Research Article

Impact of Bursty Human Activity Patterns on the Popularity of Online Content

Qiang Yan[1] and Lianren Wu[1, 2]

[1] *School of Economics and Management, Beijing University of Posts and Telecommunication, Beijing 100876, China*
[2] *School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA*

Correspondence should be addressed to Lianren Wu, lianrenw@asu.edu

The dynamics of online content popularity has attracted more and more researches in recent years. In this paper, we provide a quantitative, temporal analysis about the dynamics of online content popularity in a massive system: Sina Microblog. We use time-stamped data to investigate the impact of bursty human comment patterns on the popularity of online microblog news. Statistical results indicate that the number of news and comments exhibits an exponential growth. The strength of forwarding and comment is characterized by bursts, displaying fat-tailed distribution. In order to characterize the dynamics of popularity, we explore the distribution of the time interval Δt between consecutive comment bursts and find that it also follows a power-law. Bursty patterns of human comment are responsible for the power-law decay of popularity. These results are well supported by both the theoretical analysis and empirical data.

1. Introduction

The advent of Web 2.0 and online social media [1] is fostering web-mediated brokers such as microblog and search engines, through which anyone can easily publish and promote content online. The dynamics of online content popularity has been deeply affected by the existence of these web-mediated brokers. The Web 2.0 and online social media not only change traditional communication processes with new types of phenomena, but also generate a huge amount of time-stamped data, making it possible for the first time to study the dynamics of online content popularity and human activity patterns at the global system scale.

Many human factors may affect the popularity of online content, which include human interests [2–4], social identity [5, 6], limited attention [7], and memory effects [8]. In this

paper, we focus on the impact of bursty patterns of human comment on the popularity of Sina Microblog (http://www.weibo.com/) [9]. Temporal heterogeneous and bursty are widely observed in many human-activated systems, which may result from both endogenous mechanisms like the highest-priority first protocol and exogenous factors like the seasonality and heterogeneity of human activities [10]. This phenomenon was found in various human activity patterns such as instant messaging [11, 12], web browsing [13, 14], E-mail and surface mail [15, 16], and mobile phone calls [17, 18]. If only regarding the timing of events, these human activity patterns are often described by a power-law distribution $P(\tau) \sim \tau^{-\alpha}$, where τ is the time interval between two consecutive activities. Regarding this phenomenon, many mechanisms of human dynamics have been proposed to explain the temporal bursts, such as task priority mechanism [15], memory effect mechanism [8, 19], human interaction mechanism [20], interest-driven mechanism [2, 3], and social identity mechanism [5, 6].

Here, the popularity of microblog news is defined as the number of comments per day posted for a piece of news. It is well documented that the statistical properties of the variable are very heterogeneous, with distribution following power-law. We explore the distribution of the time interval Δt between consecutive comment bursts and find it follows power-law. Furthermore, we prove that the exponent α, which characterizes the bursty patterns of human comment, is connected to that in the decay of popularity θ by the relation $\theta = \alpha - 1$.

The rest of this paper is organized as follows. Section 2 gives the data set. In Section 3, we study the dynamics of online news popularity. Section 4 introduces the relationship between bursty patterns of human activity and online content popularity. The power-law distribution is verified in Section 5. Finally, in Section 6 we conclude the work.

2. Data Set

The data of this research is collected from Sina Microblog (http://www.weibo.com/), which is one of the biggest microblogging platforms in China. We collected all news about a public topic, dated from August 20, 2009 to September 3, 2010, with the duration of 380 days. During this period, there are totally 125,150 pieces of news released, which have been forwarded 2,260,826 times and triggered 1,786,000 comments. For each piece of news, news ID, releasing time, times of forwarding, and number of comments were recorded. Therefore, we can track the dynamics of one specific piece of news through a unique ID.

From the statistical results, we find that during the data collecting time window, the number of news and comments exhibits an exponential growth (Figure 1). At the beginning of the observation time window, only a small amount of news and comments were posted. This is because Sina Microblog was just launched on August 14, 2009, and only a small group of people knew this application at that time.

The measurements also indicate that the number of forwarding and comments possess heterogeneity and burst. In our data set, among all the 125,150 pieces of news, 65772 pieces of news were forwarded and 69440 pieces of news were commented, respectively. As shown in Figure 2, the strength of forwarding S_{forward} (the number of forwarding for a piece of news) and the strength comment S_{comment} (the number of comments for a piece of news) follow power-law with the same exponent.
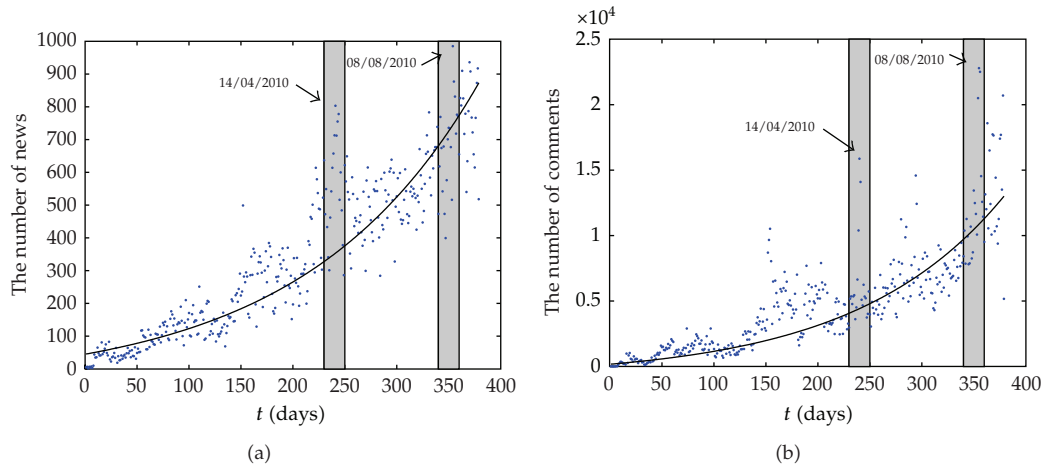
Figure 1: (a) The number of news and (b) the number of comments as the functions of time. In this figure, we find abnormal busty phenomena during some days marked by shadow. With further analysis of the data, we find that these phenomena may result from exogenous factors. For example, there are two major natural disasters happened on April 14, 2010 and August 7, 2010, which are Qinghai earthquake and Zhouqu mudslide, respectively.
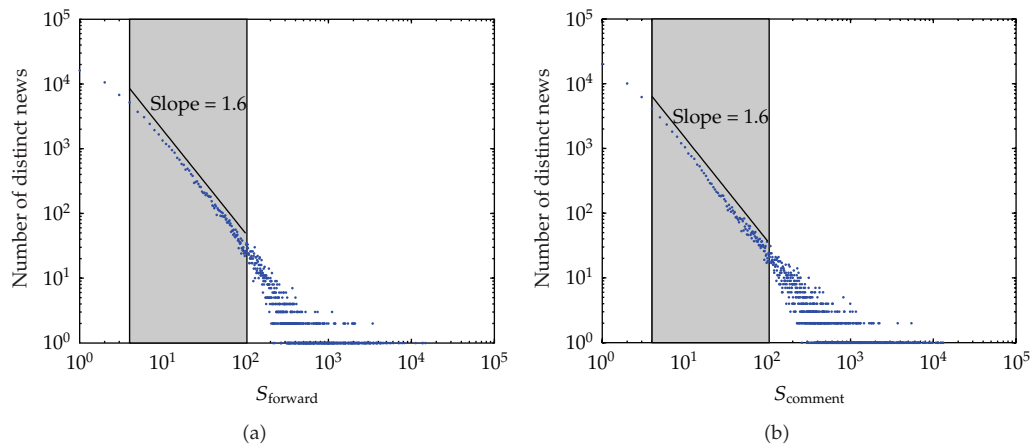

Figure 2: (a) The strength of forwarding $S_{forward}$. (b) The strength of comment $S_{comment}$. From the figures, we can find that most news received few forwarding and comments, but only a small part of news has a lot of forwarding and comment.

3. The Dynamics of Online Popularity

In order to quantitatively analyze the popularity of online microblog news, we consider the number of comments per day posted for a piece of news, expressed by $X(t)$ at time t. We study

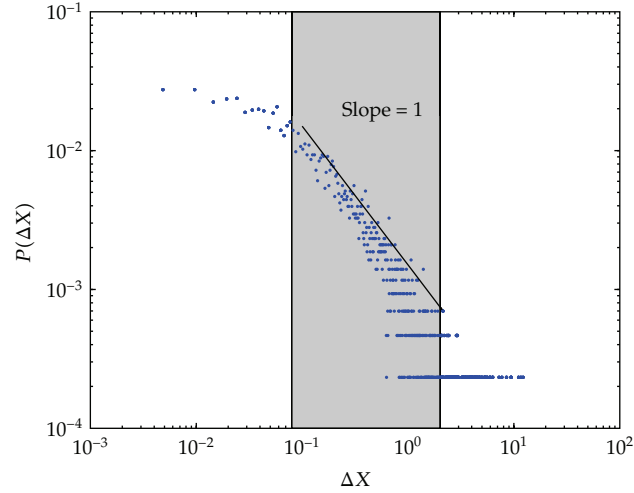$$[\Delta X]_t = \frac{[X(t) - X(t-1)]}{\overline{X}}, \tag{3.1}$$

Figure 3: Probability distribution of ΔX. We adopt the method of nonlinear least square to estimate the exponent. Slope = 1.07, SSE = 1 $\times$ 10^{-4}, R-square = 0.8999, and RMSE = 7.9 $\times$ 10^{-4}.

which represents the relative variation of the measurement in the time unit. Here, we use one day as the time unit, so $\overline{X}$ is the average value of comment strength. And $X(1)$ means the number of comments posted for a piece of news in the first day after the news released.

The relative variation of comment in the time unit is shown in Figure 3. Most news experienced a burst and received little attention thereafter. Since the relative variation may be negative, indicating a decrease in popularity, but our main concern is the positive values, so the events with negative variation are neglected.

Another way to characterize the dynamics of bursty systems is to study the distribution of time intervals between successive events. We analyzed the time distribution between consecutive comment bursts [21], namely, the time intervals between positive ΔX, shown in Figure 4. The intervals between bursts are distributed in a power-law.

We use maximum likelihood estimation (MLE) method [22] in conjunction with the Kolmogorov-Smirnov (KS) statistic table [23] to verify whether the fit is a good match to the data. In this case, the KS statistic suggests that the power-law curve is the better fit for the data, which will be explained in detail in Section 5.

4. Relationship between Bursty Human Activity Patterns and Popularity

In this section, we focus on the impact of bursty human activity patterns on the dynamics of popularity. First of all, we show that the number of comments for a piece of news $X(t)$ can be derived from the comment patterns of users.

Assume that a given piece of news is released at time t_0 and that all users can comment on it. The comment patterns are different from the browsing patterns [24]. Every user can comment on a piece of news more than once. In Figure 5, we show the comment patterns of one user, each vertical line represents a separate comment on the news. The thick line denotes the time when the user comments on the news for the first time after it was released at t_0. The release time of the news t_0 divides the time interval Δt into two consecutive comments
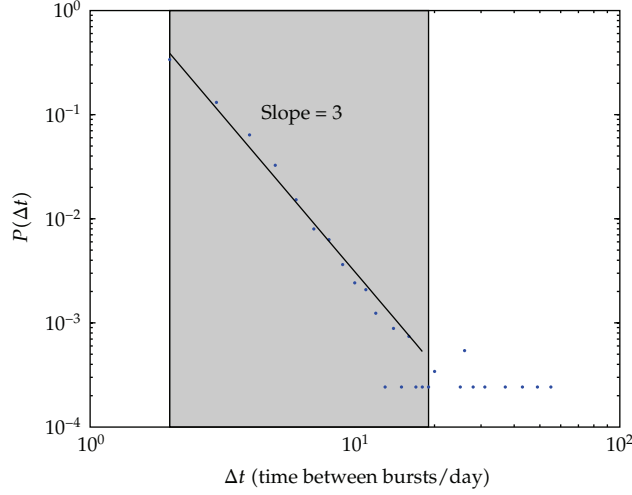
Figure 4: Distribution of popularity burst, which follows power-law with slope = 3.0.
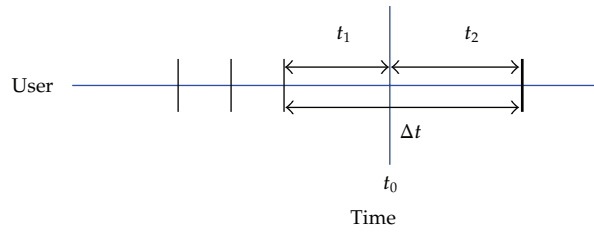

Figure 5: Comment patterns of user, each vertical line represents a separate comment on the news. The thick line denotes the time when the user comments on the news for the first time after it was released at t_0.

of length t_1 and t_2, where $t_1 + t_2 = \Delta t$. The probability that a user comments at time t_2 after the news was released is proportional to the number of possible Δt intervals. For a user characterized by a power-law intercomment time distribution with exponent α and a minimum time unit of t_m, the probability of finding an Δt interval having a length larger than t_2 is

$$p(\Delta t > t_2) = (\alpha - 1)t_m^{\alpha-1} \int_{t_2}^{\infty} (\Delta t)^{-\alpha} d\Delta t = \left(\frac{t_2}{t_m} \right)^{-\alpha+1}. \tag{4.1}$$

In (4.1) we assume that $\alpha > 1$.

For all users characterized by different exponents, the number of comment $X(t)$ can be calculated analytically as the average of (4.1) over the observed exponent values:

$$X(t) \sim \left\langle \left(\frac{t}{t_m} \right)^{-\alpha+1} \right\rangle_{\alpha}. \tag{4.2}$$
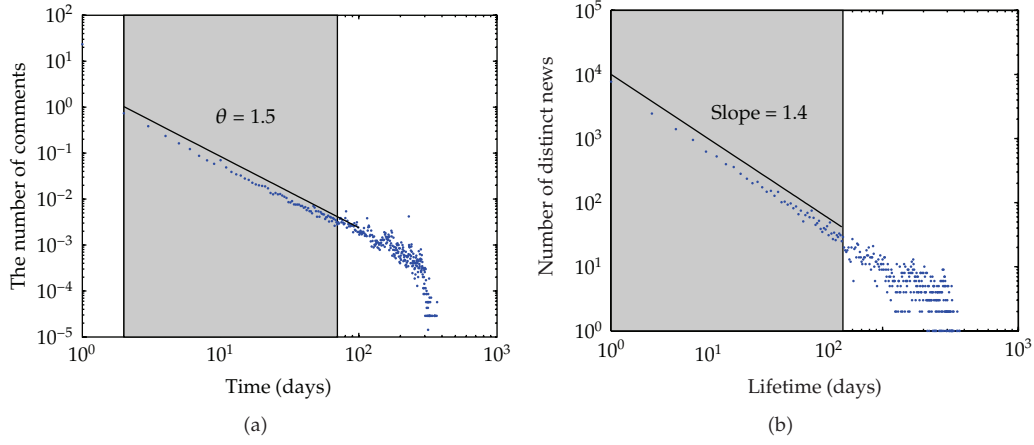
Figure 6: (a) The number of comments $X(t)$ for a piece of news as the function of time. (b) The distribution of news lifetime, which follows a power-law with slope = 1.4. All the results are the average over 69440 pieces of news, which have been commented on.

For simplicity, we assume that $t_m = 1$ and focus on the case that all users are characterized by the same exponent α. For example, the intercomment time distribution follows a power-law with exponent $\alpha = 2.5$ at the collective level (Figure 7).

Hence, (4.2) can be written as

$$X(t) \sim t^{-\alpha+1}. \tag{4.3}$$

Thus, we prove that the number of comments for a piece of news decay follows a power-law with the exponent $\alpha - 1$, namely, the decays of popularity follows a power-law with the exponent $\alpha - 1$.

In our data set, the number of comments $X(t)$ for a piece of news follows a power-law with exponent $\theta = 1.5$ (Figure 6(a)). More than 80% of comments take place within the first day, then decay to only 10% on the second day, and finally reach a small amount after five days. Meanwhile, we statistic the lifetime of all news (Figure 6(b)) and find the average lifetime of all news is 5.16 days. Distribution of interval between two consecutive comments at the collective level follows a power-law with exponent $\alpha = 2.5$ (Figure 7).

To sum up, we prove the fact that bursty human activity patterns deeply affected the popularity of news. Meanwhile, we conclude that the exponent α characterizing the bursty human activity patterns is connected to that in the decay of popularity θ by the relation:

$$\theta = \alpha - 1. \tag{4.4}$$

These results are supported by both the theoretical analysis and empirical data.
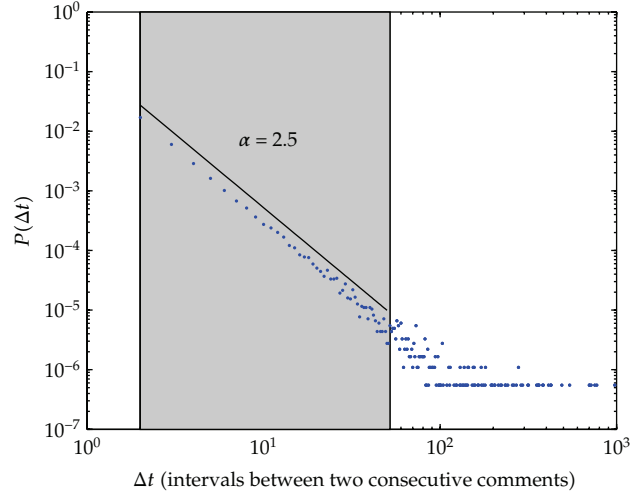
Figure 7: Distribution of intervals between two consecutive comments at the collective level, which follows power-law with exponent $\alpha = 2.5$.

5. Testing the Power-Law Hypothesis

Recent empirical observations suggested that power-law distributions occur in many natural and man-made systems. Unfortunately, most previous empirical studies of power-law distributed data have not attempted to test the power-law hypothesis quantitatively. Instead, they typically rely on qualitative appraisals of the data, for instance, based on visualizations.

In this section, we use a goodness-of-fit test to tell whether the fit is a good match to the data. First, we fit our empirical data to the power-law model using the methods of maximum likelihood estimation (MLE) and calculate the KS statistic for this fit [22]. Next, we use the KS table [23] obtaining good basis to confirm or reject the power-law distribution hypothesis.

Mathematically, a quantity x obeys a power-law if it is drawn from a probability distribution as follows:

$$p(x) \propto x^{-\alpha}, \tag{5.1}$$

where α is a constant parameter of the distribution known as the exponent or scaling parameter.

In the discrete case, power-law distribution, known as the zeta distribution [25], is expressed as

$$p(x) = \frac{\zeta(\alpha, x)}{\zeta(\alpha, x_{\min})}, \tag{5.2}$$

where $\zeta(\alpha)$ is the Riemann zeta function defined as $\sum_{k=1}^{\infty} k^{-\alpha}$.

Maximum likelihood estimation of the zeta distribution maximizes the log-likelihood function given by

$$l(\alpha \mid x) = \prod_{i=1}^{n} \frac{x_i^{-\alpha}}{\zeta(\alpha)},$$

$$L(\alpha \mid x) = \log l(\alpha \mid x)$$

$$= \sum_{i=1}^{n} (-\alpha \log(x_i) - \log(\zeta(\alpha)))$$

$$= -\alpha \sum_{i=1}^{n} \log(x_i) - n \log(\zeta(\alpha)),$$

(5.3)

where $l(\alpha \mid x)$ is the likelihood function of α, given the data $x = x_i \; 1 \le i \le n$, $L(\alpha \mid x)$ is the log-likelihood function.

When $x_{\min} > 1$, this maximum can be obtained theoretically for the zeta distribution by finding the zero of the derivative of the log-likelihood function [26]:

$$\frac{d}{d\alpha} L(\alpha \mid x) = -\sum_{i=1}^{n} \log(x_i) - n \frac{1}{\zeta(\alpha)} \frac{d}{d\alpha} \zeta(\alpha) = 0,$$

$$\frac{\zeta'(\alpha)}{\zeta(\alpha)} = -\frac{1}{n} \sum_{i=1}^{n} \log(x_i).$$

(5.4)

The most commonly used goodness-of-fit test is the Kolmogorov-Smirnov (KS) test [27], which is simply the maximum distance between the CDFs (cumulative distribution functions) of the data and the fitted model:

$$K = \max_{x \ge x_{\min}} |F(x) - P(x)|,$$

(5.5)

where $F(x)$ is the cumulative distribution function of the data, $P(x)$ is the cumulative distribution function for the power-law model that best fits the data in the region $x \ge x_{\min}$.

Firstly, we fit our empirical data to the power-law model using the MLE method and calculate the KS statistic for this fit. Secondly, we adopt the KS table, shown in Table 1, for obtaining a goodness-of-fit estimate [23]. Statistics were collected from the simulations to generate the KS quantiles. For each of the logarithmically spaced sample sizes, 10,000 power-law distributions were simulated, with random exponents from 1.5 to 4.0. Thirdly, we calculate the P value, namely, the fraction of the time that the resulting statistic is larger than the value for the empirical data. Conover [28] presented detailed instructions of how to use the KS table for obtaining a goodness-of-fit estimate.

According to the above theory and KS table, we obtain the goodness-of-fit estimates as shown in Table 2.

From Table 2, we can find that data sets described in our paper are consistent with a power-law distribution. For the data set described in Figure 3, however, most data is less than 1. Therefore, in this case we cannot use the method of maximum likelihood estimation.

Table 1: KS test table for power-law distribution. The table was created assuming MLE as the estimation method.

Sample number	Quantile			
	0.9	0.95	0.99	0.999
50	0.0826	0.0979	0.1281	0.1719
100	0.0580	0.0692	0.0922	0.1164
500	0.0258	0.0307	0.0412	0.0550
1000	0.0186	0.0216	0.0283	0.0358
2000	0.0129	0.0151	0.0197	0.0246
3000	0.0102	0.0118	0.0155	0.0202
4000	0.0102	0.0118	0.0155	0.0202
5000	0.0073	0.0086	0.0113	0.0147
10000	0.0059	0.0069	0.0089	0.0117
50000	0.0025	0.0034	0.0061	0.0077

Table 2: Basic parameters of the data sets described in our paper, along with their power-law fits and the corresponding P value.

Smaple number	n	$\langle x \rangle$	σ	$\zeta'(\alpha)/\zeta(\alpha)$	α	P value
Figure 2(a)	125 150	2147.03	2253.87	0.1936	1.617	0.95
Figure 2(b)	125 150	1696.11	2489.25	0.2061	1.621	0.95
Figure 4	4 138	478.05	752.74	0.3876	3.048	0.72
Figure 6(a)	69 440	0.24	1.55	0.1864	1.552	0.80
Figure 6(b)	69 440	966.94	1783.26	0.1573	1.395	0.90
Figure 7	1 786 000	3284.13	22829.74	0.2653	2.496	0.90

Here, we use the method of nonlinear least square and find that the data obeys a power-law distribution in certain areas.

6. Conclusion

The main goal of this paper is to explore the impact of bursty human activity patterns on the dynamics of popularity. Through theoretical analysis and empirical data, we prove that bursty human activity patterns are responsible for the power-law decay of popularity. This conclusion is consistent with previous studies [21, 24, 29–31]. Our statistical results also indicate that the number of news and comments exhibits an exponential growth. The strength of forwarding and comment is characterized by bursts, displaying fat-tailed distribution. In order to characterize the dynamics of bursty systems, we explore the distribution of the time interval Δt between consecutive comment bursts and find that it also follows a power law. Our measurements also indicate that microblog news have short lifetime. Most comments take place within the first day and the average lifetime of all news is 5.16 days. The average lifetime may vary for different social media, but the decay law of popularity is very likely generic, as they do not depend on the content, but are determined mainly by the human activity patterns. Indeed, the exponent α, which characterizes the bursty human activity patterns, is connected to that in the decay of popularity θ by the relation $\theta = \alpha - 1$.

Acknowledgments

References

[1] D. Tapscott and A. D. Williams, *Wikinomics: How Mass Collaboration Changes Everything*, Portfolio Hardcover, New York, NY, USA, 2006.

[2] Z. D. Zhao, H. Xia, M. S. Shang, and T. Zhou, "Empirical analysis on the human dynamics of a large-scale short message communication system," *Chinese Physics Letters*, vol. 28, no. 6, Article ID 068901, 4 pages, 2011.

[3] M. S. Shang, G. X. Chen, S. X. Dai, B. H. Wang, and T. Zhou, "Interest-driven model for human dynamics," *Chinese Physics Letters*, vol. 27, no. 4, Article ID 048701, 4 pages, 2010.

[4] X. P. Han, T. Zhou, and B. H. Wang, "Modeling human dynamics with adaptive interest," *New Journal of Physics*, vol. 10, Article ID 073010, 8 pages, 2008.

[5] Q. Yan, L. Wu, and L. Yi, "Research on the human dynamics in mobile communities based on social identity," *Discrete Dynamics in Nature and Society*, vol. 2012, Article ID 672756, 13 pages, 2012.

[6] Q. Yan, L. Yi, and L. Wu, "Human dynamic model co-driven by interest and social identity in the MicroBlog community," *Physica A*, vol. 391, no. 4, pp. 1540–1545, 2012.

[7] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition among memes in a world with limited attention," *Scientific Reports*, vol. 2, article 335, 2012.

[8] A. Vazquez, "Impact of memory on human dynamics," *Physica A*, vol. 373, pp. 747–752, 2007.

[9] http://www.weibo.com.

[10] T. Zhou, Z.-D. Zhao, Z. Yang, and C. Zhou, "Relative clock verifies endogenous bursts of human dynamics," *Europhysics Letters*, vol. 97, no. 1, article 18006, 2012.

[11] W. Hong, X. P. Han, T. Zhou, and B. H. Wang, "Heavy-tailed statistics in short-Message communication," *Chinese Physics Letters*, vol. 26, no. 2, Article ID 028902, 2009.

[12] Z.-D. Zhao and T. Zhou, "Empirical analysis of online human dynamics," *Physica A*, vol. 391, no. 11, pp. 3308–3315, 2012.

[13] B. Gonçalves and J. J. Ramasco, "Human dynamics revealed through Web analytics," *Physical Review E*, vol. 78, no. 2, Article ID 026123, 2008.

[14] F. Radicchi, "Human activity in the web," *Physical Review E*, vol. 80, no. 2, Article ID 026118, 2009.

[15] A. L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.

[16] J. G. Oliveira and A. L. Barabási, "Darwin and Einstein correspondence patterns," *Nature*, vol. 437, no. 7063, p. 1251, 2005.

[17] H.-H. Jo, M. Karsai, J. Kertesz, and K. Kaski, "Circadian pattern and burstiness in mobile phone communication," *New Journal of Physics*, vol. 14, Article ID 013055, 2012.

[18] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, "Uncovering individual and collective human dynamics from mobile phone records," *Journal of Physics A*, vol. 41, no. 22, Article ID 224015, 2008.

[19] A. Vázquez, J. G. Oliveira, Z. Dezsö, K. I. Goh, I. Kondor, and A. L. Barabási, "Modeling bursts and heavy tails in human dynamics," *Physical Review E*, vol. 73, no. 3, Article ID 036127, pp. 1–19, 2006.

[20] Y. Wu, C. Zhou, J. Xiao, J. Kurths, and H. J. Schellnhuber, "Evidence for a bimodal distribution in human communication," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 44, pp. 18803–18808, 2010.

[21] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Characterizing and modeling the dynamics of online popularity," *Physical Review Letters*, vol. 105, no. 15, Article ID 158701, 2010.

[22] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[23] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with fitting to the power-law distribution," *European Physical Journal B*, vol. 41, no. 2, pp. 255–258, 2004.

[24] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási, "Dynamics of information access on the web," *Physical Review E*, vol. 73, no. 6, Article ID 066132, 2006.

[25] N. L. Johnson, S. Kotz, and A. W. Kemp, *Univariate Discrete Distributions*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1992.

[26] H. Bauke, "Parameter estimation for power-law distributions by maximum likelihood methods," *European Physical Journal B*, vol. 58, no. 2, pp. 167–173, 2007.

[27] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 2nd edition, 1992.

[28] W. J. Conover, *Practical Nonparametric Statistics*, John Wiley & Sons, New York, NY, USA, 1999.

[29] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[30] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 41, pp. 15649–15653, 2008.

[31] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," *Science*, vol. 311, no. 5762, pp. 854–856, 2006.