Spatial-Temporal Modeling of Malware Propagation in Networks

Zesheng Chen, *Student Member, IEEE,* and Chuanyi Ji, *Member, IEEE*

Abstract—Network security is an important task of network management. One threat to network security is malware (malicious software) propagation. One type of malware is called topological scanning that spreads based on topology information. The focus of this work is on modeling the spread of topological malwares, which is important for understanding their potential damages, and for developing countermeasures to protect the network infrastructure. Our model is motivated by probabilistic graphs, which have been widely investigated in machine learning. We first use a graphical representation to abstract the propagation of malwares that employ different scanning methods. We then use a spatial-temporal random process to describe the statistical dependence of malware propagation in arbitrary topologies. As the spatial dependence is particularly difficult to characterize, the problem becomes how to use simple (i.e., biased) models to approximate the spatially dependent process. In particular, we propose the independent model and the Markov model as simple approximations. We conduct both theoretical analysis and extensive simulations on large networks using both real measurements and synthesized topologies to test the performance of the proposed models. Our results show that the independent model can capture temporal dependence and detailed topology information and, thus, outperforms the previous models, whereas the Markov model incorporates a certain spatial dependence and, thus, achieves a greater accuracy in characterizing both transient and equilibrium behaviors of malware propagation.

Index Terms—graphical models, malware, modeling, security, stochastic processes.

I. INTRODUCTION

PROTECTING a computer system from malicious attacks is a key challenge to network security and management. One such attack is due to malware propagation. The frequency and virulence of malware outbreaks have increased dramatically in the last few years, posing a significant threat to network infrastructure.

Malware is *malicious software* in short, which is designed specially for either damaging or disrupting a computer system. This terminology is used to cover an entire gamut of hostile softwares including viruses, Trojan Horses, and network worms [1]. There are mainly two types of malwares categorized by how they spread. Active network worms such as Sapphire and Morris exploit self-propagating malicious code [2], whereas viruses such as Melissa and Concept need human interactions to spread [3]. Spreading can take place rapidly, resulting in

potential network damages and service disruptions. Hence, an important step toward preventing such catastrophic events is to study the dynamic behavior of malware spreading.

Recent investigations of malware propagation focus mostly on modeling the spread of malwares employing random scanning scheme [2], [4], [5]. Random scanning selects targets to infect randomly. Malwares, however, can use other scanning methods. For example, Morris worm exploits topological scanning, which examines local configuration files to find potential neighbors [6]. Although only a few topological malwares are known, topological scanning is a potential thread to the network routing infrastructure, world wide web (WWW) networks, and peer-to-peer systems [7], where topologies play an important role for malware propagation [8]. Only a handful work, however, has been done on topological-scanning malwares. For instance, contact process is used to analyze the ease of propagation on different topologies [9]. The difficulty lies in characterizing the impact of topology and the interactions among nodes in both space and time [10]. Such interactions result in a complex spatial-temporal dependence, which is especially hard to model.

The goal of this work is to develop a modeling framework and mathematical models that can characterize the spread of malwares employing different scanning strategies and the impact of the underlying topology on malware propagation. To this purpose, we first abstract the problem of malware propagation using a graphical representation so that different scanning methods can be mapped to the corresponding topologies and parameters. With the help of the graphical representation, we then formulate malware propagation through a spatial-temporal random process based on the interactions among nodes. We take advantage of a discrete-time model and detailed topology information to describe the spatial and temporal statistical dependencies of malware propagation in arbitrary networks.

As the temporal dependence can be naturally modeled as Markov, spatial dependence requires calculations with a multivariate probability distribution. When the number of random variables is large, an exact solution of spatial dependence is computationally too expensive to obtain. The problem then becomes how to approximate the spatial dependence using a simple, i.e., biased model, in a general setting of machine learning. In particular, spatial approximation is studied in light of the mean-field approximation [11]. Mean-field approximation is widely studied in machine learning [11] but usually for static networks where time is not involved. Exact mean-field solutions for dynamic networks are complex. Hence, we consider in this work simple approximations. The simplest approximation assumes spatial independence, which is asserted in our *independent model*. The spatial independence assumption

factorizes an exact joint probability distribution into a form that only depends on one-node marginal probabilities. Although the independent model ignores the spatial dependence, it captures temporal dependence and detailed topology information. Simulation results show that the independent model performs better than the previous models in characterizing the transient behavior of malware propagation. A test on spatial correlation though indicates a strong spatial dependence among nodes. We therefore present the *Markov model* that incorporates the simplest spatial dependence as the *conditional* independence, motivated by the Bethe approximation used in graphical models [12]. The spatial Markov assumption factorizes an exact joint probability distribution into a form that only depends on one-node and two-node marginal probabilities. We have conducted both theoretical analysis and extensive simulations on real and synthesized topologies of large networks. Our results demonstrate that the Markov model equipped with the simple spatial dependence can achieve a greater accuracy than the independent model, especially in the sparse graphs. We then use a *relative entropy* to illustrate a performance gap between the Markov model and the reality, suggesting directions for further improvements.

We apply our proposed models to describe the *final size of infection*, which corresponds to the equilibrium solution and characterizes the potential damage of malware propagation. Simulation results show that the Markov model can characterize the final size of infection no matter whether the underlying network is a homogeneous network or a complex network.

Several approaches have been proposed to model and simulate malware spreading in different topologies. Kephart and White presented the Epidemiological model, which is suitable to analyze virus spreading in random graphs [10]. This work points out the difficulty in applying the Epidemiological model to study arbitrary topologies. Garetto *et al.* analyzed malware spreading in small-world topologies using a variation of the influence model, where the influence of neighbors is constrained to take a multilinear form [3]. Boguñá *et al.* studied epidemic spread in complex networks [13], and Wang *et al.* proposed a model for virus propagation in arbitrary topologies [14]. Both [13] and [14] are proposed to obtain the epidemic threshold of virus infection. Zou *et al.* and Wang *et al.* investigated the effect of topology and immunization on the propagation of computer virus through simulation [8], [15]. Ganesh *et al.* modeled the spread of an epidemic as a contact process [16] to study what makes an epidemic either weak or potent [9]. The model assumes that a vulnerable node can be infected by its infected neighbors at a rate that is proportional to the number of infected neighbors. Some recent investigations focus on random-scanning worms. Zou *et al.* modeled the spread of Code Red, taking into consideration of the human countermeasures and the worm's impact on Internet infrastructure [4]. Chen *et al.* studied the propagation of active worms employing random scanning and extended the proposed modeling method to investigate the spread of localized-scanning worms [5]. Moore *et al.* applied the Epidemiological model to investigate the requirements for containing the self-propagation worm with random target selection [2]. The prior work, however, has not incorporated the spatial dependence on malware propagation in networks. This

motivates the development of mathematical models to capture the spatial dependence, and the use of spatial models to characterize both transient and equilibrium behaviors of malware propagation with different scanning methods in arbitrary topologies. Furthermore, based on the models proposed in this paper we study the significance of the spatial dependence in determining epidemic thresholds and the speed of propagation [17].

The rest of this paper is organized as follows. In Section II, we provide a problem formulation of malware propagation. In Section III, we model the spread of malwares accurately through a spatial-temporal random process. To approximate the spatial dependence, we present the independent model and the Markov model in Sections IV and V, respectively. In Section VI, we apply our proposed models to estimate the final size of infection. We conclude this paper in Section VII with a brief summary and an outline of future work.

II. MALWARE PROPAGATION IN NETWORKS

In this section, we first introduce malware propagation briefly. We then abstract the problem using a susceptible → infected → susceptible (SIS) model and a graphical representation. Finally, we model different scanning mechanisms using graphical representations.

A. Malware Propagation

A computer is called *infected* if a malware is present there, and *susceptible* if it could be infected by the intrusion of the malware. If a malware cannot exist on the computer, we call this computer *insusceptible* to the malware. An infected computer is *cured* if it removes the copy of the malware and recovers to be susceptible. *Final size of infection* is defined as the number of initially susceptible computers that ultimately become infected in a network. The widespread occurrence of a malware is referred to as an *epidemic* [18]. Malware propagation is a procedure that the malware infects as many computers as possible through network connections. Those connections can be logical as to be described in the following.

A malware can propagate in many ways. For example, when a worm is released into the Internet, it scans many machines among its neighbors in an attempt to find a susceptible machine. When a vulnerable host is found, the worm sends out a probe to infect the target. If successful, a copy of this worm is transferred to the new host, which then begins to run the worm code and tries to infect other targets. Morris worm is a typical self-propagation malware and moves from node to node using only its own and the infected node's local information [6]. Specifically, Morris worm retrieves the neighbor list from the local Unix files */ect/hosts.equiv* and */.rhosts* and in individual users' *.forward* and *.rhosts* files. Another topological worm is a secure shell (SSH) worm, which locates new targets by searching its current host for the names and addresses of other hosts that are likely to be susceptible to infection [19]. E-mail virus is an other example of topological malwares. When an e-mail user receives an e-mail message and opens the attachment containing a virus program, the virus infects the user's machine and uses the recipient's address book to send copies of itself to other e-mail

addresses. The addresses in address book represent the neighborhood relationship. A *birth rate* (or an *infection rate*) is introduced to denote the rate at which an infected computer can infect a susceptible neighbor. The birth rate is affected by many factors. For example, for worms, the factors include the number of computer's susceptible neighbors, payload size of the malware copy, exploited computer vulnerability, and network congestion. For e-mail virus, the factors include e-mail checking frequency, user vigilance in opening an e-mail attachment, and mailbox configuration. Some malwares may have a large birth rate to flood the network as quickly as possible, whereas other malwares spread slowly and surreptitiously to evade detection and, thus, have a small birth rate.

An infected computer might die for encountering an unexpected resource limit on the computer. Moreover, during the spreading of a malware, some infected computers may stop functioning properly, forcing the users to reboot these machines or kill some of the processes exploited by the malware. These computers are then cured, but subject to further infection. A *death rate* (or a *cure rate*) is introduced to denote the rate at which an infected computer becomes susceptible. The death rate is affected by many factors, such as resources on the computers, user alertness, the ability of a malware to disguise, and the performance of intrusion detection system (IDS).

Combining infection and recovery, we have one of the simplest epidemiological models, the SIS model, which is widely used in the epidemiological research [18]. Such a model neglects the details of infection inside a single computer, abstracts the malware transmission and removal as probabilities per unit time in the form of birth rate and death rate, and considers a computer to be in one of the two possible discrete status, *infected* or *susceptible*. Although simple, the SIS model can capture key characteristics of malware spreading dynamics. The *susceptible → infected* (SI) model further ignores recovery and is regarded as a special case of the SIS model.

The SIS model assumes that an infected computer cannot be reinfected. The model also assumes that users do not become more vigilant after experiencing a malware infection. Therefore, the birth rate and the death rate do not change with time. Moreover, in this paper we ignore patching, which is usually employed to repair security holes at the computers. This is because the spreading of malware can be much faster compared with traditional patching techniques that need human intervention, and a patch may not be available when some malware attacks unknown vulnerabilities. Nevertheless, our proposed models can be easily extended to take patching into consideration.

B. Graphical Representation

A *malware network* consists of all nodes in a network that are either infected or susceptible. A malware network can be constructed by removing insusceptible nodes and the edges associated with these nodes in the original network. Hence, a malware network is an abstraction of vulnerable nodes, which can be either end-hosts, routers, and servers, or e-mail addresses.

We use a *directed graph* $G(V, E)$ to represent a malware network, where V is the set of nodes and E is the set of edges. As defined in Section II-A, each node has two status, susceptible or infected, as illustrated in Fig. 1. Each edge (j, i) is associated
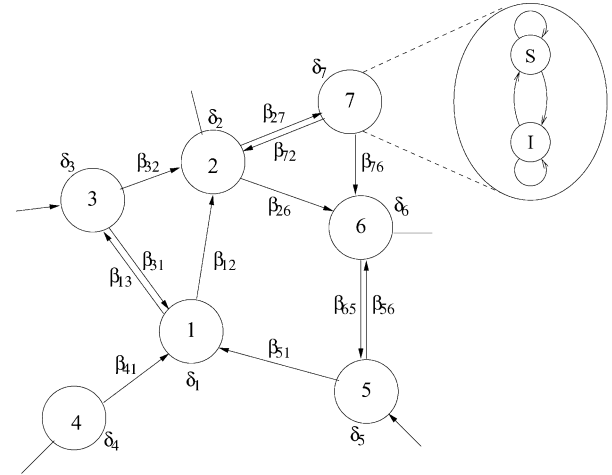

Fig. 1. Directed graph $(\text{S} = \text{Susceptible}, \text{I} = \text{Infected})$.

with β_{ji}, the *birth rate* at which an infected node j can infect a susceptible neighbor i. Similarly, each node i is associated with δ_i, the *death rate* at which an infected node i becomes susceptible. A *neighborhood* of node i, denoted by N_i, is a subset of V such that every node j in this subset has an edge from node j to node i, i.e., $N_i = \{j \,|\, (j, i) \in E\}$. Fig. 1 shows an example of directed graph wherein the neighborhood of node 1 is given as $N_1 = \{3, 4, 5\}$.

In this paper, we consider two widely used types of networks in the research of epidemic modeling: *homogeneous networks* and *complex networks* [13]. In a homogeneous network, each node has roughly the same nodal degree. A fully-connected topology, a standard hypercubic lattice, and an Erdös–Rényi (ER) random network are three typical examples of homogeneous networks [20]. In a complex network, the nodal degree complies to a particular distribution. A widely-studied representative complex network has a power-law topology, where the nodal degree distribution is characterized as $P(k) \sim k^{-r}$ with $P(k)$ being the probability that a node has a degree of k [21]. It has been shown that AS-level Internet topology, WWW networks, and some overlay topologies of peer-to-peer systems can be described by power-law characteristics [22], [23]. Moreover, the e-mail groups and network exhibit the power-law distribution, which is observed in [8] and [24]. Hence, a malware network with a power-law topology can be used to study potential malware propagation on those networks.

C. Scanning Methods

A malware spreads by employing distinct scanning mechanisms such as random, localized, and topological scanning [7]. Although the nature of scanning methods is different, they can be modeled using the same graphical representation.

Random scanning is used by some well-known malwares such as Code Red v2 and Sapphire worms. A malware that employs random scanning selects target IP addresses at random. If each IP address is visualized as a network node, random scanning results in a fully-connected topology illustrated in Fig. 2(a), where the birth rate (β) is identical for every edge.

Localized scanning is used by Code Red II and Nimda worms. Instead of selecting targets randomly, a malware preferentially
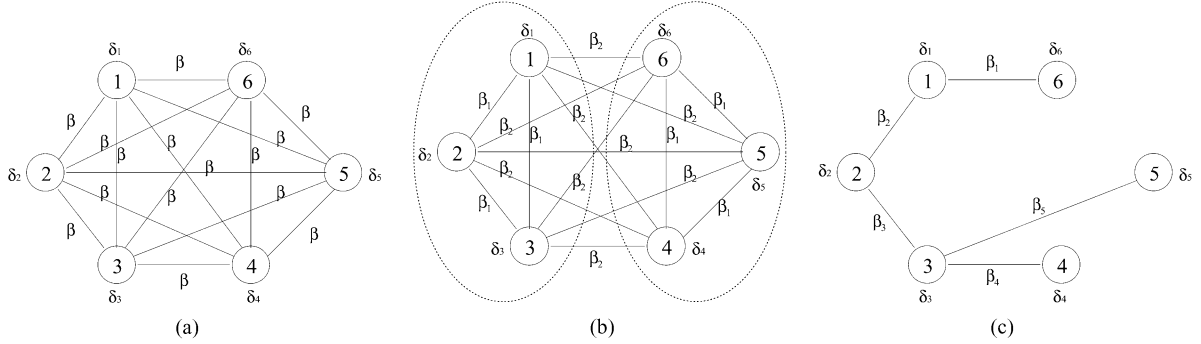
Fig. 2.　Graphical representations of scanning methods. (a) Random Scanning. (b) Localized Scanning. (c) Topological Scanning.

scans for hosts in the "local" address space. Such a scanning scheme results in a fully-connected topology such as the one illustrated in Fig. 2(b), where nodes within a group (e.g., IP addresses with the same first two octets) infect one another with the same birth rate (β_1), whereas nodes in different groups infect one another with a different birth rate (β_2).

Topological scanning is used by e-mail viruses and Morris/SSH worms. The malware relies on the information contained in the victim machine in order to locate new targets. The information may include routing tables, e-mail addresses, a list of peers, and uniform resource locations (URLs). Topological scanning scheme can result in an arbitrary topology such as an undirected power-law topology illustrated in Fig. 2(c), where β_i's and δ_i's $(i = 1, 2, \ldots, 5)$ represent different birth rates and death rates.

Although only a few topological worms are known, topological scanning is worth investigating for the following reasons. First, the network routing infrastructure, World Wide Web (WWW) networks, and peer-to-peer systems are vulnerable to topological scanning. For example, a malware attacking a website could look for neighboring websites in its URLs and use these websites as targets. Second, when IPv4 is upgraded to IPv6, the address space will be much sparser. It would be difficult for either random-scanning or localized-scanning worms to find a target in the IPv6 address space. Therefore, topological scanning may be preferred by attackers. Finally, models of such malware-propagation would provide insights for developments of countermeasures, which are lacking for topological worms.

III. SPATIAL-TEMPORAL MODEL

The problem of modeling malware propagation in networks can be stated as follows: Given a malware network topology, values of β_{ji}'s and δ_i's, and an initial infection node, what is the expected number of infected nodes at time t? To approach this problem, we formulate malware propagation through a spatial-temporal random process based on local interactions of nodes in networks.

Let $X_i(t)$ denote the status of a network node i at time t, where t represents discrete time

$$X_i(t) = \begin{cases} 1, & \text{if node } i \text{ is infected at time } t; \\ 0, & \text{if node } i \text{ is susceptible at time } t. \end{cases}$$
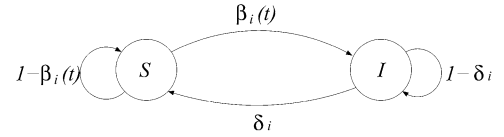

Fig. 3.　State diagram of a node i.

As node i can be infected only by its neighbors, $X_i(t)$ is statistically dependent on $X_i(t - 1)$ and the status of its neighbors. Since the status of a neighbor also depends on its own neighbors, conceptually, the status of all nodes is statistically dependent in space and time. Let vector $\mathbf{X}(t)$ denote the status of all nodes at time t, i.e., $\mathbf{X}(t) = \{X_1(t), X_2(t), \ldots, X_M(t)\}$, where M represents the total number of nodes in the network. $\mathbf{X}(t)$ is then a *spatial-temporal process*.

If node i is susceptible, it can be compromised by any of its infected neighbors, e.g., node j, with a birth rate β_{ji}. Therefore, given the status of the neighbors of node i, at the next time step the susceptible node i can get infected with probability $\beta_i(t) = 1 - \prod_{j \in N_i}(1 - \beta_{ji})^{x_j(t)}$, where $x_j(t)$ is the realization of the status of node j at time t and $x_j(t) = 0$ or 1. Otherwise, node i is infected and has a death rate δ_i to recover at the next time step. This procedure can be expressed by a Markov chain in Fig. 3. Therefore, the temporal dependence of node i can be shown as

$$P(X_i(t+1) = 0 \mid X_i(t) = 1) = \delta_i \tag{1}$$

$$P(X_i(t+1) = 1 \mid X_i(t) = 0, \mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t)) = \beta_i(t) \tag{2}$$

where vector $\mathbf{X}_{N_i}(t)$ is used to denote the status of all neighbors of node i at time t and vector $\mathbf{x}_{N_i}(t)$ is the realization of $\mathbf{X}_{N_i}(t)$, i.e., $\mathbf{X}_{N_i}(t) = \{X_j(t), j \in N_i\}$ and $\mathbf{x}_{N_i}(t) = \{x_j(t), j \in N_i\}$. If for $\forall j, \beta_{ji} \ll 1$, the birth rate (β) is identical for every edge, and the death rate (δ) is identical for every node, then $\beta_i(t) \approx \sum_{j \in N_i} \beta_{ji} x_j(t) = \beta \sum_{j \in N_i} x_j(t)$ and $\delta_i = \delta$, which are assumptions used in contact process [9].

The probability that node i recovers from infected to susceptible status at time $t+1$ is expressed by $R_i(t) = P(X_i(t+1) = 0, X_i(t) = 1)$. Thus, (1) leads to

$$R_i(t) = \delta_i P(X_i(t) = 1). \tag{3}$$

Given node i is susceptible at time t, the probability that node i remains susceptible at the next time step can be defined as
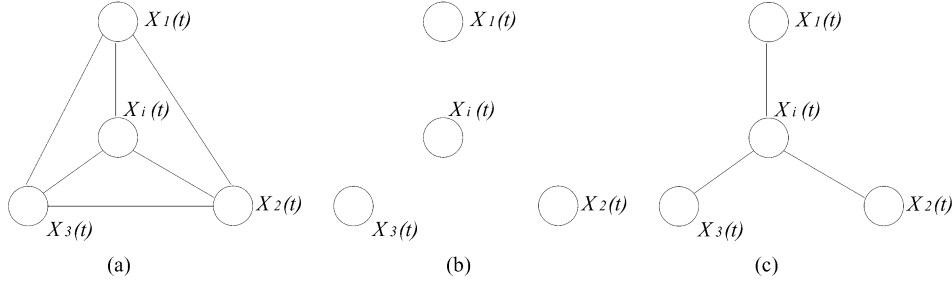
Fig. 4. Dependency graph. (a) Joint probability. (b) Independent model. (c) Markov model.

$S_i(t) = P(X_i(t+1) = 0 \mid X_i(t) = 0)$. From the local dependence of (2), we have

$$S_i(t) = \sum_{\mathbf{x}_{N_i}(t)} [P(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = 0)(1 - \beta_i(t))]. \tag{4}$$

Therefore, the definitions of $R_i(t)$ and $S_i(t)$ yield that for $\forall i \in \{1, 2, \ldots, M\}$

$$P(X_i(t+1) = 1) = 1 - R_i(t) - P(X_i(t) = 0)S_i(t). \tag{5}$$

Combined with (3) and (4), (5) provides a recursive relationship between $X_i(t+1)$ and $X_i(t), X_j(t)$ for $j \in N_i$, and gives a formal stochastic model. This model characterizes explicitly the spatial and temporal statistical dependencies. In particular, the joint probability $P(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = 0)$ characterizes the spatial dependence due to network topology and nodal interactions. The transition probabilities, $\beta_i(t)$ and δ_i, characterize the temporal evolution due to infection and recovery. Together, they describe the spatial-temporal process of malware propagation in networks. The expected number of infected nodes at time $t, n(t)$, can be easily computed from $P(X_i(t) = 1)$, i.e., $n(t) = E[\sum_{i=1}^{M} X_i(t)] = \sum_{i=1}^{M} P(X_i(t) = 1)$.

Although in principle, (5) can be used to study the behavior of malware propagation, it is challenging to model the spatial dependence. This is because the joint probability $P(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = 0)$ is computationally too expensive to obtain, especially when the size of neighborhood is large. For example, if node i has k neighbors, the total number of status needed to describe this joint probability is $O(2^k)$. Therefore, we introduce approximations for spatial dependence in Sections IV and V. An example of the dependency graph of the joint probability $P(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = 0)$ is shown in Fig. 4(a), where node i has three neighbors (i.e., nodes 1, 2, 3) and all nodes are dependent on each other.

Remark: It is noted that the mean-field methods are used to reduce the computational complexity involved in typical calculations with multivariate probability distributions when the number of random variable is large [11]. The mean-field methods, however, are difficult to be employed directly to our problem. A typical context for a mean-field approximation is to compute the marginal probabilities and expectations of a given joint distribution. Our problem, however, requires to obtain an accurate joint distribution based on the marginal probabilities. Moreover, in many cases the mean-field methods use a set

of recursions to find a stationary solution of a corresponding optimization problem in space [25], whereas the topological malware propagation involves both space and time. Although the mean-field methods are currently difficult to be grafted directly to malware propagation problem, the spirit of the mean-field theory motivates our approach for approximating spatial dependence. For example, the naive mean field assumes that each random variable acts independently and, thus, approximates the true distribution through a complete factorization [11]. This idea is adopted by our independent model.

IV. INDEPENDENT MODEL

The simplest spatial approximation is to assume independence, resulting in our independent model.

A. Model

In the independent model, we assume that the status of all nodes at time t $(t = 0, 1, 2, \ldots)$ is spatially independent. That is

$$P(\mathbf{X}(t) = \mathbf{x}(t)) = \prod_{i=1}^{M} P(X_i(t) = x_i(t)) \tag{6}$$

where $\mathbf{x}(t)$ is the realization of $\mathbf{X}(t)$, i.e., $\mathbf{x}(t) = \{x_1(t), x_2(t), \ldots, x_M(t)\}$. With the spatial independence assumption, the dependency graph shown in Fig. 4(a) is reduced to the graph shown in Fig. 4(b), which is a graph with no edges. Thus, the joint probability $P(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = 0)$ can be factorized into a form that only depends on one-node marginal probabilities. This kind of full factorization is also employed in the naive mean-field approach, where each factor is obtained through the mean-field equations [25].

Theorem 1: (Independent Model): If the status of all nodes at the same time step is spatially independent, the state evolution of node i from (5) satisfies

$$P(X_i(t+1) = 1) = 1 - R_i(t) - P(X_i(t) = 0)S_i^{\text{ind}}(t) \tag{7}$$

where

$$S_i^{\text{ind}}(t) = \prod_{j \in N_i} [1 - \beta_{ji} P(X_j(t) = 1)]. \tag{8}$$

Proof: Since the status of all nodes at time t is spatially independent, it is true that

$$P(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = 0) = \prod_{j \in N_i} P(X_j(t) = x_j(t)).$$

With this assumption, it follows from (4) that

$$
\begin{aligned}
S_i^{\mathrm{ind}}&(t) \\
&= \sum_{\mathbf{x}_{N_i}(t)} \prod_{j \in N_i} \left[P(X_j(t) = x_j(t))(1 - \beta_{ji})^{x_j(t)} \right] \\
&= \prod_{j \in N_i} \sum_{x_j(t)} \left[P(X_j(t) = x_j(t))(1 - \beta_{ji})^{x_j(t)} \right] \\
&= \prod_{j \in N_i} [P(X_j(t) = 0) + P(X_j(t) = 1)(1 - \beta_{ji})] \\
&= \prod_{j \in N_i} [1 - \beta_{ji} P(X_j(t) = 1)]
\end{aligned}
$$

where the exchange of the summation and product signs is because: Set $f(x_j(t)) = P(X_j(t) = x_j(t))(1 - \beta_{ji})^{x_j(t)}$ and $j = 1, 2, \ldots, K$, where K is the number of neighborhood of node i; thus

$$
\begin{aligned}
\sum_{\mathbf{x}_{N_i}(t)} &\prod_{j=1}^{K} f(x_j(t)) \\
&= \sum_{x_1(t)} \sum_{x_2(t)} \cdots \sum_{x_K(t)} f(x_1(t)) f(x_2(t)) \cdots f(x_K(t)) \\
&= \left(\sum_{x_1(t)} f(x_1(t)) \right) \left(\sum_{x_2(t)} f(x_2(t)) \right) \cdots \left(\sum_{x_K(t)} f(x_K(t)) \right) \\
&= \prod_{j=1}^{K} \sum_{x_j(t)} f(x_j(t)).
\end{aligned}
$$

∎

Such an independent model is intuitive. That is, node j, one of the neighbors of node i, can infect node i with probability $\beta_{ji} P(X_j(t) = 1)$. Thus, the probability that node i cannot be infected by its neighbors at time $t + 1$ is $\prod_{j \in N_i} [1 - \beta_{ji} P(X_j(t) = 1)]$, according to the independence assumption. Although ignoring the spatial dependence, the independent model maintains temporal dependence and detailed topology information. Moreover, if node i has k neighbors, the total number of status needed to describe the joint probability $P(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = 0)$ is reduced from $O(2^k)$ to $O(k)$.

Remark: It should be noted that the spatial independence assumption is implicitly used in [14]. The independent model given here, however, is different from the model proposed in [14] in the following aspects. First, our proposed model is derived from the accurate spatial-temporal process and the explicit approximation on spatial dependence. Second, our independent model only allows one event (i.e., susceptible → infected, or infected → susceptible) in one single discrete time step, whereas the model in [14] grants concurrence of infection and recovery. Finally, our model focuses on the transient behavior of malware propagation, whereas the model in [14] emphasizes on the steady-state solution and the epidemic threshold.

B. Performance

How good is the performance of this independent model? We compare the performance of the independent model with that of some well-known models and the simulation results in both homogeneous and complex networks. For simplicity, we consider the special cases where the birth rate (β) is identical for every edge and the death rate (δ) is identical for every node. Such an assumption is used in all previous models. Simulation provides a benchmark for assessing the performance of models. For the simulation, we track each node's status (infected or susceptible) in discrete time. Each simulation has 100 individual runs and is averaged over the cases that the malware survives.[1]

1) Homogeneous Networks: In homogeneous networks, the standard Epidemiological model uses a nonlinear differential equation to measure the malware population dynamics [10]

$$
\frac{dn(t)}{dt} = \beta \bar{k} n(t) \left[1 - \frac{n(t)}{M} \right] - \delta n(t) \tag{9}
$$

where $\bar{k}$ is the average nodal degree. The solution to the previous equation is

$$
n(t) = \frac{n(0)M(1 - \rho)}{n(0) + [M(1 - \rho) - n(0)]e^{-(\beta' - \delta)t}} \tag{10}
$$

where $\beta' = \beta \bar{k}$ and $\rho = (\delta / \beta')$. Another model used in homogeneous networks is the analytical active worm propagation (AAWP) model, which uses a discrete time [5]

$$
n(t + 1) = (1 - \delta)n(t) + [M - n(t)] \left[1 - \left(1 - \frac{1}{M} \right)^{sn(t)} \right] \tag{11}
$$

where the scanning rate $s = \beta \bar{k}$ and the patching rate is ignored. Both the Epidemiological model and the AAWP model have been used to model the spread of active worms that employ random scanning, and shown to perform well if the underlying graph is an ER random graph with a large $\bar{k}$ or a fully-connected topology [5], [10].

Fig. 5 shows the evolution of the average number of infected nodes for the Epidemiological model, the AAWP model, the independent model, and the simulation on a four-neighbor two-dimensional lattice with 10 000 nodes, $\beta = 0.1$, and $\delta = 0.1$. The two-dimensional (2-D)-lattice considered in this paper is wrapped around in both dimensions to form a torus. It is observed that all three models over-predict the growth of infected nodes. The independent model, however, describes the transient behavior of malware propagation better than the other two models.

2) Complex Networks: Boguñá *et al.* classify the complex networks into two types: uncorrelated and correlated complex networks, and present epidemic models for each type [13]. We name these two models as uncorrelated complex network (UCN) model and correlated complex network (CCN) model. In these models, the number of infected nodes with a degree of k at time t, $n_k(t)$, can be described by the following [13]:

$$
\frac{dn_k(t)}{dt} = \beta k \left[1 - \frac{n_k(t)}{M_k} \right] \Theta_k(t) - \delta n_k(t) \tag{12}
$$

---

[1]We focus on the transient behavior of epidemic malware and ignore the cases that the malware dies out.
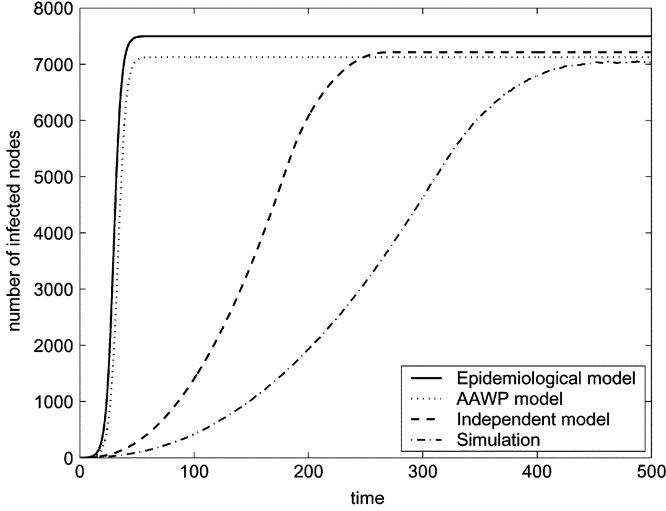
Fig. 5. Malware propagation in a 2-D-lattice with 10 000 nodes, $\beta = 0.1$, and $\delta = 0.1$.
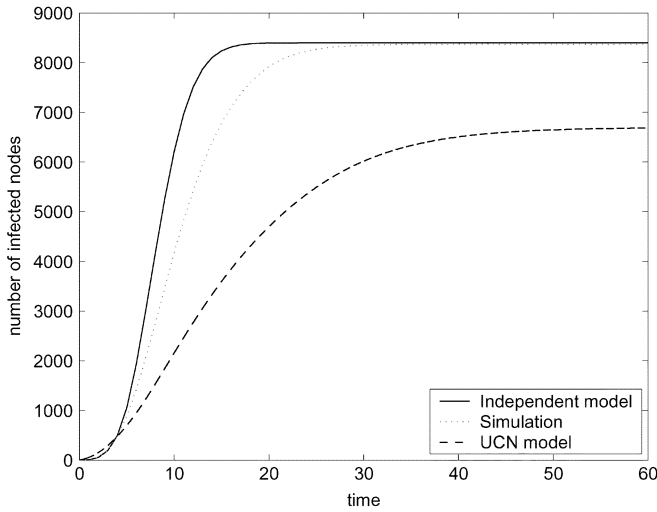

Fig. 6. Malware propagation in a BA network with 10 000 nodes, $\bar{k} = 1.9998, \beta = 0.5$, and $\delta = 0.1$.

where M_k is the total number of nodes with a degree of k and $\sum_k M_k = M$. In the UCN model, $\Theta_k(t)$ is independent of k and defined as

$$\Theta_k(t) = \frac{1}{\bar{k}} \sum_{k'} k' P(k') n_{k'}(t) \qquad (13)$$

whereas in the CCN model, the effect of the degree k is considered and the expression for $\Theta_k(t)$ is

$$\Theta_k(t) = \sum_{k'} P(k' \mid k) n_{k'}(t). \qquad (14)$$

Fig. 6 compares the predictions of the independent model against the UCN model in a Barabási–Albert (BA) network, which is a type of power-law networks [22]. BA networks considered in this paper are generated using the AS-level BA model

in Boston University Representative Internet Topology Generator (BRITE) simulator [26], which is a tool for topology generation. BRITE simulator can provide good synthetic topologies that are the base of our simulations. In Fig. 6, the BA network has 10 000 nodes, with $\bar{k} = 1.9998, \beta = 0.5$, and $\delta = 0.1$. The infection starts at a single node with a degree of 5. Since the BA networks lack correlations [27], we only consider the UCN model for BA networks. It is observed that the independent model over-predicts the spread of malware, whereas the UCN model under-predicts it. When compared with the simulation results, however, the independent model yields a greater accuracy than the UCN model.

An intuitive explanation for the results in Figs. 5 and 6 can be given as follows: The Epidemiological model, the AAWP model, and the UCN model express the propagation dynamics in terms of how many nodes are infected, without delving into the details of which nodes are infected [10], whereas the independent model considers the details of how nodes are connected to one another. Therefore, the topology information can help us obtain models better than the previous ones. Moreover, the independent model can be used in arbitrary graphs and with varying β_{ji}'s or δ_i's, whereas the other models are used in special graphs and assume that β_{ji} (or δ_i) is identical for every edge (or node).

C. Test of Spatial Independence Assumption

As the independent model achieves better performance than the previous models, Figs. 5 and 6 still show obvious performance gaps between the independent model and the simulation results. Is the spatial independence a good enough assumption? To answer this question, we consider the *correlation coefficient* $\rho_{ij}(t)$ between the status of node i and node j, which is defined as

$$\rho_{ij}(t) = \frac{E[X_i(t)X_j(t)] - E[X_i(t)]E[X_j(t)]}{\sqrt{\text{Var}[X_i(t)]\text{Var}[X_j(t)]}} \qquad (15)$$

where $E[X_i(t)X_j(t)] = P(X_i(t) = 1, X_j(t) = 1), E[X_i(t)] = P(X_i(t) = 1)$, and $\text{Var}[X_i(t)] = P(X_i(t) = 1)[1 - P(X_i(t) = 1)]$. If the status of node i is independent of that of node $j, \rho_{ij} = 0$. Otherwise, if the status of nodes i, j is positively (or negatively) correlated, $\rho_{ij} > 0$ (or $\rho_{ij} < 0$). We obtain the correlation coefficients through simulation on a four-neighbor 2-D-lattice, with 10 000 nodes, $\beta = 0.1, \delta = 0.1$, and 1 000 individual runs. In this 2-D-lattice, each node is represented by its coordinate (x, y), where x, y are integers and $0 \leq x, y \leq 99$. Node (x, y) has four neighbors $(x - 1, y), (x + 1, y), (x, y - 1)$, and $(x, y + 1)$, where the arithmetic operations are *modular* on 100. We assume that the malware begins to spread from node $(0, 0)$, and consider the correlation coefficients between the status of node $(0, 0)$ and node $(0, i)$ (denoted by $\rho_i(t)$) for $i = 1, 2, 3$. Fig. 7 shows how the correlation coefficients vary with time. It is observed that the correlation coefficients are initially close to 0, but increase with time. When $t > 50$, all coefficients are larger than 0.25. This shows a strong dependence in space among nodes, and suggests a better model that accounts for spatial dependence.
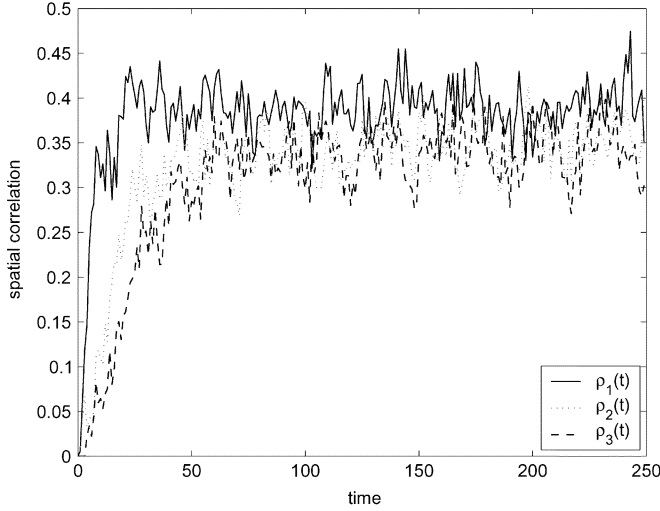
Fig. 7. Spatial correlation in 2-D-lattice with 10 000 nodes, $\beta = 0.1$, and $\delta = 0.1$.

V. MARKOV MODEL

A. Model

Our Markov model assumes a conditional independence in space [28]. That is, at time t $(t = 1, 2, 3, \ldots)$, given the status of node i the status of its neighbors is (conditionally) independent

$$
\begin{aligned}
P\left(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = x_i(t)\right) \\
= \prod_{j \in N_i} P(X_j(t) = x_j(t) \mid X_i(t) = x_i(t)). \quad (16)
\end{aligned}
$$

With the spatial Markov assumption, the dependency graph shown in Fig. 4(a) is changed to the graph shown in Fig. 4(c), where the edges between the neighbors of node i are deleted. The spatial Markov assumption is motivated by the Bethe approximation [12], a way of deriving and correcting the mean-field theory, which has been widely investigated in the area of machine learning. The Bethe approximation factorizes an exact joint probability distribution into a form that only depends on one-node and two-node marginal probabilities in a Markov network. Moreover, the Bethe approximation is shown to be equivalent to belief propagation in [12]. Here we adopt the spirit of the Bethe approximation by incorporating a simple spatial dependence into the Markov model.

Theorem 2: (Markov Model): If the status of node i's neighbors at the same time step is spatially independent given the status of node i, then the state evolution of node i from (5) satisfies

$$
P(X_i(t+1) = 1) = 1 - R_i(t) - P(X_i(t) = 0)S_i^{\mathrm{mar}}(t) \quad (17)
$$

where

$$
S_i^{\mathrm{mar}}(t) = \prod_{j \in N_i} [1 - \beta_{ji} P(X_j(t) = 1 \mid X_i(t) = 0)]. \quad (18)
$$

Proof: Since the status of node i's neighbors at time t is spatially independent given the status of node i, as shown by (16), (4) yields

$$
\begin{aligned}
S_i^{\mathrm{mar}}(t) \\
= \sum_{\mathbf{x}_{N_i}(t)} \prod_{j \in N_i} \left[P(X_j(t) = x_j(t) \mid X_i(t) = 0)(1 - \beta_{ji})^{x_j(t)} \right] \\
= \prod_{j \in N_i} \sum_{x_j(t)} \left[P(X_j(t) = x_j(t) \mid X_i(t) = 0)(1 - \beta_{ji})^{x_j(t)} \right] \\
= \prod_{j \in N_i} [1 - \beta_{ji} P(X_j(t) = 1 \mid X_i(t) = 0)]. \quad \blacksquare
\end{aligned}
$$

The computation of the conditional probability $P(X_j(t) = 1 \mid X_i(t) = 0)$ is given in the Appendix. The Markov model takes into account a part of the neglected correlations between random variables (i.e., node i and its neighbors at time t) and, thus, improves the approximation. The Markov model differs from the independent model only in the probability that one of node i's neighbors infects node i. This probability is $\beta_{ji} P(X_j(t) = 1 \mid X_i(t) = 0)$ for the Markov model, whereas it is $\beta_{ji} P(X_j(t) = 1)$ for the independent model. If the dependence between node i and its neighbors is ignored, the Markov model is reduced to the independent model. Moreover, with the spatial Markov assumption, if node i has k neighbors, the total number of status needed to describe the joint probability $P(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = 0)$ is $O(k)$.

Is it always beneficial to incorporate the spatial dependence? We investigate this issue by introducing the notion of *association* defined in [29].

Definition 1: Random variables $T_1, \ldots, T_n$ are *associated* if

$$
\mathrm{Cov}[f(\mathbf{T}), g(\mathbf{T})] = E[f(\mathbf{T})g(\mathbf{T})] - E[f(\mathbf{T})]E[g(\mathbf{T})] \geq 0
$$

for all nondecreasing functions f and g for which $E[f(\mathbf{T})]$, $E[g(\mathbf{T})]$, and $E[f(\mathbf{T})g(\mathbf{T})]$ exist, and $\mathbf{T} = \{T_1, \ldots, T_n\}$.

In most cases, if one neighbor of node i, e.g., node j, is infected, node i then has an increasing probability to be infected. That is, node i and node j are positively correlated as shown in Fig. 7. Therefore, the status of nodes i and j, $X_i(t)$ and $X_j(t)$, is associated by definition. Furthermore, if $X_i(t)$ and $\mathbf{X}_{N_i}(t)$ are associated random variables, we can show in the following theorem that the Markov model indeed achieves better performance than the independent model.

Theorem 3: (Performance Bound): If $X_i(t)$ and $\mathbf{X}_{N_i}(t)$ are associated, then

$$
S_i^{\mathrm{ind}}(t) \leq S_i^{\mathrm{mar}}(t) \leq S_i(t). \quad (19)
$$

Proof: Since $X_i(t)$ and $X_j(t)(j \in N_i)$ are *associated*, $\mathrm{Cov}[X_i(t), X_j(t)] \geq 0$. We can write

$$
P(X_i(t) = 1, X_j(t) = 1) \geq P(X_i(t) = 1) \cdot P(X_j(t) = 1)
$$

which leads to

$$
P(X_j(t) = 1) \geq P(X_j(t) = 1 \mid X_i(t) = 0). \quad (20)
$$

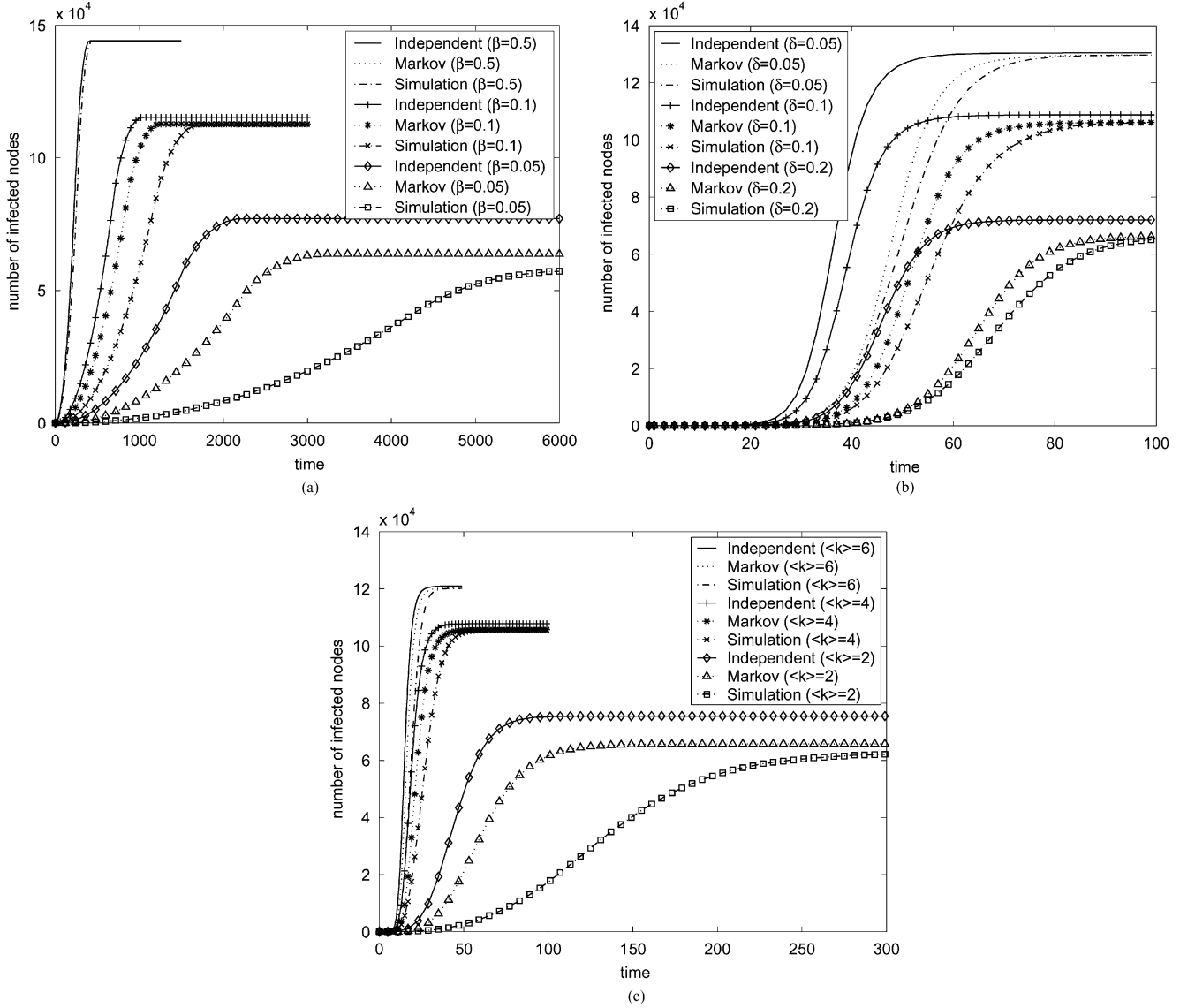Therefore, $S_i^{\mathrm{ind}}(t) \leq S_i^{\mathrm{mar}}(t)$.

Fig. 8. Malware propagation in different topologies. (a) 2-D-lattice with $160\,000$ nodes and $\delta = 0.1$. (b) ER random graph with $160\,000$ nodes, $\bar{k} = 4$, and $\beta = 0.1$. (c) BA network with $160\,000$ nodes, $\beta = 0.1$, and $\delta = 0.1$.

Given $X_i(t) = 0$, let $f(\mathbf{X}_{N_i}(t)) = -(1 - \beta_{li})^{X_l(t)}$ and $g(\mathbf{X}_{N_i}(t)) = -\prod_{j \in N_i - \{l\}} (1 - \beta_{ji})^{X_j(t)}$, where $l \in N_i$. Since $\mathbf{X}_{N_i}(t)$ are associated, from the definition of association we have

$$\text{Cov}\left[f\left(\mathbf{X}_{N_i}(t)\right), g\left(\mathbf{X}_{N_i}(t)\right) \middle| X_i(t) = 0\right] \geq 0 \qquad (21)$$

which leads to

$$E\left[f\left(\mathbf{X}_{N_i}(t)\right) \middle| X_i(t) = 0\right] \cdot E\left[g\left(\mathbf{X}_{N_i}(t)\right) \middle| X_i(t) = 0\right]$$
$$\leq E\left[f\left(\mathbf{X}_{N_i}(t)\right) g\left(\mathbf{X}_{N_i}(t)\right) \middle| X_i(t) = 0\right].$$

Repeated use of the previous argument yields

$$\prod_{j \in N_i} E\left[(1 - \beta_{ji})^{X_j(t)} \middle| X_i(t) = 0\right]$$

$$\leq E\left[\prod_{j \in N_i} (1 - \beta_{ji})^{X_j(t)} \middle| X_i(t) = 0\right]. \qquad (22)$$

That is, $S_i^{\text{mar}} \leq S_i(t)$. ∎

B. Performance

How much does the spatial Markov dependence help in improving the performance? We compare the performance of our proposed models with the simulation results in 2-D-lattice, ER random graph, BA power-law network, a real topology, and a top-down hierarchical topology. Except for the 2-D-lattice, which is a regular graph, we begin each simulation and models with a single, randomly chosen infected node on a given topology. Each plot considers 10 different initially-infected nodes, and each simulation plot also has 10 individual runs for an initially-infected node.

Fig. 8(a) compares the performance of the independent model and the Markov model with the simulation results on a four-neighbor 2-D-lattice. The number of nodes $M = 160\,000$ and the death rate $\delta = 0.1$. For the case of the birth rate $\beta = 0.5$, the three curves nearly coincide with each other. When β decreases, however, the infection spreads at a faster rate in both the independent and the Markov models than the simulation. In all three

cases ($\beta = 0.5, \beta = 0.1, \beta = 0.05$), the Markov model yields more accurate results than the independent model.

Fig. 8(b) shows the predictions of two models with the simulation results on an ER random graph, with 160 000 nodes, an average nodal degree $\bar{k} = 4$, and $\beta = 0.1$. When we constructed the ER random graph, the generated graph was disconnected for that $\bar{k}$ is small. Therefore, in this disconnected graph we chose the largest cluster with 156 763 nodes and an average degree of 4.07 as the target network. It can be seen that the Markov model yields far better performance than the independent model when compared with the simulation results.

Fig. 8(c) depicts the simulation results against two models on a BA network, with 160 000 nodes, $\beta = 0.1, \delta = 0.1$, and $\langle k \rangle = \bar{k}$. For the case when $\bar{k} = 6$, both models give precise results. When $\bar{k}$ decreases, however, the predictions of both models become worse. In all three cases ($\bar{k} = 6, \bar{k} = 4, \bar{k} = 2$), the Markov model predicts malware propagation more accurately than the independent model.

It is observed that the parameters can affect the accuracy of the models. When β or $\bar{k}$ is large, both the independent model and the Markov model perform well. When both β and $\bar{k}$ are small, however, both models fail to predict the slow growth of malware propagation. Therefore, both models are suited for dense graphs, where each node fluctuates independently about its mean value. On the other hand, the Markov model outperforms the independent model in all cases with different parameters and underlying topologies. That is, Theorem 3 is confirmed by the results shown in Fig. 8.

Another observation is that the underlying topology can affect the speed of malware propagation and the final size of infection. For the case of $\beta = 0.1$ and $\delta = 0.1$, although all three graphs (Fig. 8) have the same number of nodes and edges, the malware spreading dynamics in these graphs are significantly different. It takes the malware about 1 716 time steps to enter an equilibrium stage in the 2-D-lattice, whereas it needs about 100 time steps and 66 time steps in the ER random graph and the BA network, respectively. Moreover, after entering the equilibrium stage, the malware infects a total of 112 506 nodes in the 2-D-lattice, 106 023 nodes in the ER random graph, and 105 511 nodes in the BA network. This shows the effect of network structure on the dynamics of malware propagation.

Fig. 9 shows malware propagation in real topology, ER random graph, BA network, and four-neighbor 2-D-lattice for the special case when $\beta = 1$ and $\delta = 0.1$. The real topology is an AS graph collected at the Oregon router server route-views.oregon-ix.net, which is a site for collecting border gateway protocol (BGP) data [30]. The dataset is selected on 1 June 2004 and contains 38 086 links among 17 653 ASes ($\bar{k} = 4.3$). The constructed ER random graph has a largest cluster with 17 648 nodes and $\bar{k} = 4.3$, and the malware only propagates in this largest cluster. The BA network has 17 652 links among 17 653 nodes ($\bar{k} = 2$). The generated BA network is connected and, thus, is a tree. The 2-D-lattice is with 17 689 nodes and $\bar{k} = 4$. Among all these four topologies, the curves of both models overlap with the simulation results in this special case. Therefore, both models can achieve the best performance in the case of $\beta = 1$. Although the AS graph and the ER random graph have almost the same number of (connected) nodes and
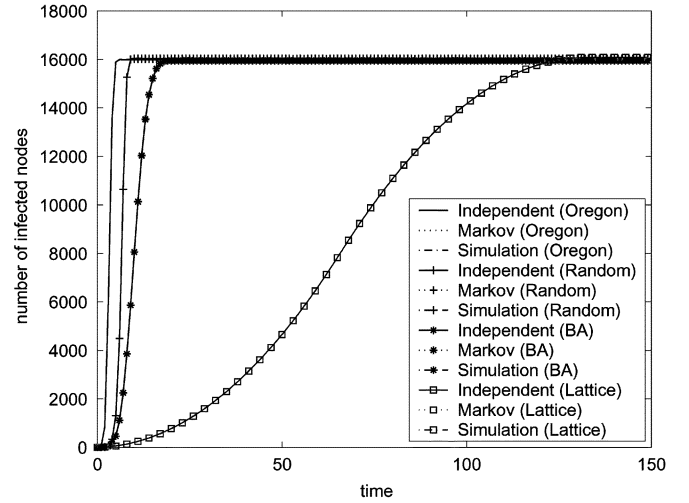

Fig. 9.   Malware propagation in real topology, ER random graph, BA network, and 2-D-lattice with $\beta = 1$ and $\delta = 0.1$.
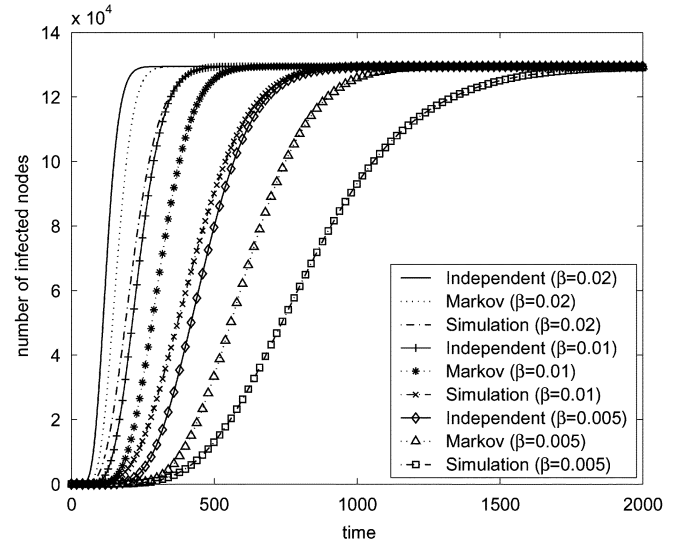

Fig. 10.   Malware propagation in a top-down hierarchical topology with 129 480 nodes, 266 005 edges, and $\delta = 0$.

average nodal degree, the malware takes only 6 time steps to enter an equilibrium stage in the AS graph, whereas it needs about 9 time steps in the ER random graph. This shows that these two topologies have different diameters and the AS graph is more vulnerable to malware propagation than the ER random graph. It is interesting to notice that although the dynamics of malware spreading in different topologies are distinct, the final sizes of infection are almost the same, i.e., $n(t) \approx 16\,000$, when $t = 150$. This reflects that for the case when $\beta = 1$ and $\delta = 0.1$, the final size of infection is not dependent on the network structure, but on the total number of nodes.

Fig. 10 demonstrates another special case when $\delta = 0$, which corresponds to the SI model. The malware spreads in a top-down hierarchical topology generated by BRITE [26]. The top AS-level topology is from NLANR on 2 January 2000 [31], with 6 474 ASes and 13 895 interconnections. The down router-level topology is generated by BRITE router-level BA model, with 20 nodes per AS. The constructed top-down hierarchical topology
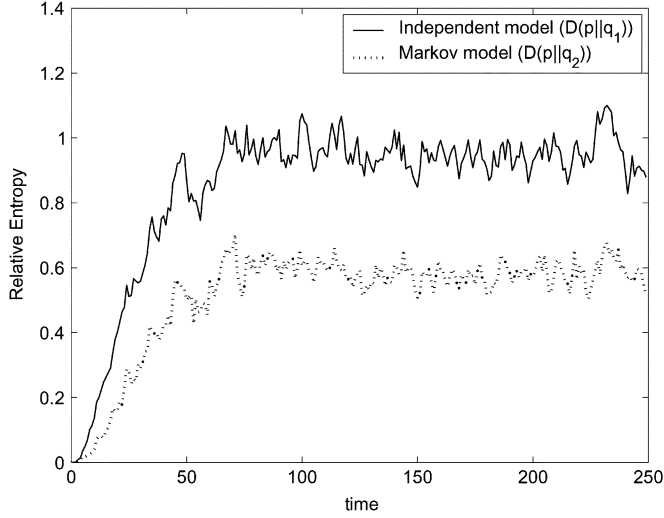
Fig. 11. Relative entropies in 2-D-lattice with 10 000 nodes, $\beta = 0.1$, and $\delta = 0.1$.

has 129 480 nodes and 266 005 edges. The merit of the Markov model can also be observed in this special case when $\delta = 0$.

C. Test of Spatial Markov Assumption

To further examine the goodness of spatial Markov assumption, we use a *relative entropy* (or Kullback-Leibler distance) between two probability mass functions $p(x, t)$ and $q(x, t)$ as defined in [32]

$$D(p \| q) = \sum_x p(x, t) \log \frac{p(x, t)}{q(x, t)}. \qquad (23)$$

The relative entropy is a measure of the distance between two distributions $p(x, t)$ and $q(x, t)$. If $q(x, t)$ is "closer" to $p(x, t)$, $D(p \| q)$ is smaller; and $D(p \| q) = 0$ if and only if $p = q$.

For our case, $p(x, t) = P(\mathbf{X}_{N_i}(t) = \mathbf{x}_{N_i}(t) \mid X_i(t) = 0)$ is the joint distribution of the status of node i's neighbors given node i is susceptible at time t. For the independent model, $q_1(x, t) = \prod_{j \in N_i} P(X_j(t) = x_j(t))$; for the Markov model, $q_2(x, t) = \prod_{j \in N_i} P(X_j(t) = x_j(t) \mid X_i(t) = 0)$. We obtain the relative entropies $D(p \| q_1)$ and $D(p \| q_2)$ through simulation on a four-neighbor 2-D-lattice, with 10 000 nodes, $\beta = 0.1, \delta = 0.1$, and 1000 individual runs. As described in Section IV-C, each node is represented by its coordinate and the malware begins to spread from node $(0, 0)$. Node i is specified at $(1, 1)$. Fig. 11 shows how the relative entropies $D(p \| q_1)$ and $D(p \| q_2)$ change with time. It is observed that the relative entropies are initially close to 0, but increase with time. $D(p \| q_2)$ is smaller than $D(p \| q_1)$ for all time t, suggesting that the spatial Markov model is indeed a better approximation than the spatial independent model. On the other hand, when $t > 60, D(p \| q_2) > 0.5$. This explains the performance gap between the Markov model and the simulation observed in Figs. 8 and 10. Hence, a model that incorporates more spatial dependence than the Markov model may result in a smaller relative entropy.
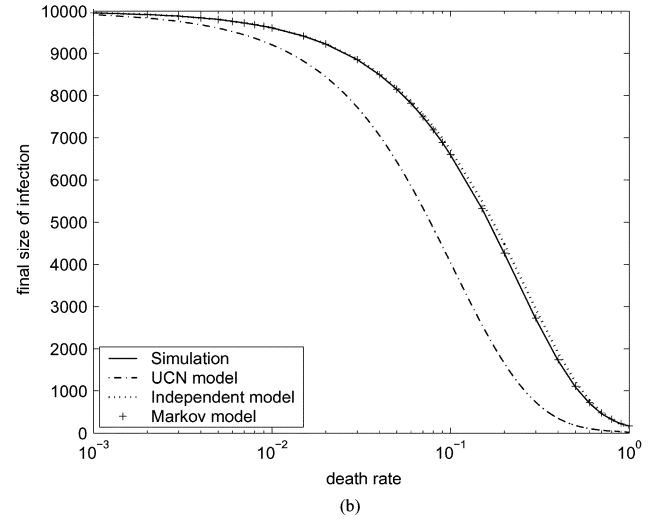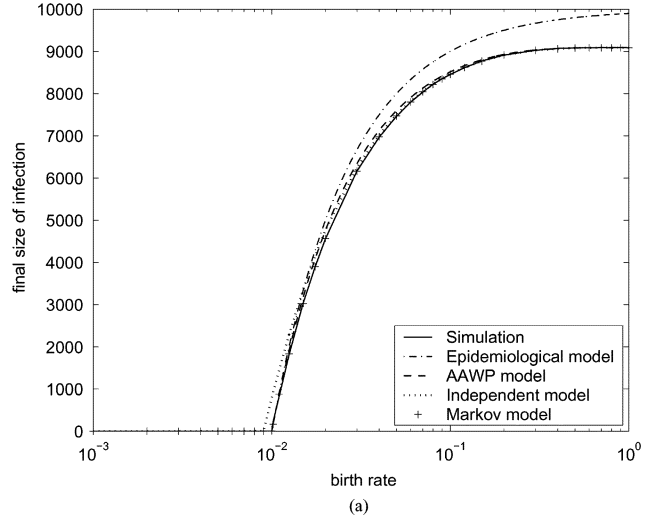

(a)

(b)

Fig. 12. Performance comparisons in estimating the final size of infection. (a) An ER random graph with 10 000 nodes, $\bar{k} = 10$, and $\delta = 0.1$. (b) A BA network with 10 000 nodes, $\bar{k} = 4$, and $\beta = 0.1$.

VI. FINAL SIZE OF INFECTION

The final size of infection corresponds to the equilibrium state of a malware network, which is the average number of infected nodes when time t approaches infinity, i.e., $\lim_{t \to +\infty} n(t)$. The final size of infection characterizes the potential damage due to malware propagation. If the final size of infection can be predicted at an early stage of malware spreading, the potential damage can be assessed and preventive actions can be taken accordingly. In this section, we compare our proposed models with the simulation results and the other models in estimating the final size of infection in homogeneous and complex networks. Each simulation scenario has 100 individual runs and is averaged over the cases that the malware survives. The final size of infection is sampled at time $t = 2000$.

Fig. 12(a) shows a comparison of the Epidemiological model, the AAWP model, the independent model, the Markov model, and the simulation results on a connected ER random graph, with 10 000 nodes, $\bar{k} = 10$, and $\delta = 0.1$. When compared with the simulation results, the Epidemiological model over-predicts the final size of infection when $\beta \geq 0.02$, whereas the AAWP model and the independent model slightly over-predict it. The results of the Markov model and the simulation overlap

for $0.001 \leq \beta \leq 1$. Therefore, the Markov model is the most accurate one among all these models.

Fig. 12(b) gives another comparison of the UCN model, the independent model, the Markov model, and the simulation results on a BA network, with 10 000 nodes, $\bar{k} = 4$, and $\beta = 0.1$. The UCN model under-predicts the final size of infection, whereas the independent model slightly over-predicts it. The results of the Markov model and the simulation overlap for $0.001 \leq \delta \leq 1$. Therefore, both the independent model and the Markov model are shown to be good estimators of the final size of infection, and the Markov model is more accurate than the independent model.

VII. Conclusion

In this paper, we have presented a spatial-temporal model to study the dynamic spreading of malwares that employ different scanning methods. Making use of this model, we have studied the impact of the underlying topology on malware propagation. We show that detailed topology information and spatial dependence are key factors in modeling the spread of malwares. The independent model incorporates detailed topology information and, thus, outperforms the previous models. Our Markov model incorporates both detailed topology information and simple spatial dependence and, thus, achieves a greater accuracy than the independent model, especially when both the birth rate and the average nodal degree are small. Moreover, when the graph is dense, each node fluctuates independently about its mean value and, thus, both models perform well. These results are validated through analysis and extensive simulations on large networks using real and synthesized topologies.

The class of models we have investigated are biased, i.e., with a reduced complexity. Hence, the performance of such models is important. Relative entropy is used as a performance measure and shows that a performance gap still exists between the Markov model and the reality. Formulations are needed to incorporate more spatial dependence into the model. Furthermore, as both models are motivated by the spirit of the mean-field approximation in machine learning, a formal treatment of the mean-field approximation to include temporal dependence will be studied in our future work. As part of the ongoing work, we also plan to estimate the parameters of malware propagation (e.g., birth rate and death rate) and use our proposed models to study the countermeasures for controlling the spread of malwares. Our modeling approach may also help to understand a wide range of information propagation behavior in Internet, such as BGP update streams and file sharing in peer-to-peer applications.

Appendix
Computation of Conditional Probability

To calculate the conditional probability $P(X_j(t) = 1 \mid X_i(t) = 0)$, we introduce a two-node joint probability $P(X_i(t) = 1, X_j(t) = 1)$. Thus

$$P(X_j(t) = 1 \mid X_i(t) = 0)$$
$$= \frac{P(X_j(t) = 1) - P(X_i(t) = 1, X_j(t) = 1)}{P(X_i(t) = 0)}. \quad (24)$$

To simply the notation, we set $P_{uv}(t) = P(X_i(t+1) = 1, X_j(t+1) = 1 \mid X_i(t) = u, X_j(t) = v)$, where $u, v \in \{0, 1\}$. The two-node joint probability can be obtained by the following equations: s

$$P(X_i(t+1) = 1, X_j(t+1) = 1)$$
$$= \sum_{u,v} [P(X_i(t) = u, X_j(t) = v)P_{uv}(t)] \quad (25)$$

where the total probability theorem is used and

$$P_{11}(t) = (1 - \delta_i)(1 - \delta_j) \quad (26)$$

since given that both node i and node j are infected at time t, they independently choose to stay in *infected* status

$$P_{01}(t) = (1 - \delta_j)[1 - S_i'(t)] \quad (27)$$

in that $S_i'(t) = P(X_i(t+1) = 0 \mid X_i(t) = 0, X_j(t) = 1)$, and thus

$$S_i'(t) = (1 - \beta_{ji}) \prod_{l \in N_i - \{j\}} [1 - \beta_{li} P(X_l(t) = 1 \mid X_i(t) = 0)]$$

where the spatial Markov assumption is used; similarly

$$P_{10}(t) = (1 - \delta_i)[1 - S_j'(t)] \quad (28)$$

in that

$$S_j'(t) = (1 - \beta_{ij}) \prod_{l \in N_j - \{i\}} [1 - \beta_{lj} P(X_l(t) = 1 \mid X_j(t) = 0)]$$
$$P_{00}(t) \approx [1 - S_i''(t)][1 - S_j''(t)] \quad (29)$$

where

$$S_i''(t) = \prod_{l \in N_i - \{j\}} [1 - \beta_{li} P(X_l(t) = 1 \mid X_i(t) = 0)]$$
$$S_j''(t) = \prod_{l \in N_j - \{i\}} [1 - \beta_{lj} P(X_l(t) = 1 \mid X_j(t) = 0)].$$

Equation (29) uses an approximation to avoid the introduction of a three-node joint probability $P(X_i(t) = 0, X_j(t) = 0, X_l(t) = x_l(t))$ if nodes i, j, l construct a triangle.

Equation (25) is obtained by replacing $P_{uv}(t)$ with the results from (26)–(29). Equations (17) and (25) provide a recursive relationship between $X_i(t+1), X_j(t+1)$ and $X_i(t), X_j(t)$ for $j \in N_i$. It is assumed that the status of all nodes is independent at time 0.

REFERENCES

[1] Malware Website [Online]. Available: http://www.malware.org/malware.htm

[2] D. Moore, C. Shannon, G. M. Voelker, and S. Savage, "Internet quarantine: Requirements for containing self-propagating code," in *Proc. INFOCOM'03*, vol. 3, San Francisco, CA, Apr. 2003, pp. 1901–1910.

[3] M. Garetto, W. Gong, and D. Towsley, "Modeling malware spreading dynamics," in *Proc. INFOCOM'03*, vol. 3, San Francisco, CA, Apr. 2003, pp. 1869–1879.

[4] C. C. Zou, W. Gong, and D. Towsley, "Code red worm propagation modeling and analysis," in *Proc. 9th ACM Conf. Computer and Communication Security*, Washington, DC, Nov. 2002, pp. 138–147.

[5] Z. Chen, L. Gao, and K. Kwiat, "Modeling the spread of active worms," in *Proc. INFOCOM'03*, vol. 3, San Francisco, CA, Apr. 2003, pp. 1890–1900.

[6] H. Orman, "The Morris worm: A fifteen-year perspective," *IEEE Security Privacy*, vol. 1, pp. 35–43, Sept./Oct. 2003.

[7] S. Staniford, V. Paxson, and N. Weaver, "how to Own the internet in your spare time," in *Proc. 11th USENIX Security Symp. (Security'02)*, 2002.

[8] C. C. Zou, D. Towsley, and W. Gong, "e-mail worm modeling and defense," in *Proc. 13th Int. Conf. Computer Communications and Networks*, Chicago, IL, Oct. 2004, pp. 409–414.

[9] A. Ganesh, L. Massoulie, and D. Towsley, "The effect of network topology on the spread of epidemics," in *Proc. INFOCOM'05*, Miami, FL, Mar. 2005.

[10] J. O. Kephart and S. R. White, "Directed-graph epidemiological models of computer viruses," in *Proc. IEEE Computer Soc. Symp. Research in Security and Privacy*, 1991, pp. 343–359.

[11] M. Opper and D. Saad, Eds., *Advanced Mean Field Methods, Theory and Practice*. Cambridge, MA: MIT Press, Feb. 2001.

[12] J. S. Yedidia, "An idiosyncratic journey beyond mean field theory," *Adv. Mean Field Methods, Theory Pract.*, pp. 21–36, Feb. 2001.

[13] M. Boguñá, R. Pastor-Satorras, and A. Vespignani *et al.*, "Epidemic spreading in complex networks with degree correlations," in *Statist. Mech. Complex Netw.*, R. Pastor-Satorras *et al.*, Eds., 2003, pp. 127–147.

[14] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic spreading: An eigenvalue viewpoint," in *Proc. Symp. Reliable and Distributed Systems*, Florence, Italy, Oct. 2003.

[15] C. Wang, J. C. Knight, and M. Elder, "On viral propagation and the effect of immunization," in *Proc. 16th ACM Annu. Computer Applications Conf.*, New Orleans, LA, Dec. 2000, pp. 343–354.

[16] T. M. Liggett, *Interacting Particle Systems*. New York: Springer-Verlag, 1985.

[17] R. Narasimha, C. Ji, and Z. Chen, "Understanding the significance of spatial dependence in modeling malware propagation on networks," IEEE Transaction on Signal Processing, 2005, to be published.

[18] H. Andersson and T. Britton, "Stochastic epidemic models and their statistical analysis," in *Lecture Notes in Statistics*. New York: Springer-Verlag, 2000, vol. 151.

[19] S. E. Schechter and J. Jung, *Inoculating SSH Against Address-Harvesting Worms*.

[20] P. Erdös and A. Rényi, "On the evolution of random graphs," in *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, 1960, pp. 17–61.

[21] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *Proc. ACM SIGCOMM'99*, 1999, pp. 251–262.

[22] A. Barabási and R. Albert, "Emergence of scaling in random networks," *Science 286*, pp. 509–512, 1999.

[23] S. Saroiu, P. K. Gummadi, and S. D. Gribble, "A measurement study of peer-to-peer file sharing systems," in *Proc. Multimedia Computing and Networking (MMCN'02)*, Jan. 2002.

[24] H. Ebel, L. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," in *Phys. Rev. E 66*, 2002. 035 103(R).

[25] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," Dept. Statist., Univ. California, Berkeley, CA, Tech. Rep. Technical Report 649, 2003.

[26] A. Medina, A. Lakhina, I. Matta, and J. Byers, "BRITE: Universal topology generation from a user's perspective,", Tech. Rep. (User Manual) BU-CS-TR-2001-003, Apr. 2001.

[27] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, "Large-scale topological and dynamical properties of the Internet," *Phys. Rev. E 65*, 2002. 066 130.

[28] S. L. Lauritzen, *Graphical Models*. London, U.K.: Oxford Univ. Press, 1996.

[29] J. D. Esary, F. Proschan, and D. W. Walkup, "Association of random variables, with applications," in *Ann. Math. Statist.*, vol. 38, Oct. 1967, pp. 1466–1474.

[30] University of Oregon Route Views Project [Online]. Available: http://routeviews.org/

[31] NLANR Routing Raw Data [Online]. Available: http://moat.nlanr.net/Routing/rawdata

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
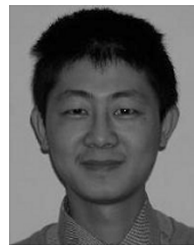
Zesheng Chen (S'03) received the B.S. and M.S. degrees from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 1998 and 2001, respectively. He is currently working toward the Ph.D. degree in the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta.

His research interests include network security and performance evaluation of computer networks.

Chuanyi Ji (S'85–M'91) received the B.S. degree (Honor) from Tsinghua University, Beijing, China in 1983, the M.S. degree from University of Pennsylvania, Philadelphia, in 1986, and the Ph.D. degree from the California Institute of Technology, Pasadena, in 1992, all in electrical engineering.

She is currently an Associate Professor in the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. She was a Faculty Member at Rensselaer Polytechnic Institute, Troy, NY, from 1991 to 2001. She spent her sabbatical at Bell-Labs Lucent in 1999, and was a visiting faculty at the Massachusetts Institute of Technology, Cambridge, MA, in 2000. Her research lies in the areas of network management and control, and adaptive learning systems. Her research interests are in understanding and managing complex networks, applications of adaptive learning systems to network management, learning algorithms, statistics and information theory.

Dr. Ji received the CAREER award from the National Science Foundation in 1995 and the Early Career Award from Rensselaer Polytechnic Institute in 2000.