Modeling and Analysis on the Propagation Dynamics of Modern Email Malware

Sheng Wen, *Student Member, IEEE*, Wei Zhou, Jun Zhang, *Member, IEEE*,
Yang Xiang, *Senior Member, IEEE*, Wanlei Zhou, *Senior Member, IEEE*,
Weijia Jia, *Senior Member, IEEE*, and Cliff C. Zou, *Senior Member, IEEE*

Abstract—Due to the critical security threats imposed by email-based malware in recent years, modeling the propagation dynamics of email malware becomes a fundamental technique for predicting its potential damages and developing effective countermeasures. Compared to earlier versions of email malware, modern email malware exhibits two new features, reinfection and self-start. Reinfection refers to the malware behavior that modern email malware sends out malware copies whenever any healthy or infected recipients open the malicious attachment. Self-start refers to the behavior that malware starts to spread whenever compromised computers restart or certain files are visited. In the literature, several models are proposed for email malware propagation, but they did not take into account the above two features and cannot accurately model the propagation dynamics of modern email malware. To address this problem, we derive a novel difference equation based analytical model by introducing a new concept of virtual infected user. The proposed model can precisely present the repetitious spreading process caused by reinfection and self-start and effectively overcome the associated computational challenges. We perform comprehensive empirical and theoretical study to validate the proposed analytical model. The results show our model greatly outperforms previous models in terms of estimation accuracy.

Index Terms—Network security, email malware, propagation modeling

1 Introduction

In the real world, email is a basic service for computer users, while email malware poses critical security threats. For a number of years, the propagation of email malware has followed the same modus operandi. A viral email is sent to the victim and appears as though it was sent by somebody the recipient trusts. The subject is also related to the recipient's business area. Once the victim is tricked into either clicking the malicious hyperlinks or opening the attachments inside such an email, the computer will be compromised. Then, the compromised computer will start to infect new targets found in its email address lists immediately. To prevent email malware, scientists have spared no effort to dissuade people from opening unexpected hyperlinks and email attachments. However, the success of recent new email malware, such as "Here you are" [1], indicates that those education measures are not very successful. A key reason is because social engineering is a tried-and-true technique in the context of security. For example, by convincing computer users that the received emails with malicious hyperlinks and attachments were from a trusted source, the technique of email-borne malware will be highly effective and is still widely adopted by current malware authors [2].

Current research on email malware [3], [4], [5], [6], [7] focuses on modeling the propagation dynamics which is a fundamental technique for developing countermeasures to reduce email malware's spreading speed and prevalence. There are a few works reported to model email malware propagation. Previous works [4], [5], [6], [8] assume that a user can be infected and send out malware copies ***only once***, no matter whether or not the user visits a malicious hyperlink or attachment again. Real instances are those early email malware like Melissa in 1999 [9] and Love letter in 2000 [10], which will check whether a victim has been compromised before the infection. However, modern email malware is far more aggressive to spread in network than before by introducing two new propagation features [7], [11], [12], [13]. First feature is "reinfection", i.e., an infected user sends out malware copies whenever this user visits the malicious hyperlinks or attachments. Second feature is "self-start", i.e., an infected user sends out malware copies when certain events (like PC restart) are triggered. Researchers in [3] stated that a user can be infected multiple times. However, their model assumes that an infected user could send out ***only one*** malware copy each time the user checks emails, even if the user visits more than one malicious hyperlinks or attachments. In short, previous works [3], [4], [5], [6] did not take the two new features into account, and hence, cannot accurately estimate the propagation of modern email malware. An empirical study to support our argument will be introduced later in Section 2.

It is a big challenge to investigate modern email malware through mathematical modeling. In fact, most email

- *S. Wen, W. Zhou, J. Zhang, Y. Xiang, and W.L. Zhou are with the School of Information Technology, Deakin University, 17 Rochdale Drive, Burwood East, Melbourne, Australia, VIC 3151.*
  E-mail: {wsheng, weiz, jun.zhang, yang, Wanlei}@deakin.edu.au.
- *W. Jia is with Shanghai Jiaotong University, Shanghai, P.R. China.*
  E-mail: weijiaj@gmail.com.
- *C.C. Zou is with the School of Electrical Engineering and Computer Science, University of Central Florida, 4000 Central Florida Blvd, Orlando, FL 32816. E-mail: czou@cs.ucf.edu.*

Manuscript received 18 Dec. 2012; revised 18 Aug. 2013; accepted 8 Nov. 2013. Date of publication 19 Nov. 2013; date of current version 16 July 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TDSC.2013.49

malware in the last decade, such as Sircam in 2001 [14], Sobig in 2003 [15], Mydoom in 2004 [16], Nyxem in 2006 [17], "Here you are" in 2010 [1] and recent unnamed email malware [2] belong to the modern email malware. The previous analytical model [4] presented the spreading procedure by an susceptible-infected-susceptible (SIS) process, while it does not consider the new features of modern e-mail malware. These observations become the motivation of our work to develop a new analytical model that can precisely present the propagation dynamics of the modern email malware. Since the spreading procedure can be characterized by an susceptible-infected-immunized (SII) process, we name our proposed model as SII.

The major contributions of this paper are listed below:

- We propose a new analytical model to capture the interactions among the infected email users by a set of difference equations, which together describe the overall propagation of the modern email malware.
- We introduce a new concept of virtual nodes to address the underestimation in previous work, which can represent the situation of a user sending out one more round of malware copies each time this user gets infected.
- We perform empirical and theoretical study to investigate why and how the proposed SII model is superior to existing models.

The rest of the paper is organized as follows. Section 2 states the problems in modeling modern email malware. In Section 3, a new SII analytical model is presented in detail. Section 4 reports a series of experiments to validate the proposed model, followed by the theoretical justification in Section 5. Further discussion and related work are presented in Sections 6 and 7. Finally, Section 8 concludes this paper.

2 Problem Statement

Choosing email as the spreading carrier of malware is not a new technique in the last decade. Early versions of email malware, such as Melissa and Love letter, work in a "naive" way. That is, a compromised user will send out malware emails only once, after which the user will not send out any further malware copies, even if she visits the malicious hyperlinks or attachments again. Take Melissa for example, the malware first checks a specific registry key in the Window OS and the malware will not do anything further when the value of this key suggests that the user has been infected before. In the following, we name this spreading mechanism as ***nonreinfection***.

However, modern email malware is far more aggressive in spreading throughout email networks than before. Without checking if a computer has been infected before, modern email malware makes use of every chance to spread itself. We characterize its propagation with two kinds of new mechanisms, namely ***reinfection*** and ***self-start***.

2.1 Problem from Technical Perspective

Reinfection, as the name suggests, indicates a user may get infected whenever the user visits malicious hyperlink or attachments. The reinfection outperforms the nonreinfection in two aspects: 1) a user can be infected again even if the user has been infected before; 2) a user will send out a malware copy each time the user gets infected. Thus, a recipient may repeatedly receive malware emails from the same compromised user.

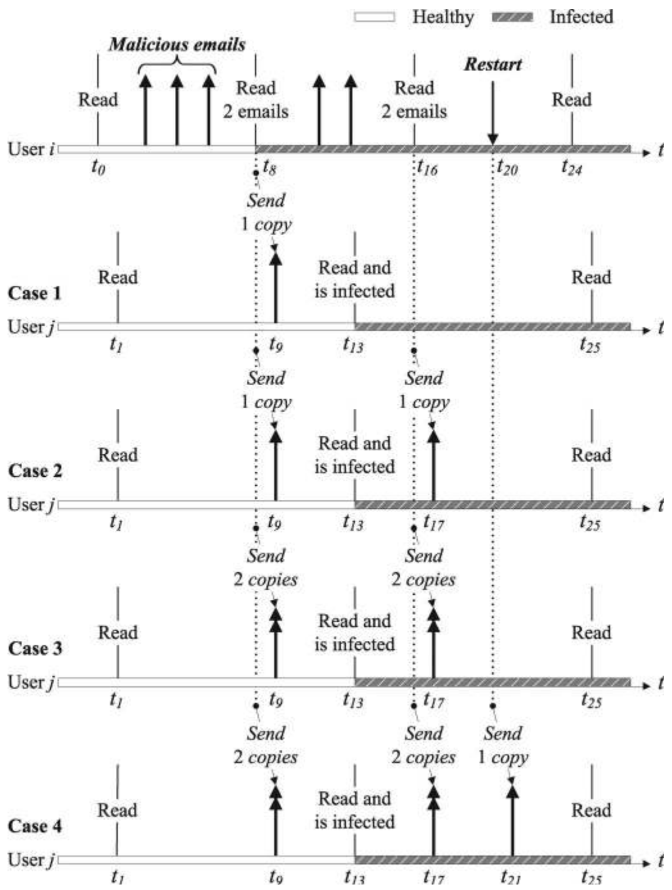


Fig. 1. Recipient user j's behavior for different types of malware emails. User i reads two of three malware emails at t_8 and another two malware emails at t_{16}, and then restarts at t_{20}. Case 1: nonreinfection; Case 2: reinfection in the work [3]; Case 3: reinfection of modern email malware; Case 4: both self-start mechanism in modern email malware. We assume a user will visit the malicious hyperlink or attachment if the user reads emails in this figure.

We illustrate the reinfection process in Fig. 1. Suppose an email user i gets infected and sends out malware email copies to another email user j. In case 1 of the nonreinfection, although user i reads two malware emails at t_8, the user will get infected and send only one malware copy to user j at t_8. The malware email arrives at user j at t_9. Then, when user j checks mailbox at t_{13} and reads the malware email from user i, user j gets infected. User j will not receive any more malware emails from user i after t_9. Nevertheless, in case 3 of the reinfection, user j will receive two malware copies from user i at t_9. Furthermore, after user j gets infected at t_{13}, when user i reads another two malware emails, user j receives another two malware copies from user i at t_{17}. Compared with case 1 of the nonreinfection, user j in case 3 of the reinfection receives totally four malware emails.

Generally, it is common for the malware emails to reuse the themes but with slight variations on the body of the message and the attachment names. This trick increases the possibility for a user to be infected and particularly prompts the spreading efficiency of the modern reinfection email malware. In Table 1, we list some types of email malware

TABLE 1
The Number of Themes in Some Email Malware

Name	Subject	Message	Attachment	Hyperlink
Sircam	random	random	random	none
SoBig	13	2	13	none
Mydoom	19	8	26	none
Nyxem	23	8	36	none
NetSky	1	1	25	none
W32.Imsolk	2	7	none	12

The statistical results come from Symantec Security Response [18].

with the number of their themes. In this paper, we assume every malicious email has different themes.

In fact, reinfection is also not enough to describe the propagation of modern email malware [12]. In many cases, they modify registry entries in Windows OS and the spreading process can be triggered whenever compromised computers restart or certain files are opened by infected users. Take Mydoom for example, it runs every time Windows starts.

We also illustrate the outperformance of the self-start in Fig. 1, case 4. User i has been infected at t_8. When the user restarts the computer at t_{20}, a malware email copy will be sent to user j in case 4 of the self-start. Compared with the nonreinfection and the reinfection, user j receives totally five malware emails. Thus, this maneuver has promoted the spreading efficiency of modern email malware.

2.2 Problem from Empirical Perspective

Currently, many models have been presented to model the propagation of email malware. For example, the works [3], [4], [5], [6], [7] present the nonreinfection; the works [3], [13] model the reinfection and the work [7] also discusses the self-start. However, the previous models are not appropriate for the modeling. We explain the reasons in the following. It is composed of two questions:

First, can we use the models of the nonreinfection to present the propagation dynamics of modern email malware? Compared with the reinfection and the self-start, to model the nonreinfection is simple. By using simulations [3], [5], [7] or analytical methods [4], [6], previous models have precisely presented the propagation dynamics of the nonreinfection. However, as stated in [7], [12], most real email malware is the self-start email malware. We simulate various spreading mechanisms and provide their differences (Δ) to the nonreinfection in the inset figure of Fig. 2. We can see that the self-start has a peak difference of 4×10^4 infected incidents. Thus, the self-start can spread ***much faster*** than the nonreinfection and the previous nonreinfection models ***cannot*** be used to present the propagation of modern email malware.

Second, can we use the previous models of the reinfection and the self-start to describe the propagation of modern email malware? The differential equation model adopted in [13] has been proven by the earlier work [3] to overestimate the spreading speed by 20 percent. Because the work [7] does not provide enough details in modeling the reinfection and the self-start, we mainly refer to the work in [3]. The model is illustrated in case 2 of Fig. 1 on the basis of their implementation [19]. We can see that user i always sends out only one malware copy to user j even if the user is infected by two malware emails at both t_8 and t_{16}. Compared with the reinfection discussed in Section 2.1, user j receives totally two malware emails in this case. We also present the numerical results from the simulations of the reinfection [3] in Fig. 2. We can see that the reinfection presented in [3] has ***noticeable differences*** to the self-start and the reinfection discussed in Section 2.1. Thus, the previous models of the reinfection and the self-start ***cannot*** be used in the propagation of modern email malware.

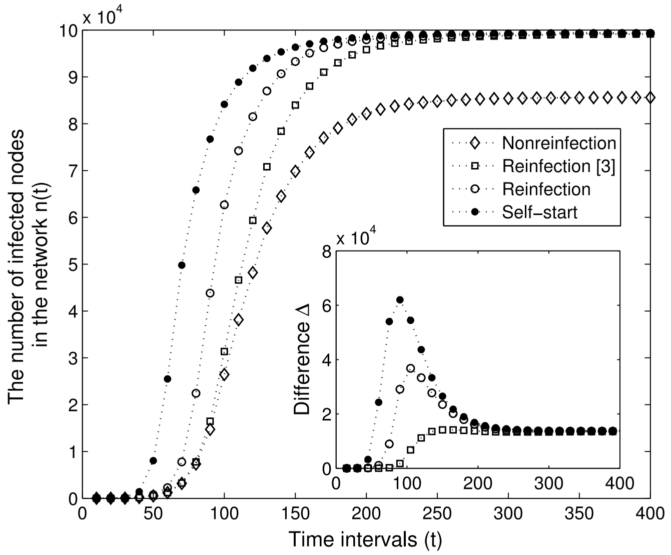


Fig. 2. The propagation of email malware in an email network with 10^6 users. The results are averaged from simulations of 100 times. The inset figure provides the differences (Δ) of various spreading mechanisms to the nonreinfection mechanism.

3 SII Model

In order to overcome the inaccuracy of previous models, we extend our previous SII model [8] for modern email malware. SII model is different from SIS and SIR models [20] because ***both*** susceptible and infected users can be immunized and never become susceptible again.

3.1 Modeling Nodes, Topology and Events of User

Nodes and topology information are the basic elements for the propagation of modern email malware. A node in the topology represents a user in the email network. Let random variable $X_i(t)$ denote the state of a node i at discrete time t. Then, we have

$$X_i(t) = \begin{cases} \boldsymbol{Hea.}, \mathit{healthy} \begin{cases} \boldsymbol{Sus.}, & \mathit{susceptible} \\ \boldsymbol{Imm.}, & \mathit{immunized} \end{cases} \\ \boldsymbol{Inf.}, \mathit{infected} \begin{cases} \boldsymbol{Act.}, & \mathit{active} \\ \boldsymbol{Dor.}, & \mathit{dormant.} \end{cases} \end{cases} \tag{1}$$

The state transition graph of an arbitrary node i in an email network is shown in Fig. 3. All nodes in networks are initially susceptible. Since infected users will send out malware copies when they are compromised, node i transits from the *susceptible* state to the *active* state after the user of node i gets infected. The infection probability is denoted by $v(i,t)$. The user is infectious at the *active* state. When a user is infected but not infectious, the node of this user transits to the *dormant* state. Besides, any user can be compromised again even if the user has been infected before. We represent the infection probabilities of an arbitrary node being at the

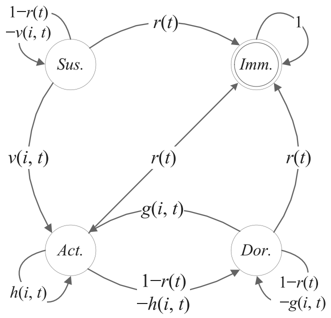


Fig. 3. State transition graph of a node in email topology. "Sus.": healthy but susceptible; "Act.": a user is infectious and will send out malware e-mail copies; "Dor.": a user is infected but not yet infectious; "Imm.": healthy and will never be infected again.

dormant state and the *active* state as $g(i,t)$ and $h(i,t)$ respectively. Whatever the state an arbitrary node is at, it may transit to the *immunized* state. The probability of immunization is denoted by $r(t)$. In fact, if the values of $g(i,t)$ and $h(i,t)$ are equal to zero, any infected node i will stay at the *dormant* state until the user of this node is immunized. In this scenario, Fig. 3 will be simplified as the state transition representation of the nonreinfection email malware. For the convenience of readers, we list major notations of this paper in Table 2.

In our SII model, we propose employing an M by M square matrix with elements p_{ij} to describe a topology consisting of M nodes, as in

$$\begin{pmatrix} p_{11} & \cdots & p_{1M} \\ \vdots & p_{ij} & \vdots \\ p_{M1} & \cdots & p_{MM} \end{pmatrix} p_{ij} \in [0,1], \tag{2}$$

wherein p_{ij} represents the probability of user j visiting a deceptive malware email received from user i. If p_{ij} is equal to zero, it means the email address of user j is not in the contact list of user i. Therefore, the matrix reflects the topology of an email network. In this paper, we assume the states of neighboring nodes are independent. This assumption has been adopted by previous models [7], [21], [22], [23]. In Section 6.3, we will analyze the impact of the independence assumption. Readers can refer to the work [4], [24] for more information of this assumption.

TABLE 2
Major Notations Used in This Paper

Symbol	Explanation
T_i	Email checking time of user i.
R_i	Event triggering period of user i.
$r(t)$	The recovery function of users, which provides the probability for any user to be immunized at time t.
$X_i(t)$	The state of a network node i at time t: "*Sus.*" susceptible, "*Imm.*" immunized, "*Act.*" active and "*Dor.*" dormant.
p_{ij}	The probability of user j visiting malware emails from user i.
$open_i(t)$	The event of user i checking newly arrived emails at time t.
$start_i(t)$	The event of user i restarting computer at time t.
τ	The arbitrary time between user i last checking emails and the current time t (excluding t).
M	The size of the Email network.
$n(t)$	The number of infected nodes in the Email network at time t.
$v(i,t)$	The infection probability of a susceptible node i at time t.
$g(i,t)$	The infection probability of a dormant node i at time t.
$h(i,t)$	The infection probability of an active node i at time t.
N_i	The set of neighboring nodes of node i.
$N_{i\mid N}$	The subset includes the real neighboring nodes of user i.
$N_{i\mid R}$	The virtual nodes caused by visiting more than one malware.
$N_{i\mid S}$	The virtual nodes caused by certain events triggered.
H_i	The extended set of N_i ($H_i = N_{i\mid N} + N_{i\mid R} + N_{i\mid S}$).
Δ	The numerical differences of various spreading mechanisms.
$\beta_{ji}(t)$	The probability of user i being infected by user j at time t.
$\zeta_i(t)$	The average value of $\beta_{ji}(t)$ for each $j \in H_i$.

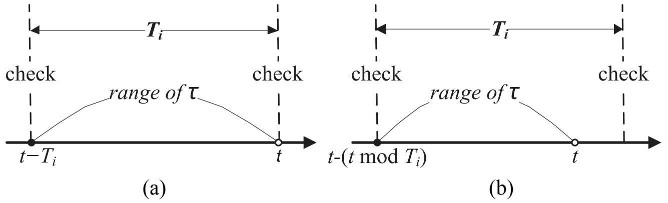


Fig. 4. Different cases of variable τ. (a) User checks new emails at current time t; (b) user does not check emails at current time t.

We have noticed that the infection of email malware depends on unwary email users checking new emails and visiting those malicious ones. In fact, this process involves two components in the modeling. First, we introduce a flag variable $open_i(t)$. We have $open_i(t) = 1$ if the user is checking new emails at time t, otherwise $open_i(t) = 0$. Let T_i denote the email checking period of user i, then we have

$$P(open_i(t) = 1) = \begin{cases} 0, & otherwise \\ 1, & t \bmod \quad T_i = 0. \end{cases} \tag{3}$$

Note that different users have different values of T_i. Readers can find more discussions about T_i at Section 6.4. An email user may receive multiple emails at different time but read all of them at one time when the user checks the mailbox. Supposing that an arbitrary user i checks new emails at time t, then those emails which will be checked at time t are the ones which arrived at user i after the user's last checking action of her mailbox. It is significant to obtain the number of such emails for our modeling. Thus, we introduce a variable τ to indicate an arbitrary time between the time of user i's last email checking action and the current time t (excluding t). As shown in Fig. 4, the value of τ has two forms depending on if user checks emails at current t or not. Then, we have

$$\begin{cases} t - T_i \le \tau \, < \, t, & \text{if } open_i(t) = 1 \\ t - (t \bmod T_i) \le \tau \, < \, t, & \text{otherwise.} \end{cases} \tag{4}$$

A compromised user can only spread malware to the neighboring users in email networks. Thus, for each email user in networks, we record and accumulate every newly arrived malicious email from neighboring users at each τ, and finally obtain the joint infection probability of each user who checks those emails.

3.2 Modeling Propagation Dynamics

We use the values 0 and 1 to substitute the *healthy* state and the *infected* state, respectively. Given a topology of an email network with M nodes, the expected number of infected users at time t, $n(t)$, is computed as in

$$\begin{aligned} n(t) &= E\left[\sum_{i=1}^{M} X_i(t)\right] = \sum_{i=1}^{M} E[X_i(t)] = \sum_{i=1}^{M} P(X_i(t) = 1) \\ &= \sum_{i=1}^{M} P(X_i(t) = Inf). \end{aligned} \tag{5}$$

The expected number of infected nodes, $n(t)$, is ascribed to the sum of the probability of each node being infected at time t, $P(X_i(t) = Inf.)$. As shown in Fig. 3, a susceptible node can be compromised and be at the infected state, and an infected node can be recovered and be at the immunized state. The state transitions help us derive the computation of $P(X_i(t) = Inf.)$ by difference equations as follows:

$$\begin{aligned} P(X_i(t) = Inf.) = &(1 - r(t)) \cdot P(X_i(t-1) = Inf.) \\ &+ \ v(i,t) \cdot P(X_i(t-1) = Sus.). \end{aligned} \tag{6}$$

For the computation of $P(X_i(t) = Sus.)$, we have

$$\begin{aligned} P(X_i(t) = Sus.) = &1 - P(X_i(t) = Inf.) \\ &- P(X_i(t) = Imm.). \end{aligned} \tag{7}$$

Moreover, for the computation of $P(X_i(t) = Imm.)$, we have

$$\begin{aligned} P(X_i(t) = Imm.) = &P(X_i(t-1) = Imm.) + r(t) \cdot \\ &[1 - P(X_i(t-1) = Imm.)]. \end{aligned} \tag{8}$$

Once we obtain the values of $v(i,t)$ and $r(t)$, the value of $P(X_i(t) = Inf.)$ can be computed by the iteration of the above equations (6), (7), (8).

In fact, there are three preconditions for an arbitrary user being infected by email malware: 1) the user has not been immunized; 2) the user checks mailbox for new emails; 3) the user unwarily visits one received malware emails. When the first and the second preconditions are satisfied, we use $s(i,t)$ to represent the probability of user i visiting malware emails from neighboring nodes. Then, the infection probability $v(i,t)$ can be derived as in

$$v(i,t) = s(i,t) \cdot P(open_i(t) = 1) \cdot [1 - r(t)]. \tag{9}$$

In our SII model, an arbitrary user i visits malicious hyperlinks or attachments with probability p_{ji} when reading malware emails from a neighboring user j. We use N_i to denote the set of neighboring nodes of node i. Then, we can compute $s(i,t)$ as in

$$s(i,t) = 1 - \prod_{j \in N_i} (1 - p_{ji} \cdot P(X_j(\tau) = Act.)), \tag{10}$$

wherein the event $X_j(\tau) = Act.$ means node j is infected and sends out a malware email copy to neighboring nodes at time τ. Considering different values that the variable τ may take, we disassemble the equation (10) by excluding $t-1$ from the range of value τ. There are two cases. First, as shown in Fig. 5a, user does not check new emails in the mailbox at time $t-1$. Thus, we have

$$\begin{aligned} &\prod_{j \in N_i} (1 - p_{ji} \cdot P(X_j(\tau) = Act.)) \\ &= \prod_{j \in N_i, \tau \neq t-1} (1 - p_{ji} \cdot P(X_j(\tau) = Act.)) \\ &\quad \times \prod_{j \in N_i} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)) \end{aligned} \tag{11}$$

$$= [1 - s(i,t-1)] \cdot \prod_{j \in N_i} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)). \tag{12}$$

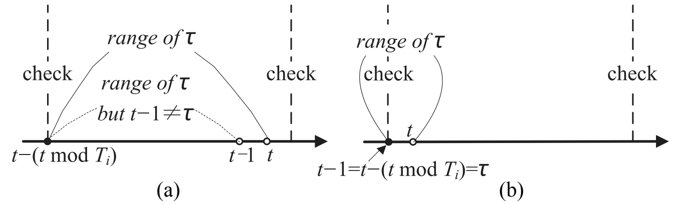


Fig. 5. Different cases for the computation of $s(i,t)$. User in (b) checks new emails at time $t-1$, but user in (a) does not.

Second, as shown in Fig. 5b, user checks new emails in the mailbox at time $t-1$. Thus, the malware email copies received at time t are those sent at time $t-1$ by the infected neighboring users. The variable τ only takes the value $t-1$. In this case, we have

$$\begin{aligned} &\prod_{j \in N_i} (1 - p_{ji} \cdot P(X_j(\tau) = Act.)) \\ &= \prod_{j \in N_i} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)). \end{aligned} \tag{13}$$

Actually, the difference of equations (12), (13) is caused by user checking newly arrived emails at time $t-1$. We then unify the computation of $s(i,t)$ as in

$$\begin{aligned} s(i,t) = 1 - \{1 - s(i,t-1) \cdot [1 - P(open_i(t-1) = 1)]\} \cdot \\ \prod_{j \in N_i} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)). \end{aligned} \tag{14}$$

In equation (14), different measures of $P(X_j(t-1) = Act.)$ and N_i may lead to different spreading performance. We show the algorithm of our SII model in the supplementary file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TDSC.2013.49 (Algorithm 1).

3.3 Virtual Nodes

For modern email malware, recall that a compromised user may send out malware email copies to neighbors every time the user visits those malware hyperlinks or attachments. Malware emails are also sent out when certain events like computer restart are triggered. Thus, at an arbitrary time t, a user may receive multiple malware email copies from an identical neighboring user who has been compromised. In order to represent the repetitious spreading process of the reinfection and the self-start, we introduce *virtual nodes* to present the kth infection caused by infected users opening the kth malware email copy.

As shown in Fig. 6, node 1, 2, 3 send malware emails to node 4. When the user of node 4 visits those emails, the user gets infected. If the user of node 4 visits two malware emails, node 4 will send malware email copies twice to node 6. If the user of node 4 visits three malware emails, node 4 will send treble malware email copies to node 6. The spreading process of extra malware email copies is ***equivalent*** to two virtual nodes sending a malware copy to node 6.

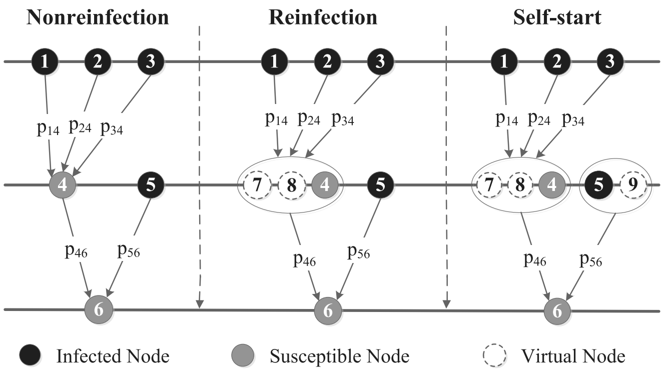


Fig. 6. An example to explain virtual nodes in the reinfection case and the self-start case. Node 1, 2, 3 send a malware copy to node 4.

We introduce virtual node 7 to denote the possible spreading if user 4 visits the second malware email. We also use virtual node 8 to denote the possible spreading if user 4 visits the third malware email. Moreover, when the user of the infected node 5 restarts computer or some specific events are triggered, this user will also send out a malware email copy to the user of node 6. It is also ***equivalent*** to a virtual node sending a malware copy to node 6. We introduce virtual node 9 to denote this process.

In order to represent the spreading process of virtual nodes, we extend N_i into a new set of neighboring nodes, H_i, which contains three subsets: $N_{i|N}$, $N_{i|R}$ and $N_{i|S}$. Then, we revise the equation (14) for modeling the propagation of modern email malware as in

$$\begin{aligned} s(i,t) = 1 - \{1 - s(i,t-1) \cdot [1 - P(open_i(t-1)=1)]\} \cdot \\ \prod_{j \in N_{i|N}} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)) \cdot \\ \prod_{j \in N_{i|S}} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)) \cdot \\ \prod_{j \in N_{i|R}} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)). \end{aligned} \tag{15}$$

First, the subset $N_{i|N}$ includes the real neighboring nodes of user i (e.g. $N_{6|N}$ = node 4, node 5). In fact, the nodes in $N_{i|N}$ represent the neighboring users who visit the first malware email copy and get infected. Since the states of neighboring nodes are independent, each node is infected by neighboring nodes regardless of the state of this node [4]. Thus, we simply consider $v(i,t) = g(i,t) = h(i,t)$ (See Section 6.3 for discussion). Then, the value of $P(X_j(t-1) = Act.)$ for user j in $N_{i|N}$ can be derived as

$$\begin{aligned} P(X_j(t-1) = Act.) = v(j,t-1) \cdot P(X_j(t-2) = Sus.) \\ + g(j,t-1) \cdot P(X_j(t-2) = Dor.) \\ + h(j,t-1) \cdot P(X_j(t-2) = Act.) \end{aligned} \tag{16}$$

$$= v(j,t-1) \cdot [1 - P(X_j(t-2) = Imm.)]. \tag{17}$$

Second, the subset $N_{i|S}$ includes the virtual nodes which present the extra spreading processes caused by certain events triggered in infected nodes (i.e., $N_{6|S}$=node 9). In this case, we introduce another flag variable $start_i(t)$. We have $start_i(t) = 1$ if the events happen at time t, otherwise $start_i(t) = 0$. Assuming the events are periodically triggered and R_i is the event triggering period of user i, we have

$$P(start_i(t) = 1) = \begin{cases} 0, & otherwise \\ 1, & \text{t mod } R_i = 0. \end{cases} \tag{18}$$

Then, we can compute the value of $P(X_j(t-1) = Act.)$ for user j who belongs to $N_{i|S}$ as in

$$\begin{aligned} P(X_j(t-1) = Act.) = P(start_j(t-1) = 1) \cdot \\ P(X_j(t-1) = Inf.). \end{aligned} \tag{19}$$

Third, the subset $N_{i|R}$ includes the virtual nodes which present the extra spreading processes caused by users visiting more than one malware copies when they check new emails (i.e., $N_{6|R}$ = node 7, node 8). In fact, this is a permutation problem. For example, node 7 in Fig. 6 presents the spreading process caused by user 4 visiting the second malware email. Thus, the value of $P(X_7(t-1) = Act.)$ indicates the probability of user 4 visiting any two or three malware emails from user 1, 2 and 3. We use $\beta_{ji}(t)$ to present the probability of user i being infected by user j at time t and $\overline{\beta_{ji}}(t)$ to indicate the negation of $\beta_{ji}(t)$ as in

$$\begin{cases} \beta_{ji}(t) = p_{ji} \cdot P(X_j(\tau) = Act.) \cdot P(open_i(t) = 1) \\ \overline{\beta_{ji}}(t) = 1 - p_{ji} \cdot P(X_j(\tau) = Act.) \cdot P(open_i(t) = 1). \end{cases} \tag{20}$$

We then derive the value of $P(X_7(t-1) = Act.)$ as in

$$\begin{aligned} P(X_7(t-1) = Act.) &= \beta_{14}(t-1)\beta_{24}(t-1)\overline{\beta_{34}}(t-1) \\ &+ \beta_{14}(t-1)\overline{\beta_{24}}(t-1)\beta_{34}(t-1) \\ &+ \overline{\beta_{14}}(t-1)\beta_{24}(t-1)\beta_{34}(t-1) \\ &+ \beta_{14}(t-1)\beta_{24}(t-1)\beta_{34}(t-1) \\ &= P(X_4(t-1) = Act.) - \beta_{14}(t-1)\overline{\beta_{24}}(t-1)\overline{\beta_{34}}(t-1) \\ &- \overline{\beta_{14}}(t-1)\beta_{24}(t-1)\overline{\beta_{34}}(t-1) \\ &- \overline{\beta_{14}}(t-1)\overline{\beta_{24}}(t-1)\beta_{34}(t-1). \end{aligned}$$

Node 8 in Fig. 6 presents the spreading process caused by user 4 visiting the third malware email. Similarly, we compute the value of $P(X_8(t-1) = Act.)$ as in

$$\begin{aligned} P(X_8(t-1) = Act.) &= \beta_{14}(t-1)\beta_{24}(t-1)\beta_{34}(t-1) \\ &= P(X_7(t-1) = Act.) \\ &- \beta_{14}(t-1)\beta_{24}(t-1)\overline{\beta_{34}}(t-1) \\ &- \beta_{14}(t-1)\overline{\beta_{24}}(t-1)\beta_{34}(t-1) \\ &- \overline{\beta_{14}}(t-1)\beta_{24}(t-1)\beta_{34}(t-1). \end{aligned}$$

If a user is popular, the node of this user may have many neighbours. For popular users, the permutation problem may become very complex. For example, we assume an arbitrary node i has m neighbours, to compute the probability of user i opening the second malware email from those m neighbours, we have to run C_m^2 combinations of multiplication of $\beta_{ji}(t)$. If the user of each node opens at most k emails, the computation for each run each node is totally

$\sum_{k=1}^{k=d} C_m^k$. Noticeably, it is computationally too expensive to obtain the result. Thus, we introduce Bernoulli approximation. The Bernoulli experiment is widely used to model the number of successes in a sample drawn from a large population. We can see that virtual nodes provide us a series of easy-derived equations which can then be used to compute the probabilities of multiple infections using the Bernoulli approximation. This helps us solve the combination problem. We use $\zeta_i(t)$ to denote the average value of $\beta_{ji}(t)$ for each neighboring node j ($j \in H_i$, $H_i = N_{i|N} + N_{i|R} + N_{i|S}$). Then, we have

$$\zeta_i(t) = \frac{1}{\| H_i \|} \sum_{j \in H_i} \beta_{ji}(t). \tag{21}$$

Moreover, we use k to denote the order of the malware emails that user j visits, such as $k = 2$ when the user visits the second malware email and $k = 3$ when user j visits the third one. We have the value of $P(X_{j\_k}(t-1) = Act.)$ for user j who belongs to $N_{i|R}$ as in

$$P(X_{j\_1}(t-1) = Act.) = P(X_j(t-1) = Act.) \qquad (k = 1), \tag{22}$$

$$\begin{aligned} P(X_{j\_k}(t-1) = Act.) &= P(X_{j\_(k-1)}(t-1) = Act.) \\ &- \left(\prod_{j=1}^{k} \beta_{ji}(t) \prod_{j=k+1}^{\|H_i\|} \overline{\beta_{ji}}(t) + \cdots + \prod_{j=1}^{\|H_i\|-k} \overline{\beta_{ji}}(t) \prod_{j=\|H_i\|-k+1}^{\|H_i\|} \beta_{ji}(t) \right). \end{aligned} \tag{23}$$

$$\begin{aligned} &\approx P(X_{j\_(k-1)}(t-1) = Act.) \\ &\quad - C_{\|H_i\|}^{k} [\zeta_i(t)]^k [1 - \zeta_i(t)]^{\|H_i\|-k} \qquad (k \geq 2). \end{aligned} \tag{24}$$

In fact, the value of k also reflects the vigilance of email users for opening malicious hyperlinks or malware attachments. If the email malware becomes more deceptive, the value of k becomes larger. We will carry out extensive discussion about k in Section 6.2. We use equations (15), (17), (19), (22) and (24) to model the propagation. Particularly, we show the iteration of $s(i, t)$ and the computation of virtual nodes in the supplementary file, available online (Algorithm 2 and 3 respectively).

4 Model Evaluation

4.1 Experiment Environment

In this field, all existing research adopts simulation to evaluate analytical models, such as [4], [6]. We follow this approach to evaluate the proposed SII model based on simulations. In real-world scenarios, the spread of most email malware is typically impossible to track given the directed, topological manner in which they spread. Some email malware, like Nyxem [17], once compromising a computer, will automatically generate a single http request for the URL of an online statistics page. However, as the report [17] said, the statistics of Nyxem also cannot present a precise investigation on the spread of email malware due to the legitimate access, repeated probes and DDoS attacks to the web page. It should be pointed out that there is no real data set available for the evaluation of models of modern email malware.

The email topology is a key component of simulation. Some existing work [3], [4], [20] shows that topological factors have strong impact on the speed and scale of the malware propagation. In this paper, we build the topology according to the previous analysis of real email networks [25], [26], which exhibit "semi-directed", scale-free and small-world properties. The topology has 100,000 nodes. We reproduce the degree for each node by the Power-law distribution [27]. Moreover, the probability of users being infected by their friends (p_{ij}), the email checking period (T_i) and the event triggering period (R_i) are mainly decided by human factors. Similar to [19], these parameters will follow the Gaussian distribution. Note that the Gaussian distribution generator may provide unrealistic values, such as $p_{ij} < 0$ and $T_i < 1$. In our experiment, we replace these values with the minimums of their realistic range. Thus, if $p_{ij} < 0$, $T_i < 1$ and $R_i < 1$, we let $p_{ij} = 0, T_i = 1$ and $R_i = 1$.

We draw an SII-compatible propagation simulator from existing simulation models [3], [5], [28], [29]. The implementation is in C++ and Matlab 7. The random numbers in the experiments are produced by the C++ TR1 library extensions. We run each simulation 100 times for an average result. The number 100 comes from the discussion "how many simulation runs are needed before we obtain a steady curve? [3]". Each run of the spread begins with two infected nodes, which are randomly chosen from the topology. These two nodes keep a topology distance of 6 (the number of edges between them) so that certain clustering coefficient [25] is implicitly implemented. For the convenience of readers, we have put the source codes of the SII model and the simulator online [30].

4.2 Comparison with Previous Models

To evaluate the accuracy of our model, we conduct experiments with different parameter settings. Most values in the settings come from previous works [3], [25], [26].

We compare our SII model with previous models [3], [4], [5], [6]. The work in [3] presents the modeling of the reinfection without virtual nodes. The models in [4], [5], [6] present the propagation of the nonreinfection. In this experiment, the topology has the properties: the power-law exponent $\alpha = 2.5$, the average degree $E(D) = 5.5$. The values of p_{ij} follow Gaussian distribution $N(0.5, 0.2^2)$. In order to exclude the impact of recovery processes, the experiment is carried out with $r(t) = 0$. We let T_i and R_i follow Gaussian distribution $N(40, 20^2)$. The vigilance k of email users is set to be 5. As shown in Fig. 7, our SII model is far more accurate than previous models. We exhibit the differences Δ in the inset of Fig. 7. We can see that the results of previous models deviate from simulations by 80 thousands less infections at maximum. There is also a minor divergence between the results of SII model and simulations. As explained in [4], this difference is caused by the independent assumption. We have presented an extensive discussion on the impact of this assumption in Section 6.3.

4.3 Impact of Parameters in the Modeling

We also evaluate the impact of various parameters on the accuracy of the modeling.

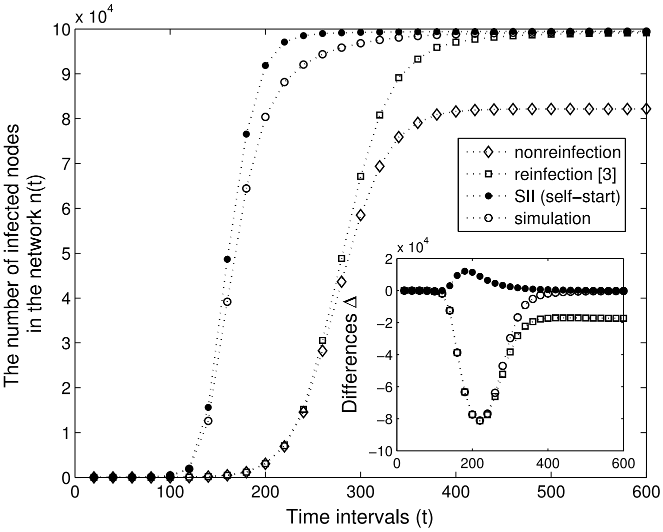


Fig. 7. The comparison between SII model and previous models. $p_{ij} \sim N(0.5, 0.2^2)$, $\alpha = 2.5$, $E(D) = 5.5$, $k = 5$. Δ denotes the differences between the results of SII model, previous models and the simulations.

First, we evaluate the accuracy with different distributions of T_i and R_i. In this experiment, the topology has the same settings as in Fig. 7. As shown in Fig. 8, the curves of our SII model are close to the simulations even if the distributions of T_i and R_i are different.

Second, we also evaluate the accuracy with different distributions of p_{ij}. The same topologies are used in this experiment. We let T_i and R_i follow Gaussian distribution $N(40, 20^2)$. As shown in Fig. 9, the results of our SII model are close to the results of simulations. In the inset figure of Fig. 9, we can also see that the SII model achieves better performance in accuracy when the infection probabilities p_{ij} are averagely higher. For the same reason of the independent assumption, we can achieve better accuracy once we relax this assumption in the future modeling.

Third, we evaluate the accuracy in different topologies. In this experiment, we let T_i and R_i follow Gaussian distribution $N(40, 20^2)$ and the infection probability p_{ij} follow $N(0.5, 0.2^2)$. As shown in Fig. 10, our SII model is effective in various topologies with different power-law exponents α and means of degrees $E(D)$.

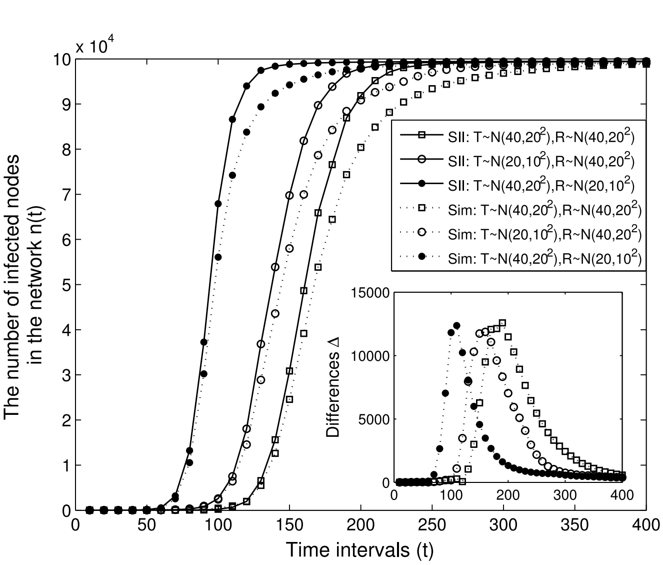


Fig. 8. The accuracy with different distributions of T_i and R_i. $p_{ij} \sim N(0.5, 0.2^2)$, $\alpha = 2.5$, $E(D) = 5.5$. Δ denotes the differences between the results of our SII model and the simulations.

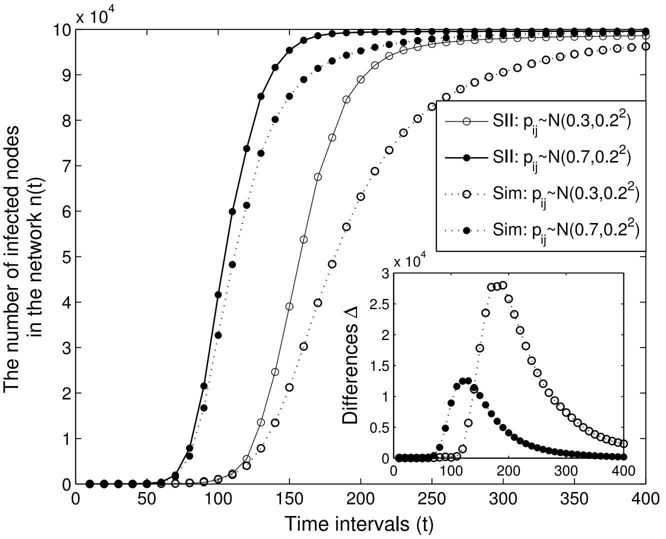


Fig. 9. The accuracy with different distributions of p_{ij}. $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$, $\alpha = 2.5$, $E(D) = 5.5$. Δ denotes the differences between the results of our SII model and the simulations.

Finally, we evaluate the accuracy with recovery functions $r(t)$. We consider two recovery functions: 1) constant recovery rate ρ ($r(t) = \rho$); 2) Qualys rate. According to the statistics of Qualys Inc. [31], after detection, the number of susceptible and infected users decreases by 50 percent of the remaining every 30 days in 2003 and 21 days in 2004. We use a variable $d1$ to denote the temporal span from the malware starting spreading to scientists having found this malware on the Internet. During the temporal period $d1$, modern email malware can spread freely on the Internet ($r(t) = 0$). We introduce another variable $d2$ to denote the temporal span of 50 percent decreasing. Then, we have the Qualys rate of the recovery functions as in

$$r(t) = \begin{cases} 0, & t < d1 \\ 1 - 0.5^{\frac{t-d1}{d2}}, & t \geq d1. \end{cases} \tag{25}$$

In this experiment, the topology has the properties: the power-law exponent $\alpha = 2.5$, the average degree $E(D) = 5.5$. The values of p_{ij} follow Gaussian distribution $N(0.5, 0.2^2)$. The values of T_i and R_i follow $N(40, 20^2)$. We can see in Fig. 11 that our SII model are accurate compared with the simulations. This means the SII model is suitable for the propagation modeling of modern email malware.

5 Theoretical Justification

The empirical study has shown our SII model is superior to previous models [3], [4], [5], [6]. We further provide the theoretical justification in modeling the spreading mechanism and state transition of the propagation.

5.1 Superiority in the Spreading Mechanisms

Recall that modern email malware has two aggressive spreading mechanisms. The first one is caused by the

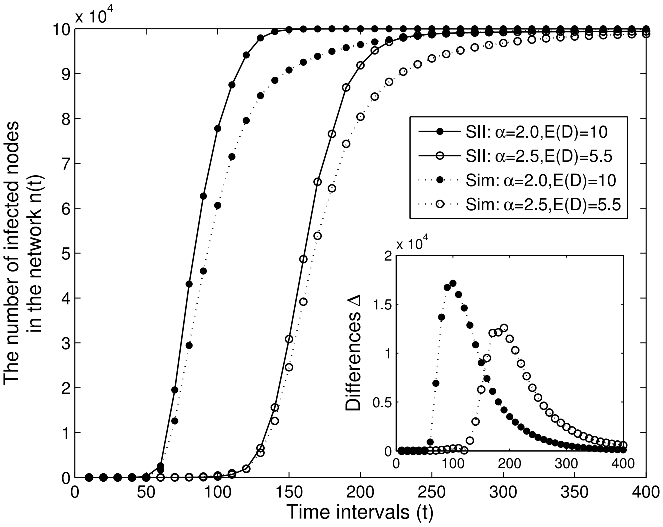


Fig. 10. The accuracy in different topologies (α and $E(D)$). $p_{ij} \sim N(0.5, 0.2^2)$, $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$. Δ denotes the differences between the results of our SII model and the simulations.

reinfection: any user can be infected again even if this node has been infected before. The second one is the repetitious spreading process caused by the reinfection and the self-start: any infected user spreads malware e-mail copies every time the user visits malware emails or the infected computer restarts.

For the first mechanism, we have the equation (17) in SII model. We can also present previous models [4], [5], [6] by setting the probabilities $g(i,t)$ and $h(i,t)$ to be zero. Considering the number of infected nodes $n(t)$, we have $n(t) \propto P(X_j(t) = Act.)$ for $\forall j \in N_i, i \in [1, M]$ by equations (5), (6), (9), (14). We use ω_1 and ω_2 to denote $P(X_i(t) = Act.)$ for previous models [4], [5], [6] and the SII model respectively. Then, we have

$$\omega_1 = v(i,t) \cdot P(X_i(t-1) = Sus.), \tag{26}$$

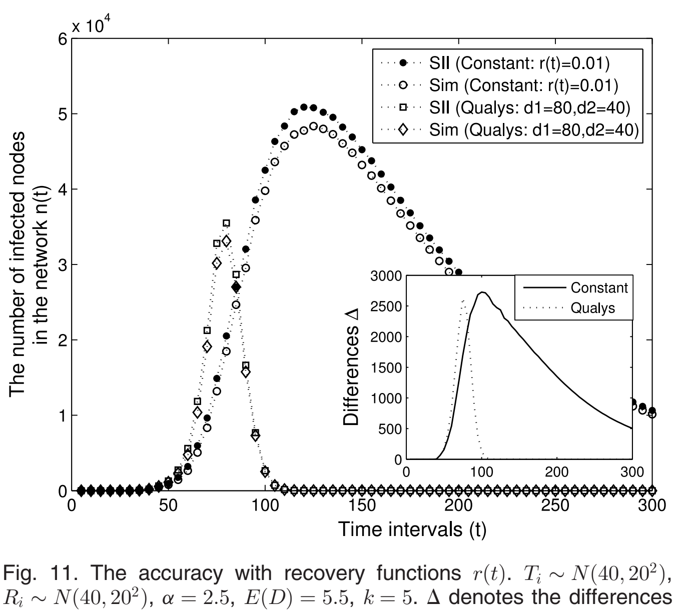


Fig. 11. The accuracy with recovery functions $r(t)$. $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$, $\alpha = 2.5$, $E(D) = 5.5$, $k = 5$. Δ denotes the differences between the results of our SII model and the simulations.

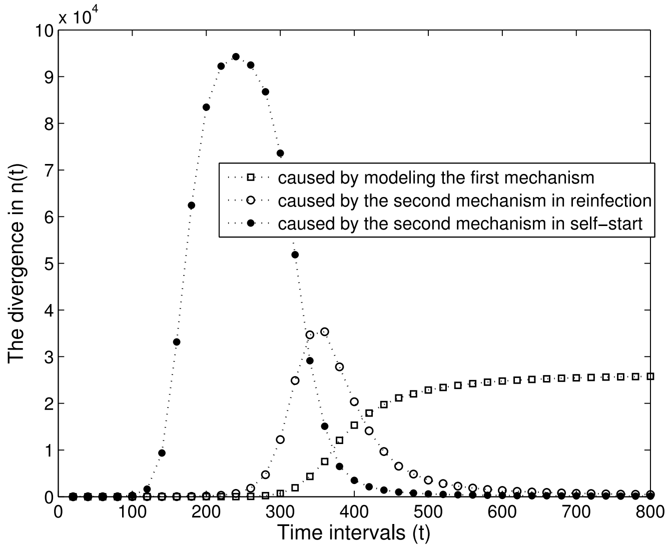


Fig. 12. The divergence of $n(t)$ for modeling various mechanisms. $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$, $\alpha = 2.5$, $E(D) = 5.5$, $k = 5$, $r(t) = 0$.

$$\omega_2 = v(i,t) \cdot [1 - P(X_i(t-1) = Imm.)]. \tag{27}$$

Obviously, we have

$$\omega_2 - \omega_1 = v(i,t) \cdot P(X_i(t-1) = Inf.) > 0. \tag{28}$$

Given an email network and a spreading case of modern e-mail malware on it, we investigate the divergence of $n(t)$ caused by $\omega_2 - \omega_1 > 0$. In order to eliminate the impact of the second mechanism, we set $N_{i|R}, N_{i|S} = \Phi$. As shown in Fig. 12, we can see the divergence caused by modeling the first spreading mechanism reaches more than 20 thousands.

For the second mechanism, we have the equation (15) in SII model. In order to present the repetitious spreading processes, we extend the set of neighboring users by virtual nodes. Note that we have $H_i > N_i$. In model [3], no matter how many malware emails an infected user visits, only one malware email copy will be sent out. Thus, we have $N_{i|R}, N_{i|S} = \Phi$ and $H_i = N_i$ for the model [3]. We use ω_3, ω_4 and ω_5 to denote the impact of modeling the spreading in [3], the reinfection process and the self-start process as in

$$\omega_3 = \prod_{j \in N_{i|N}} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)), \tag{29}$$

$$\omega_4 = \prod_{j \in N_{i|R}} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)), \tag{30}$$

$$\omega_5 = \prod_{j \in N_{i|S}} (1 - p_{ji} \cdot P(X_j(t-1) = Act.)). \tag{31}$$

Then, we can have

$$\omega_3 > \omega_3 \times \omega_4 > \omega_3 \times \omega_4 \times \omega_5. \tag{32}$$

According to the equation (15), this means nodes in the network are easier to be infected and become infectious if the nodes have larger neighboring sets. We investigate the divergence of $n(t)$ caused by the inequality (32) in modeling the

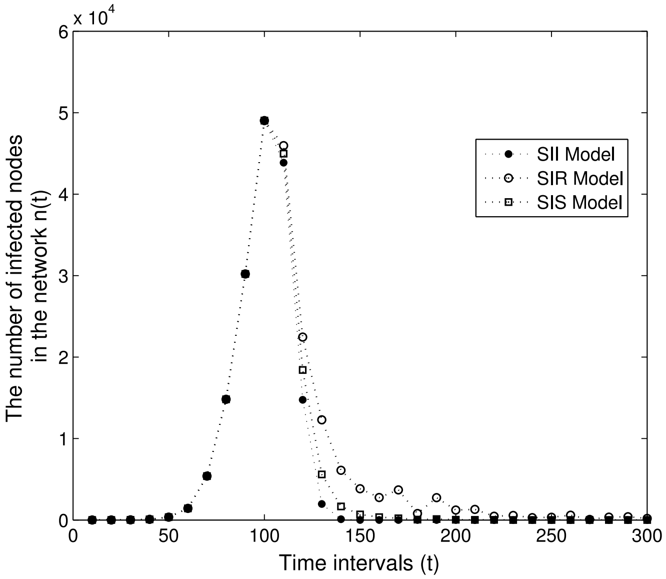


Fig. 13. The difference of SII, SIR and SIS models. $p_{ij} \sim N(0.5, 0.2^2)$, $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$, Qualys recovery function: $d1 = 80$, $d2 = 40$.

reinfection. As shown in Fig. 12, we can see the maximal divergence of modeling the second mechanism reaches 35,000. Particularly, by modeling the self-start, the divergence reaches 95,000.

On the basis of above analysis, we can see that our SII model is able to present the propagation of modern email malware. The divergence between $n(t)$ and its estimation is large in previous models [3], [4], [5], [6]. Thus, previous models cannot be used in modeling the propagation of modern email malware.

5.2 Superiority in Modeling State Transitions

We compare our SII model with SIS models [4], [21], [22], [23], [32], [33] and SIR models [20], [34]. The difference among these models is caused by different considerations on the state transition of nodes. SIS models assume infected nodes become susceptible again after recovery. If infected nodes cannot become susceptible again once they are cured, the models are called SIR models. Considering the propagation of modern email malware, after users clean their infected computers or become more vigilant against a type of malware, they are unlikely to be infected any more. Therefore, SIS models are not appropriate to model the propagation of modern email malware. SIR models may suit for modern email malware, but the real case is that a susceptible user can be immunized directly without being infected at first. Thus, the state transition of our SII model is similar to SIR model except nodes at the susceptible state can directly transit to the *immunized* state.

In order to exclude the impact of other factors, we derive the SIS and SIR models on the basis of the SII model. First, a susceptible user can be immunized in SII model, but not in SIR model. Thus, we can revise equation (8) to obtain an SIR model as in

$$\begin{aligned} P(X_i(t) = Imm.) = P(X_i(t-1) = Imm.) + r(t) \cdot \\ P(X_i(t-1) = Inf.). \end{aligned} \tag{33}$$

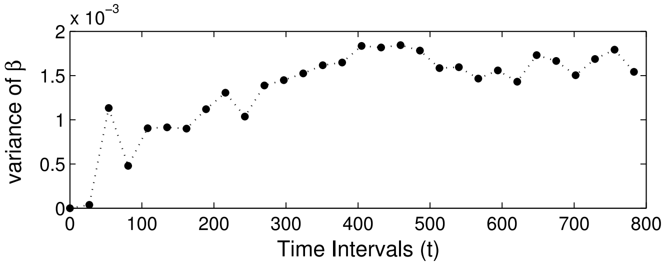


Fig. 14. The variance of β on the node with maximal degree. $p_{ij} \sim N(0.5, 0.2^2)$, $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$.

Second, an SIS model does not have the immunized state. We can have it by setting $P(X_i(t) = Imm.) = 0$ and revising equation (7) as in

$$\begin{aligned} P(X_i(t) = Sus.) = (1 - v(t)) \cdot P(X_i(t-1) = Sus.) \\ + r(t) \cdot P(X_i(t-1) = Inf.). \end{aligned} \tag{34}$$

As shown in Fig. 13, the results of SII model decrease more rapidly than SIR and SIS models. Thus, we cannot use traditional SIS and SIR models to model the propagation of modern email malware.

6 Further Discussion and Limitations

In this section, we will discuss the limitations of our proposed model. The experiments adopt typical settings: $p_{ij} \sim N(0.5, 0.2^2)$, $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$. Moreover, the recovery functions will not be considered in this section ($r(t) = 0$).

6.1 Test of Bernoulli Approximation

It is computationally too expensive to calculate the real value of the probability of virtual nodes being infectious, particularly when the number of virtual nodes is large. Thus, we use the average value $\zeta_i(t)$ to substitute each $\beta_{ji}(t)$ ($j \in H_i$) and apply the Bernoulli approximation on this probability. In this section, we adopt the variance of $\beta_{ji}(t)$, $var_i(t)$, to investigate the accuracy of the Bernoulli approximation, as in

$$var_i(t) = \sum_{j=1}^{\|H_i\|} [\beta_{ji}(t) - \zeta_i(t)]^2 / \|H_i\|. \tag{35}$$

When the values of $var_i(t)$ are small, the values of $\beta_{ji}(t)$ are close to the average value $\zeta_i(t)$, which means the Bernoulli approximation is accurate.

In our experiment, we examine the node which has maximal degree, since it has a large number of virtual nodes in the modeling. As shown in Fig. 14, it turns out that the values of the variance $var_i(t)$ are rather small, on the order of 10^{-3}. Thus, we can approximate values of $\beta_{ji}(t)$ using the average value $\zeta_i(t)$.

6.2 The Effect of Users' Vigilance (the Value k)

Modern email malware infects unwary users when they open malicious email attachments or visit infectious hyperlinks in the email content. Users' vigilance determines the number of malicious emails that are opened by the users. The higher a user's vigilance is, the less malware emails are opened. As discussed in this paper, the vigilance of users

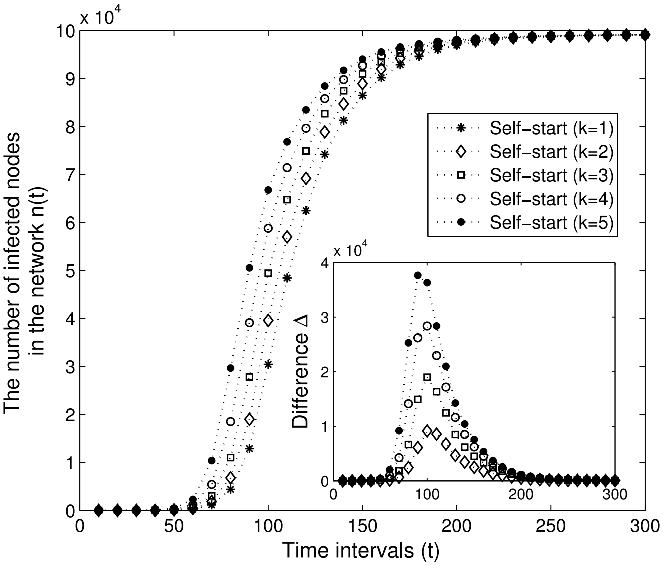


Fig. 15. The effect of users' vigilance. $p_{ij} \sim N(0.5, 0.2^2)$, $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$. **Δ** denotes the differences with varied values of k.

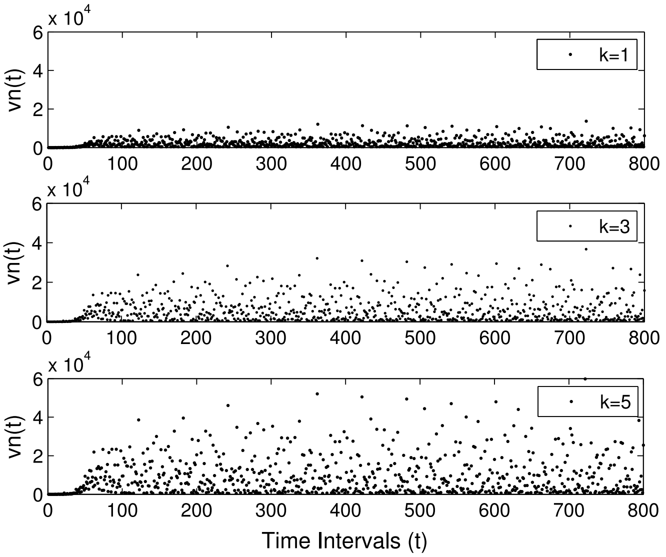


Fig. 16. The number of virtual nodes $vn(t)$ for different k. $p_{ij} \sim N(0.5, 0.2^2)$, $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$.

determines the number of virtual nodes for each user in the modeling, which greatly affects the spreading speed and scale. In this section, we analyze and quantify the effect of users' vigilance (the value k).

The value k presents the maximal number of malware emails that each user may visit. We run the SII model from $k = 1$ to $k = 5$ in self-start case. As shown in Fig. 15, the differences of $n(t)$ are large with varied values of k. For example, at time tick 100, the difference reaches about 10,000 for each increment in the value of k. Besides, we statistically investigate the number of virtual nodes that the method will add in computing the model's numerical results. We use $vn(t)$ to denote the number of virtual nodes at each time t. As shown in Fig. 16, larger values of k leads to an increment in the values of $vn(t)$. In practical terms, both the experiments suggest that lower vigilance of real-world users may speed up the outbreak of modern email malware on the Internet.

6.3 The Independent Assumption

We can see from Fig. 3 that the values $v(j,t)$, $g(j,t)$ and $h(j,t)$ denote the conditional probabilities as in

$$\begin{cases} v(j,t) = P(X_j(t) = Act. \mid X_j(t-1) = Sus.) \\ g(j,t) = P(X_j(t) = Act. \mid X_j(t-1) = Dor.) \\ h(j,t) = P(X_j(t) = Act. \mid X_j(t-1) = Act.). \end{cases} \tag{36}$$

Take $v(j,t)$ for example, we derive the equation as in

$$v(j,t) = P(open_j(t) = 1) \cdot [1 - r(t)] \cdot \left\{ 1 - \prod_{i \in N_j} (1 - p_{ij} \cdot P(X_i(t-1) = Act. \mid X_j(t-1) = Sus.)) \right\}. \tag{37}$$

However, in the above modeling, we assume the states of nodes are independent of each other. The infection of a node depends only on its neighboring nodes regardless of the state of this node. Thus, we can derive the following approximation in the equations (10) and (17) as in

$$v(j,t) = g(j,t) = h(j,t) = P(open_j(t) = 1) \cdot [1 - r(t)] \cdot \left\{ 1 - \prod_{i \in N_j} (1 - p_{ij} \cdot P(X_i(\tau) = Act.)) \right\}. \tag{38}$$

We can see the essence of the independent assumption is using marginal probability $P(X_i(t-1) = Act.)$ to substitute the conditional probabilities $P(X_i(t-1) = Act. | X_j(t-1) = Sus.)$, $P(X_i(t-1) = Act. | X_j(t-1) = Act.)$ and $P(X_i(t-1) = Act. | X_j(t-1) = Dor.)$. As stated in [4], this approximation may cause the inaccurate estimation in the modeling. In Section 4.2, we have seen the analytical results still deviate from simulations.

In order to show how many errors will be caused by the independent approximation, we introduce a variable $diff(a,b)$ to denote the difference between the marginal and the conditional probabilities as in

$$diff(a,b) = P(X_i(t-1) = a) - P(X_i(t-1) = a \mid X_j(t-1) = b) \tag{39}$$

$$= P(X_i(t-1) = a) - \frac{P(X_i(t-1) = a, X_j(t-1) = b)}{P(X_j(t-1) = b)}, \tag{40}$$

whereas $a = Act.$,$b \in \{Act., Sus., Dor.\}$. The node that has larger degree is easier to be affected by the independent assumption. Thus, we examine $diff(a,b)$ on a pair of neighboring nodes, one of which has maximal degree in the network. The probabilities in equation (40) are averaged by simulation of 1,000 individual runs. As shown in Fig. 17, $diff(a,b)$ is not equal to zero. Thus, the independent approximation may cause errors in the modeling. In addition, we examine the symmetric Kullback-Leibler divergence [35], D_{KL}, between the marginal and conditional probabilities. We can see in Table 3 that the result of D_{KL} (b="Sus.") is much larger than the ones when b="Act." or "Dor.". This means

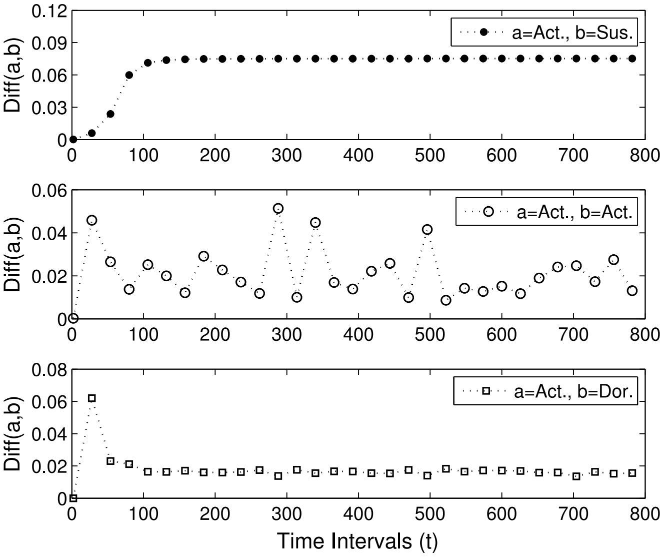


Fig. 17. The errors $diff(a,b)$ for the independent assumption. $p_{ij} \sim N(0.5, 0.2^2)$, $T_i \sim N(40, 20^2)$, $R_i \sim N(40, 20^2)$.

the case of using $P(X_i(t-1) = Act.)$ instead of $P(X_i(t-1) = Act.|X_j(t-1) = Sus.)$ will cause larger errors than another two cases.

In fact, the conditional probabilities are too expensive to obtain mathematically. Thus, most analytical models and analysis, such as [7], [21], [22], [23], assume nodes are independent of each other. In this paper, we follow the independent assumption, and mainly focus on presenting the reinfection and the self-start in the modeling. We plan to investigate how to relax the independent assumption in modeling the reinfection and the self-start in the future. Readers could refer to [4], [24] for possible solutions.

6.4 The Periodical Assumption on T_i and R_i

A premise of the above modeling is that user checks newly arrived emails and certain events, such as the restart of computers, are triggered at regular periods (T_i and R_i). However, in real-world situations, users may check new emails and trigger certain events at any time. Indeed, some people may check emails at 7 o'clock in the morning but at 17 o'clock in the afternoon of the next day. Nevertheless, most people may follow a long-term period of email checking time denoted by T_i. We can also assume a long-term period of R_i. The values of T_i and R_i depend only on users' own patterns. In the analytical modeling, it is reasonable to adopt long-term regular periods of T_i and R_i instead of varied checking time values and triggering time values for each user.

We plan to incorporate irregular checking time into our analytical SII model in the future. A possible solution is to assign new values of T_i for each user after the user checking new emails at current time t. The same operation can also be applied to the value of R_i. Then, our SII model could possibly be compatible with varied T_i and R_i for each user.

6.5 Discussion of Modeling Repetitious Infections

In this paper, we introduce virtual nodes in order to address the modeling of reinfection and self-start. Readers could also think of dividing the infected state into several sub-infected states. The kth sub-infected state indicates user having received k email malware copies ($0 < n < \|H_i\|$). However, this method is too computationally expensive as the infection probability will be calculated by $\sum_{k=1}^{k=d} C_m^k$ combinations ($m = N_i$). Moreover, the sub-infected states are difficult to be implemented when $\|H_i\|$ are large. On the contrary, the virtual node, which indicates user opening the kth malware email copy, can be easily derived and approximated. Currently, this is still an on-going work in our research. We plan to simplify the modeling the multiple infections in the future.

TABLE 3
KL Divergence in the Independent Assumption

	a=Act., b=Sus.	a=Act., b=Act.	a=Act., b=Dor.
D_{KL}	5.2390	0.0432	0.0446

7 Related Work

There have been substantial efforts in modeling the propagation dynamics of Internet malware in the last decade. ***First***, to model the epidemic spreading on topological networks, early researchers adopt differential equations to present the propagation dynamics of malware. However, as discussed in [3], the differential models [20], [32], [34] greatly overestimate the spreading speed due to the "homogeneous mixing" assumption. Additionally, Zou et al. [3] and Gao et al. [5] rely on simulations to model the spread of email malware. Their simulation models avoid the "homogeneous mixing" problem but cannot provide analytical propagation studies. The works [4], [8], [21], [23] propose mathematical models, which have captured the accurate topological information. Wen et al. [8] further addressed the temporal dynamics and the spatial dependence problem in the propagation modelling. However, all these models cannot present the reinfection and self-start processes of modern email malware. The works in [21], [22], [23] focus on threshold conditions for malware fast extinction on the Internet. Their works study the final stable state of epidemic spread based on SIS models, whereas we study the transient propagation dynamics of modern email malware.

Second, there are some works which characterize the propagation dynamics of isomorphic malware, such as P2P malware [33], mobile malware [36], [37] and malware on online social networks [28], [29]. R. Thommes and M. Coates [33] adopt differential equations to present the propagation of P2P malware through a P2P network. The models [36], [37] are proposed for the mobile environment by presuming nodes meet each other with a probability. These works assume all individual devices are homogeneously mixed, and thus, they are unlikely to work in the real mobile environment. The models [28], [29] present the propagation of online social malware by simulations. Since these models [28], [29], [33], [36], [37] are based on nonreinfection, they cannot be adopted to present the propagation of modern e-mail malware.

8 Conclusion

In this paper, we have proposed a novel SII model for the propagation of modern email malware. This model is able to address two critical processes unsolved in previous models: the reinfection and the self-start. By introducing a group of difference equations and virtual nodes, we presented the repetitious spreading processes caused by the reinfection and the self-start. The experiments showed that the result of our SII model is close to the simulations. For the future work, there are also some problems needed to be solved, such as the independent assumption between users in the network and the periodic assumption of email checking time of users.

REFERENCES

[1] M. Fossi and J. Blackbird, "Symantec Internet Security Threat Report 2010," technical report Symantec Corporation, Mar. 2011.

[2] P. Wood and G. Egan, "Symantec Internet Security Threat Report 2011," technical report, Symantec Corporation, Apr. 2012.

[3] C.C. Zou, D. Towsley, and W. Gong, "Modeling and Simulation Study of the Propagation and Defense of Internet E-Mail Worms," *IEEE Trans. Dependable and Secure Computing*, vol. 4, no. 2, pp. 105-118, Apr.-June 2007.

[4] Z. Chen and C. Ji, "Spatial-Temporal Modeling of Malware Propagation in Networks," *IEEE Trans. Neural Networks*, vol. 16, no. 5, pp. 1291-1303, Sept. 2005.

[5] C. Gao, J. Liu, and N. Zhong, "Network Immunization and Virus Propagation in Email Networks: Experimental Evaluation and Analysis," *Knowledge and Information Systems*, vol. 27, pp. 253-279, 2011.

[6] S. Wen, W. Zhou, Y. Wang, W. Zhou, and Y. Xiang, "Locating Defense Positions for Thwarting the Propagation of Topological Worms," *IEEE Comm. Letters*, vol. 16, no. 4, pp. 560-563, Apr. 2012.

[7] J. Xiong, "Act: Attachment Chain Tracing Scheme for Email Virus Detection and Control," *Proc. ACM Workshop Rapid Malcode (WORM '04)*, pp. 11-22, 2004.

[8] S. Wen, W. Zhou, J. Zhang, Y. Xiang, W. Zhou, and W. Jia, "Modeling Propagation Dynamics of Social Network Worms," *IEEE Trans. Parallel and Distributed Systems*, vol. 24, no. 8, pp. 1633-1643, Aug. 2013.

[9] (1999) Cert, advisory ca-1999-04, Melissa Macro Virus, http://www.cert.org/advisories/CA-1999-04.html, 2009.

[10] Cert, Advisory ca-2000-04, Love Letter Worm, http://www.cert.org/advisories/CA-2000-04.html, 2000.

[11] M. Calzarossa and E. Gelenbe, *Performance Tools and Applications to Networked Systems: Revised Tutorial Lectures*. Springer-Verlag, 2004.

[12] G. Serazzi and S. Zanero, "Computer Virus Propagation Models," *Proc. 11th IEEE/ACM Int'l Conf. Modeling, Analysis and Simulations of Computer and Telecomm. Systems (MASCOTS '03)*, pp. 1-10, Oct. 2003.

[13] B. Rozenberg, E. Gudes, and Y. Elovici, "SISR: A New Model for Epidemic Spreading of Electronic Threats," *Proc. 12th Int'l Conf. Information Security*, pp. 242-249, 2009.

[14] (2001) Cert, Advisory ca-2001-22, w32/sircam Malicious Code, http://www.cert.org/advisories/CA-2001-22.html, 2001.

[15] Cert, Incident Note in-2003-03, w32/sobig.f Worm, http://www.cert.org/incidentnotes/IN-2003-03.html, 2003.

[16] C. Wong, S. Bielski, J.M. McCune, and C. Wang, "A Study of Mass-Mailing Worms," *Proc. ACM Workshop Rapid Malcode (WORM '04)*, pp. 1-10, 2004.

[17] D. Moore and C. Shannon, "The Nyxem Email Virus: Analysis and Inferences," technical report, CAIDA, Feb. 2006.

[18] Symantec, A-Z Listing of Threats and Risks, http://www.symantec.com/security Response, 2012.

[19] C. Zou, Internet Email Worm Propagation Simulator, http://www.cs.ucf.edu/czou/research/emailWormSimulationhtml, 2005.

[20] M. Boguna, R. Pastor-Satorras, and A. Vespignani, "Epidemic Spreading in Complex Networks with Degree Correlations," *Lecture Notes in Physics*, vol. 625, pp. 1-23, 2003.

[21] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint," *Proc. 22nd Int'l Symp. Reliable Distributed Systems (SRDS)*, pp. 25-34, 2003.

[22] A.J. Ganesh, L. Massouli, and D.F. Towsley, "The Effect of Network Topology on the Spread of Epidemics," *Proc. IEEE INFOCOM '05*, pp. 1455-1466, 2005.

[23] D. Chakrabarti, J. Leskovec, C. Faloutsos, S. Madden, C. Guestrin, and M. Faloutsos, "Information Survival Threshold in Sensor and p2p Networks," *Proc. IEEE INFOCOM '07*, pp. 1316-1324, 2007.

[24] Y. Wang, S. Wen, S. Cesare, W. Zhou, and Y. Xiang, "Eliminating Errors in Worm Propagation Models," *IEEE Comm. Letters*, vol. 15, no. 9, pp. 1022-1024, Sept. 2011.

[25] H. Ebel, L.I. Mielsch, and S. Bornholdt, "Scale-Free Topology of E-Mail Networks," *Physical Rev. E*, vol. 66, no. 3, Sept. 2002.

[26] M.E. . Newman, S. Forrest, and J. Balthrop, "Email Networks and the Spread of Computer Viruses," *Physical Rev. E*, vol. 66, no. 3, 2002.

[27] T. Bu and D.F. Towsley, "On Distinguishing between Internet Power Law Topology Generators," *Proc. IEEE INFOCOM '02*, pp. 638-647, 2002.

[28] G. Yan, G. Chen, S. Eidenbenz, and N. Li, "Malware Propagation in Online Social Networks: Nature, Dynamics, and Defense Implications," *Proc. Sixth ACM Symp. Information, Computer and Comm. Security (ASIACCS '11),*, pp. 196-206, 2011.

[29] W. Fan and K.H. Yeung, "Online Social Networks-Paradise of Computer Viruses," *Physica A: Statistical Mechanics and Its Applications*, vol. 390, no. 2, pp. 189-197, 2011.

[30] S. Wen, "Topology Generator and Propagation Simulator of Modern Email Malware," Experement Result, http://www.deakin.edu.au/wsheng/emailpropagation.html, 2012.

[31] G. Eschelbeck, "The Laws of Vulnerabilities," *Proc. BlackHat Conf.*, 2004.

[32] R. Pastor-Satorras and A. Vespignani, "Epidemic Spreading in Scale-Free Networks," *Physical Rev. Letters*, vol. 86, pp. 3200-3203, 2001.

[33] R. Thommes and M. Coates, "Epidemiological Modelling of Peer-to-Peer Viruses and Pollution," *Proc. IEEE INFOCOM '06*, pp. 1-12, 2006.

[34] Y. Moreno, J.B. Gomez, and A.F. Pacheco, "Epidemic Incidence in Correlated Complex Networks," *Physical Rev. E*, vol. 68, Sept. 2003.

[35] D.H. Johnson and S. Sinanovic, "Symmetrizing the Kullbackleibler Distance," technical report, Rice Univ., Houston, TX, 2001.

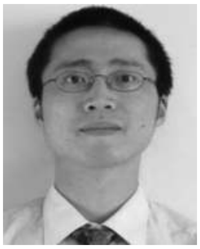
[36] G. Yan and S. Eidenbenz, "Modeling Propagation Dynamics of Bluetooth Worms (Extended Version)," *IEEE Trans. Mobile Computing*, vol. 8, no. 3, pp. 353-368, Mar. 2009.

[37] S.M. Cheng, W.C. Ao, P.Y. Chen, and K.C. Chen, "On Modeling Malware Propagation in Generalized Social Networks," *IEEE Comm. Letters*, vol. 15, no. 1, pp. 25-27, Jan. 2011.

Sheng Wen graduated with the degree in CS from Central South University of China in 2012. He is currently working toward the PhD degree at the School of Information Technology, Deakin University, Melbourne, Australia, under the supervision of Prof. Wanlei Zhou and Yang Xiang. His focus is on modeling of virus spread, information dissemination, and defence strategies of the Internet threats. He is a student member of the IEEE.

Wei Zhou received the BEng and MEng degrees from Central South University, Changsha, China, in 2005 and 2008, respectively, all in CS. She is currently working toward the PhD degree in School of Information Science and Engineering, Central South University. Her research interests include distributed systems, computer networks, and network security.

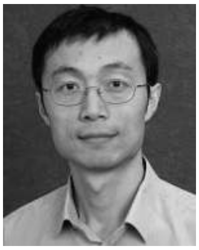
Jun Zhang received the PhD degree from the University of Wollongong, Australian, in 2011. He is currently a lecturer with the School of Information Technology, Deakin University. His research interests include network and system security, pattern recognition, and multimedia retrieval. He has published more than 30 research papers in the reputed journals and conferences. He is a member of the IEEE.



Yang Xiang received the PhD degree in CS from Deakin University, Australia. He is currently a full professor with the School of Information Technology, Deakin University. His research interests include network and system security, distributed systems, and networking. He has published more than 100 research papers in many International journals and conferences. He serves as the associate editors of *Journal of Network and Computer Applications (JNCA)*, *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, and *IEEE Transactions on Computers (TC)*. He is a senior member of the IEEE.

Wanlei Zhou received the PhD degree from the Australian National University, Canberra, in 1991. He is currently the chair professor of information technology and the head of School of Information Technology, Faculty of Science and Technology, Deakin University, Melbourne, Australia. His research interests include distributed and parallel systems and network security. He has published more than 200 papers in refereed International journals and refereed International conferences proceedings. He serves as the associate editor of *IEEE Transactions on Information Forensics and Security (TIFS)*. He is a senior member of the IEEE.

Weijia Jia received the PhD degrees from the Polytechnic Faculty of Mons, Mons, Belgium, in 1993. He is currently a full professor with the Department of CS, City University of Hong Kong (CityU), Hong Kong. His research interests include next-generation wireless communication, QoS routing protocols and multicast. He has published more than 200 papers in refereed International journals and refereed International conferences proceedings. He serves as an associate editor of the *IEEE Transactions on Parallel and Distributed Systems (TPDS)*. He is a senior member of the IEEE.



Cliff C. Zou received the PhD degree from the University of Massachusetts, Amherst, in 2005. He is an assistant professor in the School of Electrical Engineering and C.S., University of Central Florida, Orlando. His research interests include computer and network security, network modeling, and performance evaluation. He is a senior member of the IEEE.