Impact of Network Structure on Malware Propagation: A Growth Curve Perspective

HONG GUO, HSING KENNETH CHENG, AND KEN KELLEY

HONG GUO (hguo@nd.edu; corresponding author) is an assistant professor of information systems at Mendoza College of Business, University of Notre Dame. She specializes in the economic analysis of information technology policy issues such as net neutrality and public safety networks. She is also interested in characterizing key design features of emerging information systems (e.g., consumer review systems, mobile platforms, digital games, and others) and examining firms' corresponding strategies.

HSING KENNETH CHENG (hkcheng@ufl.edu) is John B. Higdon Eminent Scholar in the Department of Information Systems and Operations Management of Warrington College of Business Administration, University of Florida. He received his Ph.D. in computers and information systems from William E. Simon Graduate School of Business Administration, University of Rochester. His research interests focus on analyzing the impact of the Internet and information technology on software development and marketing, and on information systems policy issues, in particular, network neutrality. He has chaired and served on the program committee of many information systems conferences and workshops.

KEN KELLEY (kkelley@nd.edu) is professor of management in the analytics area at Mendoza College of Business, University of Notre Dame. His work focuses on the development, improvement, and evaluation of statistical methods and measurement issues in the areas of research design, effect size estimation and confidence interval formation, longitudinal data analysis, and statistical computing. He is an Accredited Professional Statistician™ of the American Statistical Association, associate editor of *Psychological Methods*, recipient of the Anne Anastasi Early Career Award of the American Psychological Association's Division of Evaluation, Measurement, and Statistics, and a fellow of the American Psychological Association.

ABSTRACT: Malicious software, commonly termed "malware," continuously presents one of the top security concerns, and causes tremendous worldwide financial losses for organizations. In this paper, we propose a structural risk model to analyze malware propagation dynamics measured by a four-parameter (asymptote, point of inflection, rate, and infection proportion at inflection) growth curve. Using both social network data and technological network infrastructure from a large organization, we estimate the proposed structural risk model based on incident-specific nonlinear growth curves. This paper provides empirical evidence for the explanatory power of the structural characteristics of the underlying networks on malware propagation dynamics. This research provides useful findings for security managers in designing their malware defense strategies. We also simulate three common malware defense strategies (preselected immunization strategies, countermeasure

Journal of Management Information Systems / 2016, Vol. 33, No. 1, pp. 296–325.

ISSN 0742–1222 (print) / ISSN 1557–928X (online)
DOI: 10.1080/07421222.2016.1172440

dissemination strategies, and security awareness programs) based on the proposed structural risk model and show that they outperform existing strategies in terms of reducing the size of malware infection.

KEY WORDS AND PHRASES: information systems security, malware defense, malware propagation, malware propagation trajectory, network analysis, social networks, technological networks.

Malicious software, commonly termed "malware," continuously presents one of the top security concerns for organizations [18]. Worldwide financial losses due to malware averaged $12.18 billion per year from 1997 to 2006 [17] and increased to $110 billion between July 2011 and the end of July 2012 [56]. Typical malware includes viruses, worms, Trojan horses, spyware, adware, and others. Since the first computer virus surfaced in the early 1980s, malware has developed into thousands of variants that differ in infection mechanism, propagation mechanism, destructive payload, and other features [22]. Among them, viruses and worms, the two most commonly seen types of malware, have drawn more industry and research attention than other types of malware due to their self-replicating nature, dramatic propagation speed, and potentially severe destructive consequences. For these reasons, in this paper we study the propagation process of viruses and worms within organizational networks.

Viruses and worms may propagate through different organizational networks, which can be divided into two categories—technological networks (TN) and social networks (SN). As the computational foundation of organizational business processes, technological networks (e.g., LANs and WANs) consisting of interconnected computers, routers, and other network devices, enable data transmissions to perform required tasks. Some malware may propagate through technological networks. For example, the Blaster starts from a local machine's IP address or a completely random address and attempts to infect sequential IP addresses. In addition to technological networks, there are social networks (e.g., e-mails, instant messaging systems, P2P networks, and social networking sites) inherently embedded within an organization. For example, business communications as well as personal contacts among employees inside and outside their departments constitute an information distribution network. Recently, social-network-based malware has become a great threat because of the popularity of social media in organizations [15, 42]. Users expose more personal information in a social networking environment and are more of a target for social-network-based malware [19]. In addition, social networking tools connect individuals who have a certain level of mutual trust, which enables malware to disguise and propagate easily over social networks [39]. For example, MyDoom is transmitted primarily via e-mail and P2P network. Koobface is another representative and revolutionary social-network-based malware considered the first to successfully propagate through social networking sites [1].

In this paper, we view the propagation process of self-replicating malware as a special type of network flow—that is, computer viruses and worms start from certain nodes and propagate through the edges within organizational networks. This paper aims to address several important questions regarding malware propagation and defense: how does malware propagate within organizational networks, that is, TNs and SNs? What are the appropriate measures of network structures in the context of self-replicating malware propagation? How can network structural characteristics be used to explain the dynamics of the malware propagation process? What are the implications of network structure for malware defense?

This paper adopts theoretical concepts and methods from the field of social network analysis, namely, centrality measures and subgroup analysis, to capture the structural characteristics of both social network and technological network. In particular, based on the unique properties of the malware propagation process, we identify random-walk betweenness as the appropriate centrality measure to evaluate the structural position of individual nodes. Modularity-maximizing decomposition is then applied to analyze the embedded subgroup structure. Based on the derived centrality measures and the discovered subgroup structure, we formulate our structural risk model to examine the impact of individual-, group-, and network-level characteristics on malware propagation dynamics. Details of the structural risk model are provided in the research model section.

In order to estimate and evaluate the proposed structural risk model, we construct real organizational networks and simulate the self-replicating malware propagation processes. We consider a real social network structure constructed from a large social networking site and then map nodes in the social network to nodes in the technological network within the organization. Details of the network construction process are provided in the research sample section.

Based on the constructed networks, we further compute random-walk betweenness (*Betweenness*) and size of the modularity-maximizing subgroups (*GroupSize*), both of which serve as independent variables in the proposed structural risk model. We then simulate the malware propagation process through the constructed networks. Absent real infection data, which are infeasible to obtain, studying these sample networks and using their structural measures to explain the spread patterns provide insights into the real malware propagation process.[1] We record the environment variables of the malware propagation simulations—*Virus, StartNum, InfRate, RecRate, VirusActRate*. These simulation environment variables serve as control variables in the proposed structural risk model. In this study, we use a generalized logistic growth curve with four parameters—asymptote (A), point of inflection (I), rate (R), and infection proportion at inflection (P), to capture the complex process of malware propagation. This four-parameter generalized logistic growth curve is used chiefly for growth of organisms (e.g., trees) in the literature, but we have adapted its use to model the cumulative growth of infected computers in a malware propagation process. These four parameters serve as dependent variables in the proposed structural risk model. Next, we use hierarchical regressions to estimate the structural risk model and show that random-walk betweenness of individual nodes, their local subgroup structures, and the type of network have

both statistically and practically significant collective inference on self-replicating malware propagation dynamics. Details of network analysis and model estimation are provided in the analysis and results section. Our simulation experiments further demonstrate that the proposed structural risk model can be integrated into existing malware defense strategies (preselected immunization strategies, countermeasure dissemination strategies, and security awareness programs), which outperform the existing defense strategies in terms of reducing the size of malware infection.

There are four major findings in this paper. First, the four-parameter generalized logistic growth curve provides a remarkably accurate approach to modeling the cumulative number of infected computers in a malware incident, as evidenced by the median R^2 (coefficient of determination) of 0.998. Second, random-walk betweenness captures the structural characteristics of individual nodes in the context of malware propagation and explains the propagation dynamics. Third, the subgroup structure of both social network and technological network and the corresponding subgroup characteristics (i.e., the size of the subgroup) have a significant impact on the malware propagation process. In a malware incident, more computers get infected if the malware starts from larger subgroups. Fourth, when a virus or worm propagates through a social network, it spreads more slowly but eventually infects a larger number of computers, compared to propagating through a technological network. These findings have important managerial implications, which are discussed in the conclusions section.

Literature Review

Malware propagation and defense has long been studied by researchers in information systems security [43]. Intrusion detection systems [10] and intrusion prevention systems [67] are used for reactive detection and proactive prevention of malware attacks. At the managerial level, different frameworks and techniques are proposed to evaluate the impact of malware [51]. Security patch management designs optimal security patching strategies by balancing the interaction between vendors' patch-release policies and firms' patch-update policies [9]. Firms may also resort to economic mechanisms, such as cyberinsurance, risk pooling arrangements, and managed security services, to manage information security risks [70, 71]. Analyzing and evaluating security risk [55, 68], leveraging system modularity [64], implementing and increasing user awareness of security countermeasures [20, 36], and optimizing security investment [8, 29, 58] are critical procedures for effective information systems security management.

Prior research on network analysis and malware propagation shows that network topology is a crucial factor for malware propagation. In a malware incident, the topology of the victim network has been shown to be one of the key determinant factors of the propagation speed and destructive consequences in various contexts such as e-mail networks [50, 51], mobile phone networks [23], supply distribution networks [69], and so on. Extant work in this area takes two distinct approaches in studying the malware propagation process [6]. Computer scientists analyze the spread of malware in

complex networks using epidemiological models from disease propagation [34] and interactive Markov chains [26]. In their work, complex networks usually display two distinct properties: a scale-free connectivity distribution in which nodes follow a power law distribution [21, 32, 37, 62], and small-world property with small average path lengths between any two nodes [31, 33, 44, 72]. In contrast, social scientists distinguish among different kinds of dyadic links, emphasizing variation in network structure across different individual nodes and using these variations to explain nodes' different outcomes [3, 7]. Our work follows the perspective of social science and examines how the variations in network properties of individual nodes account for differences in their role in the malware propagation process.

This paper contributes to both the literature of information systems security and the literature of network analysis. We study the impact of network structure on the malware propagation dynamics at the micro level. At the individual level, this study identifies an appropriate structural measure, that is, random-walk betweenness of individual nodes, and analyzes the impact of the random-walk betweenness of the starting nodes on self-replicating malware propagation dynamics. At the group level, this paper investigates the subgroup structure of the networks and demonstrates that characteristics of the local groups of the starting nodes significantly influence the malware propagation process. At the network level, there are intrinsic structural differences between social networks and technological networks. This paper proposes a holistic view of an organization's computing environment to examine malware propagation patterns within both social networks and technological networks. Our findings suggest that these different levels of structural characteristics should be incorporated in malware defense mechanisms to better secure organizational computing environments. Prior research on network topology and malware propagation relies on simulated networks with certain properties such as scale-free and small-world networks. This paper takes an alternative approach with real social and technological networks within an organization. This approach allows us to empirically estimate and evaluate the proposed structural risk model without making strict assumptions about the network structure.

In addition, this paper makes a methodological contribution by introducing growth curves, which are specific to individual malware incidents, for modeling the propagation process. We used a generalized logistic growth curve with four key parameters to capture the dynamics of malware propagation. This four-parameter generalized logistic function demonstrated an exceptionally good fit to the malware propagation data. Knowledge of these variables helps organizations to greatly improve security risk assessment prior to malware incidents and to take timely actions during malware incidents.

Research Model

Modeling the Malware Propagation Process and Dependent Variables

In this subsection, we discuss modeling the malware propagation process with an emphasis on parameter interpretation in the context of malware propagation. We

model the process of malware propagation within an organizational network as a growth curve of the *cumulative number of infected computers* as a function of time for each malware incident. We use a generalized logistic function, a type of growth curve, that offers a flexible sigmoidal form and is widely used in modeling growth in organisms [52, 53]. Because similar sigmoidal forms of growth to an asymptote have been observed in prior studies in malware propagation [23, 61, 62], we adapt the use of the generalized logistic function to model the cumulative growth of infected computers. Several parameterizations of the generalized logistic function exist and we use the following parameterization [41, 45, 57] to model the malware propagation process:

$$InfNum_t = \frac{A}{\left[1 + Se^{-R(t-1)}\right]^{1/S}}, \tag{1}$$

where t is the value of time since the process started at time zero and $InfNum_t$ represents the cumulative *inf*ection *num*ber at time t. At time zero, that is, $t = 0$ (baseline), multiple nodes within an organizational network may get infected. These nodes represent the starting nodes of a malware incident and the corresponding $InfNum_0$ represents the number of starting nodes.

As shown in Equation (1), there are four parameters in this generalized logistic function. The *asymptote* parameter, denoted by A, represents the upper limit of the cumulative number of infected computers, which measures the size of malware infection. The *point of inflection* parameter, denoted by I, represents the time point when the maximum slope of growth in the infection number occurs. The point of inflection also indicates the time point when the curve turns from convex to concave. The *rate* parameter, denoted by R, indicates the propagation speed. The *shape* parameter, denoted by S, determines the shape of the growth curve. More explicitly, the shape parameter determines the propagation proportion of the asymptote at the point of inflection. We introduce the *infection proportion at inflection* parameter, denoted by P, to represent the proportion of total number of infected computers at the point of inflection. Because the cumulative number of infected computers is $InfNum_I = \frac{A}{(1+S)^{1/S}}$ at the point of inflection, the proportion is given by Equation (2):

$$p = \frac{1}{(1+S)^{1/S}}. \tag{2}$$

There is a one-to-one mapping between P and S. The domain for S is $[-1, \infty]$ while the domain for P is [0, 1]. The infection proportion at inflection parameter (P) increases in the shape parameter (S), that is, when S increases, there is a higher proportion of infected computers at the point of inflection. For example, when $S = 1$, half of the total number of infected computers are reached at the point of inflection, that is, $P = 0.5$; when S increases to 10, P increases to 0.79. Figure 1 depicts this four-parameter generalized logistic function and its parameters.

This general growth function contains several other growth curves as special cases. Specifically, the four-parameter generalized logistic function, defined in Equation (1),

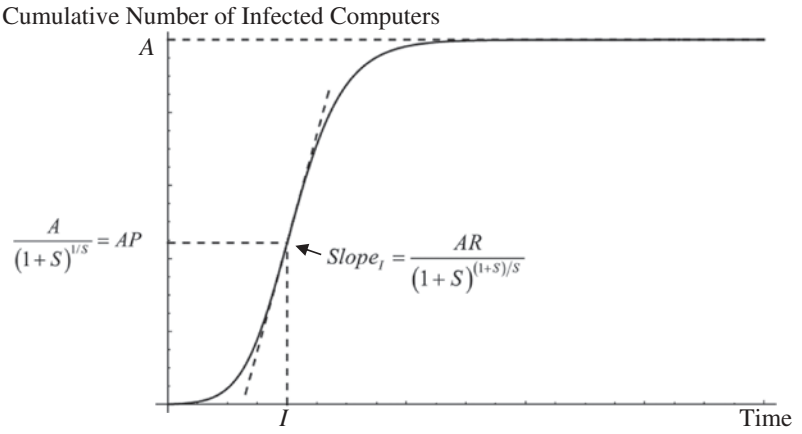


Figure 1. Modeling the Malware Propagation Process

reduces to the logistic function when $S = 1$, the Mitscherlich function when $S = -1$, and the Gompertz function when S approaches zero. This four-parameter generalized logistic function is highly flexible in capturing growth dynamics [38]. First, the point of inflection may occur *anywhere* in the process. Second, the process before the point of inflection may be either symmetric or nonsymmetric to the process after the point of inflection. Third, the proportion of the asymptote obtained at the point of inflection may be *anywhere* between 0 and 1.

The four parameters asymptote (A), point of inflection (I), rate (R), and infection proportion at inflection (P) for each instance serve as dependent variables in our model. These parameters are fundamental characteristics of a malware propagation process providing crucial knowledge about the spread of malware for security managers to make informed decisions and take timely actions in a malware incident.

Random-Walk Betweenness as Individual-Level Independent Variables

In this subsection, we propose a structural measure for the starting nodes in a malware propagation process and use it as an explanatory variable to describe the dynamics of malware propagation. Centrality reveals how influential and powerful a node is in a network. As one of the most fundamental network concepts, centrality has been examined extensively in social network analysis and many centrality measures have been proposed, such as degree, shortest-path betweenness,[2] closeness [25], flow betweenness [24], random-walk betweenness [46], and eigenvector [4] centralities. These centrality measures are developed with specific assumptions and restrictions. To correctly capture the central position of individual nodes and obtain meaningful results, appropriate centrality measures should be chosen based on the characteristics of network flows [7, 24, 54].

We next determine the appropriate centrality measure for malware propagation processes based on their network flow characteristics. First, malware propagation processes follow unconstrained walks, as opposed to trails, paths, or geodesics. Second, malware propagation processes allow parallel duplication, as opposed to serial duplication or transfer. Third, previously infected nodes may get reinfected before security patches have been applied. Fourth, infected nodes attempt to spread to their neighbors simultaneously. Based on these characteristics, random-walk betweenness [46] is appropriate for malware propagation processes.

Random-walk betweenness of a node equals the number of times that a random walk passes through the node along the way, averaged over all starting points and ending points [46]. Random-walk betweenness assumes that a network flow wanders at random until it finds its target, and counts all paths without any assumption of optimality such as geodesic paths. Nodes with high random-walk betweenness are powerful in influencing the dynamics of malware propagation across the network. In our structural risk model, random-walk betweenness is an individual-level independent variable.

Subgroup Analysis and Group-Level Independent Variables

Cohesive subgroups embedded in networks are important for the study of malware propagation. Nodes within a subgroup have more connections among each other and therefore are more likely to infect each other, whereas nodes between the subgroups have fewer connections and therefore are less likely to infect each other.[3] Traditional subgroup analysis techniques include graph partitioning, component analysis, clique analysis, core analysis, and so on. More recently, the concept of modularity, which measures the degree of variation from random network partition, has been proposed for subgroup analysis [27, 48].

Modularity has been shown to be a good indicator of the quality of network partition. New algorithms based on modularity have been developed to improve the performance of subgroup analysis [16, 27, 47, 48]. The fast algorithm proposed in [47] has become a widely used subgroup analysis that generates excellent results within a reasonable amount of time. This fast algorithm is a greedy agglomerative heuristic technique that joins pairs of communities iteratively in order to find the network partition that maximizes modularity. At the beginning, each node is considered a community. In each iteration, community pairs that would result in greatest increase or smallest decrease in modularity are joined and modularity values for the remaining communities are recalculated after each join.

In this study, we apply the fast algorithm proposed in [47] to the network to divide the nodes into subgroups. For each malware incident, we define a group-level independent variable—the average size of the subgroups containing starting nodes (*GroupSize*).

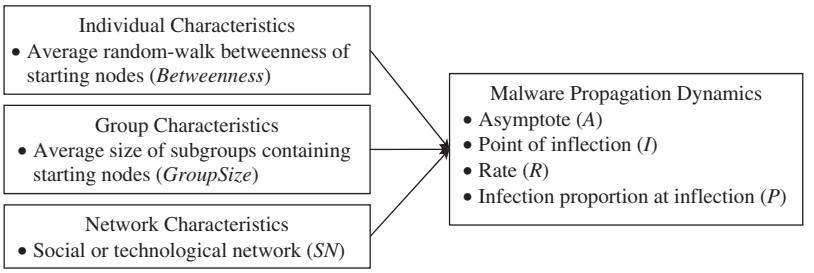


Figure 2. Structural Risk Model

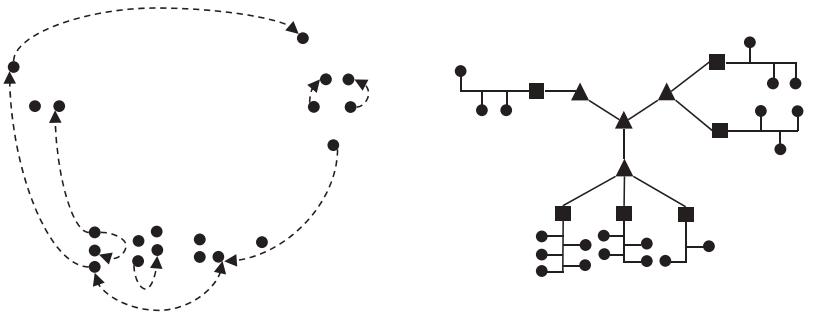
A Structural Risk Model of Self-Replicating Malware Propagation

Based on the previous three subsections, we propose a structural risk model for analyzing the propagation process of self-replicating malware. As shown in Figure 2, four critical parameters (asymptote, point of inflection, rate, and infection proportion at inflection) constitute the dependent variables in our research model. Key independent variables consist of one individual-level structural characteristic (random-walk betweenness), one group characteristic (group size), and one network characteristic (network type).

Research Sample

Social Network

In this study, we first collect social network data from a large social networking site—MySpace. Our research sample is a large-scale organizational social network on MySpace. The sample organization is one of the largest research universities in the United States with a total enrollment of approximately 50,000 students. We gathered data for all members on MySpace who were current students at this university. Among the 27,608 MySpace users affiliated with this university, there were 12,101 private users and 15,507 public users. Since the web pages of the private users were not available, this study analyzes only the public users. After removing 231 public users with invalid user IDs,[4] the number of users in our sample is 15,276 as of March 2008. Following prior studies in measuring user behavior on social networking sites [11, 30, 40, 63, 66], the relationship from one student member to another on MySpace is uncovered by mining the detailed social networking data on friend listing and communications published on each member's profile. Both friend listings and profile comments are directed. For example, a user may list another user as a friend while the other user does not return the gesture; one user may view and comment on another user's profile without any reciprocation from the other user. Figure 3a illustrates the resulting social network, represented by a directed graph. Note that there are other students who are not MySpace



a. Social Network *b*. Technological Network

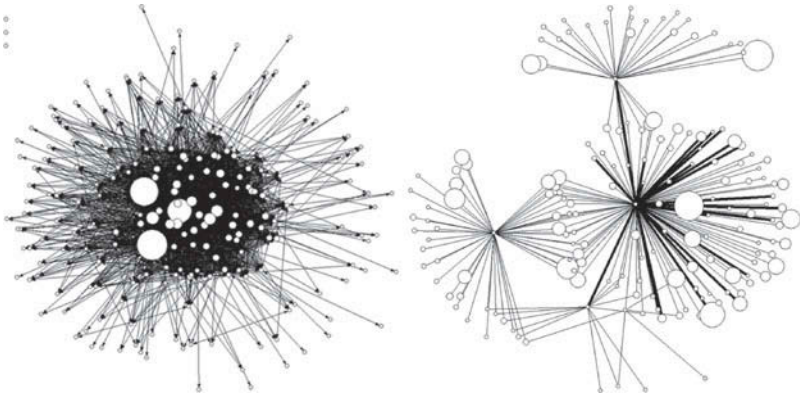
Figure 3. Topologies of an Illustrative Social Network and Technological Network
Notes: Dots represent individual nodes; triangles represent core nodes in the technological network; and squares represent major-of-study nodes in the technological network.

members and there are other social interaction tools that are not reflected on MySpace. In this paper, the constructed social network serves as a proxy of the social network of the sample organization.

Technological Network

Organizations adopt heterogeneous computing environments that involve different technological networks, the most popular being local area networks (LANs). These technological networks use different types of topologies. The three most common topologies are the star, ring, and bus. Ethernets with bus topology dominate the LAN technology application. The sample university's technological network has a typical bus topology for its local area networks which are then linked to the core network forming a tree topology. Figure 3b illustrates the resulting technological network, represented by an undirected graph. We then mapped nodes in the social network to the technological network according to the physical location of the department that hosts a given major-of-study on the technological network. Because 5,168 students out of the total 15,276 students do not reveal their major online, we remove these 5,168 nodes from our analysis. The resulting social network consists of 10,108 individual nodes. These individual nodes, combined with 15 core nodes of the campus network and 168 major-of-study nodes, constitute the technological network with a total of 10,291 nodes. A square in Figure 3b represents a major-of-study node in a building connected to the core network where the local networks in each building naturally follow the bus topology. The 168 LANs are completely connected networks that represent a worst case scenario for the exploration of malware epidemics.

The sample social and technological networks are also visualized in Figures 4a and 4b. The circles represent subgroups in social and technological networks. The size of the circles indicates the size of the subgroups. The edges represent intergroup



a. Social Network *b*. Technological Network

Figure 4. Sample Social Network and Technological Network
Notes: The circles in represent subgroups in social network and technological network. The size of the circles indicates the size of the subgroups. The edges represent intergroup

connections, with the weight of the edges indicating the relative frequency of connections between subgroups. Figures 5a and 5b are drawn using NetDraw [5]. The layouts of the figures demonstrate the subgroup structure of the networks. Compared to the social network, the technological network has much fewer intergroup connections and the sizes of its subgroups are more balanced.

Analysis and Results

The research methodology used in this study is outlined in Figure 5. After constructing the social and technological networks, we simulate the process of self-replicating malware propagation. We then conduct network analysis to compute our key structural measures as independent variables and estimate the four parameters in the generalized logistic growth curve as dependent variables. Finally, our proposed structural risk model is estimated using hierarchical regression. Details of these steps are discussed in the following subsections.

Simulation of Self-Replicating Malware Propagation

Viruses and worms are two of the most common self-replicating malwares. The key difference in their propagation mechanisms is that viruses need to be activated by the users to propagate, whereas worms can propagate without the need for users' initiation. In this study, we simulate the two different propagation processes of worms and viruses in both the social network and the technological network constructed from empirical data.

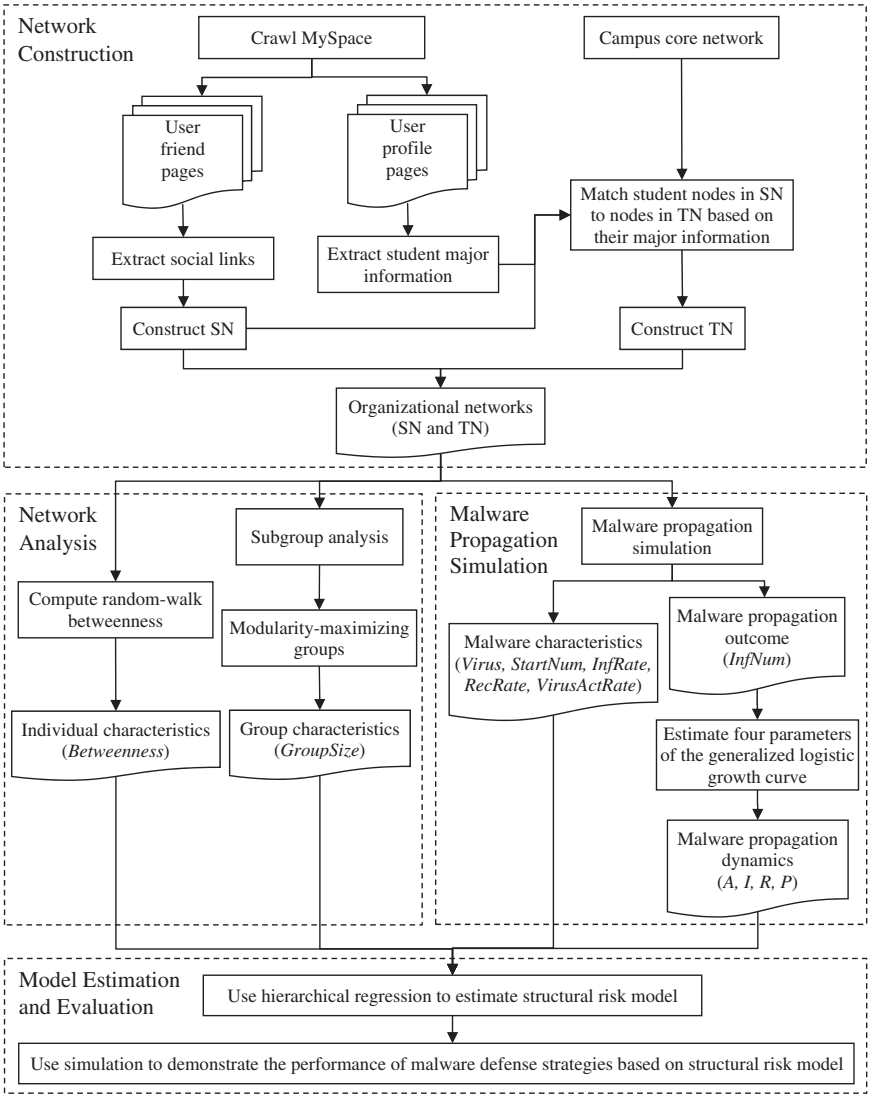


Figure 5. Overview of Research Methodology

Malware propagation has been widely studied using epidemiology models. Among these epidemiology models, the SIR (susceptible–infected–recovered) model is most commonly used. Researchers conduct computer simulations to analyze the malware propagation process [23, 35, 60]. Following this tradition, we start our analysis with simulating the worm propagation process. Based on the SIR model, there are three states for each node in the network. The node can be susceptible, infected, or recovered. A susceptible node is not infected but susceptible to malware and can be infected by its neighbors. An infected node i can infect its neighbor j according to j's infection probability α. After trying to

infect its neighbors, the infected node i may be recovered according to its recovery probability γ. If the infected node i is recovered, then it becomes immune to future infections. In practice, we consider an infected node as recovered when the malware is eliminated from the computer by the user through patching. Every infected node can try to infect its neighbors at all times before it is recovered. Viruses, on the other hand, need users' interaction to spread. We introduce a new simulation parameter, activation rate φ, to account for the probability that a particular user activates a received virus. Only when a node activates a received virus, will the node be infected and the virus propagate to its neighbors.

We employed the discrete-time simulation method to model the malware propagation process. Beginning at time 0, a set of randomly chosen nodes become infected and these nodes start the malware propagation process. The number of starting nodes is modeled as a random variable that follows a power-law distribution. We consider a general form of the probability density distribution of power function: $f(x) = \frac{k(x-a)^{k-1}}{(b-a)^k}$, where a and b are boundary parameters with $a < b$, and k is the exponent. The parameters of the power-law distribution are estimated using empirical data from the Wild List.[5] The Wild List is a monthly report of malware in the wild. There were 5,787 malware reported on the Wild List between January 1996 and June 2011. The Wild List data show diversified patterns of malware incidents. We use the number of reporting organizations as a proxy of the popularity of the malware. The numbers of initial reporting organizations for all reported malware are then used to estimate the distribution of the percentage of starting nodes. Finally, we convert the percentages to the numbers of starting nodes in the sample. The general form of power function fits this empirical data well, yielding estimated parameter values of $k = 0.036$, $a = 2$, and $b = 60$ for the distribution of the number of starting nodes.

The propagation process stops either when the malware stops spreading, that is, when the number of currently infectious nodes reduces to 0, or when the process runs long enough and reaches the maximum time epoch, which we regard as $T = 100$.[6] We assume that the three rates in the malware simulation—infection rate (α), recovery rate (γ), and activation rate (φ), follow the power-law distribution with the lower bound $a = 0$ and the upper bound $b = 1$. The power-law distribution captures the asymmetric nature of user behaviors. Most of the users have high infection rate and recovery rate while only a few of them have low infection rate and recovery rate. Thus the probability density distribution of the power function can be simplified to $f(x) = k\,x^{k-1}$ and the corresponding expected value is $k/(k + 1)$. In order to explore the impact of different malware characteristics, we examine three different levels of the infection rate, recovery rate, and activation rate values. By setting the expected value to .2 (low), .5 (medium), and .8 (high), we get the exponents for infection rate, recovery rate, and activation rate, which are .25, 1, and 4, respectively. For each simulation, we randomly assign (with equal probability) malware type (*Virus* = 1 for virus and 0 for worm) and network type (*SN* = 1 for social network and 0 for technological network). Within each simulation, we also randomly assign (with

equal probability) different levels of infection rate, recovery rate, and activation rate (*InfRate, RecRate, VirusActRate* = .2 for low, .5 for medium, and .8 for high rates) for each node in the network. As a result, the nodes in the simulations are heterogeneous in terms of tgheir *InfRate, RecRate*, and *VirusActRate*. We ran 200,000 malware propagation simulations on clusters hosted by the high performance computing facility at a research university. In other words, 200,000 sets of conditions were randomly selected based on the above parameters; within each of these 200,000 sets of conditions a simulation is run. A summary of the simulation parameters is provided in Table 1. As shown in Table 1, the frequencies of different values for these control variables are approximately the same.

Calculation of Independent Variables

As discussed in the previous subsection, there are four key simulation parameters—number of starting nodes, infection rate, recovery rate, and activation rate for viruses. These four simulation parameters along with the malware type (virus or worm) constitute the control variables in our malware propagation model. We further compute the individual-level structural measure—random-walk betweenness. For each malware incident, the individual-level independent variable is the average random-walk betweenness of the starting nodes (*Betweenness*).

Next we perform subgroup analysis using the fast algorithm proposed in Newman [47]. Our subgroup analysis gives a modularity of 0.402 for SN and 0.948 for TN, indicating significant subgroup structure for both SN and TN.[7] There are 211 SN subgroups and 172 TN subgroups. For each malware incident, we compute the average size of the subgroups that contain the starting nodes (*GroupSize*). The descriptive statistics and correlations for all quantitative variables are reported in Table 2.

Table 1. Summary of Simulation Parameters

Simulation parameter	Parameter value	Number of simulations	Total
Virus	1 (virus)	100,422	200,000
	0 (worm)	99,578	
InfRate	.2 (low)	66,682	200,000
	.5 (medium)	67,010	
	.8 (high)	66,308	
RecRate	.2 (low)	66,670	200,000
	.5 (medium)	66,468	
	.8 (high)	66,862	
VirusActRate	.2 (low)	33,706	100,422
	.5 (medium)	33,148	
	.8 (high)	33,568	
SN	1 (SN)	100,000	200,000
	0 (TN)	100,000	

Table 2. Descriptive Statistics and Correlations for Quantitative Variables

Variable	Mean	SD	StartNum	Betweenness	GroupSize	A	I	R
Number of starting nodes (StartNum)	4.010	7.409	—					
Random-walk betweenness (Betweenness)	.000201	.000108	−.012***	—				
Group size (GroupSize)	250.836	182.630	.001	−.168***	—			
Asymptote (A)	6,106.448	3,458.136	.077***	−.259***	.117***	—		
Point of inflection (I)	6.311	8.267	−.067***	.038***	−.054***	−.045***	—	
Rate (R)	5.937	32.407	.245***	.099***	−.045***	.014***	−.027***	—
Infection proportion at inflection (P)	.448	.179	.096***	.212***	−.109***	.064***	.228***	.414***

Notes: $N = 192{,}728$. ***correlation is significant at the 0.001 level (two-tailed).

Estimation Results of the Generalized Logistic Growth Curve

The estimation procedure of the four-parameter generalized logistic growth curve is implemented in statistical computing software R. We use five methods to estimate the generalized logistic curve parameters for each of the 200,000 incidents: nonlinear least squares (via the *nls* function) and optimization routines (via the *optim* function with the BFGS, Nelder-Mead, CG, and SANN methods). We used the estimated values from the procedure that yielded the minimum value of the standard deviation of the residuals between the model implied and the simulated observations. In order to quantify the fit of the generalized logistic change curve, we compare the observed and model implied (i.e., based on the estimated parameters) values. In particular, we compute a squared multiple correlation coefficient (i.e., coefficient of determination), in which the correlation between y and y-hat, where y is the observed value and y-hat is the estimated value, is squared. We calculate the R^2 values for all 200,000 incidents. Doing so yields a median R^2 of 0.998 and the values that bracketed the middle 98 percent of the distribution are 0.926 and 0.9999998 (i.e., these are the 1st and 99th percentiles).

As shown in Figure 6, the three incidents correspond to the 1st percentile R^2, the median percentile R^2, and the 99th percentile R^2. Considering the 1st percentile of R^2 is as high as .926, it is clear that the four-parameter generalized logistic growth function models the malware propagation process extremely well. We note, however, that these R^2 values exclude 7,272 values (3.64 percent) due to estimated parameter values that were inadmissible such as negative point of inflection. Given that only 3.64 percent of the sample has any issues, we are confident in the validity of our findings based on the 192,728 instances.

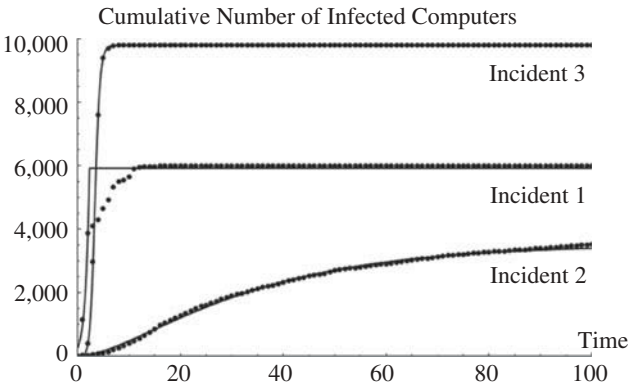


Figure 6. Examples of Fitting Growth Curves to Malware Propagation Processes
Notes: The R^2 of the fitted curve for incident 1 is 0.926, which corresponds to the 1st percentile; the R^2 of the fitted curve for incident 2 is 0.998, which corresponds to the median; and the R^2 of the fitted curve for incident 3 is 0.9999998, which corresponds to the 99th percentile.

Estimation Results of the Structural Risk Model

Based on the empirical network data (SN and TN) of the sample organization and the simulated propagation data of worm and virus, we estimate the proposed structural risk model using hierarchical regression analysis. Specifically, the following four models are estimated:

Model 1: A (or I, R, P) = $b_0 + b_1 X + \varepsilon$;
Model 2: A (or I, R, P) = $b_0 + b_1 X + b_2$ *Betweenness* + ε;
Model 3: A (or I, R, P) = $b_0 + b_1 X + b_2$ *Betweenness* + b_3 *GroupSize* + ε;
Model 4: A (or I, R, P) = $b_0 + b_1 X + b_2$ *Betweenness* + b_3 *GroupSize* + b_4 *SN* + ε,
where control variables X = *Virus, StartNum, InfRate, RecRate*, and *VirusActRate*.

In the first step of the analysis (Model 1), the control variables (*Virus, StartNum, InfRate, RecRate*, and *VirusActRate*) are entered into the regression equation. In the second step (Model 2), the individual-level independent variable (*Betweenness*) is entered into the regression equation, which already contains the Model 1 variables. In the third step (Model 3), we introduce the group-level independent variable (*GroupSize*) into the regression model, which already contains Models 1 and 2 variables. Finally, in the fourth step (Model 4), the network-level independent variable (*SN*) is entered into the regression model that already contains Models 1, 2, and 3 variables. We conduct hierarchical regression analysis on four dependent variables (A, I, R, and P) separately. The unit of analysis in our study is malware incident. Tables 3–6 present the estimation results for the four dependent variables, respectively.

Table 3 presents the step-by-step regression results for the first dependent variable—asymptote (A). The overall model, that is, Model 4, explains 78.4 percent of the variability in A. All independent variables are assessed simultaneously so that their effects can be shown in the context of overall model. The effects of all control variables (*Virus, StartNum, InfRate, RecRate*, and *VirusActRate*) on A are statistically significant. Independent variables result in a .103 increase in R^2 (F-statistic = 30,615.922, $p < .001$). Random-walk betweenness (b = 322,867.027, $p < .001$) as the individual structural characteristic of starting nodes has a statistically significant impact on A with an incremental R^2 of 0.047 (F-statistic = 33,682.462, $p < .001$). Group size, which indicates the average size of groups in which starting nodes are embedded, is significantly related to A (b = .167, $p < .001$). The type of network is also significantly related to A (b = 2,250.361, $p < .001$). These results suggest that the cumulative number of infected computers in a malware incident is positively related to random-walk betweenness of starting nodes and the average size of groups to which starting nodes belong. In addition, malware incidents in a social network results in a higher number of infected computers than in a technological network.

Table 4 presents the step-by-step regression results of control variables and independent variables on the second dependent variable: point of inflection (I). All the control variables (*Virus, StartNum, InfRate, RecRate*, and *VirusActRate*) are statistically significantly associated with I. Independent variables result in a .007 increase in R^2 (F-

Table 3. Regression Results for Asymptote (*A*)

Variables	Model 1	Model 2	Model 3	Model 4
Virus	–6,200.976***	–6,095.638***	–6,099.645***	–6,067.006***
StartNum	38.121***	36.827***	36.868***	36.476***
InfRate	8,268.107***	8,187.241***	8,191.119***	8,162.31***
RecRate	–4,634.266***	–4,560.890***	–4,566.146***	–4,535.366***
VirusActRate	7,337.548***	7,203.103***	7,201.941***	7,163.161***
Betweenness		–6,973,895.645***	–6,533,115.808***	322,867.027***
GroupSize			1.552***	.167***
SN				2,250.361***
Adjusted R^2	.681	.728	.735	.784
F Change compared to Model 1		33,682.462***	1,9626.697***	30,615.922***
F Change compared to Model 2			4,742.283***	24,756.121***
F Change compared to Model 3				43,694.782***

*$p < .05$; **$p < .01$; ***$p < .001$.

statistic = 491.501, $p < .001$) and 8.2 percent of the variability in *I* can be explained by the overall model. Random-walk betweenness ($b = –2{,}460.546$, $p < .001$) has a statistically significant effect on *I*. Both group size ($b = –.001$, $p < .001$) and network type ($b = –1.448$, $p < .001$) are related statistically significantly to *I*. These results suggest that the time point when the maximum slope of growth in the infection number occurs in a malware incident is negatively influenced by random-walk betweenness of starting nodes and the size of their local groups. The maximum slope of growth occurs earlier in a social network than in a technological network.

Regression results for rate (*R*) are presented in Table 5. The effects of all control variables (*Virus, StartNum, InfRate, RecRate*, and *VirusActRate*) on *R* are statistically significant. Random-walk betweenness ($b = 3{,}912.403$, $p < .001$) has a positive significant effect on *R*. Group size ($b = .000$, $p > .05$) is not significantly related to *R*. Network type ($b = –8.643$, $p < .001$) is negatively related to *R* at a significant level. Independent variables result in a .02 increase in R^2 (*F*-statistic = 1,422.642, $p < .001$) and 9.1 percent of the variability in *R* can be explained by the overall model. These results suggest that the propagation speed in a malware incident is affected positively by random-walk betweenness of starting nodes, but the effect of group size was not found to be statistically significant. Malware propagates faster through a technological network than through a social network.

Table 4. Regression Results for Point of Inflection (I)

Variables	Model 1	Model 2	Model 3	Model 4
Virus	3.800***	3.761***	3.766***	3.745***
StartNum	–.075***	–.075***	–.075***	–.075***
InfRate	–.4.391***	–4.361***	–4.366***	–4.348***
RecRate	–6.524***	–.6.551***	–6.543***	–6.563***
VirusActRate	–4.415***	–4.365***	–4.364***	–4.339***
Betweenness		2,581.185***	1,951.583***	–2,460.546***
GroupSize			–.002***	–.001***
SN				–1.448***
Adjusted R^2	.075	.077	.079	.082
F Change compared to Model 1		237.580***	362.818***	491.501***
F Change compared to Model 2			487.456***	617.701***
F Change compared to Model 3				746.062***

*p < .05; **p < .01; ***p < .001.

Table 5. Regression Results for Rate (R)

Variables	Model 1	Model 2	Model 3	Model 4
Virus	–6.107***	–6.585***	–6.572***	–6.697***
StartNum	1.075***	1.081***	1.081***	1.082***
InfRate	9.753***	10.120***	10.107***	10.218***
RecRate	3.124***	2.791***	2.808***	2.690***
VirusActRate	3.495***	4.106***	4.109***	4.258***
Betweenness		31,652.900***	30,244.203***	3,912.403***
GroupSize			–.005***	.000
SN				–8.643***
Adjusted R^2	.071	.082	.083	.091
F Change compared to Model 1		2,338.550***	1,249.966***	1,422.642***
F Change compared to Model 2			159.459***	953.135***
F Change compared to Model 3				1,745.368***

*p < .05; **p < .01; ***p < .001.

Table 6 presents the step-by-step regression results for infection proportion at inflection (P). As shown in Table 6, the effects of all control variables (*Virus, StartNum, InfRate, RecRate*, and *VirusActRate*) on P are significant. Random-walk betweenness (b = –70.856, p < .001) has positive significant effects on P. Group size (b = –.00001251, p < .001) and network type (b = –.138, p < .001) are negatively associated with P at a significant level. Independent variables result in a .124 increase in R^2 (F-statistic = 11,434.002 p < .001) and 30.3 percent of the variability

Table 6. Regression Results for Infection Proportion at Inflection (*P*)

Variables	Model 1	Model 2	Model 3	Model 4
Virus	–.124***	–.130***	–.129***	–.131***
StartNum	.002***	.002***	.002***	.002***
InfRate	.213***	.218***	.217***	.219***
RecRate	.156***	.152***	.152***	.151***
VirusActRate	.129***	.136***	.136***	.138***
Betweenness		369.466***	348.929***	–70.856***
GroupSize			–.0000723***	–.00001251***
SN				–.138***
Adjusted R^2	.179	.229	.234	.303
F Change compared to Model 1		12,477.281***	6,949.497***	11,434.002***
F Change compared to Model 2			1,335.325***	10,248.886***
F Change compared to Model 3				19,030.594***

*$p < .05$; **$p < .01$; ***$p < .001$.

in *P* can be explained by the overall model. These results suggest that the infection proportion at point of inflection in a malware incident is negatively related to random-walk betweenness of starting nodes and the average group size. A lower proportion of computers is infected at the point of inflection if a malware propagates through a social network rather than through a technological network.

In summary, our regression results show that in a malware incident, higher random-walk betweenness of starting nodes is associated with a larger cumulative number of infected computers (*A*), earlier occurrence of the maximum slope of growth in the infection number (*I*), faster malware propagation speed (*R*), and higher infection proportion at point of inflection (*P*). When malware starts from nodes in different groups, malware propagation dynamics vary significantly. Starting nodes from larger groups result in larger cumulative number of infected computers (*A*), earlier time point when the maximum slope of growth in the infection number occurs (*I*), and lower infection proportion at point of inflection (*P*). Network type has significant explanatory power for all four measures of malware propagation dynamics. Network type explained 4.9 percent of the variability in *A*, 0.3 percent of the variability in *I*, 0.8 percent of the variability in *R*, and 6.9 percent of the variability in *P*. Everything else being equal, a malware incident through a social network results in 13.8 percent fewer infected computers at the point of inflection and 2,250 more total infected computers than through a technological network. In addition, malware propagates more slowly and the maximum propagation speed occurs earlier in a social network than in a technological network.

Malware Defense Strategies Based on Structural Risk Model

In this section, we investigate how our proposed structural risk model can be integrated into different malware defense strategies and evaluate the effectiveness

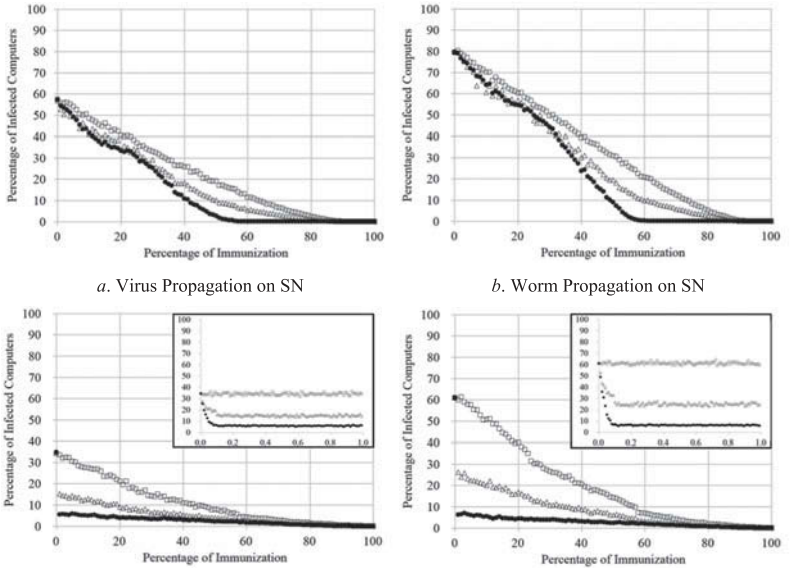
of these augmented strategies through simulations. Specifically, we conduct experiments to simulate three common defense strategies: preselected immunization strategies, countermeasure dissemination strategies, and security awareness programs based on structural risk model.

Preselected Immunization Strategies Based on Structural Risk Model

In this subsection, we simulate the malware propagation processes with preselected immunization strategies. Under a preselected immunization strategy, a certain percentage of preselected nodes adopt countermeasures (such as system updates and security patches), but these nodes do not further spread countermeasures. We compare the targeted immunization strategy based on structural risk identified using our research model to two common existing immunization strategies [14, 61]: the random immunization strategy and the targeted immunization strategy based on degree centrality. The simulation results of these three different immunization strategies are plotted in Figure 7, with the percentage of immunization varying from 0.1 percent to 100 percent. In general, structural-risk-based targeted immunization outperforms random immunization and degree-based targeted immunization in both social and technological networks, in terms of reducing the percentage of infected computers. In social networks as shown in Figures 7a and 7b, at the same immunization percentage, the size of infection under structural-risk-based targeted immunization is lower than that under random immunization and degree-based targeted immunization. To protect almost all the computers, only 55 percent of computers need to be immunized based on structural risks, whereas 90 percent of computers need to be immunized under random immunization and 80 percent for degree-based targeted immunization. In technological networks as shown in Figures 7c and 7d, when computers are randomly immunized at a very small percentage (for example, .1 percent of total computers are immunized), around 35 percent of the total computers will be infected in a virus incident, and 60 percent will be infected in a worm incident. In contrast, when the same immunization percentage is applied to computers that have high structural risks, the percentage of infected computers will drop below 10 percent.

Countermeasure Dissemination Strategies Based on Structural Risk Model

In contrast to a preselected immunization strategy, under a countermeasure dissemination strategy, only a small percentage of preselected nodes are immunized and these nodes are able to further spread countermeasures. Among common countermeasure dissemination strategies, the countermeasure competing strategy has been shown to be more realistic and more effective than other strategies [13, 14, 28]. Under the countermeasure competing strategy, countermeasures and malware propagate through separate but interlinked networks, countermeasures spread to both infected and susceptible nodes, and the receiving nodes probabilistically adopt countermeasures.



a. Virus Propagation on SN

b. Worm Propagation on SN

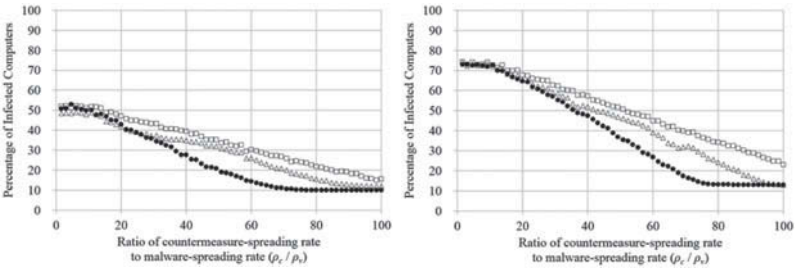
c. Virus Propagation on TN

d. Worm Propagation on TN

□ = Random immunization; Δ = Degree-based targeted immunization;
● = Structural-risk-based targeted immunization.

Figure 7. Preselected Immunization Strategies

Following prior studies in the countermeasure competing strategy [13, 14, 28], we allow malware and countermeasures to spread through different networks: malware spread through SN only and countermeasures spread through both SN and TN. We use κ to represent countermeasure adoption rate, meaning that when a node receives the countermeasure, the probability of adopting the countermeasure is κ. We denote the infection probability for countermeasures as λ and the recovery probability for countermeasures as δ. We define $\rho_c = \lambda/\delta$ as the countermeasure-spreading rate. Similarly, we define $\rho_v = \alpha/\gamma$ as the malware-spreading rate. Recall that α and γ are the infection and recovery probabilities for malware. We simulate three different countermeasure competing strategies based on different methods of selecting starting nodes: countermeasure competing strategies with random, degree-based, and structural-risk-based starts. As shown in Figure 8, we investigate how these strategies perform with the ratio of countermeasure-spreading rate to malware-spreading rate (ρ_c/ρ_v) varying from 1 to 100 and countermeasure adoption rate κ fixed at 0.1. The simulation experiments show that starting the spread of countermeasures based on structural risk identified in this study helps to mitigate the size of infection and this mitigation effect is more salient for a higher ratio of countermeasure-spreading rate to malware-spreading rate. This result demonstrates the importance of integrating structural characteristics into the dissemination strategies of countermeasures.



a. Virus Propagation *b*. Worm Propagation

□ = Countermeasure competing strategy with random start;
Δ = Countermeasure competing strategy with degree-based start;
● = Countermeasure competing strategy with structural-risk-based start.

Figure 8. Countermeasure Dissemination Strategies

Security Awareness Programs Based on Structural Risk Model

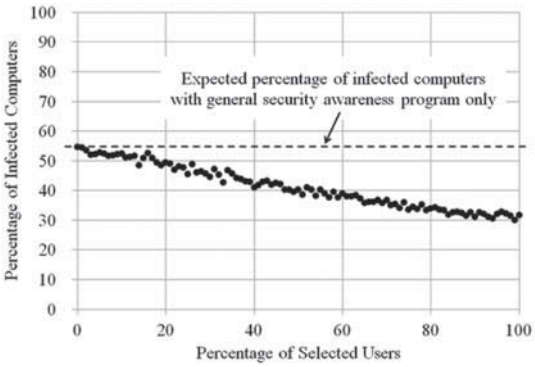
Security awareness programs inform users about proper use of information technology (IT) systems and information within an organization and provide them guidance about malware incident prevention. Organizations often offer general security awareness programs to all users. In this subsection, in order to examine the impact of offering security awareness programs on the size of malware infection, we simulate virus propagation processes through a social network and consider that security awareness programs help to reduce individual users' probability of activating the virus (i.e., activation rate φ) because of their acquired knowledge and vigilance. For example, after the deployment of a security awareness program, users are less likely to click the malicious attachment in an e-mail.

We use the case of offering a general security awareness program only to all users as a benchmark. We then investigate the impact of offering an additional security awareness program to selected users based on their structural risks with varying percentages of selected users. We assume that the general security awareness program reduces all users' activation rate φ by 10 percent and the additional security awareness program reduces the selected users' activation rate φ by 20 percent. As shown in Figure 9, when the additional security awareness program is offered to a small group of selected users based on their structural risks, the size of infection is significantly reduced.

Conclusions

Discussion

This paper takes a social network analysis perspective to examine the malware propagation problem within organizations consisting of both a social network and



● = Security awareness programs with elected users based on structural risk

Figure 9. Security Awareness Programs

technological network. Organizations' computing environment is viewed as networks where users and their computers are nodes in the networks. Edges in an organizational social network correspond to social communications such as e-mails and profile comments among users, whereas edges in an organizational technological network correspond to physical communications such as transmission of data packets among computers. Malware starts from certain nodes and propagates through the edges in a process that can be considered a dynamic network flow built on the underlying social network and technological network. We propose a novel structural risk model to explain the dynamics of malware propagation. Further, we simulate the malware propagation process using a susceptible–infected–recovered epidemic process and run hierarchical regression models to statistically test and quantify the impact of the network structures on malware propagation dynamics.

Although malware propagation exhibits an *S*-shaped growth pattern similar to other diffusion processes such as product diffusion, our proposed structural risk model establishes the important connection between network structural characteristics and the observed malware propagation pattern. Our findings suggest that nodes with larger random-walk betweenness that are embedded in larger groups are the ones with higher structural risk in terms of higher cumulative number of infected computers, earlier time point when the maximum slope of growth in the infection number occurs, and faster malware propagation speed. However, when malware starts from nodes in larger groups, we find that the proportion of total number of infected computers at the point of inflection is lower. A potential reason is that when a malware incident occurs within larger groups, due to the rapid initial spread, the occurrence of maximum growth is so early that a smaller portion of total infected computers is reached.

This study finds that the dynamics of malware propagation differ between social network and technological network. Prior studies [49] found empirical evidence of the structural differences between social networks and technological networks. For example, social networks usually demonstrate positive degree correlation (also

called assortative mixing) while most technological networks reveal negative degree correlation. Other studies [65] have identified unique activity patterns of users in online social networks. In this study, we are unable to verify what differences in network structure and/or user behavior patterns between social and technological networks account for their distinct impacts on malware propagation dynamics. Although we are unable to directly address the reasons for the differences, we do observe that there are many more total technological links than total social links, but many fewer intergroup technological links than intergroup social links within the sample organization. As a result, the mean node-to-node distance in a social network is shorter. Compared to a technological network, the sparser within-group connectivity in a social network may have led to slower propagation speed and lower infection proportion at point of inflection, while the larger intergroup connectivity may have led to a larger total number of infected computers and earlier maximum growth. One interesting direction for future research is to identify the factors that differentiate social networks and technological networks, and investigate how these factors explain the different effects of social networks and technological networks on malware propagation dynamics.

Our analysis is based on a specific social network and technological network. Since our sample social network (from a social networking site) is a typical online social network and our sample technological network (local area networks) is a widely used computer network, conclusions drawn from this specific social network and technological network can be applied to other similar networks. We realize there are many different types of social/technological networks, for example, connections through removable storage devices, to which our findings may not be directly generalizable. However, it is worth noting that the methodology proposed in this work can be generalized to other network-based processes, such as network-based diffusion of innovations and word-of-mouth information cascades.

In addition, our structural risk model and simulations are based on starting nodes, which are instrumental in malware propagation dynamics and various malware defense strategies such as countermeasure dissemination strategies. However, we acknowledge that structural risks may not necessarily be related to starting nodes. For example, in a malware incident, containment strategies aim to stop the spread of the malware and prevent further damage. Since containment actions take place after a malware incident occurs, containment strategies should focus on disconnecting all infected hosts from the network instead of just the starting nodes. Hence our analyses and findings do not apply to these cases.

Another limitation of this study is that we treat the networks (both SN and TN) as static. However, real-life networks evolve over time; for example, friend connections change over time. Due to data constraints in this study, we are unable to examine the impact of the dynamics of evolving networks. In future research, it would be interesting to study two interrelated processes—the malware propagation process and the evolving process of the underlying networks—and to examine the impact of the network dynamics on the malware propagation process.

Managerial Implications

Our findings provide useful managerial implications for malware prevention and handling strategies, as well as other information systems security decisions. In a malware incident, if the IT manager considers releasing countermeasures to reduce the size of malware infection to the minimum and ensure timely action before the maximum epidemic outbreak, the optimal starting nodes to implement the countermeasures are the ones with high individual-level random-walk betweenness that are embedded in larger cohesive subgroups. Specifically, social networking data can be used to construct the organizational social network, which can then be mapped to the technical network within the same organization. Countermeasures may start from the selected nodes and spread through both social and technological networks to achieve optimal performance. As an important malware prevention strategy, security awareness programs can be deployed. Our results demonstrate the benefit of offering additional security awareness programs for a small group of selected users with higher structural risks.

One important reason for the prevalence of malware is the highly homogeneous computing environment in organizations. When the same software is installed on multiple nodes on the network, correlated failure of these nodes may occur due to shared vulnerabilities. Software diversification strategies have been proposed [12, 59] to mitigate the risk of such correlated failure. Our findings suggest that individual nodes' structural measures and the subgroups they belong to should be incorporated into organizations' software diversification strategies. It is advisable to install heterogeneous operating systems and other critical software across nodes with high random-walk betweenness. It is also beneficial to diversify within large subgroups such that they are effectively broken into smaller subgroups with nodes sharing the same software vulnerability.

To guard against malware threats, organizations can adopt multiple layers of protection such as general-purpose and system-specific protection for internal and external attacks [68]. The consideration of structurally risky nodes is important in the installation strategy of both general-purpose and system-specific protection measures. In order to determine the optimal layered protection, experiments can be conducted to simulate the impact of multiple layers of protection simultaneously while taking into account structural risks of individual nodes.

Finally, accurate assessment of security risk is crucial for effective IT security risk management and IT security investment. As an important component of security risk, structural risk is tied to the idea of system interdependency and network externality because of the interconnectivity of assets (e.g., computers, user accounts) within organizations. Nodes with higher structural risk impose higher negative network externality on other nodes, meaning that once these nodes are compromised, they put many other nodes in danger. Our study identifies specific measures (random-walk betweenness and size of subgroups) to quantify structural risk, which captures such network externality due to system interdependency. We demonstrated empirically that not all nodes are of equal importance when attempting to secure a network. Hence, nodes with higher structural risk should be given priority when making IT security investment decisions.

Acknowledgment: This research benefitted from the Notre Dame Deloitte Center for Ethical Leadership.

Notes

1. Real malware infection data is infeasible to obtain due to the destructivity of field experiments of malware propagation [2] and the privacy issues of internal technological and social network information [15] within real organizations.

2. Freeman's betweenness is often referred to as simply betweenness. In this paper, we refer to Freeman's betweenness as shortest-path betweenness to distinguish it from other betweenness measures.

3. In social network analysis, subgroups are also referred to as communities, clusters, and so on. Although there is no consensus on the name, the essential concept is a partition of the nodes in one network into multiple subsets and connections within the subsets are dense while connections between the subsets are sparse.

4. Although not directly verified, we believe that these invalid IDs are due to the inconsistency in MySpace's databases. These 231 user IDs have been listed as current students at the sample university. However, MySpace shows that these 231 user IDs are invalid when the crawler tries to access these users' profile/friend/blog pages.

5. See WildList Archive (www.wildlist.org/WildList/t_archive.htm), maintained by The WildList Organization International.

6. Following prior studies in computer simulations of malware propagation [61], we set the maximum time epoch to 100, beyond which the malware propagation process generally stabilizes and the cumulative number of infected computers approaches the upper limit.

7. A modularity value above 0.3 is suggested to be a good indicator for significant subgroup structure [48].

References

1. Baltazar, J.; Costoya, J.; and Flores, R. *The Real Face of Koobface: The Largest Web 2.0 Botnet Explained*. Trend Micro Research, 2009. www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp_the-real-face-of-koobface.pdf.

2. Benzel, T.; Braden, R.; Kim, D.; Neuman, C.; Joseph, A.; Sklower, K.; Ostrenga, R.; and Schwab, S. Design, deployment, and use of the DETER testbed, *Proceedings of the DETER Community Workshop on Cyber-Security and Test*. Boston, MA: USENIX, 2007.

3. Bonacich, P. Power and centrality: A family of measures. *American Journal of Sociology*, *92*, 5 (1987), 1170–1182.

4. Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, *2*, 1 (1972), 113–120.

5. Borgatti, S. P. *NetDraw: Software for Network Visualization*. Lexington, KY: Analytic Technologies, 2002.

6. Borgatti, S.P.; Mehra, A.; Brass, D.J.; and Labianca, G. Network analysis in the social sciences. *Science*, *323*, 5916 (2009), 892–895.

7. Borgatti, S.P. Centrality and network flow. *Social Networks*, *27*, 1 (2005), 55–71.

8. Cavusoglu, H.; Raghunathan, S.; and Yue, W.T. Decision-theoretic and game-theoretic approaches to IT security investment. *Journal of Management Information Systems*, *25*, 2 (2008), 281–304.

9. Cavusoglu, H.; Cavusoglu, H.; and Zhang, J. Security patch management: Share the burden or share the damage? *Management Science*, *54*, 4 (2008), 657–670.

10. Cavusoglu, H.; Mishra, B.; and Raghunathan, S. The value of intrusion detection systems in information technology security architecture. *Information Systems Research*, *16*, 1 (2005), 28–46.

11. Chen, A.; Lu, Y.; Chau, P.Y.K.; and Gupta, S. Classifying, measuring, and predicting users' overall active behavior on social networking sites. *Journal of Management Information Systems*, *31*, 3 (2014), 213–253.

12. Chen, P.; Kataria, G.; and Krishnan, R. Correlated failures, diversification, and information security risk management. *MIS Quarterly*, *35*, 2 (2011), 397–422.

13. Chen, P.; Cheng, S.; and Chen, K. Optimal control of epidemic information dissemination over networks. *IEEE Transactions on Cybernetics*, *44*, 12 (2014), 2316–2328.

14. Chen, L.C., and Carley, K.M. The impact of countermeasure propagation on the prevalence of computer viruses. *IEEE Transactions on Systems, Man and Cybernetics—Part B: Cybernetics*, *34*, 2 (2004), 823–833.

15. Chi, M. *Reducing the Risks of Social Media to Your Organization*. Bethesda, MD: SANS Institute, 2011.

16. Clauset, A.; Newman, M.E.J.; and Moore, C. Finding community structure in very large networks. *Physical Review E*, *70*, 6 (2004), 066111.

17. Computer Economics. *Malware Report: The Economic Impact of Viruses, Spyware, Adware, Botnets, and Other Malicious Code*. Irvine, CA: Computer Economics, 2007.

18. Computer Security Institute. *The Fifteenth Annual CSI Computer Crime and Security Survey*. Monroe, WA: Computer Security Institute, 2010.

19. Consumer Reports. Social insecurity: What millions of online users don't know can hurt them. *Consumer Reports*, 2010.

20. D'Arcy, J.; Hovav, A.; and Galletta, D. User awareness of security countermeasures and its impact on information systems misuse: A deterrence approach. *Information Systems Research, 20* (2009), 79–98.

21. Dezső, Z., and Barabási, A. Halting viruses in scale-free networks. *Physical Review E*, 65, 5 (2002), 055103.

22. *Economist*. A thing of threads and patches. *Economist*, August 25, 2012.

23. Fleizach, C.; Liljenstam, M.; Johansson, P.; Voelker, G.M.; and Mehes, A. Can you infect me now? Malware propagation in mobile phone networks. *Proceedings of the 2007 ACM Workshop on Recurring Malcode*. New York: ACM, 2007, pp. 61–68.

24. Freeman, L.C.; Borgatti, S.P.; and White, D.R. Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, *13*, 2 (1991), 141–154.

25. Freeman, L.C. Centrality in social networks: Conceptual clarification. *Social Networks*, *1*, 3 (1979), 215–239.

26. Garetto, M.; Gong, W.; and Towsley, D. Modeling malware spreading dynamics. *Twenty-Second Annual Joint Conference of the IEEE Computer and Communications (INFOCOM 2003)*. San Francisco: IEEE, 2003, pp. 1869–1879.

27. Girvan, M., and Newman, M.E.J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 12 (2002), 7821–7826.

28. Goldenberg, J.; Shavitt, Y.; Shir, E.; and Solomon, S. Distributive immunization of networks against viruses using the "honey-pot" architecture. *Nature Physics, 1*, 3 (2005), 184–188.

29. Gordon, L.A., and Loeb, M.P. The economics of information security investment. *ACM Transactions on Information and System Security*, *5*, 4 (2002), 438–457.

30. Guo, H., Pathak, P., and Cheng, H. K. Estimating social influences from social networking sites: Articulated friendships versus communication interactions. *Decision Sciences*, *46*, 1 (2015), 135–163.

31. Guo, W.; Li, X.; and Wang, X. Epidemics and immunization on Euclidean distance preferred small-world networks. *Physica A: Statistical Mechanics and Its Applications, 380* (2007), 684–690.

32. Huang, C.; Lee, C.; Wen, T.; and Sun, C. A computer virus spreading model based on resource limitations and interaction costs. *Journal of Systems and Software*, *86*, 3 (2013), 801–808.

33. Karsai, M.; Kivelä, M.; Pan, R.K.; Kaski, K.; Kertész, J.; Barabási, A.; and Saramäki, J. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, *83*, 2 (2011), 025102.

34. Kephart, J.O., and White, S. R. Directed-graph epidemiological models of computer viruses. *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, New York, NY: IEEE, 1991, pp. 343–359.

35. Kim, J.; Radhakrishnan, S.; and Dhall, S.K. Measurement and analysis of worm propagation on Internet network topology. *Proceedings of Thirteenth International Conference on Computer Communications and Networks*. Washington, DC: IEEE Computer Society, 2004, pp. 495–500.

36. Kumar, R.L.; Park, S.; and Subramaniam, C. Understanding the value of countermeasure portfolios in information systems security. *Journal of Management Information Systems*, *25*, 2 (2008), 241–280.

37. Lloyd, A.L., and May, R.M. How viruses spread among computers and people. *Science*, *292*, (2001), 1316–1317.

38. Mahajan, V.; Muller, E.; and Bass, F. M. New product diffusion models in marketing: A review and directions for research. *Journal of Marketing*, *54*, 1 (1990), 1–26.

39. Mansfield-Devine, S. Anti-social networking: Exploiting the trusting environment of web 2.0. *Network Security*, *11* (2008), 4–7.

40. Matook, S.; Cummings, J.; and Bala, H. Are you feeling lonely? The impact of relationship characteristics and online social network features on loneliness. *Journal of Management Information Systems*, *31*, 4 (2015), 278–310.

41. McGowan, I. The use of growth curves in forecasting market development. *Journal of Forecasting*, *5*, 1 (1986), 69–71.

42. Merrill, T.; Latham, K.; Santalesa, R.; and Navetta, D. *Social Media: The Business Benefits May Be Enormous, but Can the Risks—Reputational, Legal, Operational—Be Mitigated?* Zurich, Switzerland: ACE Group, 2011.

43. Moore, T., and Anderson, R. Internet security. In *The Oxford Handbook of the Digital Economy*, ed. J. Waldfogel and M. Peitz. Oxford: Oxford University Press, 2012, pp. 572–600.

44. Moore, C., and Newman, M.E.J. Epidemics and percolation in small-world networks. *Physical Review E*, *61*, 5 (2000), 5678–5682.

45. Nelder, J.A. An alternative form of a generalized logistic equation. *Biometrics*, *18*, 4 (1962), 614–616.

46. Newman, M.E.J. A measure of betweenness centrality based on random walks. *Social Networks*, *27*, 1 (2005), 39–54.

47. Newman, M.E.J. Fast algorithm for detecting community structure in networks. *Physical Review E*, *69*, 6 (2004), 066133.

48. Newman, M.E.J., and Girvan, M. Finding and evaluating community structure in networks. *Physical Review E*, *69*, 2 (2004), 026113.

49. Newman, M.E.J., and Park, J. Why social networks are different from other types of networks. *Physical Review E*, *68*, 3 (2003), 036122.

50. Newman, M.E.J.; Forrest, S.; and Balthrop, J. Email networks and the spread of computer viruses. *Physical Review E*, *66*, 3 (2002), 035101.

51. Park, I.; Sharman, R.; Rao, H.R.; and Upadhyaya, S. Short term and total life impact analysis of email worms in computer systems. *Decision Support Systems, 43* (2007), 827–841.

52. Richards, F.J. A flexible growth function for empirical use. *Journal of Experimental Botany*, *10*, 2 (1959), 290–301.

53. Roff, D.A. *Evolution of Life Histories: Theory and Analysis*. New York: Springer, 1992.

54. Stephenson, K., and Zelen, M. Rethinking centrality: Methods and examples. *Social Networks*, *11*, 1 (1989), 1–37.

55. Straub, D.W., and Welke, R.J. Coping with systems risk: Security planning models for management decision making. *MIS Quarterly*, *22*, 4 (1998), 441–469.

56. Symantec. *The 2012 Norton Cybercrime Report*. Mountain View, CA: Symantec. 2012.

57. von Bertalanffy, L. Quantitative laws in metabolism and growth. *Quarterly Review of Biology*, *32*, 3 (1957), 217–231.

58. Wang, J.; Chaudhury, A.; and Rao, H.R. A value-at-risk approach to information security investment. *Information Systems Research*, *19*, 1 (2008), 106–120.

59. Wang, J.; Sharman, R.; and Zionts, S. Functionality defense through diversity: A design framework to multitier systems. *Annals of Operations Research*, *197*, 1 (2010), 25–45.

60. Wang, Y., and Wang, C. Modeling the effects of timing parameters on virus propagation. *Proceedings of the 2003 ACM Workshop on Rapid Malcode*. New York: ACM, 2003, pp. 61–66.

61. Wang, C.; Knight, J.C.; and Elder, M.C. On computer viral infection and the effect of immunization, *Proceedings of the Sixteenth Annual Computer Security Applications Conference (ACSAC 2000)*. New Orleans, 2000, 898879.

62. Wang, Y.; Chakrabarti, D.; Wang, C.; and Faloutsos, C. Epidemic spreading in real networks: An eigenvalue viewpoint. *Proceedings of the Twenty-Second International Symposium on Reliable Distributed Systems*. Washington, DC: IEEE Computer Society, 2003, pp. 25–34.

63. Xie, K., and Lee, Y. Social media and brand purchase: Quantifying the effects of exposures to earned and owned social media activities in a two-stage decision making model. *Journal of Management Information Systems*, *32*, 2 (2015), 204–238.

64. Xue, L.; Zhang, C.; Ling, H.; and Zhao, X. Risk mitigation in supply chain digitization: System modularity and information technology governance. *Journal of Management Information Systems*, *30*, 1, (2013), 325–352.

65. Yan, G.; Chen, G.; Eidenbenz, S.; and Li, N. Malware propagation in online social networks: Nature, dynamics, and defense implications, *Proceedings of the Sixth ACM Symposium on Information, Computer and Communications Security.* New York: ACM, 2011, pp. 196–206.

66. Yu, J.; Hu, P.J.; and Cheng, T. Role of affect in self-disclosure on social network websites: A test of two competing models. *Journal of Management Information Systems*, *32*, 2 (2015), 239–277.

67. Yue, W.T., and Çakanyildirim, M. Intrusion prevention in information systems: Reactive and proactive responses. *Journal of Management Information Systems*, *24*, 1 (2007), 329–353.

68. Yue, W.T.; Çakanyıldırım, M.; Ryu, Y.U.; and Liu, D. Network externalities, layered protection and IT security risk management. *Decision Support Systems*, *44*, 1 (2007), 1–16.

69. Zhao, K.; Kumar, A.; Harrison, T.P.; and Yen, J. Analyzing the resilience of complex supply network topologies against random and targeted disruptions. *IEEE Systems Journal*, *5*, 1 (2011), 28–39.

70. Zhao, X.; Xue, L.; and Whinston, A.B. Managing interdependent information security risks: Cyberinsurance, managed security services and risk pooling arrangement. *Journal of Management Information Systems*, *30*, 1 (2013), 123–152.

71. Zhao, X.; Fang, F.; and Whinston, A.B. An economic mechanism for better Internet security. *Decision Support Systems*, *45*, 4 (2008), 811–821.

72. Zou, C.C.; Towsley, D.; and Gong, D.W. Email worm modeling and defense. *Proceedings of the Thirteenth International Conference on Computer Communications and Networks*. New York, NY: IEEE, 2004, pp. 409–414.