Factors influencing message dissemination through social media

Zeyu Zheng,[1,2,3] Huancheng Yang,[3,1,2] Yang Fu,[1,2,*] Dianzheng Fu,[1,2] Boris Podobnik,[4,5,6,7] and H. Eugene Stanley[8]

[1]*Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China*

[2]*Key Laboratory of Network Control System, Chinese Academy of Sciences, Shenyang, 110016, China*

[3]*University of Chinese Academy of Sciences, Beijing, 100049, China*

[4]*Luxembourg School of Business, Luxembourg, L-2453, EU*

[5]*Faculty of Information Studies, SI-8000 Novo Mesto, Slovenia, EU*

[6]*Faculty of Civil Engineering, University of Rijeka, 51000 Rijeka, Croatia, EU*

[7]*Zagreb School of Economics and Management, 10000 Zagreb, Croatia, EU*

[8]*Center for Polymer Studies, Boston University, Boston, Massachusetts 02215, USA*

(Received 28 November 2017; revised manuscript received 7 March 2018; published 7 June 2018)

Online social networks strongly impact our daily lives. An internet user (a "Netizen") wants messages to be efficiently disseminated. The susceptible-infected-recovered (SIR) dissemination model is the traditional tool for exploring the spreading mechanism of information diffusion. We here test our SIR-based dissemination model on open and real-world data collected from Twitter. We locate and identify phase transitions in the message dissemination process. We find that message content is a stronger factor than the popularity of the sender. We also find that the probability that a message will be forwarded has a threshold that affects its ability to spread, and when the probability is above the threshold the message quickly achieves mass dissemination.

DOI: 10.1103/PhysRevE.97.062306

I. INTRODUCTION

Online social networks (OSNs), such as Facebook and Twitter, have a large presence on the modern information media platform [1]. OSNs transmit a huge amount of information, and traditional media sources now make widespread use of them to deliver their messages. The rapid recent development of OSNs has also lowered the threshold for the compilation of news and has blurred the boundary between media sources and users [2,3]. This instant release of information is both more rapid and more interactive than that achieved by traditional media [4,5]. Portals such as microblog quickly deliver information to target users. Online social media now play an important role in information dissemination, and examples range from targeted advertising [6], commodity recommendation systems [7], political propaganda [8] to networks of public opinion [9]. For example, using a Facetime transmission the Turkish president Erdogan called on Turkish citizens to actively oppose the coup [10]. During the United States presidential election, the Republican candidate Donald Trump tweeted frequently to quickly spread his ideas [11,12]. The rapid spread of information via OSNs has greatly changed our way of life.

Because the impact of OSNs on society is increasing, there is much interest focused on the dissemination mechanisms in social networks. To understand the mechanism underlying information diffusion, many previous studies either analyzed large amounts of empirical data, or predicted how popular a particular piece of information would become in the future [13–18]. For instance, Goel *et al.* discussed the diffusion patterns in different network and got the same results which explained in detail from the network structure level [19]. Network structure influence propagation mode, and information can also affect the dissemination during information explosion [20]. And there are a lot of models built up based on the topology [21,22]. An extension of the susceptible-infected-recovered (SIR) model of epidemics is frequently used to describe the dissemination of information [23–26]. Analyzing the communication mechanism of the spreading process is important if we are to improve our ability to disseminate our message. So long as we can make the information transmission ability exceed the threshold value accounting for the spreading of an epidemics [27], the scale-free property of social networks makes this possible [28,29]. Using a network of friend relationships among internet users, we use the SIR model in our message spreading experiment. In disease spreading, every person who comes in contact with an infected individual has the same probability of being infected, and an infected individual continues to infect others until they recover. The information dissemination process in OSNs is similar to that of epidemic spreading and also to binary-choice opinion models, but the details of the mechanism can differ.

II. DATA AND METHODS

A. Data source

Twitter is an online news and social networking website on which users post and interact through messages, "tweets," restricted to 140 characters. By the beginning of 2016, Twitter had more than 319 million monthly users. On the day of the 2016 U.S. presidential election, Twitter was the largest source of breaking news, with 40 million tweets sent by 10 p.m. (Eastern Time). The short, adaptable, rapid release

*Corresponding author: fuyang@sia.cn

of information through Twitter allowed the quick spread of information [30].

We use the opening and real-name data from Twitter provided by the Stanford Network Analysis Project (SNAP) to build our network [31]. Our network is of the relationships among Twitter users and has over 80 000 nodes, over 1 300 000 edges, and a network average degree $\langle k \rangle$ of approximately 33. We run our simulation tests on the simple undirected graph obtained by simplifying Twitter network data.

B. SIR model

We model mathematically the dissemination process of a message spread on OSNs using differential equations. In recent years extensions of the applied SIR epidemic model have been widely used in complex network research in various domains [32–37].

We use the SIR model to divide OSN Netizens into three categories.

(i) Suceptible: a Netizen who is likely to retweet a message once they receive it, and whose total number is denoted $S(t)$.

(ii) Infected: Netizens who are randomly generated from Susceptibles, who retweet the message, and whose total number is denoted $I(t)$.

(iii) Recovered: a Netizen who retweets the message once, does not retweet it again, and whose total number is denoted $R(t)$.

The total number of Netizens is denoted $N(t)$. This gives us

$$N(t) = I(t) + S(t) + R(t), \tag{1}$$

$$I(t) = pS(t-1) - \mu I(t-1), \tag{2}$$

$$R(t) = R(t-1) + \mu I(t). \tag{3}$$

A Netizen becomes an Infected (I) when they deliver a message. All users linked with that Netizen become Susceptibles (S) and can read the message. Susceptibles retweet the message and become Infected with a probability $p \overset{p}{\sim} \frac{I(t)}{S(t)}$. All Infecteds eventually become Recovereds and no longer retweet the message.

We define one retweeting step to be when a user retweets a message with probability p after receiving it. Within a few steps there is a sufficient number of Infecteds to spread the message throughout the network. We here simplify the SIR model by fixing some paraments to simulate the spreading process and lowering the impact of little factors. Reference [38] indicates most messages Twitter users share are also the most they receive, irrespective of differences between the number of users they follow and the number of users who follow them. Thus the most suitable approximation of message flow in a social network is undirected network.

We define a message to be "widely spread" when it has been viewed by greater than 80% of network users. When comparing the effects of different factors on the propagation results, we randomly select the message senders to reduce the impact of other factors on the results. When there are sufficient experimental users in our sample, the experimental environment provides all possible irrelevant factors and differ only in the variable we control.

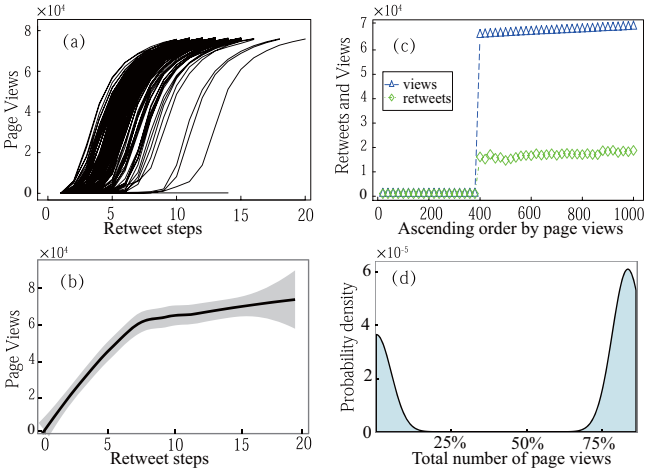

FIG. 1. Phase transition of page views and their distribution universally exist in the message dissemination. (a) Most of 1000 experiments rapidly rise within eight retweet steps. (b) Heavy line shows the fitting curve of page views which the page views rise quickly at the beginning, then start to level off. The light color interval is the standard error band of the fitting curves. (c) The green diamond shows the number of retweets while the blue triangle indicates page views, and they change synchronously. There are phase transitions exist in number of retweets and page views. More than 60% of messages can be spread widely. And others are hardly seen. (d) There are two distant peaks in the distribution of page views. Nearly 2/5 of the messages are hardly seen, and more than 3/5 of the messages could be spread across the network.

C. Kernel density estimation

We apply the Gaussian kernel density estimation method to determine the bandwidth for univariate observations. The calculation of kernel density is to display the distribution of data and to observe the possibility of the data falling on different values more smoothly. This enable us to better observe the distribution of, e.g., page views or user degree [39,40]. We also can calculate the probability distribution of X (i.e., the variable) using N univariate observations (i.e., $X_1 \tilde{} X_n$)

$$P_n(x) = \frac{1}{\sqrt{2\pi} nh} \sum_{j=1}^{n} e^{-(\frac{x-X_j}{2h^2})}. \tag{4}$$

Here, the h is the smoothing bandwidth. We scale the kernels using the standard deviation of the smoothing kernel.

III. RESULTS

In our simulation experiments we randomly select 1000 users to send the initial message, and we set the probability that the message is forwarded at 0.1. Then we record the views of the 1000 messages at the end of each retweet step. Figure 1(a) shows the 1000 curves of the viewing process. Note that some of the final views approach 0, implying that some messages are almost unseen. Most other curves rapidly increase and resemble an S shape. They rapidly grow during

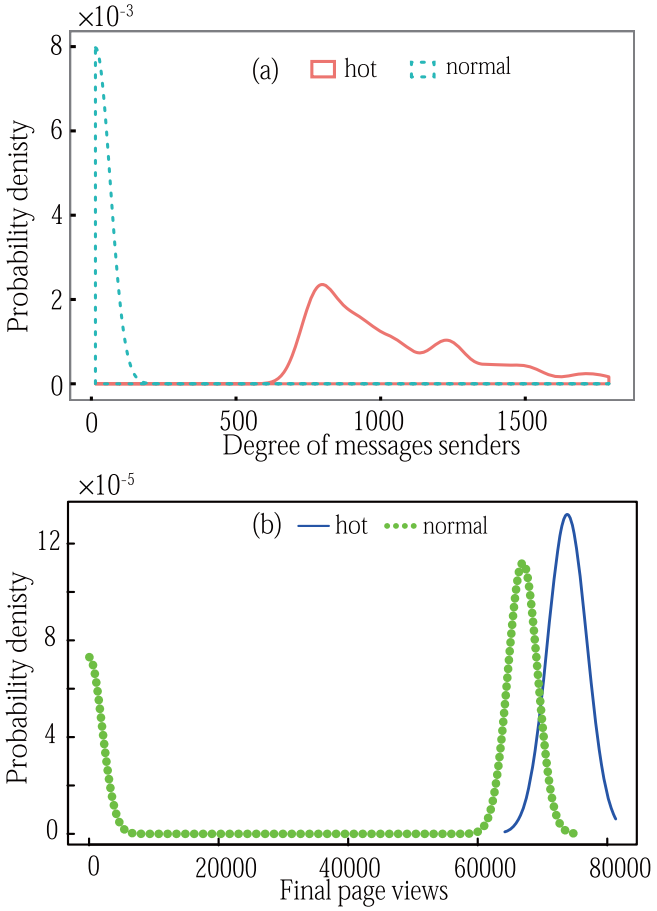

FIG. 2. Impact of message's degree distribution on infection range is very little. (a) Degree distributions between hot users and normal users differ greatly. The average degree of hot users is more than 20 times beyond the average of normal users. (b) The probability density distributions of the infection ranges between the hot users and the normal users is not very different. Nearly two thirds of the messages sent by normal users get widely spread while all the messages sent by hot users spread well. So hot users cost a lot but gain less.

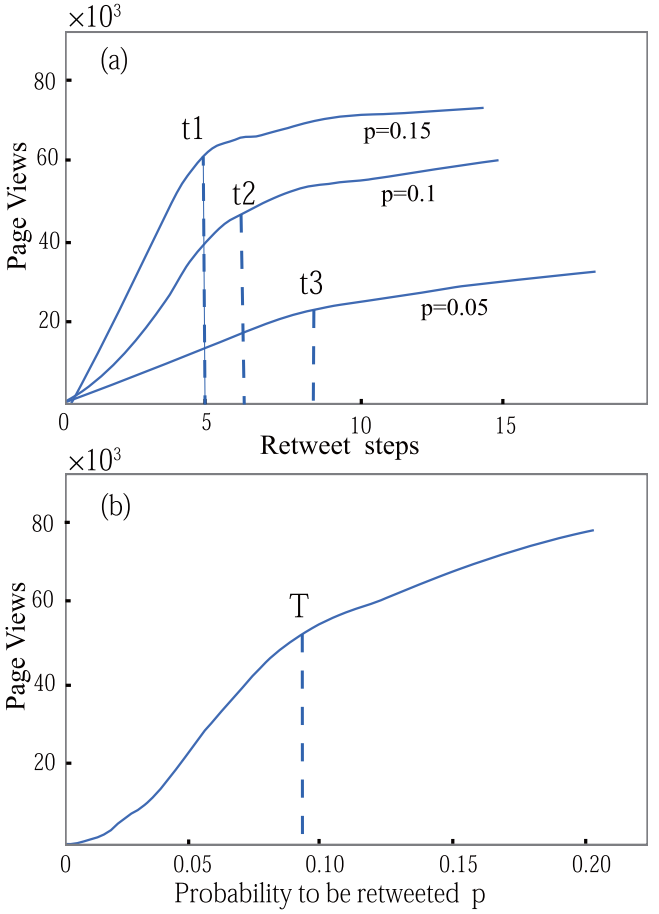

FIG. 3. The effect on message propagation is getting bigger with retweet-probability p rising. (a) It shows the page views curves of the spread process under three different retweet-probability. Each curve has a turning point that the curve's slope has decreased after the turning point. The bigger the p is, the earlier the turning points appear. (b) It indicates that the total page views of messages increase while the retweet probability p is increasing. The effect under different p values on the total page views is not linear. In general, the bigger the p value is, the smaller the influence on the infection range will be.

3–10 retweet-steps, and cease growing and stabilize after ten retweet steps. We use all of the curves in Fig. 1(a) to plot the fitting curve (heavy black line) in Fig. 1(b) applying the Loess method. Figure 1(b) shows the fitting standard error band (light gray), and also the trend of overall growth.

We determine the number of final page views and retweets and plot the scatter in ascending order by page view. Figure 1(c) shows the maximum number of page views of the 1000 messages (blue triangles) and the maximum number of retweets (green diamonds). The plot indicates that the page views and retweets of more than one-third of the messages approach zero, indicating that these messages have not spread. The numerical values of rest points suddenly increase up to maximum values, which approximately equal each other. The portion of the slope of the scatter plot curve that is very gentle indicates that when the message spreads it spreads across the entire network. It indicates the presence of a networks communicative ability threshold below which the message will not widely spread and above which it will. Because from the morphology of the scatter plot we find that the number of retweets agree with the number of page views, we can use just the message page views and not lose rigor. We calculate the distribution of the maximum number of the 1000 messages using a Gauss kernel density distribution. Figure 1(d) shows the probability density distribution curve. The higher the crest, the greater the probability that the variable falls within the range corresponding to the value of the abscissa. Note that there are two peaks, one in the vicinity of **0** and the other above 75%. Note that messages tend to either be widely viewed or almost entirely ignored. The middle part of the trough is wide and values approach 0, verifying the presence of a phase transition.

We analyze the data that fall around different peaks to determine the main factors that affect the ranges of infection. We first consider the message sender effect. We examine the messages sent from all users to locate "hot users," the 100 users with the highest degree, and 100 normal users that retweet a message with probability p to serve as a control group. The spreading process is random. Hot users and normal

users spread the same messages. Figure 2 shows user degree distributions and probability densities of the message infection ranges.

Figure 2(a) shows the contrasting curves of the two types of user degree distributions. The data are concentrated at a wave crest. The degree distribution curve of normal users (blue dotted line) shows a peak at approximately 30, i.e., on average they have 30 friends. The degree distribution curve of hot users (red solid line) shows their degree is greater than than 800, more than 20 times that of normal users. High degree users with many neighbors are the primary message spreaders.

Figure 2(b) shows probability distribution density of the total message views sent by regular users (green dotted line), and the probability distribution density of the total message views sent by hot users (blue line). The infection range distribution curve of messages sent by regular users has peaks at 0 and 65 000, and the ratio of the area is 1 : 2, implying that one-third of common user messages do not spread, but two-thirds do. The distribution of the message views sent by the hot users has only one peak, at 72 000, implying that all messages sent by hot users spread widely. Thus the communication effect of hot users is approximately 50% higher than that of normal users. Comparing Figs. 2(a) and 2(b), the degree distributions indicate that hot users have approximately 20 times more friends than regular users, but this yields only an additional 50% payback. Hence the cost-effectiveness ratio is $y = 50\%/20 = 2.5\%$.

The impact of retweet probability on the dissemination of a message is much stronger than the impact of user degree. Figure 3 shows the distribution of page message views for different retweet probability p values. Here, we set the p values at 0.05, 0.1, and 0.15 to test three groups. Apart from this p value, all other parameters are fixed. Figure 3(a) shows the three spreading process curves. Note that the message page views increase and the p values increase, and they reach the turning point sooner. Figure 3(b) shows that the slope of the curve is very high between $p = 0$ and $p = 0.1$, and then markedly slows after $p = 0.1$. Applying double devotion when p is less than 0.1 yields double revenue. The cost-effectiveness ratio is $y = 100\%$. Applying double devotion when p is greater than 0.1 yields one-half revenue. The cost-effectiveness ratio is $y = 50\%/1 = 50\%$. The average cost-effectiveness ratio is greater than 20 times higher then that provided by message sender degree.

We classify a message "hot" when it reaches 80% of the users. Figure 4(a) shows the relationship between the number of retweet steps and the retweet probability p required to make a message hot. Figure 4(b) shows a log-log curve of two regimes with differing slopes. The slope of the second regime is smaller than the slope of the first regime, indicating that the dissemination slows after reaching the turning point. The turning point in the graph is close to the one we obtained in the previous experiment. When the retweet-probability increases beyond a threshold, the page views increase more slowly. In the process of information propagation, the maximum repetition number (i.e., the number of users who receive the message more than once in the spread process) increases with the retweet-probability p, and then decreases slightly (Fig. 5). After the experiment on twitter data, we repeat the experiment based on the Facebook and Wiki-vote data. We carried out several groups of experiments under different thresholds of

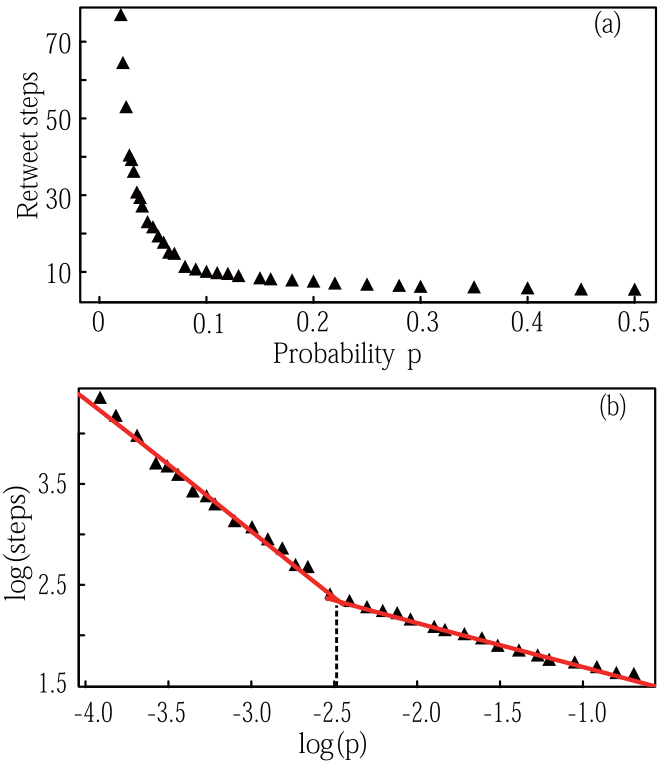

FIG. 4. (a) The retweet steps needed for the same dissemination is approximately exponential in relation to the retweet probability p. (b) Logarithmic curve of $\log(p)$ and log(steps) is divided into two sections. That means characteristics of network communication change with the two different slopes. The turning point of the curve is close to the one in the previous experiment. When the retweet probability increases beyond a threshold, the users who receive messages from different users have a great overlap, so that the page views increase more slowly.

"widely spread" to verify the relationship between $\log(p)$ and log(steps) under different scales of the network. The results show that the larger the network, the smaller the p at turning point. Also turning points get bigger with arising of thresholds. It is a universal phenomenon that there are turning points at the curve of log(steps) and $\log(p)$. The slopes of the curves change around the turning points. The slopes of the curves imply

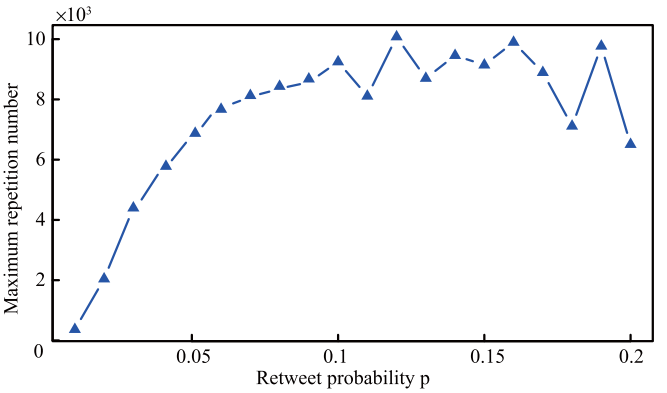

FIG. 5. The relationship between the maximum repetition number and the retweet-probability p.

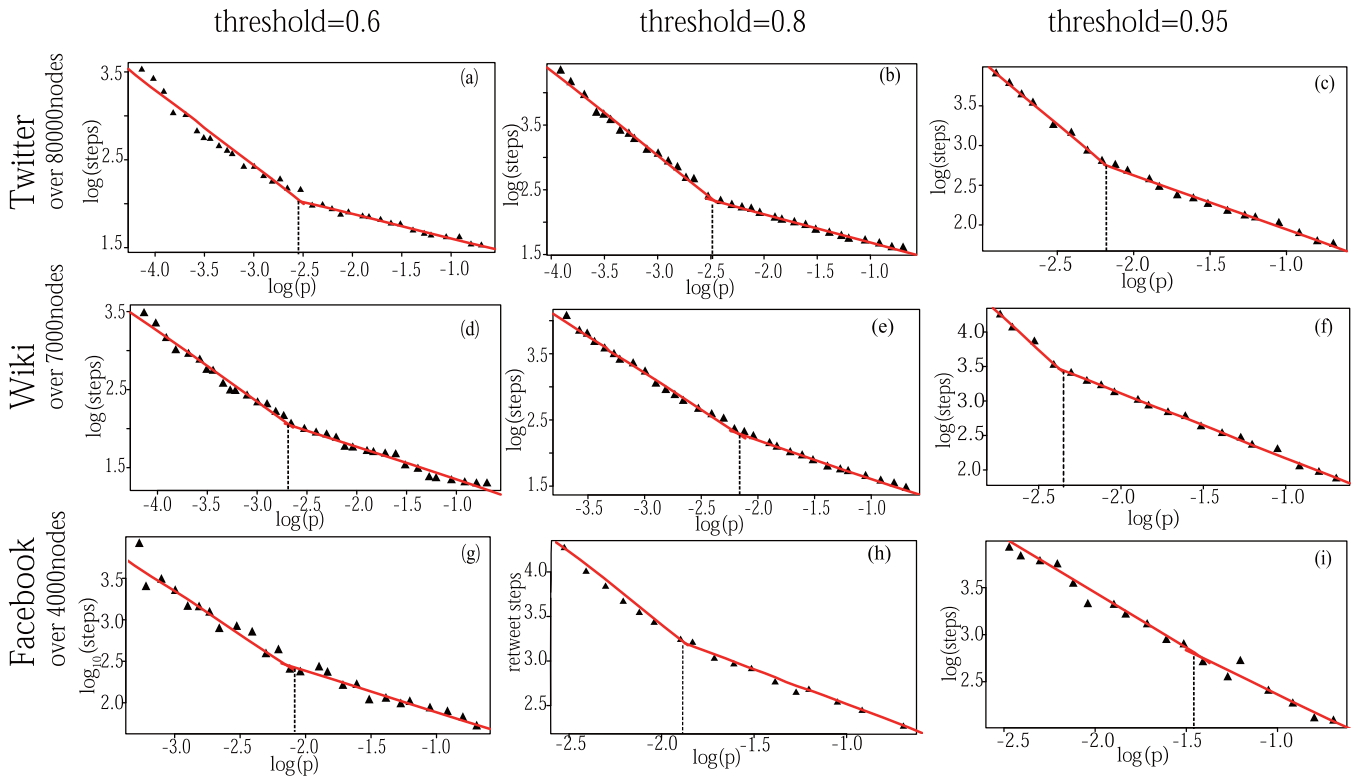

FIG. 6. The relationship between log(p) and log(steps) under different scales of the network and different thresholds of "widely spread".

the resistance of the network to the propagation of message (Fig. 6).

IV. CONCLUSION

We have simulated the process of message dissemination on OSNs, and we find a phase transition in the distribution magnitude. In the two experiments, we find that the main factor that influences the infection range of a message is the quality of the message itself (i.e., the message propagation coefficient p, which is the probability that the message is retweeted). The influence of the message sender is relatively small, less than 5% of the influence of the quality of the message. Thus to be effectively communicated a message must encourage users to retweet, increase the probability of its being forwarded, and eventually become "hot". One outstanding example was the immensely popular but short-lived "Gangnam Style" music video that quickly spread around the world—receiving more than one billion views on YouTube. There are many other current examples in pop culture and on the iInternet—subjects that quickly become known worldwide. The ability of hot users to guide public opinion through the network is also becoming progressively weaker, and the social network is becoming multipolar. Thus normal users can draw attention to themselves as long as they are able to release content that attracts people.

Our results indicate that the retweet-probability $p = 0.1$ is a turning point that may be related to network structure [41–43]. Figure 5 shows the resistance to the spread of information. The resistance to message spreading differs according to differences in network density [44]. Thus we can achieve a good communication effect with a small investment if we can find a better critical p-views point.

ACKNOWLEDGMENT

This research was supported by the Program for National Natural Science Foundation (71671182). And the Boston University Center for Polymer Studies is supported by NSF Grants PHY-1505000, CMMI-1125290, and CHE-1213217, and by DTRA Grant HDTRA1-14-1-0017.

[1] K. Kaupo, K. Lafsson, and L. Blinka, The effect of smartphone use on trends in European adolescents excessive internet use, Behav. Inf. Tech. **35**, 68 (2016).

[2] A. G. Asmolov and G. A. Asmolov, From we-media to i-media: Identity transformations in the virtual world, Psychology in Russia State of Art **1**, 101 (2009).

[3] L. Chu, Y. Zhang, G. Li, and S. Wang, Effective multimodality fusion framework for cross-media topic detection, IEEE Trans. Circuits Syst.Video Technol. **26**, 556 (2016).

[4] G. Iyer and Z. Katona, Competing for attention in social communication markets, Ssrn Electronic Journal **62**, 2304 (2016).

[5] Y. Peng, J. Li, H. Xia, S. Qi, and J. Li, The effects of food safety issues released by we media on consumers awareness and purchasing behavior: A case study in China, Food Policy **51**, 44 (2015).

[6] K. Bimpikis, A. Ozdaglar, and E. Yildiz, Competitive targeted advertising over networks, Oper. Res. **64**, 705 (2016).

[7] A. Tuzhilin, Towards the next generation of recommender systems, ICEBI-10, 2010.

[8] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, A 61-million-person experiment in social influence and political mobilization, Nature **489**, 295 (2011).

[9] F. Xiong, Y. Liu, and J. Cheng, Modeling and predicting opinion formation with trust propagation in online social networks, Commun. Nonlinear Sci. Numer. Simul. **44**, 513 (2017).

[10] H. Pope, Broadcast revolution in the air for turks; a boom in radio and television stations is shining light into turkey's dark corners, writes Hugh Pope in istanbul, https://www.independent.co.uk/news/world/europe/broadcast-revolution-in-the-air-for-turks-1567946.html.

[11] A. Bovet, F. Morone, and H. A. Makse, Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump, arXiv:1610.01587 (2016).

[12] A. Saini and N. Markuzon, Predictive modeling of opinion and connectivity dynamics in social networks, 2016.

[13] E. Adar and L. A. Adamic, Tracking information epidemics in blogspace, in *Ieee/wic/acm International Conference on Web Intelligence* (IEEE Computer Society, Compiegne, France, 2005), p. 207.

[14] A. Anagnostopoulos, R. Kumar, and M. Mahdian, Influence and correlation in social networks, in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, Las Vegas, Nevada, USA, 2008), p. 7.

[15] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, *International Conference on World Wide Web* (ACM, New York, 2004), p. 491.

[16] L. Guo, E. Tan, S. Chen, Z. Xiao, and X. Zhang, The stretched exponential distribution of internet media access patterns, in *Twenty-Seventh ACM Symposium on Principles of Distributed Computing* (ACM, Toronto, Canada, 2008), p. 283.

[17] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, and S. Moon, I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system, in *ACM SIGCOMM Conference on Internet Measurement* (ACM, San Diego, CA, 2007), p. 1.

[18] K. Lerman and R. Ghosh, Information contagion: An empirical study of the spread of news on digg and twitter social networks, Comput. Sci. **52**, 166 (2010).

[19] S. Goel, D. J. Watts, and D. G. Goldstein, The structure of online diffusion networks, in *ACM Conference on Electronic Commerce* (ACM, Valencia, Spain, 2012), p. 623.

[20] E. Massaro, A. Ganin, N. Perra, I. Linkov, and A. Vespignani, Resilience management during large-scale epidemic outbreaks, Sci. Rep. **8**, 1859 (2018).

[21] D. Borkmann, A. Guazzini, E. Massaro, and S. Rudolph, A cognitive-inspired model for self-organizing networks, in *IEEE Sixth International Conference on Self-Adaptive and Self-Organizing Systems Workshops* (IEEE, Lyon, France, 2012), p. 229.

[22] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, Information diffusion in online social networks: A survey, Acm Sigmod Record **42**, 17 (2013).

[23] M. E. J. Newman, Spread of epidemic disease on networks, Phys. Rev. E **66**, 016128 (2002).

[24] R. Pastor-Satorras and A. Vespignani, Epidemic Spreading in Scale-Free Networks, Phys. Rev. Lett. **86**, 3200 (2001).

[25] S. Meloni, A. Arenas, S. Gmez, J. Borge-Holthoefer, and Y. Moreno, *Modeling Epidemic Spreading in Complex Networks: Concurrency and Traffic* (Springer, New York, 2012).

[26] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, Theory of rumour spreading in complex social networks, Physica A **374**, 457 (2007).

[27] C. Granell, S. Gomez, and A. Arenas, Dynamical Interplay Between Awareness and Epidemic Spreading in Multiplex Networks, Phys. Rev. Lett. **111**, 128701 (2013).

[28] P. Blanchard and T. Krüger, The cameo principle and the origin of scale-free graphs in social networks, J. Stat. Phys. **114**, 1399 (2004).

[29] D. Centola, The spread of behavior in an online social network experiment, Science **329**, 1194 (2010).

[30] Twitter in wikipedia, https://en.wikipedia.org/wiki/Twitter/.

[31] Stanford network analysis project, http://snap.stanford.edu/data/.

[32] J. Woo and H. Chen, Epidemic model for information diffusion in web forums: Experiments in marketing exchange and political dialog, Springerplus **5**, 66 (2016).

[33] C. H. Li, C. C. Tsai, and S. Y. Yang, Analysis of epidemic spreading of an sirs model in complex heterogeneous networks, Commun. Nonlinear Sci. Numer. Simul. **19**, 1042 (2014).

[34] D. Shah and T. Zaman, Detecting sources of computer viruses in networks: Theory and experiment, Acm Sigmetrics Performance Evaluation Review **38**, 203 (2010).

[35] X. Chu, Z. Zhang, J. Guan, and S. Zhou, Epidemic spreading with nonlinear infectivity in weighted scale-free networks, Physica A **390**, 471 (2011).

[36] O. N. Bjørnstad, B. F. Finkenstädt, and B. T. Grenfell, Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series sir model, Ecological Monographs **72**, 169 (2002).

[37] J. Wu, Z. Gao, and H. Sun, Simulation of traffic congestion with sir model, Mod. Phys. Lett. B **18**, 1537 (2004).

[38] M. G. Rodriguez, K. Gummadi, and B. Schoelkopf, Quantifying information overload in social media and its impact on social contagions, eprint arxiv, 2014.

[39] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, Kernel density estimation via diffusion, Ann. Stat. **38**, 2916 (2010).

[40] T. Rui, Y. Zhou, and M. A. Guangyan, Target detection using kernel density estimation and gaussian model cascade mechanism, Computer Engineering & Applications **47**, 1 (2006).

[41] X. Z. Zhang and A. D. Chang, Discharge-time curve fitting method for calculating hydrogeological parameters in loess aquifer, Water Resources Protection **25**, 25 (2009).

[42] A. Alsayat and H. El-Sayed, Social media analysis using optimized k-means clustering, in *IEEE International Conference on Software Engineering Research,Management and Applications* (Song, Baltimore, MD, 2016), p. 61.

[43] J. Kleinberg, The small-world phenomenon: An algorithmic perspective, Proceedings of Acm Symposium on Theory of Computing **406**, 163 (2000).

[44] C. Stegehuis, R. V. D. Hofstad, and J. S. H. V. Leeuwaarden, Epidemic spreading on complex networks with community structures, Sci. Rep. **6**, 29748 (2016).