NET Institute*

www.NETinst.org

Working Paper #08-24

October 2008

Computer Virus Propagation in a Network Organization: The Interplay between Social and Technological Networks

Hsing Kenny Cheng and Hong Guo
Warrington College of Business Administration
University of Florida

* The Networks, Electronic Commerce, and Telecommunications ("NET") Institute, http://www.NETinst.org, is a non-profit institution devoted to research on network industries, electronic commerce, telecommunications, the Internet, "virtual networks" comprised of computers that share the same technical standard or operating system, and on network issues in general.

Computer Virus Propagation in a Network Organization: The Interplay between Social and Technological Networks[1]

Hsing Kenny Cheng
Department of Information Systems and Operations Management
Warrington College of Business Administration
The University of Florida, Gainesville, FL 32611-7169
hkcheng@ufl.edu

Hong Guo
Department of Information Systems and Operations Management
Warrington College of Business Administration
The University of Florida, Gainesville, FL 32611-7169
guohong@ufl.edu

This paper proposes a holistic view of a network organization's computing environment to examine computer virus propagation patterns. We empirically examine a large-scale organizational network consisting of both social network and technological network. By applying information retrieval techniques, we map nodes in the social network to nodes in the technological network to construct the composite network of the organization. We apply social network analysis to study the topologies of social and technological networks in this organization. We statistically test the impact of the interplay between social and technological network on computer virus propagation using a susceptible-infective-recovered epidemic process. We find that computer viruses propagate faster but reach lower level of infection through technological network than through social network, and viruses propagate the fastest and reach the highest level of infection through the composite network. Overlooking the interplay of social network and technological network underestimates the virus propagation speed and the scale of infection.

Key words: social network analysis, interplay between social and technological networks, computer viruses

[1] We gratefully acknowledge financial support from the NET Institute (www.netinst.org) and the Kauffman Foundation.

1. Introduction

Network organizations rely on various interconnected networks to achieve higher operational efficiency and more flexible management (Sproull and Kiesler 1991, Van Alstyne 1997). Business partners including different manufacturers, suppliers, and distributors form a supply chain network in order to reach their customers. Business communications as well as personal contacts among employees inside and outside their departments constitute an information distribution network. Even individuals from decentralized geographical locations can cooperate with each other through virtual teams creating a virtual collaboration network. These are examples of social networks inherently embedded within an organization. As the computational foundation of these business processes, technological networks consisting of interconnected computers, routers, and other network devices enable the data transmissions to perform required organizational tasks.

Social networks and technological networks coexist in an organization. One salient difference between social networks and technological networks is their structural topologies. In particular, empirical evidence shows that social networks usually demonstrate non-trivial clustering and positive degree correlation (also called assortative mixing) while most technological networks reveal lower level of clustering and negative degree correlation (Newman and Park 2003). These diversified structural features of social networks, technological networks, and the composite organizational networks have significant influences on the organizations' operational processes.

While the emerging networked organizational structure increases operational efficiency tremendously, it also serves as a more vulnerable channel for malware propagation. Malware, malicious software written to cause damage to infected computers, has evolved dramatically

since the first personal computer virus "Brain" surfaced in early 1986. Brain, a boot sector virus spreading through infected floppy disks, didn't spread quickly. Nor did it cause much harm. However, by showing how destructive self-replicating programs could do, Brain heralded a new era of more devastating computer viruses. Computer viruses pose a critical threat to computer users and organizations causing massive expenses in damages. It is estimated by Computer Economics that the total worldwide financial losses from malware are on average $12.18 billion per year in the period from 1999 to 2006 (Computer Economics 2007). With the ubiquitous presence of Internet, computer viruses develop into thousands of variants which differ in their infection mechanism, propagation mechanism, destructive payload and other features.

From the propagation mechanism point of view, viruses can propagate through one of several different vectors including emails, instant messaging systems, P2P networks, social networking websites, LANs, WANs, etc. Some more sophisticated viruses can even propagate through multiple vectors. These vectors can be classified into the two categories of social and technological networks as discussed above. Hence, computer viruses can be classified into three categories based on their propagation vectors – social network (SN) based, technological network (TN) based, and composite network (CN) based. For example, MyDoom is primarily transmitted via email and P2P network and therefore is a SN-based virus. Unsuspecting computer users expose more personal information and are more vulnerable to SN-based viruses. The Blaster, as an example of TN-based viruses, starts from the local machine's IP address or a completely random address and attempts to infect sequential IP addresses. Nimda, a well-known multi-vector virus, spreads itself by different propagation methods including IP probing, email, network shares, etc. and therefore is a CN-based virus.

Prior research on computer viruses shows that network topology is crucial for virus propagation. In a computer virus incident, the topology of the victim network is the determinant factor of the propagation speed and destructive consequences. Towards this direction, researchers examine network topology to enhance computer security. For example, some research uses local network measures to explain virus propagation. Kephart and White incorporate average node degree into traditional epidemic models (Kephart and White 1991, Kephart and White 1993). Other studies consider specific topologies such as small-world network (Moore and Newman 2000) and scale-free network (May and Lloyd 2001, Pastor-Satorras and Vespignani 2001). Most extant work focuses on degree distributions and assumes certain distributions such as power-law distribution. However, most real world networks are not exactly scale-free (Balthrop et al. 2004). Few research incorporates network properties of individual nodes and examine network topology empirically.

This paper empirically examines a large-scale organizational network which consists of both social network and technological network. We utilize a novel information retrieval technique to map nodes in the social network to nodes in the technological network to construct the composite network of the organization. We apply social network analysis to study the topologies of social, technological, and composite networks in an organization. We perform a comprehensive network analysis on these three networks and compare them both visually and quantitatively. Further, we statistically test the impact of the interplay of social and technological network on computer virus propagation using a susceptible-infective-recovered epidemic process. We find that computer viruses propagate faster but reach lower level of infection through technological network than through social network, and viruses propagate the fastest and reach the highest

level of infection through the composite network. Overlooking the interplay of social network and technological network underestimates the virus propagation speed and the scale of infection.

This paper takes a social network analysis perspective to examine the computer virus propagation problem in a network organization consisting of intertwined social and technological networks. A network organization is viewed as a network where individuals and their computers in the organization are nodes in the network, logical information communications among individuals form edges in the social network and physical data transmission among individuals' computers form edges in the technological network. Computer viruses start from certain nodes and propagate through the edges and the propagation process can be considered as a dynamic network flow built upon the underlying social network and technological network.

This paper aims to address three research questions. First, what are the structural differences among social network, technological network and composite network? Second, how does computer virus propagate through social network, technological network, and composite network? What are the differences in their propagation patterns? Third, can network structural differences and the interplay of social network and technological network explain different virus propagation patterns?

This work addresses some critical limitations of existing literature. Most extant research examines virus propagation either at the individual computer level or at the single network level, but not at the organizational level. Extant literature overlooks the impact of the differences and the interplay between social network and technological network on virus propagation. This paper proposes a holistic view of a network organization's computing environment to examine computer virus propagation patterns. Since all computers in an organization are potential virus victims and all networks, including social networks and technological networks, are possible

virus vectors, we view an organization as a composite of both and examine its social network and technological network simultaneously. There are intrinsic differences between social network and technological network which affect virus propagation. However, previous research of virus propagation and defense either does not distinguish between social networks and technological networks or fails to consider the combined effect of the two. This paper performs a comprehensive network analysis of the three networks and compares their structural differences which serve as the rational of the differences in their virus diffusion behavior.

The remainder of the paper is organized as follows. In Section 2, we discuss the research method and propose a multilevel structural determinants model for computer virus propagation. Various research questions and hypotheses are then proposed to examine the virus propagation in reference to the structural properties of the networks. The following section introduces our research sample by analyzing the structural properties of the sample organization's social network, technological network, and composite network. In the section of computational analyses, we compare the virus propagation on the sample organization's social network, technological network, and composite network and report the results. Finally, we conclude the paper and discuss possible extensions of this paper.

2. Research Model and Hypotheses

2.1 Dynamics of Computer Viruses Propagation

Most prior research in virus propagation focuses on the overall scale of computer epidemics measured by the number of total infected computers. However, this single measure only considers one static point in the whole virus propagation process. This simplification ignores

some important characteristics of virus propagation such as the propagation speed. In this study, we propose three additional measures to capture the dynamic features of a virus propagation process.

Insert Figure 1 about here

Let N_t be the cumulative number of infections detected at time t, $t = 1, \ldots, T$, where T is the maximum number of time epochs. Two groups of measures based on N_t are critical to evaluate a virus incidence – infection time and infection number. In particular, we are interested in two critical infection times – time to takeoff (T_d), and time to equilibrium (T_{eqm}), and the corresponding infection numbers. As shown in Figure 1, the cumulative infection number typically follows an *S*-shaped curve (Kephart and White 1991, Kephart and White 1993). Accordingly, the process of virus propagation has three stages – incubation, proliferation and equilibrium. The two infection times – time to takeoff (T_d) and time to equilibrium (T_{eqm}) – are the two cutoff points between these stages. These infection times and their corresponding infection numbers capture the fundamental characteristics of a virus propagation process and thus should be adopted as key control variables in information security management. A better understanding of when computer virus epidemics take off, when it reaches its equilibrium, and the scales in its propagation help information security managers make more informed responses as well as design improved security policies. Therefore we use these four measures as dependent variables in our models that follow. More formally, Table 1 gives both conceptual and operational definitions of these variables.

Insert Table 1 about here

2.2 Computer Viruses Propagation Through the Three Networks

For each starting node, we use one-way repeated measures analysis of variance (ANOVA) to examine how the propagation patterns of computer viruses differ in various network contexts such as social networks, technological networks, and composite networks. The following model is developed to discern whether the three network types have significantly different virus diffusion dynamics in terms of four key measures of infection time and infection number. Specifically, let

$$Y_{ij} = \mu + \alpha_j + s_i + \alpha s_{ji} + \varepsilon_{ij},$$

where j denotes network type which can be SN, TN, or CN; i denotes starting node; $Y's$ are dependent variables which can be T_d, T_{eqm} and N_d, N_{eqm}; μ denotes the overall mean; α_j denotes the main effect of network type; s_i denotes the effect of starting node; αs_{ji} represents the mean influence of starting node i and network j; and ε_{ij} is random error.

We pose the following research questions and propose hypotheses in below regarding key measures of infection time and infection number for the three networks.

Research Question 1**:** *What are the differences of SN, TN, CN in terms of the key measures of infection time?*

Since the density of TN is much higher than that of SN, and CN has the highest density, we hypothesize that computer virus propagates the fastest in CN and then TN and SN.

RESEARCH HYPOTHESIS 1A: $T_{d,SN} > T_{d,TN} > T_{d,CN}$

RESEARCH HYPOTHESIS 1B: $T_{eqm,SN} > T_{eqm,TN} > T_{eqm,CN}$

Research Question 2: *What are the differences of SN, TN, CN in terms of the key measures of infection number?*

Since the mean node-to-node distance of SN is shorter than that of TN, and CN has the shortest mean distance, we hypothesize that CN reaches the highest level of infection number and then SN and TN.

RESEARCH HYPOTHESIS 2A: $N_{d,CN} > N_{d,SN} > N_{d,TN}$

RESEARCH HYPOTHESIS 2B: $N_{eqm,CN} > N_{eqm,SN} > N_{eqm,TN}$

2.3 The Interplay between Social Network and Technological Network

Network topology has a hierarchical structure with individual nodes nested in subgroups. Hierarchical linear model (HLM) is thus used to capture the nested nature of the network topology data. In order to examine the impact of the interplay between social network and technological network on virus propagation, we consider a hierarchical linear model with two-way cross classification which enables us to simultaneously assess the interactive effect of individual-level, group-level variables of all three networks. The individual nodes are contained within a two-way cross-classification of SN group and TN group. In the research model, we have individual nodes at Level 1 and SN group and TN group are cross-classified at level 2. As shown in Figure 2 and Figure 3, we propose a two-level hierarchical linear model with individual-level variables at the first level, group-level variables at the second level while individuals can be cross-classified based on SN and TN groups.

Insert Figure 2 and Figure 3 about here

Individual Level (Level-1):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}\text{OutDegree}_{ijk} + \pi_{2jk}\text{InDegree}_{ijk} + e_{ijk}$$

where

Y_{ijk} 's denote four measures of computer virus propagation dynamics,

π_{pjk} with $p = 1, 2$ are level-1 coefficients,

e_{ijk} denote level-1 random effect.

Outdegee and indegree are two most widely used centrality measures for individual nodes. Outdegree is defined ase the number of outgoing links from the focal node and indegree is defined as the number of incoming links to the focal node. We note that these two variables only depend on the focal node and its direct neighbors and are independent of the rest of the network. Therefore, both outdegree and indegree are local structural properties. In our model, we specify *OutDegree* and *InDegree* to be the individual-level independent variables.

Each of the level-1 coefficients π_{pjk} can be further expressed as an outcome variable in the group-level model as follows:

Group Level (Level-2):

$$\pi_{0jk} = \theta_0 + \left(\beta_0 + b_{01j}\right)\text{SNGroupSize}_k + \left(\gamma_0 + c_{01k}\right)\text{TNGroupSize}_j + \delta_{0jk}\text{SNGroupSize}_k * \text{TNGroupSize}_j + b_{00j} + c_{00k} + d_{00jk}$$

$$\pi_{1jk} = \theta_1 + \left(\beta_1 + b_{11j}\right)\text{SNGroupSize}_k + \left(\gamma_1 + c_{11k}\right)\text{TNGroupSize}_j + \delta_{1jk}\text{SNGroupSize}_k \cdot \text{TNGroupSize}_j + b_{10j} + c_{10k} + d_{10jk}$$

$$\pi_{2jk} = \theta_2 + \left(\beta_2 + b_{21j}\right)\text{SNGroupSize}_k + \left(\gamma_2 + c_{21k}\right)\text{TNGroupSize}_j + \delta_{2jk}\text{SNGroupSize}_k \cdot \text{TNGroupSize}_j + b_{20j} + c_{20k} + d_{20jk}$$

where

θ_0 is the model intercept or the grand mean,

θ_p with $p = 1, 2$ is the group mean,

β_m with $m = 0, 1, 2$ are the fixed effects of the column-specific predictor, i.e., SNGroupSize,

b_{m1j} with $m = 0, 1, 2$ are the random effects of the column-specific predictor SNGroupSize which vary across rows, i.e, different TN groups,

γ_m with $m = 0, 1, 2$ are the fixed effects of the row-specific predictor, i.e., TNGroupSize,

c_{m1k} with $m = 0, 1, 2$ are the random effects of the row-specific predictor TNGroupSize which vary across columns, i.e, different SN groups,

δ_{mjk} with $m = 0, 1, 2$ are the fixed effect of the cell-specific predictor SNGroupSize $\cdot$ TNGroupSize, i.e., the cross-classification effect,

b_{m0j} is the SN-specific error,

c_{m0k} is the TN-specific error, and

d_{m0jk} is the cell-specific error.

Hence, we pose the following research questions and associated hypotheses.

Research Question 3: *Does the size of social network and technological network affect the key measures of the infection time associated with these networks?*

RESEARCH HYPOTHESIS 3A: *For* T_d, $\beta_{0,T_d}, \gamma_{0,T_d} < 0$

RESEARCH HYPOTHESIS 3B: *FOR* T_{eqm}, $\beta_{0,T_{eqm}}, \gamma_{0,T_{eqm}} < 0$

Research Question 4: *Does the size of social network and technological network affect the key measures of the infection number associated with these networks?*

RESEARCH HYPOTHESIS 4A: *FOR* N_d, $\beta_{0,N_d}, \gamma_{0,N_d} > 0$

RESEARCH HYPOTHESIS 4B: *FOR* N_{eqm}, $\beta_{0,N_{eqm}}, \gamma_{0,N_{eqm}} > 0$

3. Research Sample

We collected empirical data of a large-scale organization's social network, technological network, and constructed the composite network accordingly. The sample organization is one of the largest research universities in the U.S. with a total enrollment of more than 50,000 students. We first gathered all member data on MySpace, the largest social networking website on the Internet, with affiliation to this university. Mining the detailed friend listing data of each MySpace member enabled us to construct the social network. We obtained the core technological network of this university, and then mapped the social network to the technological network to derive a composite network according to each subject's physical location on the technological network. The following subsections give detailed descriptions of our research sample.

Insert Figure 4 and Figure 5 about here

3.1 Social Network

Using Perl and RegEx, we collected data about members of MySpace who are current students of the sample organization. The number of current student members of MySpace in February 2006 is 14,933, which accounts for more than 30% of all enrolled students of the sample university.

Another motivation for us to choose this dataset is that college students are considered the high risk group in terms of virus infection and propagation. The relationship from one research sample on MySpace to another is uncovered by mining the detailed friend listing data published in each member's online space, resulting in the social network of our research sample represented by a directed graph as shown in Figure 4(a).

3.2 Technological Network

Organizations adopt different heterogeneous computing environments which involve different technological networks, the most popular being LAN. These technological networks use different types of topologies. The three most common topologies are the star, ring, and bus. Ethernet with bus topology dominates 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 as shown in Figure 4(b). Since 2267 members out of the total 14,933 members do not reveal their subject of studying online, the resulting technological network consists of 12,666 individual nodes. Combined with 15 core nodes of the campus network and 168 subject nodes, the technological network has a combination of 12,849 total nodes. A square in Figure 4(b) represents a subject node in a building connected to the core network where the local networks (LANs) in each building naturally follow the bus topology. The 168 LANs are completely connected networks which represent a worst case scenario for the exploration of computer virus epidemics.

3.3 Composite Network

We were able to map each individual node in the social network (SN) to the technological network (TN) by locating each individual's physical presence in the technological network. The general structure of the composite network (CN) is shown in Figure 4(c). The CN is a directed graph which has the same number of nodes as TN and two different sets of edges – edges from SN and edges from TN. One salient and intuitive feature of CN is that edges from SN serve as bridges to connect LANs directly.

The sample social and technological networks are also visualized in Figure 5(a) and 5(b). Figure 5(a) is drawn using LaNet-vi (Alvarez-Hamelin et al. 2005). LaNet-vi is a k-core decomposition-based visualization tool designed for large complex networks. Figure 5(b) is drawn using Ucinet (Borgatti et al. 2002).

4. Structural and Computational Analyses

Network structure is the focus of social network analysis. Social network analysis views social entities as nodes and relationships as edges. Nodes and edges are the two fundamental elements in a network. A rich set of concepts and methods have been developed to analyze network structures. Wassermann and Faust (1994), a popular text, provides a good review for social network analysis. Social network analysis is widely used in sociology, organizational studies and other fields. For example, it is used to analyze customer networks, inter-firm alliances, and information flow networks. In this section, we utilize social network analysis tools to examine three different network context for the computer virus propagation problem. Specifically we examine three important structural properties – cohesion, degree distribution and subgroup structure of social network, technological network, and composite network.

4.1 Cohesion

Common cohesion measures include density, reciprocity rate, node-to-node distance, diameter, and reachability. We report these structural properties of SN, TN, and CN in Table 2. The SN of the sample organization has a density of 0.0006 which is much sparser than TN (with density 0.179) and CN (with density 0.184). Dyad-based reciprocity rate is very high for SN (0.9943), implying that the friendship between users are mutual. Although the diameter of social network is greater than that of technological network, the mean node-to-node distance of social network (4.406) is much shorter. Among the three networks, CN has the smallest diameter and the shortest mean distance among the nodes.

Insert Table 2 about here

4.2 Degree Distribution

The out-degree and in-degree measure how many outgoing or incoming edges a node has in a network. The degree distribution describes the variability of the nodal degrees in a network. The most commonly seen degree distribution is power-law distribution. As shown in Figure 6, the social network approximately follows a power-law distribution while the technological network has a more discrete distribution.

Insert Figure 6 about here

4.3 Subgroup Structure

There are cohesive subgroups embedded in networks. Subgroup structure of a network refers to a partition of the network into subgroups where there are far more links within subgroups than between subgroups. In other words, ties among the individual nodes are concentrated inside the

subgroups rather than outside the subgroups. Subgroup structure is a common structural property for networks including both social networks and technological networks (Newman and Girvan 2004). Traditional subgroup analysis techniques include component analysis, clique analysis, core analysis, hierarchical clustering and so on. Among them, *k*-core decomposition has two salient advantages – easy to compute and the resulting subgroups usually demonstrate a hierarchical structure (Alvarez-Hamelin, Dall'Asta, Barrat and Vespignani 2005). Therefore we apply *k*-core decomposition to analyze the subgroup structures of SN and TN, resulting in 38 SN subgroups and 168 TN subgroups. The discovered SN groups and TN groups constitute the group-level (level-2) factor in our two-way cross-classified HLM model. Following the convention that the factor with more units becoming the row factor while the factor with less units becoming the column factor, we specify TN group as the row factor and SN group as the column factor in our model. As a result, the individual-level (level-1) data are contained in cells cross-classified by the row factor (TN group) and column factor (SN group).

The size of the group is the sole level-2 variable defined for each group (either SN group or TN group), denoted by *SNGroupSize* and *TNGroupSize*. Figure 7 shows the distribution of group size of both social network and technological network.

Insert Figure 7 about here

4.4 Computer Virus Propagation Analyses

Computer virus propagation has been widely researched using epidemiology models. Among these epidemiology models, SIR (Susceptible–Infected–Recovered) model is most commonly used. Researchers conduct computer simulations to analyze the virus propagation process. Following the SIR model, we developed computer algorithms to simulate the virus propagation

in the social network, technological network, and composite network. 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 virus and can be infected by its neighbors. An infected node i can infect its neighbor j according to j's infection probability α_j. After trying to infect its neighbors, the infected node i may be recovered according to its recovery probability δ_i. 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 virus is eliminated from the computer by the user through security patching. Every infected node can try to infect its neighbors at all times before it is recovered.

We applied the discrete-time simulation method to analyze the computer virus propagation process. Beginning at time 0, a single randomly chosen node becomes infected and this node starts the virus propagation process. We randomly selected 3000 starting nodes. The propagation process stops either when the virus stops spreading, i.e., when the number of currently infectious nodes reduces to 0, or when the process runs long enough and reaches the maximum iteration number of time epoch $T = 100$. We assume a power-law distribution of the simulation parameters α_i (the probability that node i gets infected in each infection attempt) and δ_i (the probability that node i gets recovered at time $t+1$ given that node i is infected at time t). The power-law distribution captures the asymmetric nature of user behaviors. Most of the users have high infection rate and recovery rate while only few of them have low infection rate and recovery rate. We set the parameter of the exponential in the power-law distribution to 2.690 and 2.286 for infection rate and recovery rate respectively. The parameters are chosen such that the mean value of the infection rate and recovery rate are consistent with the empirical findings in the literature. Chen and Carley (Chen and Carley 2004) find the ratio of infection rate to

recovery rate is between 0.01 and 0.2 with a mean of 0.05. For each starting node, we ran the simulation 20 times. Then we calculated the mean value for the number of infections and used it as the dependent variable. We ran each simulation 20 times and used the average values for the dependent variables – T_d, T_{eqm} and N_d, N_{eqm}.

5. Research Results

5.1 Different Patterns of Computer Viruses Propagation

Table 3 summarizes the results for the four hypotheses. Comparing computer virus propagations through the three networks, we find that viruses propagate fastest through CN and reach the highest level of infection number. Although viruses propagate faster through TN than SN, the equilibrium infection number is higher for SN than TN.

Insert Table 3 about here

A repeated measures one-way ANOVA reveals that there are significant differences in propagation patterns of computer viruses among social network, technological network and composite network, with F ranging from 1399.925 to 17396.421, and $p < .001$ for all four measures of virus diffusion dynamics. LSD comparisons revealed that all three means based on three network types were significantly different from each other. We find that although computer viruses spread slower in social network than technological network ($T_{d,SN} > T_{d,TN}$), the final scale of computer infections on social network is higher than that on technological network ($N_{eqm,SN} > N_{eqm,TN}$). Compared with social network and technological network, computer viruses

spread the fastest and reach the highest infection number through composite network ($T_{d,SN} > T_{d,TN} > T_{d,CN}$ and $N_{eqm,CN} > N_{eqm,SN} > N_{eqm,TN}$).

5.2 Multilevel Structural Determinants of Computer Viruses Propagation

As indicated in high reciprocity rate, we note that individual-level outdegree and indegree are highly correlated. Hence, we derive reduced models by eliminating indegree from the individual-level model to correct the multicollinearity problem and report corresponding research findings. Table 4 presents the estimation results of our model of the multilevel structural determinants of computer viruses propagation.

Insert Table 4 about here

We find computer viruses spread faster and eventually infect more computers if the epidemic incidence starts from a node located in a larger technological group. The size of social groups, on the other hand, does not affect the speed or the scale of computer virus propagation directly. Instead, subgroup structure of social network affects the scale of computer epidemics indirectly through interaction with individual-level centrality measures such as outdegree.

6. Conclusions

Extant literature on computer virus propagation unfortunately does not examine the problem from the perspective of an organization's network structure and overlooks the interplay of social network and technological network embedded within the organization. To address this critical issue, we collected empirical social network (SN) data from MySpace, the largest social networking website, and mapped them to the technological network (TN) of a large organization

to construct the composite network (CN) that illustrates the effect of interplay of SN and TN. We then applied social network analysis techniques to compare and contrast the virus propagation in SN, TN, and CN. We further examine the impact of the interplay of social network and technological network on the computer virus propagation process. We find that viruses propagate faster but reach lower level of infection through technological network than through social network, and computer viruses propagate the fastest and reach the highest level of infection through the composite network. Overlooking the interplay of social network and technological network underestimates the virus propagation speed and the scale of infection.

References

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

Alvarez-Hamelin, I., L. Dall'Asta, A. Barrat, A. Vespignani. 2005. *Large scale networks fingerprinting and visualization using the k-core decomposition*.

Balthrop, J., S. Forrest, M.E.J. Newman, M.M. Williamson. 2004. Technological networks and the spread of computer viruses. *Science*. **304** 527-529.

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

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

Kephart, J.O., S.R. White. 1993. Measuring and modeling computer virus prevalence. *Proceedings of the 1993 IEEE Computer Society Symposium on Research in Security and Privacy* 2-15.

May, R.M., A.L. Lloyd. 2001. Infection dynamics on scale-free networks. *Physical Review E*. **64**(6) 66112.

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

Newman, M.E.J., M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E*. **69**(2) 26113.

Newman, M.E.J., J. Park. 2003. Why social networks are different from other types of networks. *Physical Review E*. **68**(3) 36122.

Pastor-Satorras, R., A. Vespignani. 2001. Epidemic spreading in scale-free networks. *Physical Review Letters*. **86**(14) 3200-3203.

Sproull, L., S. Kiesler. 1991. *Connections: New ways of working in the networked organization*. The MIT Press.

Van Alstyne, M. 1997. The state of network organization: A survey in three frameworks. *Journal of Organizational Computing*. **7**(3).

Figures and Tables:

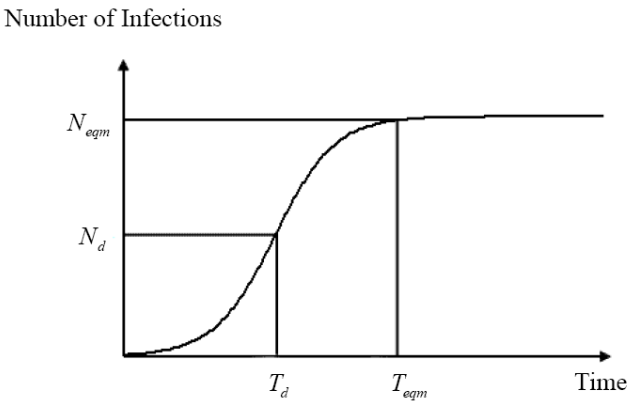


Figure 1: Dynamics of Computer Virus Propagation

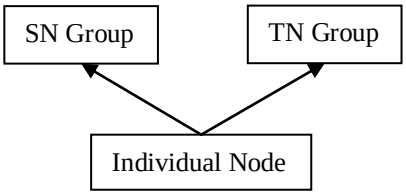


Figure 2: Classification diagram

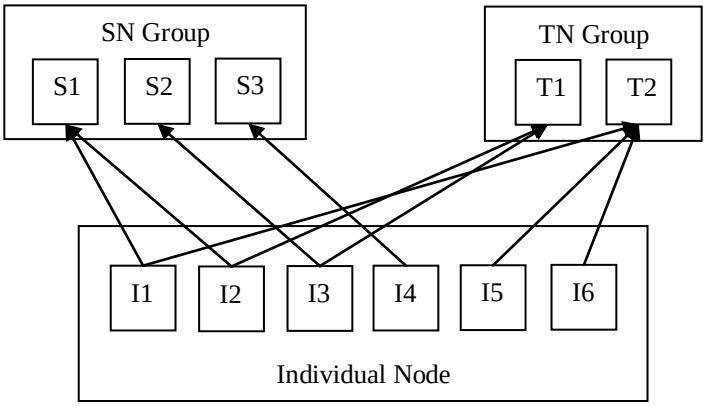
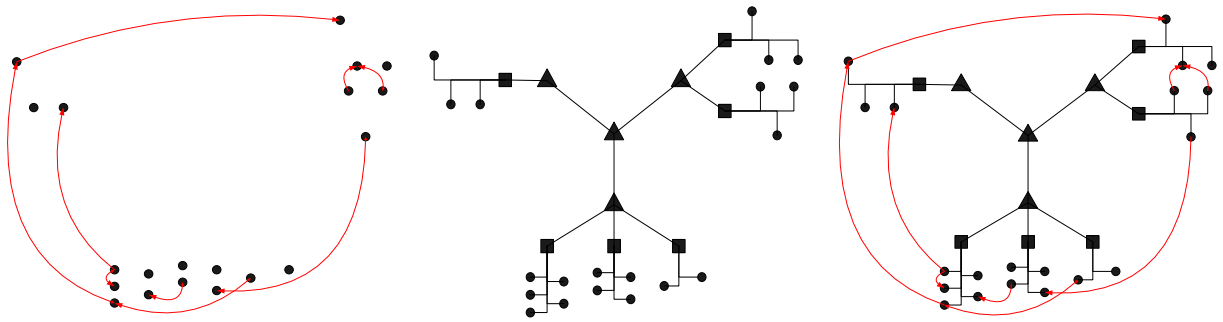


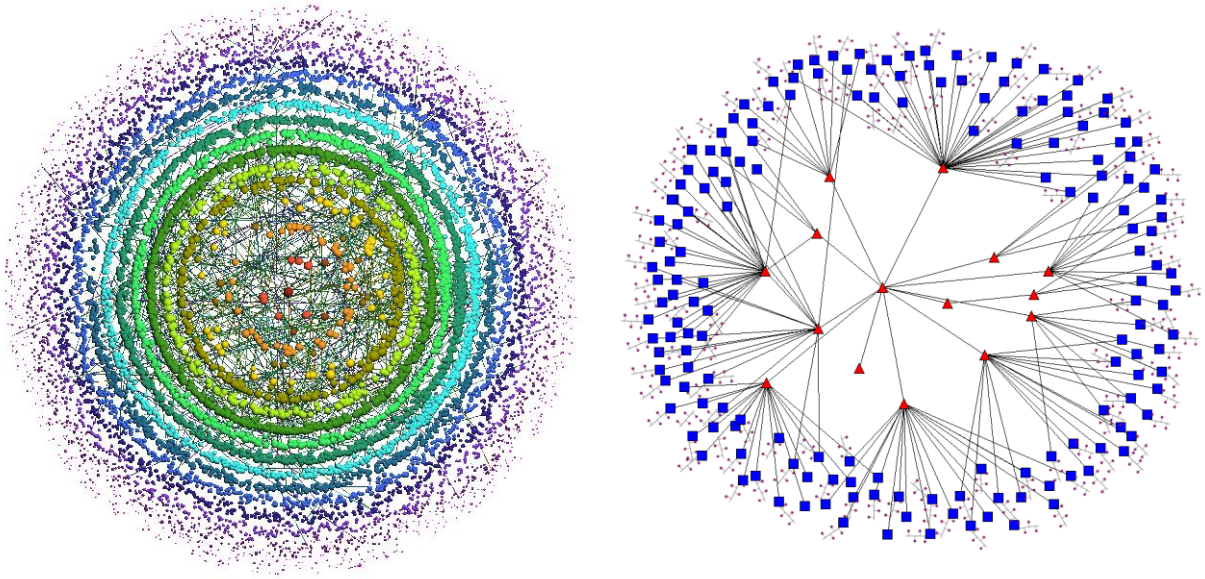
Figure 3: Unit diagram where individual nodes lie within a cross-classification of SN, TN, and CN



(a) Social Network (b) Technological Network (c) Composite Network

Figure 4: Topologies of Three Networks

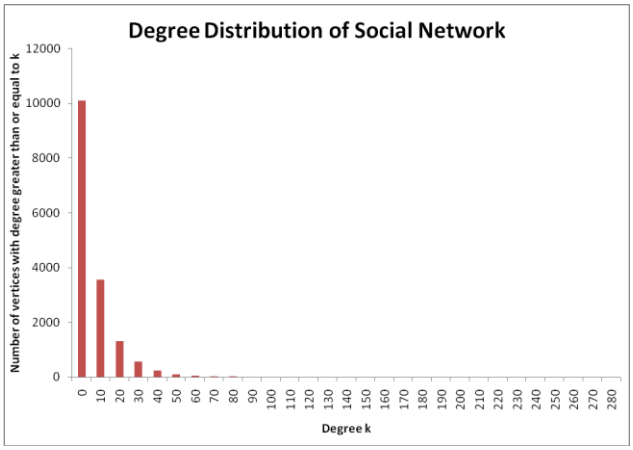
Notes: Circle represents individual node; Triangle represents core node in technological network; Square represents major node in technological network.



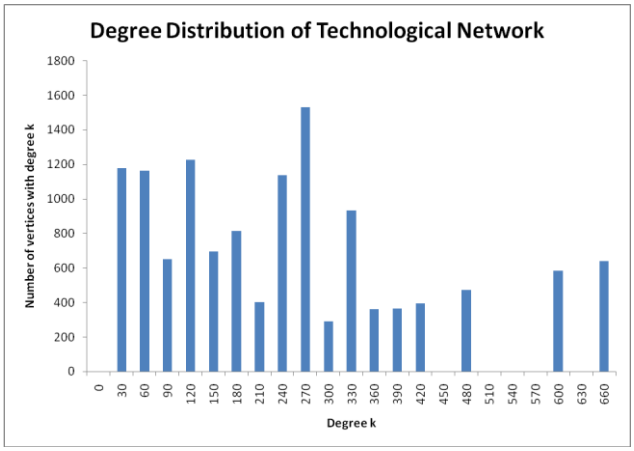
(a) Social Network (b) Technological Network

Figure 5: Research Sample Networks

Notes: In Figure 5(a), both color and size of the nodes denote nodal degree. Degree decreases when the node color changes from red to green, blue, and purple and when the node size becomes smaller. In Figure 5(b), circle represents individual node; triangle represents core node in technological network; square represents major node in technological network.

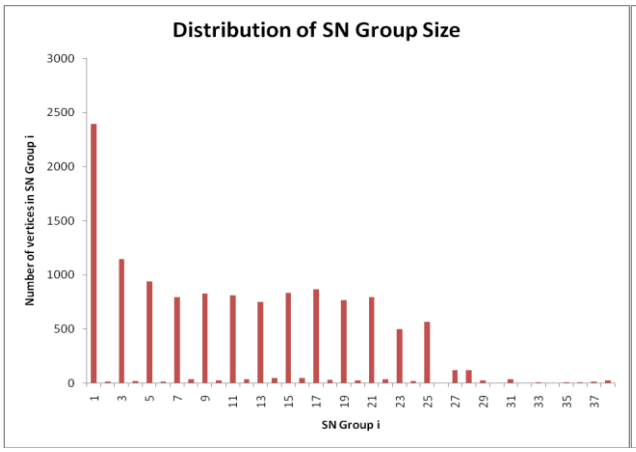


(a) Social Network

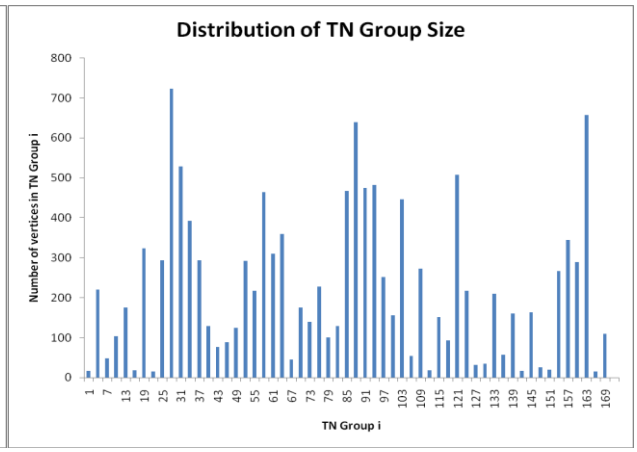


(b) Technological Network

Figure 6: Degree Distributions



(a) Social Network



(b) Technological Network

Figure 7: Distribution of Group Sizes

Table 1: Definitions of Dependent Variables

Dependent Variable		Conceptual Definition	Operational Definition
Takeoff	Time to Takeoff (T_d)	The time epoch with the fastest growth (largest slope) in number of infections	$T_d = \text{argmax}\left\{\frac{N_{t+1}-N_t}{N_t}, t \in \{1,...,T\}\right\}$
	Number of Infections at Takeoff (N_d)	The number of infections at time T_d	$N_d = N_{t=T_d}$
Equilibrium	Time to Equilibrium (T_{eqm})	The first time epoch when the number of infections reaches equilibrium state	$T_{eqm} = \min\left\{t \text{ such that } \frac{N_{t+k}-N_t}{N_t} \le 5\%, \atop k=1,...,10,\ t \in \{1,...,T\}\right\}$
	Number of Infections at Equilibrium (N_{eqm})	The number of infections at time T_{eqm}	$N_{eqm} = N_{t=T_{eqm}}$

Table 2: Structural Properties of Three Networks

Statistics	Social Network	Technological Network	Composite Network
Directed/Undirected	Directed	Undirected	Directed
Number of Nodes	12,666	12,849	12,849
Number of Edges	105,528	2,970,324	3,066,068
Mean Degree	8.332	231.172	238.623
Mean Density	0.000658	0.0180	0.0186
Reciprocity Rate	0.994	1.000	0.9998
Mean Node-to-Node Distance	4.406	5.839	2.824
Diameter	12	8	7
Reachability	0.617	1.000	1.000

Table 3: Comparison of Virus Propagation Differences on Three Networks

Research Question		SN Mean (SD)	TN Mean (SD)	CN Mean (SD)	F	p	Results
Infection Time	T_d	1.435 (.996)	.164 (.621)	.139 (.400)	3738.125	.000	$T_{d,SN} > T_{d,TN} > T_{d,CN}$
	T_{eqm}	14.287 (3.520)	7.274 (1.193)	7.668 (.293)	10038.387	.000	$T_{eqm,SN} > T_{eqm,CN} > T_{eqm,TN}$
Infection Number	N_d	32.571 (32.173)	3.396 (10.568)	7.737 (23.395)	1399.925	.000	$N_{d,SN} > N_{d,CN} > N_{d,TN}$
	N_{eqm}	8161.265 (2069.140)	6101.055 (1370.047)	12798.652 (1.352)	17396.421	.000	$N_{eqm,CN} > N_{eqm,SN} > N_{eqm,TN}$

Table 4: Hierarchical Linear Model Estimation

				Infection Time		Infection Number	
				T_d	T_{eqm}	N_d	N_{eqm}
Fixed Effect	π_{0jk}	Intercept θ_0	Coefficient	1.00754	8.107	33.196	12798.615
			SE	0.0532	0.0314	2.847	0.143
		TN Group γ_0	Coefficient	-0.00367	-0.00171	-0.111	0.000253
			SE	0.000346	0.000155	0.015	0.000205
		SN Group β_0	Coefficient	0.000013	-0.00002	0.00157	0.000008
			SE	0.000019	0.000019	0.00161	0.000126
	π_{1jk}	Intercept θ_1	Coefficient	0.000241	0.000620	0.031	0.0127
			SE	0.000726	0.000796	0.065	0.00519
		TN Group γ_1	Coefficient	0.000001	-0.000002	0.000064	-0.000033
			SE	0.000002	0.000002	0.000149	0.000012
		SN Group β_1	Coefficient	-0.000001	-0.000001	-0.000082	-0.000010
			SE	0.000001	0.000001	0.000070	0.000006
Random Effect	TN Group b_{00j}		Variance	0.222	0.0409	20.046	0.00183
			df	146	146	146	146
			χ^2	29009.271	2357.267	3749.638	170.854
	SN Group c_{00k}		Variance	0.00006	0.000	0.126	0.00008
			df	28	28	28	28
			χ^2	32.846	34.319	27.00475	23.789
	Individual-level error ε_{ijk}		Variance	0.031	0.0405	273.205	1.818