Empirical analysis of collective human behavior for extraordinary events in the blogosphere

Yukie Sano,[1,*] Kenta Yamada,[2] Hayafumi Watanabe,[3] Hideki Takayasu,[4,5] and Misako Takayasu[3]

[1]*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*

[4]*Sony Computer Science Laboratories, 3-14-13 Higashi-Gotanda, Shinagawa-ku, Tokyo 141-0022, Japan*

[5]*Meiji Institute for Advanced Study of Mathematical Sciences, 1-1-1 Higashimita, Tama-ku, Kawasaki 214-8571, Japan*

To uncover an underlying mechanism of collective human dynamics, we survey more than 1.8 billion blog entries and observe the statistical properties of word appearances. We focus on words that show dynamic growth and decay with a tendency to diverge on a certain day. After careful pretreatment and the use of a fitting method, we found power laws generally approximate the functional forms of growth and decay with various exponents values between -0.1 and -2.5. We also observe news words whose frequencies increase suddenly and decay following power laws. In order to explain these dynamics, we propose a simple model of posting blogs involving a keyword, and its validity is checked directly from the data. The model suggests that bloggers are not only responding to the latest number of blogs but also suffering deadline pressure from the divergence day. Our empirical results can be used for predicting the number of blogs in advance and for estimating the period to return to the normal fluctuation level.

I. INTRODUCTION

Collective behavior in human society has attracted considerable interest in the past decade [1–14], because developments in information technology have enabled the storage of large volumes of high-frequency human activity data, for instance, detecting bubbles in stock exchange activities [6], modeling dealer behavior using real data in foreign exchange markets [7], and empirical analysis of consumer behavior in supermarkets and convenience stores using purchase history and point-of-sale (POS) data [8,9]. Human activity data that is collected from the web, for example, YouTube videos and the social network service Facebook, are analyzed to not only explain basic individual human behavior but also elucidate hidden network structures in the society [10,11].

Here we also use the data from the Web to uncover nontrivial mechanism of collective human activities. Because word frequency on the Web is expected to immediately reflect the real social mood, it has attracted increasing attention among many academic and industrial researchers. In fact, it is stored electronically and analyzed widely. For example, the Library of Congress in the United States, which is the largest library in the world, has been archiving the entire public tweet of Twitter, a microblogging system since 2007 (http://blog.twitter.com/2010/04/tweet-preservation.html).

A blog is a type of Web site 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 was popularized around the year 2000 when free blog services began to be provided by Internet service companies. A *blogger*, who is an owner of a blog site, can easily upload his or her "entries" any time, and readers can easily post comments on the blog page. This interactive quality has contributed to the success of blogs; they are now widely used as basic social communication tools. The whole blog community is often called the *blogosphere*, and its scientific study is expected to be a promising new field of science, as huge amount of records are compiled as digital data.

In this study, we analyze the keyword appearance rate in blogs in which the functional forms of growth and decay around the peak are approximated by power-law functions of time. Similar power laws have been established in other fields of human activity. For example, a power law can describe a decrease in online book sales with an exponent that depends on endogenous or exogenous shocks [12]. Relaxation in audience number for online movies can also be described by power laws with various values of exponents that reflect the quality of the content [10]. Alfi *et al.* found that growth in conference registration numbers is also approximated by a power law diverging at the deadline [13].

In Sec. II, we describe the analyzed data and Japanese blogs. In Sec. III, we introduce our pretreatment procedures and peaked words. In Sec. IV, we focus on the time evolution of these peaked words and prove that they grow and decay with power laws. To reproduce power laws, we introduce a simple model of posting blogs in Sec. V. In Sec. VI, we discuss the predictability of our model from the standpoint of application, and the final section is devoted to conclusions.

II. DATA DESCRIPTION

The data analyzed in this study were obtained from the blogosphere written in Japanese over a period of 4 years, from November 1, 2006, to October 31, 2010. According to the technical report by the Internet search engine company
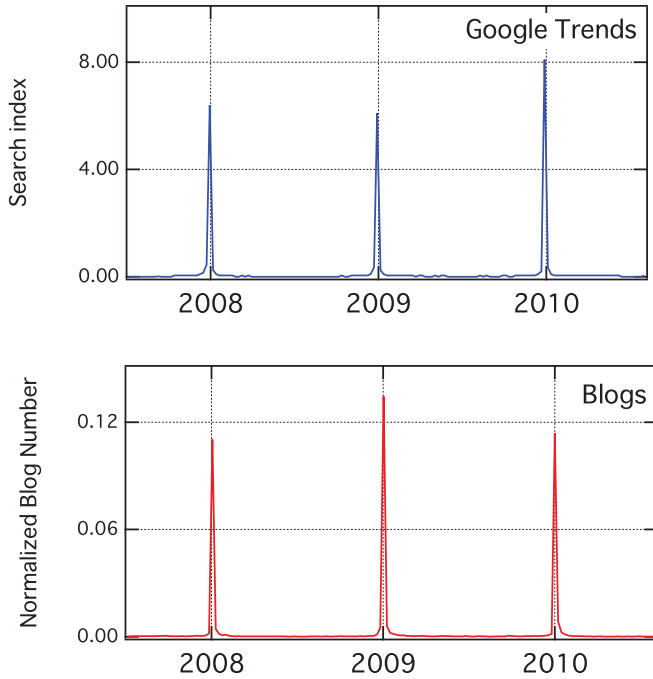
---

*sano.yukie@nihon-u.ac.jp

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

Technorati (http://technorati.com), which tracked more than 70 million blogs worldwide in 2007, the share of Japanese blogs is 37%, the largest among all languages. Although we analyze only the Japanese blogosphere, we show an example in which the dynamic properties in Japanese and English are considerably similar. Figure 1 shows the temporal change of the frequency of the English "April fool" observed by Google Trends (http://www.google.com/trends) and surveyed worldwide compared to the number of blog entries containing the corresponding Japanese. In both cases, we confirm that there is a clear peak on the week including April Fools' Day.

In blogosphere research, it is important to note the existence of spam blogs. They are automatically generated blogs in which the same words are repeated multiple times, mainly for the purpose of advertising. As the share of spams in the Japanese blogosphere is said to be 40%, it is important to exclude spams from the data. We used a new Internet service called "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 filtering and we apply the middle level, which is known to remove most of the spams while keeping most of the human blogs untouched. 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 major blog service providers. It covers more than 1.8 billion blog entries in 15 million blogs accounting for 90% of the Japanese blogosphere.

For analysis of Japanese we introduced a pretreatment to separate Japanese words that are not separated by spaces. Here we use the commonly used Japanese morphological analyzer "MeCab" (http://mecab.sourceforge.net/) to individually separate words according to a dictionary. By adding words to its dictionary, this software can treat multiword phrases such as *April Fool* as one word, *April-Fool*. Most of the words used in this study are already listed in the software's dictionary as one word, except names of people.

III. PEAKED WORDS

In the blogosphere, there are special words whose frequency grows or decays around a peak day such as *April Fool* with the peak on April 1. In the following discussion, we denote these words as "peaked words" and analyze their functional forms of growth and decay.

A. Pretreatment

We first apply the following pretreatment to the data to exclude both trivial circadian human activity patterns and systematic noise. In this subsection, we mainly focus on statistics of blogosphere itself, not peaked words.

Time shift: In the blogosphere, although a day starts at 00:00:00, 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 details of their activities time stamped in seconds. By counting the number of entries posted at every hour, a circadian activity pattern is plotted in Fig. 2. The solid line shows the 24-h activity pattern obtained directly from the data. However, we discovered there are a certain number of blog entries with time stamps that are exactly 00:00:00. We consider this time stamp to be caused by an artificial systematic spec or error, and we exclude this data from the statistics when capturing the circadian pattern. The red bars in Fig. 2 show the revised circadian activity pattern. Using a 24-h clock, we find that blogging activity is lowest around 4:00, and, thus, we consider the start of a day at 05:00 to be reasonable. Because the share of activity in the interval between 00:00 and 05:00 is
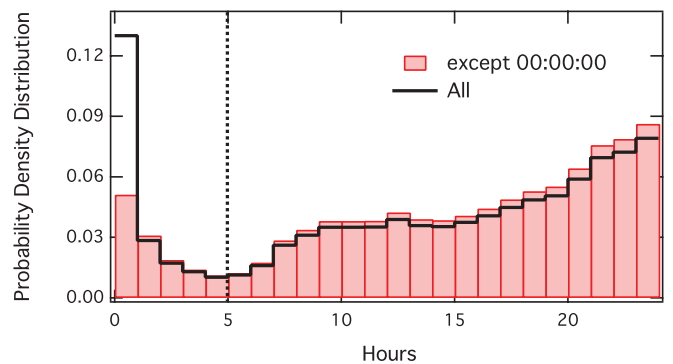

FIG. 2. (Color online) Probability density distribution of circadian activity of blogs posted by 10 000 bloggers. The solid line is calculated from all entries and the red bar is from the entries excepted that have a time stamp of 00:00:00. In both, 04:00 is the smallest ratio in a day.
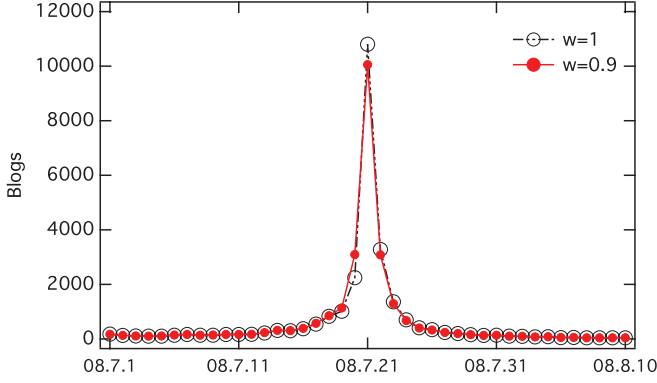
FIG. 3. (Color online) Typical example of time series of the peaked word *Marine Day* in 2008. $w = 1$ corresponds to no revision and $w = 0.9$ corresponds to modified time series introduced in Sec. III A. Because of the circadian effect, the data of the day after the peak is always higher than that before the peak without modification.

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:

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

where the weight is set as $w = 0.9$. With this modification, we can determine the time-shifted time series. In Fig. 3, open circles show the original daily data in which $w = 1.0$ in Eq. (1), and colored circles show time-shifted data in which $w = 0.9$. The time-shifted data 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 $x'(t)$. To clarify the effect of this time-shift procedure, we also show results without this time-shift procedure in Appendix A.

Normalization: There are nonuniform and nonstationary properties in the total number of entries per day [14]. For example, there was a sudden drop in February 2007 that was caused by search engine software's system maintenance. In order to reduce the systematic fluctuations caused by such nonuniform properties, we apply the following normalization procedure. There is already a method to separate internal and external noises [15], which simply deducts the external factor depending on its ratio of the total traffic. They assume that each traffic $x'_j(t)$ in a small component $j \in N$ consists of the total traffic $x'(t)$ without overlap, where $\sum_{j=1}^{N} x'_j(t) = x'(t)$. Here we simply divide $x'_j(t)$ by the total number, $x'(t)$. 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 $x'(t)$ that is averaged over the entire observation period. This normalized quantity is proportional to the probability that a blog contains the j-th word on the t-th day, and it is not necessarily an integer.

By introducing this normalization, the fluctuations caused by the aforementioned nonuniform properties can be reduced. In this study, we measure the word frequency using this normalization procedure.

B. Word selection

We determine candidates for peaked words in the following three categories.

Event: We selected the names of 14 public holidays and 16 major annual events in Japan. The appearance for these words grows and decays around the date of the event. In addition, these are words affiliated with an event, such as *Santa Claus* for *Christmas* and we can observe similar growth and decay behaviors for those words. However, in this analysis we neglected such affiliated words.

Date: We selected dates such as *May 9*, resulting in 365 words. There are many blog entries that announce some special day, e.g., birthdays and festivals. Growth and decay of these words always show a clear peak at the date.

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

IV. DYNAMICS OF PEAKED WORDS

We call the slopes before the peak day, *fore-slopes* and those after the peak day *after-slopes*, and we examine both in this section. As no standard method is known for checking the validity of approximation by a power-law time evolution for given time series, we apply a statistical test for power-law function introduced by Preis *et al.* [16] that is based on a Kolomogorov-Smirnov statistical test [17].

A. Method

We define the number of days in each slope by the number of consecutive days in which the word frequency is larger than the median value $\bar{x}_j$ from the peak. The median value is estimated throughout the entire observation period. We then approximate the functional form of the slopes using two models, a power law and an exponential law,

$$x_j(t) - \bar{x}_j = A_j|t_c - t|^{-\alpha_j}, \tag{2}$$

$$x_j(t) - \bar{x}_j = B_j \exp(-\beta_j|t_c - t|). \tag{3}$$

The parameters of these models, α_j, A_j, β_j, and B_j, are determined by use of the least-squares method. The fitting region is $[t_c \pm 1, t_c \pm n]$, where n is the number of days in slope. We then apply the Kolmogorov-Smirnov goodness-of-fit test to choose the better model. It was originally used as a statistical test for distributions. Here we apply it for evaluation of the statistical fitness of the functional form of the time series. For both models we calculate the KS statistic D, representing the deviation and defined as

$$D = \max_{t \in [t_c \pm 1, t_c \pm n]} \left| X_j^{(\text{empirical})}(t) - X_j^{(\text{model})}(t) \right|, \tag{4}$$

where $X_j^{(\text{empirical})}(t)$ is the cumulative number of the empirical value which is counted from the data and $X_j^{(\text{model})}(t)$ is the cumulative number which is calculated from the model. In both cases, numbers are normalized by $X_j(t_c \pm 1)$. By comparing the values of D for both models, the power-law model is accepted if the D value for the power law is smaller. In the

case that the power law is accepted, we check the validity of the model as introduced in Ref. [16]. First, we generated a data set of 1000 synthetic data points. One data set contains n data points. Synthetic data points are generated randomly following the normal distribution with the mean value best estimated from the model $x_j^{(model)}(t)$ and the standard deviation is $\sigma(x_j^{(model)}(t))$ as follows:

$$\sigma(\langle x_j \rangle) \simeq \sqrt{\langle x_j \rangle (1 + a^2 \langle x_j \rangle)}, \qquad (5)$$

where $a = \frac{\sqrt{\langle X^2 \rangle_c}}{\langle X \rangle} = 0.08$ is a constant parameter characterizing the fluctuation in the number of all bloggers which is determined independently of the word [see Appendix B for theoretical derivation of Eq. (5)]. For each synthetic time series, we compare its D value with that of the empirical one. We count the number of cases in which the D value for the synthetic time series is larger. If the number of such cases are less than 100 from the 1000 synthetic samples, we accept the power-law model as $q = 0.1$. Contrary to the ordinary sense of p value, the power-law hypothesis is considered to be valid for larger q. Thus, if the q is close to 1, then the difference between empirical data and the model can be attributed to statistical fluctuation alone and we accept power-law hypothesis. If the q is smaller than 0.1, we reject the power-law hypothesis. We change the border of the fitting region n from 5 days to a maximum slope length. The value of power exponent, α_j, is given by the value for the case with the largest n.

B. Results

Figure 4 is a typical result of data fitting for the word *Marine Day* in 2008 with a log-log scale, as shown in Fig. 3 with a linear scale. For all cases of power-law fitting, the distribution of the estimated power exponents are shown in Fig. 5 and summarized in Table I.

The absolute value of the power exponents of the after-slopes is larger than that of the fore-slopes in 58% of the 65 samples for Event and 80.6% of the 603 samples for Date. For Date, we confirm a significant difference between fore-slopes and after-slopes by use of a t test with $p < 2 \times 10^{-16}$ while it is rejected with $p = 0.80$ for Event. The number of days of the after-slopes is larger than that of the fore-slopes in 55% of the 65 samples for Event and 65.8% of the 603 samples for Date. For Date, we confirm significant difference between fore-slopes and after-slopes by use of the KS test with $p < 2 \times 10^{-16}$ while it is rejected with $p = 0.22$ for Event.

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. The absolute values of the exponent after the peak tend to be estimated as smaller for high impact news because of the effect of sequential broadcasts after the news. 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 that 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.
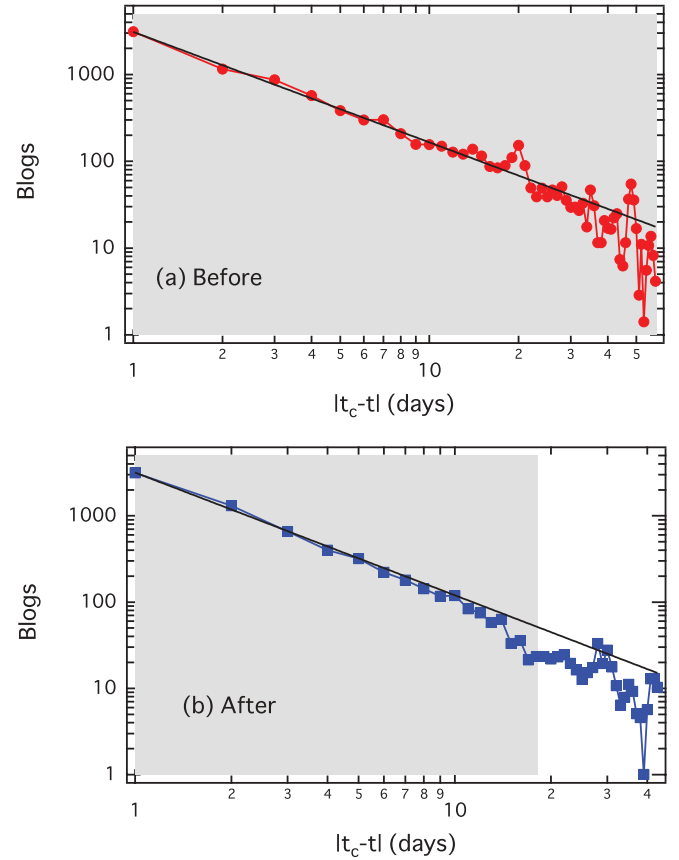

FIG. 4. (Color online) Examples of data fitting by power laws of *Marine Day* in 2008 for fore-slope (a) and after-slope (b) plotted in log-log scale. For fore-slope, models are fitted by Eq. (2) with $\alpha_j = 1.27$ and $A_j = 3100$ ($q = 0.239, n = 58$). For after-slope, $\alpha_j = 1.42$ and $A_j = 3171$ ($q = 0.108, n = 18$). The shaded area shows the interval n in which the power-law model is accepted.

C. An extreme case: *Tsunami*

The power-law decay per day of the word tsunami in the Japanese blogosphere is shown in Fig. 6(a). The peak day was March 12, 2011, the day after the quake, with 142 617 posts or 12.6% of all blog posts in raw data. After pretreatment with time shift and normalization, the estimated power exponent α_j is 0.67 with $A_j = 61788$ ($n = 50$) using Eq. (2). It is expected to take approximately 8623 days ($\sim$23.4 years) to return to the normal fluctuation level if we simply broaden the power-law function. The normal fluctuation level was 140 appearances per day, estimated from the data 1 month before the quake.

TABLE I. Mean values of power exponent α_j with standard deviations and medians of slope days n.

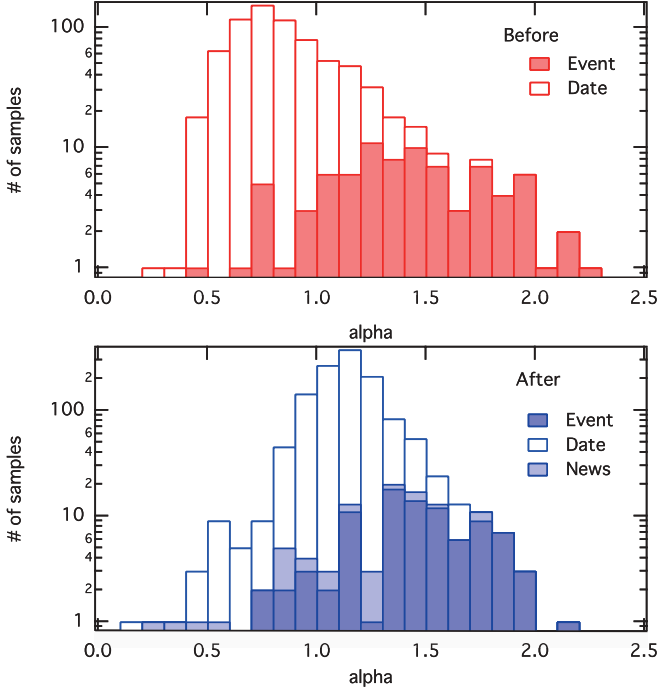|		α_j	n (days)	Samples
Event	Before	1.40 ± 0.38	10	83
	After	1.44 ± 0.28	16	91
Date	Before	0.79 ± 0.38	9	776
	After	1.11 ± 0.16	21	1229
News	After	1.09 ± 0.45	10	21
All	Before	0.85 ± 0.30	9	859
	After	1.13 ± 0.21	20	1341

FIG. 5. (Color online) Distribution of power exponent α_j of the fore-slopes (a) and the after-slopes (b). The mean value of α_j of the fore-slope is 0.85 ± 0.30 and that of the after-slope is 1.13 ± 0.21. Bars are colored by three categories of peaked words.

Although most of the news words decay in approximately 10 days, the case of *tsunami* is a rare exception because the number of entries is still 10 times higher than before the peak, even a year after the quake.

Twitter also shows a similar power-law behavior, even though the time resolution differs. Figure 6(b) shows the number of tweets measured per hour that include *tsunami*, calculated based on 1 397 783 tweets. We believe that this type of power law reflects the robustness of the empirically observed dynamics of collective human behavior.

V. THE MODEL

In this section, we propose a simple dynamic model to describe the typical power-law growth and decay of frequency of blogs with peaked words. There is already a simple model to describe people's universal behavior before a deadline by assuming pressure inversely proportional to the remaining time [13]. As this simple model can describe only the special case $\alpha = 1$, a kind of utility function that includes the tendency to postpone the action is introduced to describe the general case. Here, we introduce another approach to describe the general case. We introduce the following two assumptions for the number changes of blogs including the j-th target word at the t-th day, $\Delta x_j(t) = x_j(t+1) - x_j(t)$, increments for fore-slope and decrements for after-slope.

1. The pressure from the peak day t_c works inversely proportional to the time, $1/|t_c - t|$ [13].

2. The number of changes $\Delta x_j(t)$ is proportional to the number of blogs including the j-th target word, $x_j(t)$.
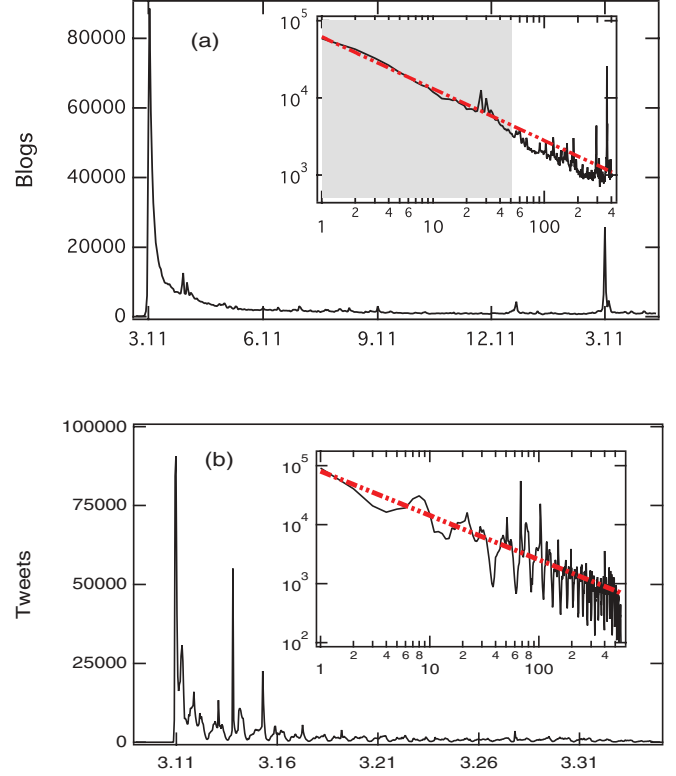

FIG. 6. (Color online) Decay of *tsunami* observed for blogs for 417 days since March 1, 2011, (a) and tweets for 648 h (27 days) since March 9, 2011 (b). Horizontal time step size is per day for blogs and per hour for tweets. Inset: Log-log plots of the time series. Red dashed lines show the slope of power law with the exponent $\alpha_j = 0.67$ for blogs and $\alpha_j = 0.75$ for tweets.

We can write these two assumptions into mathematical form in the continuous case as we assume $\Delta x_j(t) \simeq \frac{dx_j(t)}{dt}$. The time evolution of blogs for the fore-slope is given as

$$\frac{dx_j(t)}{dt} = \alpha_j^{(\text{fore})} \frac{x_j(t)}{(t_c - t)} + f(t), \qquad (6)$$

where $f(t)$ is an independent noise with a zero mean. The value $\alpha_j^{(\text{fore})} > 0$ is a proportionality factor that describes the effect of the above-mentioned two assumptions. Similarly, the decrement of the after-slope is given as

$$\frac{dx_j(t)}{dt} = -\alpha_j^{(\text{after})} \frac{x_j(t)}{(t - t_c)} + f(t), \qquad (7)$$

where $\alpha_j^{(\text{after})} > 0$ is also a proportionality factor that describes the effect of the two assumptions. Because we know that blogs decrease after t_c, we add a negative sign to Eq. (7). It is easy to confirm that both Eqs. (6) and (7) derive the power-law divergence, Eq. (2), in the case with no noise term $f(t)$. Thus, $x_j(t) \propto (t_c - t)^{-\alpha_j^{(\text{fore})}}$ for fore-slopes and $x_j(t) \propto (t - t_c)^{-\alpha_j^{(\text{after})}}$ for after-slopes. In the case that there is no pressure from the peak day t_c, blog dynamics follow Eq. (3) of the exponential law.

As a check of our assumption, we rewrite Eqs. (6) and (7) into the following form without the noise term $f(t)$, and we calculate the left-hand-side and right-hand-side values from the real data. Note that t_c is not necessarily an integer since
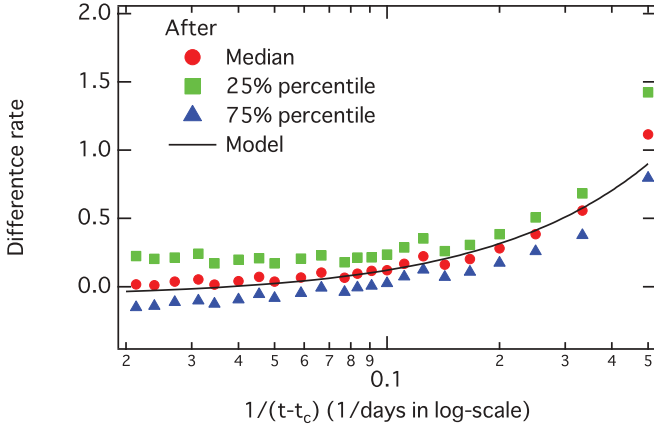
FIG. 7. (Color online) Relationship between difference rates $|\Delta x_j(t)|/x_j(t)$ and the absolute inverse number of time from the peak calculated by Eq. (8) for after-slopes, which summarized 1341 samples, shown in Table I. The solid line shows the theoretical relation, Eq. (8), with the averaged empirical value, $\alpha_j = 1.13$.

the divergence point is expected to exist in a single day time period,

$$\frac{|\Delta x_j(t)|}{x_j(t)} \simeq \alpha_j \frac{1}{|t - t_c|}. \qquad (8)$$

We surveyed all 1341 keywords for after-slopes as listed in Table I, and the median, upper, and lower quantile points are plotted in Fig. 7. As this figure shows, the median and quantile points fit well with the theoretical curve. This means that for the majority of words the relation Eq. (8) holds, implying that blog number changes are proportional to the number of recently written blogs, $x_j(t)$, and it is also promotional to $1/(t - t_c)$.

Now we know that the above relation [Eq. (8)] holds as a whole system; however, there remain two scenarios to realize this: The first case is where the main bloggers forming the peaked behavior are repeaters and the assumptions hold for each blogger individually, and the second case is where the main bloggers are newly joined bloggers and the assumptions hold for general bloggers, implying the existence of collective interaction in the blogosphere. In order to clarify which is the right scenario, we pay attention to randomly chosen 30 000 bloggers whose activities can be traced precisely. For all these bloggers we observe the days when they posted the typical keyword *Marine Day*. We count the total number of blogs including this keyword among these bloggers for each week as plotted in Fig. 8 (bottom); we also count the number of bloggers who posted the keyword for the first time, and the ratio of the number of new comers over the total number in the week is plotted in Fig. 8 (top). As known from this figure we confirm that the share of repeaters in the peaked behavior is generally less than half, namely the power-law behavior is formed mainly by newly joined bloggers. Similar results are confirmed also for some other typical keywords. This fact implies that the second scenario is correct and the factor α characterizes the strength of influence of written blogs to general bloggers representing the existence of interaction in the blogosphere.

For fore-slopes there is a natural reason of appearance of factor, $1/(t_c - t)$, in Eq. (6) explained by the deadline effect [13], that is, a blogger who plans to post the keyword *Marine*
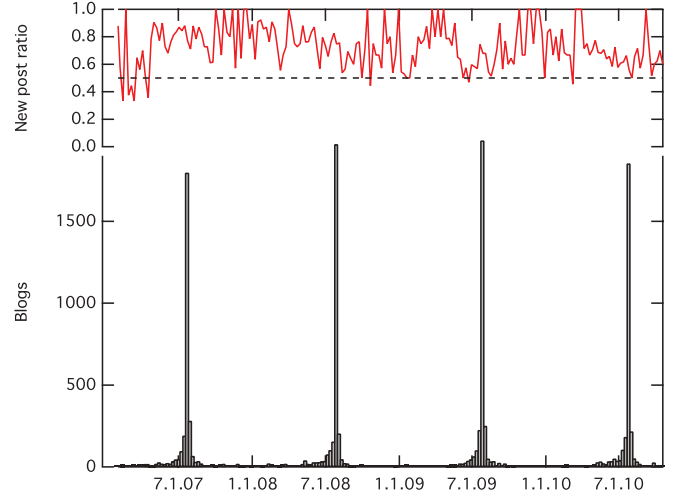

FIG. 8. (Color online) Number of blog entries including *Marine Day* per week which is calculated from randomly selected 30 000 bloggers (bottom) and the ratio of the number of newly posted bloggers in the same week (top). The dashed line shows the ratio 0.5.

Day may think that there are $t_c - t$ days before the deadline and a posting date can be chosen from $t_c - t$ candidates, so the probability of posting a blog on the day is proportional to $1/(t_c - t)$. This effect can be regarded as a universal property for each blogger individually.

On the other hand, the reason for after-slopes is less obvious. For a blogger who wants to post the keyword after the event, the probability of writing a blog including the keyword might be proportional to the decay of strength of memory. In the field of psychology, the functional form of memory decay is usually approximated by a nonlinear function [18], and here, as a simplest assumption, we introduce the inverse power law of memory decay from the deadline, $1/(t - t_c)$, which has the same functional form as the case of fore-slopes. With this assumption we can explain the nontrivial exponents of power-law behaviors of blogs by introducing the factor α that describes the strength of influence of written blogs to general bloggers.

VI. PREDICABILITY OF FREQUENCY

As an application of this study, we explore the possibility of estimating the word frequency in the near future. In Fig. 9, we show an example of the prediction of the blog frequency of *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 t_c. From the data, we find that the slope period starts on April 28, 85 days before t_c, 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. (2) is shown by the red line. In Fig. 9(b), the case of prediction for 5 days before the peak day is shown. The prediction error becomes smaller for shorter prediction period as expected. Note that a small difference in estimation of the exponent α_j makes a big difference near the peak; thus, the number of data points plays an important role in its accuracy.
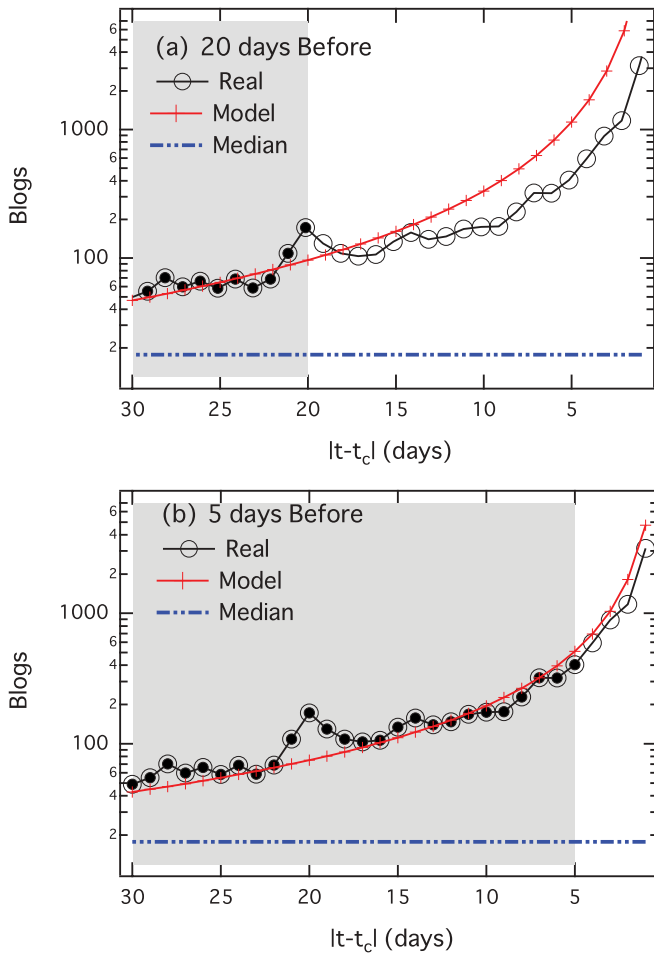
FIG. 9. (Color online) Typical examples of prediction for (a) 20 days and (b) 5 days before the peak day for *Marine Day* in 2008 in semilogarithmic scale. The red line indicates the prediction line, the blue dashed line indicates median $\bar{x}_j = 17$, and the solid line shows the real values. Open circles indicate the future values and colored circles are the known values used for prediction. The estimation is started 85 days before t_c. Estimated values are $\alpha_j = 1.78$ and $A_j = 20222$ for 20 days before the peak day and $\alpha_j = 1.38$ and $A_j = 4725$ for 5 days before t_c.

VII. CONCLUSIONS

By analyzing a large database of Japanese blogs, we showed that the functional forms of growth and decay of word appearance that peaked on a certain day are generally approximated by power laws with the various exponents values between -0.1 and -2.5. The values of the power exponents depend on the category of words such as Event, Date, and News. In the case of Event and Date, clarification of asymmetry in the power exponents of the fore-slope and after-slope is an interesting subject for future research on collective human behavior. In the case of News, the power law can be observed only after the peak, and its power exponent depends on its impact. In the case of significant news such as the March 11 earthquake in 2011, the absolute value of the power exponent is clearly smaller than 1.

We also checked the validity of our simple model that indicates that bloggers change their probability of posting proportional to the number of blogs and inversely proportional

to the time interval from the peak. By checking the data of bloggers' detailed activities, we confirmed that the peaked behavior mainly consists of newly posted bloggers. This implies that there exists a kind of global interaction between the newcomers and the keywords which makes the numbers of new comers and keywords proportional.

In addition, these power functions can be observed also in Twitter, and it suggests that these power-law behaviors are universal in social phenomena. An agent-based mathematical model will be used to reproduce these empirical properties of blogger activity in the near future [19].

APPENDIX A: THE CASE WITH NO TIME SHIFT

We show the results without the time shift; thus, Eq. (1) with $w = 1$. Figure 10 shows a typical example of data fitting
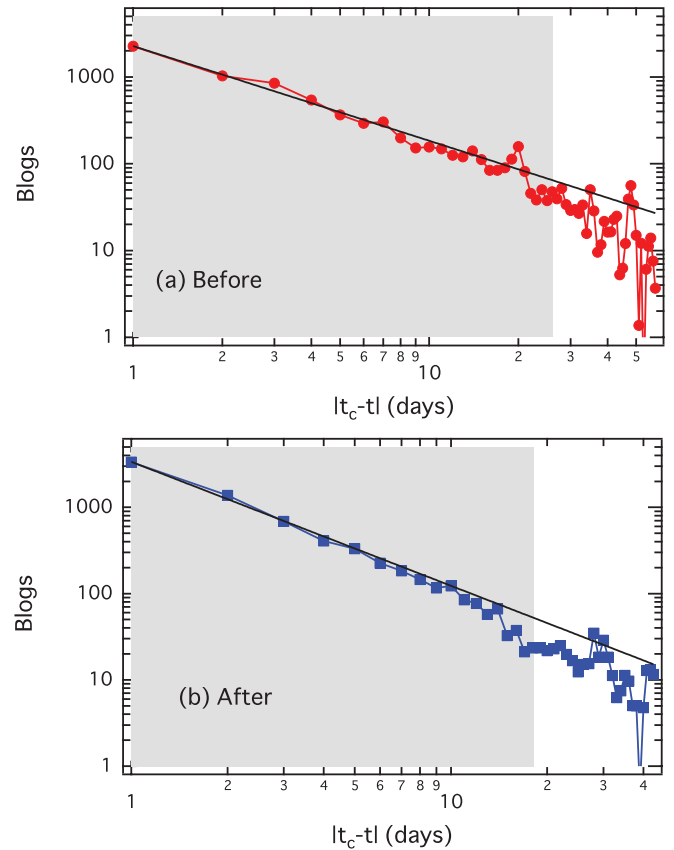

FIG. 10. (Color online) Examples of data fitting without time-shift pretreatment by power laws of *Marine Day* in 2008 for fore-slope (a) and after-slope (b) plotted in log-log scale. For fore-slope, models are fitted by Eq. (2) with $\alpha_j = 1.10$ and $A_j = 2273$ ($q = 0.103$, $n = 26$). For after-slope, $\alpha_j = 1.44$ and $A_j = 3369$ ($q = 0.104$, $n = 18$). The shaded area shows the interval n in which the power-law model is accepted.
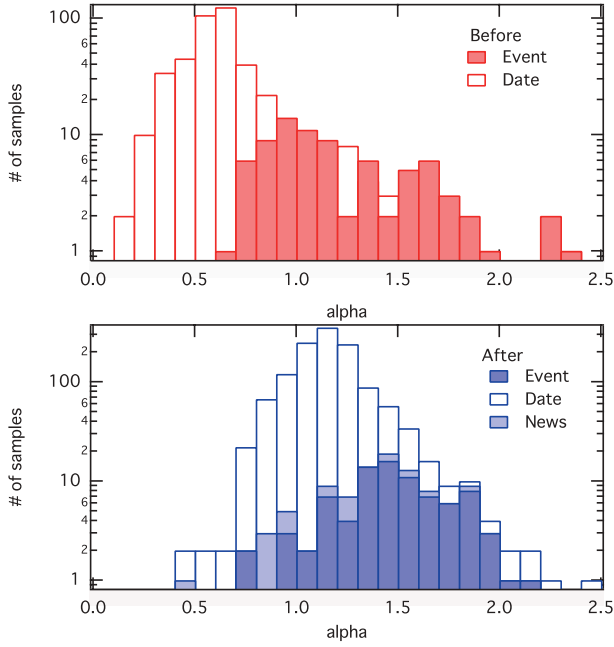
FIG. 11. (Color online) Distribution of power exponent α_j of fore-slopes (a) and after-slopes (b) without a time shift. Mean value of α_j of fore-slope is 0.73 ± 0.35 and after-slope is 1.16 ± 0.21. Bars are colored by three categories of peaked words.

without a time shift for the word *Marine Day*, as mentioned in Sec. IV B. There is no major change in power exponent α_j for fore-slope and after-slope. However, for the value of intercept A_j, we can find major deviation, especially for fore-slope ($A_j = 3171$ with a time shift and $A_j = 2273$ without a time shift). In Fig. 11 and Table II, we summarize the whole samples.

APPENDIX B: MODIFIED RANDOM DIFFUSION MODEL

We introduce a modified random diffusion model, which is used in Eq. (5). The random diffusion model was originally introduced to describe diffusion properties of random walkers on a given network [20,21], and two of the authors (Y.S. and M.T.) have modified the model to be applicable to the fluctuations in word appearance in the blogosphere [22]. In our modified random diffusion model, we assume that there are two states, active and nonactive, for each blogger, and the

TABLE II. Mean values of power exponent α_j with standard deviations and medians of slope days n in the case withno time-shift procedure, introduced in Sec. III A.

		α_j	n (days)	Samples
Event	Before	1.21 ± 0.38	15.5	80
	After	1.48 ± 0.28	16	85
Date	Before	0.64 ± 0.25	11.5	418
	After	1.14 ± 0.18	22	1176
News	After	1.21 ± 0.35	11	18
All	Before	0.73 ± 0.35	12	498
	After	1.16 ± 0.21	21	1279

number of active bloggers fluctuates randomly 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:

$$c_j = \frac{\langle x_j \rangle}{\langle X \rangle}, \tag{B1}$$

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

$$P(x_j | c_j) = \int_0^\infty \phi(X) \exp(-c_j X) \frac{(c_j X)^{x_j}}{x_j!} dX. \tag{B2}$$

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

$$P(x_j = 0 | c_j) \simeq \int_0^\infty \phi(X)(1 - c_j X) dX$$
$$\simeq 1 - c_j \langle X \rangle,$$
$$P(x_j = 1 | c_j) \simeq \int_0^\infty \phi(X)(c_j X) dX$$
$$\simeq c_j \langle X \rangle. \tag{B3}$$

For $\langle x_j \rangle \approx 2$, $P(x_j \geqslant 2 | c_j) \approx 0$, thereby $P(x_j | c_j)$ is approximated by the Poisson distribution with both the mean 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,

$$P(x_j | c_j) \simeq \int_0^\infty \phi(X) \frac{1}{\sqrt{2\pi c_j X}} \exp\left[-\frac{(x_j - c_j X)^2}{2 c_j X}\right] dX. \tag{B4}$$

By introducing a new variable, $y_j = \frac{x_j}{c_j \langle X \rangle}$, Eq. (B4) becomes

$$P(y_j | c_j) \simeq \int_0^\infty \phi(X) \frac{1}{\sqrt{2\pi \left(\frac{X}{c_j \langle X \rangle^2}\right)}} \exp\left[-\frac{\left(y_j - \frac{X}{\langle X \rangle}\right)^2}{2\left(\frac{X}{c_j \langle X \rangle^2}\right)}\right] dX. \tag{B5}$$

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

$$P(y_j | c_j) \simeq \int_0^\infty \phi(X) \delta\left(y_j - \frac{X}{\langle X \rangle}\right) dX. \tag{B6}$$

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

$$P(x_j | c_j) \simeq \frac{1}{c_j} \phi\left(\frac{x_j}{c_j}\right). \tag{B7}$$

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

$$\langle x_j \rangle = \int_0^\infty x_j P(x_j|c_j) dx_j$$

$$\simeq \int_0^\infty x_j \frac{1}{c_j} \phi\left(\frac{x_j}{c_j}\right) dx_j = c_j \langle X \rangle, \quad \text{(B8)}$$

$$\langle x_j^2 \rangle = \int_0^\infty x_j^2 P(x_j|c_j) dx_j$$

$$\simeq \int_0^\infty x_j^2 \frac{1}{c_j} \phi\left(\frac{x_j}{c_j}\right) dx_j = c_j^2 \langle X^2 \rangle. \quad \text{(B9)}$$

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

$$\sigma_j \simeq \sqrt{c_j^2(\langle X^2 \rangle - \langle X \rangle^2)}. \quad \text{(B10)}$$

By correlating both results Eqs. (B3) and (B10), we can get the following relation:

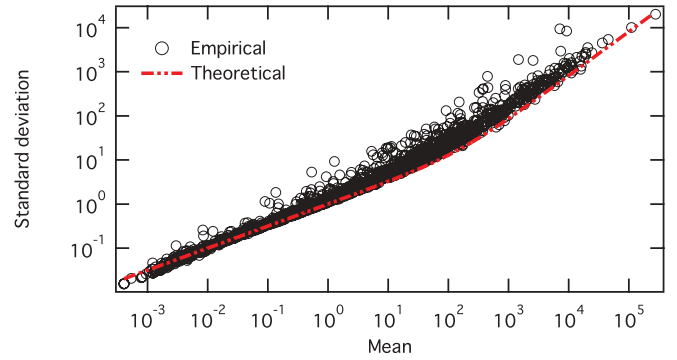$$\sigma_j \simeq \sqrt{c_j \langle X \rangle + c_j^2 \langle X^2 \rangle_c}, \quad \text{(B11)}$$

FIG. 12. (Color online) Relationship between mean and standard deviation of word frequency in the blogosphere. Empirical results of 1771 adjectives and theoretical result of Eq. (B12) are duplicated in the figure.

where $\langle X^2 \rangle_c$ denotes the second-order cumulant. By using $\langle x_j \rangle = c_j \langle X \rangle$, we rewrite Eq. (B11) into

$$\sigma_j \simeq \sqrt{\langle x_j \rangle + \left(1 + \frac{\langle X^2 \rangle_c}{\langle X \rangle^2} \langle x_j \rangle\right)}. \quad \text{(B12)}$$

Figure 12 shows empirical results using 1771 adjectives and Eq. (B12) with $\frac{\langle X^2 \rangle_c}{\langle X \rangle^2} = 0.08$.

[1] L. Conradt, Nature (London) **471**, 40 (2011).
[2] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási, Phys. Rev. E **73**, 066132 (2006).
[3] J. Ratkiewicz, S. Fortunato, A. Flammini, and F. Menczer, and A. Vespignani, Phys. Rev. Lett. **105**, 158701 (2010).
[4] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertéz, PloS One **7**, e38869 (2012).
[5] H.-H. Jo, M. Karsai, J. Kertész, and K. Kaski, New J. Phys. **14**, 013055 (2012).
[6] K. Watanabe, H. Takayasu, and M. Takayasu, Phys. Rev. E **80**, 056110 (2009).
[7] K. Yamada, H. Takayasu, T. Ito, and M. Takayasu, Phys. Rev. E **79**, 051120 (2009).
[8] H. Ueno, T. Watanabe, H. Takayasu, and M. Takayasu, Physica A **390**, 499 (2011).
[9] T. Mizuno, M. Toriyama, T. Terano, and M. Takayasu, Physica A **387**, 3931 (2008).
[10] R. Crane and D. Sornette, Proc. Natl. Acad. Sci. USA **105**, 15649 (2008).
[11] A. L. Traud, P. J. Mucha, and M. A. Porter, Physica A **391**, 4165 (2012).
[12] D. Sornette, F. Deschatres, T. Gilbert, and Y. Ageon, Phys. Rev. Lett. **93**, 228701 (2004).
[13] V. Alfi, G. Parisi, and L. Pietronero, Nat. Phys. (London) **3**, 746 (2007).
[14] Y. Sano and M. Takayasu, J. Econ. Interact. Coord. **5**, 221 (2010).
[15] M. A. de Menezes and A.-L. Barabási, Phys. Rev. Lett. **93**, 068701 (2004).
[16] T. Preis, J. J. Schneider, and H. E. Stanley, Proc. Natl. Acad. Sci. USA **108**, 7674 (2011).
[17] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[18] H. Ebbinghaus, *Memory: A Contribution to Experimental Psychology* (Dover, New York, 1987).
[19] K. Yamada, Y. Sano, H. Takayasu, and M. Takayasu, Systems, Control and Information **56**, 537 (2012) (in Japanese).
[20] M. A. de Menezes and A.-L. Barabási, Phys. Rev. Lett. **92**, 028701 (2004).
[21] S. Meloni, J. Gómez-Gardeñes, V. Latora, and Y. Moreno, Phys. Rev. Lett. **100**, 208701 (2008).
[22] Y. Sano, K. Kaski, and M. Takayasu, in *Proceedings of the 9th Asia-Pacific Complex Systems Conference (Complex'09)* (Chuo University, Tokyo, 2009), p. 195.