Email Worm Modeling and Defense

Cliff C. Zou*, Don Towsley†, Weibo Gong*
*Department of Electrical & Computer Engineering
†Department of Computer Science
University of Massachusetts, Amherst, MA 01003
{czou, gong}@ecs.umass.edu, towsley@cs.umass.edu

***Abstract*— Email worms constitute one of the major Internet security problems. In this paper, we present an email worm model that accounts for the behaviors of email users by considering email checking time and the probability of opening email attachments. Email worms spread over a logical network defined by email address relationship, which plays an important role in determining the spreading dynamics of an email worm. Our observations suggest that the node degrees of an email network are heavy-tailed distributed. We compare email worm propagation on three topologies: power law, small world and random graph topologies; and then study how the topology affects immunization defense on email worms. The impact of the power law topology on the spread of email worms is mixed: email 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 than on the other two.**

I. Introduction

"Email worms" are malicious computer programs that propagate through email: when an email user clicks a worm program in the attachments of a worm email, the worm compromises the user's computer and then finds all email addresses stored on this computer to send out worm email.

Email has become an indispensable communication medium in our life. However, email worms keep attacking us with increasing intensity and using more advanced social engineering tricks. Some famous email worms include *Melissa* in 1999, "*Love Letter*" in 2000, "*W32/Sircam*" in 2001, "*SoBig*" in 2003, "*MyDoom*", "*Bagle*" and "*Netsky*" in this year [1].

Like earthquake modeling or tornado modeling, a good email worm model gives us deep understanding of email worms, helps us evaluate the effectiveness of defense mechanisms, and provides possible early warning to help us control a worm's potential damage. In this paper, we first present a realistic email worm model that accounts for the behaviors of email users by considering users' email checking time and the probability of opening email attachments. Then we carry out extensive simulation studies. Email worms spread over a logical network defined by email addresses; our observation shows that the Internet-scale email network might be heavy-tailed distributed and we model it as a power law network. To study how topology affects an email worm's propagation, we compare worm spreading on power law, small world and random graph topologies. We derive the conclusion that email worms spread more quickly on a power law topology than on the other two topologies; the other two topologies have little differences in terms of the propagation dynamics of email worms.

Based on the above email worm model, we study the effectiveness of "selective immunization defense", i.e., selecting some most connected email users' computers to make them immune to an email worm. The results explain why selective immunization defense against email worm propagation 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 surveys related work. Section III presents the email worm propagation model. In Section IV, we explains why we choose power-law topology to represent the logical email network. We conduct extensive simulation studies and present the results in Section V. Then we study selective immunization defense against email worm attacks in Section VI. Finally, we make some discussions and conclude the paper with Section VII.

II. Related work

Kephart, White and Chess studied viral infection based on epidemiology models from 1991 to 1993 [2][3][4], where they considered virus spreading on random graph networks and local networks. After Code Red worm incident [1] in 2001, many researchers have studied worm propagation modeling [5][6][7][8]. However, these papers studied "scan-based" worms that propagate through random scanning — models of scan-based worms do not need to consider topological issues and thus not suitable for modeling email worms.

Wang *et al.* [9] simulated a simple virus propagation based on a clustered topology and a tree-like hierarchic topology. Newman *et al.* [10] collected email address book data from a university and showed that the campus-level email topology has an exponential and a stretched exponential distribution for in-degree and out-degree, respectively. However, they didn't consider email lists, which can dramatically increase edges in an email network. In addition, most email worms send out worm email to all addresses existed in a computer, not just in email address books. Thus the Internet-scale email network may follow a completely different topology. Briesemeister *et al.* [11] used simulation to study epidemic spreading on scale-free networks without considering other topologies and user interactions. Satorras *et al.* [12] used SIS epidemic model to study epidemic spreading in scale-free networks — the SIS model is not suitable for modeling the propagation of one

email worm because a cured host is not likely to become susceptible to the same email worm again.

From an email worm's point of view, the connectivity of a partly immunized email network is a "percolation" problem. Newman *et al.* [13] derived the analytical solution of the "percolation threshold" for arbitrary topologies: if nodes are removed randomly from a network, the network will be broken into pieces when the fraction of removed nodes is higher than the network's "percolation threshold". However, it is more effective and reasonable to "selectively immunize" nodes in an email network rather than the uniform immunization studied in [13]. Albert *et al.* [14] showed that a power law network is vulnerable under selective attack, which is consistent to our conclusions.

III. Email worm propagation model

We represent the topology of the logical email network by an undirected graph $G =< V, E >$. $\forall v \in V$, v denotes an email user; $\forall e = (u, v) \in E$, $u, v \in V$, represents that two users u and v have the email address of each other in their own computers. "Node degree" d of a node means that this node has d edges connecting to d other nodes. $|V|$ is the total number of email users. For a reader's convenience, Table I lists most notations used in this paper.

TABLE I

Notations used in this paper

Symbol	Explanation
G	Undirected graph representing email network, $G =< V, E >$
N_t	Number of infected users at time t
α	Power law exponent of a power law topology that has complementary cumulative degree distribution $F(d) \propto d^{-\alpha}$
N_∞^h	Number of users uninfected when a worm finishes spreading
T_i	Email checking time interval of user i, $i = 1, 2, \cdots, \|V\|$
P_i	Probability of user i to open email worm attachments
T	Gaussian-distributed random variable that generates $E[T_i]$, $T \sim N(\mu_T, \sigma_T^2)$ ($E[T_i] = 0$ when $T < 0$)
P	Gaussian-distributed random variable that generates P_i, $P \sim N(\mu_P, \sigma_P^2)$ ($P_i = 0$ when $P < 0$ and $P_i = 1$ when $P > 1$)
D_t	Average degree of those nodes that are healthy before time t but are infected at time t, $\forall t > 0$
$L(p)$	Remained link ratio — fraction of links remained after removal of the top p percent most connected nodes
$C(p)$	Connection ratio — fraction of remained nodes that are connected after removal of the top p most connected nodes

Email worms depend on email users' interaction to propagate. There are primarily two human behaviors affecting email worms: one is the *email checking time interval*, denoted by T_i, $i = 1, 2, \cdots, |V|$, which is the time interval between two consecutive email checking by user i; another is the *opening probability*, denoted by P_i, the probability with which user i opens a worm attachment.

Some email worms exploit email clients' bugs such that they can compromise computers without users to execute any attachment. Such email worms can be modeled by assigning $P_i \equiv 1$ for those vulnerable computers.

Email checking time of a user is a stochastic variable determined by the user's habit. Denote $E[T_i]$ as the mean value of the checking time interval T_i for user i, $i = 1, 2, \cdots, |V|$. T_i may follow different distributions. For example, it is a constant value when a user checks email once every morning or uses email client programs to fetch and check email at a specified time interval; it is exponentially distributed (i.e., checking action is a Poisson process) if a user checks email at a random time. In this paper we will study how different distributions of email checking time interval affect the propagation of an email worm.

Since the number of email users in the Internet is very large and users' behaviors are independent, we assume that the mean checking email time $E[T_i]$ of user i is generated by a Gaussian distributed random variable T, i.e., $T \sim N(\mu_T, \sigma_T^2)$ ($E[T_i] = 0$ when $T < 0$). We assume that when a user checks his email, he checks all new email in his mailbox.

The opening probability P_i of user i is determined by: (1) the user's security awareness; and (2) the social engineering tricks deployed by an email worm (e.g., "MyDoom" infected more users than any email worm before due to its advanced social engineering techniques [1]). Therefore, for the propagation of one email worm, we assume P_i to be constant for user i. Similar to $E[T_i]$, we assume P_i of user i is generated by a Gaussian distributed random variable P, i.e., $P \sim N(\mu_P, \sigma_P^2)$ ($P_i = 0$ when $P < 0$ and $P_i = 1$ when $P > 1$).

An email user is called *infected* once the user opens an email worm attachment; upon opening a worm attachment, an infected user immediately sends out worm email to all his neighbors. Let N_0 denote the number of initially infected users. Let random variable N_t denote the number of infected users at time t during email worm propagation, $N_0 \leq N_t \leq |V|$, $\forall t > 0$. It takes time before a recipient receives an email worm sent out by an infected user; but the email transmission time is usually much smaller compared with a user's email checking time interval, and thus, it is ignored in our model.

IV. Email network topology discussion

Let $f(d)$ be the fraction of nodes with node degree d in email network graph G. The complementary cumulative distribution function (ccdf) is denoted by $F(d) = \sum_{i=d}^{\infty} f(i)$, i.e., the fraction of nodes with degree greater than or equal to d. We have examined more than 800,000 email groups in *Yahoo!* [15], the sizes of which vary from as low as 4 to more than 100,000. Fig. 1 presents 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*, i.e., the ccdf $F(d)$ decays slower than exponentially [16].

Currently, "email groups", or so-called "email lists", have become very popular. Once a user has the address of an email group in his address book or stored in his computer, from an email worm's point of view, this user virtually has all the addresses contained in the email group. Therefore, even though a user's computer may only contains tens of email addresses, the node degree of the user in the email network graph might be as large as several thousand if one of the email address is a popular email group. Since email groups are heavy-tailed

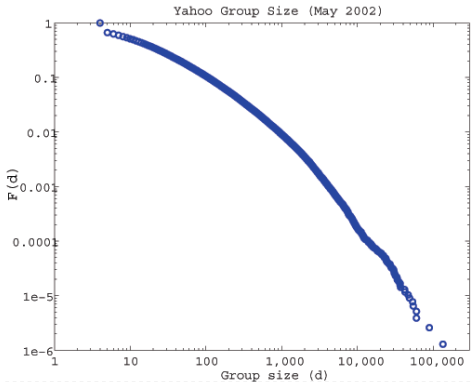


Fig. 1. Complementary cumulative distr. of Yahoo! group size (in May 2002)

distributed as shown in Fig. 1, it suggests that the Internet-scale email network is also heavy-tailed distributed.

In this paper we use the power law generator in [16] to generate power law topologies. The node degree of a "power law topology" is heavy-tailed distributed and has the power law ccdf $F(d) \propto d^{-\alpha}$, which is linear on a log-log plot [16]. Except power law topology generators, there is no other network generator available to create a heavy-tailed distributed topology. Thus a power law topology generator is the best candidate to generate the email network although the node degree of a real email network may not be strictly power law distributed.

There are other popular topologies such as random graph topology and small world topology [17]. We study worm propagation on these topologies as well in order to understand how different topologies affect email worm propagation. We generate the small world topology by using the two-dimensional small world model presented in [18].

V. Email Worm Simulation Studies

We are interested in $E[N_t]$ — the average number of infected users at any time t. In all of our simulation experiments, we derive $E[N_t]$ by averaging the results of N_t from 100 simulation runs with different seeds to generate random numbers. The underlying power law network has 100,000 nodes, average node degree 8 and power law component $\alpha = 1.7$. Except the experiments on distributions of email checking time interval, in other experiments we assume that the email checking time interval T_i of user i follows a Poisson process with rate $1/E[T_i]$, $i = 1, 2, \cdots, |V|$, where $E[T_i]$ follows $T \sim N(40, 20^2)$. Other simulation parameters are: $P \sim N(0.5, 0.3^2)$ and $N_0 = 2$. Initially infected nodes are randomly chosen in each simulation run.

A. Reinfection vs. Non-reinfection

First we consider two cases with different infection assumptions: the *reinfection* case versus the *non-reinfection* case. *Reinfection* means that the computer of a user sends out email worm copies whenever the user opens a worm attachment. *Non-reinfection* means that an infected computer of a user sends out worm copies to all its neighbors only once, after which it will not send out any worm email again even if the user opens worm attachments repeatedly. Fig. 2 illustrates the behavior of $E[N_t]$ as a function of time t on a power law email network.

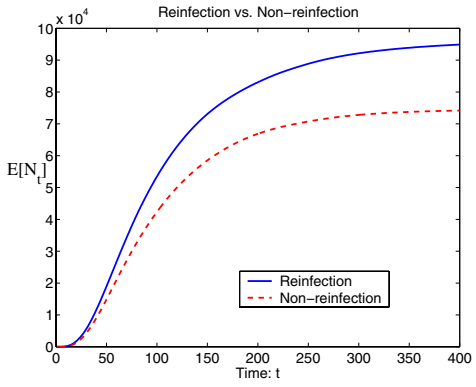


Fig. 2. Reinfection vs. non-reinfection

In our email worm model, user i opens a worm attachment with probability P_i when he checks a worm email. Thus user i has the probability $1-(1-P_i)^m$ to be infected when receiving m worm email — this is the reason why more users are infected in the reinfection case than in the non-reinfection case as shown in Fig. 2.

Since some users never open worm email attachments or open them with very low probabilities, in both cases a certain number of users will not be infected when the worm propagation is over. Let N_∞^h denote the number of users that are not infected when the worm propagation is over. In the non-reinfection 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 $(1-P_i)^{m_i}$. Let $G(x)$ denote the probability generating function of the node degrees of the email network:

$$G(x) = \sum_{k=1}^{\infty} P(d = k)x^k \tag{1}$$

where $P(d = k)$ is the probability a node has degree k. When all users are equally likely to open worm attachments, i.e., $P_i = p, \forall i \in \{1, 2, \cdots, |V|\}$, we derive the lower bound for $E[N_\infty^h]$ as:

$$E[N_\infty^h] \geq |V| \sum_{k=1}^{\infty} P(d = k)(1-p)^k = |V|G(1-p) \tag{2}$$

A reinfection email worm propagates faster and thus is the focus of our study. In the following, we only consider reinfection email worms.

B. Topology effect: Power law, Small world and Random graph topologies

The topology of email logical network plays an important role in determining the behaviors of an email worm's propagation. In this section we study the impact of different topologies on email worm propagation. Through this study, we can have

better understanding of what factors affect an email worm's spreading speed, which makes it possible for us to find out the appropriate defense mechanism by taking advantage of such properties.

We run the 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 8 and $100,000$ nodes. Fig. 3 shows the $E[N_t]$ as a function of time t of these three topologies.

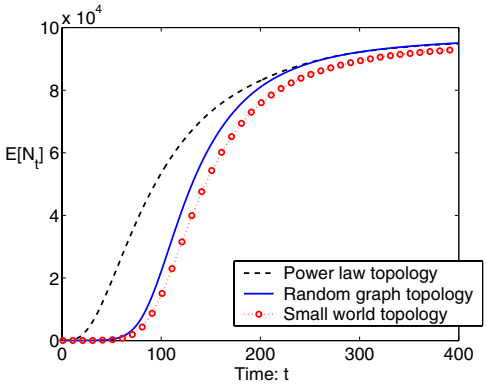


Fig. 3. Topology effect on email worm propagation

To study why worm propagates faster on a power law network, let D_t denote the average degree of those nodes that are healthy before time t but are infected at time t. D_t tells us what nodes are being infected at each time t, $t = 1, 2, 3, \cdots$. We repeat the experiment in Fig. 3 and derive D_t for each topology by averaging the results of $1,000$ simulation runs. We plot each D_t of these three networks as a function of time t in Fig. 4. Note that the D_t of a small world and a random graph networks are almost the same.

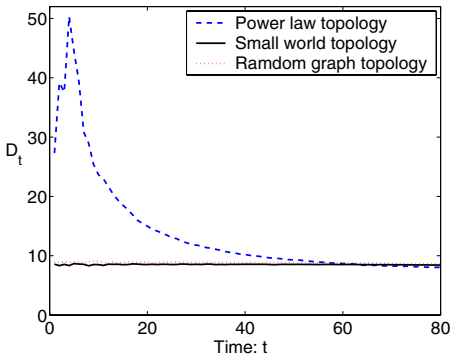


Fig. 4. Average degree of nodes that are being infected at each time tick

Fig. 4 clearly shows that on a power law network an email worm tends to first infect most highly connected nodes — these nodes send out a much larger number of worm email copies 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 the similar node degrees.

Another reason for the propagation speed difference of these three topologies is their different *"characteristic path length"*, which is defined as the number of edges in the shortest path between two vertices averaged over all pairs of vertices [17]. For an email network, a smaller characteristic path length means that an infected user needs a smaller number of steps to reach other users, and thus an email worm would propagates faster. [16][19] show that a power law topology has the smallest characteristic path length among those three topologies while the other two have the similar characteristic path lengths.

We also investigate the sensitivity of our results to the scale of an email network. We run the same experiments as shown in Fig. 3 on a $1,000,000$-node email network (the average node degree remains 8). we observe the same behaviors of worm propagation on this tenfold larger network, which shows that the behavior of worm propagation doesn't change when the network scale changes.

C. Effect of email checking time distribution

Here we study how different distributions of T_i affect an email worm's propagation. We study three distributions of T_i: exponential distribution, a 3rd-order Erlang distribution, and a constant email checking time interval for each user. For comparison, in all three cases $E[T_i]$ follows the same Gaussian distribution $T \sim N(40, 20^2)$.

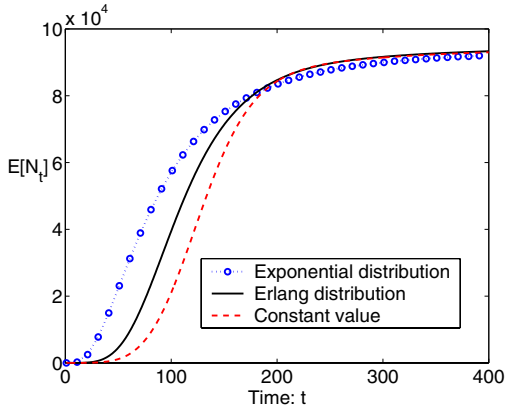
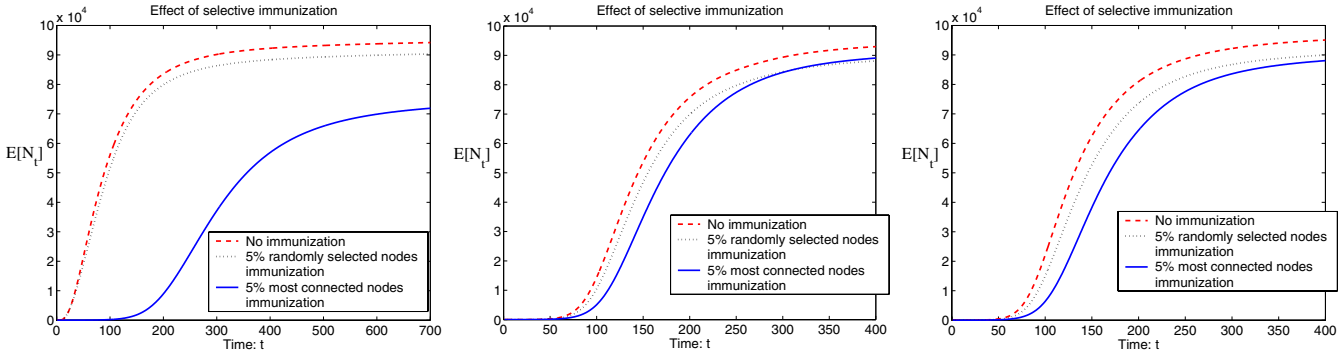


Fig. 5. Effect of the distributions of email checking time interval (on a power law email network)

Fig. 5 shows the average number of infected users, $E[N_t]$, on a power law email network. We also conduct this experiment on a small world network and a random graph network — both networks give the similar worm propagation patterns (propagation speed on these two networks is slower than on a power law network as illustrated in Fig. 3).

Given the same mean value, the exponential distribution is more *stochastically variable* [20] than the kth-order Erlang distribution where $k > 1$. Both of them are more stochastically variable than the constant value. Fig. 5 shows that an email worm propagates faster as the email checking time interval T_i becomes more "variable". We have proven this conclusion for a simplified worm propagation model and presented the proof



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

Fig. 6. Effect of selective immunization on email worm propagation

in our technical report [21]. Intuitively, it is due to so-called *snowball* effect: before worm copies in the system with less variable checking time give birth to the next generation — infecting some new users — worm copies in another system with more variable checking time have already given birth to *several* generations, although each generation's population is relatively small.

D. Other simulation studies

We have conducted many other simulation experiments, including: (1) studying the impact of the power law exponent α of a power law email network; (2) studying the impact of the network average node degree and the degree of initially infected users. For the results and details of these experiments, please see our technical report [21].

VI. Immunization Defense for Email Worms

In this section, we consider immunization defense against email worm attacks. "Immunization" means that before an email worm starts to propagate, some nodes have already been immunized such that they cannot be infected by the worm. If some email users are well educated and they never open suspicious email attachments, they can be treated as immunized nodes in the email network.

A. Effect of selective immunization

We simulate worm propagation under two different immunization defense methods: in the first case we randomly choose 5% nodes in the email network to immunize, while in the second case we choose 5% most connected nodes to immunize. We plot $E[N_t]$ as a function of time t for these two immunization methods in Fig. 6 (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 $E[N_t]$ for the original case where there is no immunization.

We observe from Fig. 6 that selective immunization is a very effective defense on a power law email network while it has little effect for a small world or a random graph network. This result is consistent with the conclusions in [14]: 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 worm propagation speed.

B. Selective percolation and worm prevention

From an email worm's point of view, the connectivity of a partly immunized email network is a "percolation" problem. The authors in [13] studied simple percolation by removing some nodes from networks *uniformly* — their approaches cannot be used here to study the selective immunization defense.

We introduce the corresponding concept *"selective percolation"*. A selective percolation value p means to remove the top p percent of the most connected nodes from a network, $0 < p < 1$. Let $C(p)$ denote the *connection ratio*, the fraction of how many remained nodes still connected after removing the top p percent of the most connected nodes from the network. Let $L(p)$ denote the *remained link ratio*, the fraction of links remained after removal the top p percent of the most connected nodes from the network.

$$\begin{cases} C(p) & = c_p/(|V| - |V|p) \\ L(p) & = (|E| - e_p)/|E| \end{cases} \quad 0 < p < 1 \tag{3}$$

where e_p is the number of removed edges and c_p is the number of nodes in the largest cluster of the remaining network.

We generate 100 networks for each type of the three topologies — power law, small world and random graph topologies. Each network has the same average degree 8 and $100,000$ nodes. For every selective percolation value p we derive $C(p)$ and $L(p)$ by averaging those 100 numbers derived by equation (3) from each network instance. The results are shown in Fig. 7. All three topologies have their own selective percolation thresholds: if the fraction of selectively immunized users exceeds the threshold, an email network will be broken into separated fragments and no email worm outbreak will occur. The selective percolation threshold of a power law topology is much smaller than the threshold for the other two topologies, which is consistent with the experiment shown in Fig. 6. When we immunize the top 5% of most connected

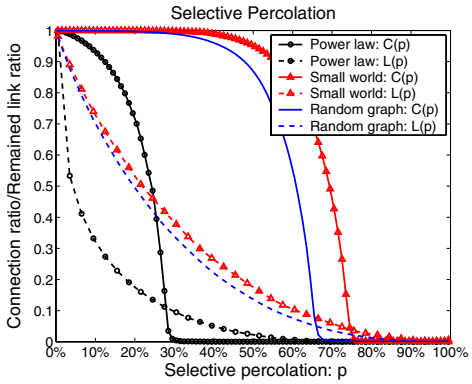


Fig. 7. Selective percolation on power law, small world, and random graph topologies

nodes in a power law network, Fig. 7 shows that although 97.5% of remained nodes in the network are still connected, 55.5% of the network edges have been removed. Thus an email worm has fewer and longer paths to reach and infect nodes in the remaining network.

VII. Conclusions

We present an email worm model that accounts for the behaviors of email users considering email checking frequency and the probability of opening an email attachment. Email worms spread over a logical network defined by email address relationship. Our observations suggest that the node degrees in an email network are heavy-tailed distributed. We compare email worm propagation on three topologies: power law, small world and random graph topologies; and then study how the topology affects immunization defense. The impact of the power law topology on the spread of email worms is mixed: email 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 than on the other two.

There are still many works to do on email worm modeling and defense. First, in this paper we have mainly used simulation to study email worm propagation. The next step is to derive mathematical model like the models for scan-based worms [5][6][7]. Second, we have only considered static immunization defense in this paper — we assume that before the break out of an email worm, part of users and computers have already been immunized of the worm and no more users or computers 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 the "dynamic immunization". Third, although we have considered the impact of email lists on the topology of Internet email network, instead of an undirected graph used in this paper, a directed graph is preferred in order to more accurately capture some one-way email address relationship (i.e., user A has the email address of user B, but user B does not have the address of user A). Finally, the privacy issue of email makes it hard for us to build up a global cooperation and defense system in the near future. Therefore, for email worm defense, in the short term we plan to study how to protect email users in a local network, such as an enterprise network, from email worm attacks.

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 F30602-00-0554, and by NSF under Grant EIA-0080119, ANI9980552, ANI-0208116. 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] CERT, "CERT/CC advisories," http://www.cert.org/advisories/.
[2] J. Kephart, D. M. Chess, and S. White, "Computers and epidemiology," *IEEE Spectrum*, vol. 30, no. 5, May 1993.
[3] 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.
[4] ——, "Measuring and modeling computer virus prevalence," in *Proceedings of IEEE Symposium on Security and Privacy*, 1993.
[5] Z. Chen, L. Gao, and K. Kwiat, "Modeling the spread of active worms," in *Proceedings of the IEEE INFOCOM*, March 2003.
[6] S. Staniford, V. Paxson, and N.Weaver, "How to own the internet in your spare time," in *Proceedings of Usenix Security Symposium*, August 2002.
[7] C. 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.
[8] C. 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.
[9] 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.
[10] M. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Phys. Rev. E.*, vol. 66, no. 035101, 2002.
[11] L. Briesemeister, P. Lincoln, and P. Porras, "Epidemic profiles and defense of scale-free networks," in *Proceedings of ACM CCS Workshop on Rapid Malcode (WORM'03)*, October 2003.
[12] R. P. Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. E.*, vol. 86, no. 14, April 2001.
[13] M. Newman, S. Strogatz, and D. Watts, "Random graphs with arbitrary degree distributions and their applications," *Phys. Rev. E.*, vol. 64, no. 026118, 2001.
[14] R. Albert, H. Jeong, and A. Barabasi, "Error and attack tolerance of complex networks," *Nature*, vol. 406, 2000.
[15] "*Yahoo!* groups," http://groups.yahoo.com.
[16] T. Bu and D. Towsley, "On distingishing between internet power law topology generators," in *Proceedings of the IEEE INFOCOM*, June 2002.
[17] D. Watts and S. Strogatz, "Collective dynamic of small-world networks," *Nature*, vol. 393, 1998.
[18] M. Newman, I. Jensen, and R. Ziff, "Percolation and epidemics in a two-dimensional small world," *Phys. Rev. E.*, vol. 65, no. 021904, 2002.
[19] 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.
[20] S. M. Ross, *Stochastic Processes*. John Wiley & Sons, Inc., 1996.
[21] C. C. Zou, D. Towsley, and W. Gong, "Email virus propagation modeling and analysis," Umass ECE Dept., Tech. Rep. TR-03-CSE-04, May 2003.