Modeling and Simulation Study of the Propagation and Defense of Internet Email Worm

Cliff C. Zou*, Don Towsley‡, Weibo Gong†
* School of Electrical Engineering and Computer Science
University of Central Florida, Orlando FL
‡Department of Computer Science, University of Massachusetts, Amherst MA
†Department of Electrical & Computer Engineering, University of Massachusetts, Amherst MA
czou@cs.ucf.edu, gong@ecs.umass.edu, towsley@cs.umass.edu

***Abstract*— As many people rely on email communications for business and everyday life, Internet email worms constitute one of the major security threats for our society. Unlike scanning worms such as Code Red or Slammer, email worms spread over a logical network defined by email address relationship, making traditional epidemic models invalid for modeling the propagation of email worms. In addition, we show that the topological epidemic models presented in [1], [2], [3], [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 email worm propagation in this paper. We present an email worm simulation model that accounts for the behaviors of email users, including email checking time and the probability of opening an email attachment. Our observations of email lists suggest that an Internet email 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 email 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 email worms is mixed: email worms spread more quickly 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, email worm, worm modeling, epidemic model, simulation**

I. Introduction

Computer viruses and worms have been studied for a long time both by 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, email, exploiting security holes in software, etc.

Currently, email 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 *email worm* in the research area—a computer program can be called an email worm as long as it can replicate and propagate by sending copies of itself through email messages.

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

- Sending malicious codes through email 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 email service.
- A large number of users have little knowledge of email worms and trust most email they receive, especially email from their friends [7].

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

The contributions of this research work are summarized in the following:

- We show in Section V that the topological epidemic models for modeling epidemic spreading in topological networks presented in [1], [2], [3], [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 email worm simulation model that accounts for the behaviors of email users, including email checking frequency and the probability of opening an email attachment.
- Our observation shows that the size of email groups follows a heavy-tailed distribution. Since email groups greatly affect the email network topology, we believe the Internet email network is also heavy-tailed distributed and we model it as a power law network.
- We carry out extensive simulation studies of email worm propagation. From these experiments we derive a better understanding of the dynamics of an email 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 email checking time affect worm infection, etc.
- We gain insight into the differences among power law, small world, and random graph topologies by comparing email worm propagation patterns. The impact of power law topology on the spread of email worms is mixed: email worms spread more quickly 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 II introduces related work. An email worm simulation model is presented in Section III. In Section IV we discuss email network topology and model it as a power law topology. In Section V we show why previous differential equation models are not accurate for email worm modeling, the primary reason why we rely on simulations to study email worms in this paper. We present extensive simulation studies of email worm propagation without considering immunization in Section VI. In Section VII, we study immunization defense against email worms and the corresponding percolation problem. Finally, Section VIII concludes this paper with some discussions.

II. Related Work

Kephart, White, and Chess published a series of papers from 1991 to 1993 on viral infection based on epidemiology models [8], [9], [10]. [8], [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 email 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], [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 V, models presented for scanning worms are not suitable for modeling the propagation of email worms, due to the topological email 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], [19] only studied the final stable state of epidemic propagation, while we study the propagation transient dynamics as an email 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. but the authors only studied the epidemic threshold in the stable state. Later, Moreno et al. [2], [3] and Boguna et al. [1] provided Susceptible-Infectious-Recovered (SIR) differential equation models to study the dynamics of epidemic spreading on topological networks. We show in Section V 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 tree-like 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, make it unsuitable for modeling the propagation of email worms over the Internet. Wong et al. [22] provided the analysis of two email worms, SoBig and MyDoom, based on the monitored trace from a campus network. Newman et al. [23] showed that the email network distribution on a campus follows exponential or stretched exponential distributions. However, such a conclusion was derived based on the number of email addresses in the address books of the campus users. It did not consider the significant impact of email lists, nor the fact that most email worms target all email addresses found on compromised computers, not just users' email 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 tree-like hierarchic topology. From an email worm's point of view, the connectivity of a partly immunized email network is a *percolation* problem. Newman et al. [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 mostly-connected nodes, the formulas presented in [24], [25] are not suitable. Albert et al. [26] were the first to explain the vulnerability of power law networks under attacks: by selectively attacking the mostly-connected nodes, a power law network tends to be broken into many isolated fragments. They concluded that the power law topology was vulnerable under deliberate attack. This conclusion is consistent with our results derived from selective immunization defense study, as described in Section VII-B.

III. Email Worm Propagation Simulation Model

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

Because an email worm spreads on a logical network defined by email address relationship, it's difficult to mathematically analyze email worm propagation. In Section V, 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 mathematical analysis in order to focus on realistic scenarios of email worm propagation.

Strictly speaking, an email logical network is a directed graph: each vertex in the graph represents an email user, while a directed edge from node A to node B means that user B's email 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 email to user B before an email worm spreads out. Thus, user B's computer has a great chance of containing the email address of user A as well. For this reason, most edges in the email logical network can be treated as undirected edges. Therefore, in this paper we model the Internet email network as an undirected graph.

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

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

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

Email checking time T_i of user i ($i = 1, 2, \cdots, |V|$) is a stochastic variable determined by the user's habit. 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 email once every morning or uses email client programs to fetch and check email 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 email at a random time. In Section VI-H we study how different distributions of email checking time affect the propagation of an email 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 email worm (for example, MyDoom infected more users than any email worm before due to its advanced social engineering techniques [6]). For the propagation of *one* email worm, we assume C_i to be constant for user i.

We assume that email 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 email checking time, $E[T_i]$, is itself a random variable, denoted by T. When a user checks email, the user checks all new email received since the last checking time.
- User i opens a worm attachment with probability C_i when the user checks a worm email. Let C denote the random variable that generates C_i, $i = 1, 2, \cdots, |V|$.
- Because the number of email 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, $T \sim N(\mu_{\mathrm{T}}, \sigma_{\mathrm{T}}^2)$, $C \sim N(\mu_{\mathrm{C}}, \sigma_{\mathrm{C}}^2)$. 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:

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

$$E[T_i] = \max\{T, 0\} \tag{2}$$

An email user is called *infected* once the user opens a worm email attachment; upon opening a worm attachment, an infected user immediately sends out a worm email to all neighbors. Let $I(0)$ denote the number of initially infected users that send out a worm email 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 email worm propagation, $I(0) \leq I(t) \leq |V|$, $\forall t > 0$.

TABLE I

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.

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

IV. HEAVY-TAILED EMAIL NETWORK TOPOLOGY

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

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

Let $f(k)$ be the fraction of nodes with degree k in an email 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 email 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 email lists, especially the popular email lists, are much larger than the number of email addresses existing in normal computers, we believe the Internet email network topology is mainly determined by the topology property of email lists. The popular Yahoo! email groups are heavy-tailed distributed, as shown in Fig. 1, which suggests that the Internet-scale email network is probably also heavy-tailed distributed.

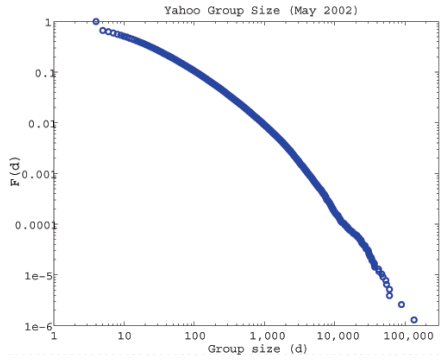


Fig. 1. Complementary cumulative distr. of Yahoo! group size (in May 2002)

In order to generate a heavy-tailed email 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 email network, although the degree of a real Internet email network may not be strictly power law distributed. In this paper we use the GLP power law generator presented in [28]. We choose the GLP power law network generator instead 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 email network because they do not provide a heavy-tailed degree distribution. However, in order to understand how a heavy-tailed email topology affects email worm propagation, we also study email 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 again.

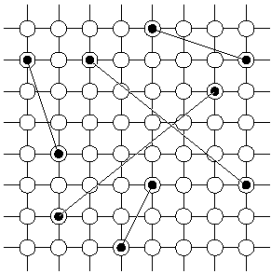


Fig. 2. Illustration of a two-dimensional small world network

We generate the small world network by using the model presented in [31], 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 $E[k] \cdot n/2$.

V. 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 III 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 *susceptible-infectious-susceptible*. If infected hosts cannot become susceptible again once they are cured, the models are called *SIR* models, hosts can only have the status transition as susceptible-infectious-recovered (or *SI* models if no infected hosts can recover). For modeling of the propagation of a single email worm incident, after an email user cleans his or her infected computer, the user is not likely to open another copy of the same worm email 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 email worms, thus we do not consider the recovery process and focus solely on SI models.

A. 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], [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 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 the 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 "hit-list" worm, "flash" worm [18], "local preference" worm (such as Blaster worm and Sasser worm) and "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 email worm can only spread hop-by-hop on an email logical network. We must consider topological issues in its modeling. Since the homogeneous assumption will not stand for email worm modeling, we cannot use the above model (3) or its extensions in this paper.

B. 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, 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]:

$$\frac{di_k(t)}{dt} = \lambda k[1 - i_k(t)]\Theta(t) \tag{4}$$

$$\Theta(t) = \frac{\sum_n nP(n)i_n(t)}{\sum_s sP(s)} = \frac{\sum_n nP(n)i_n(t)}{E[k]}$$

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 a 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 $\Theta(t)$:

$$\frac{di_k(t)}{dt} = \lambda k[1 - i_k(t)]\Theta(t) \tag{5}$$

$$\Theta(t) = \frac{\sum_n (n-1)P(n)i_n(t)}{E[k]}$$

When a node has more edges, it has a higher probability of being infected quickly by an epidemic (or an email 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 or 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.

C. 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.

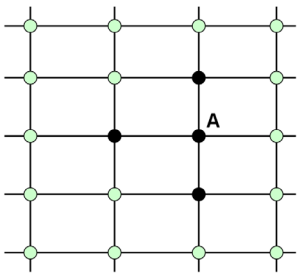


Fig. 3. Illustration of an epidemic spreading in a topological network

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 3 out of 4 neighboring nodes a moment later (labelled as black nodes). At this time, the epidemic has 10 links, called *effective infection links*, that connect infected nodes with susceptible ones. The 3 links interconnecting those 4 infected nodes have no contribution to the epidemic spreading later. On the other hand, if these 4 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.

Model (5) is better: the 3 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 would 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.

D. 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 IV. 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:

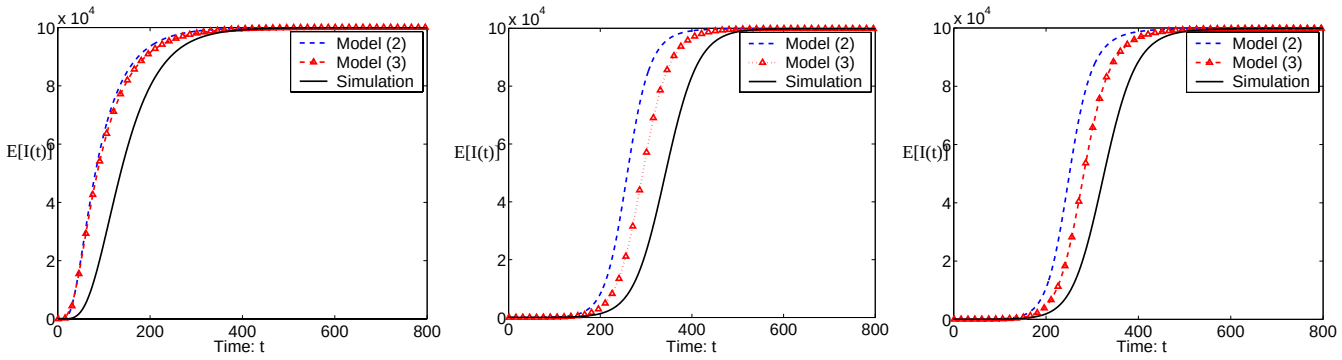
$$i_k(t) = i_k(t-1) + \lambda k[1 - i_k(t-1)]\Theta(t-1) \qquad (6)$$

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

$$I(t) = \sum_k i_k(t)P(k)N \qquad (7)$$

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 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.



a. Power law topology b. Small world topology c. Random graph topology

Fig. 4. Numerical solutions of model (4) and model (5) compared with the epidemic simulation results on three different topologies

It clearly shows that models (4) and (5) overestimate the propagation speed of 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 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 email 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 email worm propagation.

VI. Email Worm Simulation Studies

A. Description of the discrete-time email worm simulator

Discrete-time simulation has been used in many worm modeling papers [9], [33], [21], [17]. Thus, we simulate email worm propagation in discrete time, too. All events (worm infection, user checking email, etc.) are assumed to happen right at each discrete time tick. Before the start of an email 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 email 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 email at this time tick. If user i checks email at time t, the user checks all new email received after his or her last email checking. Each new worm email is opened with probability C_i. Once a worm email is opened, user i is infected (if the user has not been infected before) and the worm will send worm email to all neighbors of the user. These worm emails could be read by their recipients as soon as the next time tick, $t + 1$. Then, a new email checking time T_i is assigned to user i in order to determine when he or she will check email 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 VI-H, we specifically study how different distributions of X affect email 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 email 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 email 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 our experiments, we find it is accurate enough to choose $E[T] = 40$. The value of C depends on how deceptive an email worm is, which varies from one worm incident to another. Thus, we choose $E[C] = 0.5$ to simulate the general case of email worm propagation.

For the convenience of readers, we have put the source codes of our email worm propagation simulator and topology generators on-line [34].

B. Reinfection vs. nonreinfection

First, we consider two cases under different infection assumptions: reinfection versus nonreinfection. *Reinfection*

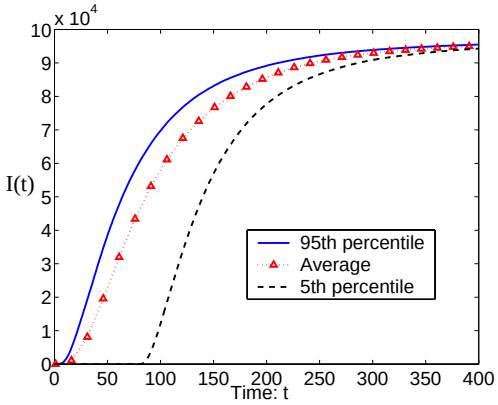


Fig. 6. 5th and 95th percentiles of 100 simulation runs

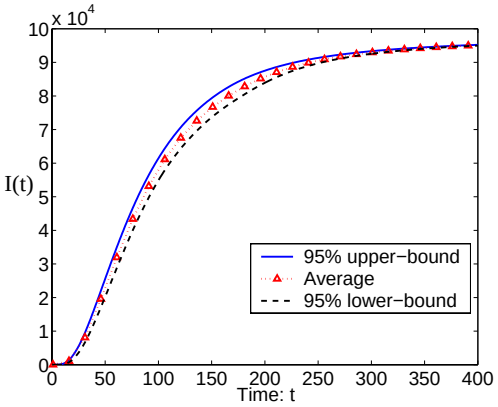


Fig. 7. 95% statistical confidence interval of 100 simulation runs

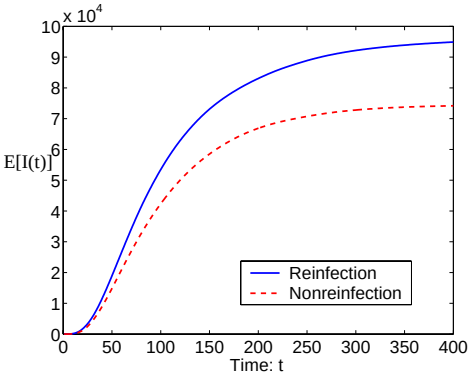


Fig. 5. Reinfection vs. nonreinfection

means that a user will send out worm email copies whenever he opens an email worm attachment. Thus, a recipient can repeatedly receive worm emails 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 email, even if he or she opens a worm attachment again. Some email 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 email network.

C. Variability in worm propagation

An email worm propagation is, in fact, a stochastic process. Under the same network condition, the same email worm could spread faster or slower in different runs. To study how variable an email worm propagation could be, we simulate the email 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 the 95th and 5th 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 5 runs the worm propagates faster than the 95th percentile curve while in another 5 runs the worm propagates slower than the 5th percentile curve. This figure shows that an email 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.

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] :

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

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.

D. Impact of user clicking probability

In our email worm model, each user i opens an email attachment with probability C_i when reading a worm email. Thus, user i has the probability $1 - (1 - C_i)^m$ to be infected when receiving m worm email—the chance of being infected increases as a function of the amount of worm email 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 probabilities of opening email 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 the nonreinfection case, user i, who has m_i edges (neighbors), will receive at most m_i copies of the worm email—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, i.e., $C_i = p, \forall i \in \{1, 2, \cdots, |V|\}$.

Let $G(x)$ denote the probability generating function of the degrees of the email 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 email users become cautious in clicking worm email attachments, a larger number of email users will stay healthy without being infected by the email worm.

The email worm has successfully spread in all 100 simulation runs. In fact, the email worm has a small chance to die 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 email attachment for the first round, then no worm email exists in the network after those neighbors finish checking their email 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 email worm has the probability $(1-p)^m$ to die before it infects any other users.

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

E. Initially infected nodes with highest degree vs. 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, while 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, while 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's important to let an email worm spread as widely as possible before people become aware of the worm. It will help an email worm to propagate faster by choosing the right initial launching points, such that those initially infected computers contain a large number of email addresses.

F. Topology effect: power law, small world, and random graph

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

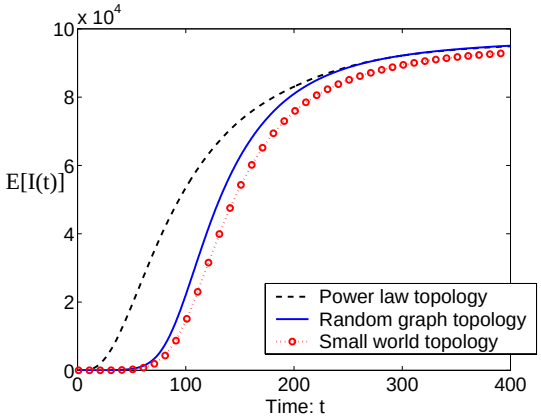


Fig. 9. Effect of topology on email worm propagation

We run our email 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 email 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 by 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.

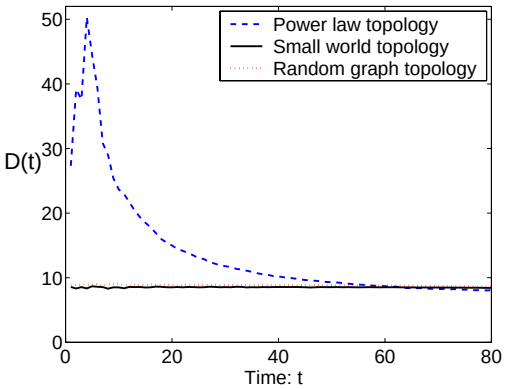
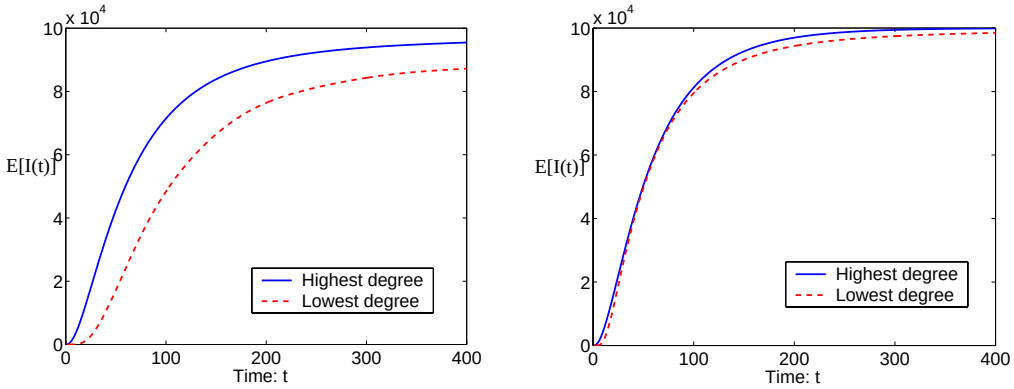


Fig. 10. Average degree of nodes that are being infected at each time tick t

We also observe from 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 has the smallest characteristic path length among those three topologies, while the other two have similar characteristic path lengths [28], [36]. *Characteristic path length* is defined as the



a. Average degree $E[k] = 8$ b. Average degree $E[k] = 20$

Fig. 8. Effect of different degrees of initially infected nodes

number of edges in the shortest path between two vertices, averaged over all pairs of vertices [30]. An email 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 email 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 varies significantly [26]. Once an email worm infects a highly-connected node, a large number of worm emails 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. 9(b) and derive $D(t)$ for each topology by averaging the results of $1,000$ simulation runs. We plot $D(t)$ 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 email worm tends to first infect some highly-connected nodes—these nodes will then send out a much larger number of worm email 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.

G. 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. 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 AS level power law topology has a power law exponent of $\alpha = 1.1475$ [28]. Using $\alpha = 1.1475$ for a $100,000$ nodes 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 email network.

On the other hand, we don't know the true value of α for the real Internet email network. In order to see how the power law exponent α affects email worm propagation, we compare worm propagation on the following 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 email 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.

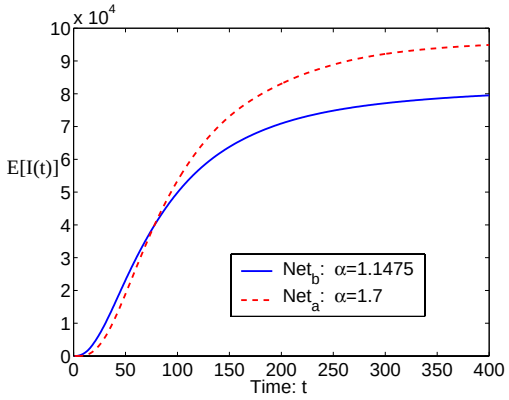
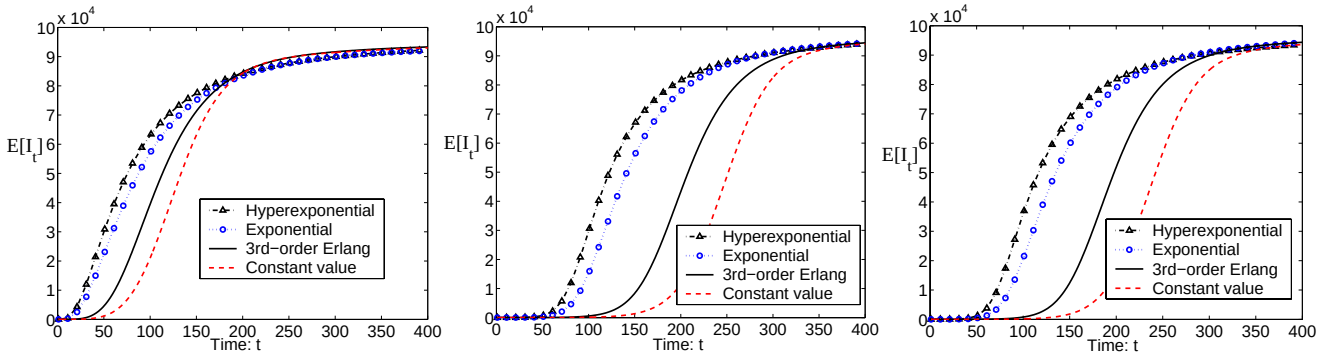


Fig. 11. Effect of power law exponent α on email worm propagation

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 email sent out than in Net_a. Those highly-connected nodes behave like amplifiers in email worm propagation (see the amplification effect explained in Section VI-F). Thus, initially an email worm spreads faster on Net_b than on Net_a.

After having infected most highly-connected nodes, the email 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 one, while in Net_a it is three. Since a node with fewer links is harder to be infected, during the second



a. Power law topology b. Small world topology c. Random graph topology

Fig. 12. Effect of the distribution of email checking time T_i

phase of worm propagation the email worm spreads slower on Net_b than on Net_a.

H. Effect of email checking time distribution

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

Fig. 12 shows $E[I(t)]$ under four different distributions of email checking time T_i: hyperexponential distribution [37], exponential distribution, 3rd-order Erlang distribution, and constant value. For comparison reason, we let each distribution to 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: $I(0) = 2$, $C \sim N(0.5, 0.3^2)$, $T \sim N(40, 20^2)$ (the average email 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 3-order Erlang distribution has $CV = 1/\sqrt{3}$, and a constant value has $CV = 0$. Fig. 12 shows that an email worm propagates faster as the email 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 *snowball* effect: before worm copies in the system with less dispersed checking time give birth to the next generation—infecting some new hosts—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.

VII. Immunization and Percolation for Email Worm Defense

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

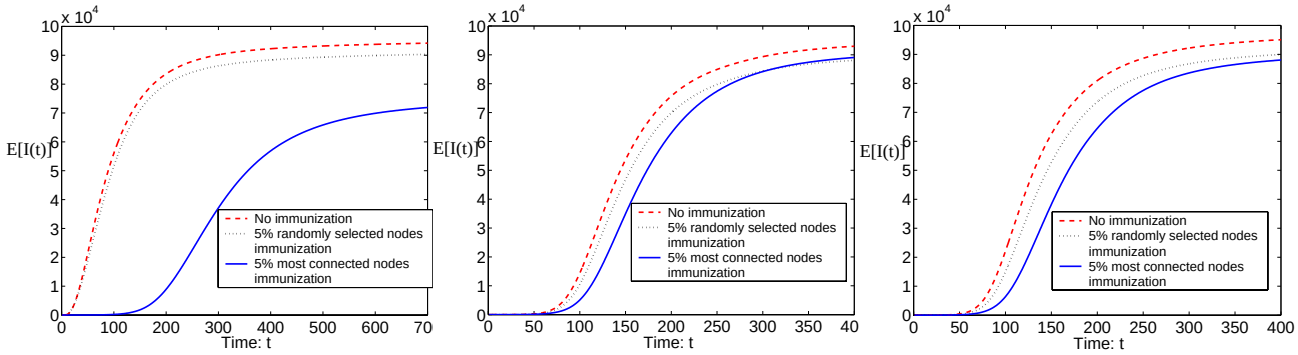
A. Effect of selective immunization

It's not possible for us to immunize all email users in the email 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 email 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 email network, selecting those most highly-connected nodes to immunize is also quite effective against email 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, while 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 from Fig. 13 that selective immunization is quite effective for a power law topology, while it has little effect for a small world topology or a random graph topology. On



a. Power law topology b. Small world topology c. Random graph topology

Fig. 13. Effect of selective immunization on email worm propagation

a power law email network, we can significantly slow down email worm propagation by selecting those most-connected nodes to immunize.

The results here are consistent with the conclusions in [26]. Albert and colleagues [26] showed that selectively attacking the most-connected nodes rapidly increases the diameter of a power law network. Since an email worm depends on the connectivity of the underlying email 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.

B. Selective percolation and email worm prevention

Having observed that selective immunization is quite effective for a power law email 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 email worm?

From an email 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 email 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 remained links after removing the top p percent of most-connected nodes from the network. Then we have:

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

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 equation (13) from those generated 100 networks for each type of topology, respectively. In 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. 14(a) 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 email network will be broken into separated fragments and no worm outbreak will occur.

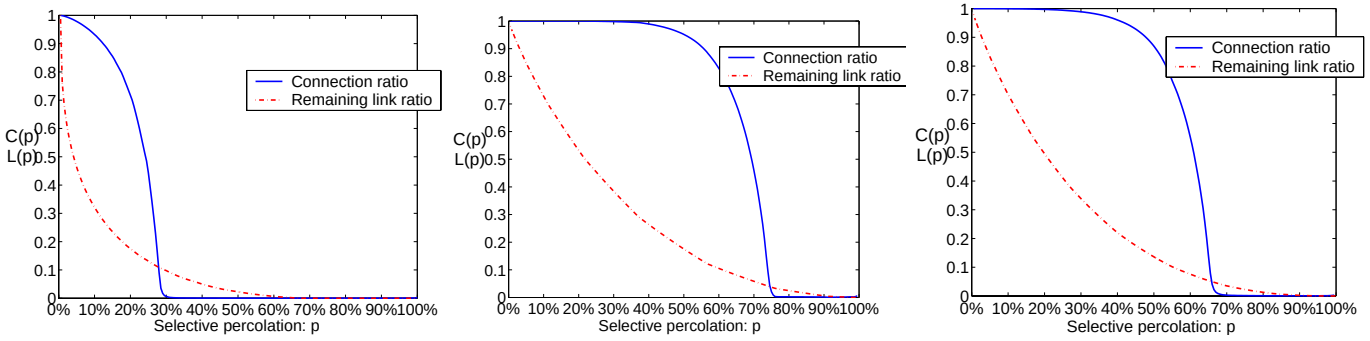
The selective percolation threshold of a power law topology is much smaller than that of either a small world topology or a random graph topology. Although a power law topology is more vulnerable under deliberate attacks [26], it benefits more from a selective immunization defense.

Fig. 14(a) shows that when we immunize the top 5 percent of most-connected nodes in a power law network, although 97.5 percent of remaining nodes are still connected, 55.5 percent of the original network edges have been removed. Thus, an email worm has fewer and longer paths to reach and infect nodes in the remaining network. Fig. 14(b)(c) 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.

VIII. Conclusion

In this paper we first show that topological epidemic models presented in [1], [2], [3], [4] largely overestimate an epidemic



a. Power law topology b. Small world topology b. Random graph topology

Fig. 14. Selective percolation on three topologies

spreading speed on a topological network due to their implicit homogeneous mixing assumption. Then, we present an email worm simulation model by considering email users' behaviors, such as email checking frequency and the probability of opening an email attachment. Given that email worms spread over a logical network defined by email address relationship, our observations of email lists suggest that the degrees in an email network are heavy-tailed distributed. To understand how the heavy-tailed topology affects email worm propagation, we compare email worm spreading on three topologies: power law, small world, and random graph; then, study how the topology affects immunization defense. From these studies, we derive a better understanding of an email 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 email worm propagation is mixed: on one hand, an email 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 email users who can send out email to a large number of users.

There is still much work to do on email worm modeling and defense. First, in this paper we have relied on simulations to study email 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 email worm propagation, since email worms do not send out random scanning. Wong and colleagues [22] only provided limited monitoring results of a campus network. Additionally, email communication traffic is hard to share due to the privacy concern. Therefore, it is hard to validate our simulation model with real email worm incidents. We plan to conduct more research on email 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 email worm, a fraction of users have already been immunized from the worm, and no more users will become immunized during the propagation of an email worm. However, the more realistic scenario is that email users and computers gradually become immunized as an email worm spreads out, which means we need to further study dynamic immunization defense against email worms.

Fourth, although we have considered the impact of email lists on the topology of Internet email network, instead of an undirected graph, a directed graph is preferred in order to more accurately capture one-way email address relationship (that is, user A has the email address of user B, but user B does not have the address of user A). In addition, there are many email lists having constraints on who can submit a broadcast messages to a mailing list (for example, only the administrator can)—such email lists need specific modeling. And finally, we need to further consider how to match the email logical network with the physical networks of email servers because a good filter on an email server will protect many email users in the logical email 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 DARPA under contract F30602-00-2-0554 and by NSF under Grant EIA-0080119, CNS-0627318. It was also supported in part by ARO contract DAAD19-01-1-0610 and contract 2000-DT-CX-K001 from the U.S. Department of Justice, Office of Justice Programs.

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," *Phys. Rev. E.*, vol. 68, 2003.
[3] Y. Moreno, R. P. Satorras, and A. Vespignani, "Epidemic outbreaks in complex heterogeneous networks," *Eur. Phys. J. B*, vol. 26, 2002.
[4] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Letters*, vol. 86, 2001.
[5] F. Cohen, "Computer viruses: theory and experiments," *Computers and Security*, vol. 6, no. 1, February 1987.
[6] CERT, "CERT/CC advisories," http://www.cert.org/advisories/.
[7] ——, "CERT advisory CA-2001-20: Continuing threats to home users," July 2001, http://www.cert.org/advisories/CA-2001-20.html.

[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," in *Proceedings of IEEE Symposium on Security and Privacy*, 1991, pp. 343–359.

[10] ——, "Measuring and modeling computer virus prevalence," in *Proceedings of IEEE Symposium on Security and Privacy*, 1993.

[11] Z. Chen, L. Gao, and K. Kwiat, "Modeling the spread of active worms," in *Proceedings of the IEEE INFOCOM*, March 2003, pp. 1890–1900.

[12] G. Kesidis, I. Hamadeh, and S. Jiwasurat, "Coupled kermack-mckendrick models for randomly scanning and bandwidth-saturating internet worms," in *Proceedings of 3rd International Workshop on QoS in Multiservice IP Networks (QoS-IP*, February 2005, pp. 101–109.

[13] D. Nicol and M. Liljenstam, "Models of Internet worm defense," January 2004, IMA Workshop 4: Measurement, Modeling and Analysis of the Internet. http://www.ima.umn.edu/talks/workshops/1-12-16.2004/nicol/talk.pdf.

[14] D. Nojiri, J. Rowe, and K. Levitt, "Cooperative response strategies for large scale attack mitigation," in *Proceedings of 3rd DARPA Information Survivability Conference and Exhibition*, April 2003.

[15] J. Wu, S. Vangala, L. Gao, and K. Kwiat, "An efficient architecture and algorithm for detecting worms with various scan techniques," in *Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS'04)*, February 2004.

[16] C. Zou, L. Gao, W. Gong, and D. Towsley, "Monitoring and early warning for Internet worms," in *Proceedings of 10th ACM Conference on Computer and Communications Security (CCS'03)*, October 2003, pp. 190–199.

[17] C. Zou, W. Gong, and D. Towsley, "Code Red worm propagation modeling and analysis," in *Proceedings of 9th ACM Conference on Computer and Communications Security (CCS'02)*, October 2002, pp. 138–147.

[18] S. Staniford, V. Paxson, and N.Weaver, "How to own the Internet in your spare time," in *Proceedings of USENIX Security Symposium*, August 2002, pp. 149–167.

[19] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic spreading in real networks: An eigenvalue viewpoint," in *Proceedings of 22nd Symposium on Reliable Distributed Computing*, October 2003.

[20] A. Ganesh, L. Massoulie, and D. Towsley, "The effect of network topology on the spread of epidemics," in *Proceedings of the IEEE INFOCOM*, March 2004.

[21] C. Wang, J. C. Knight, and M. C. Elder, "On viral propagation and the effect of immunization," in *Proceedings of 16th ACM Annual Computer Applications Conference*, December 2000.

[22] C. Wong, S. Bielski, J. M. McCune, and C. Wang, "A study of massmailing worms," in *Proceedings of ACM CCS Workshop on Rapid Malcode (WORM'04)*, October 2004.

[23] M. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Phys. Rev. E.*, vol. 66, no. 035101, 2002.

[24] C. Moore and M. Newman, "Exact solution of site and bond percolation on small-world networks," *Phys. Rev. E.*, vol. 62, 2000.

[25] M. Newman, S. Strogatz, and D. Watts, "Random graphs with arbitrary degree distributions and their applications," *Phys. 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.

[28] T. Bu and D. Towsley, "On distingishing between Internet power law topology generators," in *Proceedings of the IEEE INFOCOM*, June 2002.

[29] P. Erdos, "Graph theory and probability," *Canad. 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," *Phys. Rev. E.*, vol. 65, no. 021904, 2002.

[32] C. Zou, D. Towsley, and W. Gong, "On the performance of Internet worm scanning strategies," *Journal of 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," in *Proceedings of the IEEE INFOCOM*, March 2003.

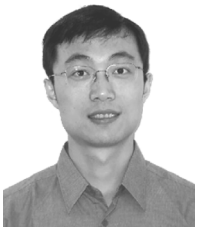
[34] C. Zou, "Internet email worm propagation simulator," http://www.cs.ucf.edu/ czou/research/emailWormSimulation.html.

[35] M. Veeraraghavan, "How long to run simulations - confidence intervals," http://www.ece.virginia.edu/ mv/edu/prob/stat/how-to-simulate.doc.

[36] M. Jovanovic, F. Annexstein, and K. Berman, "Modeling peer-to-peer network topologies through small-world models and power laws," in *Telecommunications Forum*, November 2001.

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

[38] C. Zou, D. Towsley, and W. Gong, "Email virus propagation modeling and analysis," Umass ECE Dept., http://www.cs.ucf.edu/ czou/research/emailvirus-techreport.pdf, Tech. Rep. TR-03-CSE-04, May 2003.



Cliff C. Zou (M'05) received the Ph.D degree in Department of Electrical and Computer Engineering from University of Massachusetts, Amherst, MA, in 2005.

He is an Assistant Professor in School of Electrical Engineering and Computer Science, University of Central Florida. His research interests include computer and network security, network modeling and performance evaluation.

Don Towsley (M'78-SM'93-F'95) received a B.A. degree in physics and his Ph.D. degree in computer science, both from University of Texas University. He is currently Distinguished University Professor in the Department of Computer Science at the University of Massachusetts - Amherst, where he co-directs the Networking Research Laboratory. Professor Towsley has been a Visiting Scientist at AT&T Labs - Research, IBM Research, INRIA , Microsoft Research Cambridge, and the University of Paris 6.

His research interests include network measurement, modeling, and analysis. Dr. Towsley currently serves as Editor-in-Chief of the *IEEE/ACM Transactions on Networking* and on the editorial boards of *Journal of the ACM* and *IEEE Journal of Selected Areas in Communications*. He is currently the Chair of the IFIP Working Group 7.3 on computer performance measurement, modeling, and analysis. He has also served on numerous editorial boards including those of *IEEE Transactions on Communications* and *Performance Evaluation*. He has been active in 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 Co-Chair 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 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 twice been the recipient of IBM Faculty Fellowship Awards and is a Fellow of the IEEE and the ACM.

Weibo Gong (S'87-M'87-SM'97-F'99) received his Ph.D degree from Harvard University in 1987, and have been with the Dept. of Electrical and Computer Engineering, University of Massachusetts, Amherst since then. He is also an adjunct professor in the Dept. 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.