Research on virus transmission of online social network

Min Yang*, Yaoliang Song, Qianmu Li

School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China

Received 3 April 2014, www.tsi.lv

Abstract

Online social networks (OSN) are up-and-coming complex network systems. Experiments indicate that it is difficult for simple complex network theory to describe virus transmission behaviour. Based on comprehensive research into current virus transmission, this paper combines user behaviour with social engineering theory and builds a model of virus transmission on OSN. Key factors affecting virus transmission on OSN are then analysed. Lastly, in light of public opinion transmission theory, this paper refers to social reinforcement factors concepts to describe computer virus transmission on OSN and analyses transmission disciplines in regular and random networks.

Keywords: online social network, virus transmission, epidemic spreading

1 Introduction

The classical virus transmission model SI/SIR/SIS is frequently used to simulate the spreading process of viruses, which is through nodes. On social networks, the relationship between nodes depends on the on-line status. Besides, we are more likely to receive information from friends. So, whether from the macro- or microscopic angle, virus transmissions on OSN differ greatly from those on off-line.

Using research into on-line social networks, Centola investigated the transmission process of public opinion on various network structures [1]. The research argues that social reinforcement can be a great impact factor in terms of transmission of behaviours on networks. If an individual is subject to certain behaviours or opinions from multiple friends then, from a macro point of view, this behaviour or opinion will tend to propagate faster on networks. Experiments show that a single signal cannot influence the decision of a node. Only when the node receives more signals can it receive information or produce behaviours. Information and behaviours can propagate faster on a regular network with a high clustering coefficient than on a random network, because people can receive signals more easily from other nodes on a regular network with a high clustering coefficient. This article describes the research of Centola, considers the differences between virus transmission and information behaviour transmission, and builds the social network virus transmission SEIR model combining with communication opinion [2].

2 Model description

2.1 MECHANISM OF VIRUS SPREADING IN THE MODEL

According to the model described in this article, the virus transmission process is as follows: Firstly, some virus infected nodes exist on social networks. They deliver the signal with a virus to all nodes on the friends list. However, only some of the friend nodes of this node will be infected by the virus. Then the infected nodes will spread the virus signal to all their friend nodes on their friends list. Thus these signals can be links to the virus or the Auto-run file of the virus in real life.

The nodes in social networks can be described by these four states:

1) S status. S status indicates the node has not received any signals and it can be infected [3].

2) I status. I status indicates the node has been infected by a virus and it will spread the virus signal to infect other nodes.

3) R status. R status indicates this node recovers from I status. It develops immunity, and it may receive the virus signal but not be infected.

4) E status. E status indicates the node has received the virus, but it is not infected and it will not spread the virus. E status can describe the issues that users receive malice information, but not believe and be infected. During the spread process, the node which receives more signals is more likely to be infected [4]. Through this process, the virus achieves spread on the social networks. The model is shown in Figure 1.

Firstly, randomly select a node as "seed". The seed node must be at I status, other nodes at S status. The seed node deliver virus signals to other nodes. Then it will recover to R status and will never infect or spread the virus further.

* ***Corresponding author*** e-mail: ym8670435@126.com

At time *t*, if a node at status *S* or status *E* receives the virus signal, then the probability for it to transform to status I is λ_m m is the frequency the node receiving the signals of the virus, *m* and λ_m have a positive correlation. If the node does not receive any signal, the status will not change, irrespective of how many times the node received the signals before.

At time t, if the status of the node is I, during the next Δt, the node will deliver the virus to all its neighbours, and itself will transform to status *R*.

The changes of status of all the nodes are synchronous in the model, i.e., at time *t* all nodes assess their status at the next time simultaneously and make the changes during $t+\Delta_t$.

When there is no node to change status, transmission of the virus stops.

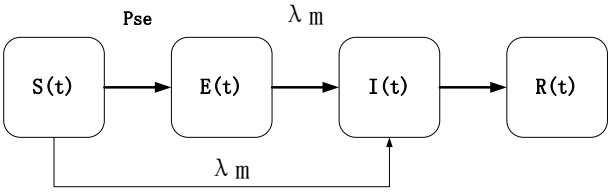


FIGURE 1 Mechanism of virus transmission in the model

The differences between the model provided by this article and other existing models are that we use the link between the nodes at most one time. For social networks, if one node delivers the information with the virus to its neighbour nodes many times, it will increase their vigilance. So, as social networks, few users will deliver the same information to their neighbours many times [5].

2.2 THE MATHEMATICAL MODEL OF THE VIRUS SPREADING PROCESS

When node *j* is at status *S*, it receives the virus signals the first time. We assume the probability of infection is λ_1, λ_1 is the initial spread rate. When node j is at status S, the rate of infection when it receives the virus signals the second time is λ_2, simply, when node j is at status *S*, the *m-th* time it receives the signals, the infected rate is λ_m, Here is the expression:

$$\begin{aligned}
&\lambda_1 = \lambda_1,\\
&\lambda_2 = \lambda_1 + \alpha(1-\lambda_1),\\
&\lambda_3 = \lambda_2 + \alpha(1-\lambda_2),\\
&\cdots\\
&\lambda_m = \lambda_{m-1} + \alpha(1-\lambda_{m-1}).
\end{aligned} \tag{1}$$

In the expression, $\alpha \in [0,1]$ it means social reinforcement factors, the bigger α, the higher and the rate at which the other node can be infected. From the Equation (1), we know the infection rate for a node after receiving *m-th* times is $\alpha(1-\lambda_{m-1})$ higher than if it receives (*m*–1)-*th* times. $\alpha(1-\lambda_{m-1})$ can be regarded as the rate when the node isn't infected at the (*m*–1)-*th* virus signal $1-\lambda_{m-1}$ transforming to I status influenced by social reinforcement factors α. There are two special situations in the Equation (1), when $\alpha = 1$, social reinforcement factors have a great influence, and when *m* >= 2, the rate of virus infection approaches 1. That means the node will be infected as soon as it receive the virus signals two times. When $\alpha = 0$, it means social reinforcement factors have no influence on the model. That is, the model provided by this article degenerates to the normal SIR model [6].

Simplifying and deforming the Equation (1) we can get the Equation (2):

$$\lambda_m = \begin{cases} \lambda_1 & \alpha = 1, m = 1 \\ 1 & \alpha = 1, m \geq 2 \\ 1-(1-\lambda_1)(1-\alpha)^{m-1} & 0 \leq \alpha < 1, m \geq 1 \end{cases}. \tag{2}$$

In the Equation (2), when α and λ_1 are fixed, λ_m will increase monotonically by *m*. That is, a node in social networks is more likely to be infected if it receives virus signals more times. As α increases, so does λ_m. In Figure 2, we describe the relationship among, *m* and α when $\lambda_1 = 0.2$.

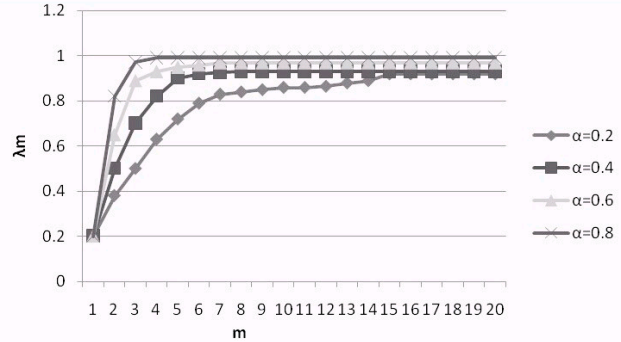


FIGURE 2 the relationship between λ_m, m and α

During the spreading process, a node turns from status *S* to status *E*, meaning the node received the virus signals but is not infected and will not spread the virus. The rate P_{SE} is decided by each neighbouring node.

$$P_{SE} = 1 - \prod_{c \in N_j} \prod_{n=1}^{m-1} (1-\lambda_n). \tag{3}$$

Node *i* is the neighbour of node *j*. $\prod_{c \in N_j} \prod_{n=1}^{m-1} (1-\lambda_n)$ means at time *m*–1, the neighbour of node *j* still has not attained the rate of status *S*. which are the neighbours of node *j*, the *j* node must receive the virus signals next time.

Combining Figure 1 with the description of the mechanism of the model, we can get the node status changing the sum formula:

$$S(t+\Delta t)=S(t)(1-\lambda_m), \quad (4)$$

$$E(t+\Delta t)=S(t)\left(1-\prod_{c\in N_j}\prod_{n=1}^{m-1}(1-\lambda_n)\right)+E(t)(1-\lambda_m), \quad (5)$$

$$I(t+\Delta t)=S(t)\lambda_m+E(t)\lambda_m, \quad (6)$$

$$R(t+\Delta t)=I(t). \quad (7)$$

3 Simulation experiment and analysis

3.1 EXPERIMENTAL NETWORK SELECTION AND CONSTRUCTION

The regular network that simulation experiments mainly select is the Moore network, with the network boundary conditions periodic. The network structure is shown in Figure 3:

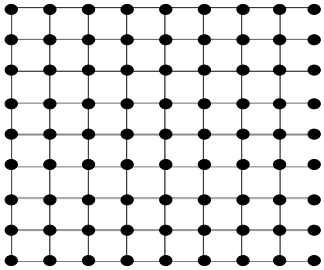


FIGURE 3 Nodal neighbour structure in the Moore network

Referring to experimentation in which Centola researched the dissemination of public behaviour in the network, in the simulation experiments in this paper, we select the network structure in which the length and width ratio is two. It is a uniform random network using the Maslov-Sneppen small world model [1]. The process of setting it up is as follows:

1) At time *t*, we randomly select a pair of edges A-B, C-D, then we reconnect this pair of edges to A-D, B-C. In the process of reconnection, we do not allow self-connection and reconnection.

2) In order to establish a completely random network structure, the reconnection process must be repeated many times. In the experiment in this article, the number of repeating the above process of selection is pN_E. In pN_E, p can be used to describe the randomness of constructed random networks, N_E is the number of connections in random networks.

3) Strictly speaking, only when $p\to\infty$, it is a random network really but as the literature shows simulations of topological information and true random network are very approximate when $p>1$, so in this paper $p=10$.

The main tool for virus transmission simulation experiments on a regular Moore network is cellular automata, so we set up a cellular automata model *T(C,Q,V,f)*: Cellular space is selected as a quadrangle cellular space; Cellular discrete state is combined as *Q={S, E, I, R}*; Cellular neighbour set's selection is as shown in Figure 1, the cellular boundary's selection cycle type; Cellular transformation rules as shown in Equation (7).

3.2 DESIGN OF SIMLATION EXPERIMENT AND ANALYSIS OF THE RESULTS

3.2.1 Experiment about the impact on the experimental results and analysis of different network topologies

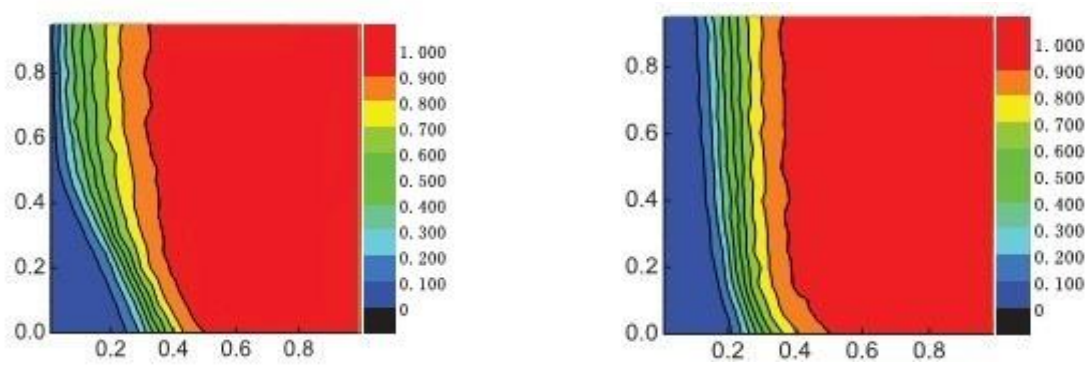
Let $\rho_t=\frac{I_t}{N_t}$, ρ_t represent the proportion of nodes whose state is I to the total nodes in the network space at time *t*, $\rho_{\infty reg}$ and $\rho_{\infty rand}$ respectively represent the proportions of nodes whose state is *I* to the total nodes in the regular networks and random networks in the final stable state. In order to compare the spread of the virus in the two different network topological structures, we define $\delta_p=\rho_{\infty reg}-\rho_{\infty rand}$. If $\delta_p>0$, the virus spread in the dissemination of a wider range in a regular network than in a random network; if $\delta_p<0$, the spread of the virus is faster in the random network.

Figure 4 shows the relation diagram of α, λ_1 and ρ_t in the regular networks and random networks. In Figure 4, the longitudinal coordinates are α the abscissa is λ_1, different colors represent different ρ_∞.

In Figure 5, when the λ_1 = 0.1, 0.2, ..., 1, we take different social enhancement factor α's values of ρ_∞; it will help us more clearly observe the change trend of ρ_∞ and provide a supporting role to explain Figure 4.

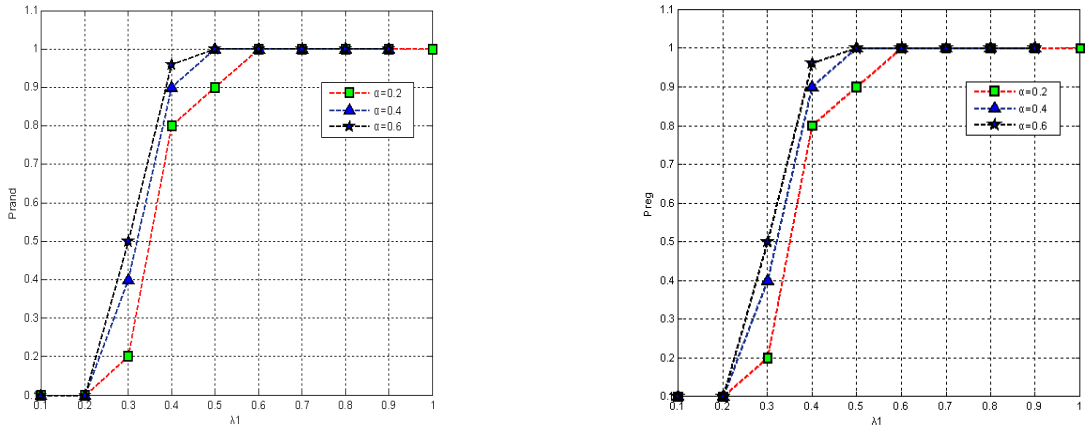
In order to more intuitively represent the spread of the virus in two different kind of networks, we use along with the changes of α and λ_1, and we use Matlab visual effects to draw Figure 6. In Figure 6 the longitudinal coordinates are α, the abscissa is λ_1, different colours represent different δ_p.

As shown as Figure 4, when α is larger but λ_1 is smaller, the range of the spread of the virus is bigger in a regular network than in a random network. However, when α is smaller and λ_1 is larger (about 0.3), the range of the spread of the virus is larger in a random network than in a regular network.



(a) (b)

FIGURE 4 ρ_∞ in the regular networks and random networks



(a) (b)

FIGURE 5 ρ_∞ in the regular networks and random networks when λ_1 and α are given

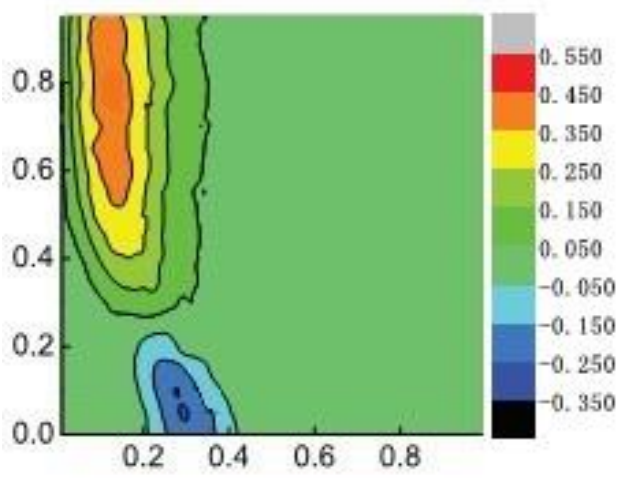


FIGURE 6 δ_p in two different kinds of networks

Figure 6 shows the distribution of δ_ρ in regular networks and random networks in the space (λ_1, α). In Figure 6, we can clearly see that there are two isolated "island" shapes in this figure: one is that when λ_1 is small and α is big, the scope of the spread of the virus is much greater in a regular network, which is consistent with Centola's experimental results. The other is when λ_1 is relatively larger and α relatively smaller, the scope of the spread of the virus is much greater in random networks than in regular networks. This is fully consistent with the conclusion of previous researchers: viruses spread faster in random networks than in regular networks. δ_ρ has a gap except in two ranges, the other times the value of δ_ρ is zero.

Analysis of two areas is worthwhile, the first region is located in the top right corner of Figure 6 areas: When λ_1 is very large, regardless of whether social reinforcement factor α is big or small, the virus will spread almost throughout all the network range, the expression of experiment shown that there have not any basic difference of virus between regular networks and random networks; the second region is located in the lower left corner of Figure 5 areas: when λ_1 is very small and social enhancement factor α is very small, the virus in the two kinds of networks are not spread very well apart, in Figure 4 we can see that in the steady state *I* node occupies a small proportion of all nodes in the network in 0.1.

3.2.2 Experiment about evolution of I state nodes in the model

In order to illustrate the effectiveness of the models, in Figure 7 we give the change curve of the infected nodes' amount with the curve of time in regular networks and random networks, and in Figure 8 we present the experimental results of Centola. In Figure 8, the black solid circular and hollow triangle respectively represent the number of individuals receiving public opinion behaviour dissemination in regular networks and random networks.

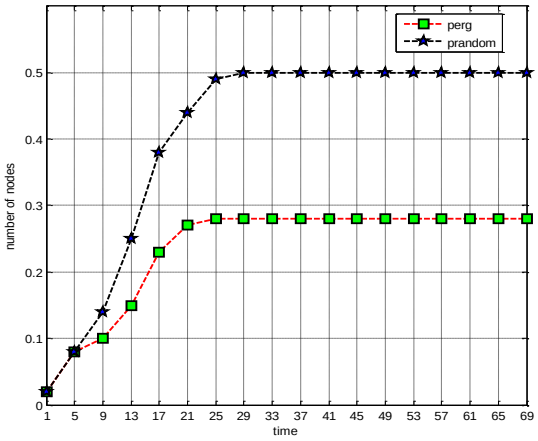


FIGURE 7 The change curve of the infected nodes' amount with the curve of time

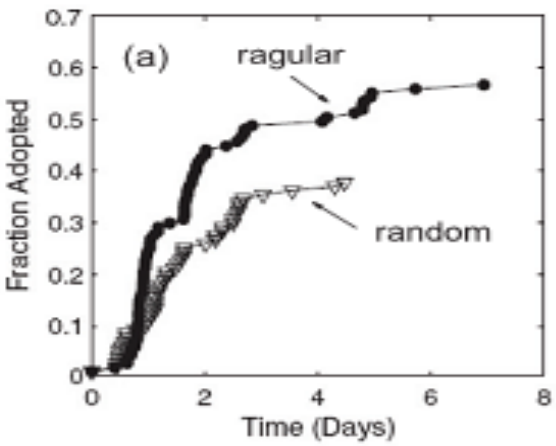


FIGURE 8 The experimental results of Centola

The parameters in Figure 7 is that the value of λ_1 is 0.18 and the value of α. By comparing Figures 7 with 8, we can find that from the preliminary transmission rate and the final steady-state communication range, regular networks spread faster and wider than random networks. In order to describe each time *t* infected nodes' density, we use $\rho_t - \rho_{t-1}$ to describe the increased amount of the *I* node's density in regular networks and random networks in each time. Thus we can draw Figure 9:

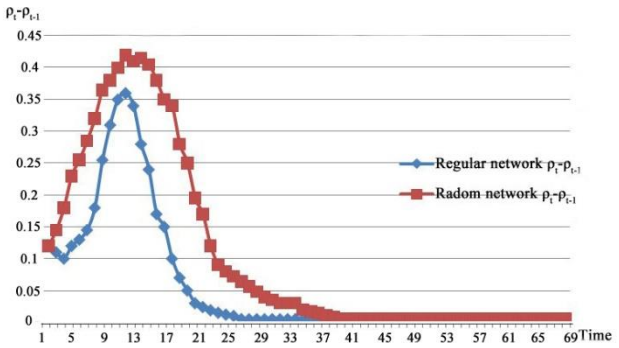


FIGURE 9 The change of $\rho_t - \rho_{t-1}$ in regular network and random networks

In Figure 9 we can see that when virus transmission process begins, the rate of spread of the virus in the two networks grows very fast. After the two curves are almost simultaneously at a peak of their own, this corresponds to virus outbreak events in the real world. After reaching the peak value, the two curves begin to decline rapidly with the passage of time. In the propagating process of the virus, $\rho_t - \rho_{t-1}$ is always bigger in regular networks than in random networks, which means that the transmission rate of the virus is bigger in regular networks than in random networks. The value of $\rho_t - \rho_{t-1}$ is zero, meaning that with the termination of the propagating process, the range of transmission of the virus reaches the maximum value.

3.2.3 Experiment about the critical value of the social enhancement factor when the virus spreads in the social network

In the propagating process of the virus, α plays an important role in the transmission process of the virus in the social network. In Experiment 2, when α is equal to 0.52 and λ_1 is equal to 0.14, finally there is a δ_ρ difference of about 0.35. Therefore, an issue emerges: what is the social enhancement factor's influence in the spread of the virus, namely, how many does a node need to receive a signal to be infected?

Parameters of Experiment 2 were selected as follows: λ_1 is 0.18, α is 0.4. The model is simulated by cellular automata, degree of nodes is four. The cellular space is 1000, the experiment is carried out 10 times, the statistic selection for the network reaches the steady state. Each cell has a count value to calculate the number of signals of the virus which are to be received to change into the *R* state, according to the rules of virus transmission, count subtract one to calculate the number of signals of the virus which are to be received to change into the *I* state [7].

We define P_m as the probability of being infected by the virus after the user receives the virus's signal m times. The P_m statistics are as follows:

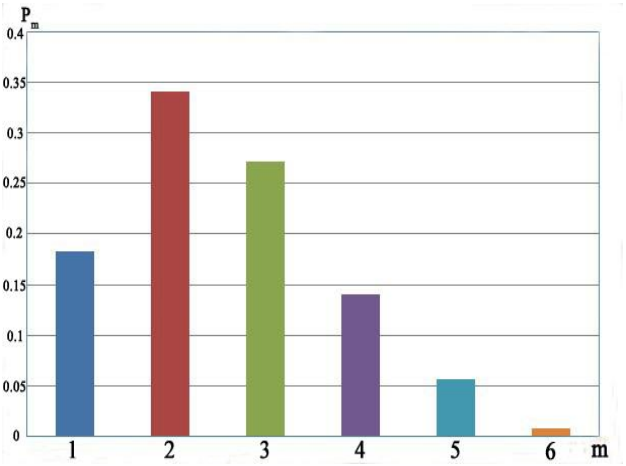


FIGURE 10 P_m statistics

As shown in Figure 10, only about 17% of people after they receive the virus's signal only once will accept and become infected with the virus. When the node receives two signals of the virus, more than 30% of people will choose to believe the information and will become infected with the virus. Although when *m* is equal to 3 or 4, the degree of users' adoption of virus information is very high, the experiment proved that the second virus signal is the most important. This conclusion is verified by the experiment of Centola [9].

4 Verify and description of the validity of the model

The first experiment forms the basis for all other tests. In Experiment 1, we tested the relationship between λ_1 and ρ_t in the case of different topologies and different values of α In Figures 3 and 4, what are shown are the values of $\rho_{\infty reg}$ and $\rho_{\infty rand}$ in the hexagonal network topology. There are two examples to prove the validity of the model presented in this paper: One is under the condition of small λ_1 and large α the scope of transmission of the virus is large in the regular network, which is consistent with the conclusions of Centola experiments; another is under the condition of relatively large λ_1 and smaller α, the scope of transmission of the virus in the random network is wider than the one in the regular network. This is in line with the research conclusions of previous scholars: A virus travels faster in a random network than in a regular network [10].

In Experiment 2, we compared the experimental results with Centola. In addition to good description of the spread of the virus on social networks, this model can also describe well the Centola experiment when $\lambda_1 = 0.18$, $\alpha = 0.4t$, which confirms the validity of this model from the side.

In Experiment 3, we obtained statistical information on the social reinforcement factor parameter that makes users infected with the virus. By averaging the testing values of repeated measurements, we can infer that the second time users receive the virus information is crucial for whether or not the node is infected with the virus, as the same point is also verified in the Centola experiment.

5. Conclusion

This paper combines innovatively with social public opinion communication and establishes a SEIR virus spread on a social network model. Using Matlab and quadrilateral cellular space and periodic boundary conditions of cellular automata as a tool, the model was put forward by the simulation experiment. In the experiments we first established regular and uniform random networks, then studied the effects of network topology, the social reinforcement factor parameters α, and the initial virus infection rate λ_1 on the process of the spread of the virus.

Studies have shown that in the process of the spread of the virus in a social network, social reinforcement factor α and the initial transmission rate λ_1 played a very important role. The main conclusions and results of the model are as follows:

1) Even when the initial transmission rate λ_1 is very small, the virus can still be spread on the regular network if social reinforcement factor α relatively large, but on the same condition, the virus cannot be spread extensively on uniform random networks. That is to say, the virus travels faster and wider in the random network than in the regular network. This conclusion supports the results of the Centola experiment in certain cases. When social reinforcement factor α is 0.4 and the initial transmission rate λ_1 is 0.18, the model proposed in this paper can be used to simulate the Centola experimental network.

2) When the initial transmission rate λ_1 is bigger, social reinforcement factor α is smaller, the spread of the virus travels faster and wider in the uniform random network than in the regular network. This conclusion is consistent with the traditional conclusion of virus spread on the network.

3) When the initial transmission rate λ_1 is very large, all nodes in the network have a high probability of infection no matter what social reinforcement factor α is. As a result, the virus spreads quickly to the whole social network, and the spread of the virus has nothing to do with the network structure.

4) When social reinforcement factor $\alpha = 0$ the proposed model can be degraded as the standard SIR model for the spread of the virus in complex networks.

References

[1] Zou C, Gong W, Towsley D 2007 *IEEE Transactions on Dependable and Secure Computing* **4**(2) 105-18
[2] Anderson R., May R M 2002 Infectious Diseases in Humans *Oxford University Press*: Oxford 56-77
[3] Yang M, Li Q, Song Y 2014 A SEIR Model Epidemic of Virus on the Online Social Network *Journal of Digital Information Management* **12**(2) 102-7
[4] Jin C, Liu J 2009 Network virus transmission model based on effects of removing time and user vigilance *International Journal of Network Security* **9**(2) 156-63
[5] Li Q, Zhang H 2012 Information Security Risk Assessment Technology of Cyberspace: A Review Information *An International Interdisciplinary Journal* **15**(11) 4677-84
[6] Li Q, Li J 2007 Rough Outlier Detection Based Security Risk Analysis Methodology *China Communications* **5**(7) 14-21
[7] Maslov S, Sneppen K 2002 Specificity and Stability in Topology of Protein Networks *Science* **296** 910-3
[8] Chen Z, Ji C 2005 *IEEE Transactions on Neural Networks* **16**(5) 1291-303
[9] Balthrop J, Forrest S, Newman M E J 2004 Computer Science: Technological Networks and the Spread of Computer Viruses *Science* **304**(5670) 527-9

Authors

Min Yang, born in April, 1978, Nanjing, Jiangsu Province, P. R. China

University studies: Ph.D at Nanjing University of Science and Technology.
Scientific interest: Computer and electronic engineering.
Experience: Teaching experience of 13 years.

Yongliang Song, born in June, 1960, Nanjing, Jiangsu Province, P. R. China

University studies: Ph.D at Nanjing University of Science and Technology.
Scientific interest: Communication and information systems.
Experience: Teaching experience of 30 years, 8 scientific research projects.

Qianmu Li, born in January, 1979, Nanjing, Jiangsu Province, P. R. China

University studies: Ph.D at Nanjing University of Science and Technology
Scientific interest: Computer applications
Experience: Teaching experience of 10 years, 10 scientific research projects