Modeling and Simulation Study of the Propagation and Defense of Internet E-mail Worms

Cliff C. Zou, *Member*, *IEEE*, Don Towsley, *Fellow*, *IEEE*, and Weibo Gong, *Fellow*, *IEEE*

Abstract—As many people rely on e-mail communications for business and everyday life, Internet e-mail worms constitute one of the major security threats for our society. Unlike scanning worms such as Code Red or Slammer, e-mail worms spread over a logical network defined by e-mail address relationships, making traditional epidemic models invalid for modeling the propagation of e-mail worms. In addition, we show that the topological epidemic models presented in [1], [2], [3], and [4] largely overestimate epidemic spreading speed in topological networks due to their implicit *homogeneous mixing* assumption. For this reason, we rely on simulations to study e-mail worm propagation in this paper. We present an e-mail worm simulation model that accounts for the behaviors of e-mail users, including e-mail checking time and the probability of opening an e-mail attachment. Our observations of e-mail lists suggest that an Internet e-mail network follows a heavy-tailed distribution in terms of node degrees, and we model it as a power-law network. To study the topological impact, we compare e-mail worm propagation on power-law topology with worm propagation on two other topologies: small-world topology and random-graph topology. The impact of the power-law topology on the spread of e-mail worms is mixed: E-mail worms spread more quickly on a power-law topology than on a small-world topology or a random-graph topology, but immunization defense is more effective on a power-law topology.

Index Terms—Network security, e-mail worm, worm modeling, epidemic model, simulation.

1 Introduction

Computer viruses and worms have been studied for a long time by both research and application communities. Cohen's work [5] formed the theoretical basis for this field. In the early 1980s, viruses spread mainly through the exchange of floppy disks. At that time, only a small number of computer viruses existed, and virus infection was usually restricted to a local area. As computer networks and the Internet became more popular, from the late 1980s, viruses and worms quickly evolved the ability to spread through the Internet by various means such as file downloading, e-mail, exploiting security holes in software, and so forth.

Currently, e-mail worms constitute one of the major Internet security problems. For example, Melissa in 1999, Love Letter in 2000, and W32/Sircam in 2001 spread widely throughout the Internet and caused tremendous damage [6]. There is, however, no formal definition of an *e-mail worm* in the research area—a computer program can be called an e-mail worm as long as it can replicate and propagate by sending copies of itself through e-mail messages.

- *C.C. Zou is with the School of Electrical Engineering and Computer Science, University of Central Florida, Eng3-335, Orlando, FL 32816-2362. E-mail: czou@cs.ucf.edu.*
- *D. Towsley is with the Department of Computer Science, University of Massachusetts, Amherst, Amherst, MA 01003-4610. E-mail: towsley@cs.umass.edu.*
- *W. Gong is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, Amherst, MA 01003-4610. E-mail: gong@cs.umass.edu.*

Manuscript received 13 Oct. 2005; revised 6 Nov. 2006; accepted 27 Nov. 2006; published online 5 Feb. 2007.
For information on obtaining reprints of this article, please send e-mail to: tdsc@computer.org, and reference IEEECS Log Number TDSC-0140-1005.
Digital Object Identifier no. 10.1109/TDSC.2007.1001.

Although spreading malicious codes through e-mail is an old technique, it is still effective and is widely used by current attackers. Sending malicious codes through e-mail has some advantages that are attractive to attackers:

- Sending malicious codes through e-mail does not require any security holes in computer operating systems or software, making it easy for attackers to program and release their malicious codes.
- Almost everyone who uses computers uses an e-mail service.
- A large number of users have little knowledge of e-mail worms and trust most e-mail they receive, especially e-mail from their friends [7].

In order to understand how worms propagate through e-mail, we focus exclusively on those that propagate solely through e-mail, such as Melissa (if we overlook its slow spreading through file exchange). An e-mail worm, as discussed in this paper, is defined as a piece of malicious code that spreads through e-mail by including a copy of itself in the e-mail attachment—an e-mail user will be infected if he or she opens the worm e-mail attachment. If the e-mail user opens the attachment, the worm program will infect the user's computer and send itself as an attachment to all e-mail addresses that can be found in the user's computer. There are a few e-mail worms that attack e-mail agents' vulnerabilities and, thus, they can infect computers by simply being read by users (with no attachments). These e-mail worms can be considered special ones that vulnerable e-mail users have 100 percent probability of being infected with, whereas nonvulnerable e-mail users have no probability of being infected.

The contributions of this research work are summarized as follows:

- We show in Section 5 that the topological epidemic models for modeling epidemic spreading in topological networks presented in [1], [2], [3], and [4] largely overestimate epidemic spreading speed due to their implicit homogeneous mixing assumption. These mean-field differential equation models have been used and referred to by many papers since 2001 without questioning their accuracy.
- We present an e-mail worm simulation model that accounts for the behaviors of e-mail users, including e-mail checking frequency and the probability of opening an e-mail attachment.
- Our observation shows that the size of e-mail groups follows a heavy-tailed distribution. Since e-mail groups greatly affect the e-mail network topology, we believe that the Internet e-mail network is also heavy-tailed distributed, and we model it as a power-law network.
- We carry out extensive simulation studies of e-mail worm propagation. From these experiments, we derive a better understanding of the dynamics of an e-mail worm spreading—how the degrees of initially infected nodes affect worm propagation, how topological properties such as the power-law exponent affect worm behavior, how the distributions of e-mail checking time affect worm infection, and so forth.
- We gain insight into the differences among power-law, small-world, and random-graph topologies by comparing e-mail worm propagation patterns. The impact of power-law topology on the spread of e-mail worms is mixed: E-mail worms spread more quickly on a power-law topology than on a small-world topology or a random-graph topology, but immunization defense is more effective on a power-law topology.
- We derive by simulations the selective percolation curves and thresholds for power-law, small-world, and random-graph topologies, respectively. The selective percolation curves can explain why selective immunization defense against epidemic spreading is quite effective for a power-law topology but not so good for the other two topologies.

The rest of the paper is organized as follows: Section 2 introduces related work. An e-mail worm simulation model is presented in Section 3. In Section 4, we discuss e-mail network topology and model it as a power-law topology. In Section 5, we show why previous differential equation models are not accurate for e-mail worm modeling, which is the primary reason why we rely on simulations to study e-mail worms in this paper. We present extensive simulation studies of e-mail worm propagation without considering immunization in Section 6. In Section 7, we study immunization defense against e-mail worms and the corresponding percolation problem. Finally, Section 8 concludes this paper with some discussions.

2 Related Work

Kephart et al. published a series of papers from 1991 to 1993 on viral infection based on epidemiology models [8], [9], [10]: Papers by Kephart et al. [8] and [9] were based on a birth-death model in which viruses were spread via activities primarily confined to local interactions. The authors further improved their model by adding a "kill signal" process, and they also considered the special model of viral spread in organizations [10]. To model local interaction and topological impact on virus spreading, they only considered the simplest random-graph topology in their modeling, making their models unsuitable for the e-mail worm modeling studied here. After the famous Code Red incident in July 2001 [6], many researchers studied how to model Internet-scale worm propagation such as [11], [12], [13], [14], [15], [16], and [17], followed by the first model work by Staniford et al. [18]. However, they focus mainly on modeling variants of *random scanning* worms. As explained in Section 5, models presented for scanning worms are not suitable for modeling the propagation of e-mail worms due to the topological e-mail network.

To derive the epidemic threshold of Susceptible-Infectious-Susceptible (SIS) models on topological networks, Wang et al. [19] first presented general formulas based on the eigenvalues of the adjacency matrix of a topological graph. Later, Ganesh et al. [20] formalized this approach and further derived the lifetime approximation of an epidemic on topological networks. In this paper, we are interested in modeling the propagation dynamics of *one* worm incident where infected hosts are not likely to become susceptible again, so the SIS models are not appropriate. In addition, [20] and [19] only studied the final stable state of epidemic propagation, whereas we study the propagation transient dynamics as an e-mail worm spreads out.

To model the epidemic spreading on topological networks, Pastor-Satorras and Vespignani [4] presented a differential equation for an SIS model by differentiating the infection dynamics of nodes with different degrees. However, the authors only studied the epidemic threshold in the stable state. Later, Moreno et al. [2], [3] and Boguna et al. [1] provided the Susceptible-Infectious-Recovered (SIR) differential equation models to study the dynamics of epidemic spreading on topological networks. We show in Section 5 that such differential equation models greatly overestimate the epidemic spreading speed due to their implicit homogeneous mixing assumption.

In 2000, Wang et al. [21] studied a simple virus propagation model based on a clustered topology and a treelike hierarchic topology. In their model, copies of the virus would activate at a constant rate without accounting for any user interactions. The lack of a user model, coupled with the simplified topologies, makes it unsuitable for modeling the propagation of e-mail worms over the Internet. Wong et al. [22] provided the analysis of two e-mail worms, SoBig and MyDoom, based on the monitored trace from a campus network. Newman et al. [23] showed that the e-mail network distribution on a campus follows exponential or stretched exponential distributions. However, such a conclusion was derived based on the number of e-mail addresses in the address books of the campus users. It did not consider the significant impact of e-mail lists, nor the fact that most e-mail worms target all e-mail addresses found on compromised computers, not just users' e-mail address books.

Some researchers have studied immunization defense against virus and worm propagation. Immunization means that a fraction of nodes in a network are immunized; hence,

they cannot be infected. Wang et al. [21] showed that selective immunization can significantly slow down virus propagation for treelike hierarchic topology. From an e-mail worm's point of view, the connectivity of a partly immunized e-mail network is a *percolation* problem. Moore and Newman [24], [25] derived the analytical solution of the percolation threshold of small-world topology and later for arbitrary topologies: If nodes are removed *uniformly* from a network and the fraction of these nodes is higher than the percolation threshold, the network will be broken into pieces. In this paper, since we study *selective immunization* by removing the most connected nodes, the formulas presented in [24] and [25] are not suitable. Albert et al. [26] were the first to explain the vulnerability of power-law networks under attack: By selectively attacking the most connected nodes, a power-law network tends to be broken into many isolated fragments. They concluded that the power-law topology was vulnerable under a deliberate attack. This conclusion is consistent with our results, which were derived from a selective immunization defense study, as described in Section 7.2.

3 E-mail Worm Propagation Simulation Model

E-mail worm, as considered in this paper, is defined as a piece of malicious code that propagates through sending a worm e-mail to all e-mail addresses it can find on compromised computers. Some previous e-mail worms such as Nimda [6] propagated through several other ways besides e-mail spreading such as through open network shares or random scanning. In this paper, we only model their propagation via the e-mail spreading mechanism.

Because an e-mail worm spreads on a logical network defined by e-mail address relationships, it is difficult to mathematically analyze e-mail worm propagation. In Section 5, we show that the differential equation models presented by others cannot accurately model an epidemic spreading in a topological graph. Therefore, in this paper, we will rely on simulation modeling rather than on mathematical analysis in order to focus on realistic scenarios of e-mail worm propagation.

Strictly speaking, an e-mail logical network is a directed graph: Each vertex in the graph represents an e-mail user, whereas a directed edge from node A to node B means that user B's e-mail address is in user A's computer. On the other hand, since user A has user B's address, user A probably has already sent some e-mail to user B before an e-mail worm spreads out. Thus, user B's computer has a great chance of containing the e-mail address of user A as well. For this reason, most edges in the e-mail logical network can be treated as undirected edges. Therefore, in this paper, we model the Internet e-mail network as an undirected graph.

We represent the topology of the logical Internet e-mail network by an undirected graph $G =< V, E >$, $\forall v \in V$, v denotes an e-mail user, and $\forall e = (u, v) \in E$, $u, v \in V$ represents two users u and v who have each others' e-mail addresses in their computers. $|V|$ is the total number of e-mail users. The *degree* of a node is defined as the number of edges connected to the node.

Let us first describe the e-mail worm propagation scenario captured by our model: First, users check their e-mail from time to time. When a user checks his e-mail and encounters a message with a worm attachment, he may discard the message (if he suspects the e-mail or detects the e-mail worm by using an antivirus software) or open the worm attachment if he is unaware of it. When a worm attachment is opened, the e-mail worm immediately infects the user and sends out a worm e-mail to all e-mail addresses found on this user's computer. The infected user will not send out a worm e-mail again unless the user receives another copy of the worm e-mail and opens the attachment again.

From the above description, we see that e-mail worms, unlike scanning worms, depend on the e-mail users' interaction to propagate. There are primarily two human behaviors affecting e-mail worm propagation: one is the *e-mail checking time* of user i, denoted by T_i, $i = 1, 2, \cdots, |V|$, which is the time interval between a user's two consecutive e-mail checking events; the other is the *opening probability* of user i, denoted by C_i, $i = 1, 2, \cdots, |V|$, which is the probability that user i opens a worm attachment. Some e-mail worms exploit the e-mail clients' vulnerabilities such that they can compromise computers without users executing any attachment; these e-mail worms can be modeled by assigning $C_i \equiv 1$ for those vulnerable users.

E-mail checking time T_i of user i ($i = 1, 2, \cdots, |V|$) is a stochastic variable determined by the user's habits. Denote $E[T_i]$ as the expectation of the random variable T_i. The checking time T_i may follow several different distributions. For example, it could be a constant value if a user checks his or her e-mail once every morning or uses e-mail client programs to fetch and check e-mail at a specified time interval. For another example, it could follow exponential distribution (that is, checking action is a Poisson process) if a user checks e-mail at a random time. In Section 6.8, we study how different distributions of e-mail checking time affect the propagation of an e-mail worm.

The opening probability C_i of user i is determined by 1) the user's security awareness and 2) the social engineering tricks deployed by an e-mail worm (for example, MyDoom infected more users than any e-mail worm before due to its advanced social engineering techniques [6]). For the propagation of *one* e-mail worm, we assume C_i to be constant for user i.

We assume that e-mail users have independent behaviors. We model T_i and C_i, $i = 1, 2, \cdots, |V|$, as follows:

- The mean value of user i's e-mail checking time $E[T_i]$ is itself a random variable, denoted by T. When a user checks his or her e-mail, the user checks all the new e-mail received since the last checking time.
- User i opens a worm attachment with probability C_i when the user checks a worm e-mail. Let C denote the random variable that generates C_i, $i = 1, 2, \cdots, |V|$.
- Because the number of e-mail users $|V|$ is very large, and their behaviors are independent, it is reasonable to assume that T and C are independent Gaussian random variables, that is, $T \sim N(\mu_T, \sigma_T^2)$ and $C \sim N(\mu_C, \sigma_C^2)$. Considering that C_i must be between 0 and 1, and $E[T_i]$ must be bigger than zero, we assign C_i and $E[T_i]$ as

TABLE 1
Major Notations Used in This Paper

Notation	Explanation
$G =< V, E >$	Undirected graph representing an email network. $v \in V$ denotes a user, $\|V\|$ is user population.
$E[X]$	The expectation of a random variable X.
k	Vertex degree of a node in a graph; the average degree of a graph is denoted by $E[k]$.
N	Total number of nodes in an email network, $N = \|V\|$.
$P(k)$	Fraction of nodes with degree k in an email network.
T_i	Email checking time of user i — the time interval between user i's two consecutive email checking, $i = 1, 2, \cdots, \|V\|$.
C_i	Opening probability of user i — the probability with which user i opens a worm attachment.
T	Gaussian-distributed random variable that generates $E[T_i]$, $i = 1, 2, \cdots, \|V\|$. $T \sim N(\mu_T, \sigma_T^2)$.
C	Gaussian-distributed random variable that generates C_i, $i = 1, 2, \cdots, \|V\|$. $C \sim N(\mu_C, \sigma_C^2)$.
$I(0)$	Number of initially infected users at the beginning of worm propagation.
$I(t)$	Number of infected users at time t, $\forall t > 0$.
$V(t)$	Number of worm emails in the system at time t, $\forall t > 0$.
α	Power law exponent of a power law topology that has the complementary cumulative degree distribution $F_c(k) \propto k^{-\alpha}$.
$N^h(\infty)$	Number of users that are not infected when an email worm propagation is over.
$D(t)$	Average degree of nodes that are healthy before time t but are infected at time t, $\forall t > 0$.
$C(p)$	Connection ratio — the percentage of remaining nodes that are still connected after removal of the top p percent of most-connected nodes from a network.
$L(p)$	Remaining link ratio — fraction of remained links after removing the top p percent of most-connected nodes.

$$C_i = \begin{cases} \max\{C, 0\} & C \leq 1 \\ 1 & C > 1, \end{cases} \quad (1)$$

$$E[T_i] = \max\{T, 0\}. \quad (2)$$

An e-mail user is called *infected* once the user opens a worm e-mail attachment; upon opening a worm attachment, an infected user immediately sends out a worm e-mail to all its neighbors. Let $I(0)$ denote the number of initially infected users that send out a worm e-mail to all their neighbors at the beginning of a worm propagation. Let random variable $I(t)$ denote the number of infected users at time t during e-mail worm propagation, $I(0) \leq I(t) \leq |V|$, $\forall t > 0$.

It takes time for a recipient to receive a worm e-mail sent out by an infected user. However, the e-mail transmission time is usually much smaller compared to a user's e-mail checking time. Thus, in our model, we ignore the e-mail transmission time. Table 1 is a list of the major notations used in this paper.

4 Heavy-Tailed E-mail Network Topology

The topology of an e-mail network plays a critical role in determining the propagation dynamics of an e-mail worm. Therefore, before we start to study e-mail worm propagation, we need to first determine the e-mail topology.

One very important fact of an e-mail network (in terms of e-mail worm propagation) is that once a computer contains the address of an e-mail list, from an e-mail worm's point of view, this computer has virtually *all* the addresses associated with the e-mail list. Therefore, even though a user's computer may only contain tens of e-mail addresses, the degree of the user in the e-mail network might be as large as several thousands if one of the e-mail addresses is a popular e-mail list. For this reason, we first study the property of e-mail lists.

Let $f(k)$ be the fraction of nodes with degree k in an e-mail network graph G. The complementary cumulative distribution function (ccdf) is denoted by $F_c(k) = \sum_{i=k}^{\infty} f(i)$, that is, the fraction of nodes with degrees greater than or equal to k. We have examined more than 800,000 e-mail groups (lists) in Yahoo! [27], the sizes of which vary from as low as 4 to more than 100,000. Fig. 1 shows the empirical ccdf of the group sizes of Yahoo! in the log-log format. From this figure, we can see that the size of Yahoo! groups is *heavy-tailed distributed*, that is, the ccdf $F_c(k)$ decays slower than exponentially [28].

Because the sizes of e-mail lists, especially the popular e-mail lists, are much larger than the number of e-mail addresses existing in normal computers, we believe that the Internet e-mail network topology is mainly determined by the topology property of e-mail lists. The popular Yahoo! e-mail groups are heavy-tailed distributed, as shown in Fig. 1, which suggests that the Internet-scale e-mail network is probably also heavy-tailed distributed.

In order to generate a heavy-tailed e-mail network, we need to find a suitable topology generator. Currently, except for power-law topology generators, there are no other suitable network generators available to create a heavy-tailed topology. The degree of a power-law topology is heavy-tailed distributed and has the power-law ccdf $F_c(k) \propto k^{-\alpha}$, which is linear on a log-log plot [28]. Therefore, a power-law topology generator is by far the best candidate to generate an e-mail network, although the degree of a real Internet e-mail network may not be strictly power-law distributed. In this paper, we use the Generalized Linear Preference (GLP) power-law generator presented in [28]. We choose the GLP power-law network generator instead

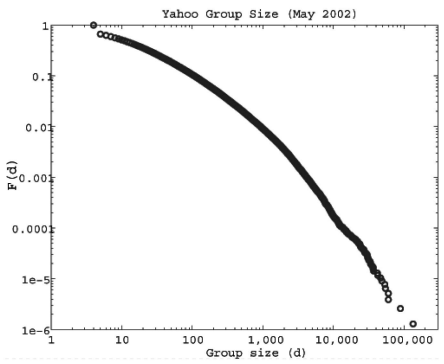


Fig. 1. Complementary cumulative distribution of Yahoo! group size (in May 2002).

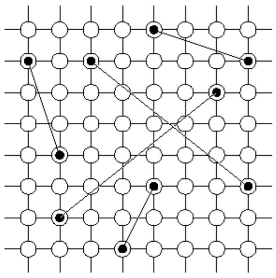


Fig. 2. Illustration of a two-dimensional small-world network.

of other generators because it has an adjustable power-law exponent α.

There are some other popular topologies such as random-graph topology [29] and small-world topology [30] that are not suitable for the e-mail network because they do not provide a heavy-tailed degree distribution. However, in order to understand how a heavy-tailed e-mail topology affects e-mail worm propagation, we also study e-mail worm propagation on both random-graph and small-world topologies.

In this paper, the random-graph network with n vertices and an average degree $E[k] \geq 2$ is constructed as follows: We start with n vertices and add n edges one by one: Edge i, $i = 1, 2, \cdots, n$, connects vertex i to another randomly chosen vertex. Then, we repeatedly connect two randomly chosen vertices with an edge until the total number of edges reaches $E[k] \cdot n/2$. If the generated network happens to be disconnected, we regenerate another network.

We generate the small-world network by using the model presented in [31], which is depicted in Fig. 2. We deploy the following two steps to construct a small-world network that has an average degree $E[k] > 4$. First, we arrange and connect all vertices so that they form the regular two-dimensional grid network, as shown in Fig. 2. Second, we repeatedly connect two randomly chosen vertices with an edge until the total number of edges reaches $E[k] \cdot n/2$.

5 Why Differential Equation Models Are Not Appropriate

Many differential equation models have been presented to model epidemic spreading [1], [3], [4], [18], [17]. In this section, we will explain why we use the simulation-based model presented in Section 3 instead of those differential equation models.

There are two major classes of epidemic models, defined by whether infected hosts can become susceptible again after recovery. If this is true, the models are called *SIS* models because hosts can change their status as *SIS*. If infected, hosts cannot become susceptible again once they are cured, the models are called *SIR* models, and hosts can only have the status transition as SIR (or Susceptible-Infectious (*SI*) models if no infected hosts can recover). For modeling of the propagation of a single e-mail worm incident, after an e-mail user cleans his or her infected computer, the user is not likely to open another copy of the same worm e-mail again. Therefore, we only consider SIR epidemic models in this paper.

SIR models are the natural extensions of SI models by adding the recovery process of infected hosts. Our major focus in this paper is to understand the propagation dynamics of e-mail worms; thus, we do not consider the recovery process and focus solely on SI models.

5.1 Epidemic Model for Homogeneous Networks

The most simple and popular differential equation model is the epidemic model shown below, which has been used by many papers (for example, [11], [18], [17], and [32]) to model random scanning worms such as Code Red and Slammer [6]:

$$\frac{dI(t)}{dt} = \frac{\eta}{\Omega} I(t)[N - I(t)], \tag{3}$$

where N is the total population, and $I(t)$ is the number of infected hosts at time t. η is the worm scan rate, and Ω is the size of Internet Protocol (IP) space scanned by the worm. All hosts are assumed to be either vulnerable or infected.

This model relies on the *homogeneous* assumption that any infected host has an equal opportunity to infect *any* vulnerable host in the system. It means that all hosts in the system can contact each other directly; hence, the system can be treated as a completely connected graph. In other words, there is no topological issue in the modeling. For scanning worms such as Code Red or Slammer [6], because they randomly generate IP addresses to scan and infect, the propagation of these worms satisfies the homogeneous assumption, and they can be accurately modeled by (3).

Some variants of random scanning worms cannot be directly modeled by (3), such as the "hit-list" worm, the "flash" worm [18], the "local preference" worm (such as Blaster worm and Sasser worm), and the "bandwidth-limited" worm. Through extending the simple epidemic model (3), these worms can still be accurately modeled [12], [32], because it is not necessary to consider topological issues in their modeling.

However, an e-mail worm can only spread hop by hop on an e-mail logical network. We must consider topological issues in its modeling. Since the homogeneous assumption will not stand for e-mail worm modeling, we cannot use the above model (3) or its extensions in this paper.

5.2 Epidemic Model for Topological Networks

Because the simple epidemic model (3) is not appropriate for modeling epidemic spreading in topological networks, some researchers [4], [3], [2], [1] have presented new topological models by distinguishing the different dynamics of nodes with different degrees.

Suppose, in a topological network, $P(k)$ is the fraction of nodes that have degree k. The average degree of the network is $E[k] = \sum_k kP(k)$. We denote $i_k(t)$ as the fraction of infected hosts in the k-degree host set. The infection rate is denoted by λ, which is the probability that a susceptible node is infected by one neighboring infected node within a unit time. The differential equation model for nodes with degree k is [3]

$$\begin{aligned} \frac{di_k(t)}{dt} &= \lambda k[1 - i_k(t)]\Theta(t), \\ \Theta(t) &= \frac{\sum_n nP(n)i_n(t)}{\sum_s sP(s)} = \frac{\sum_n nP(n)i_n(t)}{E[k]}. \end{aligned} \tag{4}$$

The factor $\Theta(t)$ is "the probability that any given link points to an infected host" [4]. $\Theta(t)$ is derived based on the conclusion that the probability a link points to an s-degree node is proportional to $sP(s)$ [3], [4].

Boguna et al. [1] improved the model (4) by considering that "since the infected vertex under consideration received the disease through a particular edge that cannot be used for transmission anymore, the correct probability must consider one less edge." They modified the formula of $\Theta(t)$ as

$$\begin{aligned} \frac{di_k(t)}{dt} &= \lambda k[1 - i_k(t)]\Theta(t), \\ \Theta(t) &= \frac{\sum_n (n-1)P(n)i_n(t)}{E[k]}. \end{aligned} \tag{5}$$

When a node has more edges, it has a higher probability of being infected quickly by an epidemic (or an e-mail worm). Since the above two models differentiate nodes with different degrees, they provide a better modeling for topological epidemic spreading than the homogeneous model (3).

Unfortunately, (4) and (5) still have flaws in modeling epidemic spreading in topological networks. The important variable in model (4), $\Theta(t)$, does not distinguish whether infected nodes are connected, clustered together, or scattered around the topological network. In fact, the calculation of $\Theta(t)$ in (4) has the implicit assumption that infected nodes are *uniformly* distributed in the topological network, which is obviously a wrong assumption for topological epidemic spreading where infected nodes must be connected with each other.

Model (5) is better than model (4) since it considers the fact that one link for an infected node should not be considered in its infection power—the node itself is infected by a previously infected node (its parent) on the other end of this link. However, this consideration is only accurate for a newly infected node that connects to no other infected ones except its parent. Thus, the accuracy of this model still relies on the assumption of homogeneous mixing of infected nodes.

5.3 Discussion of the Overestimation in Models (4) and (5)

The consequence of the so-called homogeneous mixing assumption is that models (4) and (5) overestimate the propagation speed of an epidemic in a topological network, especially at the beginning stage when a small number of nodes are infected and clustered with each other. As pointed out by [18], a worm's propagation speed is largely determined by its initial spreading speed. Therefore, the overestimation in models (4) and (5) cannot be ignored and could generate significant modeling errors.

Let us use a simple two-dimensional grid network, shown in Fig. 3, as an example to illustrate the modeling problem. Suppose node A is an initially infected node and it infects three out of four neighboring nodes a moment later (labeled as black nodes). At this time, the epidemic has 10 links, called *effective infection links*, that connect infected nodes with susceptible ones. The three links interconnecting those four infected nodes have no contribution to the epidemic spreading later. On the other hand, if these four infected nodes are scattered in the network as implicitly assumed by model (4), the epidemic would have 16 effective infection links. Therefore, model (4) overestimates the epidemic propagation speed by 60 percent for the scenario shown in Fig. 3.

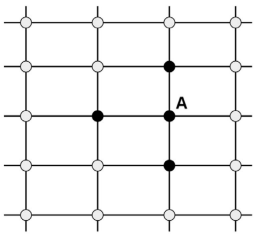


Fig. 3. Illustration of an epidemic spreading in a topological network.

Model (5) is better: The three newly infected nodes have the correct effective infection links expressed by the model since they have not infected others, but model (5) still treats node A as having $(k-1) = 3$ effective infection links. Thus, the number of effective infection links used by the model would be 12 instead of the true value of 10. Therefore, it overestimates the epidemic speed by 20 percent.

How much models (4) and (5) overestimate an epidemic spreading speed is determined by many factors. First, the overestimation will be smaller if the initially infected nodes are scattered over the network instead of clustered together. Second, if the initially infected nodes have larger degrees, their clustering effect will show up more slowly until most of their neighboring nodes have been infected; hence, the overestimation would be smaller.

5.4 Simulation Verification of the Overestimation in Models (4) and (5)

To verify our conjecture above, we first generate several large-scale topological networks, then use these network graphs to compute the numerical solutions of models (4) and (5) and compare with the epidemic spreading simulation results on these networks.

We first generate a power-law network, a small-world network, and a random-graph network, as described in Section 4. All three networks have $|V| = 100,000$ nodes with an average degree of $E[k] = 8$. The power-law network has the exponential power-law exponent $\alpha = 1.7$.

We use the discrete-time method to calculate the numerical solutions of model (4) and model (5). From time $t-1$ to t, we can derive

$$i_k(t) = i_k(t-1) + \lambda k[1 - i_k(t-1)]\Theta(t-1). \tag{6}$$

Then, the total number of infected nodes at time t and $I(t)$ would be

$$I(t) = \sum_k i_k(t)P(k)N. \tag{7}$$

Now, we describe how we conduct the epidemic spreading simulation for the scenario described by models (4) and (5). In every discrete-time unit, if a susceptible node is connected with one infected node, it has the probability λ to be infected within the time unit (λ is the infection rate). If a susceptible node is connected with n infected nodes, it has

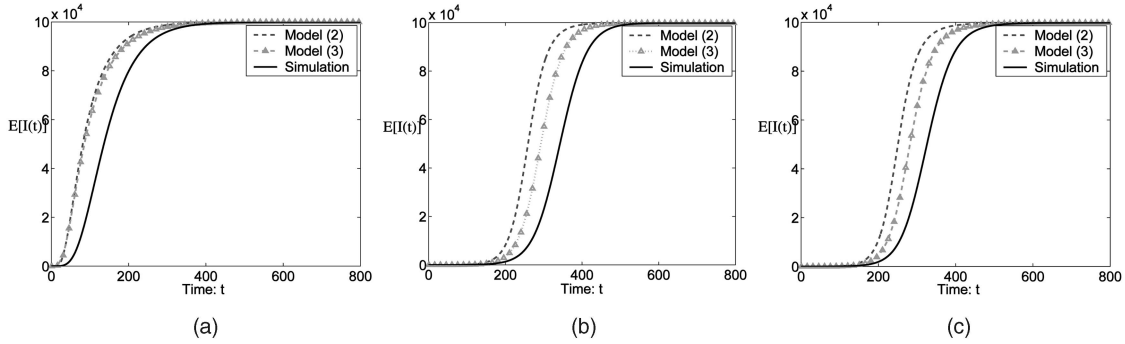


Fig. 4. Numerical solutions of models (4) and (5) compared with the epidemic simulation results on three different topologies. (a) Power law topology. (b) Small world topology. (c) Random graph topology.

the probability $1-(1-\lambda)^n$ to be infected within the time unit. If a node is infected at the discrete time t, it becomes infectious in the next time $t+1$. With the same parameter setting and different random number generator seeds, we run the simulation 1,000 times to derive the average epidemic propagation speed $E[I(t)]$.

In the experiment, $I(0)=2$ and $\lambda=1/200$. The two initially infected nodes in simulations are randomly chosen from all nodes in the network. Fig. 4 shows the comparison among the two differential equation models (4) and (5) and the simulation results on three different topological networks. It clearly shows that models (4) and (5) overestimate the propagation speed of the epidemic on all three topological networks.

We are not arguing that models (4) and (5) are wrong. In fact, they provide a better modeling for epidemic spreading in topological networks than the general epidemic model (3). However, they overestimate the epidemic spreading speed, and the overestimation is not negligible. This is the primary reason why we rely on a simulation model to study the propagation of e-mail worms.

It would be much better if we could provide a new analytical model and then mutually verify it with our simulation model. Unfortunately, we cannot present an accurate analytical model; hence, this paper will rely on a simulation model to study e-mail worm propagation.

6 E-mail Worm Simulation Studies

6.1 Description of the Discrete-Time E-mail Worm Simulator

Discrete-time simulation has been used in many worm-modeling papers [9], [33], [21], [17]. Thus, we simulate e-mail worm propagation in discrete time, too. All events (worm infection, user checking e-mail, and so forth) are assumed to happen right at each discrete time tick. Before the start of an e-mail worm simulation, user (node) i is assigned with a clicking probability C_i and average checking time $E[T_i]$, $i=1,2,\cdots,|V|$ according to (1) and (2), respectively. Each of the initially infected nodes in $I(0)$ is randomly chosen from the entire network. These nodes will send out a worm e-mail right at the first time tick, $t=1$.

At each discrete time tick t, the simulator checks all nodes (users) in the network to see if any user checks e-mail at this time tick. If user i checks e-mail at time t, the user checks all the new e-mail received after his or her last e-mail checking. Each new worm e-mail is opened with probability C_i. Once a worm e-mail is opened, user i is infected (if the user has not been infected before) and the worm will send worm e-mail to all neighbors of the user. These worm e-mail could be read by their recipients as soon as the next time tick $t+1$. Then, a new e-mail checking time T_i is assigned to user i in order to determine when he or she will check his or her e-mail again. In the discrete-time simulation, T_i is a positive integer derived by

$$T_i=\max\{\lfloor X\rfloor,1\}, \tag{8}$$

where X is a random variable. The smallest time unit in a discrete-time simulation is one; thus, T_i must be no smaller than one. In all simulation experiments, X is exponentially distributed with the mean value $E[T_i]$ derived from (2), if not otherwise defined. In Section 6.8, we specifically study how different distributions of X affect e-mail worm propagation. The simulation ends when all users are infected or when a specified simulation end time has been reached.

We are interested in $E[I(t)]$—the average number of infected users in the e-mail network at any time t. We derive $E[I(t)]$ by averaging the results of $I(t)$ from many simulation runs that have the same inputs but different random number generator seeds. For most experiments presented in the following, we perform 100 simulation runs to derive the average value $E[I(t)]$.

The underlying power-law network has $|V|=100,000$ nodes, an average degree of 8, and a power-law exponent of $\alpha=1.7$. Other simulation parameters are $T\sim N(40,20^2)$, $C\sim N(0.5,0.3^2)$, and $I(0)=2$. If not otherwise specified, initially infected nodes are randomly chosen from the entire network in each simulation run, and all simulation experiments run under the same power-law e-mail network, with the same parameters specified above.

In a discrete-time simulation, each discrete time tick can represent an arbitrary time interval in the real world, such as 1 minute, 10 minutes, or even 1 hour. Thus, the absolute time tick value used in a discrete-time simulation does not matter much, such as the mean value $E[T]=40$ used in our simulations. On the other hand, since all simulated events are assumed to happen right at discrete time ticks, a discrete-time simulation would be more accurate if a discrete time tick represents a shorter time interval. From

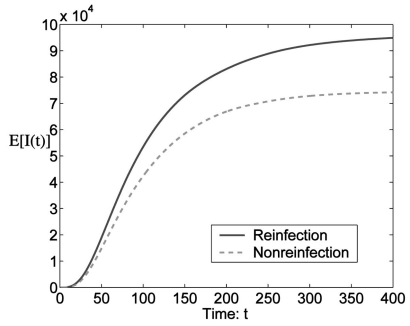


Fig. 5. Reinfection versus nonreinfection.

our experiments, we find that it is accurate enough to choose $E[T] = 40$. The value of C depends on how deceptive an e-mail worm is, which varies from one worm incident to another. Thus, we choose $E[C] = 0.5$ to simulate the general case of e-mail worm propagation.

For the convenience of readers, we have put the source codes of our e-mail worm propagation simulator and topology generators online [34].

6.2 Reinfection versus Nonreinfection

First, we consider two cases under different infection assumptions: reinfection versus nonreinfection. *Reinfection* means that a user will send out worm e-mail copies whenever he opens an e-mail worm attachment. Thus, a recipient can repeatedly receive worm e-mail from the same infected user. *Nonreinfection* means that each infected user sends out worm copies only once, after which the user will not send out any worm e-mail, even if he or she opens a worm attachment again. Some e-mail worms belong to the nonreinfection type such as Melissa and Love Letter; others are the reinfection type such as W32/Sircam.

Fig. 5 illustrates the behavior of $E[I(t)]$ as a function of time t for both reinfection and nonreinfection cases on a power-law e-mail network.

6.3 Variability in Worm Propagation

An e-mail worm propagation is, in fact, a stochastic process. Under the same network condition, the same e-mail worm could spread faster or slower in different runs. To study how variable an e-mail worm propagation could be, we simulate the e-mail worm propagation for 100 runs (the reinfection scenario) under the same simulation settings but with different seeds in the random number generator.

An intuitive measurement of the worm propagation variability is demonstrated by the 95th and fifth percentile curves of $I(t)$, first presented in [33]. Fig. 6 shows these two curves compared with the curve of $E[I(t)]$. Among those 100 simulation runs, in five runs, the worm propagates faster than the 95th percentile curve, whereas, in another five runs, the worm propagates slower than the fifth percentile curve. This figure shows that an e-mail worm spreads with the similar dynamics after around 5 percent of vulnerable hosts have been infected, but the initial propagation dynamics could be dramatically different. Therefore, the initial phase of worm spreading largely determines the overall worm propagation speed.

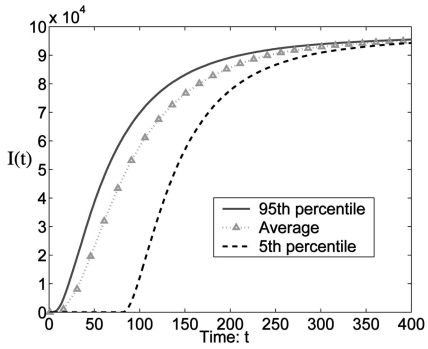


Fig. 6. The fifth and 95th percentiles of 100 simulation runs.

Another way to measure the variability of worm propagation is to use the statistics term "confidence interval" [35]. For every discrete time t $(t = 1, 2, 3, \cdots)$, $E[I(t)]$ derived from simulation is the mean value of 100 samples $I(t)$ from these 100 simulation runs. Suppose the estimated standard deviation of these 100 $I(t)$ samples is σ, then the 95 percent confidence interval of $E[I(t)]$ is [35]

$$\left(E[I(t) - t\frac{\sigma}{\sqrt{100}}, E[I(t) + t\frac{\sigma}{\sqrt{100}}\right), \tag{9}$$

where $t = 1.984$ is the value of t-distribution with 99 degrees of freedom for 95 percent confidence interval. Fig. 7 shows $E[I(t)]$ of the worm propagation in 100 simulation runs, together with its upper and lower bounds in terms of 95 percent confidence interval.

6.4 Impact of User Clicking Probability

In our e-mail worm model, each user i opens an e-mail attachment with probability C_i when reading a worm e-mail. Thus, user i has the probability $1 - (1 - C_i)^m$ to be infected when receiving m worm e-mail—the chance of being infected increases as a function of the amount of worm e-mail received. For this reason, more users are infected in the reinfection case than in the nonreinfection case, as shown in Fig. 5.

Because some users have a very low probability of opening e-mail attachments, in both cases shown in Fig. 5, a certain number of users will not be infected when the worm propagation is over. Let $N^h(\infty)$ denote the number of users that are not infected when the worm propagation is over. In

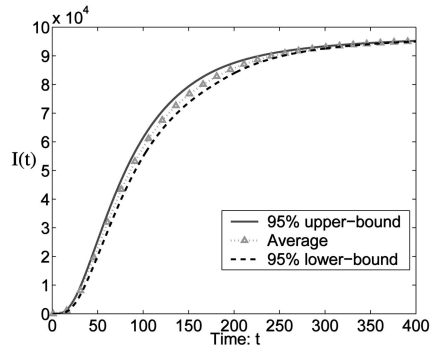


Fig. 7. Ninety-five percent statistical confidence interval of 100 simulation runs.

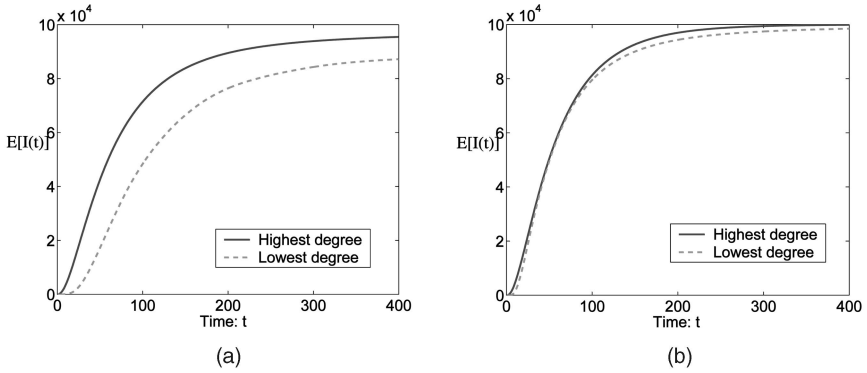


Fig. 8. Effect of different degrees of initially infected nodes. (a) Average degree $E[k] = 8$. (b) Average degree $E[k] = 20$.

the nonreinfection case, user i, who has m_i edges (neighbors), will receive at most m_i copies of the worm e-mail—the probability that user i is not infected is at least $(1 - C_i)^{m_i}$. For the nonreinfection case, we can derive a lower bound for $E[N^h(\infty)]$ if we know the network degree distribution $P(k)$ and assume that C_i is the same for all users, that is, $C_i = p, \forall i \in \{1, 2, \cdots, |V|\}$.

Let $G(x)$ denote the probability-generating function of the degrees of the e-mail network:

$$G(x) = \sum_{k}^{\infty} P(k)x^k. \tag{10}$$

Then, we can derive the lower bound for $E[N^h(\infty)]$:

$$E[N^h(\infty)] \geq |V| \sum_{k}^{\infty} P(k)(1-p)^k = |V|G(1-p), \tag{11}$$

where $|V|$ is the user population. This formula shows that, as e-mail users become cautious in clicking worm e-mail attachments, a larger number of e-mail users will stay healthy without being infected by the e-mail worm.

The e-mail worm has successfully spread in all 100 simulation runs. In fact, the e-mail worm has a small chance of dying before it spreads. For example, in the beginning, those users initially infected send out worm copies to their neighbors. If all their neighbors decide not to open the worm e-mail attachment for the first round, then no worm e-mail exists in the network after those neighbors finish checking their e-mail for the first time. If we assume that all users open worm attachments with the same probability p, and the number of worm copies sent out by those initially infected users is m, then the e-mail worm has the probability $(1 - p)^m$ to die before it infects any other users.

A reinfection e-mail worm propagates faster and is the focus of our study. In the following, we only consider reinfection e-mail worms, if not otherwise stated.

6.5 Initially Infected Nodes with the Highest Degree versus the Lowest Degree

In our previous experiment, the degree of the power-law network varies from 3 to 1,833. Because a worm propagation speed is largely determined by its initial infection speed (as shown in Fig. 6), it appears that the degrees of initially infected nodes are critical to the overall worm propagation speed. We consider two cases: In the first case, the initially infected nodes have the highest degree, whereas, in the second case, the initially infected nodes have the lowest degree. Both cases have the same number of initially infected nodes $I(0) = 2$. Fig. 8 shows the behavior of $E[I(t)]$ as a function of time t of these two cases on two power-law networks, respectively. Both power-law networks have the same $|V| = 100,000$ nodes and a power-law exponent of $\alpha = 1.7$ but different connection densities—one has an average degree of 8, whereas another has an average degree of 20.

Fig. 8 shows that the identities of the initially infected nodes are more important in a sparsely connected network than in a densely connected network. From a worm writer's point of view, it is important to let an e-mail worm spread as wide as possible before people become aware of the worm. It will help an e-mail worm propagate faster by choosing the right initial launching points, such that those initially infected computers contain a large number of e-mail addresses.

6.6 Topology Effect: Power Law, Small World, and Random Graph

In Section 4, we discussed why we believe that the e-mail logical network is a heavy-tailed network. In this section, we examine how topology affects e-mail worm propagation.

We run our e-mail worm simulation on a power-law network, a small-world network, and a random-graph network, respectively. All three networks have the same average degree $E[k] = 8$ and $|V| = 100,000$ nodes. Fig. 9 shows $E[I(t)]$ as a function of time t of these three topologies, respectively.

Fig. 9 shows that e-mail worm propagation on a small-world network is a little slower than the one on a random-graph network. This is because a small-world topology has a larger clustering coefficient than a random-graph topology [30]. *Clustering coefficient* measures how clustered together neighboring nodes are. As illustrated in Fig. 3, if a topology has a higher clustering coefficient, its infected nodes tend to have more links interconnecting themselves; thus, such a topology has fewer effective infection links than a topology that has a lower clustering coefficient.

We also observe in Fig. 9 that the worm infection speed on a power-law topology is much faster than on the other two topologies. One reason is that a power-law topology

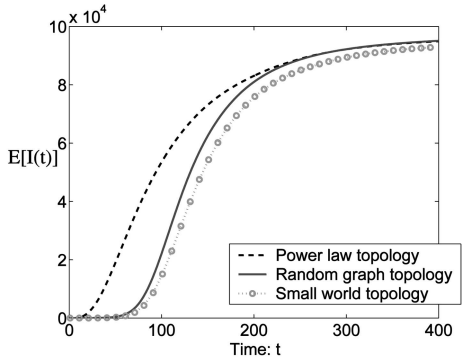


Fig. 9. Effect of topology on e-mail worm propagation.

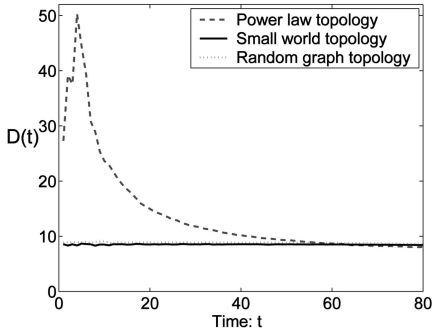


Fig. 10. Average degree of nodes that are being infected at each time tick t.

has the smallest characteristic path length among those three topologies, whereas the other two have similar characteristic path lengths [28], [36]. *Characteristic path length* is defined as the number of edges in the shortest path between two vertices, averaged over all pairs of vertices [30]. An e-mail worm can reach and infect a node more quickly by traveling through a shorter path on a power-law network than on a small-world or random-graph network.

Another reason is that an e-mail worm exhibits more firing power on a power-law network at the early stage of worm propagation. On a power-law network, the degrees of nodes vary significantly [26]. Once an e-mail worm infects a highly connected node, a large number of worm e-mail will be sent out from this infected node.

Let $D(t)$ denote the average degree of nodes that are healthy before time t but are infected at time tick t. $D(t)$ tells us what kind of nodes are being infected at each time t, $t = 1, 2, 3, \cdots$. We repeat the experiment in Fig. 9b and derive $D(t)$ for each topology by averaging the results of 1,000 simulation runs. We plot the $D(t)$ of each network as a function of time t, as shown in Fig. 10. Note that the $D(t)$ of a small-world network and a random-graph network are almost the same.

Fig. 10 clearly shows that, on a power-law network, an e-mail worm tends to first infect some highly connected nodes—these nodes will then send out a much larger number of worm e-mail than other infected nodes. Thus, the infection speed will be *amplified* by them at the beginning. Neither a small-world nor a random-graph network exhibits such amplification effect, since all nodes on them have similar degrees.

6.7 Effect of the Power-Law Exponent α

The power-law exponent α is an important parameter for a power-law topology. It is the slope of the curve of the complementary cumulative degree distribution in a log-log graph [28]—the smaller α is, the more variable the degrees of nodes in the topology become. In our previous simulations, we use $\alpha = 1.7$ to generate the power-law network with $|V| = 100{,}000$ nodes and an average degree of 8. This power-law network has the highest degree of 1,833 and the lowest degree of 3.

The Internet Autonomous System (AS)-level power-law topology has a power-law exponent of $\alpha = 1.1475$ [28]. Using $\alpha = 1.1475$ for a 100,000-node power-law network with an average degree of 8 will produce a network with the highest degree up to 28,000 and the lowest degree of 1. Thus, we think $\alpha = 1.1475$ is too small for modeling the Internet e-mail network.

On the other hand, we do not know the true value of α for the real Internet e-mail network. In order to see how the power-law exponent α affects e-mail worm propagation, we compare worm propagation on two power-law networks: one has $\alpha = 1.7$ and the other one has $\alpha = 1.1475$. Both networks have $|V| = 100{,}000$ nodes and an average degree of 8. We denote the network with $\alpha = 1.7$ as the power-law network Net_a and the network with $\alpha = 1.1475$ as the power-law network Net_b. $E[I(t)]$ is plotted for both networks as functions of time t in Fig. 11. It shows that an e-mail worm initially propagates faster on network Net_b than on Net_a. Later, however, the worm spreads more quickly on Net_a than on Net_b.

Net_b concentrates a large number of links on a small number of nodes. Once some of these nodes have been infected, there will be more copies of the worm e-mail sent out than in Net_a. Those highly connected nodes behave like amplifiers in e-mail worm propagation (see the amplification effect explained in Section 6.6). Thus, initially, an e-mail worm spreads faster on Net_b than on Net_a.

After having infected most highly connected nodes, the e-mail worm enters the second phase as shown in Fig. 10—mainly trying to infect the nodes that have small degrees. Net_b has more nodes with smaller degrees than Net_a—the smallest degree in Net_b is 1, whereas, in Net_a, it is 3. Since a node with fewer links is harder to be infected, during the second phase of worm propagation, the e-mail worm spreads slower on Net_b than on Net_a.

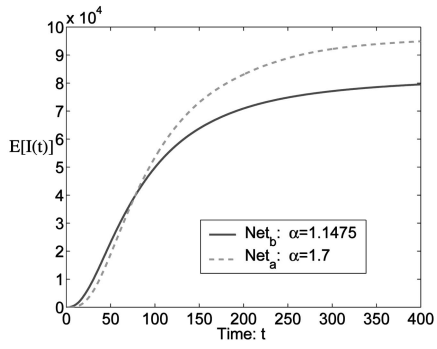


Fig. 11. Effect of power-law exponent α on e-mail worm propagation.

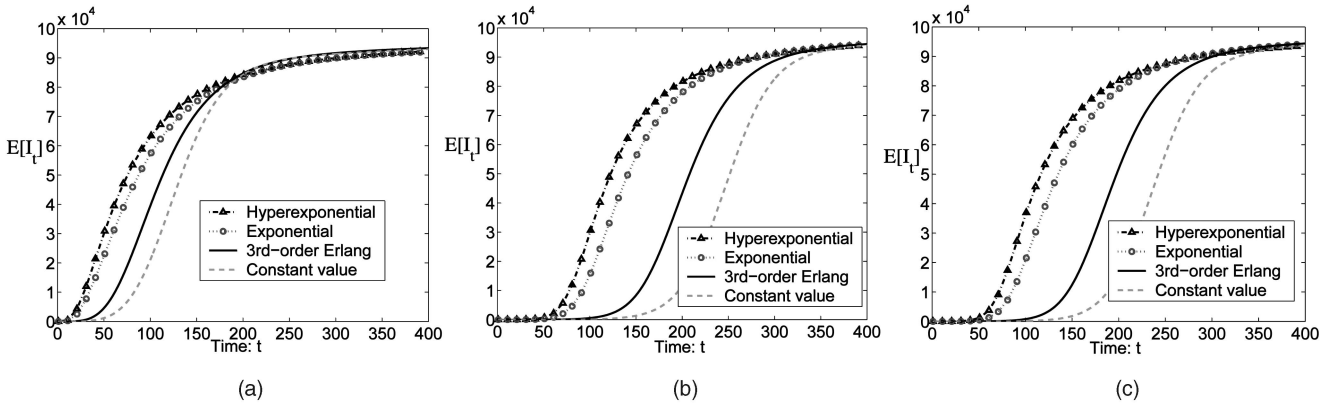


Fig. 12. Effect of the distribution of e-mail checking time T_i. (a) Power law topology. (b) Small-world topology. (c) Random graph topology.

6.8 Effect of E-mail Checking Time Distribution

In our e-mail worm simulation experiments above, we assume that user i's e-mail checking time T_i is exponentially distributed with mean $E[T_i]$, $i = 1, 2, \cdots, |V|$. What if the e-mail checking time T_i is drawn from some other distributions or is simply a fixed value? For example, some e-mail agent software used by e-mail users will automatically retrieve new e-mail from users' mailboxes at a constant time interval. In this section, we study the effect of the distribution of e-mail checking time on the propagation of an e-mail worm.

Fig. 12 shows $E[I(t)]$ under four different distributions of e-mail checking time T_i: hyperexponential distribution [37], exponential distribution, third-order Erlang distribution, and constant value. For comparison reasons, we let each distribution have the same mean value of $1/\lambda$. The probability density function of the hyperexponential checking time is chosen as

$$f_X(x) = f_{Y_1}(y)/4 + 3f_{Y_2}(y)/4, \tag{12}$$

where Y_1 and Y_2 are exponential random variables with rates $\lambda/2$ and $3\lambda/2$, respectively. Based on the formulas provided in [37], it is not hard to know that this hyperexponential distribution has the same mean value of $1/\lambda$.

The other simulation parameters are identical: $I(0) = 2$, $C \sim N(0.5, 0.3^2)$, and $T \sim N(40, 20^2)$ (the average e-mail checking time $E[T_i]$ of different users still follows a normal distribution), $i = 1, 2, \cdots, N$.

In statistics, *coefficient of variation* (CV) is a measurement of dispersion of a probability distribution [37]. It is defined as the ratio of the standard deviation σ to the mean μ of a random variable, that is, $CV = \sigma/\mu$. An exponential distribution has $CV = 1$, the hyperexponential distribution (12) has $CV = \sqrt{5/3}$, the third-order Erlang distribution has $CV = 1/\sqrt{3}$, and a constant value has $CV = 0$. Fig. 12 shows that an e-mail worm propagates faster as the e-mail checking time interval T_i becomes more dispersed.

We have proven this conclusion for a simplified worm propagation model. The detailed proof can be found in our technical report [38]. Intuitively, this phenomenon is due to the *snowball* effect: Before the worm copies in the system with less dispersed checking time give birth to the next generation—infecting some new hosts—the worm copies in another system with more dispersed checking time have already given birth to several generations, although each generation's population is relatively small.

7 Immunization and Percolation for E-mail Worm Defense

In this section, we consider immunization defense against e-mail worm attacks. For an e-mail network, immunizing a node means that the node cannot be infected by the e-mail worm under study. In this paper, we consider a *static* immunization defense. By this, we mean that, before an e-mail worm starts to propagate, a small number of nodes in the network have already been immunized. If some e-mail users are well educated and never open suspicious e-mail attachments, they can be treated as immunized nodes in the e-mail network.

7.1 Effect of Selective Immunization

It is not possible for us to immunize all e-mail users in the e-mail network. A realistic approach is to immunize a small subset of nodes. Thus, we need to know how to choose the appropriate size and membership of this subset in order to slow down or constrain the e-mail worm spreading.

Wang et al. [21] explained that selective immunization could significantly slow down virus propagation for tree-like hierarchic topology. We find that, for a power-law e-mail network, selecting those most highly connected nodes to immunize is also quite effective against e-mail worm propagation.

We simulate worm propagation under two different immunization defense methods: In the first case, we randomly choose 5 percent of the nodes to immunize, whereas, in the second case, we choose 5 percent of the most connected nodes to immunize. We plot $E[I(t)]$ as a function of time t for these two immunization methods in Fig. 13 (on a power-law network, a small-world network, and a random-graph network, respectively). In order to see the effect of immunization, we also plot $E[I(t)]$ for the original case where there is no immunization.

We observe in Fig. 13 that selective immunization is quite effective for a power-law topology, whereas it has little effect for a small-world topology or a random-graph topology. On a power-law e-mail network, we can significantly slow down e-mail worm propagation by selecting those most connected nodes to immunize.

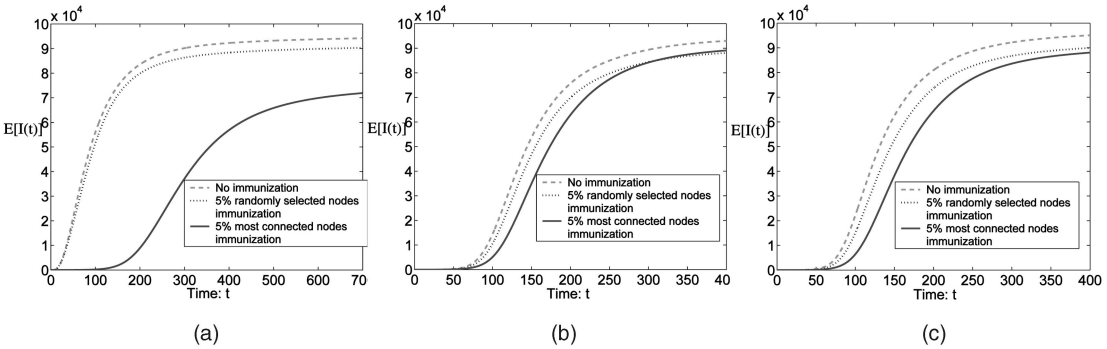


Fig. 13. Effect of selective immunization on e-mail worm propagation. (a) Power law topology. (b) Small-world topology. (c) Random graph topology.

The results here are consistent with the conclusions in [26]. Albert et al. [26] showed that selectively attacking the most connected nodes rapidly increases the diameter of a power-law network. Since an e-mail worm depends on the connectivity of the underlying e-mail network to spread, immunizing the most connected nodes has the effect of rapidly increasing the network diameter. This, in turn, significantly slows down the worm's propagation.

7.2 Selective Percolation and E-mail Worm Prevention

Having observed that selective immunization is quite effective for a power-law e-mail network, then what is the appropriate size of the subset to immunize, and how many nodes do we need to immunize in order to prevent an outbreak of an e-mail worm?

From an e-mail worm's point of view, the connectivity of a partially immunized network is a "percolation" problem. Newman et al. [25] studied the standard percolation by *uniformly* removing a fraction of nodes from networks—their approaches cannot be used here to study the selective immunization defense.

Because we want to study the effect of selective immunization, we introduce a new concept, "selective percolation." For example, a selective percolation value of p means to remove the top p percent of the most connected nodes from a network.

Suppose the e-mail graph $G =< V, E >$ has $|V|$ nodes and $|E|$ edges. For a selective percolation value of p, $0 < p < 1$, let $C(p)$ denote the *connection ratio*, the percentage of how many remaining nodes are still connected after removing the top p percent of the most connected nodes from the network. Let $L(p)$ denote the *remaining link ratio*, the fraction of the remaining links after removing the top p percent of most connected nodes from the network. Then, we have

$$\begin{cases} C(p) = c_p/(|V| - |V|p) \\ L(p) = (|E| - e_p)/|E| \end{cases} 0 < p < 1, \tag{13}$$

where e_p is the number of removed links, and c_p is the size of the largest cluster in the remaining network when we remove the top p percent most connected nodes.

We generate 100 networks for each type of the three topologies: power-law, small-world, and random-graph topologies. Each network has an average degree of 8 and $|V| = 100,000$ nodes. For every selective percolation value p chosen from $p = 0.01, 0.02, 0.03, \cdots, 1$, we calculate $C(p)$ and $L(p)$ by averaging the simulated results derived by (13) from those generated 100 networks for each type of topology, respectively. This way, $C(p)$ and $L(p)$ derived here are properties of the corresponding topology, not of one single generated network.

For each of the three topologies, we plot $C(p)$ and $L(p)$ as functions of the selective percolation value p in Fig. 14.

Fig. 14a shows that a power-law topology has a selective percolation threshold (the threshold here is about 0.29). If the fraction of selectively immunized users exceeds this threshold, the e-mail network will be broken into separated fragments, and no worm outbreak will occur.

The selective percolation threshold of a power-law topology is much smaller than that of either a small-world topology or a random-graph topology. Although a power-law topology is more vulnerable under deliberate attacks [26], it benefits more from a selective immunization defense.

Fig. 14a shows that, when we immunize the top 5 percent of most connected nodes in a power-law network, although 97.5 percent of the remaining nodes are still connected, 55.5 percent of the original network edges have been removed. Thus, an e-mail worm has fewer and longer paths to reach and infect nodes in the remaining network. Figs. 14b and 14c show that this is not the case for a small-world topology or a random-graph topology; a 5 percent selective immunization removes fewer than 20 percent of the edges.

The selective percolation threshold of the random-graph topology (0.68) is slightly smaller than the threshold in the small-world topology. This is understandable since the random-graph topology has a more variant degree distribution than the small-world topology; hence, a selective immunization will remove more edges from the random-graph topology than from the small-world topology.

8 Conclusion

In this paper, we first show that the topological epidemic models presented in [1], [2], [3], and [4] largely overestimate an epidemic spreading speed on a topological network due to their implicit homogeneous mixing assumption. Then, we present an e-mail worm simulation model by considering the e-mail users' behaviors such as e-mail checking frequency and the probability of opening an e-mail attachment. Given that e-mail worms spread over a logical network defined by an e-mail address relationship, our observations of e-mail lists suggest that

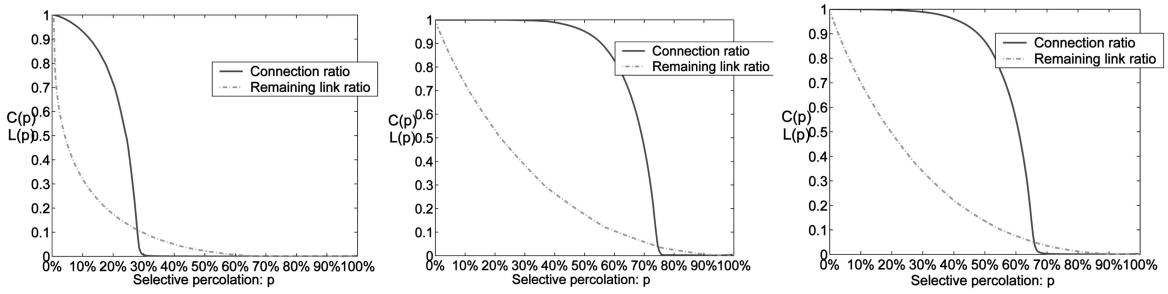


Fig. 14. Selective percolation on three topologies. (a) Power law topology. (b) Small-world topology. (c) Random graph topology.

the degrees in an e-mail network are heavy-tailed distributed. To understand how the heavy-tailed topology affects e-mail worm propagation, we compare e-mail worm spreading on three topologies—power law, small world, and random graph—and then study how the topology affects immunization defense. From these studies, we derive a better understanding of an e-mail worm's behaviors and the differences among power-law, small-world, and random-graph topologies.

Compared to small-world and random-graph topologies, the impact of power-law topology on e-mail worm propagation is mixed: On one hand, an e-mail worm spreads faster on a power-law topology than on a small-world or a random-graph topology; on the other hand, it is more effective to carry out selective immunization on a power-law topology than on the other two topologies. This conclusion shows that we could achieve an effective defense by focusing our precious defense resources and effort on the small number of e-mail users who can send out e-mail to a large number of users.

There is still much work to do on e-mail worm modeling and defense. First, in this paper, we have relied on simulations to study e-mail worm propagation and showed that previous topological epidemic models are not accurate. The next step is to derive a more accurate analytical model by relaxing the homogeneous mixing assumption.

Second, currently, there is still no accurate monitoring work of Internet-scale e-mail worm propagation since e-mail worms do not send out random scanning. Wong et al. [22] only provided limited monitoring results of a campus network. Additionally, e-mail communication traffic is hard to share due to the privacy concern. Therefore, it is hard to validate our simulation model with real e-mail worm incidents. We plan to conduct more research on e-mail worm monitoring and collaborate with others to solve this problem.

Third, we have only considered static immunization defense in this paper—we assume that, before the break out of an e-mail worm, a fraction of users have already been immunized from the worm, and no more users will become immunized during the propagation of an e-mail worm. However, the more realistic scenario is that e-mail users and computers gradually become immunized as an e-mail worm spreads out, which means that we need to further study dynamic immunization defense against e-mail worms.

Fourth, although we have considered the impact of e-mail lists on the topology of Internet e-mail networks, instead of an undirected graph, a directed graph is preferred in order to more accurately capture a one-way e-mail address relationship (that is, user A has the e-mail address of user B, but user B does not have the address of user A). In addition, there are many e-mail lists having constraints on who can submit broadcast messages to a mailing list (for example, only the administrator can)—such e-mail lists need specific modeling. Finally, we need to further consider how to match the e-mail logical network with the physical networks of e-mail servers because a good filter on an e-mail server will protect many e-mail users in the logical e-mail network.

ACKNOWLEDGMENTS

The authors would like to thank Zihui Ge and Daniel R. Figueiredo for providing the size distribution data of Yahoo! groups. This work was supported in part by the US Defense Advanced Research Projects Agency under Contract F30602-00-2-0554 and by US National Science Foundation Grants EIA-0080119 and CNS-0627318. It was also supported in part by ARO Contract DAAD19-01-1-0610 and Contract 2000-DT-CX-K001 from the Office of Justice Programs, US Department of Justice.

REFERENCES

[1] M. Boguna, R. Pastor-Satorras, and A. Vespignani, "Epidemic Spreading in Complex Networks with Degree Correlations," *Lecture Notes in Physics: Statistical Mechanics of Complex Networks,* 2003.

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

[3] Y. Moreno, R.P. Satorras, and A. Vespignani, "Epidemic Outbreaks in Complex Heterogeneous Networks," *European Physical J. B,* vol. 26, 2002.

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

[5] F. Cohen, "Computer Viruses: Theory and Experiments," *Computers and Security,* vol. 6, no. 1, Feb. 1987.

[6] CERT, "CERT/CC Advisories," 2005, http://www.cert.org/advisories/.

[7] CERT, "CERT Advisory CA-2001-20: Continuing Threats to Home Users," http://www.cert.org/advisories/CA-2001-20.html, July 2001.

[8] J. Kephart, D.M. Chess, and S. White, "Computers and Epidemiology," *IEEE Spectrum,* vol. 30, no. 5, May 1993.

[9] J. Kephart and S. White, "Directed-Graph Epidemiological Models of Computer Viruses," *Proc. IEEE Symp. Security and Privacy,* pp. 343-359, 1991.

[10] J. Kephart and S. White, "Measuring and Modeling Computer Virus Prevalence," *Proc. IEEE Symp. Security and Privacy,* 1993.

[11] Z. Chen, L. Gao, and K. Kwiat, "Modeling the Spread of Active Worms," *Proc. IEEE INFOCOM '03,* pp. 1890-1900, Mar. 2003.

[12] G. Kesidis, I. Hamadeh, and S. Jiwasurat, "Coupled Kermack-McKendrick Models for Randomly Scanning and Bandwidth-Saturating Internet Worms," *Proc. Third Int'l Workshop QoS in Multiservice IP Networks (QoS-IP),* pp. 101-109, Feb. 2005.

[13] D. Nicol and M. Liljenstam, "Models of Internet Worm Defense," *Proc. Inst. for Math. and Its Applications (IMA) Workshop 4: Measurement, Modeling and Analysis of the Internet,* http://www.ima.umn.edu/talks/workshops/1-12-16.2004/nicol/talk.pdf, Jan. 2004.

[14] D. Nojiri, J. Rowe, and K. Levitt, "Cooperative Response Strategies for Large Scale Attack Mitigation," *Proc. Third DARPA Information Survivability Conf. and Exhibition,* Apr. 2003.

[15] J. Wu, S. Vangala, L. Gao, and K. Kwiat, "An Efficient Architecture and Algorithm for Detecting Worms with Various Scan Techniques," *Proc. 11th Ann. Network and Distributed System Security Symp. (NDSS '04),* Feb. 2004.

[16] C. Zou, L. Gao, W. Gong, and D. Towsley, "Monitoring and Early Warning for Internet Worms," *Proc. 10th ACM Conf. Computer and Comm. Security (CCS '03),* pp. 190-199, Oct. 2003.

[17] C. Zou, W. Gong, and D. Towsley, "Code Red Worm Propagation Modeling and Analysis," *Proc. Ninth ACM Conf. Computer and Comm. Security (CCS '02),* pp. 138-147, Oct. 2002.

[18] S. Staniford, V. Paxson, and N. Weaver, "How to Own the Internet in Your Spare Time," *Proc. Usenix Security Symp.,* pp. 149-167, Aug. 2002.

[19] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint," *Proc. 22nd Symp. Reliable Distributed Computing,* Oct. 2003.

[20] A. Ganesh, L. Massoulie, and D. Towsley, "The Effect of Network Topology on the Spread of Epidemics," *Proc. IEEE INFOCOM '04,* Mar. 2004.

[21] C. Wang, J.C. Knight, and M.C. Elder, "On Viral Propagation and the Effect of Immunization," *Proc. 16th ACM Ann. Computer Applications Conf.,* Dec. 2000.

[22] C. Wong, S. Bielski, J.M. McCune, and C. Wang, "A Study of Massmailing Worms," *Proc. ACM Conf. Computer and Comm. Security Workshop Rapid Malcode (WORM '04),* Oct. 2004.

[23] M. Newman, S. Forrest, and J. Balthrop, "Email Networks and the Spread of Computer Viruses," *Physical Rev. E,* vol. 66, no. 035101, 2002.

[24] C. Moore and M. Newman, "Exact Solution of Site and Bond Percolation on Small-World Networks," *Physical Rev. E,* vol. 62, 2000.

[25] M. Newman, S. Strogatz, and D. Watts, "Random Graphs with Arbitrary Degree Distributions and Their Applications," *Physical Rev. E,* vol. 64, no. 026118, 2001.

[26] R. Albert, H. Jeong, and A. Barabasi, "Error and Attack Tolerance of Complex Networks," *Nature,* vol. 406, pp. 378-382, 2000.

[27] Yahoo! Groups, http://groups.yahoo.com, 2005.

[28] T. Bu and D. Towsley, "On Distinguishing between Internet Power Law Topology Generators," *Proc. IEEE INFOCOM '02,* June 2002.

[29] P. Erdos, "Graph Theory and Probability," *Canadian J. Math.,* vol. 11, 1959.

[30] D. Watts and S. Strogatz, "Collective Dynamic of Small-World Networks," *Nature,* vol. 393, 1998.

[31] M. Newman, I. Jensen, and R. Ziff, "Percolation and Epidemics in a Two-Dimensional Small World," *Physical Rev. E,* vol. 65, no. 021904, 2002.

[32] C. Zou, D. Towsley, and W. Gong, "On the Performance of Internet Worm Scanning Strategies," *J. Performance Evaluation,* vol. 63, no. 7, July 2006.

[33] D. Moore, C. Shannon, G.M. Voelker, and S. Savage, "Internet Quarantine: Requirements for Containing Self-Propagating Code," *Proc. IEEE INFOCOM '03,* Mar. 2003.

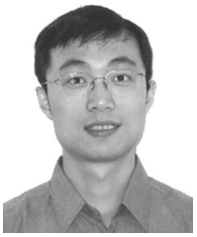
[34] C. Zou, "Internet Email Worm Propagation Simulator," http://www.cs.ucf.edu/~czou/research/emailWormSimulation.html, 2005.

[35] M. Veeraraghavan, "How Long to Run Simulations—Confidence Intervals," http://www.ece.virginia.edu/~mv/edu/prob/stat/how-to-simulate.doc, 2005.

[36] M. Jovanovic, F. Annexstein, and K. Berman, "Modeling Peer-to-Peer Network Topologies through Small-World Models and Power Laws," *Telecomm. Forum,* Nov. 2001.

[37] K. Trivedi, *Probability and Statistics with Reliability, Queuing and Computer Science Applications.* John Wiley & Sons, 2001.

[38] C. Zou, D. Towsley, and W. Gong, "Email Virus Propagation Modeling and Analysis," Technical Report TR-03-CSE-04, Electrical and Computer Eng. Dept., Univ. of Massachusetts, http://www.cs.ucf.edu/~czou/research/emailvirus-techreport.pdf, May 2003.



Cliff C. Zou (M '05) received the PhD degree from the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, in 2005. He is an assistant professor in the School of Electrical Engineering and Computer Science, University of Central Florida. His research interests include computer and network security, network modeling, and performance evaluation. He is a member of the IEEE.

Don Towsley (M '78–SM '93–F '95) received the BA degree in physics and the PhD degree in computer science, both from the University of Texas. He is currently a distinguished university professor in the Department of Computer Science at the University of Massachusetts, Amherst, where he codirects the Networking Research Laboratory. He has been a visiting scientist at the AT&T Labs-Research, IBM Research, Institut National de Recherche en Informatique et en Automatique (INRIA), Microsoft Research Cambridge, and the University of Paris 6. His research interests include network measurement, modeling, and analysis. He currently serves as editor-in-chief of the *IEEE/ACM Transactions on Networking* and on the editorial boards of the *Journal of the ACM* and *IEEE Journal of Selected Areas in Communications*. He is currently the chair of the International Federation for Information Processing (IFIP) Working Group 7.3 on computer performance measurement, modeling, and analysis. He has also served on numerous editorial boards, including those of the *IEEE Transactions on Communications* and *Performance Evaluation*. He has been active on the program committees for numerous conferences including IEEE INFOCOM, ACM SIGCOMM, ACM SIGMETRICS, and IFIP performance conferences for many years and has served as technical program cochair for ACM SIGMETRICS and performance conferences. He has received the 2007 IEEE Keji Kobayashi Computer and Communications Award, the 1999 IEEE Communications Society William Bennett Award, and several conference/workshop best paper awards. He is also the recipient of the University of Massachusetts Chancellor's Medal and the Outstanding Research Award from the College of Natural Science and Mathematics at the University of Massachusetts. He is one of the founders of the Computer Performance Foundation. He has been the recipient of the IBM Faculty Fellowship Award twice. He is a fellow of the IEEE and the ACM.

Weibo Gong (S '87–M '87–SM '97–F '99) received the PhD degree from Harvard University in 1987. He has been with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, since then. He is also an adjunct professor in the Department of Computer Science at the same campus. His major research interests include control and systems methods in communication networks, network security, and network modeling and analysis. He is a recipient of the *IEEE Transactions on Automatic Control* 's George Axelby Outstanding paper award, an IEEE Fellow, and the program committee chair for the 43rd IEEE Conference on Decision and Control.