Study of Information Spreading on Micro-blogging based on SEIR model

Abstract: Micro-blogging has become an important channel for information spreading nowadays. In this study, we build a system dynamics model to investigate the mechanism of information dissemination on micro-blogging. The model is an extension of the traditional SIR model considering the feature of micro-blogging and its users. We gathered the data from Sina micro-blogging for the case of a school bus accident happened in Nov 2011. Data shows that new posts and forwarding posts outbreak quickly after the accident. The number of new posts and forwarding posts both peaked on the second day, and then gradually reduced. Six days later, the number of posts stabilized at a relatively low level. We use the data to calibrate the SD model and then studied how different parameters affect the spread of information. We found that infectivity not only affects the speed of information spreading but also affects the wide-spread of information. Increasing the hesitation rate (change from infective state to exposed state) greatly hinder information spreading.

Keywords: system dynamics, information spreading, micro-blogging, SIR model

1. Introduction

Similar to twitter, Sina micro-blogging (also known as Sina weibo) has become one of the most influential channels for information spreading. Since its first online in August 2009, Sina micro-blogging had attracted more than 250 million users by the end of year 2011. People use Sina micro-blogging to express their opinion, share their experiences and spread information. The analyst from the People's Daily Online asserts that the most influential media in future will be micro-blogging. The great aspect of Micro-blogging is that it facilitates interaction between people, free-expression of opinions, and information spreading. During emergencies, people are using micro-blogging to help other. For example, during the Beijing torrential rain this summer, self-organized groups used Sina micro-blogging to post street flood information and where people need a ride. Yet, at the same time, micro-blogging has its own problems. Rumors spread fast on micro-blogging and could cause public panic. One example is the irrational purchase of iodized salt of the Chinese citizens after Japan’s 9-magnitude earthquake, tsunami, and nuclear crisis. Rumor from internet and micro-blogging said that consuming iodized salt helps protect against radioactive exposure and that radiation leaks from the Japan nuclear power plants could contaminate sea salt production, possibly leading to a salt shortage in the near future. From the afternoon of Mar. 16th, 2011, residents in the coastline cities began hoarding salt. The next day, this irrational buying of salt has spread all over China, causing a temporary salt shortage in shops. Consequently, the increasing demand drove up salt price, rising to more than five times of the normal price in some places. This caused huge instability for the economy and the society. For its virtue side and its vicious side, the information spread mechanism on this new media—micro blogging—needs investigation.

2. Literature review

Our model is build on previous study of the rumor models, which started in the 1960s. As rumor spreading shows great similarity to the epidemic spreading, most of the existing models of rumor spreading are based on the epidemic models [1][2]. Daley and Kendall [3] first proposed the basic DK model of rumor spreading, which described rumors spreading from the spreader to other people in a crowd through a two-way communication link, and following the mass action law. In this model, individuals are divided into three groups: susceptible, infected, and recovered (*SIR*). The *SIR* model reflects the transitions from the susceptible state to the infected state at the infection rate α. Meanwhile, infected individuals are transferred to the recovery state at the recovery rate β. The transformation chart is described in Figure 1:

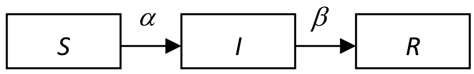

Figure 1. The transformation chart of *SIR* model

Maki and Thomson [4] focused on the analysis of the rumor spreading model based on mathematical theory and developed the MK model. The DK and MK models have been used extensively for quantitative studies of rumor spreading [5]-[11]. Sudbury [12] studied the dynamic mechanisms of information transmission on social networks, including public information dissemination, and suggested that the dynamic behavior of rumor spreading matched the SIR model. Brown et al. [13] extended the SIR model by adding a new group of people, the exposed individuals, to study epidemic spreading. They proposed an algorithm for symbolic deduction of the basic reproductive rate by a local analysis of the disease-free state and analyzed the epidemic threshold in SEIR model. The transformation chart is described in Figure 2:

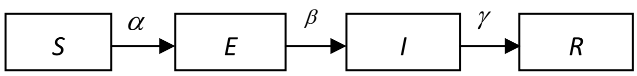

Figure 2. The transformation chart of *SEIR* model

SEIR models are widely used in epidemic contagion research. Information spreading process on micro-blogging has similarities to what described by the *SEIR* model in that people who read a post (contact an infected people) might not start to spread the information (forward the post) immediately. They are exposed but not yet become a spreader. Even so, the SEIR model has some pities. For example, in *SEIR* model, when a susceptible contacts with a spreader, the susceptible may maintain the state of susceptible or accept the rumor to become a spreader. This assumption ignores the possibility that the ignorant become a recovered directly. This possibility lies in that some susceptible have a strong background knowledge and some have little interest in the rumor.

As a result, we extend the classical SEIR rumor spreading model by ① considering the situation when a susceptible contacts an infected, the susceptible becomes a recovered directly with a certain probability; ② the exposed could directly turn into recovered with forgetting mechanism. The new *SEIR* model will be introduced in section 3. In section 4, we introduce a case of school bus accident happened in Gansu in Nov. 2011. In section 5, we use the data from the case to calibrate the model and use the model to investigate various scenarios. We conclude our paper in section 6.

3. Model introduction

People are divided into four groups according to their behavior: 1) the susceptible individuals, denoted by S, who have not heard the information; 2) the exposed individuals, denoted by E, who have heard the information but are not yet spreading the information; 3) the infected individuals, denoted by I, who are forwarding posts to other people; 5) the recovered individuals, denoted by R, who previously forwarded post but no longer do that either because lost interest in "old" news or forgetting mechanism. They will be not spread such related information any more. The total users of the micro-blogging is denoted by N. Therefore, S, E, I, R and N satisfy the condition:

$$S + E + I + R = \mathrm{N} \tag{1}$$

N represents the total number of active members using Sina micro-blogging. We assume no inflow and outflow rates in the short time spam for one event.

The transformation chart of these four groups is described in Figure 3:

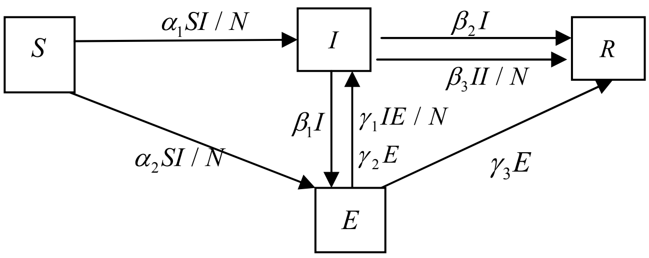

Figure 3. *The information spreading model based on SEIR model*

α_1 and α_2 denotes the probability a susceptible become an infected and an exposed upon the contact with an infected. β_1 denotes the probability of an infected to become an exposed with temporally forgetting mechanism. β_2 denote the probability of an infected to become a recovered with forgetting mechanism. B_3 denote the probability of an infected to become a recovered when the infected found the information is not new and lost interest in spreading it. The realization that the information is not new is caused by contacting another infected person. γ_1 denotes the probability of the exposed individuals transferring to the infected state by contacting an infected while γ_2 denotes the probability of exposed individuals transferring to the infected state naturally. Γ_3 denotes the probability of the exposed individuals transferring to the recovered state with forgetting mechanism.

4. The case of school bus incidents

On November 16, 2011, a school bus run by a private kindergarten collided head-on with a coal truck in Gansu province, China. 19 children were killed, as were 2 adults. The bus was originally a nine-seat van, but it had been modified to carry more passengers, and was severely overcrowded; 62 children were on board at the time of the crash, along with the 2 adults. Shortly after the event, a great deal of outcry on the Chinese internet outburst, with many posters criticizing the small amount of money spent on education.

We use Sina micro-blogging "Advanced Search" function to collect the number of posts about this accident from November 16 to November 25. The information we collect is public microblogging. For those users who set their posts as private, their posts are not able to be reached.

Various keywords have been tried for the search, such as Gansu school bus, school bus accident, kindergarten school bus safety regulations. Yet we found many posts associated with this incident didn’t mention “Gansu” or “kindergarten”. Therefore, in the end we decided to use "school bus" as the keywords for the search. It is broader than the ones we mentioned early. As we check, during Nov 16 to 25, the posts that mention school bus mostly are connected to this school bus accident.

When collect data, we focused on not only the number of total posts, but also the number of new posts, forwarded posts, posts with video, and posts with URL (link to the other pages short chain). Table 1 presents the data we collected.

Table 1 Number of posts for Gansu school bus incident

Date	Total posts	New posts	Forwarded posts	Posts with video	Posts with URL
11.16	492,100	68,628	423,472	46,284	104,272
11.17	1,078,896	96,824	982,072	83,524	245,252
11.18	573,496	95,760	477,736	22,876	163,856
11.19	332,500	37,408	295,092	7,221	65,436
11.20	221,312	33,516	187,796	2,858	50,008
11.21	101,080	26,068	75,012	1,908	23,408
11.22	90,440	18,088	72,352	10,762	38,836
11.23	47,880	8,733	39,147	377	7,980
11.24	31,388	8,512	22,976	254	9,576
11.25	22,344	5,320	17,024	303	7,448

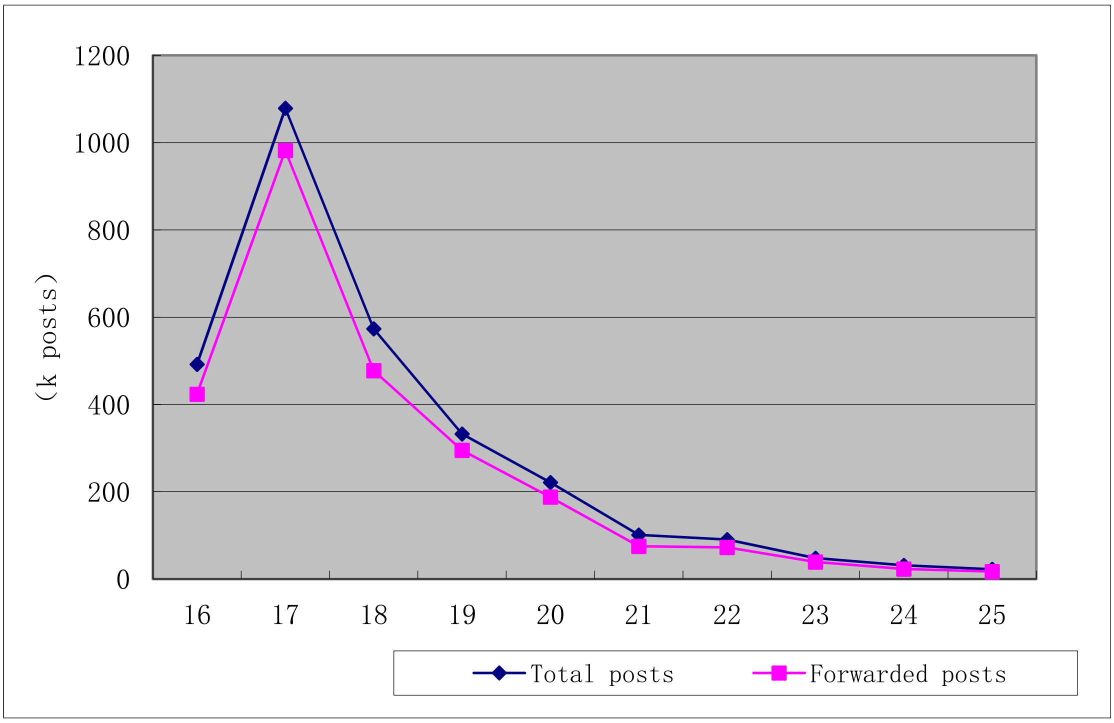

Figure 4. *The number of total posts and forwarded posts*

We can see that the number of posts increases quickly after the accident happened and peaked on the second day. Most posts are forwarded posts, with less than 20% posts are new posts. The forward function of micro-blogging has great facilitated the spread of information. With one click, it is almost effortless that information is forwarded. With more mobile connection, people can spread information anytime, anywhere. The number of posts reduces after the second day and gradually stabilizes at a rather low level after 6 days. The pattern is extremely similar to epidemic spreading.

New posts, posts with video and posts with URL, also exhibit the same pattern. Among these three categories, the posts with URL have greatest number of posts. Micro-blogging has word limits of 140 words, which is far from enough to reveal detailed information. Therefore, almost 25% of the posts have a URL. Less than 10% of posts have a video. This finding is surprising as most people assume that posts with video are more attractive to audience. One possible reason for so few posts with videos is that the videos in the posts are from local TV news. There is no on-the-spot recording of the event. Therefore, the videos are not so interesting to audience. We will further investigate the spread of post with videos in the future.

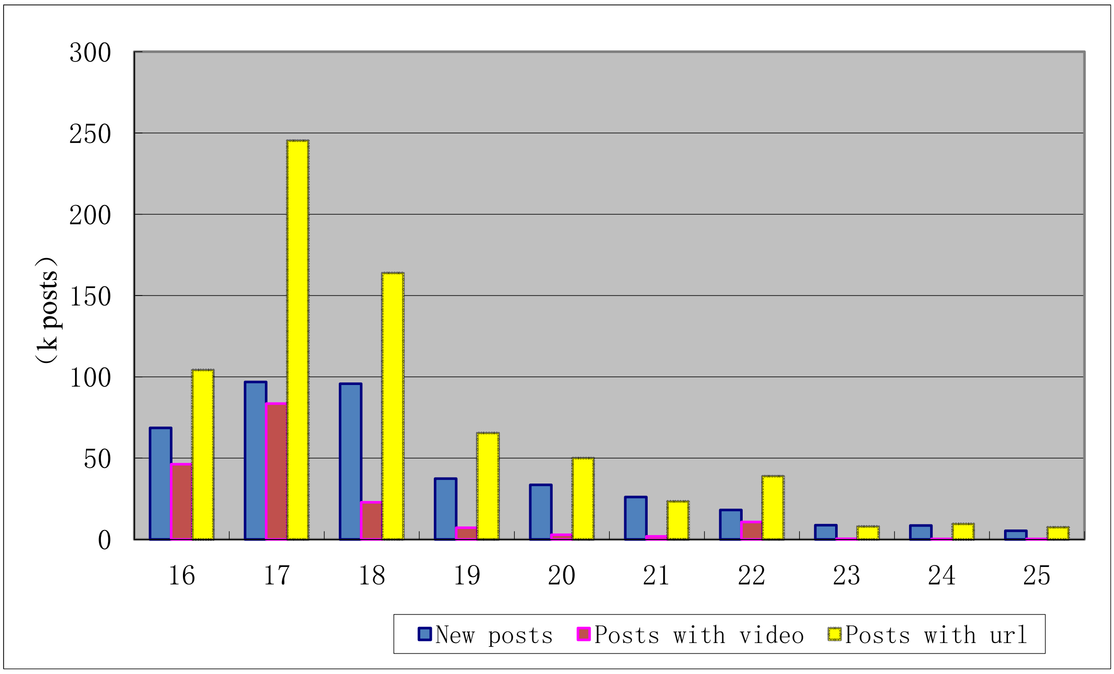

Figure 5. *The number of new posts, posts with video and posts with URL*

5. Simulation

5.1 Parameter setting and model calibration

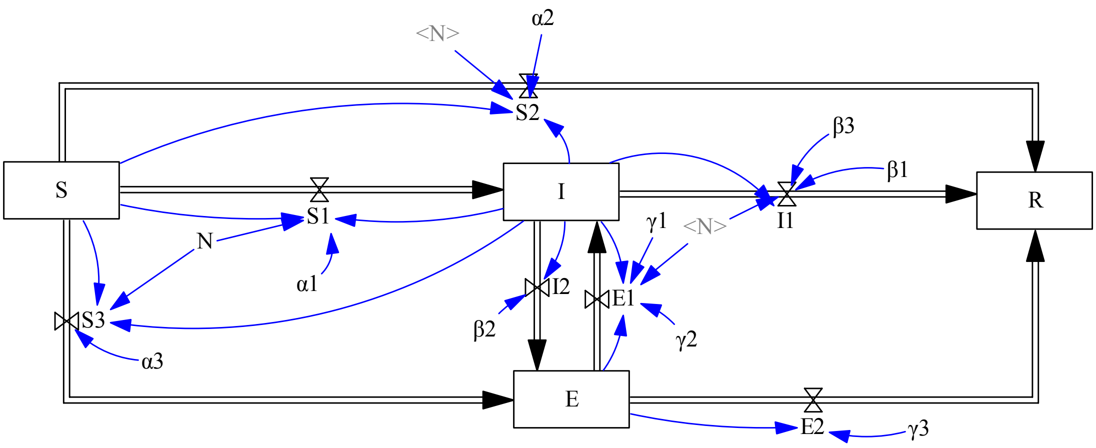

Figure 6. *The SEIR model for information spreading*

When the accident happened, the membership on Sina micro-blogging is 200 million people. However, only 10% are active members. Therefore, we set the total population N as 20 million persons. The post about this school bus accident started at 13:49 Nov 16th. Untial 15:00 Nov 16th, there are 3,480 posts on Sina micro-blogging. Therefore, we use 3.48K persons as the number of initial infected person. There is an assumption that each blog is written by a different individual. People will not write or forward two posts for the same events. This assumptions needs to be further investigated. Initial exposed, initial recovered are both set to be 0. Using equation 1, we can calculate the initial susceptible is 19996.52 k persons.

Then we use the total number of posts as a reference mode for the infected ones and calibrated the model to get α_1, α_2, α_3, β_1, β_2, β_3, and γ_1, γ_2, γ_3. We have to make some adjustment on the calibration results as the model calibrate didn't count the following constrains for the model:

$$\alpha_1+\alpha_2+\alpha_3=1 \quad (2)$$

$$\beta_1+\beta_2+\beta_3<=1 \quad (3)$$

$$\gamma_1+\gamma_2+\gamma_3<=1 \quad (4)$$

Equation 2 means that when a susceptible meet an infected, the person will change to infected or exposed or recovered, not able to be a susceptible person any more. For the exposed and infected, they could change other state except susceptible person or they could remain in their state. The final result lies as following:

$$\alpha_1 = 0.28,\ \alpha_2 = 0.06,\ \alpha_3 = 0.66$$

$$\beta_1 = 0.78,\ \beta_2 = 0.12,\ \beta_3 = 0.002$$

$$\gamma_1 = 0.03,\ \gamma_2 = 0.01,\ \gamma_3 = 0.02$$

We found the result also fit to our empirical result. We did some interview on person's behavior on micro-blogging and found that: 1) when encountered with an infected, people mostly likely to change into the exposed state. Because on micro-blogging, it is hard for person to tell whether the post reveal true information or not and most people will not spread information, therefore α_3 is the largest; 2) people also expressed the feeling that when the information has already been widely spread, they will not forward the post again. This is in consistent with the calibrate result that β_1 is the largest. β_1 represents the probability a person become a stifle when encountered with another infected. This probability should be quite high as more information lost its novelty; people are less likely to spread it.

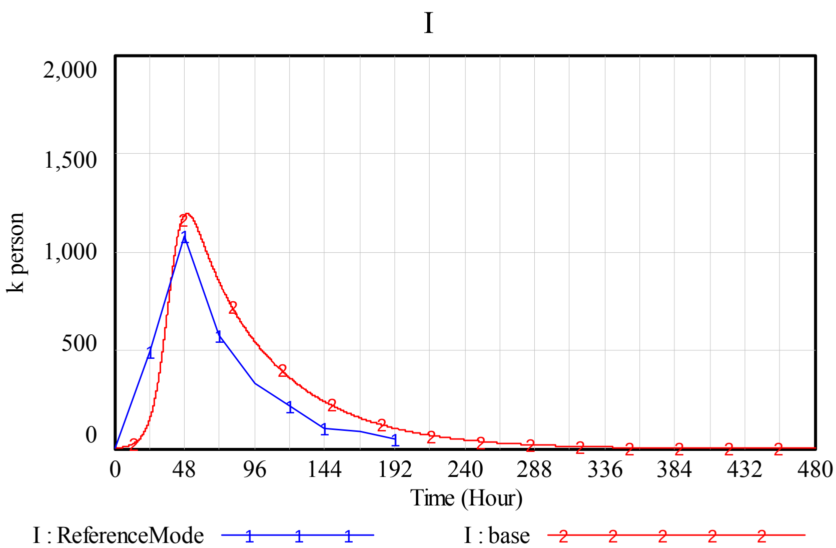

Figure 7. *The simulation result of SEIR model comparing with the reference mode*

We can see from the simulation result that the number of posts increases quickly at the beginning of the accidents and peaks on the second day. Then the number of posts gradually dies out over time. This shows exactly the same pattern as the data we collected. Such a pattern fit adds our confidence that the model could be used to study the information spreading on micro-blogging.

5.2 Scenario investigation

Based on this model and the calibrated parameters, we simulated three scenarios to investigate how various parameters affect the information spreading.

Table 2 Parameter setting for three scenarios compared with base run

Scenarios	N	α_1	B_2
1. Base	2000	0.28	0.12
2. Increase total population 25%	2500	0.28	0.12
3. Increase infectivity 25%	2000	0.35	0.12
4. Increase hesitation rate	2000	0.28	0.18

Note that when α_1 is changed, α_2, and α_3 will change proportionally because of the constrains mentioned above.

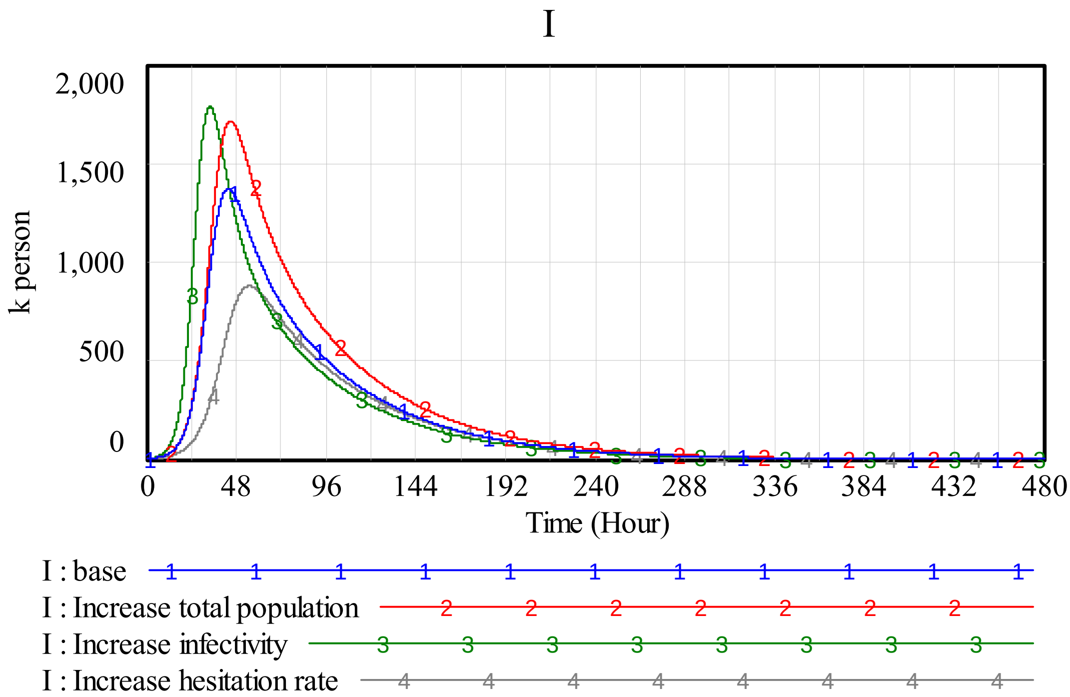

Figure 8. *The simulation result of SEIR model comparing with the reference mode*

From the simulation result, we can see that the four scenarios are exhibit similar pattern, with the number of infected person increases quickly at the beginning and peaks early (first day to second day). Then it takes some days for the discussion of this event to die out. For scenario 2, more population (red line with number 2), the number of infected person peaks at the same time as the base scenario, but the number of people infected is high then the base scenario. For the scenario of high infectivity (green line with number 3, the number of infected person increases fastest among the four scenarios, resulting the earliest peak of the number of infected persons. At the same time, the peak is the highest. This means, increase the infectivity not only make the information spread faster, but also wider. Compare with scenario 2, both scenarios increase the two variables 25%, increase infectivity has more impact on information spreading. For the scenario of high hesitation rate, the number of infected person peaks a little bit later than the other scenario but more lower. This variable has a big impact on the information spreading as well.

6. Conclusion

In this study, we adjusted the SEIR model to fit the information spread mechanism on micro-blogging. After we build the structure of the model, we use the data from a case to calibrate the model. And then we use the calibrated model to study various scenarios in order to see how these parameters affect the information spreading on micro-blogging. We found that increase infectivity greatly facilitate information spreading while increase hesitation rate highly hamper information spreading.

The work reported here is one part of ongoing research. There are many aspects to improve in the future. First, in this model, we didn't differentiate post and person. The assumption is that most person only forward one post for an event. But we highly suspect this assumption. More research should be done in this direction. And we might consider changing the model structure to better fit the reality. Second, in this model, we only use one case to calibrate the model. More cases should be tried. It is sure that different case will lead to different parameters. The idea is whether we can find any common rule for the same type of cases.

References

[1] R. M. Anderson, R. M. May, Infectious diseases of humans, Oxford University Press, Oxford, UK, 1992.

[2] H. W. Hethcote, The mathematical of infectious diseases, SIAM Review 42 (2000) 599–653.

[3] D. J. DALEY and D. G. KENDALL. Stochastic Rumors. Journal of Applied Mathematics 1965, 1 (1):42-55.

[4] D. Maki, M. Thomson, Mathematical models and applications, Prentice-Hall, Englewood Cliffs, 1973.

[5] M. E. J. Newman, S. Forest, J. Balthrop, Email networks and the spread of viruses, Physical Review E 66 (2002) 035101.

[6] N. Smith, Instant messaging as a scale-free network, Arxiv: cond-mat/0206378, 2002.

[7] G. Csanyi, B. Szendroi, Structure of a large social networks, Physical Review E 69 (2004) 036131.

[8] F. Wang, Y. Moreno, Y. Sun, Structure of peer-to-peer social network, Physical Review E 73 (2006) 036123.

[9] B. Pittel, On a Daley–Kendall model of random rumours, Journal of Applied Probability 27 (1990) 14-27.

[10] C. Lefevre, P. Picard, Distribution of the final extent of a rumour process, Journal of Applied Probability 31 (1994) 244-249.

[11] J. Gu, W. Li, X. Cai, The effect of the forget-remember mechanism on spreading, European Physical Journal B 62 (2008) 247-255.

[12] A. Sudbury, The proportion of population never hearing a rumour, Journal of Applied Probability 22 (1985) 443-446.

[13] Brown, C., El Kahoui, M., Novotni, D., Weber, A: Algorithmic methods for investigating equilibria in epidemic modelling. Journ. Symb. Comp. 41, 1157–1163 (2006)