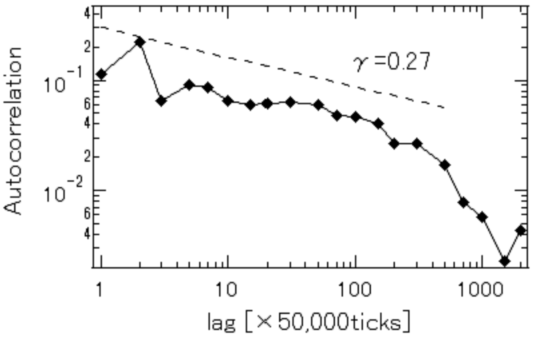

Fig. 6. Autocorrelation function of news reports. The frequency of news reports is defined as the number of news reports per 50,000 ticks, which is about half a day. The dashed reference line represents $\gamma = 0.27$.

We now turn to estimating the autocorrelation for the news frequency $f_{i,t}$ of company i using an autocorrelation function of the form

$$\rho_i(\tau) = \frac{\langle f_{i,t} f_{i,t+\tau}\rangle - \langle f_{i,t}\rangle \langle f_{i,t+\tau}\rangle}{\sigma(f_{i,t})\sigma(f_{i,t+\tau})}, \tag{3·1}$$

where τ is a time lag, $\langle\cdot\rangle$ denotes the time average over the sample period, and σ is the standard deviation. We continue to measure the frequency by the number of news reports per 50,000 ticks, and pool observations for the top 100 companies. Figure 6 presents the estimated autocorrelation function, showing that it follows a power law of the form

$$\rho(\tau) \propto \tau^{-\gamma}, \tag{3·2}$$

where $\rho(\cdot)$ is the average of $\rho_i(\cdot)$ over the 100 companies. Note that the exponent γ is about 0.27, as represented by the reference line in the figure, and that the estimated autocorrelation decays along the reference line up to $\tau = 600$, which is equivalent to approximately one year. This indicates the presence of long memory in the frequency of news reports. Similar long memory properties have also been observed for price volatility and trading volumes in stock markets (e.g., Ref. 15)).

§4. Scaling laws for the frequency of news

In this section, we investigate correlations in the frequency of news across different companies. A useful method for examining such cross correlations in the context of complex networks, such as the internet, is to look at the average flux and fluctuations at individual nodes.[16)–18)] It has been found that the coupling of the flux fluctuations with the total flux on individual nodes obeys a unique scaling law for a wide variety of complex networks, including the internet (i.e., a network of routers linked by physical connections), highways, river networks, and the World Wide Web of web pages and links.[17)] Specifically, it has been shown that the average flux $\langle f\rangle$

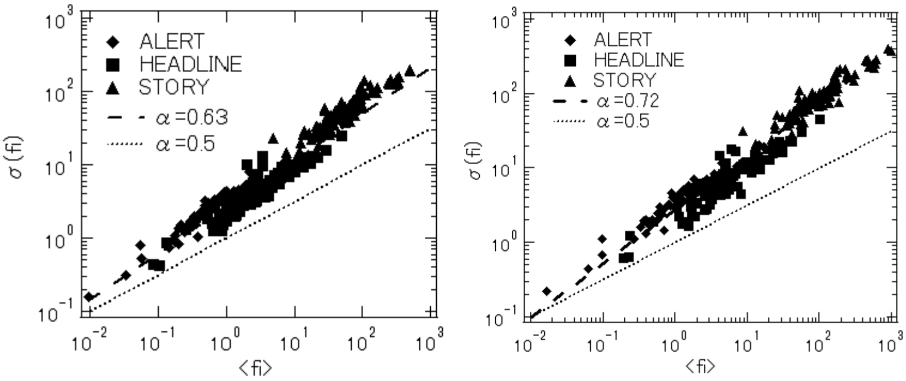

Fig. 7. The relationship between the mean and the standard deviation of the frequency of news reports for the top 100 companies in the world in terms of market capitalization in 2003. The frequency of news is defined as the number of news reports per 50,000 ticks in the left panel, while it is defined as the number of news reports per day in the right panel.

and the standard deviation σ of those individual nodes are related by[17)]

$$\sigma \sim \langle f \rangle^{\alpha}, \tag{4·1}$$

where α is a scaling exponent. The scaling exponent is equal to 1/2 if the flux on individual nodes follows a Poisson process or is governed mainly by internal dynamics. On the other hand, the scaling exponent is not 1/2 if the flux does not follow a Poisson process, and is equal to 1 if the flux on individual nodes is governed completely by external dynamics. For example, for river networks, the exponent α has been found to be quite close to unity, because the stream of rivers in different locations is mainly driven by weather patterns.

We apply this method to the frequency of news on individual companies by calculating the mean and standard deviation of the frequency of news for each company. Figure 7 plots $\sigma(f_i)$ for each of the top 100 companies as a function of the average $\langle f_i \rangle$ of the company. The frequency of news is defined as the number of news reports per 50,000 ticks in the left panel and as the number of news reports per day in the right panel. We see that in both cases the dots are not on the dotted line denoted by $\alpha = 1/2$. The estimate for α is 0.63 in the case of tick time (left panel) and even higher in the case of actual time (right panel). These results suggest that the frequency of news is governed, at least partially, by external dynamics, such as the outbreak of an economic crisis that results in a simultaneous increase in the number of news reports for each company. Note that the higher estimate of α in the right panel can be interpreted as reflecting a closer co-movement across companies due to intraday seasonality.

To see whether the scaling exponent α depends on the time scale, we estimate α for different time scales. Specifically, we count the number of news reports per s ticks, with s ranging between 5 and 100,000. Figure 8 shows that α is close to 1/2 for small values of s, indicating that the frequency of news is governed by internal dynamics on shorter time scales such as minutes. However, α increases monotonically with the time scale s and exceeds 0.6 for sufficiently large values of s, indicating that

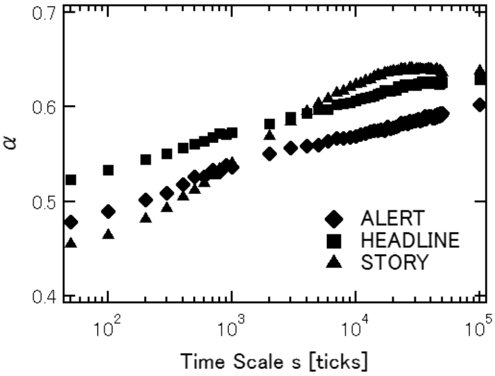

Fig. 8. Scaling exponent α for different time scales. The value of α is estimated using the observations for the top 100 companies in the world in terms of market capitalization in 2003.

the frequency of news is governed, at least partially, by external dynamics on a daily or longer time scale. Interestingly, a similar statistical property was found for transaction values on the New York Stock Exchange, namely that α is close to 1/2 on a time scale of minutes, while it is higher and close to unity on a daily or longer time scale.[16)]

§5. Extraction of common movements across companies

To learn more about the cross correlation detected in the previous section, we extract common movements across companies by applying random matrix theory (RMT) techniques to the cross-correlation matrix for the frequency of news reports. The cross-correlation matrix $\boldsymbol{C}$ is defined by

$$C_{i,j} = \frac{\langle f_{i,t} f_{j,t}\rangle - \langle f_{i,t}\rangle \langle f_{j,t}\rangle}{\sigma(f_{i,t})\sigma(f_{j,t})}, \tag{5·1}$$

and can be decomposed as

$$\boldsymbol{C} = \sum_{n=1}^{N} \lambda_n \boldsymbol{A}_n \boldsymbol{A}_n^T, \tag{5·2}$$

where λ_n is the n-th largest eigenvalue and $\boldsymbol{A}_n$ is the associated eigenvector. It has been shown that, if a cross-correlation matrix is generated from finite uncorrelated time series, the eigenvalue distribution of $\boldsymbol{C}$ is given by

$$p(\lambda) = \begin{cases} \frac{Q}{2\pi} \frac{\sqrt{(\lambda_{\max}-\lambda)(\lambda-\lambda_{\min})}}{\lambda} & \text{if} \quad \lambda_{\min} \le \lambda \le \lambda_{\max}, \\ 0 & \quad\quad \text{otherwise}, \end{cases} \tag{5·3}$$

where Q is defined as the ratio between the length of a time series L and the cross sectional dimension N (namely, $Q = L/N$), $\lambda_{\min} = \left(1 - \sqrt{1/Q}\right)^2$, and $\lambda_{\max} = \left(1 + \sqrt{1/Q}\right)^2$.[19),20)]

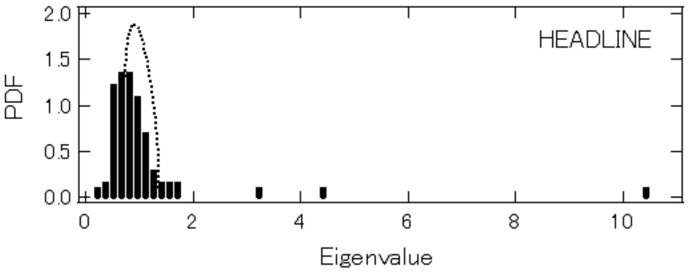

Fig. 9. The eigenvalue distribution for the case of headline news. This is estimated using the observations for the top 100 companies in the world in terms of market capitalization in 2003. The frequency of news is defined by the number of news reports per 50,000 ticks. The dotted line represents the eigenvalue distribution predicted for finite uncorrelated time series, as given by Eq. (5·3).

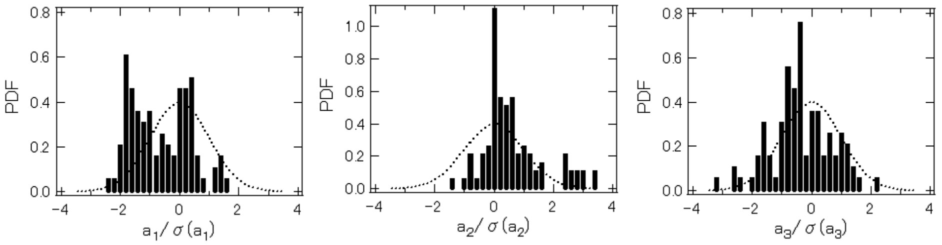

Fig. 10. The probability density functions of the eigenvector components associated with the first, second, and third largest eigenvalues. The horizontal axis shows the normalized component size (i.e., the size of the component divided by the standard deviation). The dotted line represents the standard normal distribution, which is predicted for finite uncorrelated time series.

The sample period we analyze covers seven years (January 2003 to December 2009), so that the length L of the time series is 3,274 (i.e. $3,274 \times 50,000$ ticks). As before, we pick the top 100 companies in terms of market capitalization in 2003. Given that $L = 3,274$ and $N = 100$, we have $\lambda_{\min} = 0.68$ and $\lambda_{\max} = 1.38$. Figure 9 shows the probability density function for the eigenvalues estimated from the cross-correlation matrix for the frequency of headline news, with the dotted line representing the eigenvalue distribution predicted for finite uncorrelated time series, as given by Eq. (5·3). There are eight eigenvalues exceeding $\lambda_{\max}$, with three of them exceeding $\lambda_{\max}$ by a large margin.

Figure 10 presents the probability density functions for the eigenvector components associated with the largest, second largest, and third largest eigenvalues. We see that they deviate significantly from a standard normal distribution, which is predicted for finite uncorrelated time series. Figure 11 shows the degree to which each company contributes to each of the eigenvectors associated with the three largest eigenvalues. The horizontal axis represents the 100 companies sorted by industry code. The three panels, each of which corresponds to the three largest eigenvalues, show that companies belonging to the financial sector contribute greatly to the eigenvector for the second largest eigenvalue (see the middle panel), and companies belonging to the information technology sector contribute greatly to the eigenvector

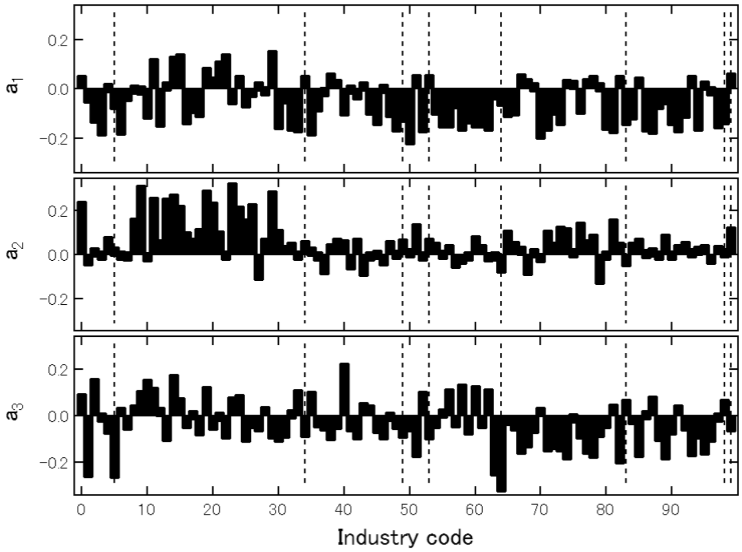

Fig. 11. Contributions of each company to the eigenvector components associated with the three largest eigenvalues of the correlation matrix. The upper, middle, and lower panels present the eigenvector components for the first, second, and third largest eigenvalues. The horizontal axis represents the 100 companies sorted by industry code. Industry codes 0-4 represent basic materials industries; 5-33 financial services industries; 34-48 consumer goods industries; 49-52 conglomerates; 53-63 services industries; 64-82 information technology industries; 83-97 health-care industries; 98 the industrial goods industry; and 99 utilities. The industry coding we use is available at <http://biz.yahoo.com/ic/ind_index.html>.

for the third largest eigenvalue (the bottom panel). On the other hand, the top panel shows that almost all non-financial companies contribute evenly to the eigenvector for the largest eigenvalue, which is similar to the result in Ref. 21) that the largest eigenvalue of the stock return correlation matrix is attributed to the "market mode" in financial markets.

Finally, we examine how the scaling exponent α changes when we eliminate common movement across companies. We start by defining F_t as follows:

$$F_t = \sum_i a_{1,i} f_{i,t}, \tag{5·4}$$

where $a_{1,i}$ denotes the eigenvector component i for the largest eigenvalue. A similar variable has been used to summarize common movement of stock prices (see, e.g. Ref. 20)). We then eliminate the common movement by regressing $f_{i,t}$ on F_t:

$$f_{i,t} = b_i + d_i F_t + \epsilon_{i,t}, \tag{5·5}$$

where b_i and d_i are regression coefficients. Using the residual term $\epsilon_{i,t}$ rather than $f_{i,t}$ itself, we estimate the scaling exponent $\alpha^{'}$ that satisfies a relationship of the form

$$\sigma(\epsilon_{i,t}) \propto \langle f_i \rangle^{\alpha^{'}}, \tag{5·6}$$

where $\sigma(\epsilon_{i,t})$ is the standard deviation of the residual term $\epsilon_{i,t}$. We find that the scaling exponent, which is equal to 0.63 when estimated using the original data, decreases to 0.61 when the common movement represented by F_t is removed. This result suggests that the deviation of α from 1/2 shown in the previous section stems, at least partially, from the common movement across companies captured by the largest eigenvalue of the correlation matrix. It is natural to suggest that the scaling exponent would approach 1/2 when one further eliminates common movements represented by other eigenvalues. One of our future research tasks is to see whether or not this is true by developing a method to eliminate the common movement represented by these eigenvalues.

§6. Conclusion

We have empirically investigated temporal and cross correlations in the frequency of news reports on companies using a dataset of more than 100 million news articles reported in English by around 500 press agencies during the period 2003–2009. Our main findings are as follows. First, the frequency of news reports on a company does not follow a Poisson process, but is instead characterized by long memory with a positive autocorrelation lasting more than a year. Second, there exist significant correlations in the frequency of news across companies. Specifically, on a daily or longer time scale, the frequency of news is governed by external dynamics such as an increase in the number of news reports due, for example, to the outbreak of an economic crisis, while it is governed by internal dynamics on a time scale of minutes. These two findings indicate that the frequency of news on a company has similar statistical properties as trading activities in stock markets, measured by trading volumes or price volatility, suggesting that the flow of information through news on companies plays an important role in price dynamics in stock markets.

Acknowledgements

We would like to thank Aki-Hiro Sato for helpful comments on an earlier version of this paper. We also thank the Yukawa Institute for Theoretical Physics at Kyoto University, where this work was completed during the YITP-W-11-04 on "Econophysics 2011". This work was supported in part by a Grant-in-Aid for Young Scientists (A) from the Ministry of Education, Culture, Sports, Science and Technology (No. 23686019).

References

1) D. M. Cutler, J. M. Poterba and L. H. Summers, J. of Portfolio Management **15** (1989), 4.
2) P. Balduzzi, E. J. Elton and T. C. Green, J. of Financial and Quantitative Analysis **36** (2001), 523.
3) M. L. Mitchell and J. H. Mulherin, J. of Finance **49** (1994), 923.
4) J.-P. Bouchaud, Y. Gefen, M. Potters and M. Wyart, Quantitative Finance **4** (2004), 176.
5) J.-P. Bouchaud, J. Kockelkoren and M. Potters, Quantitative Finance **6** (2006), 115.
6) A. Joulin, A. Lefevre, D. Grunberg and J.-P. Bouchaud, arXiv:0803.1769.
7) G. Harras and D. Sornette, Swiss Finance Institute Research Paper Series No. 08-16 (2008).
8) R. Lambiotte, M. Ausloos and M. Thelwall, J. of Informetrics **1** (2007), 277.

9) Y. Sano, K. Kasaki and M. Takayasu, *Proceedings of the 9th Asia-Pacific Complex Systems Conference, 2009*, p. 195.
10) D. Newman, C. Chemudugunta, P. Smyth and M. Steyvers, “Analyzing Entities and Topics in News Articles Using Statistical Topic Models”, in *Lecture Notes in Computer Science* Vol. 3975 (Springer, 2006), p. 93.
11) D. Leinweber and J. Sisk, “Relating News Analytics to Stock Returns”, in *The Handbook of News Analytics in Finance* (John Wiley & Sons, 2011), Chap. 6.
12) R. Cahan, Y. Luo, J. Jussa and M. Alvarez, Deutsche Bank Quantitative Strategy Report, July 2010.
13) D. A. Dickey and W. A. Fuller, J. of American Statistical Association **74** (1979), 427.
14) E. S. Said and D. A. Dickey, Biometrika **71** (1984), 599.
15) J.-P. Bouchaud and M. Potters, *Theory of Financial Risks and Derivative Pricing*, First edition (Cambridge University Press, 2000).
16) Z. Eisler, J. Kertesz, S. H. Yook and A. L. Barabasi, Europhys. Lett. **69** (2005), 664.
17) M. Argollo de Menezes and A. L. Barabasi, Phys. Rev. Lett. **92** (2004), 028701.
18) M. Argollo de Menezes and A. L. Barabasi, Phys. Rev. Lett. **93** (2004), 068701.
19) V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr and H. E. Stanley, Phys. Rev. Lett. **83** (1999), 1471.
20) V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr and H. E. Stanley, Phys. Rev. E **65** (2002), 066126.
21) C. Biely and S. Thurner, Quantitative Finance **8** (2008), 705.