Network Virus Propagation Model Based on Effects of Removing Time and User Vigilance

Cong Jin, Jun Liu, and Qinghua Deng
(Corresponding author: Cong Jin)

Department of Computer Science, Central China Normal University
Wuhan, 430079, China (Email: jincong@mail.ccnu.edu.cn)
(Received Mar. 9, 2008; revised and accepted Dec. 4, 2008)

Abstract

Network virus propagation is influenced by various factors, and some of them are neglected in most of the existed models. So, mathematical model of network virus propagation is simplified. In fact, many factors are very important during the virus propagation. In this paper, we investigate epidemiological models to reason about email virus propagation. The paper extended the classical virus propagation model SEIR for incorporating two new parameters: *User Vigilance* and *Removing Time*. We show that these parameters greatly influence the virus propagation. The fruitful simulations will demonstrate that this developed model can be used for describing email virus propagation and calculating the costs of virus outbreak. We also prove that the time of anti-virus technique appearing plays an important role in controlling virus propagation.

Keywords: Virus propagation model, epidemiology, anti-virus technique, removing time, user vigilance

1 Introduction

Currently, email has become one of the most basic applications in the Internet with the development of networked computer. Email security problem plays an important role on the security and reliability of the whole Internet because of its extensive users and close binding with credit card or account. Usually a virus email has an attachment file that contains copy of the virus. The virus hides the attachment file's executable property by forging it to be any type of files, like image, word document, etc. When an email user clicks on this attachment, the virus program will be activated and infect the local computer. Due to the facility, hackers mostly tend to choose the email as the measures of spreading their email virus.

Before some effective strategies are present to control the spreading of various email viruses, we must understand clearly how the email virus spread in the email network. Lots of efforts have been devoted to develop various mathematical models in order to simulate the real case of virus spreading. Some researchers have studied for years about email virus propagation model. In 2003, Zou et. al. [11] presented an email virus model that accounts for behaviors of email users, such as email checking frequency and the probability of opening an email attachment. Jintao Xiong [10] proposed an automated email virus detection and control scheme using attachment chain tracing in 2004.

For years, the email virus propagation models based on epidemiological theories of human epidemic disease have been researched. The model usually divides the user into many states. The studies often relate to the Susceptible-Infected-Susceptible (SIS) infection model, and the Susceptible-Infected-Removed (SIR) model. In the SIS model, a user is infected and then cured, while in the SIR model, the user probably is removed by anti-virus technique after infecting. SIS and SIR models are researched perfectly and widely used to study email virus propagation. In 1991 and 1993, Kephart and White studied SIS virus propagation model on homogeneous networks respectively [2, 3]. While Satorras and Vespignani [5, 6, 7] focused on SIS and SIR model for the spreading of epidemics in complex networks by analytical methods and large scale simulations.

We know that, in the email transmission, the email user may not open or activate the email with virus when receiving an email. The situation is named Exposed state. Anderson and May [1] investigated spreading characters of various infectious disease, and added Exposed state to SIR model. Then SEIR virus model appeared, and typical states of SEIR model in Table 1.

Table 1: Typical states of SEIR model

S	Susceptible
E	Exposed to Infection
I	Infected
R	Removed

Although the SEIR email virus propagation model

achieved a better performance than the SIR model on describing the user's behavior of information processing and anti-virus technique, the SEIR model has three important shortcomings as follows:

1) We know that email virus propagation is influenced by various parameters, and these parameters are usually regarded as constants in most of the existed models [7, 8]. In these models, time is divided into some discrete steps to describe the model. Transitions between individuals in each state of the models are described by simple probabilities in every time steps. In fact, the simplified email virus propagation model can't reflect the actual situation of virus diffusing. For example, the spreading rate of virus ($S \rightarrow E$) is small at the beginning of virus breakout, because of more infected users appearing, this parameter will become larger.

2) Virus propagation is simulated by same virus model from email virus breakout to immunization. Thus, such model can't reflect the actual situation accurately. We know that, when a serious email virus break out, current anti-virus techniques may not cleanup the email virus because few software companies would develop a technique to remove virus before email virus appearing. Based on the above analysis, in this paper, we attempt to develop two phase virus models according to the time of appearing anti-virus software. In other words, *Removing Time* parameter should be incorporated to the virus propagation model.

3) Traditional virus SEIR model neglected difference of the email users. Quite a lot Internet users less understand virus hidden in email attachment. Email users usually give an appropriate trust to emails from their friends. Email with virus may be opened without suspiciously, and not be scanned by anti-virus software. The situation is called that users have little vigilance. In addition, most Internet users install or update the anti-virus software. Suspicious email may be deleted or scanned. The high vigilance of Internet users will reduce probability of the local computer infected. Thus, *User Vigilance* parameter also should be incorporated to the virus propagation model.

In this paper, two new parameters, *i.e. Removing Time* and *User Vigilance*, are incorporated to SEIR model for improving SEIR model. Two new parameters of the email virus propagation model have been not researched almost. However, they play an important role on improving the model performance.

2 Email Virus Propagation Model

The general process of the email virus infection is described as follows.

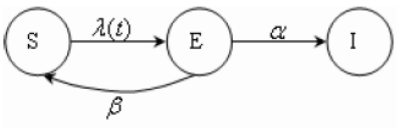


Figure 1: $t < T$ virus model

First, the virus is released into the wild by its creator. The virus is spreads freely, infecting user's machines in the network. In the beginning of the virus spreading, the serious virus is not noticed or alerted. Meanwhile, anti-virus techniques are not developed. So email users haven't abilities to remove the virus. After the virus has spread for some time, anti-virus company works to isolate the virus and generates an anti-virus technique used to detect the presence of the virus. This process can keep on some time. The time T of anti-virus technique used is called as Removing Time. So, our email virus spread model contains two phases.

1) Time $t < T$: Virus spreads freely.

2) Time $t \geq T$: Anti-virus technique presents, and the most users start to remove virus.

2.1 The Phase of Virus Spreading Freely

Before the virus can spread unchallenged, the user state only has three cases: *Susceptible* (S), *Exposed* (E) and *Infected* (I), no remove states. In this situation, the infected users become more and more because of no appearing anti-virus software. Figure 1 represents this state.

In following discussion, the meanings of the some signs are as follows:

- $P(S \rightarrow E) = \frac{rI(t)}{N}$: Rate of a susceptible user becoming *Exposed* state; N is total number of users.
- r: Clustering coefficient.
- $P(S \rightarrow I) = \alpha$: Rate of a *Exposed* susceptible user becoming *Infected* state.
- $P(E \rightarrow S) = \beta$: Rate of an *Exposed* user becoming *Susceptible* such as, by deleting email with virus.

Therefore, the state of Figure 1 can be described by the following equations [4]:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\lambda(t)S(t) + \beta E(t). \\ \frac{dE(t)}{dt} &= \lambda(t)S(t) - (\beta + \alpha)E(t). \\ \frac{dI(t)}{dt} &= \alpha E(t). \\ \frac{dR(t)}{dt} &= 0. \end{aligned} \quad (1)$$

Where, $S(t)$, $E(t)$, $I(t)$, and $R(t)$ are number of users with *Susceptible*, *Exposed*, *Infected* and *Removed* state respectively at every time step. Thus

$\frac{dS(t)}{dt}, \frac{dE(t)}{dt}, \frac{dI(t)}{dt}, \frac{dR(t)}{dt}$ are the increasing rates of *Susceptible*, *Exposed*, *Infected* and *Removed* users respectively at every time step. $\lambda(t)$ is a virus propagation function and vary with the time. In the beginning, the value of $\lambda(t)$ is small, then it increases mildly with virus spreads, and more infected users appear. In Equation (1), $\lambda(t)S(t)$ is number of susceptible users changing into exposed at time t. Consider probability of some user may discard suspicious email with virus attachment, $\beta E(t)$ means the number of increasing susceptible users at time t. Therefore $-\lambda(t)S(t)+\beta E(t)$ is changer rate of susceptible users. Furthermore, owing to no presence of anti-virus software, so the value of $\frac{dR(t)}{dt}$ keeps 0.

It should be noted that $\lambda(t)$ doesn't reveal the factor of network congestion when many computers are infected.

2.2 The Phase of Removing Users

When $t \geq T$, anti-virus technique has developed to cleanup or isolate virus. In fact, not all users install or update the anti-virus software for isolating the virus. If a user adopts the anti-virus software with a high probability for detecting and removing the virus, he can obtain a safer environment when connecting to the Internet. This user is called with a high vigilance. However, many Internet users haven't much understanding about the importance of anti-virus software. So, they don't install or update the anti-virus software in time on their computers. These users have great threat to other users. The portion of users called low vigilance will continue to infect computers from their email address book. Therefore, the virus spreading parameter *User Vigilance* should models by *User Vigilance* δ defined as follows:

$$\text{User Vigilance } \delta = \frac{\text{The number of installing anti-virus software}}{\text{The total number } N \text{ of users}}$$

User Vigilance δ indicates user rate of installing anti-virus software. $\delta \in [0,1]$. 0 indicates all users don't install or update anti-virus software, and 1 indicates all users install or update anti-virus software [9].

For high vigilance δ users, the anti-virus technique is distributed, and email virus is cleaned up. Therefore, email virus spreading and cleanup can be modelled as shown in Figure 2. *Susceptible*, *Exposed*, and *Infected* states directly become *Removed* state with a high rate. The model reflects that high vigilance users cause email virus accelerates to die.

For low vigilance $1-\delta$ users, the virus continues to spread and infect others users because of no installing anti-virus software in time. So, the virus propagation model may be simplified into anti-virus technique ($t < T$). This phase considers infected users becoming *Removed* state at quite a small rate $k(k \ll a)$. Similarly, users are thought to be one of four states: *Susceptible*, *Exposed*, *Infected*, and *Removed*. Figure 3 represents this state.

Based on the above analytic, for ($t \geq T$) phase, virus propagation model can be described by the following

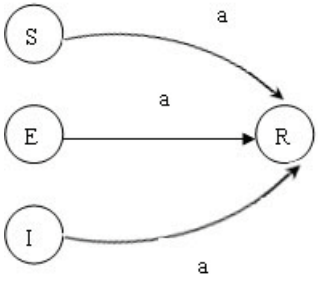


Figure 2: High vigilance users virus model

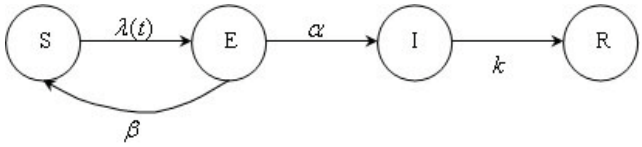


Figure 3: Low vigilance users virus model

equations:

$$\frac{dS(t)}{dt} = (1-\delta)(-\lambda(t)S(t)+\beta E(t)) - \delta a S(t).$$

$$\frac{dE(t)}{dt} = (1-\delta)(-\lambda(t)S(t)-(\alpha+\beta)E(t)) - \delta a E(t).$$

$$\frac{dI(t)}{dt} = (1-\delta)(\alpha E(t) - kI(t)) - \delta a I(t). \quad (2)$$

$$\frac{dR(t)}{dt} = (1-\delta)kI(t) + \delta a(S(t)+E(t)+I(t)). \quad (3)$$

In Equation (3), $\delta a(S(t)+E(t)+I(t))$ is total number of changing into *Removed* users from high vigilance users. Note that *Removed* rate a is usually large quantity. $(1-\delta)kI(t)$ and $\delta a(S(t)+E(t)+I(t))$ is interpreted as all removed users at time t. Equation (2) reveals that the population of infected users at time t will reduce at big rate, since most high vigilance users installed anti-virus software. We know that, from these equations, user vigilance play an important role for controlling the virus propagation.

3 Simulation Experiment

Email virus propagation is affected by many parameters in the email virus model. The influence of some parameters $\lambda(t)$, β, γ, for virus propagation behavior, has already been researched. In this paper, we only concern another two key factors, *i.e.*, *Removing Time* t and *User Vigilance* δ.

In simulation experiment, let email viruses be only transferred by users email address books. Thus email address relationship between users' address books forms a logical network for email viruses. We let the email network has 10000 email users, *i.e.*, N=10000, user clustering coefficient γ=10, and initial infected users are 10 ($I(0)$=0).

Other parameters β, α, a are set as 0.0088, 0.0022, and 0.2, respectively.

3.1 Initial Results

The Figures 4(a) and (b) provide a simple comparison between proposed email virus model after incorporating new parameters *User Vigilance* and *Removing Time* and traditional SEIR.

Figure 4(a) assumes that anti-virus technique presents at time 50, *i.e.* T=50. It shows that the number of infected users increases quickly and accumulates a high value at time 50 before anti-virus software appearing, while removed users keep 0. In other words, email virus would infect freely all email users without anti-virus software. After the anti-virus software is available (*i.e.* $t > T$), the number of infected users drops and removed users increases when quite a lot users install or update new anti-virus software. But the speed of infected users going up and removed users going down are determined by one vital parameter, *i.e.*, *User Vigilance*, and which will be deeply discussed in Section 3.2.

Figure 4(b) reveals that removed users immediately appear when virus attempts to spread, and the size of infected users is smaller than Figure 4(a). This is because the traditional SEIR virus model assumed that, as long as email virus has break out, email users have strategy to control the virus spreading further. However, this assumption is not consistent with objective fact.

Since the paper's work is extensions to traditional virus model, in following Sections 3.1 and 3.2, we will discuss the effects of *User Vigilance* and *Removing Time*.

3.2 Effect of User Vigilance δ

In proposed virus model, *User Vigilance* δ is a vital parameter and is related to how many users install the new anti-virus software. To perform effects of *User Vigilance*, δ is set as three different value (0.1, 0.2, 0.5), and let T=50 (This means that anti-virus technique is developed at time 50). The numerical curve of $I(t)$($R(t)$,$E(t)$,$S(t)$) with different δ will be discussed in Figure 5.

Figure 5(a) clearly shows the outbreak size, *i.e.*, number of infected users, for varying *User Vigilance* after anti-virus technique takes action. The general trend of curve is that $I(t)$ goes down gradually at time $t > 50$. The higher the *User Vigilance* is, the faster speed infected users decrease. This effect is probably interpreted as email users with higher vigilance accelerate the virus fading away. That is to say, the duration of outbreak is more short for a big δ (Please see Table 2), and it results in a weak cost.

Figure 5(b) illustrates that number of removed users keep zero ($t < 50$) and then increase gradually, later tend to stable state in general. Solid line marked with asterisk raises more slowly than other two lines. Thus, increasing δ means increased efficiency of cleanup (It indicates that more users installed or updated the new anti-virus software for centralized immunization). As showed in Table 2, when δ=0.1, the email virus is fully removed at time 310, but δ reaches 0.5, removing virus wins a success just at 105. This result would help us further understanding about why *User Vigilance* is great important to remove email virus.

Table 2: Some discrete data about virus spreading

User Vigilance δ	0.5	0.2	0.1
Time of Fully Removed t	105	182	310
Outbreak Duration Δt	55	132	260

Figures 5(c) and (d) give the change of exposed users, *i.e.*, susceptible users, with varied δ. According to experiment Figure 5(d), susceptible user already disappeared before anti-virus technique appearing, so the value of $S(t)$ obviously continues to keep zero. Figure 5(c) shows that the number of susceptible users drops dramatically. This is because these users become removed at a high rate.

By above simulation results, we may obtain some useful operation suggestions:

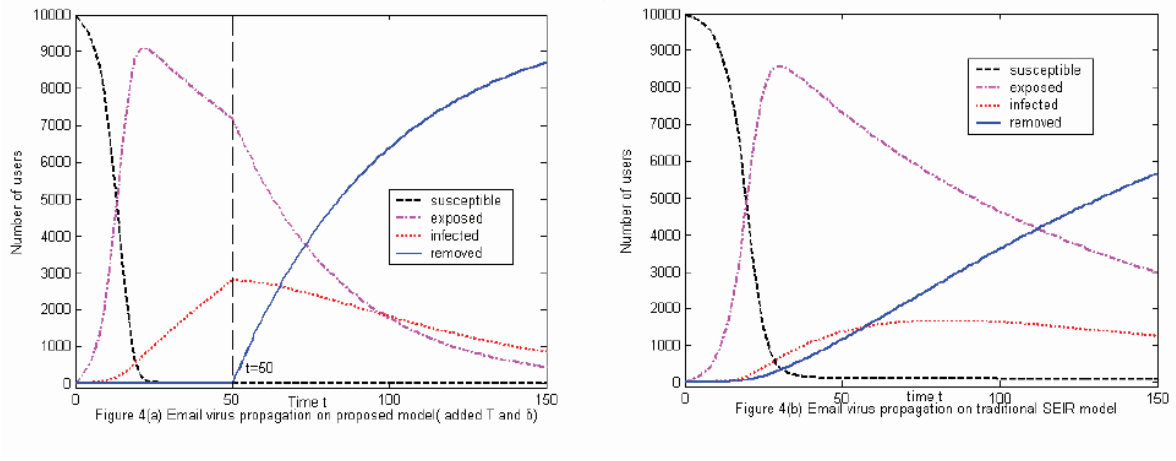
1) To a certain extent, removing viruses may be a combat between removed users and infected users. Once growth speed of the former greatly exceeds the latter, the virus spreading could not a threat to most email users. Therefore, email users with high *User Vigilance* (*i.e.*, more removed users obtained) will defeat terrible email viruses, and experience a weak loss.

2) The importance of *User Vigilance* may suggest us that email users had better actively receive a training of network information security, and relevant organizations, such as government and anti-virus company, propose a virus early-warning. These positive measures would make users possess a high anti-virus consciousness.

3.3 Effect of Removing Time T

One of the questions that can be addressed by proposed virus model is "Whether the *User Vigilance* is the most important parameter". Another parameter *Removing Time* T is simulated in Figure 6. Figure 6 shows the costs, *i.e.*, number of infected users, from varying T =5, 10, 20, and 50 respectively. The general trend is that number of infected users decrease gradually as anti-virus software is available ($t > T$). From Figure 6, the effects of *Removing Time* mainly embody following two points.

1) For a small T, the maximum of infected users are not significant, while larger T results in increasing outbreak (Table 3). That's to say, generating anti-virus technique quickly can greatly reduce costs, but it will produce opposite situation if the new anti-virus software presents lately.



(a) Proposed Virus Model (b) Traditional SEIR model

Figure 4: Comparison between proposed virus model and traditional SEIR

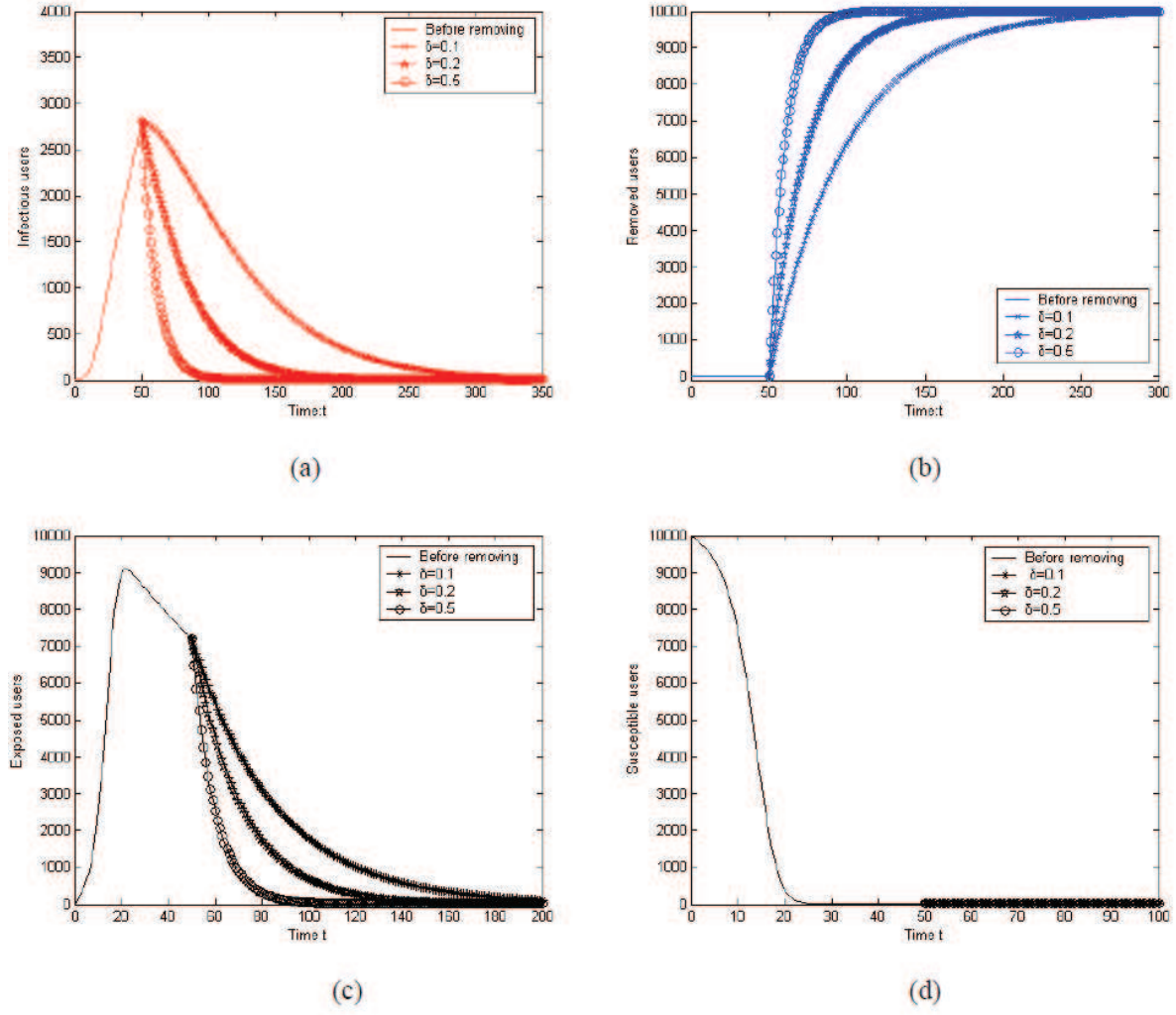


Figure 5: Effects of *User Vigilance* δ

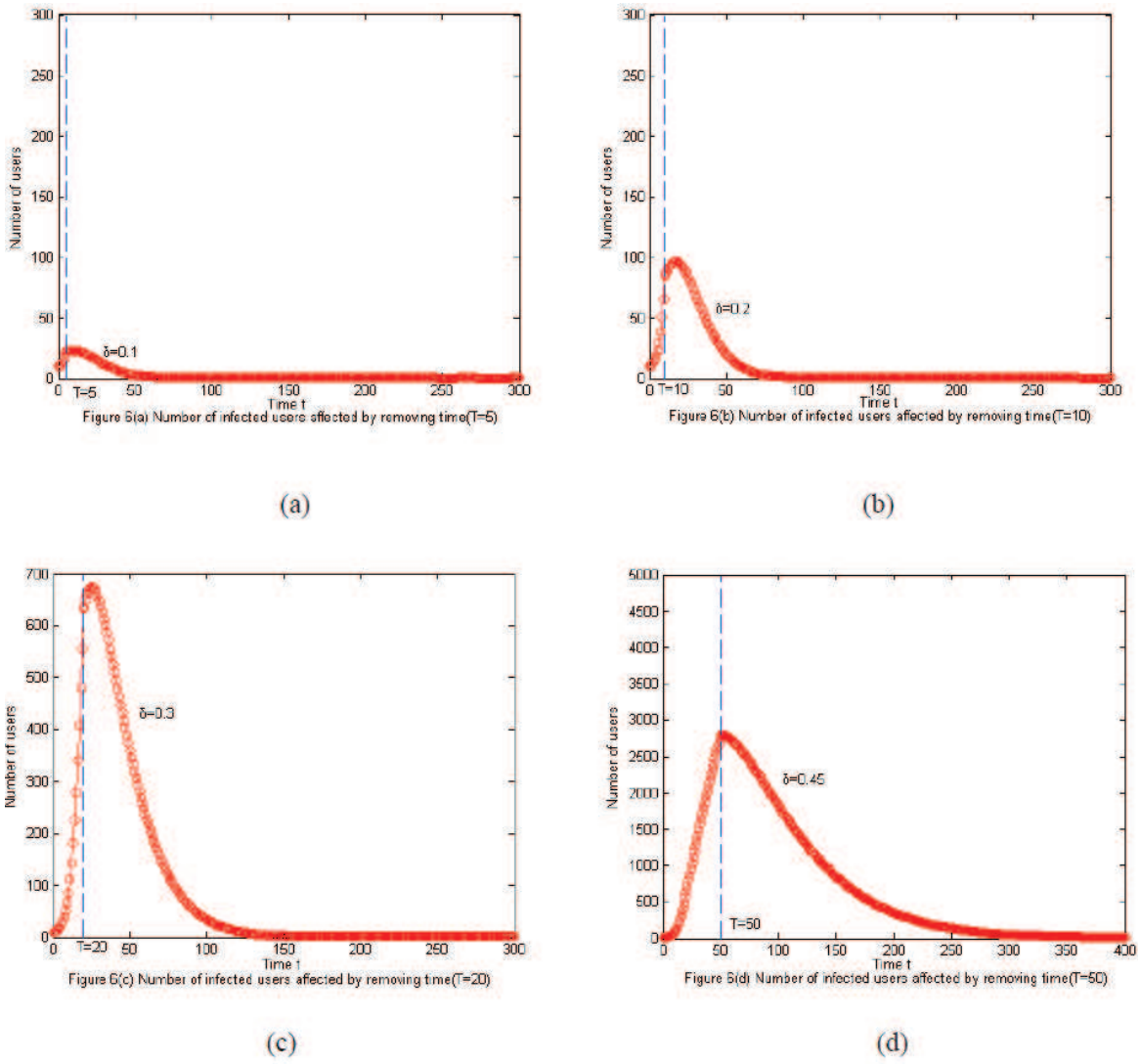


Figure 6: Number of infected users affected by *Removing Time T*

Table 3: Outbreak degree of virus with different *Removing Time T*

Removing Time T	5	10	20	50
Time t	9	18	30	62
Maximum rate of infected users	<0.23 %	<0.96 %	6.74 %	28.02 %

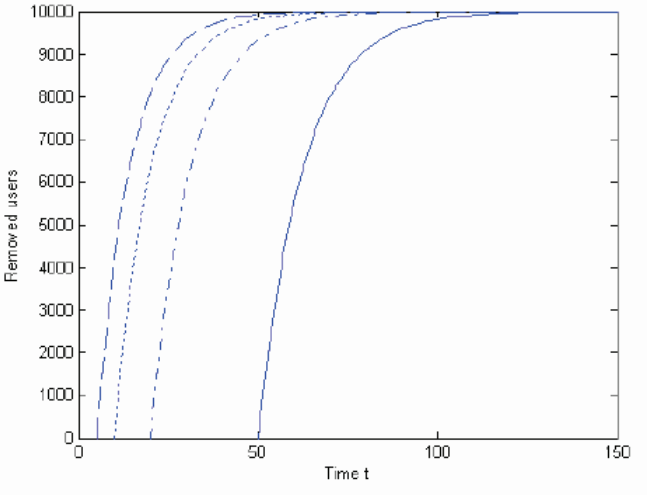


Figure 7: Number of removed users affected by removing time T

In particular, in Figures 6(a), (b), (c), we notice that the number of infected users don't decay right away when $t > T$, but to experience a slow growth. Figure 6(d) should have a similar effect, however this kind of phenomenon don't appear obviously because the proportion of vertical axis is big. Before anti-virus technique is distributed, the virus spreads unhindered quickly. The size of infected users may accumulate a high degree during virus spreading freely. So anti-virus software will take some time to make the infected users became small. This fact gives a well understanding why the maximum of infected users don't appear at time T (See Table 3).

2) For a large T, the outbreak duration (Please see Figure 7 and Figure 6) will last a long time in despite of email users with high *User Vigilance* δ. Conversely (*i.e.*, a small T), the system may be not suffer the great losses even though δ is not quite large. This phenomenon is not surprising because the network may accumulate more infected users for a large T. It indicates that anti-virus technique present lately, and email virus already spreads a long time. Therefore, the time that infected users are made immune ($I \rightarrow R$) can be long.

In summary, simulations of the model reveal that the time of anti-virus technique appearing, *i.e.*, *Removing Time T*, a vital factor to control virus email propagation. If anti-virus software is used before the large-scale outbreak of email virus, it is easy to defeat virus propagation; otherwise, it is hard to defeat despite of higher *User Vigilance*. Based on this important conclusion, one feasible measure may suggest that mail server enhances a mechanism through increasing probability of filter spam, prolonging average time of sending or receiving emails before anti-virus technique is available. The much time of researching anti-virus software will win for workers of network information security.

4 Conclusions

In this paper, we propose an email virus model based on the epidemiological viewpoint. The research extends previous work by incorporating two new parameters, *i.e.*, *User Vigilance* and *Removing Time*, to classical virus propagation model SEIR. Previous papers have focused on parameters, such as rate of cured, rate of virus spreading and coefficient clustering, which exactly give a high level understanding of the system dynamics. While these key parameters used in SEIR model reveals concrete situation in real virus propagation. For example, anti-virus technique appears early or late, and users have high or low vigilance to fight against email virus. Thus, new factors are helpful to understanding the real case of email virus propagation.

By analysis of the model and simulation studies, it reveals that the time of anti-virus technique and user vigilance are vital factors in controlling virus spreading. If email users have higher user vigilance, virus spreading can be easily defeated in spite of infected computers accumulated a high value. Furthermore, the system may only suffer a small damage because outbreak lasts for a short time. In this paper, we may obtain the following important conclusions:

The more promptly anti-virus industry develops corresponding anti-virus technique, the less outbreak email viruses result in. However, even if people have ability to remove this email virus, system will undergo a great loss (*i.e.*, lots of users infected) if most email users have much low vigilance. Finally, the paper also gives insight into the relative importance of the research of new anti-virus technique for the entire anti-virus industry.

Email virus propagation strongly depends on email network topology. However, it isn't considered for simplifying the model in this paper. In future, one important work is to consider the effects of different network topology for more exactly simulate email virus propagation.

Acknowledgments

This research was supported by the National Science Foundation of Hubei (China) under Grant No. 2007ABA119.

References

[1] R. M. Anderson and R. M. May, "Infectious diseases of humans: dynamics and control," Oxford: Oxford University Press, 1991.

[2] J. O. Kephart and S. R. White, "Directed-graph epidemiological models of computer viruses", *IEEE Computer Society Symposium on Research in Security and Privacy*, pp. 343-359, May 1991.

[3] J. O. Kephart and S. R. White, "Measuring and modeling computer virus prevalence," *IEEE Computer Society Symposium on Research in Security and Privacy*, May 1993.

[4] B. K. Mishra and D. Saini, "Mathematical models on computer viruses," *Applied Mathematics and Computation*, vol. 187, no. 2, pp. 929-936, 2007.

[5] R. P. Satorras and A. Vespignani, "An epidemic spreading in scale-free networks," *Physical Review Letters*, vol. 86, no. 14, pp. 3200-3203, 2001.

[6] R. P. Satorras and A. Vespignani, "Epidemic dynamics and endemic states in complex networks," *Physical Review E*, vol. 63, pp. 066117, 2001.

[7] R. P. Satorras and A. Vespignani, "Epidemic dynamics in finite size scale-free networks," *Physical Review E*, vol. 65, pp. 035108, 2002.

[8] R. P. Satorras and A. Vespignani, "Epidemics and immunization in scale-free networks," *Handbook of Graphs and Networks: From the Genome to the Internet*, Wiley-VCH, Berlin, May 2002.

[9] Y. Wang and C. X. Wang, "Modeling the effects of timing parameters on virus propagation," *ACM workshop on Rapid Malcode*, pp. 61-66, Oct. 2003.

[10] J. T. Xiong, "ACT: Attachment chain tracing scheme for Email virus detection and control", *ACM Workshop on Rapid Malcode*, Washington DC, USA, Oct. 2004.

[11] C. C. Zou, D. Towsley, W. Gong, "Email Virus Propagation Modeling and Analysis", *Technical Report*, TR-CSE-03-04, University of Massachusetts, Amherst, 2003.

Cong Jin received the M.S. degrees in applied mathematics from Harbin Institute of Technology, Harbin, Heilongjiang, China, in 1990. She received the Ph.D. in Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 2006. From 1993 to 2003, she was a Lecturer and then become as Professor at the Hubei University, Wuhan, Hubei, China. From 2003 to now, she is Professor of the department of computer science, Central China Normal University, Wuhan, Hubei, China. She has published more than 100 papers on information security, signal processing and algorithm design. Her main research interests include computer network security, digital copyright protection, and intelligence information processing, etc.

Jun Liu is a M.S. candidate at Central China Normal University. His research interests include: information security, computer virus, and digital image processing. He has published 4 scientific papers.

Qinghua Deng is a M.S. candidate at Central China Normal University. Her research interests include: information security, digital watermark, and digital image processing. She has published 3 scientific papers.