CHIN. PHYS. LETT. Vol. 34, No. 6 (2017) 068901

Modeling Information Popularity Dynamics via Branching Process on Micro-Blog Network *

Jin-Jie Li(李瑾颉)[1,2], Lian-Ren Wu(吴联仁)[3**], Jia-Yin Qi(齐佳音)[3], Qi-Ming Sun(孙启明)[1]

[1]*School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876*
[2]*Interdisciplinary Center for Network Science and Applications, University of Notre Dame, IN 46556, USA*
[3]*School of Management, Shanghai University of International Business and Economics, Shanghai 201620*

(Received 20 February 2017)

Predicting and modeling of items popularity on web 2.0 have attracted great attention of many scholars. From the perspective of information competition, we propose a probabilistic model using the branching process to characterize the process in which micro-blogging gains its popularity. The model is analytically tractable and can reproduce several characteristics of empirical micro-blogging data on Sina micro-blog, the most popular micro-blogging system in China. We find that the information competition on micro-blog network leads to the decay of information popularity obeying power law distribution with exponent about 1.5, and the value is similar to the exponent of degree distribution of micro-blog network. Furthermore, the mean popularity is decided by the probability of innovating a new message. Our work presents evidence supporting the idea that two distinct factors affect information popularity: information competition and social network structure.

PACS: 89.75.Da, 05.10.−a, 29.85.−c DOI: 10.1088/0256-307X/34/6/068901

The wide adoption of online social networks has increased the competition among information for our limited attention. Every day, we receive numerous messages from various online social networks. However, we do not have enough time and attention to disseminate each message that we received. It is an interesting question whether such competition may affect the items popularity. Despite the progress that has been made,[1−4] the empirical study of information popularity is still in its infancy. Studies in this direction have been mostly hindered by the shortcoming of available large-scale data. However, the availability of large-scale data from online social networks has recently created unprecedented opportunities to explore the impact of human behavior and characteristics on the information popularity.

The difficulty of estimating information popularity lies in the co-existence of various confounding factors.[5−9] Recently, some scholars have put forward different opinions on this issue. Cetin and Bingol suggested that individual attention has an important influence on the popularity of information, and proposed that decay of visibility combined with divided attention explain why most information cascades in social media fail to reach epidemic proportions.[10] Weng *et al.* employed a parsimonious agent-based model to study whether competition among memes for our finite attention affect the popularity of different memes, and the results suggested the massive heterogeneity in the popularity and persistence of memes as deriving from a combination of the competition for our limited attention and the structure of the social network.[11] Gleeson *et al.* suggested that competition between messages for the limited resource of user attention constrains the popularity of messages,[12,13] leads naturally to some messages becoming extremely popular, and others are only moderately popular, or are ignored.

Although some research works have been carried out, how limited attention affects the dynamics process of information popularity in the social network is still unclear. To address this problem, we analyze the real data from Sina micro-blog, and the results reveal that the information popularity decay obeys scaling laws. Secondly, we propose a probabilistic model based on the branching process to characterize the process by which micro-blogging gains its popularity. Thirdly, by the simulation of the proposed model, we find that the model can reproduce several characteristics of real data. Furthermore, the exponent of information popularity distribution is related to the exponent of degree distribution of the micro-blog network.

Our model is as follows: in online communication platforms like the micro-blogging system (e.g. Twitter, Sina Weibo), users follow other users. Each user has a 'screen' that records all messages received by the user, time stamped by their arrival time. In fact, only a fraction of the messages received are deemed 'interesting' by the user, and only the interesting messages are considered for possible retweeting by that user. The popularity $n(t)$ of a message is the total number of times it has been retweeted by age t, i.e., by a time after its first appearance (when it was tweeted as an innovation). Here $q_n(t)$ is the probability that a message has popularity $n(t)$ at the age t.

Firstly, we assume that each screen has capacity

*Supported by the National Natural Science Foundation of China under Grant Nos 71601005 and 71231002, and the National Basic Research Program of China under Grant No 2013CB329604.
**Corresponding author. Email: lianrenwu@suibe.edu.cn
© 2017 Chinese Physical Society and IOP Publishing Ltd

for only one message (see Fig. 1). At each time step Δt, one node is chosen at random. With probability μ, the selected node invents a new message, it appears on its screen, and is tweeted to all the node's followers. Otherwise, with probability $(1-\mu)$ the selected node retweets the message currently on its screen to all its followers, and the screen is unchanged. If there is no message on the node's screen, nothing happens. When a message M is retweeted, the popularity of message M is increased by 1.

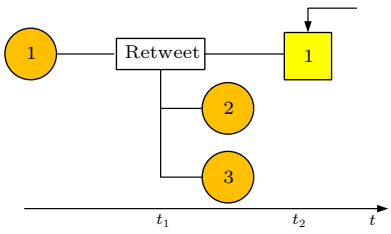

Fig. 1. Schematic diagram of the model. At time t_1, node 1 retweets the circular message to its followers (nodes 2 and 3). At time t_2, node-1 screen is overwritten by the square message, which was tweeted by one of the nodes followed by node 1.

Secondly, we introduce a branching process to describe the model, where the competitive environment causes each message to follow a critical branching process.[14,15] Here $H(t,x)$ is the probability generating function (PGF) of the popularity distribution, with

$$H(t;x)=\sum_{n=1}^{\infty}q_n(t)x^n. \tag{1}$$

We define $G_{jk}(t,\Omega;x)$ as the probability generating function of the popularity distribution for a message seeded by (i.e., first tweeted by) a node with out-degree k and in-degree j, as observed at time Ω. Meanwhile, we introduced $G(t,\Omega;x)$ as the probability generating function (PGF) of the excess popularity distribution. Using the out-degree distribution p_k of the network, we have the relation

$$G(t,\Omega;x)=\sum_k p_k G_{jk}(t,\Omega;x). \tag{2}$$

At $t=\tau$, there is random selection of a node, and we call the node screen S. At the next time step $t=\tau+\Delta t$, there are four possible outcomes for the particular screen S that contributes to $G_{jk}(t,\Omega;x)$, (a) the screen S is overwritten by some other message that is tweeted by another node; (b) during the period of Δt, with the probability μ innovates a new message; (c) during the period of Δt, with the probability $1-\mu$ retweets the message; and (d) keep the message, with none of the other outcomes (a)–(c) occurring. The results are listed in Table 1.

Table 1. Summary of the single-time-step outcomes that contribute to the PGF G_{jk}.

Outcome for screen S	Probability	Contribution to $G_{jk}(t,\Omega;x)$
(a)	$k\Delta t$	1
(b)	$\mu\Delta t$	1
(c)	$(1-\mu)\Delta t$	$xG_{jk}(t-\Delta t,\Omega;x)\times[G(t-\Delta t,\Omega;x)]^k$
(d)	$1-(k+1)\Delta t$	$G_{jk}(t-\Delta t,\Omega;x)$

Putting all four outcomes together, we obtain the expression of $G_{jk}(t,\Omega;x)$ as follows:

$$\begin{aligned}G_{jk}(t,\Omega;x)&=k\Delta t+\mu\Delta t\\&+(1-\mu)\Delta t x G_{jk}(t-\Delta t,\Omega;x)[G(t-\Delta t,\Omega;x)]^k\\&+1-(k+1)\Delta t G_{jk}(t-\Delta t,\Omega;x),\end{aligned} \tag{3}$$

and taking the limit $\lim\Delta t\to 0$ yields an ordinary differential equation for $G_{jk}(t,\Omega;x)$, parameterized by x,

$$\frac{\partial G_{jk}}{\partial t}=k+\mu-(k+1)G_{jk}+(1-\mu)xG_{jk}[G]^k. \tag{4}$$

It is straightforward to generalize the above derivation to allow each user's screen to have capacity c, meaning that the screen can simultaneously hold c messages. We then consider each user to be composed of c distinct messages, and suppose that the message may be deemed interesting and retweeted by followers with probability λ, independently. By Eq. (4) and Ref. [12], we obtain the mean popularity $m(t)$ of age-t messages,

$$\begin{aligned}m(t)&=\sum_{n=1}^{\infty}nq_n(t),\\c\frac{dm(t)}{dt}&=(\lambda k+1)(1-\mu m).\end{aligned} \tag{5}$$

With $m(0)=1$, the solution is

$$m(t)=\begin{cases}1+\frac{(\lambda k+1)}{c}t, & \mu=0,\\ \frac{1}{\mu}-\frac{1-\mu}{\mu}e^{-\frac{\mu(\lambda k+1)}{c}t}, & \mu>0.\end{cases} \tag{6}$$

If the out-degree distribution with power-law $p_k\propto k^{-\gamma}$ for large k with $1<k<3$, in this case we find that as $n\to\infty$,

$$q_n\sim\begin{cases}Bn^{-\frac{\gamma}{\gamma-1}}, & \mu=0,\\ Cn^{-\gamma}, & \mu>0.\end{cases} \tag{7}$$

With prefactors $B=-(\lambda k+1)\frac{(D\Gamma(1-\gamma))^{-\frac{1}{\gamma-1}}}{\lambda\Gamma(\frac{1}{1-\gamma})}$ and $C=\frac{\lambda z+1}{\lambda}(\frac{\lambda(1-\mu)}{\mu(\lambda z+1)})$.

From Eq. (7), we find that in the zero-innovation limit, the popularity distribution has a power-law exponent $\gamma/(\gamma-1)$ that is smaller than the exponent γ of the out-degree distribution.

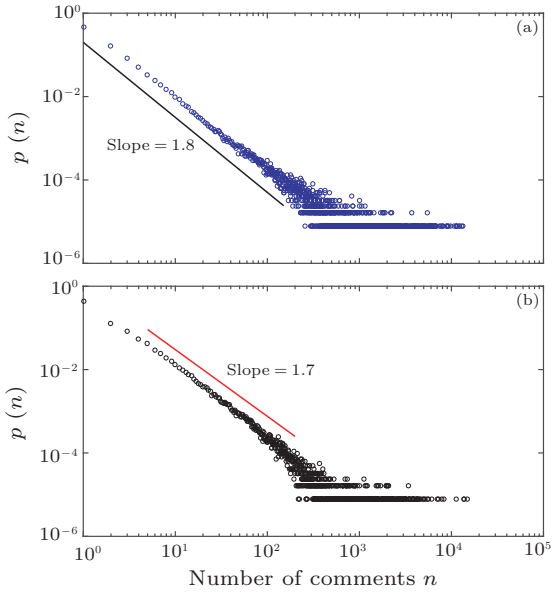

Fig. 2. The distribution of information popularity: (a) the number of retweets follows the power-law distribution with an exponent 1.8, and (b) the number of comments follows the power-law distribution with an exponent 1.7.

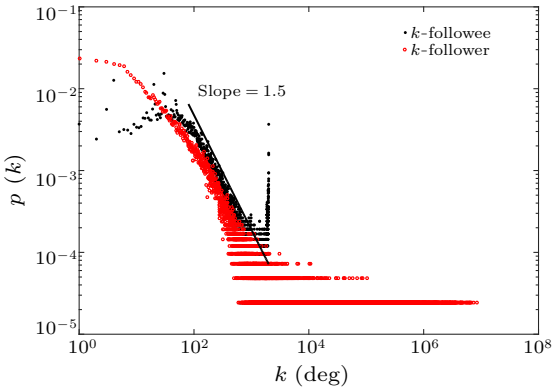

Fig. 3. The distributions of followee and follower.

Two data sets are included in this work, and collected from Sina micro-blog (www.weibo.com) by its open platform API. Sina micro-blog is one of the most popular micro-blogging platforms in China at present. The first data set includes 125 and 139 messages from 174 users during 2009/8/20 to 2010/9/3, collected by snowball sampling. These messages have been retweeted 2, 260, 826 times and commented 1, 822, 450 times. For each message, the ID of the publisher, the ID of the message, release time, the content of message, times of forwarding and number of comments were recorded. We also obtained the time when the message was commented on. As the micro-blogging system does not support this data, message forwarding time cannot be obtained. However, retweet behavior and comment behavior have the same characteristics, as shown in Fig. 2, which present a power-law distribution similar to Chinese income[16] and terrorist attacks[17] with the exponents of 1.8 and 1.7, respectively. Therefore, the popularity of comments can represent information popularity. The second data set includes 41667 users. For each user, user ID, number of followees (out-degree) and number of followers (in-degree) were recorded. The degree distributions (shown in Fig. 3) follow the power-law with the slope equals to 1.5 (namely, $\gamma = 1.5$).

From Fig. 3, we obtain that the number of followers follows a power law distribution with an exponent of 1.5, and has obvious fat tail characteristic. The number of followees also follows a power law distribution with an exponent of 1.5, but the tail up. This is mainly due to the fact that the upper limit of the followee of the Sina micro-blog system is 2000. To confirm the accuracy of the branching-process approximation and to explore the interactions of the network structure and the limited attention, here we compare numerical simulations of the model with real data and the theoretical predictions. We generate the configuration-model directed networks with prescribed out-degree distribution $p_k \sim k^{\gamma}$, and $\gamma = 1.5$. Model parameters are $N = 10^5$ nodes, and

$$\mu = \frac{\text{number of distinct messages used in the first data set}}{\text{total number of messages retweeted by users}} = \frac{125139}{2260826} \approx 0.055, \tag{8}$$

$$\lambda = \frac{\text{number of messages retweeted more than once}}{\text{number of distinct messages used in the dataset}} = \frac{59370}{125139} \approx 0.474, \tag{9}$$

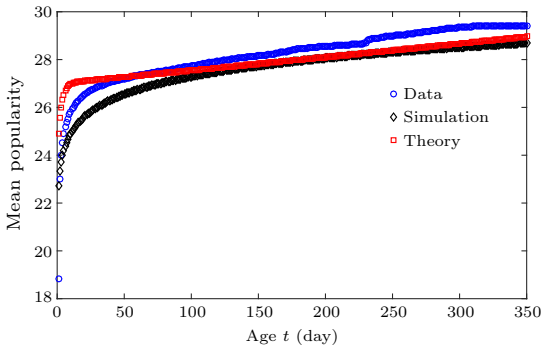

Fig. 4. Comparison of the numerical simulation, theory and with Sina micro-blog data. Numerical simulation of the mean popularity $m(t)$ of messages at age t (blue circle line), compared with empirical data (black diamond line) and theoretical analysis of Eq. (6) (red square line).

Figures 4 and 5 compare the theoretical and simulation results for the model on the real Sina micro-blog data. The theory matches the simulation results rather well. When $n \to \infty$, especially (in Fig. 5), the power-law exponent of popularity distribution and degree distribution are equal, and approximate the theoretical analysis results.

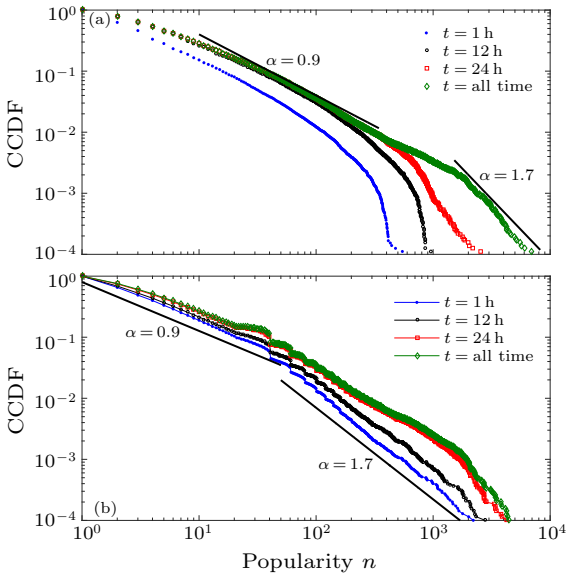

Fig. 5. Complementary cumulative distribution functions (CCDFs): the fraction of messages with popularity $\geqslant n$ for numerical simulations, compared with the theory of Eq. (7) and empirical data. (a) Simulation results of the model. Dashed lines correspond to CCDFs for power-law popularity distributions $p_n \propto n^{-\alpha}$, $p_k \propto k^{-\gamma}$ with $\gamma = 1.5$, $N = 10^5$, $\mu = 0.055$, $\lambda = 0.474$. (b) Sina micro-blog network with 125139 messages.

In summary, we have introduced a probabilistic model of information spreading that is analytically tractable yet realistic enough to reproduce several characteristic features of empirical data. In contrast to purely empirical studies or simulation-based models, we clearly analyzed the roles of two factors affecting information popularity: the competition of information and the structure of micro-blog network. We also compare theoretical and simulation results for the model on the real Sina micro-blog data, and the results match the real data rather well. However, the model has some limitations. It is required to make a number of assumptions to obtain the analytical results. There are many factors that affect the popularity of information, for example, predicting and modeling news popularity on social media[18] and quantifying paper popularity.[1] Meanwhile, it poses many challenges for researchers, including structural features,[19] the identification of the causes[20–23] and the time effect.[24] In the future, we hope for some possible extensions to the model.

References

[1] Wang D, Song C and Barabási A L 2013 *Science* **342** 127
[2] Sasahara K, Hirata Y, Toyoda M, Kitsuregawa M and Aihara K 2013 *PLoS ONE* **8** e61823
[3] Quattrociocchi W, Caldarelli G and Scala A 2014 *Sci. Rep.* **4** 4938
[4] Kim Y, Park S and Yook S H 2016 *Sci. Rep.* **6** 23484
[5] Song B, Jiang G P, Song Y R and Xia L L 2015 *Chin. Phys.* B **24** 100101
[6] Wang J R, Wang J P, He Zhen and Xu H T 2015 *Chin. Phys.* B **24** 060101
[7] Min L, Liu Z, Tang X Y, Chen M and Liu S Y 2015 *Acta Phys. Sin.* **64** 088901 (in Chinese)
[8] Wang J L, Liu F A and Zhu Z F 2015 *Acta Phys. Sin.* **64** 050501 (in Chinese)
[9] Li Y J, Yin C, Yu H and Liu Z 2016 *Acta Phys. Sin.* **65** 020501 (in Chinese)
[10] Cetin U and Bingol H O 2014 *Phys. Rev.* E **90** 032801
[11] Weng L, Flammini A, Vespignani A and Menczer F 2012 *Sci. Rep.* **2** 335
[12] Gleeson J P, Ward J A, O'sullivan K P and Lee W T 2014 *Phys. Rev. Lett.* **112** 048701
[13] Gleeson J P, O'sullivan K P, Banos R A and Moreno Y 2016 *Phys. Rev.* X **6** 021019
[14] Athreya K B and Ney P E 2012 *Branching Processes* (New York: Springer)
[15] Iribarren J L and Moro E 2011 *Phys. Rev.* E **84** 046116
[16] Xu Y, Guo L P, Ding N and Wang Y G 2010 *Chin. Phys. Lett.* **27** 078901
[17] Zhu J F, Han X P and Wang B H 2010 *Chin. Phys. Lett.* **27** 068902
[18] Zhang P, Li M H, Gao L, Fan Y and Di Z R 2014 *PLoS ONE* **9** e89192
[19] Zhang Z K, Liu C, Zhan X X, Lu X, Zhang C X and Zhang Y C 2016 *Phys. Rep.* **651** 1
[20] Yan Q, Yi L L and Wu L R 2012 *Physica* A **391** 1540
[21] Yan Q, Wu L R and Zheng L 2013 *Physica* A **392** 1712
[22] Hong W, Han X P and Wang B H 2009 *Chin. Phys. Lett.* **26** 028902
[23] Shang M S, Chen G X, Dai S X, Wang B H and Zhou T 2010 *Chin. Phys. Lett.* **27** 048701
[24] Miotto J M and Altmann E G 2014 *PLoS ONE* **91** e111506