Short Papers

Dynamics of Malware Spread in Decentralized Peer-to-Peer Networks

Krishna K. Ramachandran and
Biplab Sikdar, *Senior Member*, *IEEE*

Abstract—In this paper, we formulate an analytical model to characterize the spread of malware in decentralized, Gnutella type peer-to-peer (P2P) networks and study the dynamics associated with the spread of malware. Using a compartmental model, we derive the system parameters or network conditions under which the P2P network may reach a malware free equilibrium. The model also evaluates the effect of control strategies like node quarantine on stifling the spread of malware. The model is then extended to consider the impact of P2P networks on the malware spread in networks of smart cell phones.

Index Terms—Malware propagation, peer-to-peer networks, Internet worms and viruses.

1 Introduction

The use of peer-to-peer (P2P) networks as a vehicle to spread malware offers some important advantages over worms that spread by scanning for vulnerable hosts. This is primarily due to the methodology employed by the peers to search for content. For instance, in decentralized P2P architectures such as Gnutella [1] where search is done by flooding the network, a peer forwards the query to it's immediate neighbors and the process is repeated until a specified threshold time-to-live, TTL, is reached. Here TTL is the threshold representing the number of overlay links that a search query travels. A relevant example here is the $Mandragore$ worm [2], that affected Gnutella users. Having infected a host in the network, the worm cloaks itself for other Gnutella users. Every time a Gnutella user searches for media files in the infected computer, the virus *always* appears as an answer to the request, leading the user to believe that it is the file the user searched for. The design of the search technique has the following implications: first, the worms can spread much faster, since they do not have to probe for susceptible hosts and second, the rate of failed connections is less. Thus, rapid proliferation of malware can pose a serious security threat to the functioning of P2P networks.

Understanding the factors affecting the malware spread can help facilitate network designs that are resilient to attacks, ensuring protection of the networking infrastructure. This paper addresses this issue and develops an analytic framework for modeling the spread of malware in P2P networks while accounting for the architectural, topological, and user related factors. We also model the impact of malware control strategies like node quarantine.

The rest of the paper is organized as follows: Section 2 presents the related work and the analytic framework is presented in Section 3. We analyze the model and study the impact of quarantine in Section 4. Simulation results validating our model are presented in Section 5 and Section 6 concludes the paper.

2 Relationship to Prior Work

Though the initial thrust in P2P research was measurement oriented, subsequent works, [3], [4], [5], have proposed analytical models for the temporal evolution of information in the network. The focus of these works is on transfer of regular files and they do not apply to malware that spread actively. In addition, they are specialized to BitTorrent like networks and cannot be extended for P2P networks such as Gnutella or KaZaa.

The issue of worms in peer-to-peer networks is addressed in [6], [7] using a simulation study of P2P worms and possible mitigation mechanisms. Epidemiological models to study malware spread in P2P networks are presented in [8], [9]. These studies assume that a vulnerable peer can be infected by any of the infected peers in the network. This assumption is invalid since the candidates for infecting a peer are limited to those within TTL hops away from it and not the entire network. Another important omission is the incorporation of user behavior. Typically, users in a P2P network alternate between two states: the on state, where they are connected to other peers and partake in network activities and the off state wherein they are disconnected from the network. Peers going offline result in fewer candidates for infection thereby lowering the intensity of malware spread.

An empirical model for malware spreading in BitTorrent is developed in [10] while models for the number of infected nodes by dynamic hit list-based malware in BitTorrent networks is presented in [11], [12]. However, these models ignore node dynamics such as online-offline transitions and are applicable only to BitTorrent networks.

In [13], [14], the authors use hypercubes as the graph model for P2P networks and derive a limiting condition on the spectral radius of the adjacency graph, for a virus/worm to be prevalent in the network. The models do not account for the fact that once a peer is infected, any susceptible peer within a TTL hop radius becomes a likely candidate for a virus attack.

In the current work, we formulate a comprehensive model for malware spread in Gnutella type P2P networks that addresses the above shortcomings. We develop the model in two stages: first, we quantify the average number of peers within TTL hops from any given peer and in the second stage incorporate the neighborhood information into the final model for malware spread.

3 Malware Propagation Model for P2P Networks

This section presents our framework for modeling malware spread in P2P networks. Our model's focus is on the propagation of *malware* and **not** *regular files*.

3.1 Search Mechanism

The transfer of information in a P2P network is initiated with a search request for it. This paper assumes that the search mechanism employed is $flooding$, as in Gnutella networks. In this scenario, a peer searching for a file forwards a query to all its neighbors. A peer receiving the query first responds affirmatively if in possession of the file and then checks the TTL of the query. If this value is greater than zero, it forwards the query outwards to its neighbors, else, the query is discarded. In our scenario, it suffices to distinguish any file in the network as being either malware or otherwise. This is because, as noted earlier, an infected peer replies affirmatively to *all* the queries that it receives with the malware being substituted for the file being searched for. Thus to model malware spread, it is imperative to determine the average

- *The authors are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180. E-mail: krishna.k.ramachandran@gmail.com, bsikdar@ecse.rpi.edu.*

Manuscript received 14 Apr. 2010; accepted 10 Aug. 2010; published online 23 Nov. 2010.
For information on obtaining reprints of this article, please send e-mail to: tdsc@computer.org, and reference IEEECS Log Number TDSC-2010-04-0062.
Digital Object Identifier no. 10.1109/TDSC.2010.69.

rate at which queries reach a node, which in turn depends on the search neighborhood.

We now use the generating function approach as in [15] to quantify the search neighborhood. Define the generating function for the probability mass function (pmf) of the vertex degree as $G_0(x) = \sum_{i=0}^{\infty} p_i x^i$, where p_i is the probability that a randomly chosen vertex has degree i. Since the Gnutella network has a power law degree distribution [16], we have $p_i = Ci^{-\tau}$, where C and τ are constants. The heterogeneity of the connectivity distribution inherent in power law distributions significantly affects the search region of nodes with different degrees. Thus, we evaluate the neighborhood size of a vertex as a function of its degree k.

The distribution of the degree of a vertex that we arrive at by following an edge from a vertex is different from that of an arbitrary vertex in the graph. An edge arrives at a vertex with probability proportional to the degree of the vertex. Thus, the probability that a randomly chosen edge leads to a vertex with degree i is proportional to ip_i. The pmf of the degree of the vertex can then be obtained from the pmf of an arbitrary vertex by normalizing it with $\sum_i ip_i$ and its probability generating function (pgf) is then

$$\frac{\sum_i ip_i x^i}{\sum_i ip_i} = \frac{xG_0'(x)}{G_0'(1)}.$$

As we follow a randomly chosen edge to reach a vertex and then continue on each of the edges of that vertex, and so on, to reach all the m-hop neighbors, the number of vertices arrived at from each vertex has the degree distribution above, less one power of x to compensate for the edge we arrived on. The pgf of the number of outgoing edges at each vertex is then

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)}.$$

With N nodes in the network, the probability of any of these outgoing edges connecting to the original vertex we started at or to any of its immediate neighbors falls as N^{-1} and can thus be neglected as $N \rightarrow \infty$. The number of 2-hop neighbors is the sum of the neighbors of each 1-hop neighbor. Since the generating function for sum of random variables is the product of the individual generating functions, the pgf for the 2-hop neighbors is given by

$$\sum_k p_k[G_1(x)]^k = G_0(G_1(x)).$$

Similarly, the distribution of the m-hop neighbors is given by $G_0(G_1(G_1(\cdots G_1(x))))$, with $m-1$ iterations of the function G_1 acting on itself. Now, given that a node has degree k, the pgf of its degree is given by $G_0^{(k)}(x) = x^k$. Then, the pgf of the number of m-hop neighbors of a node with degree k can be defined in terms of the recursive convolution:

$$G_m^{(k)}(x) \triangleq \begin{cases} x^k & \text{for } m = 1 \\ [\underbrace{G_1(G_1(\cdots(G_1(x))))}_{m-1}]^k & \text{for } m \geq 2. \end{cases} \tag{1}$$

Differentiating the pgf and substituting $x = 1$ yields the average number of m-hop neighbors. For example, the average number of one and two hop neighbors of a peer with degree k are given by $z_1^{(k)} = G_0^{(k)\prime}(1) = k$ and $z_2^{(k)} = \frac{d}{dx}G_0^{(k)}(G_1(x))|_{x=1} = G_0^{(k)\prime}(1)G_1'(1) = kG_0''(1)/G_0'(1)$, respectively, (the expression for $z_2^{(k)}$ uses: $G_1(1) = 1$). The average number of m-hop neighbors is then

$$z_m^{(k)} = \left.\frac{dG^{(m)}}{dx}\right|_{x=1} = G_0^{(k)\prime}(1)[G_1'(1)]^{m-1} = k\left[\frac{z_2}{z_1}\right]^{m-1}, \tag{2}$$

TABLE 1
Notation and P2P Model Parameters

$\lambda_{on}, \lambda_{off}$	rate at which off and on peers switch on and off
λ	rate at which a peer generates queries
$1/\mu$	average download time for a particular file
r_1	rate at which peers terminate ongoing downloads
r_2	rate at which peers renew interest in downloading a file after having deleted it
$1/\delta$	average time for which a peer stores a file

where $z_2 = G_0''(1)$ and $z_1 = G_0'(1)$. Since the search neighborhood of a peer extends up to TTL hops, the average neighborhood size is given by

$$z_{av}^{(k)} = \sum_{i=1}^{TTL} z_i^{(k)} = k\frac{z_1}{z_2 - z_1}\left[\left(\frac{z_2}{z_1}\right)^{TTL} - 1\right]. \tag{3}$$

3.2 Compartmental Model

We formulate our model as a compartmental model, with the peers divided into compartments, each signifying it's state at a time instant. In addition, to account for power-law topologies, we develop the compartmental model in terms of the node degree [17]. For each possible node degree k, the network is partitioned into four classes:

- $P_S^{(k)}$: Number of peers wishing to download a file.
- $P_E^{(k)}$: Number of peers currently downloading the malware.
- $P_I^{(k)}$: Number of peers with a copy of the malware.
- $P_R^{(k)}$: Number of peers who either have deleted the malware or are no longer interested downloading any file.

Further, each class has two components: one comprising of peers of that class that are currently online, while the second represents the offline peers. For instance, $P_{I_{on}}^{(k)}$ denotes the peers with degree k infected by the malware that are currently online and $P_{I_{off}}^{(k)}$, the offline infected peers. Note that since we consider networks with a finite number of nodes, the number of classes is finite, even with power-law topologies. We denote by N_P the total number of peers in the network and by $N_P^{(k)}$ the total number of nodes with degree k, both online and offline. Table 1 defines the parameters used in our model.

Our formulation is based on the principle of mass action, where the behavior of each class is approximated by the mean number in the class at any time instant. By employing the mean-field approach, we make the following assumptions about the system:

- The number of members in a compartment is a differentiable function of time. This holds true in the event of large compartment sizes and since P2P networks comprise of tens of thousands of users, assuming this is quite reasonable.
- By abstracting the P2P graph through differential equations, the emphasis is more on the *numbers* of each class, rather than the particulars of each member of the respective classes.
- The spread of files in the P2P network is *deterministic*, i.e., the behavior is completely determined by the rules governing the model. In other words, the properties of a class are dictated by the *number* of members present.
- The size of the network does not vary over the time during which the spread of malware is modeled.

We first determine the probability that a susceptible peer with degree k is infected when it tries to download an arbitrary file. Following the discussion in Section 3.1, the probability that a neighbor of an arbitrary node has a degree j is given by $\frac{jp_j}{\bar{z}}$, with $\bar{z} = \sum_i ip_i$. Now, when a query reaches a node with degree j, it is infected and responds positively to the query with probability $P_{I_{on}}^{(j)}/N_P^{(j)}$. Then the probability that an arbitrary neighbor is infected, p_{inf}, is given by

$$p_{inf} = \sum_j \frac{jp_j}{\bar{z}} \frac{P_{I_{on}}^{(j)}}{N_P^{(j)}}. \tag{4}$$

Now, a search initiated by a node with degree k, on an average, reaches $z_{av}^{(k)}$ peers. The probability that at least one of the $z_{av}^{(k)}$ peers responds to the query and the susceptible node gets infected is thus $(1 - (1 - p_{inf})^{z_{av}^{(k)}})$.

The dynamics of the spread of malware in peers with degree k can then be represented in terms of the constituent classes by the following deterministic system of equations:

$$\frac{dP_{S_{on}}^{(k)}}{dt} = -\lambda P_{S_{on}}^{(k)}\left(1 - \left(1 - p_{inf}\right)^{z_{av}^{(k)}}\right) + r_1 P_{E_{on}}^{(k)} + r_2 P_{R_{on}}^{(k)} - \lambda_{off} P_{S_{on}}^{(k)} + \lambda_{on} P_{S_{off}}^{(k)} \tag{5}$$

$$\frac{dP_{E_{on}}^{(k)}}{dt} = \lambda P_{S_{on}}^{(k)}\left(1 - \left(1 - p_{inf}\right)^{z_{av}^{(k)}}\right) - r_1 P_{E_{on}}^{(k)} - \mu P_{E_{on}}^{(k)} - \lambda_{off} P_{E_{on}}^{(k)} + \lambda_{on} P_{E_{off}}^{(k)} \tag{6}$$

$$\frac{dP_{I_{on}}^{(k)}}{dt} = \mu P_{E_{on}}^{(k)} - \delta P_{I_{on}}^{(k)} - \lambda_{off} P_{I_{on}}^{(k)} + \lambda_{on} P_{I_{off}}^{(k)} \tag{7}$$

$$\frac{dP_{R_{on}}^{(k)}}{dt} = \delta P_{I_{on}}^{(k)} - r_2 P_{R_{on}}^{(k)} - \lambda_{off} P_{R_{on}}^{(k)} + \lambda_{on} P_{R_{off}}^{(k)} \tag{8}$$

$$\frac{dP_{S_{off}}^{(k)}}{dt} = \lambda_{off} P_{S_{on}}^{(k)} - \lambda_{on} P_{S_{off}}^{(k)} \tag{9}$$

$$\frac{dP_{E_{off}}^{(k)}}{dt} = \lambda_{off} P_{E_{on}}^{(k)} - \lambda_{on} P_{E_{off}}^{(k)} \tag{10}$$

$$\frac{dP_{I_{off}}^{(k)}}{dt} = \lambda_{off} P_{I_{on}}^{(k)} - \lambda_{on} P_{I_{off}}^{(k)} \tag{11}$$

$$\frac{dP_{R_{off}}^{(k)}}{dt} = \lambda_{off} P_{R_{on}}^{(k)} - \lambda_{on} P_{R_{off}}^{(k)}. \tag{12}$$

Note that we have strived to arrive at a generic formulation of the problem encompassing all possible scenarios. Different flavors of the model can be obtained by appropriately choosing the parameter values. For instance, $\mu = \infty$, $P_{E_{off}}^{(k)}(t) = 0, \forall\, t, k$ results in an SIR epidemic model. Also, the offline rates for the various classes have been kept same in order to reduce the number of variable and ease of analysis. Different rates for each class can easily be accommodated in the model.

We now describe the rationale behind the equations of the model above. A transition out of class $P_{S_{on}}^{(k)}$ occurs if either a peer goes offline or initiates a search query that is successful. The former occurs at rate λ_{off} while the latter is contingent on the rate λ at which requests for file download are generated, multiplied by the probability that the query reaches at least one infected node in the online state. Thus, the rate at which the transitions from $P_{S_{on}}^{(k)}$ into $P_{E_{on}}^{(k)}$ occur is given by $\lambda P_{S_{on}}^{(k)}(1 - (1 - p_{inf})^{z_{av}^{(k)}})$. Now, membership of class $P_{S_{on}}^{(k)}$ increases if:

- An offline peer of class $P_S^{(k)}$ comes online: a transition from class $P_{S_{off}}^{(k)}$ which occurs at rate λ_{on}.
- A peer currently downloading terminates the process, say due to unsatisfactory download speeds: a transition from state $P_{E_{on}}^{(k)}$ to $P_{S_{on}}^{(k)}$ at rate r_1.
- A peer that previously had the file, either accidentally or intentionally deletes the file, and wishes to download it again: a transition from state $P_{R_{on}}^{(k)}$ which occurs at rate r_2.

The peers per unit time exiting class $P_{S_{on}}^{(k)}$ total

$$\left(\lambda_{off} + \lambda\left(1 - (1 - p_{inf})^{z_{av}^{(k)}}\right)\right)P_{S_{on}}^{(k)}$$

and those entering number $r_1 P_{E_{on}}^{(k)} + r_2 P_{R_{on}}^{(k)} + \lambda_{on} P_{S_{off}}^{(k)}$. Combining the two gives the rate of change of membership of class $P_{S_{on}}^{(k)}$ as given in (5). Equations characterizing the rates of change for the remaining compartments can be derived in a similar fashion. Note that the transition rates among the various compartments are assumed to be known.

The model presented above represents an upper bound on the number of infected nodes. This is because the model neglects the correlations in the neighborhoods of nodes that are within TTL hops of each other. Also, since malware sizes are typically small (less than a few kilobytes), the download times are expected to be smaller than the on-off transition times of peers which are of the order of hours. Thus, the mean-field approximations used in our analysis are acceptable.

4 Model Analysis

In this section, we analyze the model presented in the previous section and obtain the expressions governing the global stability of the malware free equilibrium (MFE).

4.1 Malware Free Equilibrium

We now proceed with the derivation of the *basic reproduction number*, $\mathcal{R}_0$, a metric that governs the global stability of the MFE. Here, $\mathcal{R}_0$ quantifies the number of vulnerable peers whose security is compromised by an infected host during it's lifetime. It is an established result in epidemiology that $\mathcal{R}_0 < 1$ ensures that the epidemic dies out fast and does not attain an endemic state [18]. Stability information of the MFE is important since this guarantees that the system continues to be malware free even if newly infected peers are introduced.

We follow the methodology presented in [19], [20], where "next generation matrices" have been proposed to derive the basic reproduction number. In this method, the flow of peers between the states are written in the form of two vectors $\mathcal{F}$ and $\mathcal{V}$. The ith element of $\mathcal{F}$ is the rate of appearance of *new* infections in compartment i and the ith element of $\mathcal{V}$ is defined as $\mathcal{V}_i = \mathcal{V}_i^- - \mathcal{V}_i^+$, where $\mathcal{V}_i^+$ is the rate of transfer of peers into compartment i by all other means and $\mathcal{V}_i^-$ is the rate of transfer of peers out of compartment i. These vectors are then differentiated with respect to the state variables, evaluated at the malware free equilibrium, and only the part corresponding to the infected classes are then kept to form the matrices F and V, i.e.,

$$F = \left[\frac{\partial \mathcal{F}_i}{\partial x_j}(x_0)\right],\ V = \left[\frac{\partial \mathcal{V}_i}{\partial x_j}(x_0)\right],\ 1 \leq i, j \leq m, \tag{13}$$

where $\mathcal{F}_i$ and $\mathcal{V}_i$ are the ith entries of $\mathcal{F}$ and $\mathcal{V}$, x_i is the ith system state variable with $\dot{x}_i = \mathcal{F}_i(x) - \mathcal{V}_i(x)$, (x_0) is the malware free equilibrium and m is the number of infectious states. For calculating F and V, the column vectors $\mathcal{F}$ and $\mathcal{V}$ may be considered to consist of m rows, each corresponding to an infectious state. In our model, we have $m = 4K$ corresponding to $P_{E_{on}}^{(k)}$, $P_{E_{off}}^{(k)}$, $P_{I_{on}}^{(k)}$ and $P_{I_{off}}^{(k)}$. Here K is the largest node degree in the network, with $K \leq N_P$. Ordering the infectious states as $P_{E_{on}}^{(1)}, \ldots, P_{E_{on}}^{(K)}$, $P_{E_{off}}^{(1)}, \ldots, P_{E_{off}}^{(K)}$, $P_{I_{on}}^{(1)}, \ldots, P_{I_{on}}^{(K)}$, $P_{I_{off}}^{(1)}, \ldots, P_{I_{off}}^{(K)}$, from (5-12) we have

$$\mathcal{F} = \begin{bmatrix} \lambda P_{S_{on}}^{(1)} \left(1 - (1 - p_{inf})^{z_{av}^{(1)}}\right) \\ \vdots \\ \lambda P_{S_{on}}^{(K)} \left(1 - (1 - p_{inf})^{z_{av}^{(K)}}\right) \\ \bar{\mathbf{0}} \\ \bar{\mathbf{0}} \\ \bar{\mathbf{0}} \end{bmatrix},$$

$$\mathcal{V} = \begin{bmatrix} r_1 P_{E_{on}}^{(1)} + \mu P_{E_{on}}^{(1)} + \lambda_{off} P_{E_{on}}^{(1)} - \lambda_{on} P_{E_{off}}^{(1)} \\ \vdots \\ r_1 P_{E_{on}}^{(K)} + \mu P_{E_{on}}^{(K)} + \lambda_{off} P_{E_{on}}^{(K)} - \lambda_{on} P_{E_{off}}^{(K)} \\ \lambda_{on} P_{E_{off}}^{(1)} - \lambda_{off} P_{E_{on}}^{(1)} \\ \vdots \\ \lambda_{on} P_{E_{off}}^{(K)} - \lambda_{off} P_{E_{on}}^{(K)} \\ \delta P_{I_{on}}^{(1)} + \lambda_{off} P_{I_{on}}^{(1)} - \lambda_{on} P_{I_{off}}^{(1)} - \mu P_{E_{on}}^{(1)} \\ \vdots \\ \delta P_{I_{on}}^{(K)} + \lambda_{off} P_{I_{on}}^{(K)} - \lambda_{on} P_{I_{off}}^{(K)} - \mu P_{E_{on}}^{(K)} \\ \lambda_{on} P_{I_{off}}^{(1)} - \lambda_{off} P_{I_{on}}^{(1)} \\ \vdots \\ \lambda_{on} P_{I_{off}}^{(K)} - \lambda_{off} P_{I_{on}}^{(K)} \end{bmatrix},$$

with $\bar{\mathbf{0}}$ representing a K-row zero vector. Note that only state $E_{on}^{(k)}$ in the set of (5-12) has inflow of new infections and thus only its terms in $\mathcal{F}$ have a nonzero entry. Now, at the malware free equilibrium, we have

- $$\frac{dP_{S_{on}}^{(k)}}{dt} = \frac{dP_{S_{off}}^{(k)}}{dt} = \frac{dP_{E_{on}}^{(k)}}{dt} = \frac{dP_{E_{off}}^{(k)}}{dt} = \frac{dP_{I_{on}}^{(k)}}{dt} = \frac{dP_{I_{off}}^{(k)}}{dt} = \frac{dP_{R_{on}}^{(k)}}{dt} = \frac{dP_{R_{off}}^{(k)}}{dt} = 0$$
- $P_{I_{on}}^{(k)} = P_{I_{off}}^{(k)} = P_{E_{on}}^{(k)} = P_{E_{off}}^{(k)} = 0$

for all $1 \leq k \leq K$. Substituting the above values in (5) and (9), we get $r_2 P_{R_{on}}^{(k)} = 0 \ \Rightarrow\ P_{R_{on}}^{(k)} = 0$. Again, using this result in (12) yields $P_{R_{off}}^{(k)} = 0$. Note that the total number of peers with degree k, given by $N_P^{(k)} = P_{S_{on}}^{(k)} + P_{S_{off}}^{(k)} + P_{I_{on}}^{(k)} + P_{I_{off}}^{(k)} + P_{E_{on}}^{(k)} + P_{E_{off}}^{(k)} + P_{R_{on}}^{(k)} + P_{R_{off}}^{(k)}$, is a constant. Thus, at the MFE we have $N_P^{(k)} = P_{S_{on}}^{(k)} + P_{S_{off}}^{(k)}$, and using the relation from (9), the peer distribution for degree k at the MFE evaluates to the vector: $\{\hat{P}_{S_{on}}^{(k)}, \hat{P}_{S_{off}}^{(k)}, 0, 0, 0, 0, 0, 0\}$, where

$$\hat{P}_{S_{on}}^{(k)} = \frac{\lambda_{on} N_P^{(k)}}{\lambda_{on} + \lambda_{off}}, \quad \hat{P}_{S_{off}}^{(k)} = \frac{\lambda_{off} N_P^{(k)}}{\lambda_{on} + \lambda_{off}}.$$

Differentiating $\mathcal{F}$ and $\mathcal{V}$ with respect to $P_{E_{on}}^{(1)}, \ldots, P_{E_{on}}^{(K)}$, $P_{E_{off}}^{(1)}, \ldots, P_{E_{off}}^{(K)}$, $P_{I_{on}}^{(1)}, \ldots, P_{I_{on}}^{(K)}$, $P_{I_{off}}^{(1)}, \ldots, P_{I_{off}}^{(K)}$ and evaluating at the malware free equilibrium $\{\hat{P}_{S_{on}}^{(k)}, \hat{P}_{S_{off}}^{(k)}, 0, 0, 0, 0, 0, 0\}$ for all $1 \leq k \leq K$, we have

$$F = \begin{bmatrix} \mathbf{0} & G \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, V = \begin{bmatrix} A & \mathbf{0} \\ -C & B \end{bmatrix}, \tag{14}$$

with $\mathbf{0}$ representing a $2K \times 2K$ zero matrix and

$$G = \begin{bmatrix} \frac{\lambda z_{av}^{(1)} \lambda_{on} 1 \cdot p_1}{\bar{z}(\lambda_{on} + \lambda_{off})} & \cdots & \frac{\lambda z_{av}^{(1)} \lambda_{on} K \cdot p_1}{\bar{z}(\lambda_{on} + \lambda_{off})} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\lambda z_{av}^{(K)} \lambda_{on} 1 \cdot p_K}{\bar{z}(\lambda_{on} + \lambda_{off})} & \cdots & \frac{\lambda z_{av}^{(K)} \lambda_{on} K \cdot p_K}{\bar{z}(\lambda_{on} + \lambda_{off})} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{15}$$

$$A = \begin{bmatrix} r_1 + \mu + \lambda_{off} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & r_1 + \mu + \lambda_{off} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \lambda_{on} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \lambda_{on} \end{bmatrix} - \tilde{M}, \tag{16}$$

$$B = \begin{bmatrix} \delta + \lambda_{off} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \delta + \lambda_{off} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \lambda_{on} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \lambda_{on} \end{bmatrix} - \tilde{M}, \tag{17}$$

$$C = \begin{bmatrix} \mu & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \mu & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{18}$$

$$\tilde{M} = \begin{bmatrix} 0 & \cdots & 0 & \lambda_{on} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \lambda_{on} \\ \lambda_{off} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{off} & 0 & \cdots & 0 \end{bmatrix}. \tag{19}$$

Note that A, B, C, G, and $\tilde{M}$ are all $2K \times 2K$matrices. The basic reproduction number, $\mathcal{R}_0$, is then the largest absolute eigenvalue (spectral radius), $\rho()$, of the matrix FV^{-1}, i.e., $\mathcal{R}_0 = \rho(FV^{-1})$. Using elementary matrix algebra and rearranging the terms, it can be easily verified that the product FV^{-1} can be broken down into $GB^{-1}CA^{-1}$, with the constituent matrices as enumerated above. Thus,

$$\mathcal{R}_0 = \rho(GB^{-1}CA^{-1}). \tag{20}$$

4.2 Illustrative Example

We now use a simple scenario to illustrate the effectiveness of the model at isolating the impact of system parameters on the dynamics of malware propagation. Specifically, we show how the impact of user behavior on $\mathcal{R}_0$ can be evaluated by the model. The simplified model makes the following assumptions:

- Instead of modeling the network by grouping nodes according to their degree, we use a single compartment (i.e., $P_{S_{on}}$, $P_{S_{off}}$, $P_{E_{on}}$, $P_{E_{off}}$, $P_{I_{on}}$, $P_{I_{off}}$, $P_{R_{on}}$, and $P_{R_{off}}$) to include all nodes, irrespective of their degree. This model is more appropriate for random graph networks, but is used here for illustrative purposes since unlike the model presented in Section 3.2, leads to closed form solutions that are easier to visualize. Thus in (5-12), we drop the superscript (k) and use $p_{inf} = \frac{P_{I_{on}}}{P_N}$ and $z_{av}^{(k)} = z_{av} = z_1 \frac{z_1}{z_2 - z_1}\left[(z_2/z_1)^{TTL} - 1\right]$.
- Peers do not spend time in the exposed state, i.e., transition occurs directly from $P_{S_{on}}$ to $P_{I_{on}}$.
- Only susceptible peers go offline, i.e., $P_{I_{off}} = P_{R_{off}} = 0$.

This essentially reduces the systems of (5-12) to

$$\frac{dP_{S_{on}}}{dt} = r_2 P_R - \lambda P_{S_{on}}\left[1 - \left[1 - \frac{P_I}{N_P}\right]^{z_{av}}\right] - \lambda_{off} P_{S_{on}} + \lambda_{on} P_{S_{off}} \quad (21)$$

$$\frac{dP_I}{dt} = \lambda P_{S_{on}}\left[1 - \left[1 - \frac{P_I}{N_P}\right]^{z_{av}}\right] - \delta P_I \quad (22)$$

$$\frac{dP_R}{dt} = \delta P_I - r_2 P_R \quad (23)$$

$$\frac{dP_{S_{off}}}{dt} = \lambda_{off} P_{S_{on}} - \lambda_{on} P_{S_{off}}. \quad (24)$$

Using the methodology described above, the basic reproduction number can be calculated as

$$\mathcal{R}_0 = \frac{\lambda z_{av} \lambda_{on}}{\delta(\lambda_{on} + \lambda_{off})}. \quad (25)$$

Now, consider the basic reproduction number (say $\mathcal{R}_0'$) for a model that neglects online-offline transitions, i.e., a peer is always on and in one of the following three states: susceptible, infected or immune. It can be seen that in this case:

$$\mathcal{R}_0' = \frac{\lambda z_{av}}{\delta}. \quad (26)$$

The ratio of (26) and (25) gives us

$$\frac{\mathcal{R}_0'}{\mathcal{R}_0} = \frac{(\lambda_{on} + \lambda_{off})}{\lambda_{on}}. \quad (27)$$

Indeed, if one assumes that a peer strictly alternates between online and offline behavior, the probability that a peer is online at any given time can be derived as

$$p_{on} = \frac{\lambda_{on}}{(\lambda_{on} + \lambda_{off})}. \quad (28)$$

Thus, if we assume $p_{on} = 0.5$, then (27) tells us that models not incorporating peer behavior, such as in [9], end up overestimating the epidemic threshold metric by a factor of two.

4.3 Quarantine

As a form of damage control, the intensity of malware spread can be limited by quarantining infected nodes. This section quantifies the impact of the quarantine rate on the basic reproduction ratio $\mathcal{R}_0$. Quarantine is introduced in the system as follows: we assume that an infected node is taken off the network with probability η. We also assume that this operation does not result in the P2P network being split into disconnected components. The quarantined peers comprise a new compartment $P_Q^{(k)}$ and when rid of malware, enter the recovered state at rate ϑ. This introduces the following changes to the system of (5-12):

- Additional terms to the classes $P_{I_{on}}^{(k)}$ and $P_{R_{on}}^{(k)}$ reflecting the departure of quarantined peers and addition of recovered peers, respectively.
- An additional equation describing the evolution of $P_Q^{(k)}$.

Thus (7) and (8) are, respectively, modified to

$$\frac{dP_{I_{on}}^{(k)}}{dt} = \mu P_{E_{on}}^{(k)} - \delta P_{I_{on}}^{(k)} - \lambda_{off} P_{I_{on}}^{(k)} + \lambda_{on} P_{I_{off}}^{(k)} - \eta P_{I_{on}}^{(k)} \quad (29)$$

$$\frac{dP_{R_{on}}^{(k)}}{dt} = \delta P_{I_{on}}^{(k)} - r_2 P_{R_{on}}^{(k)} - \lambda_{off} P_{R_{on}}^{(k)} + \lambda_{on} P_{R_{off}}^{(k)} + \vartheta P_Q^{(k)} \quad (30)$$

and the dynamics of $P_Q^{(k)}$ are described by

$$\frac{dP_Q^{(k)}}{dt} = \eta P_{I_{on}^{(k)}} - \vartheta P_Q^{(k)}. \quad (31)$$

The addition of class $P_Q^{(k)}$ does not change the equilibrium distribution of peers at the malware free equilibrium. The only change is the addition of an extra infectious state, i.e., $m = 5K$. Accordingly, ordering the states as $P_{E_{on}}^{(1)}, \ldots, P_{E_{on}}^{(K)}$, $P_{E_{off}}^{(1)}, \ldots, P_{E_{off}}^{(K)}$, $P_{I_{on}}^{(1)}, \ldots, P_{I_{on}}^{(K)}$, $P_{I_{off}}^{(1)}, \ldots, P_{I_{off}}^{(K)}$, $P_Q^{(1)}, \ldots, P_Q^{(K)}$, the relevant matrices for computing $\mathcal{R}_0$ are modified as

$$F = \begin{bmatrix} \mathbf{0} & G & \tilde{\mathbf{0}} \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{0}} \\ \tilde{\mathbf{0}}^T & \tilde{\mathbf{0}}^T & \hat{\mathbf{0}} \end{bmatrix}, \quad V = \begin{bmatrix} A & \mathbf{0} & \tilde{\mathbf{0}} \\ -C & \bar{B} & \tilde{\mathbf{0}} \\ \tilde{\mathbf{0}} & D & E \end{bmatrix}, \quad (32)$$

where $\tilde{\mathbf{0}}$ is a $2K \times K$ zero matrix, $\tilde{\mathbf{0}}^T$ is the transpose of $\tilde{\mathbf{0}}$, G, A, C, and $\tilde{M}$ are given in (15), (16), (18), and (19), and

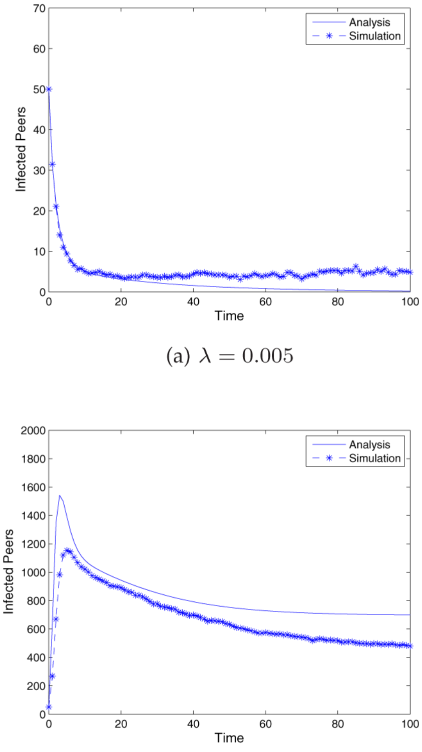
$$\bar{B} = \begin{bmatrix} \eta + \delta + \lambda_{off} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \eta + \delta + \lambda_{off} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \lambda_{on} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \lambda_{on} \end{bmatrix} - \tilde{M}, \quad (33)$$

$$D = \begin{bmatrix} -\eta & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\eta & 0 & \cdots & 0 \end{bmatrix}, \quad (34)$$

$$E = \begin{bmatrix} \vartheta & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \vartheta \end{bmatrix}. \quad (34)$$

Note that D is a $K \times 2K$ matrix and E is a $K \times K$ matrix. We then have $\mathcal{R}_0 = \rho(FV^{-1})$. Again, to illustrate the impact of quarantining infected nodes, we work with the simplified scenario introduced in Section 4.2 and evaluate the dependency of $\mathcal{R}_0$ on the quarantine rate. The equations for the model, (21-24) now become

$$\begin{aligned} \frac{dP_{S_{on}}}{dt} =& r_2 P_R - \lambda P_{S_{on}}\left[1 - \left[1 - \frac{P_I}{N_P}\right]^{z_{av}}\right] - \lambda_{off} P_{S_{on}} \\ &+ \lambda_{on} P_{S_{off}} \end{aligned} \quad (36)$$



(a) $\lambda = 0.005$

(b) $\lambda = 2.0$

Fig. 1. Impact of λ on malware intensity for the system in (5-12).

$$\frac{dP_I}{dt} = \lambda P_{S_{on}}\left[1 - \left[1 - \frac{P_I}{N_P}\right]^{z_{av}}\right] - \delta P_I - \eta P_I \tag{37}$$

$$\frac{dP_Q}{dt} = \eta P_I - \vartheta P_Q \tag{38}$$

$$\frac{dP_R}{dt} = \delta P_I - r_2 P_R + \vartheta P_Q \tag{39}$$

$$\frac{dP_{S_{off}}}{dt} = \lambda_{off} P_{S_{on}} - \lambda_{on} P_{S_{off}}. \tag{40}$$

Now, following the procedure outlined in Section 4.1

$$\mathcal{F} = \begin{bmatrix} \lambda P_{S_{on}}\left[1 - \left[1 - \frac{P_I}{N_P}\right]^{z_{av}}\right] \\ 0 \end{bmatrix}, \mathcal{V} = \begin{bmatrix} (\delta + \eta) P_I \\ \vartheta P_Q - \eta P_I \end{bmatrix} \tag{41}$$

$$\begin{aligned} F &= \begin{bmatrix} \lambda z_{av} p_{on} & 0 \\ 0 & 0 \end{bmatrix} \text{ and } V = \begin{bmatrix} (\delta + \eta) & 0 \\ -\eta & \vartheta \end{bmatrix} \\ FV^{-1} &= \frac{1}{\vartheta(\delta + \eta)} \begin{bmatrix} \vartheta \lambda z_{av} p_{on} & 0 \\ 0 & 0 \end{bmatrix}, \end{aligned} \tag{42}$$

and therefore

$$\mathcal{R}_0 = \frac{\lambda z_{av} p_{on}}{(\delta + \eta)}. \tag{43}$$

The malware spread does not reach epidemic proportions provided $\mathcal{R}_0 < 1$ and hence, the required rate for quarantining infected peers need is $\eta > \lambda z_{av} p_{on} - \delta$. Such a measure takes the node off the P2P network and thus it would not be able to participate in any further file transfers until the malware has been completely removed. This is indeed necessary since an infected peer *always* responds positively to *any* query with the malware cloaked as the file being searched for. Thus, the only way to prevent the node from infecting others is to take it off the network.

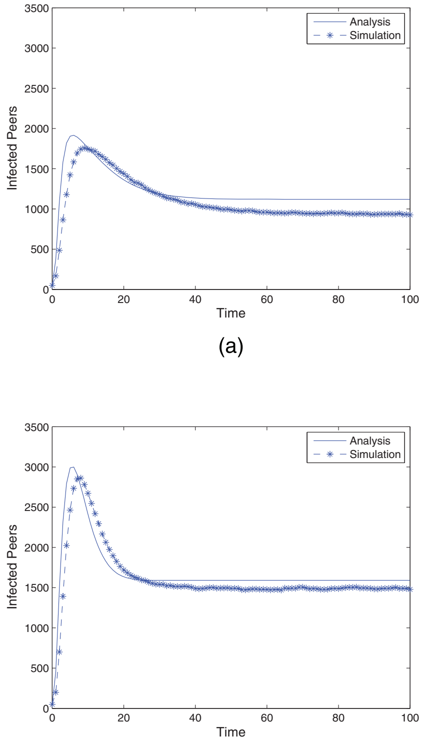
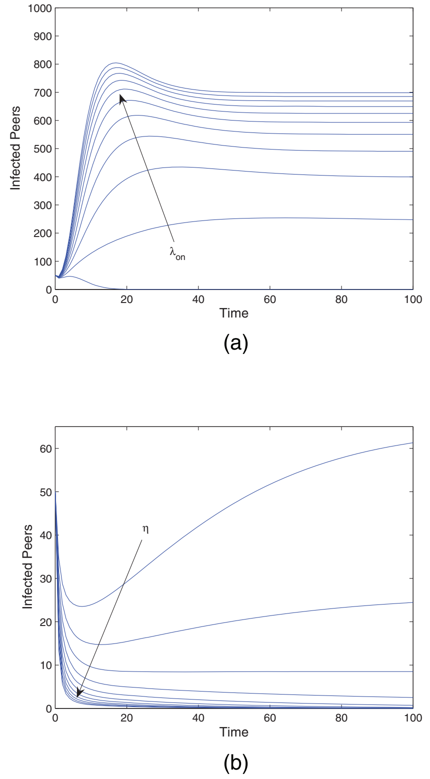


Fig. 2. Influence of offline duration on malware intensity for the system in (5-12). (a) $\lambda_{on} = 0.27$, $\lambda = 1.0$. (b) $\lambda_{on} = 0.75$, $\lambda = 1.0$.

5 Results

In this section, we validate our model using simulations and also demonstrate its capability to illustrate the effect of various system parameters on malware dynamics. The simulations were conducted using a custom built simulator. Results are reported for a 10,000 node network with a power-law graph topology with $\tau = 3.4$. The initial network state for all simulations consisted of 4,950 randomly selected nodes in the susceptible online state, 5,000 randomly selected nodes in the susceptible offline state, and 50 randomly selected nodes in the infected online state. Other parameters that stayed constant in all simulations (unless otherwise noted) were $\lambda_{on} = 0.1$, $\lambda_{off} = 0.2$, $\mu = 0.5$, $\delta = 0.3$, $r_1 = 0.1$, $r_2 = 0.1$, and $\vartheta = 0.1$. The results for each parameter setting are averaged over 20 runs and the 90 percent confidence interval was within 10 percent of the mean.

Figs. 1a and 1b substantiate our analytical result that requires the basic reproduction number to be greater than 1 for an epidemic to prevail. We see that if $\mathcal{R}_0 < 1$, the number of infected peers drops down to zero (Fig. 1a), else it reaches endemic proportions (Fig. 1b). From (20), we see that $\mathcal{R}_0$ is directly proportional to λ_{on}. This implies that nodes staying online for long periods as compared to their offline durations result in a higher intensity of malware presence. Simulations concur with the above observation and are shown in Figs. 2a and 2b. The analytic model tends to overestimate the steady-state number of infected nodes when $\mathcal{R}_0 > 1$. This is because our model does not take into account the correlation in the neighborhoods of nodes that are within TTL hops of each other. The sensitivity of malware intensity to λ_{on} (varied from 0.0 to 1.0 in steps of 0.1) is shown in Fig. 3a for $\lambda = 0.02$ and the intensity of the epidemic increases monotonically with λ_{on}. Simulations results have been omitted from Fig. 3a to avoid clutter.



(a)

(b)

Fig. 3. Simulation results for the system in (5-12) (top) and (29-31) (bottom). (a) Effect of λ_{on} on malware intensity. (b) Effect of quarantine on malware intensity.

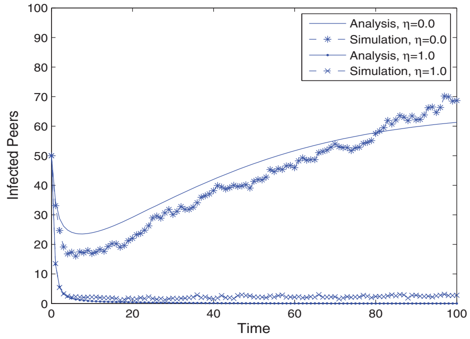


Fig. 4. Effect of quarantine on the system in (29-31) for $\lambda = 0.02$.

The effectiveness of quarantine in controlling the spread of malware is shown in Fig. 4 which shows the infected population in the network with and without quarantine. Also, (43) depicts an inverse relationship between $\mathcal{R}_0$ and the quarantine rate η. Analytic results for increasing values of η (from 0.0 to 1.0 in steps of 0.1) for $\lambda = 0.02$ for the power-law topology are presented in Fig. 3b which shows that the malware intensity is inversely proportional to η.

6 Conclusion

In this paper, we developed an analytic model to understand the dynamics of malware spread in P2P networks. The need for an analytic framework incorporating user characteristics (e.g., offline to online transitional behavior) and communication patterns (e.g., the average neighborhood size) was put forth by quantifying their influence on the basic reproduction ratio. It was shown that models that do not incorporate the above features run the risk of grossly overestimating $\mathcal{R}_0$ and thus falsely report the presence of an epidemic.

Acknowledgments

This work was supported in part by the US National Science Foundation (NSF) grant 0347623.

References

[1] Clip2, "The Gnutella Protocol Specification v0.4," http://www.clip2.com/GnutellaProtocol04.pdf, Mar. 2001.

[2] E. Damiani, D. di Vimercati, S. Paraboschi, P. Samarati, and F. Violante, "A Reputation-Based Approach for Choosing Reliable Resources in Peer-to-Peer Networks," *Proc. ACM Conf. Computer and Comm. Security (CCS),* pp. 207-216, Nov. 2002.

[3] X. Yang and G. de Veciana, "Service Capacity in Peer-to-Peer Networks," *Proc. IEEE INFOCOM '04*, pp. 1-11, Mar. 2004.

[4] D. Qiu and R. Srikant, "Modeling and Performance Analysis of BitTorrent-Like Peer-to-Peer Networks," *Proc. ACM SIGCOMM,* Aug. 2004.

[5] J. Mundinger, R. Weber, and G. Weiss, "Optimal Scheduling of Peer-to-Peer File Dissemination," *J. Scheduling,* vol. 11, pp. 105-120, 2007.

[6] A. Bose and K. Shin, "On Capturing Malware Dynamics in Mobile Power-Law Networks," *Proc. ACM Int'l Conf. Security and Privacy in Comm. Networks (SecureComm),* pp. 1-10, Sept. 2008.

[7] L. Zhou, L. Zhang, F. McSherry, N. Immorlica, M. Costa, and S. Chien, "A First Look at Peer-to-Peer Worms: Threats and Defenses," *Int'l Workshop Peer-To-Peer Systems,* Feb. 2005.

[8] F. Wang, Y. Dong, J. Song, and J. Gu, "On the Performance of Passive Worms over Unstructured P2P Networks," *Proc. Int'l Conf. Intelligent Networks and Intelligent Systems (ICINIS),* pp. 164-167, Nov. 2009.

[9] R. Thommes and M. Coates, "Epidemiological Models of Peer-to-Peer Viruses and Pollution," *Proc. IEEE INFOCOM '06,* Apr. 2006.

[10] J. Schafer and K. Malinka, "Security in Peer-to-Peer Networks: Empiric Model of File Diffusion in BitTorrent," *Proc. IEEE Int'l Conf. Internet Monitoring and Protection (ICIMP '09),* pp. 39-44, May 2009.

[11] J. Luo, B. Xiao, G. Liu, Q. Xiao, and S. Zhou, "Modeling and Analysis of Self-Stopping BT Worms Using Dynamic Hit List in P2P Networks," *Proc. IEEE Int'l Symp. Parallel and Distributed Processing (IPDPS '09),* May 2009.

[12] W. Yu, S. Chellappan, X. Wang, and D. Xuan, "Peer-to-Peer System-Based Active Worm Attacks: Modeling, Analysis and Defense," *Computer Comm.,* vol. 31, no. 17, pp. 4005-4017, Nov. 2008.

[13] A. Ganesh, L. Massoulie, and D. Towsley, "The Effect of Network Topology on the Spread of Epidemics," *Proc. IEEE INFOCOM,* 2005.

[14] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint," *Proc. IEEE Int'l Symp. Reliable Distributed Systems (SRDS),* 2003.

[15] M. Newman, S. Strogatz, and D. Watts, "Random Graphs with Arbitrary Degree Distribution and Their Applications," *Physical Rev. E,* vol. 64, no. 2, pp. 026118(1-17), July 2001.

[16] D. Stutzbach and R. Rejaie, "Characterizing the Two-Tier Gnutella Topology," *Proc. ACM Int'l Conf. Measurement and Modeling of Computer Systems (SIGMETRICS),* pp. 402-403, June 2005.

[17] R. Pastor-Satorras and A. Vespignani, "Epidemic Dynamics in Scale-Free Networks," *Physical Rev. E,* vol. 65, no. 3, p. 035108(1-4), Mar. 2002.

[18] O. Diekmann and J. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation.* Wiley, 1999.

[19] P. van den Driessche and J. Watmough, "Reproduction Numbers and Sub-Threshold Endemic Equilibria for Compartmental Models of Disease Transmission," *Math. Biosciences,* vol. 180, pp. 29-48, 2002.

[20] J. Arnio, J. Davis, D. Hartley, R. Jordan, J. Miller, and P. van den Driessche, "A Multi-Species Epidemic Model with Spatial Dynamics," *Math. Medicine and Biology,* vol. 22, pp. 129-142, Mar. 2005.