Inf Technol Manag (2013) 14:257–268
DOI 10.1007/s10799-013-0163-z

Effective immunization of online networks: a self-similar selection approach

Byung Cho Kim · Sunghwan Jung

Published online: 19 July 2013

Abstract This paper proposes a self-similar selection method as an alternative to existing immunization strategies for online networks. Given the self-similar characteristics of online networks which are shown to have fractal and scale-free structure, we presume that the self-similar selection which is well developed in physics outperforms random or targeted vaccination based on incoming or outgoing connections. We examine the effectiveness of the proposed self-similar selection method with random vaccination and other different types of targeted vaccination strategies in terms of delaying the spread of computer virus over a scale-free computer network constructed using real-world World Wide Web data. Our computer simulation results indicate that the self-similar selection method is more effective in deterring virus propagation than the existing vaccination strategies. In addition, vaccination based on self-similar selection is practical since it does not require detailed information about network morphology at the individual node level, which is often not easy to observe. Our findings have significant implications for both policy makers and network security providers.

Keywords Information security · Virus propagation · Network immunization · Self-similar selection

B. C. Kim (✉)
Department of Logistics, Service and Operations Management,
Korea University Business School, Seoul 136-701, Korea
e-mail: bkim@korea.ac.kr

S. Jung
Department of Engineering Science and Mechanics,
Virginia Tech, Blacksburg, VA 24061, USA
e-mail: sunnyjsh@vt.edu

1 Introduction

In the past couple of decades, the rapid development of information technology and significant drop in the cost of computing devices have increased Internet accessibility. More recently, the proliferation of electronic commerce and social networks in a fast-evolving mobile world, where people enjoy wireless and mobile access to the Internet, has led to intensive interconnectivity among people, computers, and web sites. As the Internet becomes a crucial part of everyone's daily life and online networks get bigger and more complex, network security becomes more important since the outcomes of security failures can be disastrous. Accordingly, computer scientists and software developers have been working on developing security countermeasures including anti-virus software, software patches, firewalls, and encryption software.

Despite these efforts on the technical side, recent statistics show that online networks still suffer from dangerous security problems. According to a recent *New York Times* article, a new digital plague, called Conficker or Downadup, has infected millions of computers, seeming to be just the first step of a multistage attack [20]. George Stathakopoulos, a general manager for Microsoft's Security Engineering and Communications group, said, "It's all about defense in depth," when we need to deal with this type of massive and well organized virus and worm attacks. Then, the critical questions become how we should manage security, what form of security management policy is the most effective, and under what conditions. Thus, security management should be a holistic approach which incorporates technical, behavioral, and managerial aspects of information security.

Online networks are formed based on people's interest-seeking behavior. Shared interests among people breed connections and eventually form a network in the borderless and fast-growing cyber space. People make a decision on what web sites they link to their web documents based on their own interests, which determines the network structure of the World Wide Web. It has been expected by scientists that these networks formed online follow random structure due to enormous diversity among people's interests and countless number of web sites from which they choose. Surprisingly, Albert et al. [2] find evidence that defies this expectation. They show that World Wide Web is a scale-free network, in which the connection (or degree) distribution follows a power law (i.e., $P(n) \propto n^{\alpha}$ where $P(n)$ is the probability of node connections, n is the number of node connections and α is the exponent). On the other hand, the random network has the Poisson distribution of node connection. Topologically, the scale-free network has a few highly connected nodes and many nodes have very few connections. Simply speaking, the nodes of a scale-free network are not randomly or evenly connected. Instead, the scale-free networks have hubs that have many connections and the ratio of hubs to the number of other nodes stays the same as the network size changes.

Researchers in various fields such as physics, biology and computer science have studied effective virus prevention strategies over various types of networks. Although it is ideal to immunize the entire network, it is often not feasible due to a number of reasons including cost, poor user awareness, and defective installation [29]. Given the barriers to the blanket virus control policy, it is crucial to find an effective way of preventing virus propagation over the online network. Currently, the most widely adopted virus prevention strategy for online networks is random vaccination of computers with anti-virus software or random removal, which turned out to be an ineffective way of managing security [22]. On the other hand, targeted vaccination, in which the highest degree vertices are immunized, is proven to be more effective than random vaccination in theory, although knowing the degree of connection among nodes is often not feasible in reality. Most of the existing algorithms developed to immunize computer networks do not consider the scale-free structure of the networks.

Grounded on well-established theory in physics, this study proposes self-similar selection method as a possibly more effective and practical alternative to other existing vaccination strategies [3, 5, 9]. Our presumption about the effectiveness of the proposed method in terms of protecting online networks was grounded on two findings in the physics literature. First, various online networks were proven to be scale-free [6]. Second, due to the self-similar structure of scale-free networks, we presume that protecting nodes based on the self-similarity may outperform incoming or outgoing connection-based approaches to determining hubs. Given the scale-free structure of the online networks such as the World Wide Web and online social networks [2], the self-similar selection approach is assumed to be effective in deterring virus propagation. We compare the effectiveness of the proposed self-similar selection method with random vaccination and other different types of targeted vaccination strategies in terms of delaying the spread of computer viruses over a scale-free computer network reconstructed using real-world World Wide Web data. Consistent with existing studies (e.g., [3, 22]), the results indicate that targeted vaccination is more effective than random vaccination. Our simulation results support our premise that the proposed self-similar method outperforms random and other existing targeted vaccination strategies in terms of delaying virus propagation. Our findings also show that an optimal group distance under the self-similar selection method exists to minimize the number of infections in networks. When the number of vaccines is limited, the self-similar selection and targeted vaccination based on incoming connection turn out to be similarly effective. At higher levels of vaccination, self-similar selection is shown to be significantly more effective than any other existing vaccination strategy.

We aim to make contributions to the literature in the following ways. First, we propose self-similar selection method as an alternative to existing computer vaccination strategies. Given the self-similarity of scale-free networks such as the World Wide Web, we presume that a self-similar selection approach is effective in terms of delaying virus propagation. Our simulation results show strong evidence to support our argument. Second, the proposed self-similar selection method is more practical than targeted vaccination strategies that require observation of node-level connection. In reality, it is often a challenge to select target nodes since the number of incoming and/or outgoing connections is often not observable. Also, it is not realistic to assume that a single authority has control over the network like the Internet or the World Wide Web. The self-similar selection method does not require such information on the network morphology. Due to self-similarity among layers of the scale-free network, an administrator is able to select the target nodes without knowing the detailed structure of connection at the individual node level. As long as the network structure at any upper level is known, target nodes can be chosen due to self-similar structure at different levels of the network. Third, our study captures reality by examining the impact of the time lag between virus detection and vaccine development while existing studies do not consider this time lag. Once a new virus is discovered, spreads out, and reaches a critical mass, scientists are able to develop a vaccine against the identified virus. Thus, vaccines can be only applied for a short period of time after an outbreak. In order to reflect such reality, we need to model the case in which we can start vaccinating vertices after a certain part in the network has already been infected. We consider the time lag and examine its impact which may give significant implications from a practical standpoint. Our findings have ramifications for network providers and policy makers who want to establish effective security management policies and efficiently deal with computer viruses, and give them a guideline to optimally vaccinate their networks given their budget size.

This paper is organized as follows. In Sect. 2, we provide background for our study including security management policy, network topology, and virus infection. Section 3 discusses the theoretical interpretation of self-similar (fractal) structures, and provides the detailed methodology of a self-similar selection method. Simulation procedures and the results are presented in Sect. 4. Section 5 discusses implications of our findings and suggests future research directions.

2 Research background

Network security is becoming an emerging research topic in the domain of information systems [4, 15, 16]. Until recently, the most exclusively studied network was a random network which consists of nodes with random connected neighbors [11]. This randomness of node connections produces the Poisson distribution for the number of node connections. Lately, Barabasi and Albert [6] discovered a new structure of network while studying how Web pages are connected to one another by hyperlinks. They found that this new structure of the World Wide Web consists of hubs, that is, nodes with a large number of links, and that the distribution of node connections follows a power law. This power law distribution does not depend on the number of node connections, therefore is called “scale-free network”. Thus, they named it a “scale-free” network. Other examples of a scale-free network provided in their paper include a network of sexual relationships, an e-mail network, and a cellular metabolic network. The discovery of a scale-free network has significant implications in that random-network theory is not able to explain the existence of hubs. In the domain of information systems, Oh and Jeon [23] introduced the concept of scale-free network in their study of membership herding in open-source software community. They examine the impact of external factors on the stability of a network and the mediating effect of different network topologies. Complex network topologies of real-work networks and algorithms to generate synthetic graphs have been extensively studied by computer scientists and physicists. Leskovec and Faloutsos [18] proposed an algorithm to generate a synthetic graph from a large network, while matching the original network’s properties such as degree distribution, diameter and spectrum. They found that the proposed method fits the target graph well. Akoglu and Faloutsos [1] presented a graph generator model, Random Typing Generator, and showed that it meets desirable properties of a graph generator model. Coupled with the idea of dissemination of information, recent research models complicated network structure and examines its impact on the spread of epidemics. Chakrabarti et al. [8] developed a Non-linear dynamical system (NLDS) that models virus propagation in any arbitrary network. They demonstrated that the proposed model has predictive power for arbitrary graphs. Ganesh et al. [12] investigated how network topology affects epidemic spread which they model as a contact process over a finite undirected graph. They identified the topology properties that determine the persistence of epidemics.

Our paper is relevant to the literature on virus propagation over computer networks. Virus propagation over computer networks has been widely studied for the last couple of decades [13, 21]. Recently, controlling virus and worm propagation has captured interest of researchers with managerial and economic perspectives [24, 30]. The most widely adopted vaccination strategies are random vaccination using anti-virus software which turned out to be inefficient [22]. Several existing studies show that targeted vaccination, in particular, simple removal of the nodes with highest degree of connections is very effective in most networks [3, 7, 9]. Newman et al. [22] analyzed a directed social network over which an email virus spreads and validated the effectiveness of targeted vaccination over random vaccination. They examined the spread of viruses within a community due to the limitation of data which omits connections from outside the community of the identified users. Tong et al. [28] examined the vulnerability of the large real graphs and proposed a near-optimal and scalable algorithm for effective immunization. They found that the proposed algorithm outperformed existing methods. Our study is closely related to the stream of research that examines immunization strategies over a scale-free network. Balthrop et al. [5] studied the dynamics of network and the effective virus prevention strategies over a network which is not necessarily scale-free. They also argued that virus propagation over a scale-free network is resilient to random vaccination, while targeted vaccination is more effective than random vaccination. Their results showed that throttling, an alternative dynamic mechanism for the control of contagion, is most effective in the case that viruses propagate as traffic with the higher density than normal network communications. Cohen et al. [10] proposed acquaintance immunization which requires no knowledge of the node degrees unlike targeted immunization strategies. They show that the proposed acquaintance immunization method is efficient for networks of any broad degree.

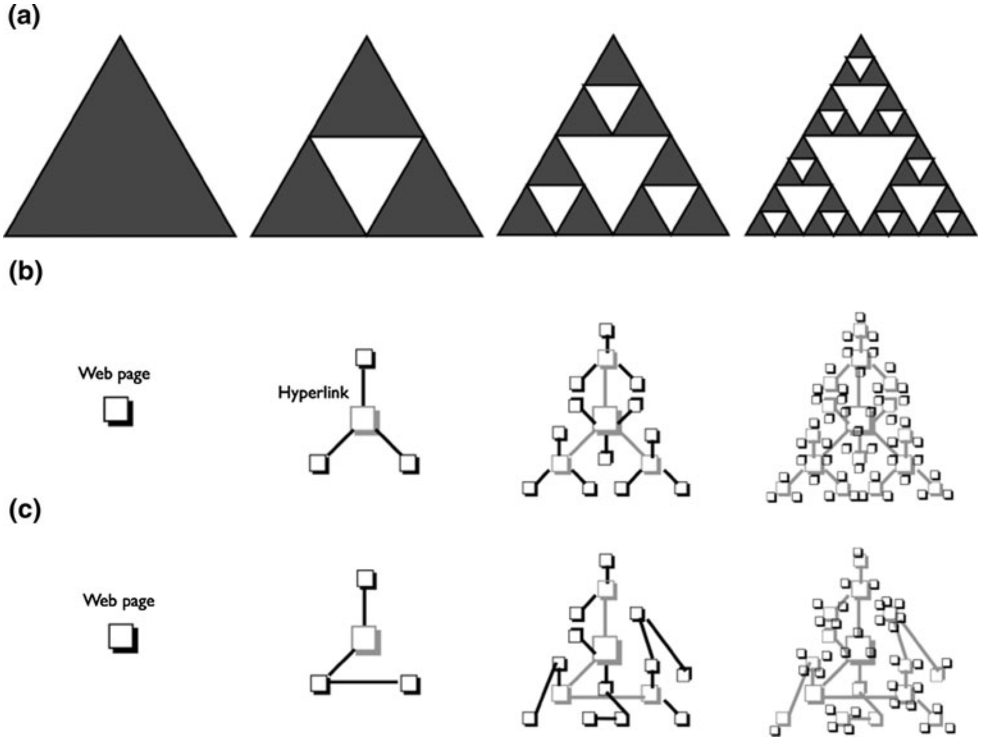
3 Self-similarity of complex networks

Physical systems out of equilibrium appear to be self-similar in different length scales. This self-similar structure has a conventional measure, such as dimension, characterizing its morphology. Commonly, only dimensions of integers have been known such as a 1D line, a 2D disk, and a 3D sphere. Since the dimension can be a fractional number for a complicated structure, this pattern is often called "fractal" as shown in Fig. 1a. In other words, the self-similarity (the property of being the same topological or geometrical patterns at different scales) is a fundamental property of fractals. Fractal examples are found in various morphologies and scientific systems; coastlines, clouds, electrochemical deposition, viscous fingering, porous rocks, turbulence flows, and many more. Not only physical objects but also abstract structures have been identified in dynamical systems such as strange attractors in phase space as illustrated in Fig. 1b [14].

The most common method for determining fractal dimension is a box-counting method [19]. A geometrical interpretation of dimension is as an exponent in the scaling of an object's content with the size as (Content ~ size$^{\text{dimension}}$). For an example, of 2D objects, the content (area) of a disk is proportional to the square (2) of its size (radius). For a 3D object, the content (volume) of a sphere is proportional to the cube (3) of its size (radius). Therefore, mathematical definition of dimension becomes (dimention = $\lim_{\text{size}\to 0} \log(\text{content})/\log(\text{size})$). However, this simple definition is not easily applied for complex objects, such as coastline, viscous figuring, and so on. Box-counting methods, therefore, provide the practical numerical tools to calculate the dimension of fractal objects. In these methods, the content is the number of boxes to cover the object with a ball of the equal size in a radius.

Mathematical description of the box counting method is summarized as follows. Suppose that set A is the fractal of our interest. Let $C(A,r) = \{B_1, B_2\cdots,B_n\}$ be a finite covering of A into sets whose box has a diameter, that is, $A \subset \bigcup_{i=1}^{k} B_i$ Then, the covering function $\Gamma(A,D,r)$ is defined as $\inf_{C(A,r)} \sum_i \delta_i^D$ where inf is the minimum over all coverings satisfying $\delta_i < r$ For fractal objects, this covering function shows a power-law with r. The box-counting method is to use the grid boxes rather than to evaluate the infimum over all coverings. For the box-counting methods, the covering function becomes $\Gamma(A,D,r) = \sum_i r^D = n(r)r^D$ where $n(r)$ is the number of nonempty grid boxes. One can measure the fractal (Haudourff) dimension given as $D = \lim_{r\to 0} \log(1/n(r))/\log(r)$. This measured dimension

Fig. 1 Schematics of **a** fractal structure, **b** self-similar scale-free network, and **c** random network



represents how self-similar the structure is. In general, a system with the higher fractal dimension forms a more geometrically complex structure than a system with the lower dimension.

Different from indentifying fractal structure out of the physical systems, it is not simple to find fractal features in the abstract object such as networks. Recently, Song et al. [25] found fractal structure in complex scale-free networks such as the World-Wide Web, metabolic networks, and protein interaction network. This study shows that complex networks embed self-similar structures through the renormalization procedure in various network scales. Using the box-counting methods inspired by renormalization procedure, a finite self-similar scaling exponent is obtained which leads to fractal dimension (i.e., 4.1 dimension for the World Wide Web) network [25–27]. As illustrated in Fig. 2, this box-counting method is defined as follows; at a given value of the box size l_B (we call it "the group distance" in our context), count the number of groups (boxes) required to cover the entire network under a condition that each group (box) contains nodes separated by a distance less than l_B. For example, Fig. 2a illustrates the box-counting method with $l_B = 2$. First, starting with the node 1, one selects a new group (a dashed circle) having all nodes connected within 2 connections. Then, one repeats this process to find other exclusive groups starting from the node 2 and 3 as in Fig. 2a. Even though the node 4 cannot find other node within this rule, it is assigned as individual box [25]. When $l_B = 3$ (see in Fig. 2b), one finds nodes connected to the node 1 within 3 connections and repeat this over other nodes. Consequently, one finds three distinct groups ($N_B = 3$) when $l_B = 3$. As shown in Fig. 2, the total number of groups (N_B) decreases with increasing the box size (l_B). Later, several box-counting methods using random sequencing are introduced for their convenience to implement in codes and shorter process time [17, 27]. In this study, we use the method introduced in Rozenfeld et al. [27], which gives the same results as others (as shown in Fig. 3). This box-counting method allows us to capture self-similar structures in complex networks. Based on the resultant structures, various strategies preventing virus spreading are tested in the following sections.

4 Simulation

4.1 Empirical data and simulation procedure

The proliferation of the World Wide Web has significantly contributed to the massive adoption of the Internet, which is now a commodity. Although many people perceive the World Wide Web as the synonym of the Internet, they are different. The Internet is a global system of interconnected computer networks; whereas the World Wide Web is a system of interlinked hypertext documents contained on the Internet (Wikipedia). The World Wide Web is just one of the ways of accessing information over the Internet. The World Wide Web has been growing rapidly for the past decade and its size estimated by Google's index is more than 30 billion web pages as of March 2013 (www.worldwidewebsize.com). Understanding how the World Wide Web is structured and has evolved is important since any other network on the Internet cannot be independently formulated from the structure of the World Wide Web. To examine the effectiveness of different vaccination

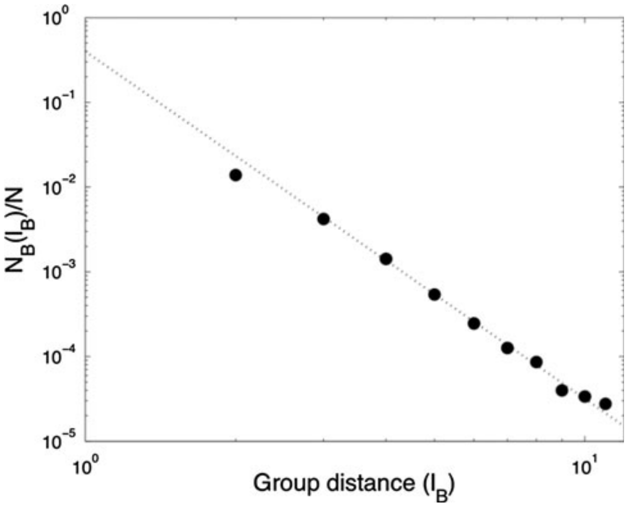
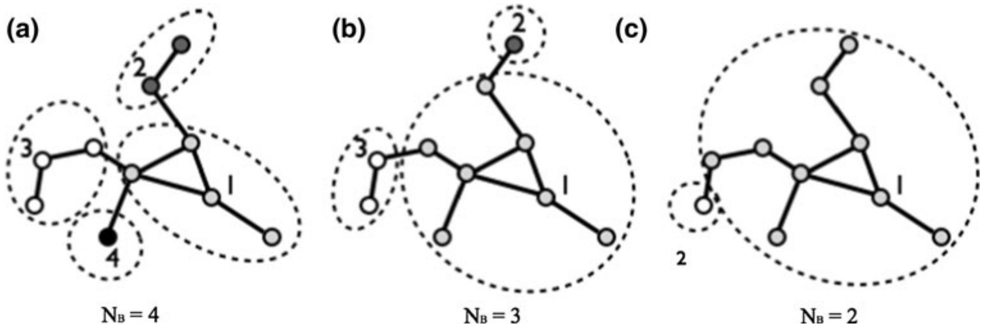
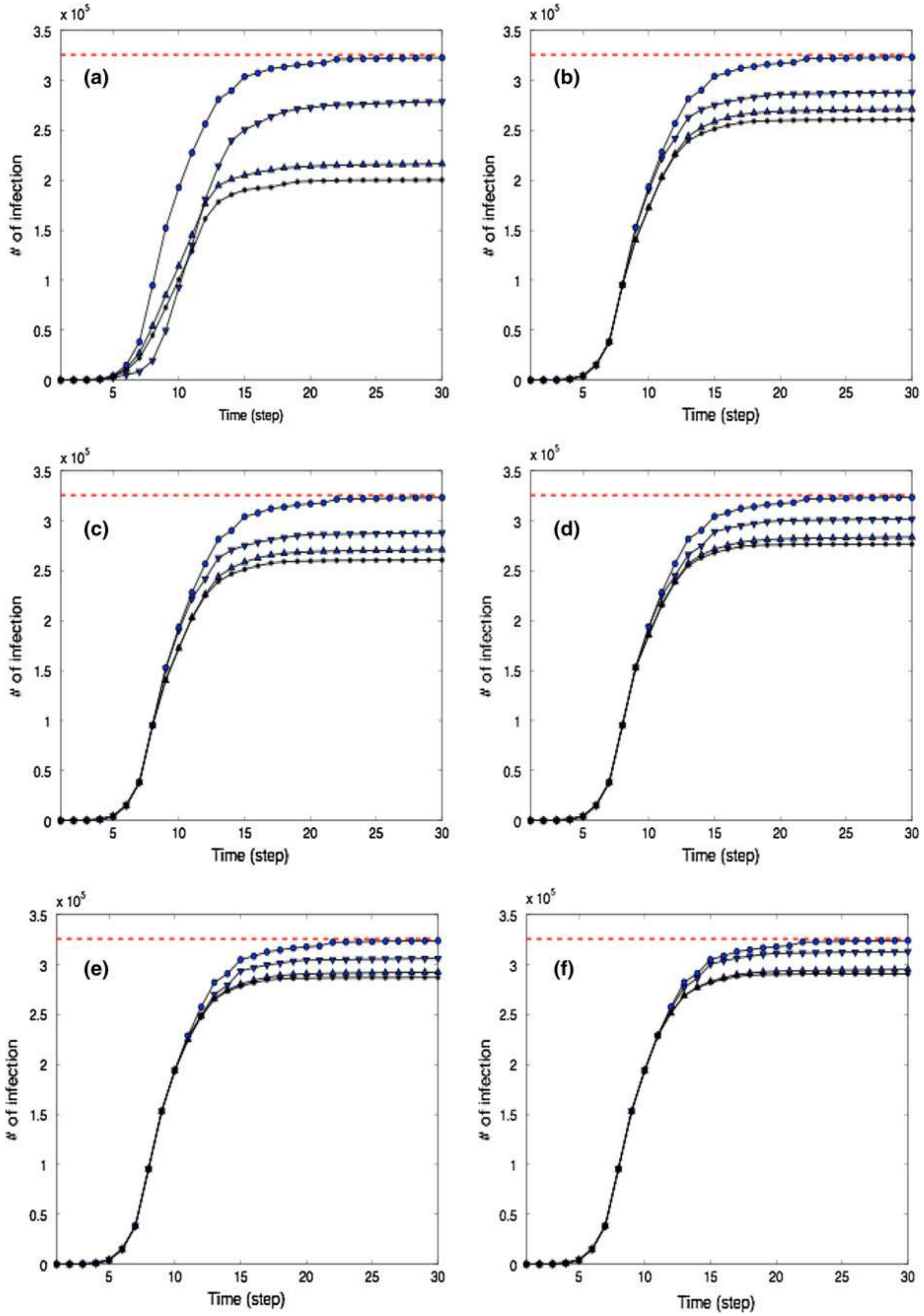


Fig. 3 A log–log plot of a normalized number of groups (N_B/N) versus a group distance (l_B). The *slope* in this figure means the fractal dimension (D). The *dotted line* is a slope with the fractal dimension (D = 4.1)

Fig. 2 Illustration of the box-counting method in a simple network with **a** $l_B = 2$, **b** $l_B = 3$, and **c** $l_B = 4$




(a)
(b)
(c)
(d)
(e)
(f)
x 10^5
# of infection
Time (step)
0
0.5
1
1.5
2
2.5
3
3.5
5
10
15
20
25
30

◀**Fig. 4** Virus propagation over complex network vaccinated at different time stages; 1,000 nodes are vaccinated (**a**) before the virus spreads, or after **b** 5 %, **c** 10 %, **d** 20 %, **e** 30 %, **f** 50 % infection. *Circles* are from random selection, *downward triangles* are from out-going selection, *upward triangles* are from in-coming selection, and *asterisks* are from the self-similar selection

strategies on the real-world network, we use the real World Wide Web structure data to reconstruct the network. The data was obtained from an online source [www.nd.edu/~networks]. The data contains 325,729 documents and 1,469,680 links. To determine the topological structure of the World Wide Web, the graphical representation has been used containing nodes (documents) and connections (directional links between two documents). More details about the data can be found in Albert et al. [2].

Based on the original data, we reconstructed the graphical representation of the World Wide Web and tested the effectiveness of various vaccination strategies. The effectiveness of vaccination is measured by investigating the trend of virus propagation on nodes in time and by looking at its asymptotic spreading at a given number of vaccinations. We propose a self-similar selection method as an alternative to existing immunization strategies, including random vaccination and targeted vaccination based on incoming or outgoing connection. While vaccinating the nodes with highest number of connections often works well for undirected networks, the effectiveness of vaccination based on incoming or outgoing connections turns out to be significantly different for directed networks such as the World Wide Web or an email network [22]. Under random vaccination, nodes are randomly selected and vaccinated. Targeted vaccination involves selection of nodes in order of incoming or outgoing connections. Existing studies show that targeted vaccination is more effective than the random vaccination in terms of delaying the spread of computer viruses [22, 29].

In this study, we propose the self-similar selection method as an effective vaccination strategy for computer networks and compare it with the aforementioned existing vaccination strategies. Details of simulation based on the self-similar selection method are the following. Using the box-counting algorithm (described in Sect. 3) to characterize the self-similar network, we classify a network into a smaller number of groups. A number of groups (N_B) are determined by the distance (l_B) between nodes, which is analogous to the box size in the fractal analysis. We choose the same number of nodes within each group structured by the box-counting method, named as "*self-similar selection*". This method will allow us to identify which groups and nodes are crucial in the network by having many connections to other nodes. Self-similar selection is defined as selective vaccinations to the nodes identified through the box-counting method. We first test the effectiveness of various vaccination strategies, including (1) random vaccination, (2) targeted vaccination based on incoming connections, (3) targeted vaccination based on outgoing connections, and (4) vaccination based on self-similar selection, in terms of delaying virus propagation over the network constructed using the World Wide Web data. Under each strategy, N nodes are selected, then vaccinated. Infected nodes spread the virus to nodes connected to them at each time step. This time step is not a physical unit, but a unit performed in the computer simulation. Our simulation model reflects reality by considering a time lag between virus initiation and introduction of a vaccine. Existing studies in physics and computer science literature focus on the complete vaccination that is vaccinating the infected nodes at the time of virus detection. This way of modeling lacks grounding in reality since vaccine developers need time to develop and test a vaccine after a new virus is detected, which holds for both computer viruses and biological viruses. Vaccine can be applied after the identified virus disperses over the network to a certain extent. For example, a vaccine to protect against 2009 H1N1 first became available early October of 2009, 6 months after a public health emergency of international concern was declared by the World Health Organization in April, when the first two cases of the H1N1 virus were reported in the United States along with hundreds in Mexico. We examine the effectiveness of different vaccination strategies at different stages in terms of severity of virus spread. We analyze 6 different cases with 0, 5, 10, 20, 30 and 50 % of virus infection ratio and examine virus propagation over the network with 1,000 vaccinated nodes. We then investigate how an increasing number of vaccinations effectively delays virus propagation. We count the number of infected nodes after 30 time steps given different numbers of vaccinations (10, 100 and 1,000) under each of the four vaccination strategies.

4.2 Results of the simulation

We first compare the proposed vaccination based on self-similar selection with three existing vaccination strategies: random vaccination, targeted vaccination strategies based on incoming connections, and targeted vaccination strategies based on outgoing connections. We test how a computer virus spreads over the World Wide Web network. On each simulation, we examined four different strategies: 1,000 nodes are vaccinated based on the random selection, the selections after sorting nodes in the descending order of incoming or outgoing connections, and the selection based on self-similar structures with $l_B = 8$. We then observe the number of infected nodes in the network. To understand the impact of the extent of virus infection which may increase with the lag between virus initiation and vaccine introduction on the effectiveness of vaccination strategies, we examine six different cases based on the timing of vaccination. The results are illustrated in Fig. 4.

We test how a virus spreads at different timing of vaccination under each vaccination strategy as shown in Fig. 3. Since virus spreading is simulated on the computer, not physical time but rather simulation time steps are considered. All six graphs show that the self-similar selection outperforms the existing vaccination strategies. Consistent with the literature in physics and computer science, random vaccination turns out to be the least effective in protecting the network from virus infection and targeting nodes based on incoming connections is more effective than vaccinating nodes with a higher number of outgoing connections. Interestingly, our proposed vaccination based on self-similar selection is shown to be the most effective regardless of the timing of vaccination.

One crucial parameter in the self-similar selection is group distance l_B. In general, self-similar selections with the larger group distance have a lesser number of groups. This is depicted in Fig. 2. Here, we test the impact of group distance on the effectiveness of self-similar selection in terms of delaying virus spreading. In this test, we applied vaccinations to 1,000 nodes selected in different structures of grouping. Figure 5 illustrates our findings.

Figure 5 shows that there exists an optimal group distance in networks in terms of preventing virus propagation. As the group distance increases, the number of infected nodes changes nonlinearly with showing the best performance with $l_B = 8$. The existence of an optimal group distance has implications from both economic and policy perspectives. Given the limited resources (e.g., number of vaccines available), the proposed vaccination based on self-similar selection can be optimized in a way to maximize its cost-efficiency. This is not the case for other vaccination strategies under which vaccines are applied on the nodes with the highest number of connections or randomly, meaning there is no flexibility or optimization involved. Policy makers or network providers may manipulate the group distance so as to effectively protect the network from threats by applying corresponding policies. For example, an operator of a social networking site may allow its users to choose different layers of friends with respect to their intimacy in order to get grouping trends with the desired distance. This way of helping members formulate the network with a generically effective shield will save significant social cost of securing cyber space.

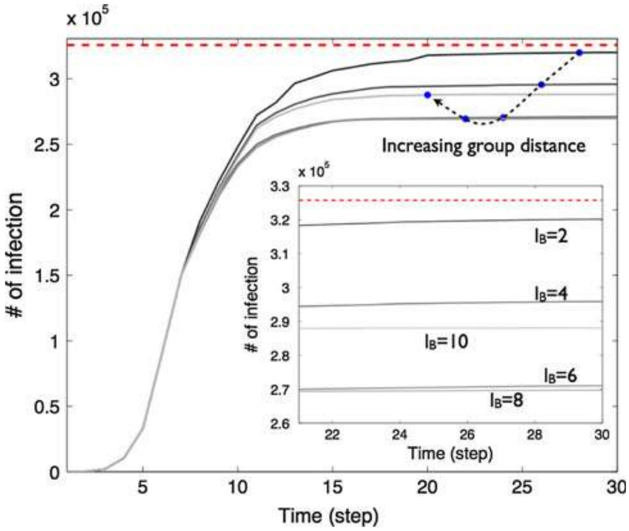
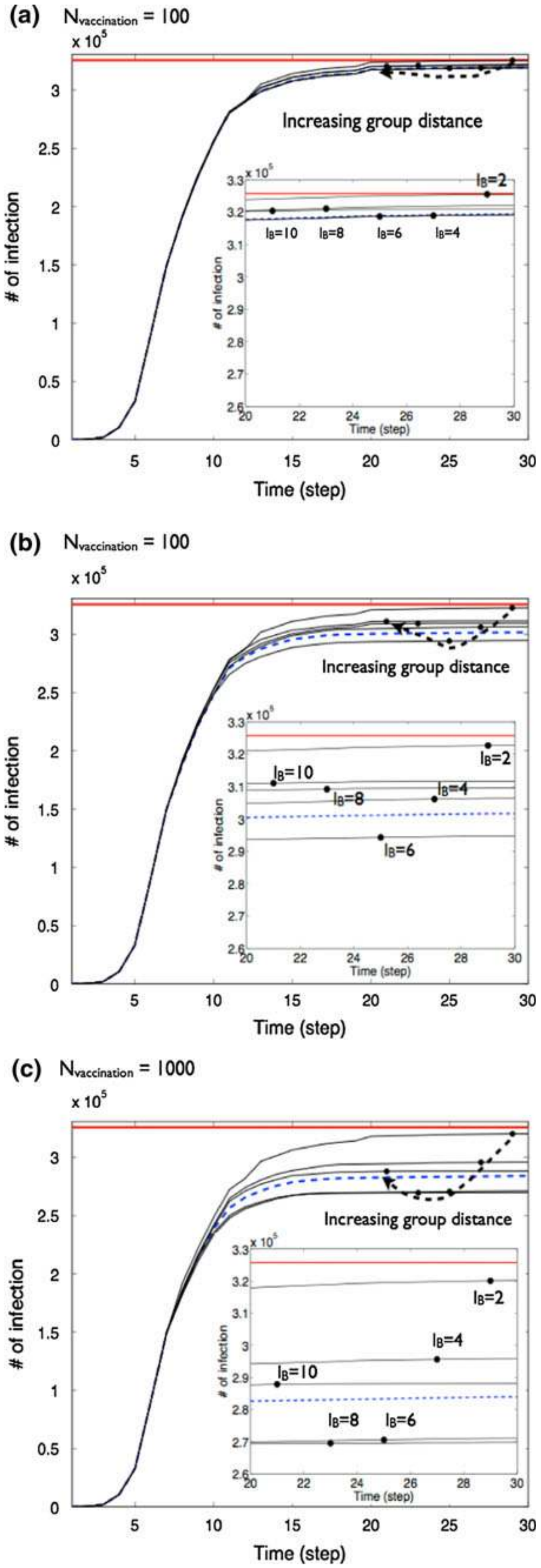


Fig. 5 Virus spreading versus vaccinations selected based on a self-similarity with different group distances; A group distance is increased from 2 to 10 with an interval 2. As the virus spreads over a network, vaccination selected with $l_B = 8$ gives the less number of infection in a complex network

We then compare the self-similar selection with targeted vaccination based on incoming connections at different levels of vaccination. Recall that incoming-connection-based targeting is verified to be most effective among three existing vaccination strategies in the previous simulation. We count the number of infected nodes after 30 time steps at three different numbers of vaccination: 10, 100 and 1,000. For the self-similar selection method, we use different levels of group distance to see the impact of optimization on effectiveness. Figure 6 illustrates our simulation results.

Figure 6 shows how many nodes are infected after the specified numbers of vaccinations are given in the network. It turns out that the vaccination based on self-similar selection outperforms the existing vaccination strategies. When a small number of vaccinations is given (N = 10), the optimized self-similar selection ($l_B = 6$) and incoming selection are similarly effective. When more vaccinations are given (N = 100), the optimized self-similar selection ($l_B = 6$) outperforms incoming selection while not optimized self-similar selection methods do not. We find the consistent result at the level of 1,000 vaccinations is that the vaccination based on self-similar selection with optimal group distance is better than other existing vaccination strategies. These findings imply that the vaccination based on self-similar selection method is better than or equal to targeted vaccination based on incoming connections. Especially, when the perfect security is prioritized with flexible budget, meaning that if sufficient numbers of vaccines can be given, policy makers can expect the best outcome from the self-similar selection method. Even with limited availability of vaccines, the vaccination based on self-similar selection method is similarly effective to targeted vaccination based on incoming connection, which is the most effective among existing methods.



◀**Fig. 6** Virus spread versus vaccinations selected based on a self-similarity with different group distances; A group distance is increased from 2 to 10 with an interval 2. As the virus spreads over a network, vaccination selected with $l_B = 6$ or $l_B = 8$ (when (1) 10, **b** 100, and **c** 1,000 vaccinations are distributed) leads to the minimal infection in a complex network, outperforming the selection based on incoming nodes (*blue lines* in plots) (Color figure online)

5 Discussion

In this study, we propose the vaccination based on self-similar selection as an alternative approach to vaccinating a computer network. We find the evidence to support the effectiveness of the proposed self-similar selection method. The most widely used virus prevention strategy in practice is random vaccination, the effect of which in terms of delaying virus propagation is minimal [5]. Recently, physicists and computer scientists find that targeted vaccination significantly outperforms random vaccination on various networks [22]. Physicists also have shown that online networks have fractal and scale-free structures [25]. Reflecting on these generic features of online networks, the vaccination incorporating fractal and self-similar structure should be effective in preventing virus propagation rather than other targeted vaccinations. Our findings have the following implications to policy makers and online network providers.

First, the vaccination based on self-similar selection method has shown to outperform the existing vaccination strategies. We find that there exists an optimal group distance for the self-similar selection method. Even when not optimized, the self-similar selection is more effective than the existing vaccination methods. Especially when a sufficient number of vaccines can be applied, the self-similar selection gives significantly better results than any other vaccination strategy in terms of delaying virus propagation. Second, the self-similar selection method is practically applicable. The relative effectiveness of a targeted vaccination compared to random approach has been proven by a number of researchers. Currently, the most widely adopted targeted vaccination is removing or immunizing the nodes with the highest number of connections. From a practical stand point, it is often difficult if not infeasible to discover the connections from one node to another. In the context of the World Wide Web, hyperlinks from a website to other sites are often not easy to observe due to enormous size of the network and the dynamical changes. Thus, despite the relative superiority of the targeted vaccination to random vaccination proven in theory, targeted vaccination has not been widely adopted in practice. It is significantly easier to observe partial networks at any level of the entire network, each of which can be considered as a group formed based on self-similarity. Due to the scale-free nature of the online networks, selection of target nodes is feasible with observation of the network structure from a partial network. Third, our study models reality of vaccination. While the existing studies do not consider the time lag between virus initiation and vaccine development, we model the lag assuming that a vaccine becomes available only when the virus reaches a critical point. Under this more realistic scenario, we examine the impact of the time lag on the effectiveness of various different strategies. Practically, our findings give implications for the anti-virus software vendors and policy makers about the timeline of vaccination with different strategies. Finally, with the structural data of the World Wide Web network, we empirically investigate the effectiveness of various vaccination strategies. We capture reality by using empirical network structure data of the World Wide Web and considering the time lag between the identification of a virus and the availability of a vaccine. We examine the effectiveness of various ways of targeted vaccination on this asymmetric network where the flow of data is bi-directional (that is, incoming and outgoing).

In this paper, we examine the effectiveness of different vaccinations strategies from a physical perspective. Certain managerial or policy issues including security often require an interdisciplinary approach. We aim to shed light on security research by looking at the policy issues from a different angle. Security is not the only such area where the self-similar selection method can be used. Given the scale-free structure of online network, this method can be tested in the context of knowledge distribution and network seeding. This will posit an interesting question for future research.

Appendix

Matlab code used to generate Fig. 3 is presented here.

```
clear all
cd('Data') %% Directory containing a raw data
Rdata = load('www.dat'); %% Read a data

A = Rdata+1; %% To prevent zero node index
T0 = min(min(A));
T1 = max(max(A));

%% Turn-onInfection in time
clear Purity Infe Infect N_infect
if 0
   n1 = histc(A(:,1), [T0:T1]); %% Outgoing
   n2 = histc(A(:,2), [T0:T1]); %% Incoming

   [B1,I1] = sort(n1,'descend'); %% Outgoing
   [B2,I2] = sort(n2,'descend'); %% Incoming
   Ir = randperm(T1);      %%% Random drawing

   %% Read simulation results from box-counting methods
   cd ..
   cd('Results_MEMB');
   l = 8; %% Grouping length
     R  =  load(['boxes_stage0_rb_'  int2str(l)  '.dat']); %%  Read  data  from  box-couting
methods
   cd ..

   N_g = 10; %%% number of group to be selected
   ttI = R(:,2);
   nn = histc(ttI, [1:max(ttI)]);
    [Vt, It] = sort(nn,'descend'); %% Sort based on how many nodes in a group from high
to low
   for i = 1:N_g
      clear temp
      temp = find(ttI(I1) == It(i));
      eval(['In' int2str(i) '= I1(temp);'])
   end

   N_out = 1e3; tstep = 30;
   TT = randperm(T1);

   for Percent = [0 0.05 0.1 .2 .3 .5]; %%% Above than this percentage, vaccin is on
      for i = 1
         for k = 1:4
            clear IIInd
            if k == 1, II = I1; Ind = II(1:N_out);
            elseif k ==2, II = I2; Ind = II(1:N_out);
            elseif k == 3, II = Ir; Ind = II(1:N_out);

            elseif k == 4,
               NN_g = ceil(N_out/N_g);
               II = [];
               for tempi = 1:N_g %% How many groups considered
                  clear In
                  eval(['In = In' int2str(tempi) '(1:NN_g);'])
                  II = [II; In];
               end
               Ind = II(:);
            end
```

```
                Infect = TT(i); %% infection point
                Purity = zeros(T1,1); Purity(Infect) = 1;
                %% Spread virus in computational time.
                for t = [1:tstep]
                    clear temp
                    Old_Infect = unique(Infect');
                    temp = ismember(A(:,1)', Old_Infect);

                    Infect = unique(A(temp,2));
                    if length(find(Purity)) > Percent*T1
                        Infect = Infect(~ismember(Infect, Ind));
                    end

                    Purity(Infect) = ones(length(Infect),1);
                    N_inf(k,t,i) = length(find(Purity));
                end
            end
        end
         save(['Infection_log_time' num2str(Percent) 'percent_0410.mat'], 'N_inf', 'Percent',
'tstep', 'T1');
    end
    %%%%%%%% Make a plot %%%%%%%%
    for Percent =  [0 0.05 0.1 .2 .3 .5];

        load(['Infection_log_time' num2str(Percent) 'percent_0410.mat']);
        for k = [1:5]
            Mn(k,:) = N_inf(k,:,1); %52
        end

        figure(12); clf; sym = ['^vo*'];
        %%% 1-incoming 2-out-going 3-random_infection 4-self-similar
        hold on; plot([1 tstep], T1*[1 1], 'r--', 'linewidth', 2);
        for k = [1:4]
            plot([1:tstep], Mn(k,:), ['k' sym(k) '-'], 'linewidth', 1.3, 'markerfacecolor', 'b');
        end
        box on;
        set(gca,'yscale', 'linear', 'xscale' , 'linear', 'fontsize', 15, 'xlim',[1 tstep]);
        xlabel('Time (step)'); ylabel('# of infection');
                print(12,    '-depsc',    ['Plot_infection_time_turnon_'    num2str(Percent)
'percent_0410.eps'])
    end
end
```

References

1. Akoglu L, Faloutsos C (2009) RTG: a recursive realistic graph generator using random typing. Data Min Knowl Disc 19(2):194–209
2. Albert R, Jeong H, Barabasi AL (1999) Diameter of the world-wide web. Nature 401:130–131
3. Albert R, Jeong H, Barabasi AL (2000) Attack and error tolerance of complex networks. Nature 406:378–382
4. Bai X, Airoldi EM, Malin B (2011) An entropy approach to disclosure risk assessment: lessons from real applications and simulated domains. Decis Support Syst 51(1):10–20
5. Balthrop J, Forrest S, Newman ME, Willamson M (2004) Technological networks and the spread of computer viruses. Science 304:527–529
6. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
7. Callaway DS, Newman ME, Strogatz SH, Watts DJ (2000) Network robustness and fragility: percolation on random graphs. Physical Review Letter 85:5468–5471
8. Chakrabarti D, Wang Y, Wang C, Leskovec J, Faloutsos C (2008) Epidemic thresholds in real networks. ACM Transactions on Information and System Security 10(4):1–26
9. Cohen R, Erez K, ben-Avraham D D, Havlin S (2000) Resilience of the Internet to random breakdowns. Physical Review Letter 85:4629–4632
10. Cohen R, Havlin S, ben-Avraham D (2003) Efficient immunization strategies for computer networks and populations. Phys Rev Lett 91(24):247901
11. Erdos P, Renyi A (1960) On the evolution of random graphs. Publication of the Mathematical Institute of Hungarian Academy of Sciences 5:17–61
12. Ganesh A, Massoulié L, Towsle D (2005) The effect of network topology on the spread of epidemics. Procedings of IEEE INFOCOM 2005:1455–1466
13. Gleissner W (1989) A mathematical theory for the spread of computer viruses. Computer and Security 8:35–41
14. Jung S, Kim S, Kahng B (2002) Geometric fractal growth model for scale-free networks. Phys Rev E 65:056101
15. Kim BC, Chen P, Mukhopadhyay T (2011) The effect of liability and patch release on software security: the monopoly case. Production and Operations Management 20(4):603–617
16. Kim BC, Chen P, Mukhopadhyay T (2010) An economic analysis of the software market with a risk-sharing mechanism. International Journal of Electronic Commerce 14(2):7–39
17. Kim JS, Goh KI, Kahng B, Kim D (2007) Fractality and self-similarity in scale-free networks. New J Phys 9:177
18. Leskovec J, Faloutsos C (2007) Scalable modeling of real graphs using Kronecker multiplication. Proceedings of the 24th international conference on machine learning 497–504
19. Mandelbrot BB (1983) The fractal geometry of nature. Freeman Publisher, WH
20. Markoff J (2009) Worm infects millions of computers worldwide. New York Times Jan 22, 2009, available online at http://www.nytimes.com/2009/01/23/technology/internet/23worm.html
21. Murray W (1988) The application of epidemiology to computer viruses. Computers and Security 7:35–41
22. Newman ME, Forrest S, Baltrop J (2002) Email networks and the spread of computer viruses. Phys Rev 66(3):035101
23. Oh W, Jeon S (2007) Membership herding and network stability in the open source community. Manage Sci 53(7):1086–1101
24. Park I, Sharman R, Rao HR, Upadhyaya S (2007) Short term and total life impact analysis of email worms in computer systems. Decis Support Syst 43(3):827–841
25. Song C, Havlin S, Makse HA (2005) Self-similarity of complex networks. Nature 433:392–395
26. Rozenfeld HD, Gallos LK, Song C, Makse HA (2008) Fractal and transfractal scale-tree networks. arXiv:0808.2206v1
27. Rozenfeld HD, Song C, Makse HA (2010) Small-world to fractal transition in complex networks: a renormalization group approach. Phys Rev Lett 104:025701

28. Tong H, Prakash BA, Tsourakakis C, Eliassi-Rad R, Faloutsos C, Chau DH (2010) On the vulnerability of large graphs. Proceedings of IEEE international conference on data mining 1091–1096
29. Wang C, Knight J, Elder M (2000) On computer viral infection and the effect of immunization, In proceedings of the 16th annual computer security applications conference 246–256
30. Yuan H, Chen G, Wu J, Xiong H (2009) Towards controlling virus propagation in information systems with point-to-group information sharing. Decis Support Syst 48:57–68