Knowl Inf Syst (2011) 27:253–279
DOI 10.1007/s10115-010-0321-0

REGULAR PAPER

Network immunization and virus propagation in email networks: experimental evaluation and analysis

Chao Gao · Jiming Liu · Ning Zhong

Received: 10 November 2009 / Revised: 4 June 2010 / Accepted: 25 June 2010 /
Published online: 14 July 2010

Abstract Network immunization strategies have emerged as possible solutions to the challenges of virus propagation. In this paper, an existing interactive model is introduced and then improved in order to better characterize the way a virus spreads in email networks with different topologies. The model is used to demonstrate the effects of a number of key factors, notably nodes' degree and betweenness. Experiments are then performed to examine how the structure of a network and human dynamics affects virus propagation. The experimental results have revealed that a virus spreads in two distinct phases and shown that the most efficient immunization strategy is the node-betweenness strategy. Moreover, those results have also explained why old virus can survive in networks nowadays from the aspects of human dynamics.

Keywords Immunization strategies · Virus propagation · Human dynamics · Email networks · Enron

1 Introduction

Recent research has shown that many real-world systems can be modeled as complex networks, such as the food chain network, the neural network, the World Wide Web (WWW),

A preliminary version of this paper was presented at the International Conference on Advanced Data Mining and Applications (ADMA'09), August 17–19, Beijing, China.

C. Gao · J. Liu · N. Zhong
International WIC Institute, Beijing University of Technology, 100124 Beijing, China

C. Gao
Beijing Key Laboratory of Multimedia and Intelligent Software, 100124 Beijing, China

J. Liu (✉)
Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong
e-mail: jiming@comp.hkbu.edu.hk

N. Zhong
Department of Life Science and Informatics, Maebashi Institute of Technology, Maebashi, Japan

the Internet, the scientific collaboration network and the social network [15,32]. In these networks, nodes denote individuals (e.g. computers, Web pages, email-boxes, people, or species) and edges represent the connections between individuals (e.g. network links, hyperlinks, relationships between two people or species) [26]. There are many research topics related to network-like environments [23,34,46]. One interesting and challenging subject is how to control virus propagation in physical networks (e.g. trojan viruses) and virtual networks (e.g. email worms) [26,30,37]. Currently, one of the most popular methods is network immunization where some nodes in a network are immunized (protected) so that they can not be infected by a virus or a worm. After immunizing the same percentages of nodes in a network, the best strategy can minimize the final number of infected nodes.

Valid propagation models can be used in complex networks to predict potential weaknesses of a global network infrastructure against worm attacks [40] and help researchers understand the mechanisms of new virus attacks and/or new spreading. At the same time, reliable models provide test-beds for developing or evaluating new and/or improved security strategies for restraining virus propagation [48]. Researchers can use reliable models to design effective immunization strategies which can prevent and control virus propagation not only in computer networks (e.g. worms) but also in social networks (e.g. SARS, H1N1, and rumors).

Today, more and more researchers from statistical physics, mathematics, computer science, and epidemiology are studying virus propagation and immunization strategies. For example, computer scientists focus on algorithms and the computational complexities of strategies, i.e. how to quickly search a short path from one “seed” node to a targeted node just based on local information, and then effectively and efficiently restrain virus propagation [42]. Epidemiologists focus on the combined effects of local clustering and global contacts on virus propagation [5]. Generally speaking, there are two major issues concerning virus propagation:

1. How to efficiently restrain virus propagation?
2. How to accurately model the process of virus propagation in complex networks?

In order to solve these problems, the main work in this paper is to (1) systematically compare and analyze representative network immunization strategies in an interactive email propagation model, (2) uncover what the dominant factors are in virus propagation and immunization strategies, and (3) improve the predictive accuracy of propagation models through using research from human dynamics.

The remainder of this paper is organized as follows: Sect. 2 surveys some well-known network immunization strategies and existing propagation models. Section 3 presents the key research problems in this paper. Section 4 describes the experiments which are performed to compare different immunization strategies with the measurements of the immunization efficiency, the cost and the robustness in both synthetic networks (including a synthetic community-based network) and two real email networks (the Enron and a university email network), and analyze the effects of network structures and human dynamics on virus propagation. Section 5 concludes the paper.

2 Related work

In this section, several popular immunization strategies and typical propagation models are reviewed. An interactive email propagation model is then formulated in order to evaluate different immunization strategies and analyze the factors that influence virus propagation.

2.1 Immunization strategies

Network immunization is one of the well-known methods to effectively and efficiently restrain virus propagation. It cuts epidemic paths through immunizing (injecting vaccines or patching programs) a set of nodes from a network following some well-defined rules. The immunized nodes, in most published research, are all based on node degrees that reflect the importance of a node in a network, to a certain extent. In this paper, the influence of other properties of a node (i.e. betweenness) on immunization strategies will be observed.

2.1.1 Degree-based strategies

Pastor-Satorras and Vespignani have studied the critical values in both random and targeted immunization [39]. The *random immunization* strategy treats all nodes equally. In a large-scale-free network, the immunization critical value is $g_c \rightarrow 1$. Simulation results show that 80% of nodes need to be immunized in order to recover the epidemic threshold. Dezso and Barabasi have proposed a new immunization strategy, named as the *targeted immunization* [9], which takes the actual topology of a real-world network into consideration.

The distributions of node degrees in scale-free networks are extremely heterogeneous. A few nodes have high degrees, while lots of nodes have low degrees. The targeted immunization strategy aims to immunize the most connected nodes in order to cut epidemic paths through which most susceptible nodes may be infected. For a BA network [2], the critical value of the targeted immunization strategy is $g_c \sim e^{\frac{-2}{m\lambda}}$. This formula shows that it is always possible to obtain a small critical value g_c even if the spreading rate λ changes drastically. However, one of the limitations of the targeted immunization strategy is that it needs to know the information of global topology, in particular the ranking of the nodes must be clearly defined. This is impractical and uneconomical for handling large-scale and dynamic-evolving networks, such as P2P networks or email networks. In order to overcome this shortcoming, a local strategy, namely the *acquaintance immunization* [8,16], has been developed.

The motivation for the acquaintance immunization is to work without any global information. In this strategy, p % of nodes are first selected as “seeds” from a network, and then one or more of their direct acquaintances are immunized. Because a node with higher degree has more links in a scale-free network, it will be selected as a “seed” with a higher probability. Thus, the acquaintance immunization strategy is more efficient than the random immunization strategy, but less than the targeted immunization strategy. Moreover, there is another issue which limits the effectiveness of the acquaintance immunization: it does not differentiate nodes, i.e. randomly selects “seed” nodes and their direct neighbors [17].

Another effective distributed strategy is the *D-steps immunization* [12,17]. This strategy views the decentralized immunization as a graph covering problem. That is, for a node v_i, it looks for a node to be immunized that has the maximal degree within d steps of v_i. This method only uses the local topological information within a certain range (e.g. the degree information of nodes within d steps). Thus, the maximal acquaintance strategy can be seen as a 1-step immunization. However, it does not take into account domain-specific heuristic information, nor is it able to decide what the value of d should be in different networks.

2.1.2 Betweenness-based strategies

The immunization strategies described in the previous section are all based on node degrees. The way different immunized nodes are selected is illustrated in Fig. 1. Besides removing

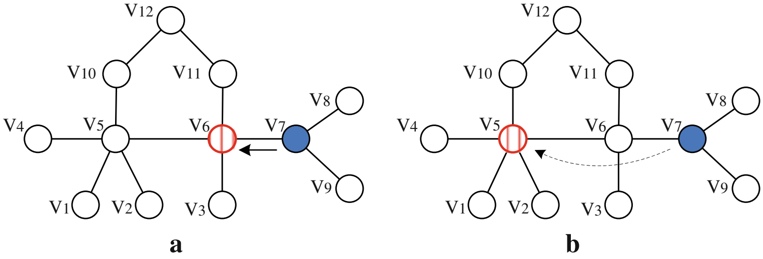


Fig. 1 An illustration of different strategies. The *targeted immunization* will directly select v_5 as an immunized node based on the degrees of nodes. Suppose that v_7 is a "seed" node. v_6 will be immunized based on the maximal *acquaintance immunization* strategy, and v_5 will be indirectly selected as an immunized node based on the *D-steps immunization* strategy, where $d = 2$

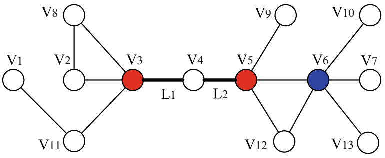


Fig. 2 An illustration of betweenness-based strategies. If we select one immunized node, the targeted immunization strategy will directly select the highest-degree node, v_6. The node-betweenness strategy will select v_5 as it has the highest node betweenness. The edge-betweenness strategy will select one of v_3, v_4 and v_5 because the edges, L_1 and L_2, have the highest edge betweenness

the highest-degree nodes from a network, many approaches cut epidemic paths by means of increasing the average path length of a network, for example by partitioning large-scale networks based on betweenness [4,36]. For a network, node (edge) betweenness refers to the number of the shortest paths that pass through a node (edge). A higher value of betweenness means that the node (edge) links more adjacent communities and will be frequently used in network communications. Although [19] have analyzed the robustness of a network against degree-based and betweenness-based attacks, the spread of a virus in a propagation model is not considered, so the effects of different measurements on virus propagation is not clear.

Is it possible to restrain virus propagation, especially from one community to another, by immunizing nodes or edges which have higher betweenness. In this paper, two types of betweenness-based immunization strategies will be presented, i.e. the node-betweenness strategy and the edge-betweenness strategy. That is, the immunized nodes are selected in the descending order of node- and edge-betweenness, in an attempt to better understand the effects of the degree and betweenness centralities on virus propagation. Figure 2 shows that if v_4 is immunized, the virus will not propagate from one part of the network to another. The node-betweenness strategy will select v_5 as an immunized node, which has the highest node betweenness, i.e. 41. The edge-betweenness strategy will select the terminal nodes of L_1 or L_2 (i.e. v_3, v_4 or v_4, v_5) as they have the highest edge betweenness.

As in the targeted immunization, the betweenness-based strategies also require information about the global betweenness of a network. The experiments presented in this paper is to find a new measurement that can be used to design a highly efficient immunization strategy. The efficiency of these strategies is compared both in synthetic networks and in real-world networks, such as the Enron email network described by [4].

2.2 Propagation models

In order to compare different immunization strategies, a propagation model is required to act as a test-bed in order to simulate virus propagation. Currently, there are two typical models: (1) the epidemic model based on population simulation and (2) an interactive email model which utilizes individual-based simulation.

Lloyd and May have proposed an epidemic propagation model to characterize virus propagation, a typical mathematical model based on differential equations [26]. Some specific epidemic models, such as SI [37,38], SIR [1,30], SIS [14], and SEIR [11,28], have been developed and applied in order to simulate virus propagation and study the dynamic characteristics of whole systems. However, these models are all based on the mean-filed theory, i.e. differential equations. This type of black-box modeling approach only provides a macroscopic understanding of virus propagation—they do not give much insight into microscopic interactive behavior. More importantly, some assumptions, such as a fully mixed (i.e. individuals that are connected with a susceptible individual will be randomly chosen from the whole population) [33] and equiprobable contacts (i.e. all nodes transmit the disease with the same probability and no account is taken of the different connections between individuals) may not be valid in the real world. For example, in email networks and Instant Message (IM) networks, communication and/or the spread of information tend to be strongly clustered in groups or communities that have more closer relationships rather than being equiprobable across the whole network. These models may also overestimate the speed of propagation [49].

In order to overcome the above-mentioned shortcomings, [49] have built an interactive email model to study worm propagation, in which viruses are triggered by human behavior, not by contact probabilities. That is to say, the node will be infected only if a user has checked his/her email-box and clicked an email with a virus attachment. Thus, virus propagation in the email network is mainly determined by two behavioral factors: email-checking time intervals (T_i) and email-clicking probabilities (P_i), where $i \in [1, N]$, N is the total number of users in a network. T_i is determined by a user's own habits; P_i is determined both by user security awareness and the efficiency of the firewall. However, the authors do not provide much information about how to restrain worm propagation.

In this paper, an interactive email model is used as a test-bed to study the characteristics of virus propagation and the efficiency of different immunization strategies. It is readily to observe the microscopic process of worm propagating through this model, and uncover the effects of different factors (e.g. the power-law exponent, human dynamics and the average path length of the network) on virus propagation and immunization strategies. Unlike other models, this paper mainly focuses on comparing the performance of degree-based strategies and betweenness-based strategies, replacing the critical value of epidemics in a network. A detailed analysis of the propagation model is given in the following section.

2.3 A general model for interactive email networks

An email network can be viewed as a typical social network in which a connection between two nodes (individuals) indicates that they have communicated with each other before [35, 49]. Generally speaking, a network can be denoted as $\mathrm{E} = \langle V, L \rangle$, where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of nodes and $L = \{\langle v_i, v_j \rangle \mid 1 \leq i, j \leq n\}$ is a set of undirected links (if v_i in the hit-list of v_j, there is a link between v_i and v_j). A virus can propagate along links and infect more nodes in a network.

In order to give a general definition, each node is represented as a tuple *<ID, State, NodeLink, P_{behavior}, B_{action}, VirusNum, NewVirus>*.

- ***ID***: the node identifier, $v_i.ID = i$.
- ***State***: the node state:

$$v_i.State = \begin{cases} healthy = 0, & if\ the\ node\ has\ no\ virus, \\ danger = 1, & if\ the\ node\ has\ virus\ but\ not\ infected, \\ infected = 2, & if\ the\ node\ has\ been\ infected, \\ immunized = 3, & if\ the\ node\ has\ been\ immunized. \end{cases}$$

- ***NodeLink***: the information about its hit-list or adjacent neighbors, i.e. $v_i.NodeLink = \{\langle i, j\rangle \mid \langle i, j\rangle \in L\}$.
- $P_{behavior}$: the probability that a node will to perform a particular behavior.
- B_{action}: different behaviors.
- ***VirusNum***: the total number of new unchecked viruses before the next operation.
- ***NewVirus***: the number of new viruses a node receives from its neighbors at each step.

In addition, two interactive behaviors are simulated according to [49], i.e. the email-checking time intervals and the email-clicking probabilities both follow Gaussian distributions, if the sample size goes to infinity. For the same user *i*, the email-checking interval $T_i(t)$ in [49] has been modeled by a Poisson distribution, i.e. $T_i(t) \sim \lambda e^{-\lambda t}$. Thus, the formula for P_{behavior} in the tuple can be written as $P^1_{\text{behavior}} = ClickProb$ and $P^2_{\text{behavior}} = CheckTime$.

- ***ClickProb*** is the probability of an user clicking a suspected email,

$$v_i.P^1_{behavior} = v_i.ClickProb = normalGenerator(\mu_p, \sigma_p) \sim N(\mu_p, \sigma_p^2).$$

- ***CheckRate*** is the probability of an user checking an email,

$$v_i.CheckRate = normalGenerator(\mu_t, \sigma_t) \sim N(\mu_t, \sigma_t^2).$$

- ***CheckTime*** is the next time the email-box will be checked,

$$v_i.P^2_{behavior} = v_i.CheckTime = expGenerator(v_i.CheckRate).$$

B_{action} can be specified as $B^1_{action} = receive\_email$, $B^2_{action} = send\_email$, and $B^3_{action} = update\_email$. If a user receives a virus-infected email, the corresponding node will update its state, i.e. $v_i.State \leftarrow danger$. If a user opens an email that has a virus-infected attachment, the node will adjust its state, i.e. $v_i.State \leftarrow infected$, and send this virus email to all its friends, according to its hit-list. If a user is immunized, the node will update its state to $v_i.State \leftarrow immunized$.

In order to better characterize virus propagation, some assumptions are made in the interactive email model:

- If a user opens an infected email, the node is infected and will send viruses to all the friends on its hit-list;
- When checking his/her mailbox, if a user does not click virus emails, it is assumed that the user deletes the suspected emails;
- If nodes are immunized, they will never send virus emails even if a user clicks an attachment.

3 Problem statement

The most important measurement of the effectiveness of an immunization strategy is the total number of infected nodes after virus propagation. The best strategy can effectively restrain virus propagation, i.e. the total number of infected nodes is kept to a minimum. In order to evaluate the efficiency of different immunization strategies and find the relationship between local behaviors and global dynamics, two statistics are of particular interest:

1. *SID*: the sum of the degrees of immunized nodes that reflects the importance of nodes in a network

$$SID = \sum v_i.degree, \quad where \;\; v_i.state = immunized.$$

2. *APL*: the average path length of a network. This is a measurement of the connectivity and transmission capacity of a network

$$APL = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i>j} d_{ij}$$

   where d_{ij} is the shortest path between *i* and *j*. If there is no path between *i* and *j*, $d_{ij} \rightarrow \infty$. In order to facilitate the computation, the reciprocal of d_{ij} is used to reflect the connectivity of a network:

$$APL' = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i>j} d_{ij}^{-1}$$

   if there is no path between *i* and *j*, $d_{ij}^{-1} = 0$.

Based on these definitions, the interactive email model given in Sect. 2.3 can be used as a test-bed to compare different immunization strategies and uncover the effects of different factors on virus propagation. The specific research questions addressed in this paper can be summarized as follows:

1. How to evaluate network immunization strategies? How to determine the performance of a particular strategy, i.e. in terms of its efficiency, cost and robustness? What is the best immunization strategy? What are the key factors that affect the efficiency of a strategy?
2. What is the process of virus propagation? What effect does the network structure have on virus propagation?
3. What effect do human dynamics have on virus propagation?

The simulations in this paper have two phases. First, a existing email network is established in which each node has some of the interactive behaviors described in Sect. 2.3. Next, the virus propagation in the network is observed and the epidemic dynamics are studied when applying different immunization strategies. More details can be found in Sect. 4.

4 Experimental evaluations

In this section, the simulation process and the structures of experimental networks are presented in Sects. 4.1 and 4.2. Section 4.3 uses a number of experiments to evaluate the performance (e.g. efficiency, cost and robustness) of different immunization strategies. Specifically, the experiments seek to address whether or not betweenness-based immunization strategies can restrain worm propagation in email networks, and which measurements can reflect and/or characterize the efficiency of immunization strategies. Finally, Sects. 4.4 and 4.5 presents an in-depth analysis in order to determine the effect of network structures and human dynamics on virus propagation.

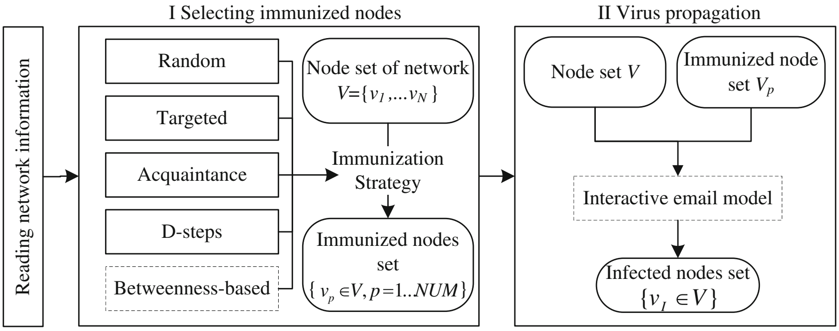

Reading network information
I Selecting immunized nodes
Random
Targeted
Acquaintance
D-steps
Betweenness-based
Node set of network $V=\{v_1,...v_N\}$
Immunization Strategy
Immunized nodes set $\{v_p \in V, p=1..NUM\}$
II Virus propagation
Node set V
Immunized node set V_p
Interactive email model
Infected nodes set $\{v_I \in V\}$

Fig. 3 The process of our experiments

4.1 Experimental design

The experimental process is illustrated in Fig. 3. Some nodes are first immunized (protected) from the network using different strategies. The viruses are then injected into the network in order to evaluate the efficiency of those strategies by comparing the total number of infected nodes. Two methods are used to select the initial infected nodes: random infection and malicious infection, i.e. infecting the nodes with maximal degrees. The user behavior parameters are based on the definitions in Sect. 2.3, where $\mu_p = 0.5$, $\sigma_p = 0.3$, $\mu_t = 40$, and $\sigma_t = 20$. Since the process of email worm propagation is stochastic, all results are averaged over 100 runs. The virus propagation algorithm is specified in Alg. 1.

4.2 The structures of experimental email networks

Many common networks have presented the phenomenon of scale-free [2,21], where nodes' degrees follow a power-law distribution [42], i.e. the fraction of nodes having k edges, $p(k)$, decays according to a power law $p(k) \sim k^{-\alpha}$ (where α is usually between 2 and 3) [29]. Recent research has shown that email networks also follow power-law distributions with a long tail [35,49]. Therefore, in this paper, three synthetic power-law networks and a synthetic community-based network, generated using the GLP algorithm [6] where the power can be tuned. The three synthetic networks all have 1000 nodes with α =1.7, 2.7, and 3.7, respectively. The statistical characteristics and visualization of the synthetic community-based network are shown in Table 1 and Fig. 4c, f, respectively. In order to reflect the characteristics of a real-world network, the Enron email network[1] which is built by Andrew Fiore and Jeff Heer, and the university email network[2] which is complied by the members of the University Rovira i Virgili (Tarragona) will also be studied. The structure and degree distributions of these networks are shown in Table 2 and Fig. 4. In particular, the cumulative distributions are estimated with maximum likelihood using the method provided by [7]. The degree statistics are shown in Table 9.

4.3 The effect of network immunization on virus propagation

In this section, a comparison is made of the effectiveness of different strategies in an interactive email model. Experiments are then used to evaluate the cost and robustness of each strategy.

[1] Http://bailando.sims.berkeley.edu/enron/enron.sql.gz.

[2] Http://deim.urv.cat/~aarenas/data/welcome.htm.

Algorithm 1 The algorithm of virus propagation

Input: NodeData[NodeNum] stores the topology of an email network. Timestep is the system clock. v_0 is the set of initially infected nodes.
Output: SimNum[timeStep][k] stores the number of infected nodes in the network in the k^{th} simulation.

```
For k=1 to Runtime //We run 100 times to obtain an average value
  NodeData[NodeNum] ← Initializing an email network as well as users'
                      checking time and clicking probability;
  NodeData[NodeNum] ← Choosing immunized nodes based on different
                      immunization strategies and adjusting their states;
  While Timestep < EndSimul //There are 600 steps at each time
    For i=1 to NodeNum
      If NodeData[i].Checktime==0
        prob← computing the probability of opening a virus-infected
              email based on user's ClickProb and VirusNum
        If NodeData[i].State==danger ||$rand() < prob$
          NodeData[i].State=infected
          $v_t \leftarrow v_t + 1$
          Send a virus to all friends according to its hit-list
        EndIf
      EndIf
    EndFor
    For i=1 to NodeNum
      Update the next CheckTime based on user's CheckRate
      NodeData[i].VirusNum+=NodeData[i].NewVirus
    EndFor
    SimNum[TimeStep][k] ← $v_t$
    TimeStep++
  EndWhile
EndFor
```

4.3.1 Immunization efficiency

The immunization efficiency of the following immunization strategies are compared: the targeted and random strategies [39], the acquaintance strategy (random and maximal neighbor) [8,16], the *D*-steps strategy ($d = 2$ and $d = 3$) [12,17] (which is introduced in Sect. 2.1),

Table 1 The structural characteristics of different communities in the synthetic community-based network

	Nodes	Edges	α	$\langle k \rangle$
Community 1	1,000	4,139	1.7	8.278
Community 2	1,000	4,107	2.7	8.214
Community 3	1,000	4,289	2.2	8.578
Community 4	1,000	4,099	1.5	8.198
The bridges between different communities: 100				
The whole network: α=1.77, $\langle k \rangle$=8.34				

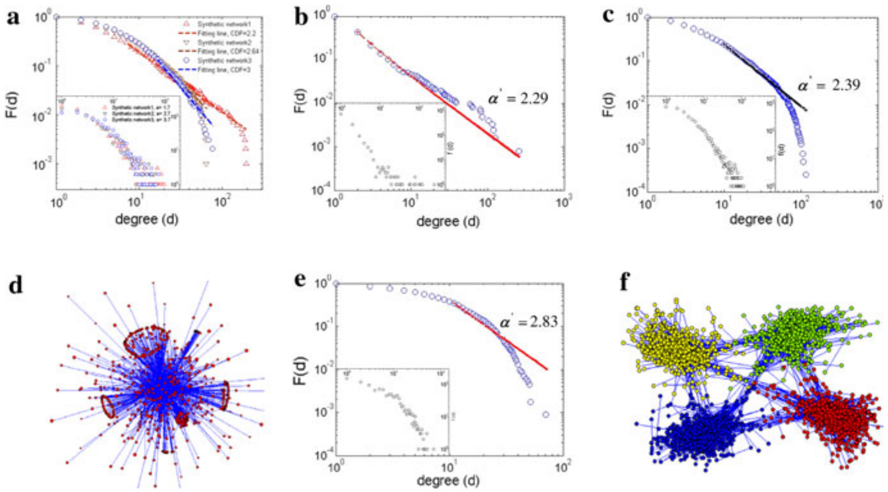


Fig. 4 The degree distributions and structures of networks. The *small plots* are the probability distributions of node degrees where α is their fitted power. The *big plots* are cumulative distributions (F(d)) and their maximum likelihood power-law fits for empirical data sets, where $\alpha^{'}$ is the fitted power. **a** Three synthetic networks. **b** Enron email network. **c** Synthetic community network. **d** Structure of synthetic network1. **e** University email network. **f** Structure of community network

Table 2 The structures of experimental networks

	α	Nodes	Edges	$\langle k\rangle$
Synthetic networks	1.7	1,000	4,224	8.456
	2.7	1,000	4,226	8.452
	3.7	1,000	4,232	8.464
Synthetic community network	1.77	4,000	16,734	8.340
Enron email network	1.3	1,238	2,106	3.400
University email network	1.46	1,133	5,451	9.620

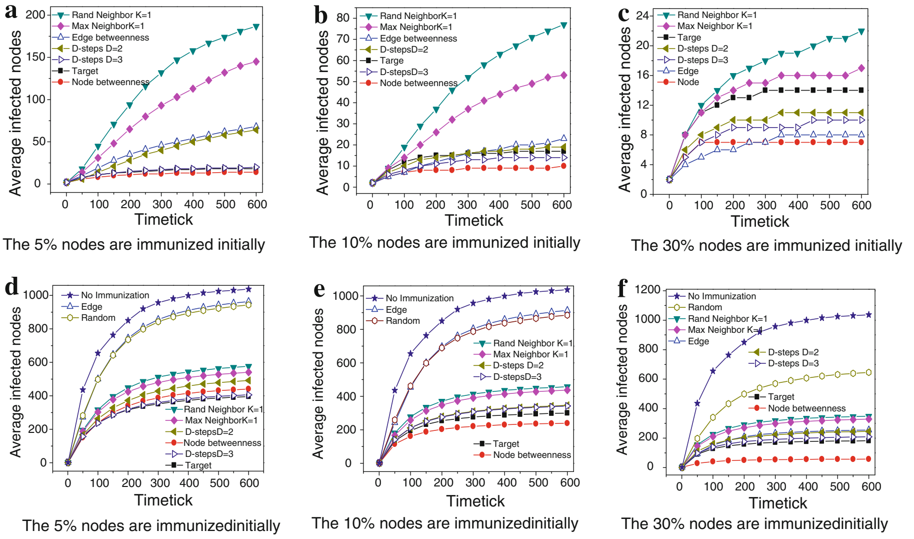
and the proposed betweenness-based strategy (node- and edge-betweenness). In the initial set of experiments, the proportion of immunized nodes (5, 10, and 30%) are varied in the synthetic networks and the Enron email network. Table 3 shows the simulation results in the Enron email network which is initialized with two infected nodes. Figure 5 shows the average numbers of infected nodes over time. Tables 4, 5, and 6 show the numerical results in three synthetic networks, respectively.

The simulation results show that the node-betweenness immunization strategy yields the best results (i.e. the minimum number of infected nodes, F) except for the case where 5% of the nodes in the Enron network are immunized under a malicious attack. The average degree of the Enron network is $\langle k\rangle = 3.4$. This means that only a few nodes have high degrees, others have low degrees (see Table 9). In such a network, if nodes with maximal degrees are infected, viruses will rapidly spread in the network and the final number of infected nodes will be larger than in other cases. The targeted strategy therefore does not perform any better than the node-betweenness strategy. In fact, as the number of immunized nodes increases, the efficiency of the node-betweenness immunization increases proportionally

Table 3 The Enron email network with different attack modes

	5%			10%			30%		
	F	SID	APL'	F	SID	APL'	F	SID	APL'
Random attack									
Targeted	19	1,995	30.58	17	2,351	23.12	14	3,017	20.98
Node-betweenness	14	1,929	23.76	10	2,154	20.99	7	2,652	19.51
Edge-betweenness	68	1,748	49.51	23	2,104	612.04	8	2,647	19.49
Random	822	169	681.35	784	328	606.13	541	985	375.03
Random neighbor	187	1,441	111.99	77	1,752	46.47	22	2,076	24.68
Max neighbor	145	1,512	84.63	53	1,803	41.23	17	2,159	22.37
D-steps $D = 2$	64	1,761	60.03	19	2,204	30.88	11	2,952	18.96
D-steps $D = 3$	20	1,931	32.44	14	2,208	23.50	10	2,751	22.80
Malicious attack									
Targeted	396	1,641	401.28	300	1,992	326.68	182	2,658	176.20
Nodes-betweenness	441	1,483	358.83	239	1,780	243.89	58	2,259	39.42
Edge-betweenness	964	80	635.73	914	147	540.10	255	1,916	138.29
Random	944	151	689.67	885	316	624.27	646	924	406.91
Random neighbor	577	1,092	491.86	458	1,376	444.62	349	1,697	403.65
Max neighbor	541	1,154	474.25	437	1,427	432.08	329	1,797	383.85
D-steps $D = 2$	492	1,417	412.29	346	1,843	333.41	209	2,591	190.95
D-steps $D = 3$	406	1,580	412.53	342	1,839	364.81	246	2,375	227.19

If there is no immunization, the final number of infected nodes is 1,037 with a random attack and 1,052 with a malicious attack, and $APL' = 756.79(10^{-4})$. The total simulation time $T = 600$



The 5% nodes are immunized initially

The 10% nodes are immunized initially

The 30% nodes are immunized initially

The 5% nodes are immunizedinitially

The 10% nodes are immunizedinitially

The 30% nodes are immunizedinitially

Fig. 5 Comparing different immunization strategies in the Enron email network. **a**–**c** Correspond to random attacks, whereas **d**–**f** are malicious attacks. In each plot, the curve labels are ranked by the total numbers of infected nodes from *top* to *bottom*

Table 4 NET1 is a synthetic network with $\alpha = 1.7$

	5%			10%			30%		
	F	SID	APL'	F	SID	APL'	F	SID	APL'
Random attack									
Targeted	679	3,697	111.56	413	4,694	57.88	7	6,666	16.20
Node-betweenness	653	3,680	109.80	368	4,588	52.31	3	6,418	12.89
Edge-betweenness	682	3,619	116.73	398	4,518	58.58	39	5,829	21.47
Random	879	414	658.05	803	845	570.01	563	2,538	295.57
Random neighbor	805	1,828	399.11	673	2,981	224.57	232	5,339	39.99
Max neighbor	721	2,897	203.75	556	4,025	95.60	34	5,946	23.28
D-steps $D = 2$	663	3,625	118.19	456	4,635	60.51	9	6,617	16.60
D-steps $D = 3$	658	3,690	112.82	431	4,684	58.93	7	6,664	16.18
Malicious attack									
Targeted	739	3,370	179.88	588	4,344	103.46	110	6,307	27.48
Nodes-betweenness	735	3,343	176.50	581	4,253	91.89	10	5,956	15.37
Edge-betweenness	829	1,586	410.34	673	3,241	185.44	236	5,301	50.72
Random	882	407	658.72	822	819	577.29	587	2,443	307.10
Random neighbor	825	1,640	438.51	716	2,758	268.94	367	4,974	72.61
Max neighbor	773	2,619	268.96	644	3,673	154.08	242	5,570	44.79
D-steps $D = 2$	741	3,305	185.04	590	4,295	105.22	109	6,258	27.14
D-steps $D = 3$	738	3,367	179.15	586	4,342	102.09	102	6,306	25.92

There are two infected nodes with different attack modes. If there is no immunization, the final number of infected nodes is 937 with a random attack and 942 with a malicious attack, and $APL' = 751.36(10^{-4})$. The total simulation time $T = 600$

more than the targeted strategy. Therefore, if global topological information is available, the node-betweenness immunization is the best strategy.

The maximal SID is obtained using the targeted immunization. However, the final number of infected nodes (F) is consistent with the average path length (APL) but not with the SID. That is to say, controlling a virus epidemic does not depend on the degrees of immunized nodes but on the path length of a whole network. This also explains why the efficiency of the node-betweenness immunization strategy is better than that of the targeted immunization strategy. The node-betweenness immunization selects nodes based on the average path length, while the targeted immunization strategy selects based on the size of degrees.

A more in-depth analysis is undertaken by comparing the change of the APL' with respect to the different strategies used in the synthetic networks. The results are shown in Fig. 6. Figure 7a, b compare the change of the final number of infected nodes over time, which correspond to Fig. 6c, d, respectively. These numerical results validate the previous assertion that the average path length can be used as a measurement to design an effective immunization strategy. The best strategy is to divide the whole network into different sub-networks and increase the average path length of a network, hence cut the epidemic paths.

In this paper, all comparative results are the average over 100 runs using the same infection model (i.e. the virus propagation is compared for both random and malicious attacks) and user behavior model (i.e. all simulations use the same behavior parameters, as shown in Sect. 4.1). Thus, it is more reasonable and feasible to just evaluate how the propagation of a

Table 5 NET2 is a synthetic network with $\alpha = 2.7$.

	5%			10%			30%		
	F	*SID*	APL'	F	*SID*	APL'	F	*SID*	APL'
Random attack									
Targeted	812	2,270	240.03	661	3,528	120.60	24	6,120	24.11
Node-betweenness	795	2,266	235.19	677	3,439	113.78	5	5,826	18.64
Edge-betweenness	803	2,189	246.67	674	3,299	129.79	177	5,210	36.54
Random	888	414	556.54	812	855	480.23	584	2,574	248.00
Random neighbor	840	924	463.68	765	1,752	335.80	393	4,348	83.63
Max neighbor	826	1,666	330.30	725	2,847	182.58	194	5,453	34.95
D-steps $D = 2$	813	2,211	247.29	683	3,469	126.62	30	6,063	24.75
D-steps $D = 3$	779	2,268	239.19	667	3,524	121.66	21	6,121	24.08
Malicious attack									
Targeted	835	2,207	253.36	721	3,443	131.30	103	6,019	26.69
Nodes-betweenness	831	2,194	248.82	710	3,368	124.83	15	5,719	20.19
Edge-betweenness	831	2,068	271.06	720	3,248	140.63	258	5,191	40.13
Random	888	409	557.33	832	835	484.45	601	2,496	257.58
Random neighbor	869	896	469.60	791	1,722	341.51	473	4,274	91.06
Max neighbor	850	1,608	342.62	747	2,788	192.53	323	343	39.54
D-steps $D = 2$	840	2,151	259.53	724	3,385	136.42	122	5,964	27.33
D-steps $D = 3$	835	2,204	252.08	722	3,441	130.88	94	6,018	26.23

There are two infected nodes with different attack modes. If there is no immunization, the final number of infected nodes is 936 with a random attack and 949 with a malicious attack, and $APL' = 634.01(10^{-4})$. The total simulation time $T = 600$

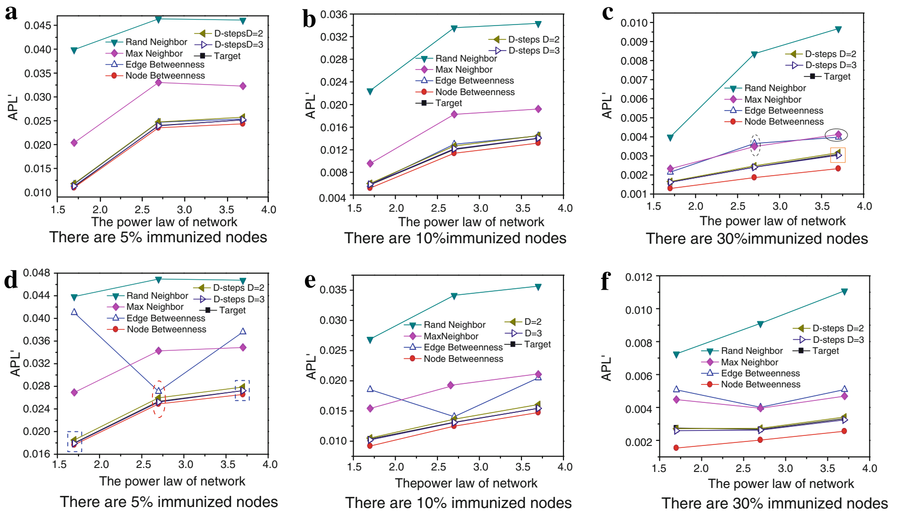
virus is affected by immunization strategies, i.e. avoiding the effects caused by the stochastic process, the infection model and the user behavior.

It can be seen that the edge-betweenness strategy is able to find some nodes with high degrees of centrality and then integrally divide a network into a number of sub-networks (e.g. v_4 in Fig. 2). However, compared with the nodes (e.g. v_5 in Fig. 2) selected by the node-betweenness strategy, the nodes with higher edge betweenness can not cut the epidemic paths as they can not effectively break the whole structure of a network. In Fig. 2, the synthetic community-based network and the university email network are used as examples to illustrate why the edge-betweenness strategy can not obtain the same immunization efficiency as the node-betweenness strategy. To select two nodes as immunized nodes from Fig. 2, the node-betweenness immunization will select $\{v_5, v_3\}$ by using the descending order of node betweenness. However, the edge-betweenness strategy can select $\{v_3, v_4\}$ or $\{v_4, v_5\}$ because the edges, L_1 and L_2, have the highest edge betweenness. This result shows that the node-betweenness strategy can not only effectively divide the whole network into two communities, but also break the interior structure of communities. Although the edge-betweenness strategy can integrally divided the whole network into two parts, viruses can also propagate in each community. Many networks commonly contain the structure shown in Fig. 2, for example, the Enron email network and university email networks. Table 7 and Fig. 8 present the results of the synthetic community-based network. Table 8 compares different strategies in the university email network, which also has some self-similar community structures [18]. These results further validate the analysis stated above.

Table 6 NET3 is a synthetic network with $\alpha = 3.7$

	5%			10%			30%		
	F	SID	APL'	F	SID	APL'	F	SID	APL'
Random attack									
Targeted	855	2,115	251.78	747	3,252	140.15	58	5,827	30.86
Node-betweenness	833	2,068	243.59	729	3,177	131.90	10	5,526	23.30
Edge-betweenness	845	2,014	253.61	741	3,048	144.59	198	5,001	39.85
Random	892	416	532.26	827	855	459.98	604	2,530	242.39
Random neighbor	875	803	461.16	781	1,562	343.45	433	4,059	96.80
Max neighbor	843	1,562	322.68	747	2,653	191.94	242	5,241	41.25
D-steps $D = 2$	848	2,042	257.46	752	3,196	144.94	66	5,776	31.66
D-steps $D = 3$	827	2,111	251.91	728	3,249	140.48	56	5,823	30.35
Malicious attack									
Targeted	861	2,021	272.38	766	3,141	154.38	203	5,701	33.27
Nodes-betweenness	856	1,970	265.49	766	3,089	147.47	33	5,479	25.59
Edge-betweenness	871	1,237	375.89	786	2,527	204.79	344	4,778	50.89
Random	894	414	533.00	837	817	466.80	611	2,483	249.01
Random neighbor	880	778	467.50	812	1,502	356.75	524	3,924	110.87
Max neighbor	864	1,454	348.61	778	2,552	210.56	390	5,100	46.92
D-steps $D = 2$	859	1,954	278.58	769	3,084	160.61	224	5,644	34.14
D-steps $D = 3$	859	2,018	272.39	768	3,139	154.50	175	5,697	32.45

There are two infected nodes with different attack modes. If there is no immunization, the final number of infected nodes is 936 with a random attack and 949 with a malicious attack, and $APL' = 607.27(10^{-4})$. The total simulation time $T = 600$



There are 5% immunized nodes

There are 10%immunized nodes

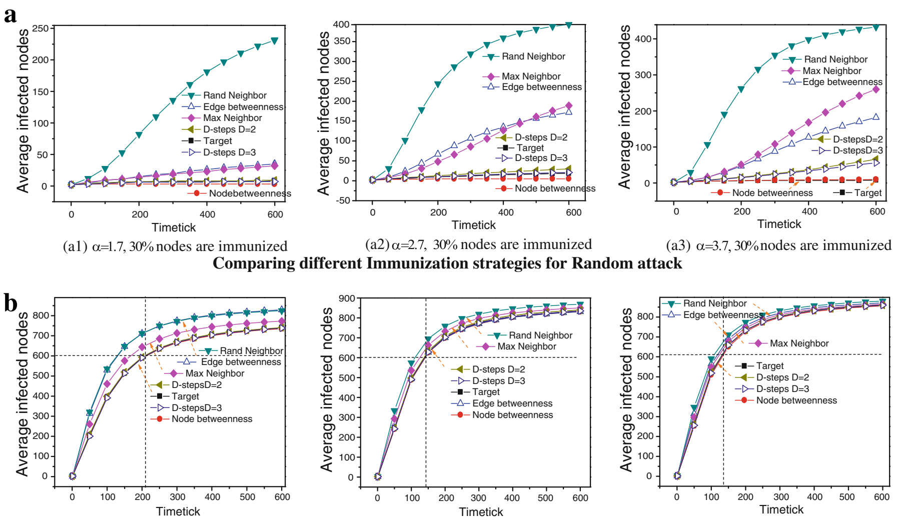
There are 30%immunized nodes

There are 5% immunized nodes

There are 10% immunized nodes

There are 30% immunized nodes

Fig. 6 APL' with respect to immunization strategies in three synthetic networks. **a**–**c** correspond to random attacks, whereas **d**–**f** are malicious attacks



(a1) α=1.7, 30% nodes are immunized (a2) α=2.7, 30% nodes are immunized (a3) α=3.7, 30% nodes are immunized

Comparing different Immunization strategies for Random attack

(b1) α=1.7, 5% nodes are immunized (b2) α=2.7, 5% nodes are immunized (b3) α=3.7, 5% nodes are immunized

Comparing different Immunization strategies for Maliciousattack

Fig. 7 Comparing immunization strategies for both random and malicious attacks in synthetic networks with different proportions of immunized nodes

Table 7 The synthetic community-based network with different attack modes

	5%			10%			30%		
	F	SID	APL'	F	SID	APL'	F	SID	APL'
Random attack									
Targeted	3,028	10,686	55.68	2,272	15,564	30.14	12	24,983	12.70
Node-betweenness	2,986	10,562	54.07	2,038	15,213	28.53	4	23,751	11.53
Edge-betweenness	3,056	9,782	62.90	2,300	14,393	36.31	386	21,312	23.18
Malicious attack									
Targeted	3,154	10,528	58.58	2,583	15,385	32.24	145	24,789	13.28
Nodes-betweenness	3,156	10,391	57.05	2,534	15,057	30.68	10	23,533	11.67
Edge-betweenness	3,238	8,390	79.15	2,710	13,003	50.96	13	20,168	35.82

If there is no immunization, the final number of infected nodes is 3,764 with a random attack and 3,778 with a malicious attack. The total simulation time $T = 600$

From the above experiments, the following conclusions can be made:

1. As shown in Tables 4–8, APL' can be used as a measurement to evaluate the efficiency of an immunization strategy. Thus, when designing a distributed immunization strategy, attentions should be paid on those nodes that have the largest impact on the APL value.
2. If the final number of infected nodes is used as a measure of efficiency, then the node-betweenness immunization strategy is more efficient than the targeted immunization strategy.
3. The power-law exponent (α) affects the edge-betweenness immunization strategy, but has a little impact on other strategies.

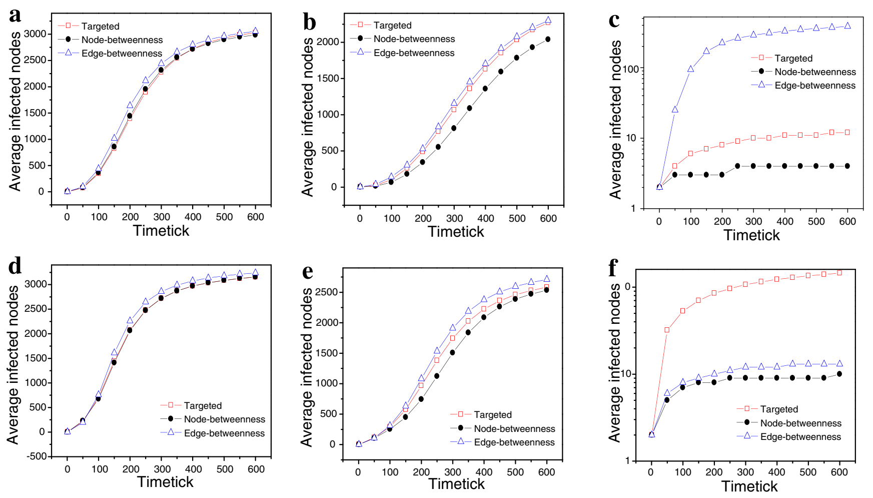


Fig. 8 Comparing immunization strategies for both random and malicious attacks in synthetic community-based networks with different proportions of immunized nodes. (**c, f**) are semilog charts. **a** The 5% nodes are immunized for random attack. **b** The 10% nodes are immunized for random attack. **c** The 30% nodes are immunized for random attack. **d** The 5% nodes are immunized for malicious attack. **e** The 10% nodes are immunized for malicious attack. **f** The 30% nodes are immunized for malicious attack

Table 8 The university email network with different attack modes

	5%			10%			30%		
	F	*SID*	APL'	*F*	*SID*	APL'	*F*	*SID*	APL'
Random attack									
Targeted	995	1,900	271.22	886	3,231	170.25	250	6,658	41.42
Node-betweenness	991	1,887	265.35	862	3,150	159.94	59	6,401	31.36
Edge-betweenness	1,001	1,759	292.36	911	2,899	196.02	492	5,610	81.22
Malicious attack									
Targeted	997	1,838	280.77	909	3,159	180.68	429	6,568	44.12
Nodes-betweenness	995	1,830	279.15	898	3,078	171.15	149	6,351	34.18
Edge-betweenness	1,010	1,332	346.80	929	2,416	251.06	581	5,091	113.18

If there is no immunization, the final number of infected nodes is 1,075 with a random attack and 1,088 with a malicious attack. The total simulation time $T = 600$

4.3.2 Immunization cost and robustness

In the previous section, the **efficiency** of different immunization strategies is evaluated in terms of the final number of infected nodes when the propagation reaches an equilibrium state. By doing experiments in synthetic networks, synthetic community-based network, the Enron email network and the university email network, it is easily to find that the node-betweenness immunization strategy has the highest efficiency. In this section, the performance of the different strategies will be evaluated in terms of cost and robustness, as in [20]. It is well known that the structure of a social network or an email network constantly evolves. It is therefore interesting to evaluate how changes in structure affect the efficiency of an immunization strategy.

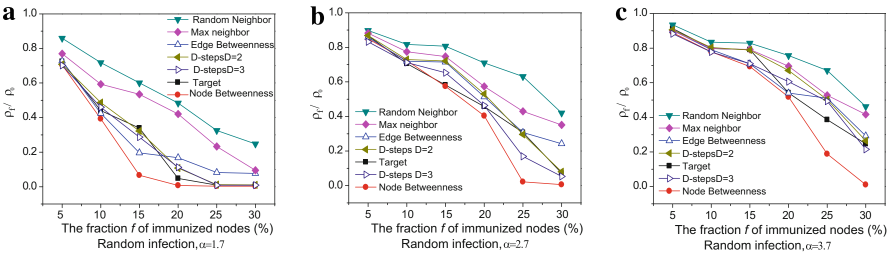
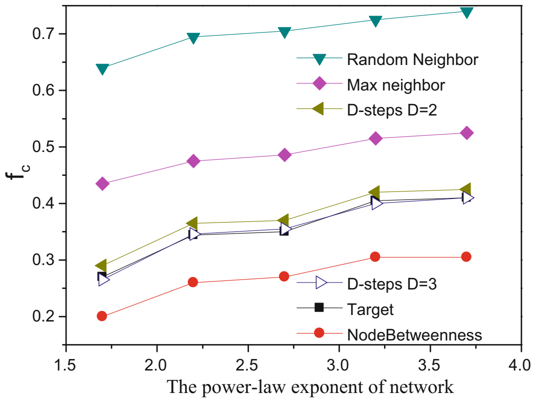


Fig. 9 The reduced prevalence ρ_f/ρ_0 as function of the fraction f of immunized nodes

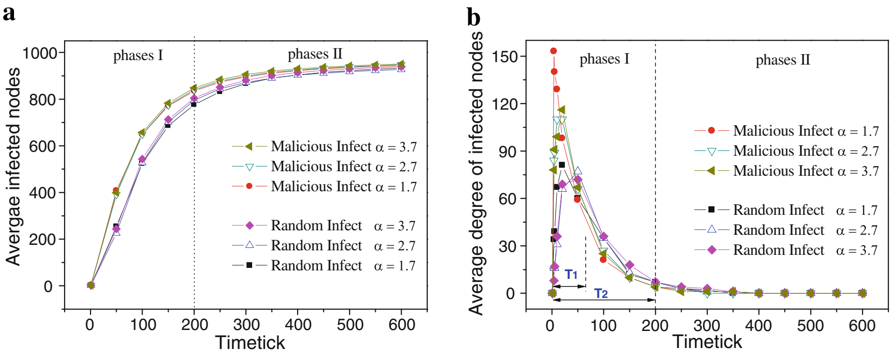
Fig. 10 The immunization critical value, f_c, as a function of the power-law exponent



- The **cost** can be defined as the number of nodes that need to be immunized in order to achieve a given level of epidemic prevalence ρ. Generally, $\rho \to 0$. There are some parameters which are of particular interest: f is the fraction of nodes that are immunized; f_c is the critical value of the immunization when $\rho \to 0$; ρ_0 is the infection density when no immunization strategy is implemented; ρ_f is the infection density with a given immunization strategy.
  Figure 9 shows the relationship between the reduced prevalence ρ_f/ρ_0 and f. It can be seen that the node-betweenness immunization has the lowest prevalence for the smallest number of protected nodes. The immunization cost increases as the value of α increases, i.e. in order to achieve epidemic prevalence $\rho \to 0$, the node-betweenness immunization strategy needs 20, 25, and 30% of nodes to be immunized, respectively, in the three synthetic networks. This is because the node-betweenness immunization strategy can effectively break the network structure and increase the path length of a network with the same number of immunized nodes.
- The **robustness** shows a plot of tolerance against the dynamic evolution of a network, i.e. the change of power-law exponents (α).
  Figure 10 shows the relationship between the immunized threshold f_c and α. A low level of f_c with a small variation indicates that the immunization strategy is robust. The robustness is important when an immunization strategy is deployed into a scalable and dynamic network (e.g. P2P and email networks). Figure 10 also shows the robustness of the D-steps immunization strategy is close to that of the targeted immunization; the node-betweenness strategy is the most robust.

Table 9 Degree statistics in the Enron email network and synthetic networks

	The size of degree >80	80–60	60–50	50–40	40–30	30–20	20–10	<10	Maximal degree
$\alpha = 1.7$	7	16	7	11	18	38	128	775	99
$\alpha = 2.7$	0	4	5	20	25	34	166	746	73
$\alpha = 3.7$	0	2	7	14	20	60	174	723	70
Enron	9	2	1	0	6	9	31	1,180	259



Virus spreading in the network without any immunization strategies

Fig. 11 Changes in the average number of infected nodes and the average degree of infected nodes over time, with virus propagation in different networks without applying any immunization strategies under different modes of attacks

4.4 The effect of network structure on virus propagation

[49] have compared virus propagation in synthetic networks with $\alpha = 1.7$ and $\alpha = 1.1475$, and pointed out that initial worm propagation has two phases. However, they do not give a detailed explanation of these results nor do they compare the effect of the power-law exponent on different immunization strategies during virus propagation.

Table 9 presents the detailed degree statistics for different networks, which can be used to examine the effect of the power-law exponent on virus propagation and immunization strategies.

First, virus propagation in non-immunized networks is discussed. Figure 11a shows the changes of the average number of infected nodes over time; Fig. 11b gives the average degree of infected nodes at each time step. From the results, it can be seen that

1. The number of infected nodes in non-immunized networks is determined by attack modes but not the power-law exponent.
   In Figs. 11a, b, three distribution curves ($\alpha = 1.7$, 2.7, and 3.7) overlap with each other in both random and malicious attacks. The difference between them is that the final number of infected nodes with a malicious attack is larger than that with a random attack, as shown in Fig. 11a, reflecting the fact that a malicious attack is more dangerous than a random attack.
2. A virus spreads more quickly in a network with a large power-law exponent than that with a small exponent.

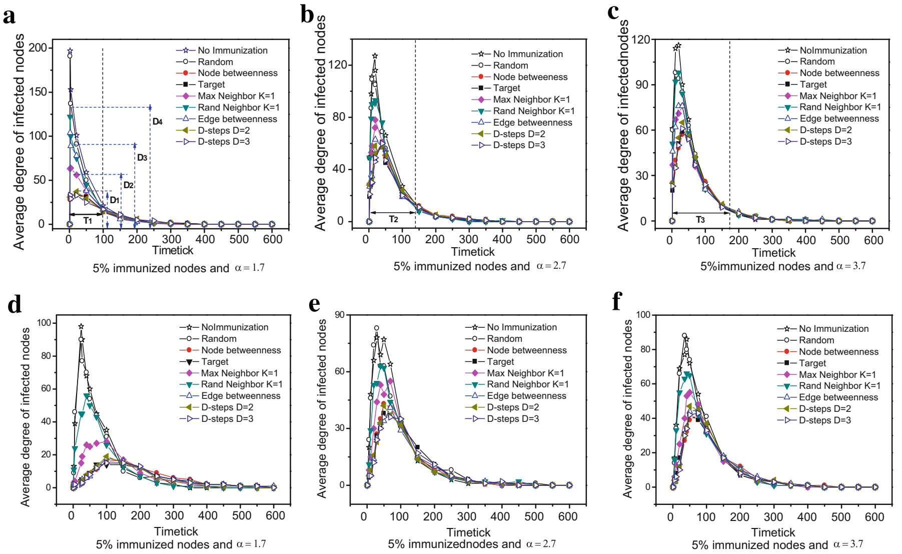


Fig. 12 Changes in the average degree of infected nodes over time, with virus propagation in different networks when applying different immunization strategies. **a**–**c** and **d**–**f** correspond to a malicious attack and random attack, respectively

Because a malicious attack initially infects highly connected nodes, the average degree of the infected nodes decreases in a shorter time comparing to a random attack ($T1 < T2$). Moreover, the speed and range of the infection is amplified by those highly connected nodes. In phase I, viruses propagate very quickly and infect most nodes in a network. However, in phase II, the number of total infected nodes grows slowly (Fig. 11a), because viruses aim to infect those nodes with low degrees (Fig. 11b), and a node with fewer links is more difficult to be infected.

In order to observe the effect of different immunization strategies on the average degree of infected nodes in different networks, 5% of the nodes are initially protected against random and malicious attacks. Figure 12 shows the simulation results. From this experiment, it can be concluded that

1. The random immunization has no effect on restraining virus propagation because the curves of the average degree of the infected nodes are basically coincident with the curves in the non-immunization case.
2. Comparing Fig. 12a, b, c and d, e, f, respectively, it can be seen that the peak value of the average degree is the largest in the network with α=1.7 and the smallest in the network with α=3.7. This is because the network with a lower exponent has more highly connected nodes (i.e. the range of degrees is between 50 and 80), which serve as amplifiers in the process of virus propagation.
3. As α increases, so does the number of infected nodes and the virus propagation duration ($T1 < T2 < T3$). Because a larger α implies a larger APL', the number of infected nodes will increase; if the network has a larger exponent, a virus need more time to infect those nodes with medium or low degrees.

Figures 13 and 14 use two cases to analyse the relationship between the average number of the infected nodes and the average degree of the infected nodes during virus propagation.

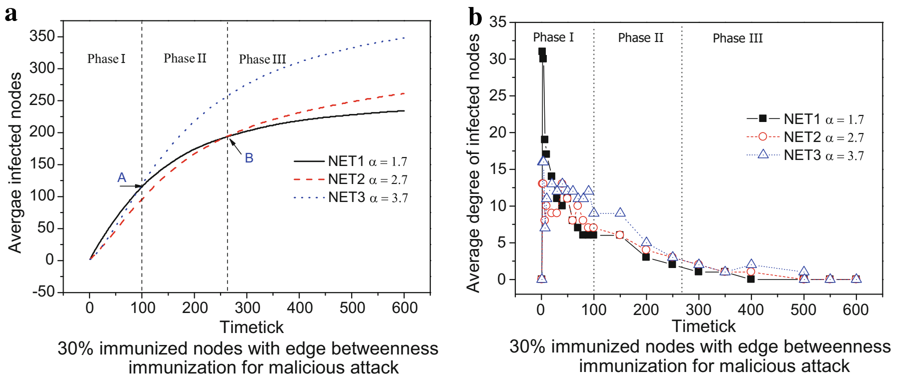


Fig. 13 The average number of infected nodes and the average degree of infected nodes, with respect to time when virus spreading in different networks. We apply the edge-betweenness immunization to protect 30% nodes in the network

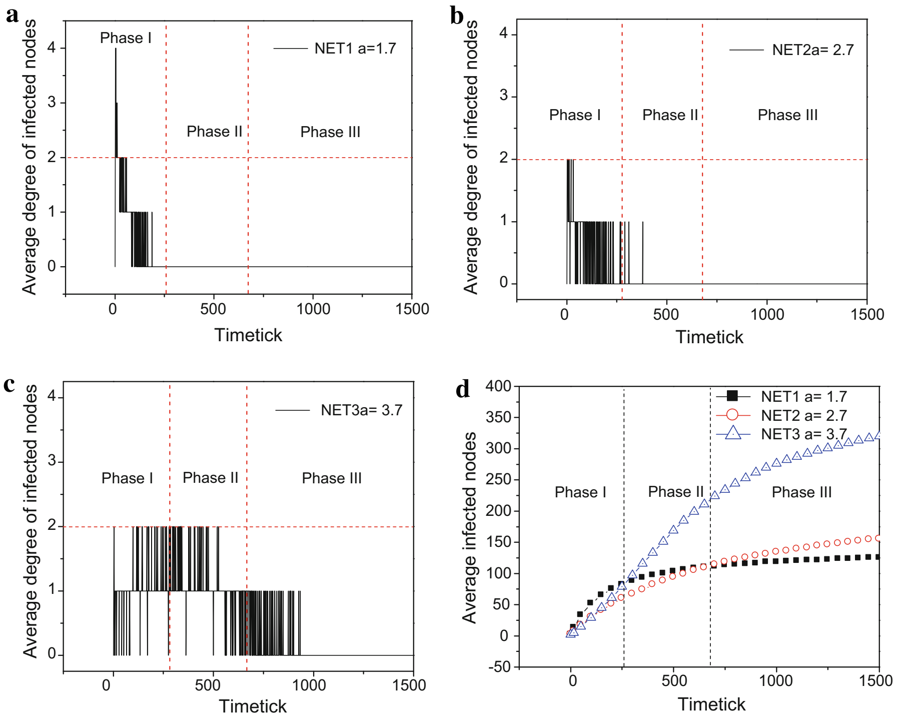


Fig. 14 The average number of infected nodes and the average degree of infected nodes, with respect to time when virus spreading in different networks. We apply the targeted immunization to protect 30% nodes in the network

First, consider the process of virus propagation in the case of a malicious attack where 30% of the nodes are immunized using the edge-betweenness immunization strategy. There are two intersections in Fig. 13a. Point *A* is the intersection of two curves NET1 and NET3, and point *B* is the intersection of NET2 and NET1. Under the same conditions, Fig. 13a shows that the total number of infected nodes is the largest in NET1 in Phase I. Corresponding to Fig. 13b, the average degree of infected nodes in NET1 is the largest in Phase I. As time goes on, the rate at which the average degree falls is the fastest in NET1, as shown in Fig. 13b. This is because there are more highly connected nodes in NET1 than in the others (see Table 9). After these highly connected nodes are infected, viruses attempt to infect the nodes with low degrees. Therefore, the average degree in NET3 that has the smallest power-law exponent is larger than those in Phases II and III. The total number of infected nodes in NET3 continuously increases, exceeding those in NET1 and NET2. The same phenomenon also appears in the targeted immunization strategy, as shown in Fig. 14.

4.5 The impact of human dynamics on virus propagation

The email-checking intervals in the above interactive email model (see Sect. 2.3) is modeled using a Poisson process. The Poisson distribution is widely used in many real-world models to statistically describe human activities, e.g. in terms of statistical regularities on the frequency of certain events within a period of time [25,49]. Statistics from user log files to databases that record the information about human activities, show that most observations on human behavior deviate from a Poisson process. That is to say, when a person engages in certain activities, his waiting intervals follow a power-law distribution with a long tail [27,43].

Vazquez et al. [44] have tried to incorporate an email-sending interval distribution, characterized by a power-law distribution, into a virus propagation model. However, their model assumes that a user is instantly infected after he/she receives a virus email, and ignores the impact of anti-virus software and the security awareness of users. Therefore, there are some gaps between their model and the real world.

In this section, the statistical properties associated with a single user sending emails is analyzed based on the Enron dataset [41]. The virus spreading process is then simulated using an improved interactive email model in order to observe the effect of human behavior on virus propagation.

4.5.1 Collective behaviors in email communication

Research results from the study of statistical regularities or laws of human behavior based on empirical data can offer a valuable perspective to social scientists [45,47]. Previous studies have also used models to characterize the behavioral features of sending emails [3,13,22], but their correctness needs to be further empirically verified, especially in view of the fact that there exist variations among different types of users. In this paper, the Enron email dataset is used to identify the characteristics of human email-handling behavior.

Due to the limited space, Table 10 presents only a small amount of the employee data contained in the database. As can be seen from the table, the interval distribution of email sent by the same user is respectively measured using different granularities: Day, Hour, and Minute. Figure 15 shows that the waiting intervals follow a heavy-tailed distribution. The power-law exponent as the Day granularity is not accurate because there are only a few data points. If more data points are added, a power-law distribution with long tail will emerge. Note that, there is a peak at $\Delta t = 16$ as measured at an Hour granularity. Eckmann et al. [13] have explained that the peak in a university dataset is the interval between the time people leave work and the time they return to their offices. After curve fitting, see Fig. 15, the waiting interval exponent is close to 1.3, i.e. $\alpha \approx 1.3 \pm 0.5$.

Although it has been shown that an email-sending distribution follows a power-law by studying users in the Enron dataset, it is still not possible to assert that all users' waiting

Table 10 Some information about individuals in the Enron dataset

ID	Name	Job	Messages	Start time	End time
44	John Arnold	Vice President	1,587	2000-02-27	2002-01-18
53	John Lavorato	CEO, Enron America	1,122	2001-01-26	2001-06-08
73	Jeff Dasovich	Government Relations Executive	6,272	1999-12-03	2002-09-22
107	Louise Kitchen	President, Enron Online	1,504	1999-05-24	2002-02-06
109	Vince Kaminski	Manager, Risk Management Head	1,219	2001-05-15	2002-01-30
122	Sally Beck	Employee, Chief Operating Officer	1,596	1999-12-13	2002-02-06

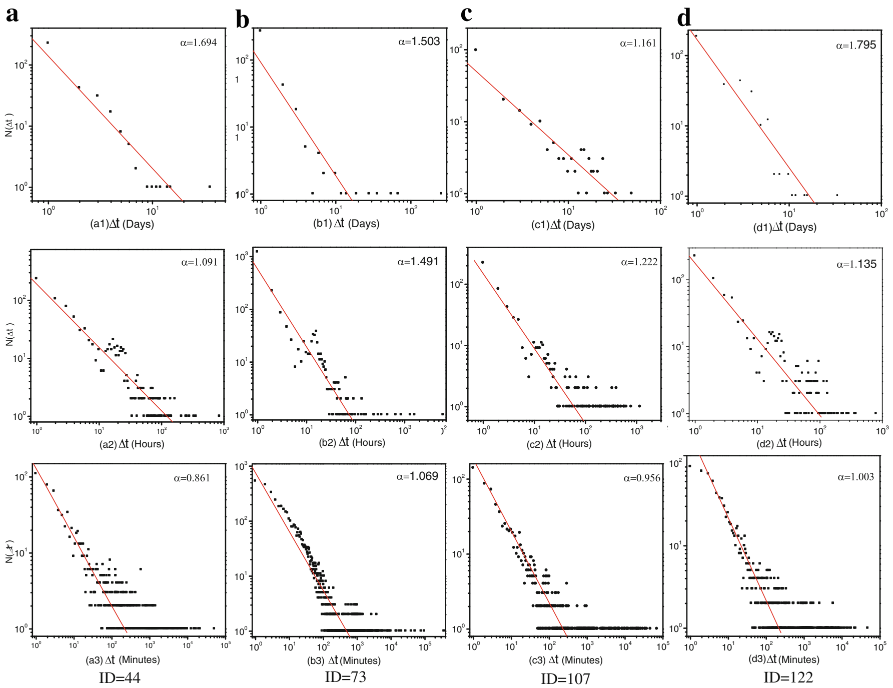


Fig. 15 Some typical inter-event distributions of different users in the Enron dataset. The x-scales correspond to Days, Hours, and Minutes, respectively. The *black points* are the raw data and the *lines* are fitted curves (logarithmic charts)

intervals follow a power-law distribution. It can only be stated that the distribution of waiting intervals has a long-tail characteristic. It is also not possible to measure the intervals between email checking since there is no information about login time in the Enron dataset. However, combing research results from human Web browsing behavior [10] and the effect of non-Poisson activities on propagation in the Barabasi group [44], it can be found that there are similarities between the distributions of email-checking intervals and email-sending intervals. The following section uses a power-law distribution to characterize the behavior associated with email-checking in order to observe the effect human behavior has on the propagation of an email virus.

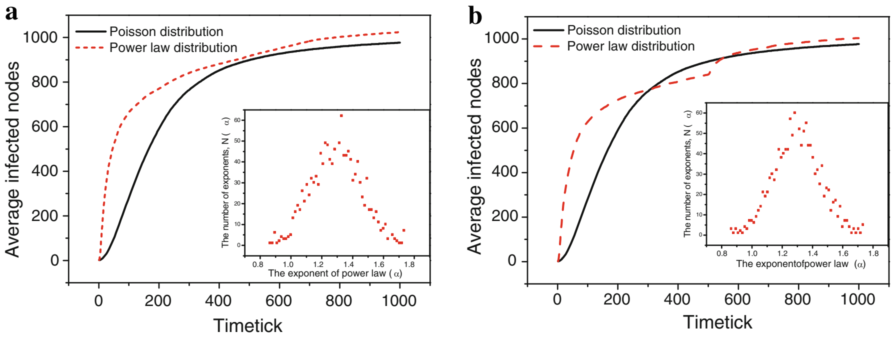


Fig. 16 The processes of virus propagation in the Enron email network. The *dashed line* and *solid line* indicate that the email-checking intervals follow a power-law distribution and a Poisson distribution, respectively. The inserted plot is a power-law exponent distribution of email-checking intervals for different users. **a** denotes that the latent viruses randomly break out, while **b** denotes that the latent viruses suddenly break out at $t = 500$

4.5.2 The effect of human behavior on virus propagation

Based on the above discussions, a power-law distribution is used to model the email-checking intervals of a user i, instead of the Poisson distribution used in [49], i.e. $T_i(\tau) \sim \tau^{-\alpha}$. An analysis of the distribution of the power-law exponent (α) for different individuals in Web browsing [10] and in the Enron dataset shows that the power-law exponent is approximately 1.3. In order to observe and quantitatively analyze the effect that the email-checking interval has on virus propagation, the email-clicking probability distribution (P_i) in our model is consistent with the one used by [49], i.e. the security awareness of different users in the network follows a normal distribution, $P_i \sim N(0.5, 0.3^2)$.

Figure 16 shows that following a random attack viruses quickly propagate in the Enron network if the email-checking intervals follow a power-law distribution. The results are consistent with the observed trends in real computer networks [31], i.e. viruses initially spread explosively, then enter a long latency period before becoming active again following user activity. The explanation for this is that users frequently have a short period of focused activity followed by a long period of inactivity. Thus, although old viruses may be killed by anti-virus software, they can still intermittently break out in a network. That is because some viruses are hidden by inactive users, and cannot be found by anti-virus software. When the inactive users become active, the virus will start to spread again.

The effect of human dynamics on virus propagation in three synthetic networks is also analyzed by applying the targeted [9], D-steps [17] and AOC-based strategy [24]. The numerical results are shown in Table. 11 and Fig. 17.

From the above experiments, the following conclusions can be made:

1. Based on the Enron email dataset and recent research on human dynamics, the email-checking intervals in an interactive email model should be assigned based on a power-law distribution.
2. Viruses can spread very quickly in a network if users' email-checking intervals follow a power-law distribution. In such a situation, viruses grow explosively at the initial stage and then grow slowly. The viruses remain in a latent state and await being activated by users.

Table 11 The final infected nodes in the different scale networks with random attack

	100		150		200	
	Power-law	Poisson	Power-law	Poisson	Power-law	Poisson
$\alpha = 1.7$						
Targeted	613	450	348	209	90	34
D-steps $D = 3$	625	505	375	220	90	38
AOC-based	558	466	328	164	50	16
$\alpha = 2.7$						
Targeted	767	706	653	564	467	378
D-steps $D = 3$	791	713	689	562	512	394
AOC-based	757	712	608	552	416	335
$\alpha = 3.7$						
Targeted	802	755	704	643	576	531
D-steps $D = 3$	827	762	735	668	602	538
AOC-based	797	758	706	630	545	507

The user's behaviors follow different distributions. There are 100, 150 and 200 immunized nodes, respectively. The total simulation time $T = 1{,}000$

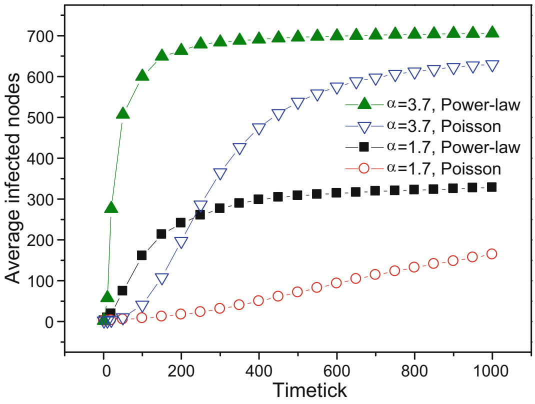


Fig. 17 Virus propagation in synthetic networks with different power-law exponents $\alpha = 1.7$ and $\alpha = 3.7$. There are 150 immunized nodes based on the AOC-based strategy. The final number of infected nodes is the largest when the network has a large power-law exponent and the checking email intervals follow a power-law distribution

5 Conclusion

In this paper, a simulation model for studying the process of virus propagation has been described, and the efficiency of various existing immunization strategies has been compared. In particular, two new betweenness-based immunization strategies have been presented and validated in an interactive propagation model, which incorporates two human behaviors based on [49] in order to make the model more practical. This simulation-based work can be regarded as a contribution to the understanding of the inter-reactions between a network structure and local/global dynamics. The main results are concluded as follows:

1. Some experiments are used to systematically compare different immunization strategies for restraining epidemic spreading, in synthetic scale-free networks including the community-based network and two real email networks. The simulation results have shown

that the key factor that affects the efficiency of immunization strategies is *APL*, rather than the sum of the degrees of immunized nodes (*SID*). That is to say, immunization strategy should protect nodes with higher connectivity and transmission capability, rather than those with higher degrees.

2. Some performance metrics are used to further evaluate the efficiency of different strategies, i.e. in terms of their cost and robustness. Simulation results have shown that the *D*-steps immunization is a feasible strategy in the case of limited resources and the node-betweenness immunization is the best if the global topological information is available.
3. The effects of power-law exponents and human dynamics on virus propagation are analyzed. More in-depth experiments have shown that viruses spread faster in a network with a large power-law exponent than that with a small one. Especially, the results have explained why some old viruses can still propagate in networks up till now from the perspective of human dynamics.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China Grant (60673015), Beijing Natural Science Foundation (4102007), and The Doctoral Student Innovation Program of Beijing University of Technology (bcx-2009-062). The authors would like to thank Dr. R. Dale-Jones and Dr. Hongjun Qiu for their great help in polishing the language.

References

1. Bailey NTJ (1975) The mathematical theory of infectious diseases and its applications. Hafner Press, New York
2. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
3. Barabasi AL (2005) The origin of bursts and heavy tails in human dynamics. Nature 435(7039):207–211
4. Bar-Yossef Z, Guy I, Lempel R, Maarek YS, Soroka V (2008) Cluster ranking with an application to mining mailbox networks. Knowl Inf Syst 14:101–139
5. Boots M, Sasaki A, Ben-Averaham D (1999) 'Small worlds' and the evolution of virulence: infection occurs locally and at a distance. Proc R Soc Lond B Biol Sci 266(1432):1933–1938
6. Bu T, Towsley D (2002) On distinguishing between internet power law topology generators. In: Lee D, Orda A (eds) Proceedings of the twenty first annual joint conference of the IEEE computer and communications societies (INFOCOM'02). IEEE Press, New York, pp 638–647
7. Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distribution in empirical data. SIAM Rev 51(4):661–703
8. Cohen R, Havlin S, Ben-Averaham D (2003) Efficient immunization strategies for computer networks and populations. Phys Rev Lett 91(24):247901
9. Dezso Z, Barabasi AL (2002) Halting viruses in scale-free networks. Phys Rev E 65(5):055103
10. Dezso Z, Almaas E, Lukacs A, Racz B, Szakadat I, Barabasi AL (2006) Dynamics of information access on the web. Phys Rev E 73(6):066132
11. Earn DJD, Rohani P, Bolker BM, Grenfell BT (2000) A simple model for complex dynamical transitions in epidemics. Science 287(5453):667–670
12. Echenique P, Gomez-Gardenes J, Moreno Y, Vazquez A (2005) Distance-d covering problem in scale-free networks with degree correlation. Phys Rev E 71(3):035102
13. Eckmann JP, Moses E, Sergi D (2004) Entropy of dialogues creates coherent structure in email traffic. Proc Natl Acad Sci U S A 101(40):14333–14337
14. Eguiluz VM, Klemm K (2002) Epidemic threshold in structured scale-free networks. Phys Rev Lett 89(10):108701
15. Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. ACM SIGCOMM Comput Commun Rev 29(4):251–262
16. Gallos LK, Liljeros F, Argyrakis P, Bunde A, Havlin S (2007) Improving immunization strategies. Phys Rev E 75(4):045104
17. Gomez-Gardenes J, Echenique P, Moreno Y (2002) Immunization of real complex communication networks. Eur Phys J B 49(2):259–264
18. Guimera R, Danon L, Diaz-Guilera A, Giralt F, Arenas A (2004) Self-similar community structure in a network of human interactions. Phys Rev E 68(6):065103

19. Holme P, Kim BJ, Yoon CN, Han SK (2002) Attack vulnerability of complex networks. Phys Rev E 65(5):056109
20. Huang XL, Zou FT, Ma FY (2007) Targeted local immunization in scale-free peer-to-peer networks. J Comput Sci Technol 22(3):457–468
21. Jeong H, Tombor B, Albert R, OItvai ZN, Barabasi AL (2000) The large scale organization of metabolic networks. Nature 407(6804):651–654
22. Johansen A (2004) Probing human response times. Physica A 338(1–2):286–291
23. Lahiri M, Berger-Wolf TY (2009) Periodic subgraph mining in dynamic networks. Knowledge and information systems. doi:10.1007/s10115-009-0253-8
24. Liu JM, Gao C, Zhong N (2010) Autonomy-oriented search in dynamic community networks: a case study in decentralized network immunization. Fundam Informaticae 99(2):207–226
25. Liu JM, Zhang SW, Yang J (2004) Characterizing web usage regularities with information foraging agents. IEEE Trans Knowl Data Eng 16(5):566–584
26. Lloyd AL, May RM (2001) How viruses spread among computers and people. Science 292(5520):1316–1317
27. Malmgren RD, Stouffer DB, Campanharo ASLO, Amaral LAN (2009) On universality in human correspondence activity. Science 325(5948):1696–1700
28. May SR (2000) Enhanced: simple rules with complex dynamics. Science 287(5453):601–602
29. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs simple building blocks of complex networks. Science 298(5594):824–827
30. Moore C, Newman MEJ (2000) Epidemics and percolation in small-world network. Phys Rev E 61(5):5678–5682
31. Moore D, Shannon C, Brown J (2002) Code-red: a case study on the spread and victims of an internet worm. Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurement(IMW'02), Marseille, France, pp 273–284
32. Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci U S A 98(2):404–409
33. Newman MEJ (2002) The spread of epidemic disease on networks. Phys Rev E 66(1):016128
34. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
35. Newman MEJ, Forrest S, Balthrop J (2002) Email networks and the spread of computer viruses. Phys Rev E 66(3):035101
36. Narasimhamurthy A, Greene D, Hurley N, Cunningham P (2009) Partitioning large networks without breaking communities. Knowl Inf Syst. doi:10.1007/s10115-009-0251-x
37. Pastor-Satorras R, Vespignani A (2001a) Epidemic spreading in scale-free networks. Phys Rev Lett 86(14):3200–3203
38. Pastor-Satorras R, Vespignani A (2001b) Epidemic dynamics and endemic states in complex networks. Phys Rev E 63(6):066117
39. Pastor-Satorras R, Vespignani A (2002) Immunization of complex networks. Phys Rev E 65(3):036104
40. Serazzi G, Zanero S (2004) Computer virus propagation models. In: Calzarossa M, Gelenbe E (eds) Performance tools and applications to networked systems, revised tutorial lectures, LNCS 2965. Springer, Heidelberg pp 26–50
41. Shetty J, Adibi J (2004) The Enron email dataset database schema and brief statistical report. Technical report, Information Sciences Institute
42. Strogatz SH (2001) Exploring complex networks. Nature 410(6825):268–276
43. Vazquez A, Oliveira JG, Dezso Z, Goh KI, Kondor I, Barabasi AL (2006) Modeling bursts and heavy tails in human dynamics. Phys Rev E 73(3):036127
44. Vazquez A, Racz B, Lukacs A, Barabasi AL (2007) Impact of non-poissonian activity patterns on spreading process. Phys Rev Lett 98(15):158702
45. Vespignani A (2009) Predicting the behavior of techno-social systems. Science 325(5939):425–428
46. Wang D, Tse QCK, Zhou Y (2009) A decentralized search engine for dynamic web communities. Knowl Inf Syst. doi:10.1007/s10115-009-0270-7
47. Watts DJ (2007) A twenty-first century science. Nature 445(7127):489
48. Whalley I, Arnold B, Chess D, Morar J, Segal A, Swimmer M (2000) An environment for controlled worm replication and analysis. Virus Bulletin 1–20
49. Zou CC, Towsley D, Gong W (2007) Modeling and simulation study of the propagation and defense of internet e-mail worms. IEEE Trans Dependable Secur Comput 4(2):105–118

Author Biographies

Chao Gao is currently a PhD student in the International WIC Institute, College of Computer Science and Technology, Beijing University of Technology. He has been an exchange student in the Department of Computer Science, Hong Kong Baptist University. His main research interests include Web Intelligence (WI), Autonomy-Oriented Computing (AOC), complex networks analysis, and network security.

Jiming Liu is the Chair Professor and Head of Computer Science Department at Hong Kong Baptist University. He was a Professor and the Director of School of Computer Science at University of Windsor, Canada. His current research interests include: Autonomy-Oriented Computing (AOC), Web Intelligence (WI), and self-organizing systems and complex networks, with applications to: (i) characterizing working mechanisms that lead to emergent behavior in natural and artificial complex systems (e.g., phenomena in Web Science, and the dynamics of social networks and neural systems), and (ii) developing solutions to large-scale, distributed computational problems (e.g., distributed scalable scientific or social computing, and collective intelligence). Prof. Liu has contributed to the scientific literature in those areas, including over 250 journal and conference papers, and 5 authored research monographs, e.g., *Autonomy-Oriented Computing: From Problem Solving to Complex Systems Modeling* (Kluwer Academic/Springer) and *Spatial Reasoning and Planning: Geometry, Mechanism, and Motion* (Springer). Prof. Liu has served as the Editor-in-Chief of *Web Intelligence and Agent Systems*, an Associate Editor of *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Systems, Man, and Cybernetics—Part B*, and *Computational Intelligence*, and a member of the Editorial Board of several other international journals.

Ning Zhong is currently Head of the Knowledge Information Systems Laboratory and is a Professor in the Department of Systems and Information Engineering at Maebashi Institute of Technology, Japan. He is also an Adjunct Professor in the International WIC Institute. He has conducted research in the areas of knowledge discovery and data mining, rough sets and granular-soft computing, Web Intelligence (WI), intelligent agents, brain informatics, and knowledge information systems, with more than 250 journal and conference publications and 10 books. He is the Editor-in-Chief of *Web Intelligence and Agent Systems* and *Annual Review of Intelligent Informatics*, an Associate Editor of *IEEE Transactions on Knowledge and Data Engineering*, *Data Engineering*, and *Knowledge and Information Systems*, a member of the editorial board of *Transactions on Rough Sets*.