Empirical analysis of collective human behavior for extraordinary events in blogosphere

Yukie Sano[1],* Kenta Yamada[2], Hayafumi Watanabe[3], Hideki Takayasu[4], and Misako Takayasu[3]
[1] *Laboratory of Physics, College of Science and Technology,*
Nihon University, 7-24-1 Narashinodai, Funabashi, Chiba, 274-8501, Japan
[2] *Waseda Institute for Advanced Study, 1-6-1 Nishi Waseda, Shinjuku-ku, Tokyo 169-8050, JAPAN*
[3] *Department of Computational Intelligence and Systems Science,*
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology,
4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8502, Japan and
[4] *Sony Computer Science Laboratories, 3-14-13 Higashi-Gotanda, Shinagawa-ku, Tokyo 141-0022, Japan*

To explain collective human behavior in blogosphere, we survey more than 1.8 billion entries and observe statistical properties of word appearance. We first estimate the basic properties of number fluctuation of ordinary words that appear almost uniformly. Then, we focus on those words that show dynamic growth with a tendency to diverge on a certain day, and also news words, that are typical keywords for natural disasters, grow suddenly with the occurrence of events and decay gradually with time. In both cases, the functional forms of growth and decay are generally approximated by power laws with exponents around -1 for a period of about 80 days. Our empirical analysis can be applied for the prediction of word frequency in blogosphere.

PACS numbers: 89.75.Da, 89.20.Hh, 89.65.Ef

1. INTRODUCTION

Investigations on macroscopic collective behavior of autonomous elements have attracted scientists' attention in many fields such as self-locomotive materials propelled by camphor [1], bird flocks [2] and fish shoals [3, 4]. Collective motion of these phenomena is recognized as a typical example of self-organization caused by neighbor interactions. In particular, collective behavior in human society has attracted considerable interest in the last decade because development in information technology enabled storage of large volumes of high-frequency human activity data. For instance, detecting bubbles in stock exchange rates [5], modeling dealer behavior using real data [6] in the foreign exchange markets, and the empirical analysis of consumer behavior in supermarkets [7] and convenience stores [8] using purchase history and point of sales (POS) data. Human activity data collected from the web, such as YouTube and social network services e.g., Facebook, are analyzed [9] [10] to not only explain basic individual human behavior but also elucidate hidden network structures in the society.

A blog is a type of website that is maintained by an individual with entries displayed chronologically with time stamps. The term "blog" originated from the combination of "web" and "log," and popularized around 2000 when free blog services began to be provided by internet service companies. A "blogger," who is the owner of a blog site, can easily upload their "entries" any time, and readers can easily post comments on the blog page. This interactive property has contributed to the success of blogs; they are now widely used as a basic social communication tool. The collective community of blogs is often called the "blogosphere," and scientific research on the blogosphere has already started.

The blogosphere is related to many science fields, such as statistical physics, engineering, sociology, linguistics, and psychology. A basic pioneering study of blog word appearance in the field of physics was carried out by Lambiotte et. al., who analyzed the appearance intervals of common words [11]. A naive Poissonian assumption was reported to generally fail event for low frequency words. However, two of the authors (Y.S. and M.T.) recently carefully examined similar time series of blog word appearances and showed that the Poissonian assumption holds for low frequency words after appropriate noise reduction procedure [12]. In the engineering field, Fujiki et. al., defined "burst" in the blogosphere from their own crawling data using natural language processing knowledge, and detected booming words automatically [13]. Shibata et. al., analyzed the network structure of trackback in the blogosphere and found that it to be sparse but highly clustered [14]. There are some theoretical and numerical models of individual bloggers that focus on network creation [15]. Grabowski et. al., determined power law distributions of individual activities of blogs and other web services, and introduced a simple model that is similar to SIR model of infectious diseases [16, 17]. Dodds et. al., defined the degree of "happiness" quantitatively by using the number frequency of target words related to happiness in blogs [18]. Bollen et. al., predicted direction changes in the Dow Jones Industrial Average for three days in the future with 87% accuracy by analyzing words in Twitter (http://twitter.com) [19] -a microblogging service started in 2006 suitable for mobile phones.

*E-mail: yukie.sano@gmail.com

In this study, we analyze the keyword appearance rate in blogs while focusing the functional forms of growth and decay around the peak, which are approximated by power laws. For earthquake research, the frequency of aftershocks is reported to decrease following a power law of time after the main shock, this is known as Omori's Law [20]. Similar power laws have been established in various fields of human society, for example, decrease in online book sales can be described by a power law with an exponent that depends on the types of the boom [21], for the change in number of audience for online movies, relaxation can also be reportedly described by the power laws with various values of the exponents that reflect the quality of the contents [9], and word frequencies in blogosphere among US presidential elections [22]. Alfi et. al., found that growth in registration numbers to conferences is also approximated by a power law diverging at the due date [23]. In the study of financial markets, price indices are known to diverge following the power laws in historical examples of hyperinflation [24]-[26], and the number density of transactions grows and decays following the power laws at the turning points of upward and downward trends [27].

In Sec. 2, we describe the analyzed data and in Sec. 3, we define peaked words and compare their statistics with those of ordinary words. In Sec. 4, we focus on the time evolution of these peaked words and prove that power law relaxations similar to Omori's law and power law divergences similar to the registration numbers can be observed for these words. In Sec. 5, we discuss the contents of the peaked words to clarify the difference before and after the peak. As an application of our empirical formulation, we introduce a method of estimating the number of words in the near future in Sec. 6, and conclusions are summarized in Sec. 7.

2. DATA DISCRIPTION

The data analyzed in this study is obtained from the blogosphere written in Japanese over a period of four years from November 1, 2006 to October 31, 2010. According to the technical report by the internet search engine company Technorati (http://technorati.com) that tracked more than 70 million blogs world-wide in 2007, the share of Japanese blogs is 37%, the largest among all languages [28]. The another report found that 78% of internet users in Japan regularly browsed blogs, which was also the highest among the other countries [29].

In the blogosphere research, it is important to note the existence of spam blogs called "splogs." Splogs are automatically generated blogs in which the same words are repeated multiple times mainly for the purpose of advertisement. The share of splogs in the Japanese blogosphere is about 40% [30]; therefore, it is important to exclude splogs from the data in order to study collective human activities.

We use a new internet service named "Kuchikomi@kakaricho" (http://kakaricho.jp) to collect the data. This service provides an application programming interface (API) that counts the number of entries in which a given target word appeared in a given period by using a search engine technology with a "spam filter." There are three levels of spam filters in this system and we apply the middle level that is known to remove most of the splogs and keep most of the human blogs untouched [31]-[36]. The API counts the number of entries in the blogs such that if one entry includes the target word multiple times, the word is counted only once. The API started crawling the blogosphere on November 1, 2006 and covered 22 major blog service providers with more than 1.8 billion blog entries in 15 million blogs covering 90% of the Japanese blogosphere. Note that there are bloggers who simultaneously maintain multiple blogs; thus, the number of blogs is not exactly equal to the number of people, but is considered to be about the same.

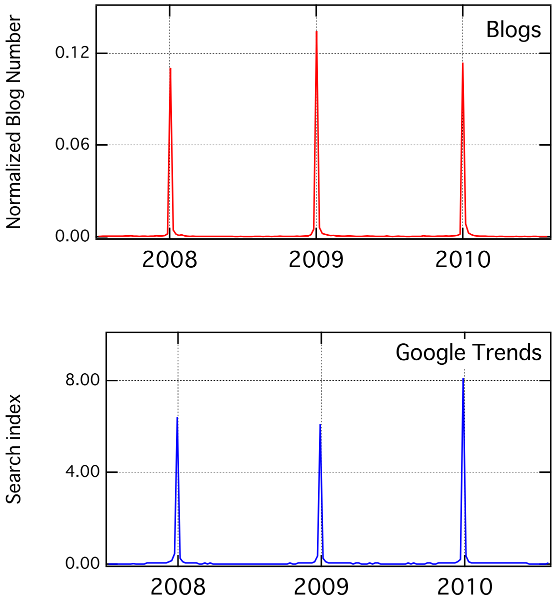

FIG. 1: (Color online) Temporal change of the frequency of the phase "April fool" per week; the results are from our blog data "Kuchikomi@kakaricho," which is targeted only in Japan and Google Trends, which is targeted worldwide. Here, the number of blogs is normalized by the whole number.

Although we only analyze the Japanese blogosphere, we show an example in which the dynamic properties of the blogospheres in Japanese and English are considerably similar. Figure 1 shows the temporal change of the frequency of the English target phase "April Fool" observed by Google Trends (http://www.google.com/trends) surveyed worldwide compared to the number of blog entries containing the corresponding Japanese phase in our Japanese blog data for the period of three years. In both cases, we confirm that there is a clear peak on the week including April fools' day. The result of Google Trends is calculated by a closed proprietary method [37], whereas our data and method are open for the public. We believe that our data is suitable for the scientific study of human behavior in the blogosphere.

3. PEAKED WORDS

In the blogosphere, there are special words whose frequency grow or decay around a peak day such as "April Fool" with the peak on April 1 as shown in Fig. 1. We denote these words as "peaked words" in this paper, and analyze the functional forms of growth and decay.

3.1. Pretreatment

We apply the following pretreatment to the data to exclude both trivial circadian human activity patterns and systematic noises.

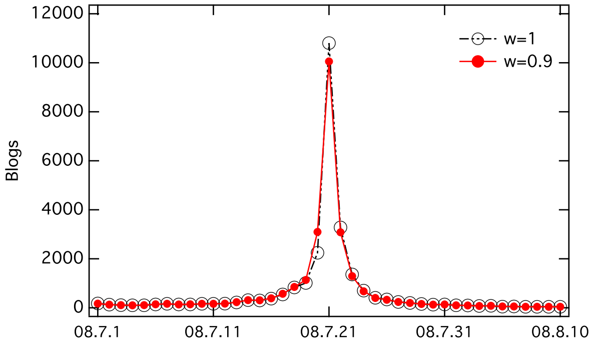

FIG. 2: (Color online) Typical example of time series with peak that includes the phase "Marine Day." $w = 1$ corresponds to no revision and $w = 0.9$ corresponds to modified time series by Eq. (1). Due to circadian effect, position of the day after peak is always higher than that before the peak without modification.

Time-Shift Figure 2 shows an example of the daily number of entries including the phase "Marine Day," ("Umi-no-hi" in Japanese), which is a national holiday in Japan falling on the third Monday of July. In Fig. 2, open circles show the original daily data. As shown in the figure, the numbers of entries before and after the peak are not symmetric. In the blogosphere, it is not trivial that a day starts at 0:00 because there are many bloggers who are active at midnight; therefore, we examine the complete circadian activity pattern and introduce a type of correction pretreatment for our daily data. For this purpose, we randomly chose the data of 10,000 bloggers with the detail of their activities time stamped in seconds. By counting the number of all submitted entries at every hour, a circadian activity pattern is plotted in Fig. 3. The solid line shows the 24-hour-activity pattern obtained directly from the data. However, we discovered certain amount of data with time stamps that are exactly 00:00:00. We consider this time stamp to be caused by an artificial systematic error and we exclude this data from the statistics when capturing the circadian pattern; the revised circadian activity pattern is shown by the red bars. From this result, we find that the activity is lowest around 4:00, and consider the start of a day at 5:00 to be reasonable. Because the share of activity in the interval between 0:00 to 5:00 is approximately 10% of the complete activity of a day, we can correct the daily number of blog entries including the j-th target word at the t-th day, $\tilde{x_j}(t)$, by the following equation:

$$x'_j(t) = w\tilde{x_j}(t) + (1-w)\tilde{x_j}(t+1), \qquad (1)$$

where the weight is set as $w = 0.9$. With this modification, we can determine time-shifted time series in Fig. 2, shown by colored circles, which show a more symmetric pattern than the original data. We also apply this procedure to determine the time series of the total number of blog entries per day, $X'(t)$.

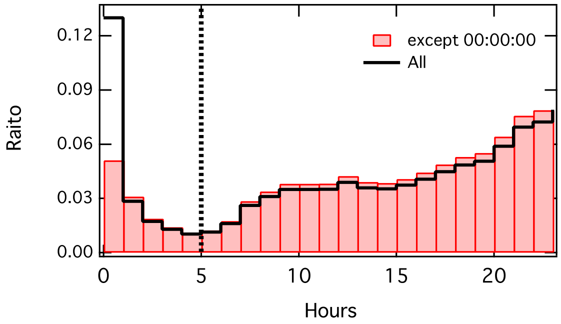

FIG. 3: (Color online) Ratio for circadian activity of blog posted by 10000 bloggers. Solid line is calculated from all entries and the red bar is from the entries excepted that have time stamp of 00:00:00. In both, 4:00 is the smallest ratio in a day.

Normalization There are non-uniform and non-stationary properties in the total number of entries per day [12]. For example, there was a sudden drop in February 2007 that was caused by the crawling system's maintenance. The reason for this has not been has not been clarified yet; however, the total number of entries per day tends to gradually increase after May 2007. In order to reduce the systematic fluctuations caused by such non-uniform properties, we apply the following normalization procedure. The normalized number of entries for the j-th word on the t-th day is defined by $x_j(t) = x'_j(t)\frac{\langle X'\rangle}{X'(t)}$, where $\langle X'\rangle$ denotes the mean value of $X'(t)$ averaged over the entire observation period. By introducing this normalization, the fluctuations caused by the aforementioned non-uniform properties can be reduced. In this study, we measure the word frequency with this normalization procedure. Note that the number of entries is not necessarily an integer.

3.2. Word Selection

Next, we discuss the method for determining the peaked words. For this purpose, we compare the basic properties of the typical peaked words with those of the "ordinary words," which are expected to appear uniformly. Two authors have already confirmed that adjective and conjunctions can be viewed as typical ordinary words [38]; therefore, by determining the basic properties of the appearance of the ordinary words, we can quantitatively specify the peculiarity of the peaked words.

We selected 1781 adjectives from the Japanese language morphological analysis dictionary of MeCab (http://mecab.sourceforge.net/) and examined the statistics of appearance of these words. We confirmed that 1771 out of 1781 words were used multiple times during our observation period, and we focused on those words. In Fig. 4, we plotted the average of frequency of the j-th word, $\langle x_j\rangle$, and its standard deviation, σ_j, in a log-log scale. A naive Poisson assumption gives the square root relation between the standard deviation and the average. However, this relation holds only for the words with a small value of average. For larger values of average, we can confirm a linear relation between the average and the standard deviation caused by the number fluctuation of all bloggers, which can be approximated by the following simple relation [38],

$$\sigma_j \simeq \sqrt{\langle x_j\rangle\left(1 + a^2\langle x_j\rangle\right)}, \qquad (2)$$

where $a = \frac{\sqrt{\langle X^2\rangle_c}}{\langle X\rangle}$ a is constant parameter characterizing the number fluctuation of all bloggers that is determined independent of the word (see Appendix A for theoretical derivation). Therefore, we can get $a = 0.3$ for the raw time series, $\tilde{x_j}(t)$, while $a = 0.08$ for the time series that is normalized by the total number, $x_j(t)$. Note that although after the normalization by the total number, $X'(t)$, there is still a non-trivial fluctuation in each time series, $x_j(t)$.

This relation between the average and the standard deviation can be used to determine the peaked words. In Fig. 4, the average and the standard deviation for the typical peaked phase "Marine Day" is plotted by a red circle. It is clear that the point is located far above the fluctuation level for the ordinary words. Using this plot, we can estimate whether the number fluctuation of a peaked word can be considered to be within that of an ordinary word. For abnormal cases, we determine the time evolution individually to determine our research target of the peaked words.

After an intensive survey of the peaked words, we chose 435 words. Here, we categorize these words into the following three types.

Date We considered the date such as "9th May," including 366 words. There are many blog entries that announce the day, for example, "my birthday," "festival," and "concert." As explained in the next section, growth and decay of these words always show a clear peak at the date.

Event We selected the names of 14 public holidays and 16 major annual evens in Japan, such as "April fool." The appearance rate for this word also grows and decays around the date of the event. These are the words affiliated with an event such as "Santa Claus" for "Christmas;" we can observe similar growth and decay behavior. However, we neglected such affiliated words in this analysis.

News A word such as "earthquake" occurs suddenly right after the events, and the word appearance rate generally decays slowly. In order to observe the functional form of such decay caused after a significant event, we selected names of the places attacked by earthquakes. We also selected 33 names of famous persons who died suddenly. In addition, we included the names of the Japanese scientists who recieved a Nobel Prize during our observation period.

4. FREQUENCY OF PEAKED WORDS

4.1. Slope Periods

To characterize the temporal behavior of number of entries of the peaked word, we categorize the entrie period into "slope periods" and "normal periods." A slope period of the j-th word is defined as the period before or after the peak day in which the daily number of entries of this word continuously exceeds a certain threshold of the fluctuation level. A normal period is defined as the remainder of the period during which the word is expected to appear following the statistics identical to those of the ordinary words.

To determine this threshold, we calculate median value of $x_j(t)$ for the entire period and define it as $\bar{x_j}$, and for the slope period, we require condition $x_j(t) \geq \bar{x_j} - \sigma_j(t)$.

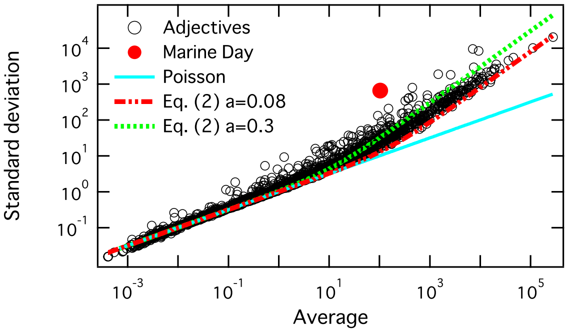

FIG. 4: (Color online) Relation between average and standard deviation of word frequency in the blogosphere. 1771 adjectives (open circles) and phase "Marine Day" (red circle), which is a typical example that has clear peaks. Eq. (2) with different parameters; $a = 0.30$ and $a = 0.08$ duplicated in the figure.

We apply the median value instead of the mean value for estimating the fluctuation level of the normal period, because the estimation of the mean value tends to considerably higher than that of a typical value in the normal period. The mode value is another candidate for characterizing the fluctuation level of the normal period. However, it depends on the bin size, and the statistics are not stable; thus, we chose the median. We have attempted to use $\bar{x_j} + \sigma_j(t)$, $\bar{x_j}$ and $\bar{x_j} - \sigma_j(t)$ as the candidates of the threshold value, and found that threshold $\bar{x_j} - \sigma_j(t)$ separates the slope periods and the normal periods most accurately for almost all the peaked words. In the cases of $\bar{x_j}+\sigma_j(t)$ and $\bar{x_j}$, the slope period tends to be cut shorter than that in the case when the trend actually converges to the normal fluctuation level. By applying this definition of slope period analysis, we can measure the length of the slope periods for each peaked word and the results are summarized in Tab. II and Fig. 11 in Appendix B. Here, we denote the slopes before the peak day as fore-slopes and those after the peak day as after-slopes. The average length is 79 days for the fore-slopes and 75 days for the after-slopes. For the peaked words categorized in news, there is no fore-slope in general; thus, we consider only the after-slopes. In the case of news words, the median value might be changed before and after the peak. In order to exclude this effect, we only consider the median value of the after-slope in $\bar{x_j}$. There are cases in which the slope periods are terminated by the limitation of the entire data period; those cases are omitted. In general, the lengths of the fore-slopes and the after-slopes are distributed roughly in the same manner around 80 days with the standard of deviation about 20 days, together with a mild tendency for the after-slopes to be shorter than the fore-slopes.

4.2. Method

Next, we consider the functional form of the slopes. As candidates for the functional forms of growth and decay around the peak day, we considered an exponential function, a power function, and a stretched exponential function; we found that the power function fit well for most of the cases. Following is the functional form that we fit with the data;

$$x_j(t) = A_j|t_c - t|^{-\alpha_j} + \bar{x_j}, \tag{3}$$

where the parameters are α_j and A_j. Next, we estimate power exponent α_j and A_j in Eq. (3) for the fore-slope and the after-slope. The maximum-likelihood method is applied to the data points plotted in the log-log scale with an average taken over each box whose bin size is uniform in the log scale. For distribution of $x_j(t)$, we assume the normal distribution with the average to be $\langle x_j(t_c \pm t)\rangle = \langle x_{jt}\rangle$ and the standard deviation to be $\sigma_j(t_c \pm t) = \sigma_{jt}$. We estimate a set of parameters that maximize the logarithm of the likelihood function, $\log L$, as defined by the following equations.

$$P(x_{jt}; \vec{\theta_{jt}}) = \frac{1}{\sqrt{2\pi\sigma_{jt}{}^2}} exp\Bigg\{-\frac{(x_{jt} - \langle x_{jt}\rangle)^2}{2\sigma_{jt}{}^2}\Bigg\}, \tag{4}$$

where $P(\vec{\theta_{jt}})$ is given by normal distribution with parameter set $\vec{\theta_{jt}} = (\langle x_{jt}\rangle, \sigma_{jt})$. As a result,

$$L(x_{j1}, \cdots x_{jn}; \vec{\theta_{j1}}, \cdots \vec{\theta_{jn_j}}) = \prod_{t=1}^{n_j} P(x_{jt}; \vec{\theta_{jt}}), \tag{5}$$

where n_j is the number of data points in the slope period. $\langle x_{jt}\rangle$ is given by Eq. (3) and σ_{jt} by Eq. (2) with tuned parameters α_j and A_j.

4.3. Results

For example, for the observed peak day, t_p, April 1 for the phase "April fool," we assume the following relation with the mathematical divergence point, $t = t_c$, in Eq. (3). Here, we consider that the divergence point, t_c, is the point that people change their feeling about the peaked words; namely, before t_c, people expect the peaked words as future events. However, after t_c, people use the words to remember the events, which we call the forgetting phase. We consider that t_p is very close to t_c with time interval ϵ_j; namely, if t_p is located before t_c, $t_p = t_c - \epsilon_j$, otherwise $t_p = t_c + \epsilon_j$, where ϵ_j is expected to take a value less than 1. We estimate three parameters, α_j, A_j, and ϵ_j, simultaneously by minimizing the sum of residual errors with the estimated power function and the real data. Note that there are cases where ϵ_j takes a value higher than 1. For the example, if $\epsilon_j \geq 1$ for "April fool," it indicates that the number of blogs posted on March 31 belongs to the forgetting phase; while, in reality, the event is not finished. To avoid this contradiction, we only consider the case of $\epsilon_j < 1$ in the following analysis.

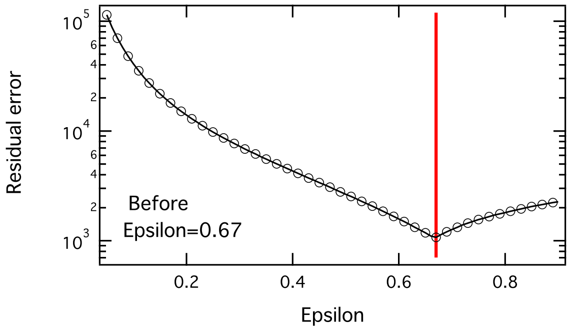

FIG. 5: (Color online) Typical result of residual error with ϵ_j changing before Marine Day in 2008. Final value of ϵ_j is 0.67 of the smallest residual error in the semi-logarithmic scale.

Figure 5 is a typical result of residual error as a function of ϵ_j for the phase "Marine Day" in 2008. As shown in Fig. 12(a) in Appendix C, the estimated values of ϵ_j typically lie in the range $0 < \epsilon_j < 1$ before the event with the average of 0.4 ± 0.5. In the case of after the event, there are cases with $\epsilon_j \geq 1$ as shown in Fig. 12(b).

Figure 6 is a typical result of data fitting for the phase, "Marine Day" in 2008. The functional form of Eq. (3) fits well for the entire range of both before and after peaks. Distributions of the estimated power exponents are shown in Fig. 7 and the average values are summarized in Tab. I. The power exponent, α_j, of the after-slope tends to be slightly higher than that of the fore-slope. We confirmed that the value of the power exponents of the after-slopes are slightly higher than those of the fore-slopes for 87% of 1402 samples. This result implies that growth and decay of people's interest is not symmetric. This asymmetry is easily recognized by considering the case of "Christmas," namely, television and social media use the word "Christmas" for advertising of bargain sales that end on December 25. Thus, people tend to easily forget the word "Christmas" after the peak day. We also expect that there are more chances for bloggers to interact with each other before Christmas by mutually reading each others blogs to plan their holidays.

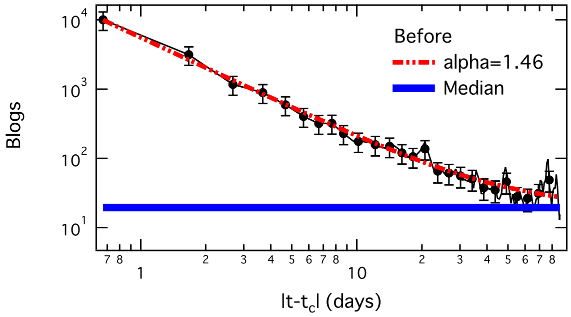

FIG. 6: (Color online) Typical result of data fitting for the same sample as Fig. 2 before "Marine Day" in 2008 in logarithmic scale. Solid line is $x_j(t)$ and dashed line is median value $\bar{x_j} = 14.0$ which we use to define slope period. Markers indicate estimated points that averaged in boxes with logarithm size. Slope length is 85 days.

TABLE I: Average values of power exponent α_j.

		Average α_j	#Samples
Event	Before	1.45± 0.38	107
	After	1.38 ± 0.29	116
Date	Before	0.89± 0.18	1381
	After	1.19± 0.15	1374
News	After	1.28± 0.32	39
All	Before	0.93 ± 0.24	1488
	After	1.21 ± 0.18	1529

In the case of the news words, there is no fore-slope and we cannot compare the values of the exponents before and after the peak. Noted that the values of the exponent after the peak tend to be estimated smaller for high impact news because of the effect of sequential broadcasts of the aftershocks. For example, in the case of the sudden death of the world famous entertainer Michael Jackson, which marked the peak day, there was a funeral service after a few days and a memorial CD released after a few weeks. Both can be regarded as aftershocks and remind us of the main news. Because of such repetition, the keyword appearance rate after the peak day is enhanced, the decay of the word appearance becomes slower, and the power exponent tends to take a smaller value.

5. CONTENTS OF PEAKED WORDS

In Sec. 4, we showed that growth and decay of the peaked words are well described by the power functions. In this section, we focus on the contents of the peaked words for better understanding the divergence point, t_c. In the study of phase transition phenomena in statistical physics, the divergence point of the control parameter is given by the point of power law divergence of an order parameter. Using this analogy, we consider that focus of people's attention changes at the divergence point, t_c, implying the presence of a type of dynamic phase transition from an expecting phase to a forgetting phase. Moreover, we examined the context in the blogs before and after t_c. We refined the number of blogs that included the

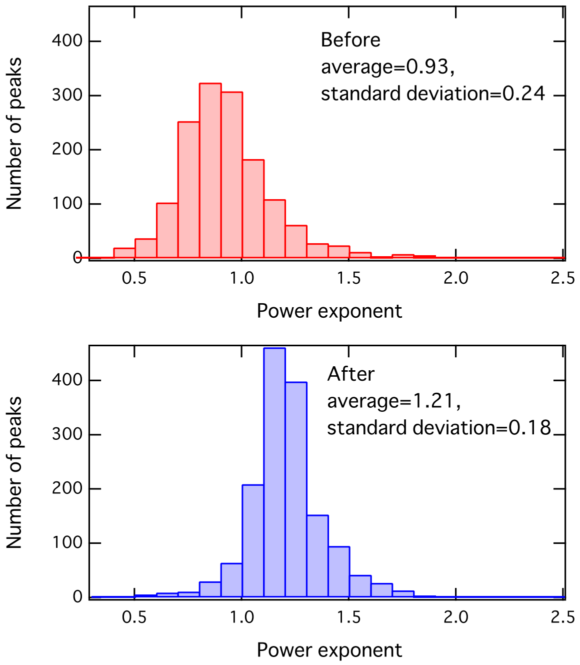

FIG. 7: (Color online) Distribution of power exponent α_j before and after t_c. Average value of α_j of after the peak day is higher than before in many cases.

j-th word by adding the condition, whether it appears with the following keywords; "yesterday," "today," and "tomorrow." For example, a typical blog context before Marine Day is "Tomorrow is holiday! Because it's Marine Day!!!." On the other hand, a typical blog context after Marine Day is "I went to BBQ with my friends yesterday, because it was a holiday of Marine Day." As expected, the conditional probability of blogs including "tomorrow" under the condition of including "Marine Day" has the highest peak on the day before Marine Day, whereas "yesterday" has the highest peak on the day after Marine Day, as shown in Fig. 8. This result demonstrates that bloggers actually change the context of the blogs from the expecting phase to the forgetting phase.

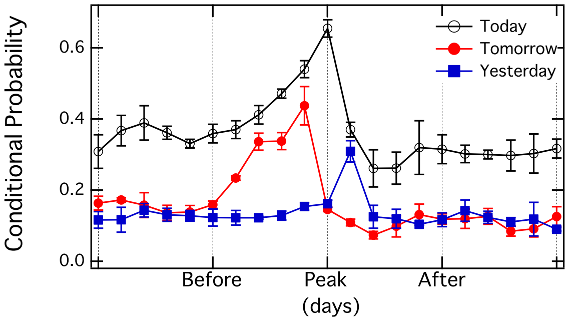

FIG. 8: (Color online) Conditional probability of finding a blog including "today" (open circles), "tomorrow" (red circles), or "yesterday" (blue squares) under the condition of including "Marine Day". The average is taken over four years from 2007 to 2010, and the error bars show the standard deviations.

6. PREDICTABILITY OF FREQUENCY

As an application of our empirical formulation, we explore the possibility of estimating the word frequency in the near future. In Fig. 9, we show an example of prediction of blog frequency with the phase "Marine Day" in 2008. In this case, we already have the information about the peak days to be July 21, 2008; thus, we can fix the divergence point on this day. From the data, we find that the slope period starts on April 28, 85 days before the critical point, as the normalized frequency continuously exceeds the median value from this day. In Fig. 9(a), the case of prediction for 20 days before the divergence point using 65 data points with Eq. (5) is shown by the dashed line. It is remarkable that the peak level can be estimated accurately at 20 days before the peak day, while the peak value is about 100 times higher than the value of the frequency level on the estimated day. In the right figure of Fig. 9(b), the case of prediction for 10 days before the peak day is shown. In this case, the fitting is very good for all the days before the peak day. However, note that a small difference in estimation of the exponent makes a big difference near the peak; thus, the number of data points plays an important role in accuracy.

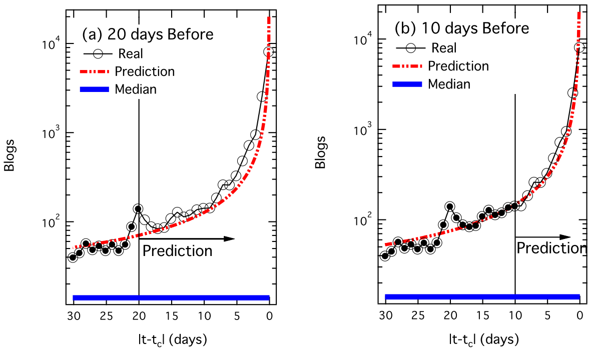

FIG. 9: (Color online) Typical examples of prediction for (a) 20 days and (b) 10 days before the peak day for the phase "Marine Day" in 2008 in semi-logarithmic scale. Red dashed line indicates the prediction line, blue bold line indicates median value $\bar{x_j} = 14.0$, and solid line shows the observed values. Open circles indicate the future values and colored circles are the known values used for prediction. Because the estimation is started 85 days before t_p, the median values are located far below the observed values. Estimated values are $\alpha_j = 1.00$, $A_j = 1131$ and $\epsilon_j = 0.14$ for 20 days before the peak day and $\alpha_j = 1.12$, $A_j = 1753$ and $\epsilon_j = 0.09$ for 10 days before the peak day.

We can also apply our method for after-slope prediction. On March 11, 2011, a huge earthquake occurred near the Northeast coast of Japan and more than 20,000 people were killed by a Tsunami. We can confirm the power law decay of the word "Tsunami" in the Japanese blogosphere as shown in Fig. 10(a). The peak day was March 12, i.e., the day after the quake. Because the estimated power exponent is 0.87 and the peak frequency is 142,678, it is expected to take approximately two years to return to the normal fluctuation level, that is, about 300 word appearances per day. Although most of the news words decay in about 80 days, as mentioned in Sec. 4.3, the case of "Tsunami" is a rare exception because the number of entries is seven times higher than the normal level even for 100 days after March 11.

Twitter also shows similar power law behavior even though the time resolution is different. Figure 10(b) shows the time evolution of the number of tweets including "Tsunami" measured per hour. Interestingly, the same power exponent, 0.87, is duplicated in the Twitter database in which 1,397,783 tweets are found. We expect that this type of power law reflects robust empirical dynamics of collective human behavior.

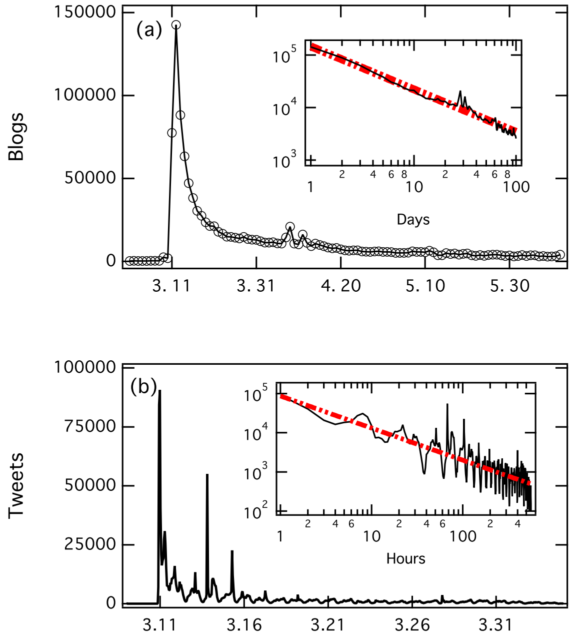

FIG. 10: (Color online) Decay of "Tsunami" observed for (a) blogs and (b) tweets. Horizontal time step size is per day for blogs and per hour for tweets. (Inset) Log-log plots of the time series. In both cases, red dashed line shows the slope of power law with the exponent $\alpha_j = 0.87$.

7. CONCLUSIONS

By analyzing a large database of Japanese blogs, we showed that the functional forms of growth and decay of word appearance peaked at a certain day and are generally approximated by the power laws with the exponents around 1. The values of the power exponents depend on the category of words such as event, date, and news. Clarification of this asymmetry in the power exponents of the fore-slope and the after-slope can be an interesting future subject for research on collective human behavior. It is likely that in the expecting phase, interactions among bloggers might be stronger with the effects of mass media, while in the forgetting phase individual human principles such as "Ebbinghaus Forgetting Curve" [39], which describes quantitatively how people forget, might play an important role.

As an application, we demonstrated the predictive power of our empirical laws for the estimation of both the peak values in advance and how people forget the events. These empirical properties of bloggers activity will be reproduced by an agent-based mathematical model in the near future [42].

Acknowledgments

The authors thank to Dentsu Kansai Inc. and Hottolink Inc. for useful discussions and providing the data. This work is partly supported by Grant-in-Aid for JSPS Fellows Grant No. 219685 (H.W.) and Grant-in-Aid for Scientific Research No. 22656025 (M.T.) from the Research Foundations of the Japan Society.

Appendix A: Modified Random Diffusion Model

Here, we introduce a modified random diffusion model. The random diffusion model was originally introduced to describe diffusion properties of random walkers on a given network [40, 41], and two of the authors (Y.S. and M. T.) have modified the model to be applicable to the number fluctuations of word appearance in the blogosphere [38]. In our modified random diffusion model, we assume that there are two states, active and non-active for each blogger, and the number of active bloggers fluctuates randomly with each day. Each active blogger randomly decides to post a blog including the j-th word. There is a key parameter in this stochastic process; the share of the j-th word, c_j, is defined by the following equation;

$$c_j = \frac{\langle x_j \rangle}{\langle X \rangle}, \tag{A1}$$

where $x_j(t)$ is the number of blog entries including the j-th word at the t-th day. $X(t)$ is the number of active bloggers at the t-th day and the brackets show the average over all realizations. We assume that the number of active bloggers, $X(t)$, $X(t) \geq 0$, fluctuates randomly following an independent probability density distribution, $\phi(X)$, with finite moments. Probability of posting x_j entries is calculated using Poisson distribution with average number $c_j X$ given as follows:

$$P(x_j|c_j) = \int_0^\infty \phi(X) \exp\left(-c_j X\right) \frac{\left(c_j X\right)^{x_j}}{x_j!} dX. \qquad \text{(A2)}$$

When $\langle x_j \rangle$ is small, a Poisson distribution is approximated by a Bernoulli distribution that assumes $x_j = 0$ with probability $1 - c_j X$, and $x_j = 1$ with probability $c_j X$. Thus, we have the following evaluations for an arbitrary distribution of $\phi(X)$.

$$\begin{aligned} P(x_j = 0|c_j) &\simeq \int_0^\infty \phi(X) \left(1 - c_j X\right) dX \\ &\simeq 1 - c_j \langle X \rangle, \\ P(x_j = 1|c_j) &\simeq \int_0^\infty \phi(X) \left(c_j X\right) dX \\ &\simeq c_j \langle X \rangle. \end{aligned} \qquad \text{(A3)}$$

For $\langle x_j \rangle \approx 2$, $P(x_j \geq 2|c_j) \approx 0$, thereby $P(x_j|c_j)$ is approximated by the Poisson distribution with both the average and the variance given by $c_j \langle X \rangle$.

For $\langle x_j \rangle \gg 1$, the Poisson distribution can be approximated by a normal distribution,

$$P(x_j|c_j) \simeq \int_0^\infty \phi(X) \frac{1}{\sqrt{2\pi c_j X}} \exp\left[-\frac{(x_j - c_j X)^2}{2 c_j X}\right] dX. \qquad \text{(A4)}$$

By introducing a new variable, $y_j = \frac{x_j}{c_j \langle X \rangle}$, Eq. (A4) becomes

$$P(y_j|c_j) \simeq \int_0^\infty \phi(X) \frac{1}{\sqrt{2\pi \left(\frac{X}{c_j \langle X \rangle^2}\right)}} \exp\left[-\frac{(y_j - \frac{X}{\langle X \rangle})^2}{2\left(\frac{X}{c_j \langle X \rangle^2}\right)}\right] dX. \qquad \text{(A5)}$$

When $\langle x_j \rangle = c_j \langle X \rangle \gg 1$, the weight function in the integral can be approximated by Dirac's delta function as,

$$P(y_j|c_j) \simeq \int_0^\infty \phi(X) \delta\left(y_j - \frac{X}{\langle X \rangle}\right) dX. \qquad \text{(A6)}$$

Therefore, we have the following simple evaluation, for x_j,

$$P(x_j|c_j) \simeq \frac{1}{c_j} \phi\left(\frac{x_j}{c_j}\right). \qquad \text{(A7)}$$

Calculating the first and second moments of $P(x_j|c_j)$, we now have the general results,

$$\begin{aligned} \langle x_j \rangle &= \int_0^\infty x_j P(x_j|c_j) dx_j \\ &\simeq \int_0^\infty x_j \frac{1}{c_j} \phi\left(\frac{x_j}{c_j}\right) dx_j = c_j \langle X \rangle, \qquad \text{(A8)} \\ \langle x_j^2 \rangle &= \int_0^\infty x_j^2 P(x_j|c_j) dx_j \\ &\simeq \int_0^\infty x_j^2 \frac{1}{c_j} \phi\left(\frac{x_j}{c_j}\right) dx_j = {c_j}^2 \langle X^2 \rangle. \qquad \text{(A9)} \end{aligned}$$

From these results, standard deviation $\sigma_j = \sqrt{\langle x_j^2 \rangle - \langle x_j \rangle^2}$ can be expressed as

$$\sigma_j \simeq \sqrt{{c_j}^2 \left(\langle X^2 \rangle - \langle X \rangle^2\right)}. \qquad \text{(A10)}$$

By correlating both results where $\langle x_j \rangle$ is small and large, we can assume the following relation;

$$\sigma_j \simeq \sqrt{c_j \langle X \rangle + {c_j}^2 \langle X^2 \rangle_c}, \qquad \text{(A11)}$$

where $\langle X^2 \rangle_c$ denotes the second order cumulant.

Appendix B: Slope length

In Tab. II, we show the typical values for the averages and the standard deviations of the fore-slopes and the after-slopes in different word categories. Corresponding distributions are plotted in Fig. 11.

TABLE II: Average of slope length in different word categories.

		Average slope (days)	#Samples
Event	Before	65 ± 23	108
	After	65 ± 25	116
Date	Before	80 ± 21	1381
	After	76 ± 20	1373
News	After	103 ± 76	39
All	Before	79 ± 21	1489
	After	75 ± 24	1528

Appendix C: Position of the peak

Here we show the averages and the standard deviations of ϵ_j, where ϵ_j is the gap between the divergence point, t_c, and the observed peak day, t_p. ϵ_j is calculated by minimizing the residual errors between the observed values and the estimated power function.

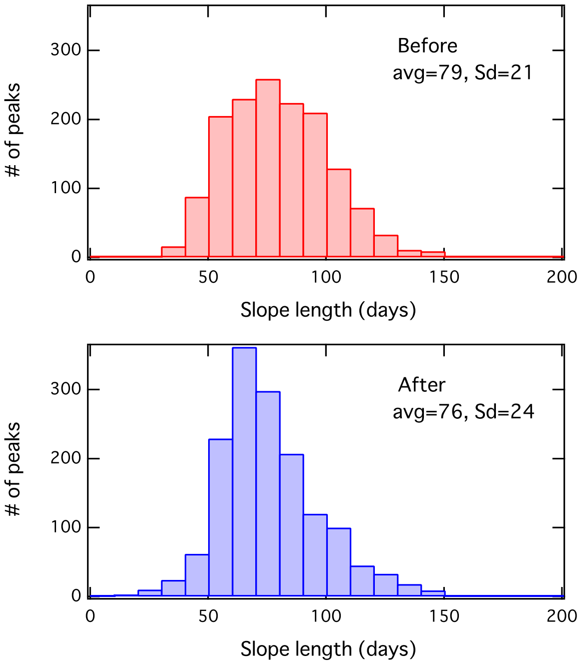

FIG. 11: Distribution of slope length of before and after peaks. In both cases, distribution continues around 80 days. Distribution of the before peak is larger than that of the after peak; possible causes are interactions of bloggers and external media effect.

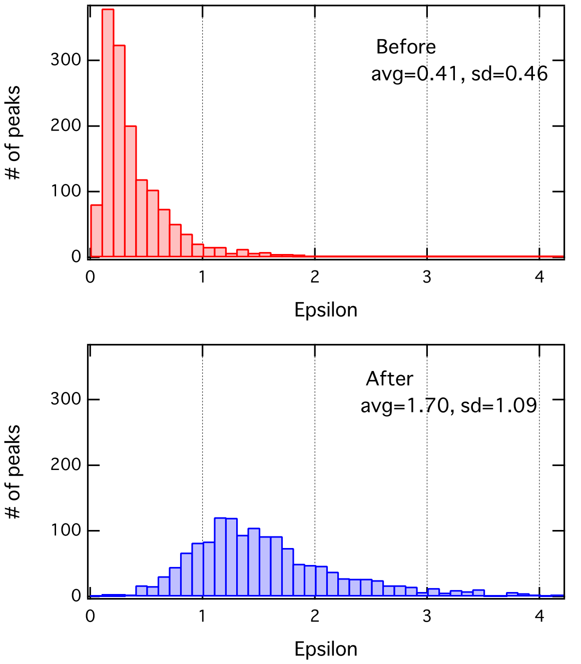

FIG. 12: Distribution of ϵ_j for before and after peaks.

TABLE III: Averaged values of ϵ_j in different word categories.

		Average ϵ_j	#Samples
Event	Before	0.84± 0.83	107
	After	1.86 ± 1.86	116
Date	Before	0.38± 0.40	1378
	After	1.65± 0.89	1362
News	After	2.98± 2.51	39
All	Before	0.41 ± 0.46	1485
	After	1.70 ± 1.09	1517

[1] N. J. Suematsu, S. Nakata, A. Awazu, and H. Nishimori, Phys. Rev. E **81**, 056210 (2010).

[2] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale Proc. Natl. Acad. Sci. USA **107**, 11865 (2010).

[3] L. Conradt, Nature (London) **471**, 40 (2011).

[4] A. J. W. Warda, J. E. Herbert-Read, D. J. T. Sumpter, and J. Krause, Proc. Natl. Acad. Sci. USA **108**, 2312 (2011).

[5] K. Watanabe, H. Takayasu, and M. Takayasu, Phys. Rev. E **80**, 056110 (2009).

[6] K. Yamada, H. Takayasu, T. Ito, and M. Takayasu, Phys. Rev. E **79**, 051120 (2009).

[7] H. Ueno, T. Watanabe, H. Takayasu, and M. Takayasu, Physica A, **390**, 499 (2011).

[8] T. Mizuno, M. Toriyama, T. Terano, and M. Takayasu, Physica A **387**, 3931 (2008).

[9] R. Crane and D. Sornette, Proc. Natl. Acad. Sci. USA **105**, 15649 (2008).

[10] A. L. Traud, P. J. Mucha, and M. A. Porter, e-print arXiv:1102.2166.

[11] R. Lambiotte, M. Ausloos, and M. Thelwall, J. of Informetrics, **1**, 277 (2007).

[12] Y. Sano and M. Takayasu, J. of Economic Interaction and Coordination, **5**, 221 (2010).

[13] T. Fujiki, T. Nanno, Y. Suzuki, and M. Okumura, *in Proceedings of the First International Workshop on Knowledge Discovery on Data Streams*, (2004).

[14] N. Shibata, M. Uchida, Y. Kajikawa, Y. Takeda, S. Shirayama, and K. Matsushima, *in Proceedings of the International Workshop and Conference on Network Science*, (2007).

[15] M. Goetz, J. Leskovec, M. Mcglohon, and C. Faloutsos, *in Proceedings of AAAI Conference on Weblogs and Social Media*, (2009).

[16] A. Grabowski and R. A. Kosinski, Phys. Rev. E **82**, 066108 (2010).

[17] A. Grabowski, Eur. Phys. J. B **69**, 605 (2009).

[18] P. S. Dodds and C. M. Danforth, J. of Happiness Studies **11**, 441 (2010).

[19] J. Bollen, H. Mao, and X.-J. Zeng, e-print arXiv:1010.3003.

[20] F. Omori, J. of the College of Science (Imperial University of Tokyo), **7**, 111 (1894).

[21] D. Sornette, F. Deschatres, T. Gilbert, and Y. Ageon, Phys. Rev. Lett. **93**, 228701 (2004).

[22] P. Klimek, W. Bayer, and S. Thurner, Physica A (in print), (2011).

[23] V. Alfi, G. Parisi, and L. Pietronero, Nature Physics (London), **3**, 746 (2007).

[24] T. Mizuno, M. Takayasu, and H. Takayasu, Physica A **308**, 411 (2002).

[25] D. Sornette, H. Takayasu, and W.-X. Zhou, Physica A **325**, 492 (2003).

[26] M. A. Szybisz and L. Szybisz, Phys. Rev. E **80**, 026116 (2009).

[27] T. Preis, J, J. Schneider, and H. E. Stanley Proc. Natl. Acad. Sci. USA (Early edition) doi:10.1073/pnas.1019484108 (2011).

[28] The State of the Live Web, April 2007, http://www.sifry.com/alerts/archives/000493.html

[29] Ministry of Internal Affairs and Communications, Report of institute for information and communication policy (in Japanese) (2009).

[30] [http://gigazine.net/index.php?/news/comments/20080326 spam blog/] (in Japanese).

[31] Y. Kato, T. Utsuro, Y. Murakami, T. Fukuhara, H. Nakagawa, Y. Kawada, and N. Kondo, *in Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, (2008).

[32] M. Utiyama and H. Isahara, J. of the Japanese Society for Artificial Intelligence, **43**, 1 (in Japanese) (2002).

[33] T. Fukuhara, T. Utsuro, H. Nakagawa, and H. Takeda, *in Proceedings of The 21st Annual Conference of the Japanese Society for Artificial Intelligence*, (in Japanese) (2008).

[34] Y. Sato, T. Utsuro, T. Fukuhara, Y. Kawada, Y. Murakami, H. Nakagawa, and N. Kondo, *in Proceedings of The 22nd Annual Conference of the Japanese Society for Artificial Intelligence*, (in Japanese) (2008).

[35] T. Takeda, and A. Takasu, J. of Information Processing Society of Japan, **49**, 4 (in Japanese) (2008).

[36] T. Takeda, IPSJ (Information Processing Society of Japan) SIG (Spoken Language Processing) Technical Reports, **36**, 89 (in Japanese) (2009).

[37] About Google Trends, http://www.google.com/intl/en/trends/about.html

[38] Y. Sano, K. Kaski, and M. Takayasu, *in Proceedings of the 9th Asia-Pacific Complex Systems Conference*, (2009).

[39] H. Ebbinghaus, Dover, New York, (1963).

[40] M. A. de Menezes and A.-L. Barabàsi, Phys. Rev. Lett. **92**, 028701 (2004).

[41] S. Meloni, J. Gomez-Gardenes, V. Latora, and Y. Moreno, Phys. Rev. Lett. **100**, 208701 (2008).

[42] K. Yamada, Y. Sano, H. Takayasu, and M. Takayasu, (in preparation)