Computer Virus Propagation Model Based on Variable Propagation Rate

Cong Jin, Qing-Hua Deng, Jun Liu

Department of Computer Science, Central China Normal University, Wuhan 430079, China
E-mail: jincong@mail.ccnu.edu.cn

Abstract. In this paper, two different propagation models based on different topologies of email network are proposed. By analyzing the means and the characters of email virus spreading, the function of email virus propagation is given, and the maximum time of email virus propagation before the anti-virus software is calculated. The condition in which email virus propagation stops is also proved. The relation between average node degree and power law exponent is discussed later. The models have been testified its rationality through simulation experiments.

1. Introduction

Computer virus propagation is influenced by various factors, and these factors are regarded as constants in most of the existed models. So, some detail information of computer virus propagation is neglected, and mathematical model is simplified. In fact, many factors are changed during the virus propagation. In this paper, the email virus propagation rate is designed as a variable for simulating exactly.

2. Preliminary Knowledge

We describe the logical email network as a directed graph $G=<V, E>$, where V is the set of nodes denotes the email users and E is the set of links. If node A has the email address of node B in its email address book then there is a link from node A point to node B point and vice versa. If A and B have the email address of each others then there is an undirected link between A and B. A remarkable property of email virus propagation is that the email virus must be expanded through email address. A must have the email address of B before it transfer the email virus to B. The directed nature of the email network makes the spread of email viruses qualitatively different from the spread of human diseases. The in-degree of a user is k_{in} means that there are k_{in} users have the email address of the user. The out-degree of a user is k_{out} means that there are k_{out} email addresses in the user's email address book. Apparently, the bigger the in-degree is, the higher the probability of being infected is. The bigger the out-degree is, the higher the probability of infecting others is.

Cliff C. Zou *et al*. points out that the nodes degrees of email network satisfied power law distribution [1]. That is $p(k) \propto k^{-\gamma}$, where γ is the power law exponent. The in-degree satisfied the power law distribution as well as the out-degree. The users that have a large of email contacts are fewer. Most of the users have a small-scale email address book. Power law distribution is an important property of email network. Another equally important property is local aggregation. It is common that somebody have the email addresses of each others. They consist of a cluster or a group. The logic email network of a group can be regard as a completely connected graph. Actually, the email network is a social network that indicates the relationship between email users. Anybody belongs to a group or more and all the big or small groups compose the whole email network. The users in the same group connect closely.

3. Email Virus Propagation Model

Email network topology deeply affected email virus propagation. To found email virus propagation model, many aspects of email virus are captured. The topology of email group is different from the whole email network. Thus two models adapt to dissimilar topologies are presented respectively.

(1) Email Virus Propagation in the Group

Let the email virus propagation be a discrete time process, *i.e.*, $t = 0, 1, 2, 3, \ldots\ldots$. The unit of time is day (24 hours). The size of the group is M . I_t is the number of infected users at time t in the group. δ is the probability of cleanup virus in the group. Users open the unsafe email with the probability α and the interval of checking email is μ . Therefore, the opening probability in unit time is $\frac{\alpha}{\mu}$. At time $t+1$ the number of infected users I_{t+1} is composed of two parts. One is the users that have been infected at time t but have not been clean at time $t+1$. The other is the newly infected users, *i.e.*, the users who are healthy at time t but infected at time $t+1$. Because of having the email addresses of each other within a group, all the other users receive the email virus copies as long as one of them has been infected. Here, the restriction of network bandwidth isn't considered, *i.e.*, there are $(M - I_t)$ users are infected newly at any time. Whether the suspicious users would be infected or not is determined by whether they would open the email. Some hackers embed virus in the email text but not the attachment. Email users are infected after checking the email in despite of not opening the attachment. Email virus like this is more covert than others. So we let that the users be infected once they open the email. The model applied to email group is given as follows:

$$I_{t+1} = (1-\delta)\, I_t + \frac{\alpha}{\mu}(M - I_t). \qquad \textbf{(1)}$$

$$I_t = \left(1 - \frac{\alpha M}{\delta\mu + \alpha}\right) e^{-(\delta + \frac{\alpha}{\mu})t} + \frac{\alpha M}{\delta\mu + \alpha}. \qquad \textbf{(2)}$$

Where $I_0 = 1$. Equation (2) shows that the maximum number of infected users is depended on the proportion of opening probability and cleanup probability. Smaller value of opening probability and bigger value of cleanup probability imply a smaller number of maximal infected users. Let the size of group is 20, *i.e.*, $M = 20$. Cleanup probability $\delta = 0.2$ and opening probability $\alpha = 0.7$. According to the habits of email users, the interval of checking email is $\mu = 1$. Experiment shows that the infected virus number increases greatly within a short time and then tends to a steady state in general. Instead of spreading continually, email virus propagation terminates at an equilibrium point result in some users remain healthy at the end of the propagation. Email virus outbreak quickly and also terminate quickly in the group.

(2) Email Virus Propagation in the Internet

It is often the case that the anti-virus software is updated only after a virus has spread for some time. In the beginning, email users know so little about the new virus that none of strategy can be use to stop the spreading of virus. The new virus propagates unrestrictedly until the malicious activities caught the attention of people. Once the anti-virus software appearing, it can be used to throttle the further propagation of the virus from the infected users. So, the virus propagation is classified into two phases.

1) The Initial Phase

Suppose that the anti-virus software starts to be available at the time T_0. Before the time T_0, *i.e.*, $t < T_0$, the spreading of email virus is modeled as follows $I_{t+1} = I_t + \frac{\alpha}{\mu} \beta(t) I_t$. Where, $\beta(t)$ is the function of virus propagation. Rather than all the email users are infected with the same probability, the users are infected by the infected contacts in the email addresses. The pervasion of email virus is implemented by spreading the virus copy to the contacts in the email address. The spreading of email virus is active but not passive. Exactly, the users who may be infected at time $t+1$ are the users that link with the user who have been infected at time t. This model takes the initiative of email virus propagation into account and believes that the number of email virus copies is $\beta(t)I_t$. Thus the number of newly infected users is $\frac{\alpha}{\mu}\beta(t)I_t$.

The function of email virus propagation $\beta(t)$ is varied with time and related with the average node degree of email users. The average node degree is greater, the $\beta(t)$ is bigger. Because of the feature of cluster email virus likely transfers the email virus copies to the infected users. The number of infected user increase sharply when it infects a healthy group in the first time. If most of the users in a group have been infected, email virus propagates mildly. Only the healthy users are favor of the spreading of email virus. Thus $\beta(t)$ is also related with the proportion of healthy users. We design the definition of $\beta(t)$ based on the two factors analyzed above. $\beta(t)=\bar{k}\frac{N-I_t}{N}$, where $\bar{k}$ is the average node degree of email users, and $\frac{N-I_t}{N}$ is the proportion of healthy users to total email users. Replace $\beta(t)$ with $\bar{k}\ \frac{N-I_t}{N}$, and we obtain $I_{t+1} = I_t + \frac{\alpha}{\mu}\ \bar{k}\ \frac{N-I_t}{N}\ I_t$. Furthermore, the differential of I_t indicates the increasing rate of email virus and we can obtain the differential of I_t described by $\frac{dI_t}{dt}=-\frac{\alpha\bar{k}}{\mu N}(I_t-\frac{N}{2})^2+\frac{N\alpha\bar{k}}{4\mu}$. Where the infected users is 5 at the initial time, namely $I_0=5$. While $I_t=\frac{N}{2}$, *i.e.*, $t=\frac{\mu}{\alpha\bar{k}}\ln\frac{N-5}{5}$, $\frac{dI_t}{dt}$ takes maximum value $\frac{N\alpha\bar{k}}{4\mu}$. In other words, email virus propagates most quickly when half of the email users are infected before the anti-virus program is available. In order to restrain the large-scale outbreak of email virus we should try our best to run the anti-virus software before the time $t=\frac{\mu}{\alpha\bar{k}}\ln\frac{N-5}{5}$. That is to say, the bigger the value of t is, there are more time for the anti-virus experts to research the anti-virus software. So email users should open the email with long interval and small probability to delay the time t. To store as small email addresses as possible in the email address book is also helpful to delay t.

2) The Latter Phase

After the anti-virus software is available, *i.e.*, $t>T_0$, the cleanup probability is not zero anymore. The case of email virus propagation is $I_{t+1} = (1-\delta)\ I_t + \frac{\alpha}{\mu}\ \bar{k}\ \frac{N-I_t}{N}\ I_t$. Furthermore,

$$\frac{dI_t}{dt}=-\frac{\alpha\bar{k}}{\mu N}[I_t-\frac{(\alpha\bar{k}-\delta\mu)N}{2\alpha\bar{k}}]^2+\frac{N(\delta\mu-\alpha\bar{k})^2}{4\mu\alpha\bar{k}} \quad \textbf{(3)}$$

$$\frac{1}{I_t}=[\frac{1}{I_0}-\frac{\alpha\bar{k}}{N(\alpha\bar{k}-\delta\mu)}]e^{(\delta-\frac{\alpha}{\mu}\bar{k})t}+\frac{\alpha\bar{k}}{N(\alpha\bar{k}-\delta\mu)} \quad \textbf{(4)}$$

There are 5000 infected email users in the Internet when the anti-virus software appears, *i.e.*, $I_0=5000$, and $\frac{dI_t}{dt}$ is the increasing rate of email virus in unit time. While $\frac{dI_t}{dt}<0$, the number of infected users lessen and the email virus no longer spreads. From Equation (3), we know that when $I_t > \frac{(\alpha\bar{k}-\delta\mu)N}{\alpha\bar{k}}$, $\frac{dI_t}{dt} < 0$ .Thus,

$$\delta > (1 - \frac{I_0}{N})\frac{\alpha\bar{k}}{\mu}. \tag{5}$$

Inequality (5) points out the restriction among various factors. The users who have large email address book should cleanup virus frequently to control virus propagation. Some users are accustomed to check email with short interval. These users should also cleanup virus with a high frequency. If users open email with low probability, a low cleanup probability is also useful to control propagation. During the process of email virus propagation, if the cleanup probability δ, the opening probability in unit time $\frac{\alpha}{\mu}$ and the average degree $\bar{k}$ satisfy the inequality (5), $\frac{dI_t}{dt} < 0$, *i.e.*, email virus will disappear gradually.

4. Discussion of Average Node Degree

The average node degree is a crucial factor of email virus propagation. To a great extent, the speed of email virus spreading depends on the average node degree. However, it is really difficult to decide the value of average node degree by statistic data due to the hugeness of email network. Thus, we discuss the relativity of average node degree and the power law exponent for ascertaining the value. The average node degree can be expressed as $\bar{k} = \sum kp(k)$, where $p(k)$ is the probability of any given node with degree *k*. The degree of email network satisfied the power law distribution, thus $p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$, where γ is the power law exponent and $\zeta(\gamma)$ is the Riemann zeta function, and $\zeta(\gamma) = \sum_{1}^{\infty} k^{-\gamma}$ [2]. Power law exponent of many actual complex networks are different from each other and the range is $2 \le \gamma \le 3$. So, we have

$$\bar{k} = \frac{\gamma - 1}{\gamma - 2}. \tag{6}$$

Most users have a small-scale email address book, so the value of $\bar{k}$ is impossible to be infinite and γ is not equal to 2, *i.e.*, γ is greater than 2. When the value of γ increases, the value of $\bar{k}$ decreases. The value of $\bar{k}$ gets the minimum 2 while γ reaches the maximum 3. If we know the value of exponent power law exactly, the value of average node degree $\bar{k}$ can be figured out from Equation (6). We established the basis for selecting the value of $\bar{k}$. It is helpful for designing the function of propagation and then further developing the propagation model.

5. Simulation Experiment

Let the unit of time be 24 hours. The parameters are set as, the size of email users is $N = 10000$, the interval of checking email μ =1, and the average of contacts $\bar{k} = 6$. Figure 1 shows that email virus spread freely before anti-virus software appearing and the speed is fast. Email virus would infect all the email users without anti-virus software. The larger the opening probability is, the higher the speed of spreading is. The time at which email virus propagates fastest is pointed out through the dashed line. Figure 2 clearly shows that email virus propagation has two cases after anti-virus software is used. Either it increase sharply and tend to a stable state or decrease and tend to zero. α is smaller and δ is greater, email virus propagation is slower. When the inequality (5) is tenable, email virus propagation goes down and the number of infected users reduce gradually. When the inverse case is tenable, email virus propagation goes up and the number of infected users adds. Let $\Delta = |\delta\mu - \alpha\bar{k}|$.

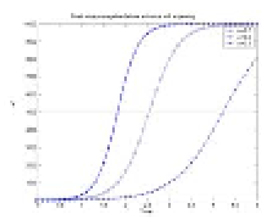

Fig. 1. Email virus propagation on different α and δ

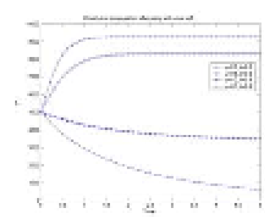

Fig. 2. Different α and δ

6. Conclusions

The terminative condition of email virus propagation plays a significant role on control. Highly-connected users request large cleanup probability. Low opening probability and large checking interval request a comparatively small cleanup probability. Instead of a fixed value, δ is different for different users to stop spreading. Average node degree is inversely proportional to power law exponent. Considering the relation between $\bar{k}$ and γ bring the model to be self-adaptive. By adjusting the power law exponent automatically, the model is suitable for different topologies. The email network is less likely to be BA scale-free network. The equation can be used to evaluate the email network model.

References

1. C.C.Zou, D.Towsley, and W.B.Gong. Email virus propagation modeling and analysis. Technical Report: TR-CSE-03-04, University of Massachusetts, Amherst 2003
2. J.T.Xiong. ACT: attachment china tracing scheme for email virus detection and control. Proc. of the 2004 ACM Workshop on Rapid Malcode, October 29-29, Washington DC, USA, 2004, 11-22