Physica A 390 (2011) 189–197

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Online social networks—Paradise of computer viruses[☆]

W. Fan [*], K.H. Yeung

City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 28 April 2010
Received in revised form 6 September 2010
Available online 29 October 2010

Keywords:
Virus propagation
Online social networks
Complex network

ABSTRACT

Online social network services have attracted more and more users in recent years. So the security of social networks becomes a critical problem. In this paper, we propose a virus propagation model based on the application network of Facebook, which is the most popular among these social network service providers. We also study the virus propagation with an email virus model and compare the behaviors of a virus spreading on Facebook with the original email network. It is found that Facebook provides the same chance for a virus spreading while it gives a platform for application developers. And a virus will spread faster in the Facebook network if users of Facebook spend more time on it.

1. Introduction

Computer viruses can spread through the Internet in many ways, such as email, instant messengers (IM), and Peer-to-Peer (P2P) file sharing. Currently, as online social network services (SNS) become popular, the users of this service turn into the targets of virus writers. People communicate and share files with their friends on these platforms, and they also take part in some activities or join groups online. These characteristics give hackers opportunities to attack the users. Virus spreading in an SNS is similar to that in an email network or an instant messenger network. All of them can spread a virus through sending or sharing files which contain malicious codes. If a user of these networks gets infected, the infected account will automatically send the same email or file to the people in the contact list of this user. As a result, the virus can spread quickly.

There have been some models to simulate the virus propagation in those networks. C. Zou et al. described their email virus propagation model in Ref. [1–3]. They assumed that the virus spreads through an email network by containing viruses in email attachments. Users' email checking time intervals and the probabilities of users opening these attachments were referred in this model. And it was found that as users' email checking time becomes more variable, the virus spreads faster. In Ref. [4,5], Komninos et al. proposed a worm propagation model for email, IM and Peer-to-Peer (P2P) networks. They considered that as time grows, users' behaviors should change. So the probability that a user opens an attachment is not fixed at a certain value, but will decrease as time goes, which is different from Zou's model. The model of virus propagation in P2P networks in Ref. [6] assumed that the probability of a user to download an infected file is relevant to the ratio of number of infected files to total files. And this ratio can be affected by users' downloading and executing behavior.

However, the models of virus propagation in email or IM networks are not suitable for that in an SNS networks. Besides sending messages like using email, users of SNS networks can upload files to their accounts. The activities that a user takes will appear in his/her friends' news feed, so all the friends can read the news when they are online. While different from email or IM networks, some people use SNS for entertainment, so plenty of them spend hours on SNS every day. It accelerates virus propagation. As the behavior of SNS users is more complex than that of other networks, it is necessary to construct a new model for virus propagation in SNS networks. Recently, it has been reported that some SNS websites, such as Facebook

[☆] This work was supported by Hong Kong Government General Research Fund (grant No. CityU-123608).

[*] Corresponding author. Tel.: +852 92796694.
E-mail address: aaelaine@gmail.com (W. Fan).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.09.034

and MySpace, were attacked by hackers [7,8]. Among SNS providers, Facebook interests the largest population and owns an application platform, so we choose Facebook to analyze its characteristics and model the virus propagation in it.

In this paper, we propose two models for virus propagation on Facebook. One is based on the Facebook's application platform. Hackers may utilize this platform to post applications with viruses. As will be reported later, malicious applications spread through the network faster than normal applications which have the same initial conditions. Another model is based on sending messages to friends. It is a traditional way, just like sending email with malicious attachments. Actually, hackers can post pictures or links that contain Trojans. For simplicity, we assume that these methods are similar and we will describe them as sending messages. As will be reported later from our simulation results, we find that as some people take Facebook as an entertainment tool, a virus will therefore spread faster than that in an email network. We also find that, the parameters of the network model can affect the propagation of a virus significantly.

2. Facebook user network topology

We define the users of Facebook as a network with N nodes in this paper. Users are nodes in the network and each node is assigned a number i, $i = 1, 2, \ldots, N$. An edge between two nodes i and j means these two users are friends. In Facebook, user i becomes a friend of user j with their names in each other's friends list, so this network is undirected. We also define that the node degree is the number of friends a user has. Some results have showed that email networks have scale-free topology [9,10]. And recently, some researchers have studied the structures of online social network topologies, such as MySpace, orkut, cyworld, and so on [11–13]. They found that all of these networks have power-law degree distributions. The degree distributions can be described as $P(k) \sim k^{-\gamma}$. Here the probability that a node connects to k nodes is $P(k)$, and $\gamma > 0$. Most nodes of these networks have low degrees while a few nodes have many connections. This structure has been demonstrated to be vulnerable and the well connected nodes are crucial in epidemic spreading [14–17].

In our simulations, we assume that the nodes' degrees of users' network of Facebook exhibit the power-law distribution, because it is also one of the online social networks. And we construct the network with two models, Barabasi–Albert scale-free network model (BA model) [18], and Generalized Linear Preference model (GLP model) [19]. In the first model, nodes are added to the network continuously, and $\gamma = 3$. $\gamma = 3$ is a constant value in the BA scale-free model, but most researched online social networks have $\gamma < 3$. Therefore we also use the GLP network in our simulations. In the second one, nodes are added with probability $1 - p$ and additional edges are added with probability p. Then we can obtain $\gamma < 3$. In both of them, ends of edges are chosen with preferential attachment, so the nodes with higher degrees will get more connections. In the following simulations, both of our virus propagation models are studied based on these two scale-free network models.

3. A model based on the Facebook application platform

One of Facebook's successes is its application platform. By using the platform, companies and individuals can develop third-party applications. Users of Facebook can add these applications to their accounts. More than 95% of users have used at least one application built on the Facebook Platform. And everyday there are 140 new applications added [20]. However, it has been reported that some new applications became available with a virus [8]. Unlike the spreading of normal applications, if a user installs this kind of application, his/her account is infected. Then fake messages are forwarded to all his/her friends to persuade them to install the same application, which will increase the probability of installation. Considering the great number of daily installations, it is necessary to construct a virus propagation model based on these third-party applications. It has been shown that the installations of applications have a preferential characteristic [21]. That means an application with greater number of users attracts more new users. Moreover, a user who has installed more applications has a higher possibility to install new applications. The distribution of the number of installations of applications is shown in Fig. 1. The data is obtained from Adonomics [22].

Gjoka developed a model to simulate the user coverage of Facebook applications [14]. With the input of the list of applications, the number of installations per application and the number of users, this model generates a graph which shows the power-law distribution of users' installations. This model is helpful for us to model the existing number of installations per user, but it cannot reflect the behavior when a user encounters an application invitation. In this paper we proposed a Facebook virus propagation model based on a third-party application, which not only follows the spreading law of normal applications, but also contains the characteristics of virus spreading.

Our Facebook user network has N_{user} nodes, and each node is assigned a number i, $i = 1, 2, \ldots, N_{user}$. We assume that the number of available applications is N_{app} and each one is denoted by k, $k = 1, 2, \ldots, N_{app}$. Each application has a number of installations to show how many users have installed it. Because there are new installations every day, the number of installations per application is not a fixed value. It is defined that the number of installations of application k at time step t is $Install_k(t)$, $k = 1, 2, \ldots, N_{app}$. From the data of Adonomics we can get to know of the initial number of installations per application before the virus begins attacking. But we will run our model in a network much smaller than the real Facebook network, so we do not use the original data of installations. We scale down the network size N_{user}, the total number of applications N_{app}, and the initial number of installations per application $Install_k(t_0)$. So we can create a list of applications and assign $Install_k(t_0)$ for each application. The distribution of $Install_k(t_0)$ is similar to the curve in Fig. 1. Next we can model the behaviors of applications as described below.

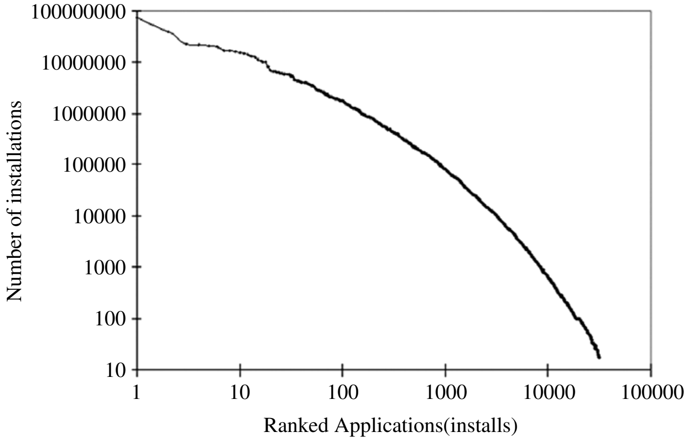


Fig. 1. The distribution of number of installations of each application.

The first step is to construct the initial number of installations per user: With the list of existing applications and $Install_k(t_0)$, $k = 1, 2, \ldots, N_{app}$, we can construct the initial number of installations per user using the model in Ref. [14]. In the beginning, we have $Install_k(t_0)$, but all the users in our network have not installed any application. Then at every step, each installation of $Install_k(t_0)$ of all the N_{app} applications is assigned with a probability to one of the users. The probability of one installation to be installed by user i is

$$P_{user}(i, t) = \frac{Apps_i(t)^\rho + init_{user}}{\sum_{j=1}^{N_{user}} (Apps_j(t)^\rho + init_{user})}. \tag{1}$$

Here $Apps_i(t)$ is the number of applications that user i has installed at time step t. The parameter ρ reflects the effect of preferential installation. $init_{user}$ is used to show the initial probability $P_{user}(i, t)$ of a user who does not install any application. That is, if $Apps_i(t) = 0$, the initial probability of user i is $\frac{init_{user}}{\sum_{j=1}^{N_{user}} (Apps_j(t)^\rho + init_{user})}$. In our simulation we have $\rho = 1$, $init_{user} = 1$. The value of $init_{user}$ and ρ are both assumed values. We only focus on the preferential installation. The values we use are of the simplest case. In Ref. [14] researchers found that the number of installations per user has a power-law distribution. That means a user who installs more applications has more chances to be selected to install other applications in our simulations. So we use preferential selection in this equation. This step is repeated until all $Install_k(t_0)$ installations of all N_{app} applications are exhausted. So the initialization is completed. After we finish the initialization step, $Apps_i(t_0)$ for each user i can be obtained. And the value of $Apps_i(t)$ will increase if user i installs new applications in simulations.

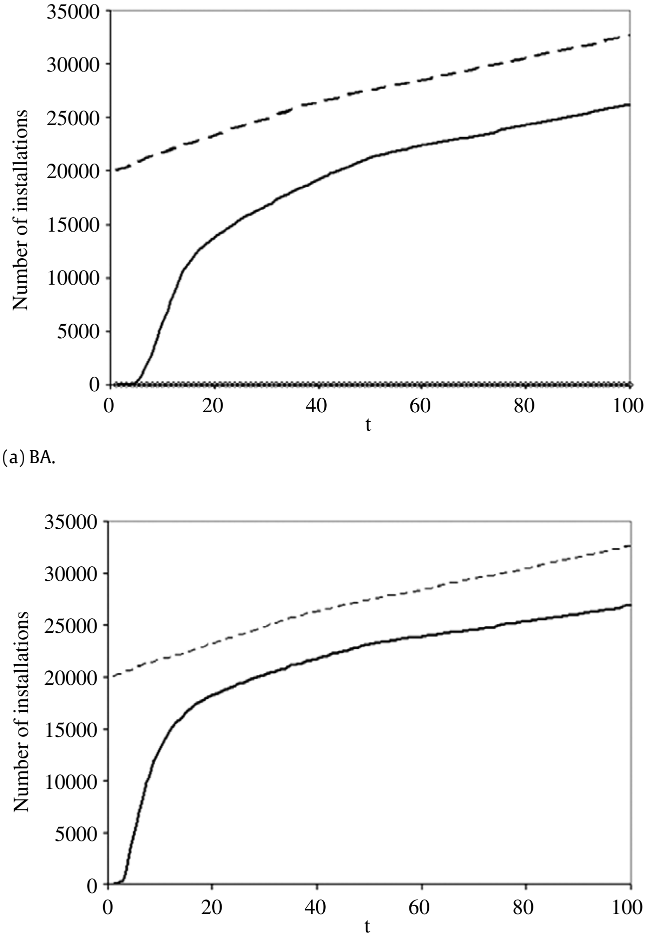
Virus propagation follows the steps below:

(a) Select the users who are infected in the beginning: The virus spreading starts from I_0 infected users at time step t_0. We randomly pick the I_0 infected users from the network. Now the total number of applications N_{app} is added by 1 and the order number of the malicious application is N_{app}. And $I(t_0) = I_0$, here $I(t)$ is the number of infected users at tth time step. They will send invitation messages to their friends.

(b) Maintain the installations' distribution: The statistics of Facebook shows that there are many new installations every day, but the curve showed in Fig. 1 does not change significantly. The distributions of installations of applications are similar from time to time. So we should maintain the preferential characteristic of installations. From Adonomics we know that millions of installations are taken every day, so we assume that in our simulations the number of new installation per day is m. The value of m is scaled down due to the small size of our network. We select one application from the application list, the probability of application k being selected is

$$P_{app}(k, t) = \frac{Install_k(t) + init_{app}}{\sum_{j=1}^{N_{app}} (Install_j(t) + init_{app})}. \tag{2}$$

Here $init_{app}$ defines the initial probability $P_{app}(k, t)$ of an application without any installation. In our simulation we have $init_{app} = 1$. Then this selected application is installed by a user i who is selected with $P_{user}(i, t)$. If user i has installed this application, we will pick other users. Fig. 1 in this paper shows that the number of installations per application also has a power-law distribution. Therefore, this equation in our model uses preferential selection, as well. It means that an application with more installations has a higher probability to be selected to be installed by users. We select an application



(a) BA.

(b) GLP.

Fig. 2. The behavior of three applications over time. Figure (a) is for the BA model while figure (b) is for the GLP model. In this simulation, $N = 50000$, $N_{app} = 100$ before the virus spreads, $m = 1000$ and $I_0 = 10$.

and assign it to a user for m times in this step. So we have m new installations. And if the malicious application is installed, we change the value of $I(t)$, and the infected user sends invitations to his/her friends.

(c) Users deal with the invitations: Every user who has received c invitation(s) at time step t will install the malicious application with the probability $P_{virus} = \frac{\sigma}{(1-\frac{Install_{N_{app}}(t)}{N_{user}}\cdot\frac{Apps_i(t)}{N_{app}})^c}$. In the third equation, the numerator is the real data we obtained from an application. We found an application which can show which user has accepted our invitations. About 5% of the users we invited accepted the invitations. In our simulation we have $\sigma = 0.05$. And we use the denominator to reflect the number of installations' influence on the probability of accepting invitations. The reason is similar to that of the first two Eqs. (1), (2): the increments of installations of application/user can raise the possibility that a user accepts invitations. $I(t)$ will be changed if necessary. From the behavior of the virus, it is reasonable that if a user receives more invitation messages, he/she has a higher risk to be infected.

(d) Repeat step (b) and (c) for the next time step t.

In this simulation, we record $I(t)$ at each time step t to show the virus spreading process in the network and the number of infected users in the end. In Fig. 2 we plot the behavior of the malicious application, as the solid lines in both figures show. The dashed lines show the behavior of the application which attracts the most users. In each figure, the number of installations of the malicious application increases rapidly. Then its increment rate is similar to that of the top application. That is because the step (c) helps the virus spread. But after the number of infected users reaches a certain value, its growth mainly comes from the step (b). So we can see from each figure that the growth rate is the same as that of the top application after $t = 50$. And in Fig. 2(a), we also plot the behavior of the application which has the same number of installations as the malicious one at $t = t_0 = 1$, as the diamond points show. This implies that the number of installations of the application which has the same initial condition does not change obviously.

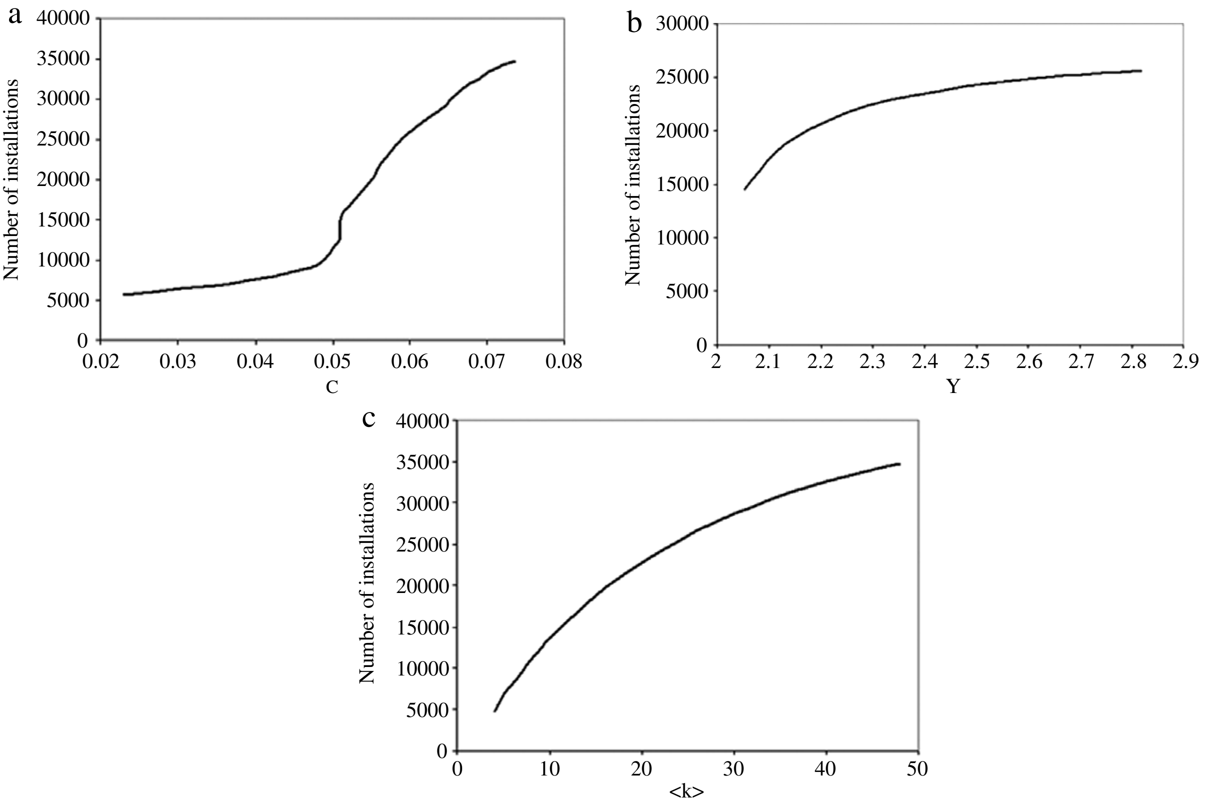


Fig. 3. The behavior of malicious application in networks with different parameters. The number of installations is taken at time step 100. In this simulation $m = 500$.

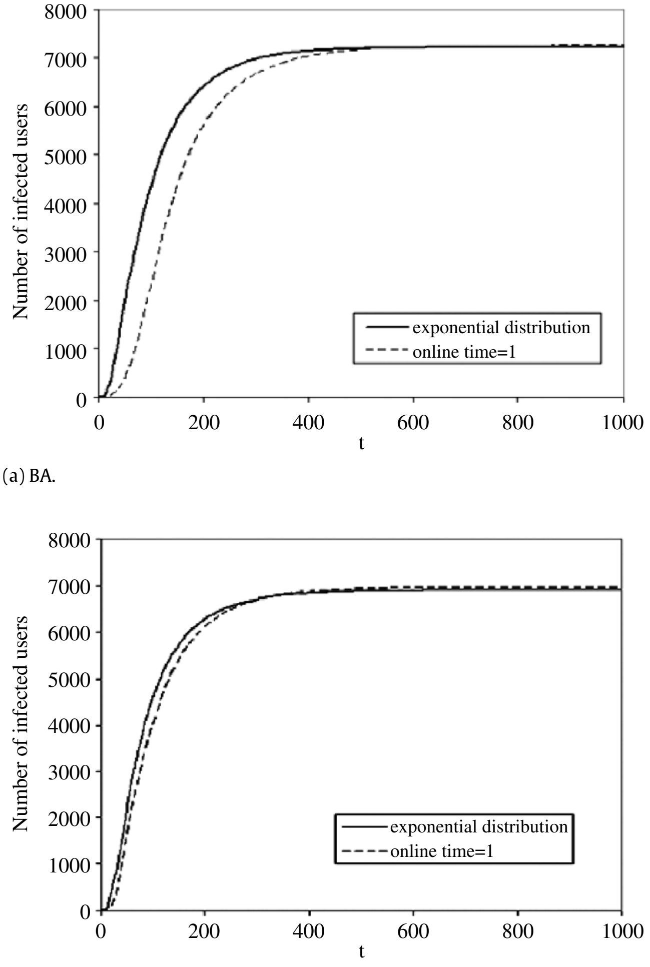
Behaviors of users can affect their friends' decisions. In the GLP scale-free network we can change some parameters of the network. In this work the effect of the clustering coefficient c, the power law exponent γ, and the average user's degree $\langle k \rangle$ are investigated. Fig. 3(a) plots the number of installations of the malicious application in the network with different clustering coefficients. Here larger c means that the network has more small-world characteristics and the friends of one user have more interactions. In this figure, more users are infected in a network with a larger clustering coefficient. Fig. 3(b) shows that the virus can infect more users in a more homogeneous network. In a scale-free network, as the exponent γ increases, the network becomes more homogeneous. And in Fig. 3(c), it is shown that the virus can spread faster in a network with greater $\langle k \rangle$, but not so obviously as Fig. 3(a). Therefore we can make a conclusion that the friendship between a user's friends helps the virus spread faster.

In our model, infected users send invitations to their friends. This influences the behaviors of users significantly as it is shown in the results. Users who are friends on Facebook may also be friends in real life, so they easily trust the invitations they receive. Then the virus can spread quickly even if there are only a few users installed it in the beginning. Our study also found that, the spreading rate of a virus is different in various networks.

4. A model based on sending messages

Another model which is based on sending messages is similar to email virus propagation. An email virus spreads by sending mails which contain malicious attachments to users. If users open these attachments, they will be infected. But on Facebook, users cannot add attachments to messages, so hackers try to lead users to third-party websites which urge users to download the virus [8]. The spreading process is like that: users log in to Facebook once in a while. When a user is online and receives a message with a link of a malicious website, he/she deletes this message without clicking this link or clicks it and then become infected. If a user is infected, the virus will send the same messages to all the friends of this user.

This process seems to be the same as the email virus [1]. Both of them depend on users' interactions. And users check their accounts with dynamic time intervals. However, they have differences. In this paper we only consider the case of Web mails. In normal cases, while using an email application, people only check that if there are new mails and then log out. But people spend more time on Facebook. Each month more than 700 billion minutes are spent on Facebook that has 500 million active users [20], which means on average every user spends more than 40 min on it per day. If Facebook users get new messages when they are online, they can check mailboxes immediately. As a result, besides the time between two



(a) BA.

(b) GLP.

Fig. 4. The behavior of a number of infected users. In this simulation, $N = 10000$, $N_{infect}(0) = 10$, and $\mu_{\gamma}(t) = \mu_0$. The data is averaged over 20 simulations.

log-in attempts, and the probability that a user clicks the direct malicious URL, we need the online time as another factor that affects virus propagation.

We still present the network as a scale-free network with N nodes. Users of Facebook are described as $i, i = 1, 2, \ldots, N$. In our model nodes have three kinds of status—susceptible, intermediate, and infected. In the beginning, all nodes are susceptible. If a node gets a message with a malicious link in the inbox but it is offline, this node becomes intermediate. An intermediate node can return to susceptible if it ignores this message, or become infected with a click probability. The three users interaction factors are as follows:

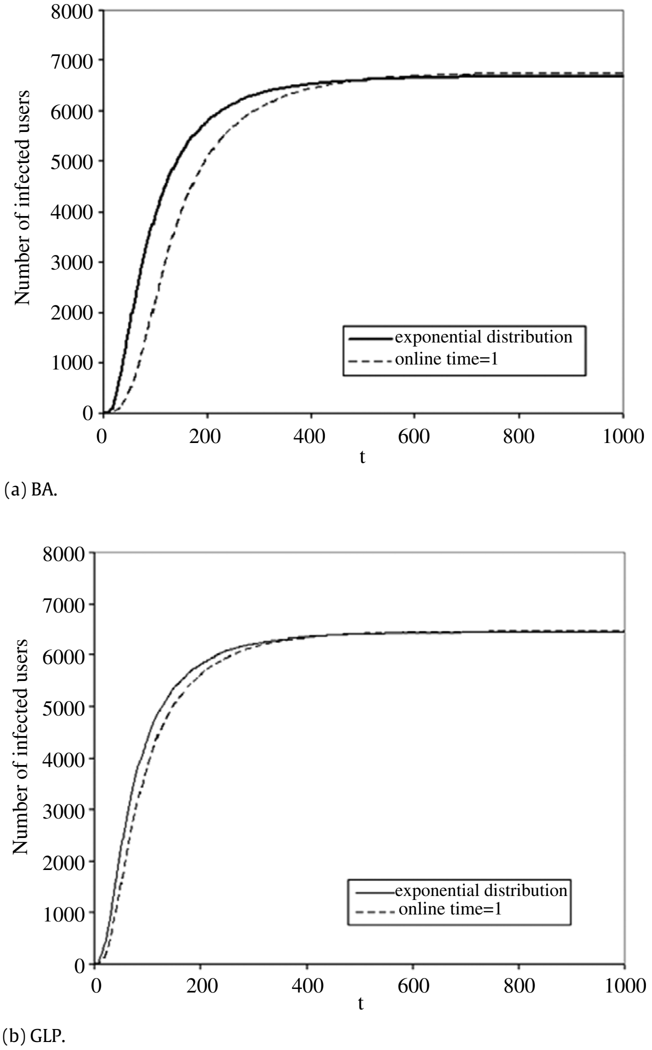
(1) Facebook log-in time $T_{login}(i)$ of node $i, i = 1, 2, \ldots, N$, follows an exponential distribution. Its mean $E[T_{login}(i)]$ is the independent Gaussian random variable $E[T_{login}] \sim N(\mu_{Tl}, \sigma^2_{Tl})$. In our simulation $E[T_{login}] \sim N(40, 400)$.

(2) Facebook online time $T_{online}(i)$ of node $i, i = 1, 2, \ldots, N$, is the independent Gaussian random variable $T_{online} \sim N(\mu_{To}, \sigma^2_{To})$. It is obvious that $T_{online}(i) < T_{login}(i)$. In our simulation $T_{online} \sim N(1, 100)$, and we assume that in the original email network $T_{online}(i) \equiv 1, i = 1, 2, \ldots, N$.

(3) The probability that a user i clicks a malicious link $P_{click}(i, t)$ is also an independent Gaussian random variable. We assume $P_{click}(t) \sim N(\mu_p(t), \sigma_p^2)$, in which $\mu_p(t)$ is monotonically decreasing with time, because more users are aware of this scam as the number of infected users increases. We have $\mu_P(t) = \mu_0(1 - \frac{N_{infect}(t)}{N})$, here μ_0 has a constant value and $N_{infect}(t)$is the number of infected users at time step t. In our simulation $\mu_0 = 0.5, \sigma_p^2 = 0.09$.

Virus propagation follows the steps below:

(a) The propagation begins at $t_0 = 1$. We randomly select $N_{infect}(t_0)$users who have malicious mails in their mailboxes in the beginning. These mails contain malicious links. If a user clicks these links, he/she will be infected and send the same mails to his/her friends.



(a) BA.

(b) GLP.

Fig. 5. The behavior of a number of infected users. In this simulation, $N = 10000$, $N_{infect}(0) = 10$, and $\mu_\gamma(t) = \mu_0(1 - \frac{N_{infect}(t)}{N})$. The data is averaged over 20 simulations.

(b) At each time step, we check all the users who log in to Facebook or are online at this step. If they have malicious mails in their mailboxes, they will click the links with $P_{click}(i)$. After a user i logs out, we will generate new $T_{login}(i)$ and $T_{online}(i)$ for this user's next log-in. All our simulations are of the non-reinfection case, that is, infected users will not send the virus to their friends twice if they click the malicious link again.

(c) Repeat step (b) for the next time step t.

4.1. Comparison of online time on a Facebook network

In this simulation, we record the number of infected users $N_{infect}(t)$ at each time step. As the solid lines show, Fig. 4 plots the $N_{infect}(t)$ of the Facebook network with a constant $\mu_p(t) \equiv \mu_0$ and $T_{online}(i)$, $(T_{online}(i) > 1)$. The dashed lines are both the number of infected users in a network with $T_{online}(i) \equiv 1$, which does not consider the user online time, or the changes of μ_p. By comparing the lines in both Fig. 4(a) and (b), we find that the virus will spread faster as some users spend lots of time on Facebook.

However, users may be aware of this virus as more and more users are infected. So the probability that they click the links would decrease, and the number of infected users will be different. We plot it in Fig. 5, in which the $\mu_p(t)$ changes over time. We find that the sizes of infected users are smaller than those in Fig. 4. In Fig. 5, the solid lines both rise faster than the dashed lines. The same results as Fig. 4 are obtained.

Our results indicate that with the same conditions, a virus spreads faster on the Facebook network if more time is spent on it. But if we assume that the probability of a user to click a malicious link is not constant, the size of infected users will be smaller.

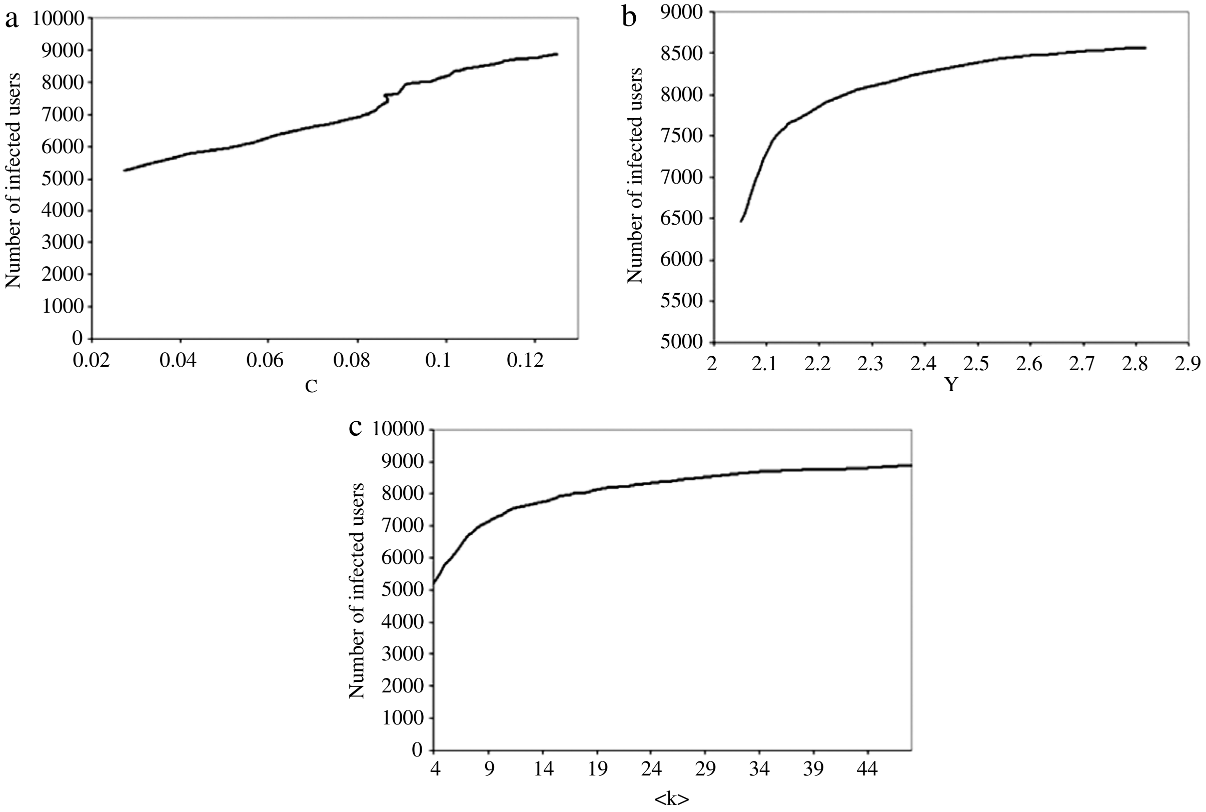


Fig. 6. The function of the final number of infected users and different parameters. In this simulation, $N = 10000$, $N_{infect}(0) = 10$, and $\mu_\gamma(t) = \mu_0(1 - \frac{N_{infect}(t)}{N})$. The data is averaged over 20 simulations.

4.2. Comparison of Facebook networks with different parameters

Virus' behaviors in GLP networks with different clustering coefficient c, power law exponent γ, and the average degree $\langle k \rangle$ are also compared. Fig. 6 plots these virus propagation processes. We also arrive at the same conclusion that in a network with greater c, γ, and $\langle k \rangle$, more users are infected. Fig. 6(a) shows that the final number of infected users increases linearly as a function of the clustering coefficient. While in Fig. 6(c), the curve is like a logarithmic function. So the conclusion is that the virus will spread to more users in a closer network.

5. Conclusion

In this paper, two models for virus propagation on the Facebook network were proposed. In the model based on the Facebook application platform, if a virus is installed by more users, it will become more popular and cause more installations. The result of our simulations showed that, even if the malicious application attracts only a few users in the beginning, it can still spread rapidly. That is because users may trust their friends on Facebook and install the malicious application. While in the second model, which is similar to email virus propagation, the probability that a user is infected becomes smaller as more and more users get infected. We found that the virus spreads faster on Facebook than in the original email network, as people use Facebook for entertainment and spend more time on it. And the network generation models and their parameters can affect the behaviors of virus. The simulation results of our two virus propagation models showed that, a virus would spread faster in a network with a greater clustering coefficient, power-law exponent or average degree.

These years more and more people use online social networks for entertainment or communication. Social networks will become a paradise of computer viruses if these service providers do not check the behaviors of their applications.

References

[1] C.C. Zou, D. Towsley, W. Gong, Email virus propagation modeling and analysis, Umass ECE Dept., Tech. Rep. TR-03-CSE-04, May (2003).

[2] C.C. Zou, D. Towsley, W. Gong, Email worm modeling and defense, in: Proc. 13th Int. Conf. Computer Communications and Networks, ICCCN'04, 2004, pp. 409–414.

[3] C.C. Zou, D. Towsley, W. Gong, Modeling and simulation study of the propagation and defense of Internet e-mail worms, IEEE Trans. Dependable Secure Comput. 4 (2) (2007) 105–118.

[4] T. Komninos, Y.C. Stamatiou, G. Vavitsas, A worm propagation model based on people's email acquaintance profiles, in: Wine, Patras, Greece, 2006.
[5] T. Komninos, P. Spirakis, Y.C. Stamatiou, G. Vavitsas, A worm propagation model based on scale free network structures and people's email acquaintance profiles, IJCSNS 7 (2) (2007) February.
[6] R.W. Thommes, M.J. Coates, Modeling virus propagation in peer to peer networks, Information, Communications and Signal Processing, in: 2005 Fifth International Conference, 981–985.
[7] http://news.bbc.co.uk/2/hi/technology/7918839.stm.
[8] http://www.kaspersky.com/news?id=207575670.
[9] H. Ebel, L.-I. Mielsch, S. Bornholdt, Scale-free topology of e-mail networks, Phys. Rev. E 66 (2002) 035103 (R).
[10] M.E.J. Newman, Stephanie Forrest, Justin Balthrop, Email networks and the spread of computer viruses, Phys. Rev. E 66 (2002) 035101.
[11] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, S. Bhattacharjee, Measurement and analysis of online social networks, in: IMC'07: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, 2007.
[12] R. Kumar, J. Novak, A. Tomkins, Structure and evolution of online social networks, in: KDD'06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.
[13] Y. Ahn, S. Han, H. Kwak, S. Moon, H. Jeong, Analysis of topological characteristics of huge online social networking services, in: WWW'07: Proceedings of the 16th International Conference on World Wide Web, 2007.
[14] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (14) (April 2 2001) 3200–3203.
[15] Z. Dezsö, A.-L. Barabási, Halting viruses in scale-free networks, Phys. Rev. E 65 (21 May 2002) 055103(R).
[16] R.M. May, A.L. Lloyd, Infection dynamics on scale-free networks, Phys. Rev. E 64 (2001) 066112.
[17] R. Pastor-Satorras, A. Vespignani, Epidemic dynamics and endemic states in complex networks, Phys. Rev. E 63 (2001) 066117.
[18] A.L. Barabasi, R.A. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[19] T. Bu, D. Towsley, On distinguishing between Internet power law topology generators, Infocom (2002).
[20] http://www.facebook.com/press/info.php?statistics#/press/info.php?statistics.
[21] M. Gjoka, M. Sirivianos, A. Markopoulou, X.W. Yang, Poking Facebook: characterization of OSN applications, in: ACM SIGCOMM Workshop on Social Networks, WOSN'08, August 2008.
[22] Adonomics. http://www.adonomics.com, 2009.