Association for Information Systems
AIS Electronic Library (AISeL)

ICIS 2007 Proceedings

International Conference on Information Systems (ICIS)

December 2007

Computer Virus Propagation in Social Networks

Hong Guo
University of Florida

Hsing Cheng
University of Florida

COMPUTER VIRUS PROPAGATION IN SOCIAL NETWORKS

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

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

Abstract

This paper applies social network analysis techniques to study computer virus propagation. We propose a novel multilevel hierarchical linear model to simultaneously evaluate the impact of both individual-level and group-level variables on virus propagation process. In this model, we propose centrality and brokerage measures as explanatory variables. We estimate our model based on empirical data from the largest social networking website and find that closeness centrality (individual level) and brokerage (group level) jointly explain 82.5% of the variance in the number of infections. This research contributes to both the literature of computer virus propagation and defense and the literature of centrality measure comparison in the filed of social network analysis by: (1) performing subgroup analysis and considering multiple levels of network characteristics to capture the intrinsic nested feature of the networks and (2) comparing different structural measures (centrality and brokerage) in terms of their performance to explain the propagation of computer viruses.

Keywords: Computer viruses, social network analysis, social networking websites, hierarchical linear model

Introduction

Computer virus attacks constantly cause the greatest financial loss and present a top security concern for organizations (Gordon et al. 2006). In response, security researchers and practitioners have developed different virus defense mechanisms including incoming attack protections and local containments (Brumley et al. 2006). However, most of these defense techniques such as intrusion detection and scanning focus on local behavior. Only recently have efforts been made to examine global virus propagation processes and the corresponding global defense strategies. Considering the entire network instead of individual nodes, researchers investigate the impact of different network topologies on virus propagation such as random graph, small-world (Moore and Newman 2000), scale-free (Pastor-Satorras and Vespignani 2001; May and Lloyd 2001), and giant strongly connected component (Newman et al. 2002) with random graphs usually used as a benchmark case. The topology used in these studies is either simulated or real network data. Built upon the underlying network, virus proliferation is modeled as a stochastic process. SIS (Susceptible–Infected–Susceptible) and SIR (Susceptible–Infected–Recovered) are the two most popular models which originate from mathematical epidemiology. In this paper we adopt a different approach by applying techniques from social network analysis to the computer virus propagation problem.

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. The focus of social network analysis is network structure. Position of a node in a network has significant impact on its influences, opportunities and constraints. Individual nodes are nested in subgroups and subgroups are intertwined with each other through inter-group connections. Centrality is a widely used structural measure in social network analysis. Among all the centrality measures, Freeman's degree centrality, betweenness centrality, and closeness centrality are the most popular ones. Researchers have studied and compared these centrality measures from different angles. Costenbader and Valente (2003) evaluate the stability of these centralities and find that indegree centrality is relatively stable. Borgatti (2005) views dynamic processes built on networks as network flows and compare centrality measures based on different types of network flows through conceptual analysis and computer simulation. However, Borgatti (2005) finds no centrality measures appropriate for virus propagation processes.

This paper takes a social network perspective to examine the computer virus propagation problem. An organization is viewed as a network where individuals in the organization are nodes in the network and communications among individuals form edges in the network. Computer viruses start from certain nodes and propagate through the edges and the propagation process can be considered as a special type of network flow. Then the questions become: Does the position of the starting node in a virus incident affect the infection result? Can we model the impact of starting node using structural measures? This paper proposes a novel solution to address these questions. In particular, we conduct a multilevel analysis using a hierarchical linear model to simultaneously examine the impact of both individual-level and group-level network characteristics on the computer virus propagation process. Instead of asking which single centrality measure is appropriate for the virus propagation process, this paper investigates whether structural measures other than centrality help explain virus propagation process. Specifically, in addition to centrality, this paper also examines brokerage, another important structural measure. We use Burt's concept of structural holes and related measure of aggregate constraints to evaluate brokerage.

Our work contributes to both the literature of computer virus propagation and the literature of centrality measure comparison in the field of social network analysis. Extant literature on virus propagation turns to simulations to find the most risky nodes in a network without questioning what characteristics make these nodes more risky than others. Instead, we propose a multilevel starting node effect model of computer virus propagation and therefore provide a way to identify risky nodes based on structural characteristics of both the focal individual node and the local group it belongs to. By performing a subgroup analysis and considering multiple levels of network characteristics, we are able to capture the intrinsic nested features of the networks.

This paper also contributes to the literature of centrality measure comparison in the field of social network analysis. We provide empirical evidence of the relationship between structural measures of the underlying network and computer virus propagation processes. Using a hierarchical linear model we find that it is misleading to claim one single structural measure is better than another for a complicated dynamic process like computer virus propagation.

Instead, we find that several measures interconnect in a more sophisticated way and jointly explain the dynamic process.

The remainder of the paper is organized as follows: In next section, we discuss four centrality measures and one brokerage measure for both individual level and group level. We then develop our research model step by step through four research questions to examine the impact of starting node on computer virus propagation. The section of research sample describes the structural characteristics of our sample network. In the following section, we conduct computational analysis and present our research results. Finally, we conclude the paper.

Research Model

In this section, we investigate the impact of the starting nodes on the spread of computer viruses. We first introduce centrality and brokerage measures from the field of social network analysis and use them as individual-level variables. We then perform a subgroup analysis and define group-level variables. A hierarchical linear model is proposed to explore the impact of structural characteristics of the starting node and its local group on virus propagation dynamics.

Centrality and Brokerage as Individual-level Independent Variables

The centrality measures reveal how influential and powerful a node is in a social network. In this paper, we use the most popular Freeman's degree, betweenness, and closeness centralities. Formal definitions of centralities can be found in Freeman (1978/1979). For directed networks, degree centrality can be further divided into outdegree centrality and indegree centrality. Degree centrality measures how many outgoing or incoming edges a node has in a network. Betweenness centrality measures how often a node falls on the geodesic paths among pairs of other nodes. Closeness centrality gauges how far on average a node is to all other nodes. Intuitively, the more central a node is, the more risky it is in the context of computer virus proliferation.

Burt (1992) proposes the concept of structural hole which refers to the gap between social groups. Brokerage represents the ability to fill in the structural hole in a network by connecting nodes which are originally not connected. Burt (1992) further defines a quantitative measure called aggregate constraint which can be used as an inverse measure for brokerage. A higher aggregate constraint corresponds to a lower brokerage. In the context of computer virus propagation, nodes with higher constraint have less chance to spread viruses while nodes with lower constraint have higher chance to spread viruses. We hypothesize that centrality and brokerage of the starting node in a computer virus incident can be used to predict the infection result. In our starting node effect model, individual-level independent variables consist of four centrality measures and one brokerage measure.

Subgroup Analysis and Group-level Independent Variables

There are cohesive subgroups[1] embedded in networks which are important for the study of virus propagation because nodes within a subgroup have more connections among each other and therefore are more likely to infect each other; nodes between the subgroups have fewer connections and therefore are less likely to infect each other. Traditional subgroup analysis techniques include component analysis, clique analysis, core analysis, etc. More recently, new algorithms (Newman and Girvan 2004; Clauset et al. 2004) have been proposed to improve the performance of subgroup analysis. Newman and Girvan (2004) introduce the concept of modularity which measures the degree of variation from random network partition. It is also suggested that any value above 0.3 is a good indicator for significant subgroup structure. They also propose a greedy algorithm to maximize the modularity of a network. However, their algorithm is very time consuming. In this study, we apply a simple k-core partition[2] on the network and divide the nodes into subgroups. Our subgroup analysis gives a modularity of 0.246 indicating a fairly well network partition. Based on the resulting subgroups, we define group-level independent variables. The most

[1] In social network analysis, subgroups are also referred to as communities, clusters, etc. 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.

[2] "k-core" is a subgroup of a given network where each node has at least k neighbors in the same core.

straightforward group-level variable is the size of a subgroup. In addition, we also consider the average value of all the centrality and brokerage measures in each group. The individual-level and group-level structural variables constitute the independent variables in our model. Figure 1 provides the conceptual model with the independent variables in the dashed line rectangles and the dependent variable in the double line rectangle.

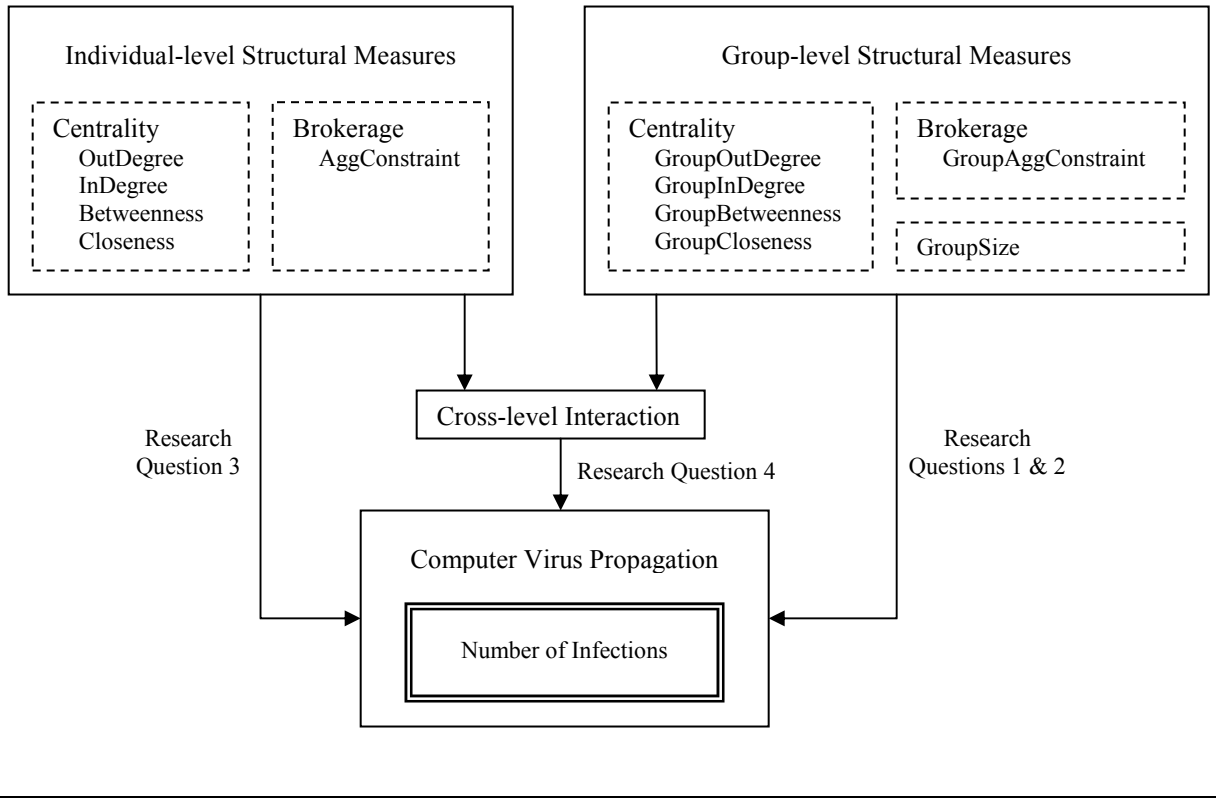

Figure 1. Conceptual Model

A Hierarchical Linear Model of Computer Virus Propagation

In this subsection, we propose a hierarchical linear model to examine the impact of starting node on computer virus propagation. The dependent variable in our model is the number of infected computers, a common measure of the severeness of a virus incident. We examine how the position of the starting node in a network affects the number of infections. We can utilize the starting node effect model to predict the propagation pattern once we know the starting node and the network structure. This model can also be used to identify the set of high risk nodes and has important implications for virus prevention, treatment, and containment strategies. For example, IT managers may adopt heterogeneous software for high risk nodes as a virus prevention strategy, target high risk nodes as starting points for automatic patching as a virus treatment strategy, and/or vaccinate high risk nodes as a containment strategy.

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 and simultaneously assess the interactive effect of both individual-level and group-level variables. To examine the starting node effect on virus propagation, we propose a two-level hierarchical linear model with individual-level variables at the first level and group-level variables at the second level. We provide a complete model specification below to demonstrate the hierarchical linear model designed to predict the number of infections. As we will show later, the independent variables in the complete model are highly correlated. Hence, we derive reduced models to correct the multicollinearity problem and report corresponding research findings.

Individual level:

$$\text{Number of infections}_{ij}\ (Y_{ij}) = \beta_{0j} + \beta_{1j} * \text{OutDegree}_{ij} + \beta_{2j} * \text{InDegree}_{ij} + \beta_{3j} * \text{Betweenness}_{ij} + \beta_{4j} * \text{Closeness}_{ij} + \beta_{5j} * \text{AggConstraint}_{ij} + r_{ij}$$

Group level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * \text{GroupSize}_j + \gamma_{02} * \text{GroupOutDegree}_j + \gamma_{03} * \text{GroupInDegree}_j + \gamma_{04} * \text{GroupBetweenness}_j + \gamma_{05} * \text{GroupCloseness}_j + \gamma_{06} * \text{GroupAggConstraint}_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} * \text{GroupSize}_j + \gamma_{12} * \text{GroupOutDegree}_j + \gamma_{13} * \text{GroupInDegree}_j + \gamma_{14} * \text{GroupBetweenness}_j + \gamma_{15} * \text{GroupCloseness}_j + \gamma_{16} * \text{GroupAggConstraint}_j$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} * \text{GroupSize}_j + \gamma_{22} * \text{GroupOutDegree}_j + \gamma_{23} * \text{GroupInDegree}_j + \gamma_{24} * \text{GroupBetweenness}_j + \gamma_{25} * \text{GroupCloseness}_j + \gamma_{26} * \text{GroupAggConstraint}_j$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31} * \text{GroupSize}_j + \gamma_{32} * \text{GroupOutDegree}_j + \gamma_{33} * \text{GroupInDegree}_j + \gamma_{34} * \text{GroupBetweenness}_j + \gamma_{35} * \text{GroupCloseness}_j + \gamma_{36} * \text{GroupAggConstraint}_j$$

$$\beta_{4j} = \gamma_{40} + \gamma_{41} * \text{GroupSize}_j + \gamma_{42} * \text{GroupOutDegree}_j + \gamma_{43} * \text{GroupInDegree}_j + \gamma_{44} * \text{GroupBetweenness}_j + \gamma_{45} * \text{GroupCloseness}_j + \gamma_{46} * \text{GroupAggConstraint}_j$$

$$\beta_{5j} = \gamma_{50} + \gamma_{51} * \text{GroupSize}_j + \gamma_{52} * \text{GroupOutDegree}_j + \gamma_{53} * \text{GroupInDegree}_j + \gamma_{54} * \text{GroupBetweenness}_j + \gamma_{55} * \text{GroupCloseness}_j + \gamma_{56} * \text{GroupAggConstraint}_j$$

where β_{0j} = mean number of infections for group *j*, i.e., group mean;

β_{pj} = differentiating effect of individual-level measure p for group *j*, with $p = 1, 2, 3, 4, 5$ corresponding to OutDegree, InDegree, Betweenness, Closeness, and AggConstraint respectively;

r_{ij} = error terms at the individual level;

γ_{00} = mean number of infections across all groups and all individuals, i.e., grand mean;

γ_{0q} = differentiating effect of group-level measure q on group mean number of infections, with $q = 1, 2, 3, 4, 5, 6$ corresponding to GroupSize, GroupOutDegree, GroupInDegree, GroupBetweenness, GroupCloseness, and GroupAggConstraint respectively;

u_{0j} = error terms at the group level;

γ_{p0} = mean differentiating effect of individual-level measure p across the groups;

γ_{pq} = differentiating effect of group-level measure q on individual-level measure p, i.e., differentiating effect of cross-level interactions.

Essentially the goal of our hierarchical linear model is to explain the number of infections by structural characteristics of starting nodes including individual node and local group. We pose the following four research questions to develop our starting node effect model step by step.

- **Research Question 1:** Are there differences in the impact of starting nodes from different groups?
- **Research Question 2**: Do group-level variables (GroupSize, GroupOutDegree, GroupInDegree, GroupBetweenness, GroupCloseness, and GroupAggConstraint) explain differences in the group mean number of infections? Which group-level variable(s) best explain between-group variance?

- **Research Question 3**: Do individual-level variables (OutDegree, InDegree, Betweenness, Closeness, and AggConstraint) explain differences in the number of infections? Which individual-level variable(s) best explain within-group variance?
- **Research Question 4**: Do the group-level variables (GroupSize, GroupOutDegree, GroupInDegree, GroupBetweenness, GroupCloseness, and GroupAggConstraint) influence the magnitude of individual-level variables (OutDegree, InDegree, Betweenness, Closeness, and AggConstraint) on explaining the number of infections? Are there significant cross-level interactions?

As indicated in Figure 1, research question 1 tests whether a multilevel analysis is necessary. Based on the result of research question 1, research question 2 identifies the best explanatory variable(s) at the second level (group level). Research question 3 then compares explanatory variables at the first level (individual level). Finally, research question 4 examines the cross-level interactions between individual-level and group-level predictors.

Research Sample

In order to examine the computer virus propagation process and further answer the four research questions raised in the previous section, we collected empirical data of a large-scale organization's social network from the largest social networking website – MySpace. Social networking websites like MySpace provide many communication tools through which computer viruses can spread. In fact, viruses which spread through social networks are more prevalent than other traditional viruses. According to TechWeb Technology News, nine of the top ten most destructive PC malware of all time are mass-mailing worms[3]. In addition, users in social networks expose more personal information and are thus more vulnerable to computer viruses. For instance, Orkut, Google's social networking website, was recently hit by a virus that stole users' financial information and passwords (Boudreau 2006). Analyzing virus propagation in social networks can give us insights into the virus diffusion process.

The sample organization is the fourth largest research university in the U.S with a total enrollment of more than 48,000 students. We first gathered all member data on MySpace with affiliation to this university. Mining the detailed data of each MySpace member enabled us to construct the corresponding social network. The following subsections give detailed descriptions of our research sample.

Social Networking Websites

In general, social network is a set of actors and the relations defined on them (Wasserman and Faust 1994). There are many different social networks embedded in an organization such as functional divisions, project teams, employee email network, and so on. In this paper, for the purpose of studying computer virus propagation, we focus on computer-mediated social networks. Common computer-mediated social networks include email network, instant messaging network, P2P network, etc. More recently, social networking websites (MySpace, Facebook, and LinkedIn, just to name a few) become more and more popular among Internet users. Social networking websites integrate email, instant messaging, and other traditional computer-mediated socialization tools and provide a sophisticated online social environment where users spend on average 1 hour and 22 minutes a day. Among the social networking websites, MySpace, ranked number 15 in the entire U.S. in terms of page hits, has the largest membership base of 20.6 million, and accounts for 10% of all advertisements viewed online in October 2005 (Hempel and Lehman 2005).

Using Perl and RegEx, we collected data about members of MySpace who are current students of the sample university. The number of current student members of this university in MySpace in February 2006 is 14,933, which accounts for more than 30% of all enrolled students. One motivation for us to choose this dataset is that college students are considered the high risk group in terms of computer virus infection and propagation. The relationship from one student member to another on MySpace is uncovered by mining the detailed data published in each member's profile. The resulting network of our research sample is then represented by a directed graph. This network is a social network of a large organization with a well defined boundary according to the members'

[3] "The 10 Most Destructive PC Viruses of All Time" by George Jones, July 05, 2006. TechWeb Technology News. http://www.techweb.com/showArticle.jhtml?articleID=160200005

affiliation. This paper is among the first to investigate the computer virus propagation using real empirical data from a social networking website.

Network Characteristics

We calculate the structural properties of the social network of our research sample using two most widely used social network analysis software, UCINET and Pajek. The results are summarized in Table 1. The formal definitions of these statistics can be found in Wassermann and Faust (1994). These statistics are all network-level measures which give us a snap shot of the structure of the social network. The sample social network is very sparse with a density of 0.0006. Reciprocity rate (0.9833) is very high for this social network which implies most of the relationships between student members are mutual. Although the diameter of the network is 13, the mean node-to-node distance (4.284) is much shorter. Because of high reciprocity rate of the sample network, indegree and outdegree measures share many similarities including a similar centralization measure and similar performance in predicting the spread of computer viruses (which we will discuss later). Closeness varies much more than the other three centrality measures. In summary, our sample social network shows significant clustering and salient subgroup structure which are distinguishing features of social networks (Newman and Park 2003).

Table 1. Network Characteristics

Category	Statistics Name	Statistics Value
Network Type	Directed/Undirected	Directed
Network Size	Number of Nodes	14933
	Number of Edges	140228
	Mean Degree	9.390
Cohesion	Mean Density	0.0006
	Transitivity	0.151
	Clustering Coefficient	0.252
	Reciprocity Rate	0.983
	Mean Node-to-node Distance	4.284
	Diameter	13
	Reachability	0.617
Centralization	InDegree Centralization	0.0200
	OutDegree Centralization	0.0202
	Betweenness Centralization	0.0377
	Closeness Centralization	0.263

Computational Analysis and Results

Virus Propagation Simulation

Computer virus propagation has been widely researched using epidemiology models (Kephart and White 1991, 1993). Among these epidemiology models, SIR (Susceptible–Infected–Recovered) model is most commonly used (May and Lloyd 2001; Chen and Carley 2004). 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 constructed from empirical data to address the research questions posed in the previous section. 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 patching. Every infected node can try to infect its neighbors at all times before it is recovered.

We applied the discrete-time simulation method. 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[4]. For each starting node, we ran the simulation 50 times. So the total number of simulations run is 150000 ($= 3000 \times 50$). Among these simulations, the virus dies out 9774 (6.52%) times and epidemics occur 140226 (93.48%) times. Then we calculated the mean value for the number of infections and used it as the dependent variable. Appendix A gives the pseudo-code for our virus propagation simulation.

The entire network has 14933 nodes in 44 subgroups. We randomly chose 3000 nodes as starting nodes. After excluding the starting nodes whose local subgroup has less than 10 members, our sample consists of 2974 starting nodes and 31 subgroups. The data set is unbalanced with the size of the subgroups ranging from 11 to 1360.

Calculation of Independent Variables

We used Pajek to calculate values for individual-level independent variables – outdegree, indegree, betweenness, closeness and brokerage. Subgroup analysis was also carried out in Pajek. Table 2 and Table 3 provide descriptive statistics and correlations for the individual-level and group-level independent variables, respectively. Most of the correlations are significant and have the expected signs. These high correlations cause multicollinearity problems. In next subsection, we report the reduced HLM model that removes some independent variables to correct the multicollinearity problems.

Table 2. Individual-level Descriptive Statistics and Correlations

Variable	Mean	SD	OutDegree	InDegree	Betweenness	Closeness	AggConstraint
OutDegree	11.77	12.06	—				
InDegree	11.70	11.99	.999**	—			
Betweenness	0.00015	0.00030	.801**	.801**	—		
Closeness	0.19	0.03	.667**	.667**	.508**	—	
AggConstraint	0.28	0.28	-.576**	-.576**	-.371**	-.736**	—

Notes: N = 2974 for individual level.

** Correlation is significant at the 0.01 level (2-tailed).

[4] Chen and Carley (2004) found the ratio of infection rate to recovery rate is between 0.01 and 0.2 with a mean of 0.05.

Table 3. Group-level Descriptive Statistics and Correlations								
Variable	Mean	SD	GroupSize	Group OutDegree	Group InDegree	Group Betweenness	Group Closeness	Group AggConstraint
GroupSize	383.81	455.22	—					
Group OutDegree	21.82	20.50	-.433*	—				
Group InDegree	21.56	20.38	-.427*	1.000**	—			
Group Betweenness	0.00026	0.00025	-.355	.860**	.862**	—		
Group Closeness	0.20	0.02	-.482**	.861**	.864**	.845**	—	
Group AggConstraint	0.18	0.20	.445*	-.600**	-.603**	-.594**	-.889**	—

Notes: N = 31 for group level.

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Research Results

We present our research findings in the order of research questions as follows.

- **Research Question 1**: Are there differences in the impact of starting nodes from different groups?

Table 4 summarizes the results of a basic unrestricted model for research question 1 where the individual-level model is $Y_{ij} = \beta_{0j} + r_{ij}$ and the group-level model is $\beta_{0j} = \gamma_{00} + u_{0j}$. The results indicate that there are significant differences in the impact of starting nodes from different groups with $\chi^2 = 10408.837$ and *p*-value close to 0. In other words, a two-level model is necessary in order to assess individual-level and group-level variables simultaneously.

Table 4. Research Question 1 Results			
Fixed Effect	Group mean γ_{00}	Coefficient	10454.737
		SE	168.172
Random Effect	Group-level error u_{0j}	Variance	850456.929
		df	30
		χ^2	10408.837
		p-value	0.000
	Individual-level error r_{ij}	Variance	470929.033

- **Research Question 2**: Do group-level variables (GroupSize, GroupOutDegree, GroupInDegree, GroupBetweenness, GroupCloseness, and GroupAggConstraint) explain differences in the group mean number of infections? Which group-level variable(s) best explain between-group variance?

To evaluate and compare the impact of group-level variables on group infection numbers, we specify the group-level model as

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * \text{GroupSize}_j + \gamma_{02} * \text{GroupOutDegree}_j + \gamma_{03} * \text{GroupInDegree}_j + \gamma_{04} * \text{GroupBetweenness}_j + \gamma_{05} * \text{GroupCloseness}_j + \gamma_{06} * \text{GroupAggConstraint}_j + u_{0j}$$

while keeping the individual-level model the same as $Y_{ij} = \beta_{0j} + r_{ij}$. Since there are significantly high correlations among these six group-level explanatory variables, we further evaluate a reduced form of this model with only one group-level independent variable to eliminate the multicollinearity problems. We find GroupSize is not significant in explaining group mean virus propagation measures. Snijders and Bosker (1999) discuss a multilevel version of R-squared. Within-group R-squared type statistic is defined as the proportional reduction of error for predicting an individual outcome. Between-group R-squared type statistic is defined as the proportional reduction of error for predicting a group mean. Table 5 compares the model fit using between-group (group-level) variance explained for the remaining five group-level independent variables. We find that the abilities to explain differences in the group means of these five variables vary dramatically. Among them, GroupAggConstraint has the strongest explanation power.

Table 5. Research Question 2 Results
Comparison of Group-level Explanatory Variables

Dependent Variable	Group-level Independent Variable	Group-level Restricted Error	Between-group Variance Explained
Number of Infections	GroupOutDegree	759993.152	0.106
	GroupInDegree	757969.985	0.109
	GroupBetweenness	765245.217	0.100
	GroupCloseness	461364.546	0.458
	GroupAggConstraint	108556.076	0.872

- **Research Question 3**: Do individual-level variables (OutDegree, InDegree, Betweenness, Closeness, and AggConstraint) explain differences in the number of infections? Which individual-level variable(s) best explain within-group variance?

We next investigate how the five individual-level variables affect infection numbers. We find similar multicollinearity problems for the individual-level variables as for the group-level variables. Therefore we reduce the model to only one individual-level variable. Table 6 contains the comparison results for all five independent variables and four dependent variables. We find that individual-level closeness measure far outperforms all the other structural measures.

Table 6. Research Question 3 Results
Comparison of Individual-level Explanatory Variables

Dependent Variable	Individual-level Independent Variable	Individual-level Restricted Error	Within-group Variance Explained
Number of Infections	OutDegree	471101.435	0.000
	InDegree	471100.914	0.000
	Betweenness	471087.654	0.000
	Closeness	257903.545	0.452
	AggConstraint	469533.985	0.003

- **Research Question 4**: Do the group-level variables (GroupSize, GroupOutDegree, GroupInDegree, GroupBetweenness, GroupCloseness, and GroupAggConstraint) influence the magnitude of individual-level variables (OutDegree, InDegree, Betweenness, Closeness, and AggConstraint) on explaining the number of infections? Are there significant cross-level interactions?

Finally we examine the possible cross-level interactions in our starting node effect model. We report the estimation results of the HLM model in Table 7. We find significant interactive effect between individual-level closeness centrality (Closeness) and group-level brokerage (GroupAggConstraint) in explaining the number of infections. Our

final estimation results show that the hierarchical linear models we proposed explain 82.5% of the total variance for the number of infections.

Table 7. Research Question 4 Results			
Fixed Effect	Group mean γ_{00}	Coefficient	12705.333
		SE	344.553
	GroupAggConstraint γ_{01}	Coefficient	-12757.799
		SE	488.344
	Closeness γ_{10}	Coefficient	-8184.990
		SE	1602.908
	Closeness * GroupAggConstraint γ_{11}	Coefficient	52795.386
		SE	1644.882
Random Effect	Group-level error u_{0j}	Variance	71019.594
		df	29
		χ^2	650.658
	Individual-level error r_{ij}	Variance	160271.871
Variance Explained: 82.5%			

Concluding Remarks

As Internet access becomes ubiquitous, so is the spread of computer viruses, causing serious worldwide damage to computer users and organizations in the order of tens of billions of dollars per year. Research to understand how virus spreads and how to effectively defend it is more important than ever. This paper takes the social network perspective to study the computer virus problem and proposes a novel multilevel starting node model for virus propagation. We use hierarchical linear models to simultaneously evaluate the impact of both individual-level and group-level variables on virus propagation processes. We find that individual-level closeness centrality (Closeness) and group-level brokerage (GroupAggConstraint) are the best predictors for the individual-level and group-level variances respectively. Although our research focuses on one particular network flow process – virus propagation, the methodology proposed in this work can be generalized to other processes such as network-based diffusion and so on. The significant impact of both individual-level and group-level structural measures found in this paper can also be applied into firms' decisions on security investment (Cavusoglu et al. 2004) and socio-organizational policies (Baskerville and Siponen 2002).

This paper uses a unique real network data set obtained from the largest social networking website representing an integrated social network of a large organization. We estimate the starting node model based on real empirical network data. The resulting model can be used for computer virus proliferation prediction as well as virus prevention, treatment, and containment strategies.

One key contribution of this paper is the introduction of a hierarchical linear model to explain dynamic network processes using static structural characteristics. We believe there are many interesting extensions to this work. For example, in addition to predicting the number of infected computers, it is useful to know how much time it takes for a computer virus to break out and the time it takes to reach the plateau of infection.

Acknowledgements

We gratefully thank the AE and three anonymous reviewers for their very detailed and helpful suggestions and comments. Any remaining error belongs to the authors.

References

Baskerville, R., and Siponen, M. "An Information Security Meta-policy for Emergent Organizations," *Logistics Information Management* (15), 2002, pp. 337-346.

Borgatti, S. P. "Centrality and Network Flow," *Social Networks* (27), 2005, pp. 55-71.

Boudreau, J. "Social Networks a Hacker's Paradise," *San Jose Mercury News*, June 22, 2006.

Burt, R. S. *Structural holes: The social structure of competition*, Harvard University Press, Cambridge, Massachusetts, 1992.

Brumley, D., Liu, L., Poosankam, P., and Song, D. "Design Space and Analysis of Worm Defense Strategies," in *ACM Symposium on InformAtion, Computer and Communications Security* (ASIACCS'06), Taipei, Taiwan, March 21-24, 2006.

Cavusoglu, H., Mishra, B., and Raghunathan, S. "A Model for Evaluating IT Security Investments," *Communications of the ACM* (47:7), 2004, pp. 87-92.

Clauset, A., Newman, M. E. J., and Moore, C. "Finding Community Structure in Very Large Networks," *Physical Review E.* (70:066111), 2004.

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), pp. 823-833, 2004.

Costenbader, E., and Valente, T. W. "The Stability of Centrality Measures When Networks are Sampled," *Social Networks* (25), 2003, pp. 283-307.

Freeman, L. C. "Centrality in Social Networks: Conceptual Clarification," *Social Networks* (1), 1978/79, pp. 215-239.

Ganesh, A., Massoulie, L., and Towsley, D. "The Effect of Network Topology on the Spread of Epidemics," in *Proceedings of IEEE Infocom*, Miami, Florida, March 13-17, 2005.

Gordon, L. A., Loeb, M. P., Lucyshyn, W., and Richardson, R. *CSI/FBI Computer Crime and Security Survey*, Computer Security Institute, 2006.

Hempel, J., and Lehman, P. "The MySpace Generation," *BusinessWeek* (3963), December 12, 2005, pp. 86-96.

Hofmann, D. A. "An Overview of the Logic and Rationale of Hierarchical Linear Models," *Journal of Management* (23:6), 1997, pp. 723-744.

Kephart, J. O., and White, S. R. "Directed-graph Epidemiological Models of Computer Viruses," in *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, 1991, pp. 343-359.

Kephart, J. O., and White, S. R. "Measuring and Modeling Computer Virus Prevalence," in *Proceedings of the 1993 IEEE Computer Society Symposium on Research in Security and Privacy*, Oakland, California, 1993, pp. 2-15.

May, R. M., and Lloyd, A. L. "Infection Dynamics on Scale-free Networks," *Physical Review E*. (64:066112), 2001.

Moore, C., and Newman, M. E. J. "Epidemics and Percolation in Small-world Networks," *Physical Review E*. (61:5), 2000.

Newman, M. E. J., Forrest, S., and Balthrop, J. "Email Networks and the Spread of Computer Viruses," *Physical Review E*. (66:035101), 2002.

Newman, M. E. J., and Girvan, M. "Finding and Evaluating Community Structure in Networks," *Physical Review E.* (69:026113), 2004.

Newman, M. E. J., and Park, J. "Why Social Networks Are Different from Other Types of Networks," *Physical Review E.* (68: 36122), 2003.

Pastor-Satorras, R., and Vespignani, A. "Epidemic Spreading in Scale-free Networks," *Physical Review Letters* (86:14), 2001.

Snijders, T. A. B., and Bosker, R. J. *Multilevel Analysis*, SAGE Publications, London, UK, 1999.

Wang, Y., Chakrabati, D., Wang, C., and Faloutsos, C. "Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint," *Symposium on Reliable Distributed Systems* (SRDS), 2003, pp. 25-34.

Wang, C., Knight, J., and Elder, M. "On Computer Viral Infection and the Effect of Immunization," *Proceedings of the 16th Annual Computer Security Applications Conference (ACSAC'00)*, New Orleans, Louisiana, 2000, pp. 246-256.

Wassermann S., and Faust, K. *Social Network Analysis*, Cambridge University Press, Cambridge, UK, 1994.

Zou, C. C., Towsley, D., and Gong, W. "Email Worm Modeling and Defense," *13th International Conference on Computer Communications and Networks (ICCCN'04)*, Chicago, Illinois, 2004, pp. 409-414.

Appendix A. Pseudo-code for Virus Propagation Simulation (SIR)

```
t = 0
InfNum = 1
CurrentInfNum = 1
StartingNode.blnInfected = True
StartingNode.blnCurrentRound = True
WHILE (t < MaxIter AND CurrentInfNum > 0)
        FOR each individual node
                IF (blnInfected = True AND blnRecovered = False AND blnCurrentRound = True) THEN
                        FOR each neighbor
                                IF (the neighbor NOT infected) THEN
                                        Try to infect the neighbor
                                        IF (the neighbor gets infected) THEN
                                                Set its blnInfected = True
                                                InfNum = InfNum + 1
                                                CurrentInfNum = CurrentInfNum + 1
                                        END IF
                                END IF
                        NEXT neighbor
                        Try to recover the individual node
                        IF (the individual node gets recovered) THEN
                                Set blnRecovered = True
                                CurrentInfNum = CurrentInfNum - 1
                        END IF
                END IF
        NEXT individual node
        Set blnCurrentRound to True for all the infected nodes
        t = t + 1
END WHILE
```