Collective Online Clicking Pattern on BBS as Geometric Brown Motion

Zhenpeng Li[1(✉)] and Xijin Tang[2(✉)]

[1] Department of Applied Statistics, Dali University, Dali 671003, China
lizhenpeng@amss.ac.cn
[2] Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100190, China
xjtang@amss.ac.cn

Abstract. In this paper, we focus on massive clicking pattern on BBS. We find that the frequency of clicking volumes on BBS satisfies log-normal distribution, and both the lower-tail and upper-tail demonstrate power-law pattern. According to the empirical statistical results, we find the collective attention on BBS is subject to exponential law instead of inversely proportional to time as suggested for Twitter [4]. Furthermore we link the dynamical clicking pattern to Geometric Brown Motion (GBM), rigorously prove that GBM observed after an exponentially distributed attention time will exhibit power law. Our endeavors in this study provide rigorous proof that log-normal, Pareto distributions, power-law pattern are unified, most importantly this result suggests that dynamic collective online clicking pattern might be governed by Geometric Brown Motion, embodied through log-normal distribution, even caused by different collective attention mechanisms.

Keywords: Geometric Brown Motion · Log-normal distribution · Power-law · Collective behaviors over BBS

1 Introduction

Humans complex social behavior patterns are displayed through the cumulative effects of individual behaviors. One of the most common strategies in studying the social behaviors is to investigate and interpret whether any "pattern" is presented by fitting observed statistical regularities via data analysis. If the observed pattern can be described by a model characterized by related social psychological factors, that means we are close to the mechanisms that generate the collective regularity. As the main communication and information transmission tools in Web 1.0 era, bulletin board systems (BBS) and online communities were the main platforms for online activities in the whole Chinese cybersphere before 2005. BBS such as Tianya Forum expose digital traces of social discourse with an unprecedented degree of resolution of individual behaviors, and are characterized quantitatively through countless number of clicks, comments, replies and updates. Thanks to the different working functional designs, comparing with

H.T. Nguyen and V. Snasel (Eds.): CSoNet 2016, LNCS 9795, pp. 345–353, 2016.
DOI: 10.1007/978-3-319-42345-6_30

micro-blogging systems such as Twitter, long-time dynamics of human collective patterns on BBS are more stably showed out. Here we focus on massive clicking pattern on BBS. We analyze a large-scale record of Tianya Forum activity and find that the frequency of clicking volumes satisfies log-normal distribution, and both the lower-tail and upper-tail demonstrate power-law behavior. Furthermore we prove that the power-law behavior is caused by collective attention exponential decay. According to the empirical statistical results, we link the dynamical clicking pattern to Geometric Brown Motion (GBM), and rigorously provide a quantitative interpretation for the collective clicking phenomenon on BBS.

2 Data Source

Tianya Forum, as one of the most popular Internet forums in China, was founded on March, 1, 1999[1]. Till 2015, it was ranked by Alexa[2] as the 11th most visited site in the People's Republic of China and 60th overall. It provides BBS, blog, microblog and photo album services. With more than 85 million registered users, it covers more than 200 million users every month [1]. Tianya BBS, composed of many different boards, such as Tianya Zatan, entertainment gossip, emotional world, Media Jianghu, etc. is a leading focused online platform for important social events and highlights in China. We obtain the data by using automatic web mining tool - gooSeeker[3] and collect 22,760 posts from the Media Jianghu Board (MJB) of Tianya Forum during the replying time span from 13 June, 2003 to 16 September, 2015. The layout of MJB is shown in Fig. 1. Each post can be described by a 5-tuple: <title, author, clicking volumes, replying volumes, and replying time>. The 5-tuple dynamic is the feedback of user community behavior, and reflects collective online patterns. For example, posting represents that users release posts and want to be concerned, posting volumes reflect the active level of MJB, clicking means that visitors are interested in the posts or reflects the posts attraction level, while replying activities represent that users have intention to join the collective action compared with simple browsing (clicking), since replying behaviors indicate joiners have more in-deep thinking and enthusiasm towards the forum topics.

As for certain title (i.e. topic), the ratio between clicking volume and replying volume reflects the attention rate of the post and public participation degree. These cumulative micro individual behaviors (such as the number of posts, clicks and replies, the ratio between clicking volume and relying volume for each post) contribute to the global collective patterns, which could be measured by quantitative data analysis and modeling methods. Based on the above ideas, in this study, we take the replying and clicking volumes as the quantitative indexes to describe online group behaviors in the forum.

[1] http://bbs.tianya.cn/.

[2] Alexa Internet, Inc. is a California-based company that provides commercial web traffic data and analytics. https://en.wikipedia.org/wiki/Alexa-Internet.

[3] http://www.gooseeker.com/.

Fig. 1. Layout of Media Jianghu Board

3 The Distribution of Replies

Replying behaviors indicate visitors have more deep thinking and enthusiasm towards the forum topics. In order to study the pattern of replies, for 22,760 posts, we count each post replying volume. Around 34 % of the total posts, or 6,828 posts have no replies. After removing the no-reply records, we investigate the posts replying pattern. The statistics result is as shown in Fig. 2.

The inset in Fig. 2 suggests that the replies after taking logarithm follow exponential distribution, and log-log scale plot demonstrates power law pattern (take the logarithm for both replies and the corresponding number of posts). Next we fit the power-law distribution $f(x) \propto x^{\alpha}, x > x_{min}$.

We estimate the lower bound of the power-law behavior x_{min}, and scaling exponent based on the method described in [2]. We find that when $ln(\text{replying}) > 3.4340$, or replying volume > 31 (the estimate of the lower bound of the power-law behavior), the distribution of replies at MJB demonstrates power-law pattern, and maximum likelihood estimate of the scaling exponent $\alpha = -1.51$.

More replying activities represent the users have more active intention to join the collective action, meanwhile replying volumes show the topics' attraction or novelty levels, which means collective attention on MJB can be described by exponential distribution of replies. It is worth to note that as a function of time t, based on exponential form novelty decay, we will unify log-normal, Power law, and Pareto distribution by Geometric Brown Motion (GBM), and provide rigorous mathematic proofs in next section.

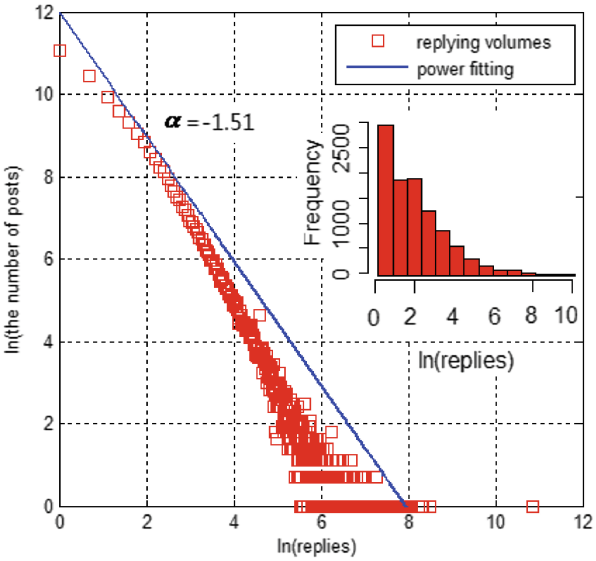

Fig. 2. The distribution of replies at MJB (the inset gives the actual histograms of replies after taking logarithm).

We use the replying number of a new post within 24 h as an index to measure collective attentions for the post. We randomly select 1000 samples from the total 22,760 posts, and average the number of replies in the first 24 h. We plot the average density distribution of replying volumes in Fig. 3. Different topics attract different users and show different attention characteristics. However the average results confirm that the collective attention is subject to exponential law.

According to the observation as shown in Fig. 3, we introduce d_t as a function of time t to account for the novelty decay [3], where d_t is defined as $d_t = \lambda e^{-\lambda t}$, $t > 0, \lambda > 0$. Here we set novelty decay d_t as exponential form instead of inversely proportional to time t in Ref. [4], where $d_t \propto 1/t$ is used to describe novelty decay of collective attentions on Twitter, as Twitter has the properties of instant arriving and fast transmission. On the contrary, usually with clearly defined title and no limitation of post length, BBS allow registered visitors to drop comments on the posts, thus generate interaction and discussion about the topics at hand. We estimate $\hat{\lambda} = 0.4888 (R^2 = 0.965)$ by nonlinear least squares method according to the selected 1000 samples. The empirical observation is not consistent with [4]. In contrast to BBS, we see Twitter has higher degree of attention as the new themes released, but the attention level declines more rapidly with the comparison as shown in Fig. 3(b). At any time of the first 24 h, Twitter users' attention level decay rate (the derivatives of the curves) is faster than that of BBS users. The characteristic is more dominant in the first 5 h, the slope $k_{Twitter}$ is obviously larger than k_{BBS} This result suggests that BBS and Twitter might have some different collective attention features also demonstrates the focusing

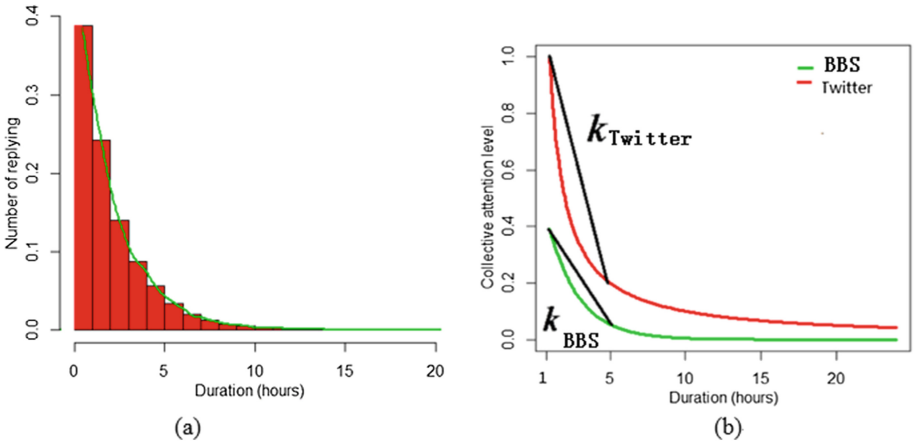

Fig. 3. Average density distribution of replying on 1000 samples in the first 24 h. (The curve line is the kernel density estimation)(Color figure Online)

patterns of behaviors emerged from the platforms between Web 1.0 and Web 2.0 are different.

4 The Distribution of Clicking Volume

We measure all the 22,760 posts on MJB with replying time span from 13 June, 2003 to 16 September, 2015. Replying and clicking time accurate to the second. We count C_t^q the clicking volumes for each post q on the Board at its corresponding replying time stamp t. The replying time stamp t is continuous, C_t^q describes the collective users' browsing pattern. At first we analyze all the 22,760 posts clicking volumes distribution in the given replying time span.

Figure 4(a) immediately suggests that the clicking volumes for the total $N = 22,760$ posts are distributed according to log-normal distribution. Since the horizontal axis is logarithmically rescaled, the histograms appear to be Gaussian function. A Kolmogorov-Smirnov normality test of ln(N) with mean 4.94826 and standard deviation 1.4427 yields a p-value of 0.0536 and testing statistic $D = 0.0895$, suggests that the frequency of clicking volumes follows a log-normal distribution. Since p-value is at the critical point of rejection region, we need to check normal distribution significance further with Quantile-Quantile (Q-Q) plots. If the random variable of the data is a linear transformation of normal variate, the points will line up on the straight lines shown in the plots. Consider Fig. 4(c), it is obvious that the empirical distributions are apparently more skewed than in the normal case. However, we observe that the (logarithmically rescaled) empirical distributions exhibit normality with the exception of the high and low end of the distributions. These tail outliers occur more frequently than could be expected for a normal distribution. We estimate $ln(N) = 4.4486$ by

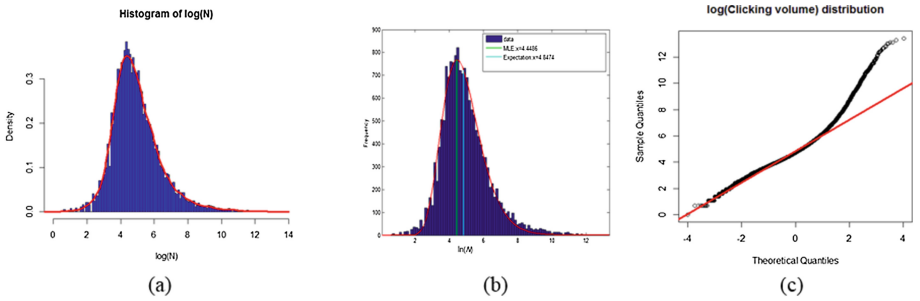

Fig. 4. Clicking volumes distribution on MJB (The solid line in the plots shows the density estimates using a kernel smoother)(Color figure online)

MLE method, e.g. the average clicking volume is about 86 times for each post, the result is as shown in Fig. 4(b).

About the tails distributions, we compute both lower tail (clicking volumes cumulative frequency below a given level) and upper tail (clicking volumes cumulative frequency above a given level) distributions. Figure 5 shows the cumulative frequency (in logarithmic scale) above (a) and below (b) a given level (in logarithmic scale), and demonstrates the upper-tail power-law behaviors, long recognized in the laws of Pareto and Zipf.

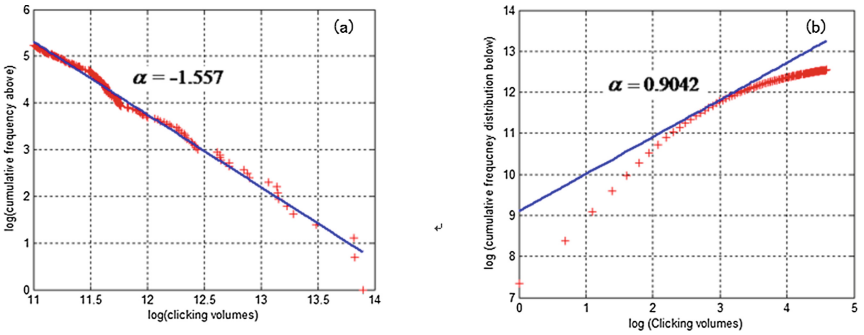

Fig. 5. Clicking volumes distribution on MJB (The "+" symbol refers to real data, and solid line in the plots is real data fitting line)

5 The Unified Stochastic Modeling on Users Clicking Pattern

As observed in Sect. 4, the frequencies of users clicking volumes satisfy log-normal distribution, in addition, both the lower-tail and upper-tail demonstrate power-law behaviors. From stochastic processes (Geometric Brown Motion) perspective,

this paper contains a quantitative interpretation for this collective phenomenon. Our endeavors will focus on mathematic rigorous proofs why log-normal, Pareto distributions, lower and upper tails power-law pattern are unified. With the analytic results of visitors' clicking records on MJB, our aim is to bridge the results of theoretical modeling and empirical data analysis. In statistics, the generalized Pareto distribution (GPD) is a family of continuous probability distributions. It is often used to model the tails of another distribution. The power law distribution and Pareto distribution (sometimes called Zipf's law) are unified based on the fact that Cumulative Distributions Function (CDF) of Probability Density Function (PDF) with a power-law form follows Pareto distribution (Zipf's law) [5].

Next we model visitors' clicking patterns per unit time as Geometric Brown Motion and prove that under the condition of visitors' clicking volumes $C_{r(t)}$ as the function of attention level r a random variable subject to exponential distribution, double tails power-law characteristic is obtained for visitors' clicking volumes evolving as GBM.

The temporal evolution of many phenomena exhibiting power-law characteristic is often considered to involve a varying but size independent proportional growth rate, which mathematically can be modelled by Geometric Brownian Motion (GBM). We set the clicking volumes fluctuation for all posts on MJB as a function of random variable that is subject to exponential distribution instead of directly as a function of fixed time stamp. The general explanations root in the fact that new topics will compete with old interesting ones, due to the limited attention of visitors or novelty decay of new topics, new posts can usurp the positions of earlier topics of interest, and soon older contents attentions are replaced by newer ones, but all these are random. We use the empirical result in Sect. 4 that the novelty decay or attention level is defined as $d_t = \lambda e^{-\lambda t}$. In other words, if we look d_t as a probability density function, then collective attention time can be seen as a random variable that satisfies exponential distribution. It seems more reasonable that the attention time to one post would be considered as a random variable, and hence, attention time is assumed as an exponential distributed random variable might be more accurate for browsing scenarios of BBS post. That is why we define stochastic fluctuation of clicking volume on the BBS post as a function of exponential distributed random variable T. According to the above analysis, firstly, we define stochastic fluctuation of clicking volumes on the forum as GBM

$$dC_T = \mu C_r dr + \sigma C_r dW_r \tag{1}$$

where W_r is the standard Wiener process with $W_0 = 0, W_t - W_s \sim N(0, t-s)$ (for $0 \leqslant s < t$) and $N(\mu, \sigma^2)$ denotes the normal distribution with expected value μ and variance σ^2. With the initial state C_{r_0} after some fixed time r, by using Ito integral to (1), we have

$$C_r = C_{r_0} e^{(\mu - \frac{\sigma^2}{2})r + \sigma W_r} \tag{2}$$

Taking logarithmic on both sides of Eq. (2), we have the following logarithmic form

$$log(C_r) = log(C_{r_0}) + (\mu - \frac{\sigma^2}{2})r + \sigma W_r \tag{3}$$

Equation (3) shows that given initial state C_{r_0} and fixed r since W_r is subject to normal distribution, $log(C_{r_0}) + (\mu - \frac{\sigma^2}{2})r$ is constant, $log(C_r)$ is subject to normal distribution, with $E(log(C_r)) = log(C_{r_0}) + (\mu - \frac{\sigma^2}{2})r$ and $var(log(C_r)) = \sigma^2 r$. Hence, we rigorously prove that C_r subject to log-normal distribution, but we could not confirm if it exhibits power-law behavior.

If we regard C_r as a function of an exponential distributed random variable instead of fixed r, we prove that GBM will exhibit power law characteristic as following. Without losing generality, for the computation simplicity, we set $C_{r_0} = 1, \sigma_2 = 1, \mu = \frac{1}{2}$, i.e. $log(C_r) \sim N(0, r)$. Since

$$f(C_r) = \int_0^\infty f(C_r, r)dr = \int_0^\infty f(C_r|r)f(r)dr, \tag{4}$$

then if we stop the process at an exponentially distributed time with mean $\frac{1}{\lambda}$, i.e. $f(r) = \lambda e^{-\lambda r}, r > 0$, the density function of C_r is

$$f(C_r) = \int_0^\infty f(C_r, r)dr = \int_0^\infty \lambda e^{-\lambda r} \frac{1}{\sqrt{2\pi r} C_r} e^{\frac{-(lnC_r)^2}{2r}} dr. \tag{5}$$

Using the substitution $r = u^2$, gives

$$f(C_r) = \frac{2\lambda}{\sqrt{2\pi} C_r} \int_0^\infty e^{-\lambda u^2 - \frac{(lnC_r)^2}{2u^2}} du. \tag{6}$$

we have the integral result for $C_r \geq 1$

$$f(C_r) = \frac{\lambda}{\sqrt{2\pi} C_r} \sqrt{\frac{\pi}{\lambda}} e^{-2\sqrt{\frac{\lambda (lnC_r)^2}{2}}} = \sqrt{\frac{\lambda}{2}} C_r^{-1-\sqrt{2\lambda}} \tag{7}$$

which is named Pareto distribution, and exhibits power-law behavior in both tails as observed in Fig. 5.

With this we end the proof. Interestingly, the result shows that clicking dynamics on the forum yields power law behavior. The above results also suggest a generic conclusion that although the GBM is used to generate log-normal distributions, only a small change from the lognormal generative process might yield a different distributed pattern.

6 Conclusions

To study the dynamics of collective attention in social media, in this paper we conduct a study on the cumulative micro individual behaviors, such as clicking volume and relying volume for each post on Media Jianghu Board of Tianya

Forum. Data analysis result shows that the frequency of clicking volumes follows a log-normal distribution. In order to explain the phenomenon, we use Geometric Brownian Motion to model the collective clicking fluctuation and the model is well matched with our empirical result. Moreover we rigorously prove that the emergence of users' collective clicking volumes double tails power-law pattern is caused by the collective attention exponential decay. This result suggests that dynamic collective online clicking pattern on BBS posts might be governed by Geometric Brown Motion, embodied through log-normal distribution, and rooted in collective attention exponential decay mechanism.

Acknowledgments. This research was supported by National Natural Science Foundation of China under Grant Nos. 71171187 and 61473284, 71462001, the Scientific Research Foundation of Yunnan Provincial Education Department under Grant No. 2015Y386, and No. 2014Z137, and the Open Project Program of State Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences.

References

1. Tianya Forum. http://help.tianya.cn/about/history/2011/06/02/166666.shtml
2. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Rev. **51**(4), 661–703 (2009)
3. Wu, F., Huberman, B.A.: Novelty and collective attention. Proc. Natl. Acad. Sci. U.S.A. **104**(45), 17599–17601 (2007)
4. Asur, S., et al.: Trends in social media: persistence and decay. SSRN 1755748 (2011)
5. A Ranking Tutorial. http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html