Malware Propagation in Large-Scale Networks

Shui Yu, *Senior Member, IEEE*, Guofei Gu, *Member, IEEE*, Ahmed Barnawi, *Member, IEEE*, Song Guo, *Senior Member, IEEE*, and Ivan Stojmenovic, *Fellow, IEEE*

Abstract—Malware is pervasive in networks, and poses a critical threat to network security. However, we have very limited understanding of malware behavior in networks to date. In this paper, we investigate how malware propagates in networks from a global perspective. We formulate the problem, and establish a rigorous two layer epidemic model for malware propagation from network to network. Based on the proposed model, our analysis indicates that the distribution of a given malware follows exponential distribution, power law distribution with a short exponential tail, and power law distribution at its early, late and final stages, respectively. Extensive experiments have been performed through two real-world global scale malware data sets, and the results confirm our theoretical findings.

Index Terms—Malware, propagation, modelling, power law

1 Introduction

Malware are malicious software programs deployed by cyber attackers to compromise computer systems by exploiting their security vulnerabilities. Motivated by extraordinary financial or political rewards, malware owners are exhausting their energy to compromise as many networked computers as they can in order to achieve their malicious goals. A compromised computer is called a bot, and all bots compromised by a malware form a botnet. Botnets have become the attack engine of cyber attackers, and they pose critical challenges to cyber defenders. In order to fight against cyber criminals, it is important for defenders to understand malware behavior, such as propagation or membership recruitment patterns, the size of botnets, and distribution of bots.

To date, we do not have a solid understanding about the size and distribution of malware or botnets. Researchers have employed various methods to measure the size of botnets, such as botnet infiltration [1], DNS redirection [3], external information [2]. These efforts indicate that the size of botnets varies from millions to a few thousand. There are no dominant principles to explain these variations. As a result, researchers desperately desire effective models and explanations for the chaos. Dagon et al. [3] revealed that time zone has an obvious impact on the number of available bots. Mieghem et al. [4] indicated that network topology has an important impact on malware spreading through their rigorous mathematical analysis. Recently, the emergence of mobile malware, such as Cabir [5], Ikee [6], and Brador [7], further increases the difficulty level of our understanding on how they propagate. More details about mobile malware can be found at a recent survey paper [8]. To the best of our knowledge, the best finding about malware distribution in large-scale networks comes from Chen and Ji [9]: the distribution is non-uniform. All this indicates that the research in this field is in its early stage.

The epidemic theory plays a leading role in malware propagation modelling. The current models for malware spread fall in two categories: the epidemiology model and the control theoretic model. The control system theory based models try to detect and contain the spread of malware [10], [11]. The epidemiology models are more focused on the number of compromised hosts and their distributions, and they have been explored extensively in the computer science community [12], [13], [14]. Zou et al. [15] used a susceptible-infected (SI) model to predict the growth of Internet worms at the early stage. Gao and Liu [16] recently employed a susceptible-infected-recovered (SIR) model to describe mobile virus propagation. One critical condition for the epidemic models is a large vulnerable population because their principle is based on differential equations. More details of epidemic modelling can be find in [17]. As pointed by Willinger et al. [18], the findings, which we extract from a set of observed data, usually reflect parts of the studied objects. It is more reliable to extract theoretical results from appropriate models with confirmation from sufficient real world data set experiments. We practice this principle in this study.

In this paper, we study the distribution of malware in terms of networks (e.g., autonomous systems (AS), ISP domains, abstract networks of smartphones who share the same vulnerabilities) at large scales. In this kind of setting, we have a sufficient volume of data at a large enough scale to meet the requirements of the SI model. Different from the

- *S. Yu is with the School of Information Technology, Deakin University, Burwood, Victoria 3125, Australia. E-mail: syu@deakin.edu.au.*
- *G. Gu is with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843-3112. E-mail: guofei@cse.tamu.edu.*
- *A. Barnawi is with the Faculty of Computing and IT, King Abdulaziz University, Jeddah, Saudi Arabia. E-mail: ambarnawi@kau.edu.sa.*
- *S. Guo is with the School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan. E-mail: sguo@u-aizu.ac.jp.*
- *I. Stojmenovic is with the School of Information Technology, Deakin University, Australia; King Abdulaziz University, Jeddah, Saudi Arabia; and the School of EECS, University of Ottawa, Ottawa, ON K1N 6N5, Canada. E-mail: ivan@site.uottawa.ca.*

Manuscript received 1 Jan. 2014; revised 14 Apr. 2014; accepted 15 Apr. 2014. Date of publication 28 Apr. 2014; date of current version 1 Dec. 2014.
Recommended for acceptance by F. Bonchi.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2014.2320725

1041-4347 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

traditional epidemic models, we break our model into two layers. First of all, for a given time since the breakout of a malware, we calculate how many networks have been compromised based on the SI model. Second, for a compromised network, we calculate how many hosts have been compromised since the time that the network was compromised. With this two layer model in place, we can determine the total number of compromised hosts and their distribution in terms of networks. Through our rigorous analysis, we find that the distribution of a given malware follows an exponential distribution at its early stage, and obeys a power law distribution with a short exponential tail at its late stage, and finally converges to a power law distribution. We examine our theoretical findings through two large-scale real-world data sets: the Android based malware [19] and the Conficker [20]. The experimental results strongly support our theoretical claims. To the best of our knowledge, the proposed two layer epidemic model and the findings are the first work in the field.

Our contributions are summarized as follows.

- We propose a two layer malware propagation model to describe the development of a given malware at the Internet level. Compared with the existing single layer epidemic models, the proposed model represents malware propagation better in large-scale networks.
- We find the malware distribution in terms of networks varies from exponential to power law with a short exponential tail, and to power law distribution at its early, late, and final stage, respectively. These findings are first theoretically proved based on the proposed model, and then confirmed by the experiments through the two large-scale real-world data sets.

The rest of the paper is structured as follows. Related work is briefly listed in Section 2. We present the preliminaries for the proposed model in Section 3. The studied problem is discussed in Section 4. A two layer malware propagation model is established in Section 5, and followed by a rigorous mathematical analysis in Section 6. Experiments are conducted to confirm our findings in Section 7. In Section 8, we provide a further discussion about the study. Finally, we summarize the paper and present future work in Section 9.

2 Related Work

The basic story of malware is as follows. A malware programer writes a program, called bot or agent, and then installs the bots at compromised computers on the Internet using various network virus-like techniques. All of his bots form a botnet, which is controlled by its owners to commit illegal tasks, such as launching DDoS attacks, sending spam emails, performing phishing activities, and collecting sensitive information. There is a command and control (*C&C*) server(s) to communicate with the bots and collect data from bots. In order to disguise himself from legal forces, the botmaster changes the url of his *C&C* frequently, e.g., weekly. An excellent explanation about this can be found in [1].

With the significant growing of smartphones, we have witnessed an increasing number of mobile malware. Malware writers have develop many mobile malware in recent years. Cabir [5] was developed in 2004, and was the first malware targeting on the Symbian operating system for mobile devices. Moreover, it was also the first malware propagating via Bluetooth. Ikee [6] was the first mobile malware against Apple iPhones, while Brador [7] was developed against Windows CE operating systems. The attack victors for mobile malware are diverse, such as SMS, MMS, Bluetooth, WiFi, and Web browsing. Peng et al. [8] presented the short history of mobile malware since 2004, and surveyed their propagation models.

A direct method to count the number of bots is to use botnet infiltration to count the bot IDs or IP addresses. Stone-Gross et al. [1] registered the URL of the Torpig botnet before the botmaster, and therefore were able to hijack the *C&C* server for ten days, and collect about 70G data from the bots of the Torpig botnet. They reported that the footprint of the Torpig botnet was 182,800, and the median and average size of the Torpig's live population was 49,272 and 48,532, respectively. They found 49,294 new infections during the ten days takeover. Their research also indicated that the live population fluctuates periodically as users switch between being online and offline. This issue was also tacked by Dagon et al. in [3].

Another method is to use DNS redirection. Dagon et al. [3] analyzed captured bots by honypot, and then identified the *C&C* server using source code reverse engineering tools. They then manipulated the DNS entry which is related to a botnet's IRC server, and redirected the DNS requests to a local sinkhole. They therefore could count the number of bots in the botnet. As discussed previously, their method counts the footprint of the botnet, which was 350,000 in their report.

In this paper, we use two large scale malware data sets for our experiments. Conficker is a well-known and one of the most recently widespread malware. Shin et al. [20] collected a data set about 25 million Conficker victims from all over the world at different levels. At the same time, malware targeting on Android based mobile systems are developing quickly in recent years. Zhou and Jiang [19] collected a large data set of Android based malware.

In [2], Rajab et al. pointed out that it is inaccurate to count the unique IP addresses of bots because DHCP and NAT techniques are employed extensively on the Internet ([1] confirms this by their observation that 78.9 percent of the infected machines were behind a NAT, VPN, proxy, or firewall). They therefore proposed to examine the hits of DNS caches to find the lower bound of the size of a given botnet.

Rajab et al. [21] reported that botnets can be categorized into two major genres in terms of membership recruitment: worm-like botnets and variable scanning botnets. The latter weights about 82 percent in the 192 IRC bots that they investigated, and is the more prevalent class seen currently. Such botnets usually perform localized and non-uniform scanning, and are difficult to track due to their intermittent and continuously changing behavior. The statistics on the lifetime of bots are also reported as 25 minutes on average with 90 percent of them staying for less than 50 minutes.

TABLE 1
Notations of Symbols in This Paper

Notation	Description
$I(t)$	Number of infected hosts at time t
$R(t)$	Number of recovered hosts at time t
N	The total number of vulnerable hosts
$\beta(t)$	The infection rate at time t
$S(L_i, t)$	Number of infected hosts of network L_i at time t
$L_{k_i}^j$	The j^{th} network compromised at round i

Malware propagation modelling has been extensively explored. Based on epidemiology research, Zou et al. [15] proposed a number of models for malware monitoring at the *early* stage. They pointed out that these kinds of model are appropriate for a system that consists of a large number of vulnerable hosts; in other words, the model is effective at the *early* stage of the outbreak of malware, and the accuracy of the model drops when the malware develops further. As a variant of the epidemic category, Sellke et al. [12] proposed a stochastic branching process model for characterizing the propagation of Internet worms, which especially focuses on the number of compromised computers against the number of worm scans, and presented a closed form expression for the relationship. Dagon et al. [3] extended the model of [15] by introducing time zone information $\alpha(t)$, and built a model to describe the impact on the number of live members of botnets with diurnal effect.

The impact of side information on the spreading behavior of network viruses has also been explored. Ganesh et al. [22] thoroughly investigated the effect of network topology on the spead of epidemics. By combining Graph theory and a SIS (susceptible—infective—susceptible) model, they found that if the ratio of cure to infection rates is smaller than the spectral radius of the graph of the studied network, then the average epidemic lifetime is of order $\log n$, where n is the number of nodes. On the other hand, if the ratio is larger than a generalization of the isoperimetric constant of the graph, then the average epidemic lifetime is of order e^{n^a}, where a is a positive constant. Similarly, Mieghem et al. [4] applied the N-intertwined Markov chain model, an application of mean field theory, to analyze the spread of viruses in networks. They found that $\tau_c = \frac{1}{\lambda_{max}(A)}$, where τ_c is the sharp epidemic threshold, and $\lambda_{max}(A)$ is the largest eigenvalue of the adjacency matrix A of the studied network. Moreover, there have been many other methodologies to tackle the problem, such as game theory [23].

3 Preliminaries

Preliminaries of epidemic modelling and complex networks are presented in this section as this work is mainly based on the two fields.

For the sake of convenience, we summarize the symbols that we use in this paper in Table 1.

3.1 Deterministic Epidemic Models

After nearly 100 years development, the epidemic models [17] have proved effective and appropriate for a system that possesses a large number of vulnerable hosts. In other words, they are suitable at a macro level. Zou et al. [15] demonstrated that they were suitable for the studies of Internet based virus propagation at the early stage.

We note that there are many factors that impact the malware propagation or botnet membership recruitment, such as network topology, recruitment frequency, and connection status of vulnerable hosts. All these factors contribute to the speed of malware propagation. Fortunately, we can include all these factors into one parameter as *infection rate* β in epidemic theory. Therefore, in our study, let N be the total number of vulnerable hosts of a large-scale network (e.g., the Internet) for a given malware. There are two statuses for any one of the N hosts, either infected or susceptible. Let $I(t)$ be the number of infected hosts at time t, then we have

$$\frac{dI(t)}{dt} = \beta(t)[N - R(t) - I(t) - Q(t)]I(t) - \frac{dR(t)}{dt}, \quad (1)$$

where $R(t)$, and $Q(t)$ represent the number of removed hosts from the infected population, and the number of removed hosts from the susceptible population at time t. The variable $\beta(t)$ is the infection rate at time t.

For our study, model (1) is too detailed and not necessary as we expect to know the propagation and distribution of a given malware. As a result, we employ the following susceptible-infected model:

$$\frac{dI(t)}{dt} = \beta I(t)[N - I(t)], \quad (2)$$

where the infection rate β is a constant for a given malware for any network.

We note that the variable t is continuous in model (2) and (1). In practice, we measure $I(t)$ at discrete time points. Therefore, $t = 0, 1, 2, \ldots$. We can interpret each time point as a new round of malware membership recruitment, such as vulnerable host scanning. As a result, we can transform model (2) into the discrete form as follows:

$$I(t) = (1 + \alpha\Delta)I(t-1) - \beta\Delta I(t-1)^2, \quad (3)$$

where $t = 0, 1, 2, \ldots$, Δ is the unit of time, $I(0)$ is the initial number of infected hosts (we also call them *seeds* in this paper), and $\alpha = \beta N$, which represents the average number of vulnerable hosts that can be infected by one infected host per time unit.

In order to simplify our analysis, let $\Delta = 1$, it could be one second, one minute, one day, or one month, even one year, depending on the time scale in a given context. Hence, we have a simpler discrete form given by

$$I(t) = (1 + \alpha)I(t-1) - \beta(I(t-1))^2. \quad (4)$$

Based on Equation (4), we define the increase of infected hosts for each time unit as follows.

$$\Delta I(t) \triangleq I(t) - I(t-1), t = 1, 2, \ldots. \quad (5)$$

To date, many researches are confined to the "early stage" of an epidemic, such as [15]. Under the early stage condition, $I(t) << N$, therefore, $N - I(t) \approx N$. As a result, a closed form solution is obtained as follows:

$$I(t) = I(0)e^{\beta N t}. \quad (6)$$

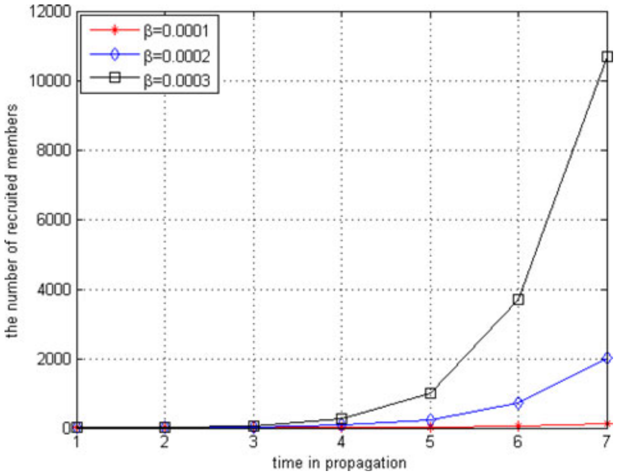


Fig. 1. The impact from infection rate β on the recruitment progress for a given vulnerable network with $N=$ 10,000.

When we take the $\ln$ operation on both sides of Equation (6), we have

$$\ln I(t) = \beta N t + \ln I(0). \tag{7}$$

For a given vulnerable network, β, N and $I(0)$ are constants, therefore, the graphical representation of Equation (7) is a straight line.

Based on the definition of Equation (5), we obtain the increase of new members of a malware at the early stage as

$$\begin{aligned}\Delta I(t) &= (e^{\beta N} - 1) I(t-1)\\ &= (e^{\beta N} - 1) I(0) e^{\beta N(t-1)}.\end{aligned} \tag{8}$$

Taking the $\ln$ operation on both side of (8), we have

$$\ln \Delta I(t) = \beta N(t-1) + \ln\left((e^{\beta N} - 1) I(0)\right). \tag{9}$$

Similar to Equation (7), the graphical representation of equation (9) is also a straight line. In other words, the number of recruited members for each round follows an exponential distribution at the early stage.

We have to note that it is hard for us to know whether an epidemic is at its early stage or not in practice. Moreover, there is no mathematical definition about the term early stage.

In epidemic models, the infection rate β has a critical impact on the membership recruitment progress, and β is usually a small positive number, such as 0.00084 for worm Code Red [12]. For example, for a network with $N = 10{,}000$ vulnerable hosts, we show the recruited members under different infection rates in Fig. 1. From this diagram, we can see that the recruitment goes slowly when $\beta = 0.0001$, however, all vulnerable hosts have been compromised in less than 7 time units when $\beta = 0.0003$, and the recruitment progresses in an exponential fashion.

This reflects the malware propagation styles in practice. For malware based on "contact", such as blue tooth contacts, or viruses depending on emails to propagate, the infection rate is usually small, and it takes a long time to compromise a large number of vulnerable hosts in a given network. On the other hand, for some malware, which take active actions for recruitment, such as vulnerable host scanning, it may take one or a few rounds of scanning to recruit all or a majority of the vulnerable hosts in a given network. We will apply this in the following analysis and performance evaluation.

3.2 Complex Networks

Research on complex networks have demonstrated that the number of hosts of networks follows the power law. People found that the size distribution usually follows the power law, such as population in cities in a country or personal income in a nation [24]. In terms of the Internet, researchers have also discovered many power law phenomenon, such as the size distribution of web files [25]. Recent progresses reported in [26] further demonstrated that the size of networks follows the power law.

The power law has two expression forms: the Pareto distribution and the Zipf distribution. For the same objects of the power law, we can use any one of them to represent it. However, the Zipf distributions are tidier than the expression of the Pareto distributions. In this paper, we will use Zipf distributions to represent the power law. The Zipf expression is as follows:

$$Pr\{x = i\} = \frac{C}{i^{\alpha}}, \tag{10}$$

where C is a constant, α is a positive parameter, called the *Zipf index*, $Pr\{x=i\}$ represents the probability of the ith $(i = 1, 2, \ldots)$ largest object in terms of size, and $\sum_i Pr\{x = i\} = 1$.

A more general form of the distribution is called the Zipf-Mandelbrot distribution [27], which is defined as follows:

$$Pr\{x = i\} = \frac{C}{(i+q)^{\alpha}}, \tag{11}$$

where the additional constant q $(q \geq 0)$ is called the *plateau factor*, which makes the probability of the highest ranked objects flat. The Zipf-Mandelbrot distribution becomes the Zipf distribution when $q = 0$.

Currently, the metric to say a distribution is a power law is to take the loglog plot of the data, and we usually say it is a power law if the result shows a straight line. We have to note that this is not a rigorous method, however, it is widely applied in practice. Power law distributions enjoy one important property, scale free. We refer interested readers to [28] about the power law and its properties.

4 Problem Description

In this section, we describe the malware propagation problem in general.

As shown in Fig. 2, we study the malware propagation issue at two levels, the Internet level and the network level. We note that at the network level, a *network* could be defined in many different ways, it could be an ISP domain, a country network, the group of a specific mobile devices, and so on. At the Internet level, we treat every network of the network level as one element.

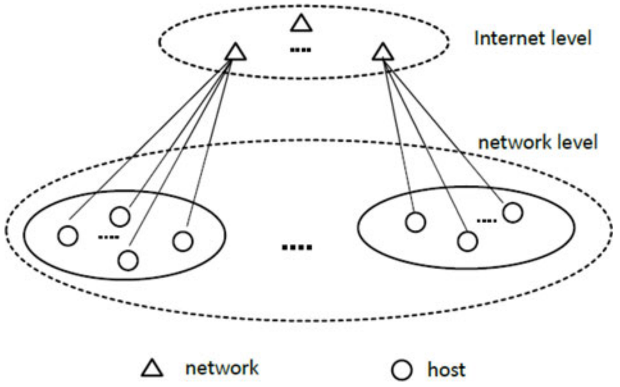


Fig. 2. The system architecture of the studied malware propagation.

At the Internet level, we suppose, there are M networks, each network is denoted as $L_i (1 \leq i \leq M)$. For any network L_i, we suppose it physically possesses $\mathbb{N}_i$ hosts. Moreover, we suppose the possibility of vulnerable hosts of L_i is denoted as $p_i (0 \leq p_i \leq 1)$. In general, it is highly possible that $\mathbb{N}_i \neq \mathbb{N}_j$, and $p_i \neq p_j$ for $i \neq j, 1 \leq i, j \leq M$. Moreover, due to differences in network topology, operating system, security investment and so on, the infection rates are different from network to network. We denote it as β_i for L_i. Similarly, it is highly possible that $\beta_i \neq \beta_j$ for $i \neq j, 1 \leq i, j \leq M$.

For any given network L_i with $p_i \cdot \mathbb{N}_i$ vulnerable hosts and infection rate β_i. We suppose the malware propagation starts at time 0. Based on Equation (4), we obtain the number of infected hosts, $I_i(t)$, of L_i at time t as follows:

$$\begin{aligned} I_i(t) &= (1+\alpha_i) I_i(t-1) - \beta_i (I_i(t-1))^2 \\ &= (1+\beta_i p_i \mathbb{N}_i) I_i(t-1) - \beta_i (I_i(t-1))^2. \end{aligned} \tag{12}$$

In this paper, we are interested in a global sense of malware propagation. We study the following question.

For a given time t since the outbreak of a malware, what are the characteristics of the number of compromised hosts for each network in the view of the whole Internet. In other words, to find a function F about $I_i(t) (1 \leq i \leq M)$. Namely, the pattern of

$$F(I_1(t), I_2(t), \ldots, I_M(t)). \tag{13}$$

For simplicity of presentation, we use $S(L_i, t)$ to replace $I_i(t)$ at the network level, and $I(t)$ is dedicated for the Internet level. Following Equation (13), for any network $L_i (1 \leq i \leq M)$, we have

$$S(L_i, t) = (1+\beta_i p_i \mathbb{N}_i) S(L_i, t-1) - \beta_i (S(L_i, t-1))^2. \tag{14}$$

At the Internet level, we suppose there are $k_1, k_2, \ldots, k_t$ networks that have been compromised at each round for each time unit from 1 to t. Any $k_i (1 \leq i \leq t)$ is decided by Equation (4) as follows:

$$k_i = (1+\beta_n M) I(i-1) - \beta_n (I(i-1))^2, \tag{15}$$

where M is the total number of networks over the Internet, and β_n is the infection rate among networks. Moreover, suppose the number of seeds, k_0, is known.

At this time point t, the landscape of the compromised hosts in terms of networks is as follows.

$$\begin{aligned} &\underbrace{S(L^1_{k_1}, t), S(L^2_{k_1}, t), \ldots, S(L^{k_1}_{k_1}, t)}_{k_1} \\ &\underbrace{S(L^1_{k_2}, t-1), S(L^2_{k_2}, t-1), \ldots, S(L^{k_2}_{k_2}, t-1)}_{k_2} \\ &\ldots \\ &\underbrace{S(L^1_{k_t}, 1), S(L^2_{k_t}, 1), \ldots, S(L^{k_t}_{k_t}, 1)}_{k_t}, \end{aligned} \tag{16}$$

where $L^j_{k_i}$ represents the jth network that was compromised at round i. In other words, there are k_1 compromised networks, and each of them have progressed t time units; k_2 compromised networks, and each of them has progressed $t-1$ time units; and k_t compromised networks, and each of them have progressed 1 time unit.

It is natural to have the total number of compromised hosts at the Internet level as

$$\begin{aligned} I(t) = &\underbrace{S(L^1_{k_1}, t) + S(L^2_{k_1}, t) + \cdots + S(L^{k_1}_{k_1}, t)}_{k_1} \\ &+ \underbrace{S(L^1_{k_2}, t-1) + \cdots + S(L^{k_2}_{k_2}, t-1)}_{k_2} \\ &+ \cdots \\ &+ \underbrace{S(L^1_{k_t}, 1) + S(L^2_{k_t}, 1) + \cdots + S(L^{k_t}_{k_t}, 1)}_{k_t} \end{aligned} \tag{17}$$

Suppose $k_i (i = 1, 2, \ldots)$ follows one distribution with a probability distribution of p_n (n stands for number), and the size of a compromised network, $S(L_i, t)$, follows another probability distribution of p_s (s stands for size). Let p_I be the probability distribution of $I(t) (t = 0, 1, \ldots)$. Based on Equation (18), we find p_I is exactly the convolution of p_n and p_s.

$$p_I = p_n \circledast p_s, \tag{18}$$

where $\circledast$ is the convolution operation.

Our goal is to find a pattern of p_I of Equation (18).

5 Malware Propagation Modelling

As shown in Fig. 2, we abstract the M networks of the Internet into M basic elements in our model. As a result, any two large networks, L_i and L_j ($i \neq j$), are similar to each other at this level. Therefore, we can model the studied problem as a homogeneous system. Namely, all the M networks share the same vulnerability probability (denoted as p), and the same infection rate (denoted as β). A simple way to obtain these two parameters is to use the means:

$$\begin{cases} p &= \dfrac{1}{M} \displaystyle\sum_{i=1}^{M} p_i \\ \beta &= \dfrac{1}{M} \displaystyle\sum_{i=1}^{M} \beta_i. \end{cases} \tag{19}$$

For any network L_i, let N_i be the total number of vulnerable hosts, then we have

$$N_i = p \cdot \mathbb{N}_i, i = 1, 2, \dots, M, \tag{20}$$

where $\mathbb{N}_i$ is the total number of computers of network L_i.

As discussed in Section 3, we know that $\mathbb{N}_i (i = 1, 2, \dots, M)$ follows the power law. As p is a constant in Equation (20), then $N_i (i = 1, 2, \dots, M)$ follows the power law as well. Without loss of generality, let L_i represent the ith network in terms of total vulnerable hosts (N_i). Based on the Zipf distribution, if we randomly choose a network X, the probability that it is network L_j is

$$Pr\{X = L_j\} = p_z(j) = \frac{N_j}{\sum_{i=1}^{M} N_i} = \frac{C}{j^{\alpha}}. \tag{21}$$

Equation (21) shows clearly that a network with a larger number of vulnerable hosts has a higher probability to be compromised.

Following Equation (18), at time t, we have $k_1 + k_2 + \cdots + k_t$ networks that have been compromised. Combining with Equation (21), in general, we know the first round of recruitment takes the largest k_1 networks, and the second round takes the k_2 largest networks among the remaining networks, and so on. We therefore can simplify Equation (18) as

$$\begin{aligned} I(t) = & \sum_{j=1}^{k_1} S(N_j, t) p_z(j) \\ & + \sum_{j=1}^{k_2} S(N_{k_1+j}, t-1) p_z(k_1 + j) \\ & + \dots \\ & + \sum_{j=1}^{k_t} S(N_{k_1+\cdots+k_{t-1}+j}, 1) \\ & \quad \cdot p_z(k_1 + \cdots + k_{t-1} + j). \end{aligned} \tag{22}$$

From Equation (22), we know the total number of compromised hosts and their distribution in terms of networks for a given time point t.

6 Analysis on the Proposed Malware Propagation Model

In this section, we try to extract the pattern of $I(t)$ in terms of $S(L_i, t^{'})$, or p_I of Equation (18).

We make the following definitions before we progress for the analysis.

1) *Early stage*. An early stage of the breakout of a malware means only a small percentage of vulnerable hosts have been compromised, and the propagation follows exponential distributions.
2) *Final stage*. The final stage of the propagation of a malware means that all vulnerable hosts of a given network have been compromised.
3) *Late stage*. A late stage means the time interval between the early stage and the final stage.

We note that many researches are focused on the early stage, and we define the early stage to meet the pervasively accepted condition, we coin the other two terms for the convenience of our following discussion. Moreover, we set variable T_e as the time point that a malware's progress transfers from its early stage to late stage. In terms of mathematical expressions, we express the early, late and final stage as $0 \leq t < T_e, T_e \leq t < \infty$, and $t = \infty$, respectively.

Due to the complexity of Equation (22), it is difficult to obtain conclusions in a dynamic style. However, we are able to extract some conclusions under some special conditions.

Lemma 1. *If distributions $p(x)$ and $q(x)$ follow exponential distributions, then $p(x)q(x)$ follows an exponential distribution as well.*

Due to the space limitation, we skip the proof and refer interested readers to [29].

At the early stage of a malware breakout, we have advantages to obtain a clear conclusion.

Theorem 1. *For large scale networks, such as the Internet, at the early stage of a malware propagation, the malware distribution in terms of networks follows exponential distributions.*

Proof. At a time point of the early stage ($0 \leq t < T_e$) of a malware breakout, following Equation (6), we obtain the number of compromised networks as

$$I(t) = I(0)e^{\beta_n M t}. \tag{23}$$

It is clear that $I(t)$ follows an exponential distribution.

For any of the compromised networks, we suppose it has progressed $t^{'} (0 < t^{'} \leq t < T_e)$ time units, and its size is

$$S(L_i, t^{'}) = I_i(0)e^{\beta N_i t^{'}}. \tag{24}$$

Based on Equation (24), we find that the size of any compromised network follows an exponential distribution. As a result, all the sizes of compromised networks follow exponential distributions at the early stage.

Based on Lemma 1, we obtain that the malware distribution in terms of network follows exponential distributions at its early stage. □

Moreover, we can obtain concrete conclusion of the propagation of malware at the final stage.

Theorem 2. *For large scale networks, such as the Internet, at the final stage ($t = \infty$) of a malware propagation, the malware distribution in terms of networks follows the power law distribution.*

Proof. At the final stage, all vulnerable hosts have been compromised, namely,

$$S(L_i, \infty) = N_i, i = 1, 2, \dots, M.$$

Based on our previous discussion, we know $N_i (i = 1, 2, \dots, M)$ follows the power law. As a result, the theorem holds. □

Now, we move our study to the late stage of malware propagation.

Theorem 3. *For large scale networks, such as the Internet, at the late stage ($T_e \leq t < \infty$) of a malware breakout, the malware distribution include two parts: a dominant power law body and a short exponential tail.*

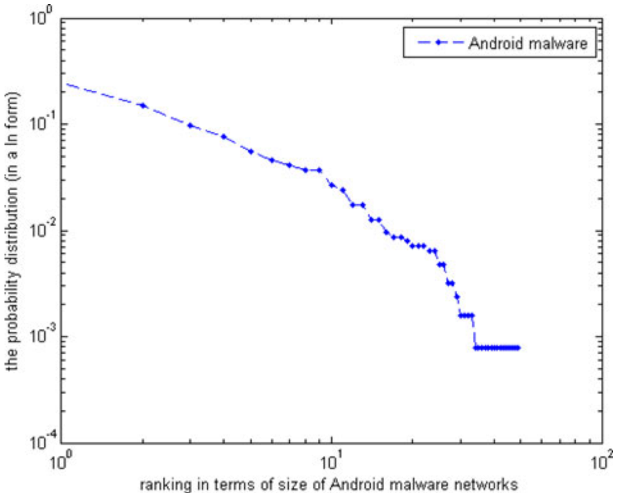


Fig. 3. The probability distribution of Android malware in terms of networks.

Proof. Suppose a malware propagation has progressed for $t(t >> T_e)$ time units. Let $t' = t - T_e$. If we separate all the compromised $I(t)$ hosts by time point t', we have two groups of compromised hosts.

Following Theorem 2, as $t' >> T_e$, the compromised hosts before t' follows the power law. At the same time, all the compromised networks after t' are still in their early stage. Therefore, these recently compromised networks follow exponential distributions.

Now, we need to prove that the networks compromised after time point t' are at the tail of the distribution. First of all, for a given network L_i, for $t_1 > t_2$, we have

$$S(L_i, t_1) \geq S(L_i, t_2). \tag{25}$$

For two networks, L_i and L_j, if $N_i \geq N_j$, then L_i should be compromised earlier than L_j. Combining this with (25), we know the later compromised networks usually lie at the tail of the distribution.

Due to the fact that $t' >> T_e$, the length of the exponential tail is much shorter than the length of the main body of the distribution. □

7 Performance Evaluation

In this section, we examine our theoretical analysis through two well-known large-scale malware: Android malware and Conficker. Android malware is a recent fast developing and dominant smartphone based malware [19]. Different from Android malware, the Conficker worm is an Internet based state-of-the-art botnet [20]. Both the data sets have been widely used by the community.

From the Android malware data set, we have an overview of the malware development from August 2010 to October 2011. There are 1,260 samples in total from 49 different Android malware in the data set. For a given Android malware program, it only focuses on one or a number of specific vulnerabilities. Therefore, all smartphones share these vulnerabilities form a specific network for that Android malware. In other words, there are 49 networks in the data set, and it is reasonable that the population of each network is huge. We sort the malware subclasses according to their size (number of samples in the data set), and present them in a loglog format in Fig. 3, the diagram is roughly a straight line. In other words, we can say that the Android malware distribution in terms of networks follows the power law.

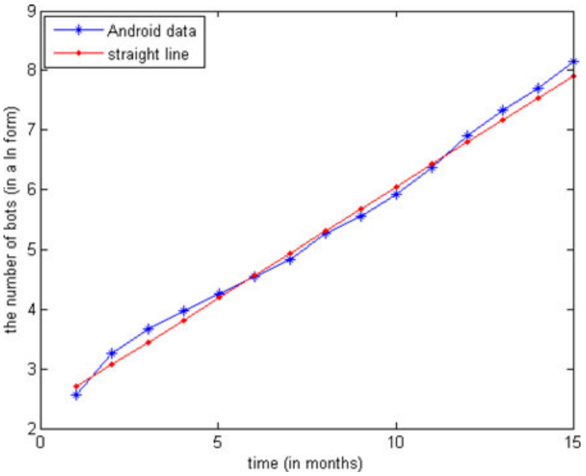


Fig. 4. The growth of total compromised hosts by Android malware against time from August 2010 to October 2011.

We now examine the growth pattern of total number of compromised hosts of Android malware against time, namely, the pattern of $I(t)$. We extract the data from the data set and present it in Table 2. We further transform the data into a graph as shown in Fig. 4. It shows that the member recruitment of Android malware follows an exponential distribution nicely during the 15 months time interval. We have to note that our experiments also indicate that this data does not fit the power law (we do not show them here due to space limitation).

In Fig. 4, we match a straight line to the real data through the least squares method. Based on the data, we can estimate that the number of seeds ($I(0)$) is 10, and $\alpha = 0.2349$. Following our previous discussion, we infer that the propagation of Android malware was in its early stage. It is reasonable as the size of each Android vulnerable network is huge and the infection rate is quite low (the infection is basically based on contacts).

We also collected a large data set of Conficker from various aspects. Due to the space limitation, we can only present a few of them here to examine our theoretical analysis.

First of all, we treat *AS* as networks in the Internet. In general, ASs are large scale elements of the Internet. A few key statistics from the data set are listed in Table 3. We

TABLE 2
The Number of Different Android Malware against Time (Months) in 2010-2011

Time point	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Variants	13	26	39	53	71	94	127	193	259	374	583	986	1,513	2,191	3,451

TABLE 3
Statistics for Conficker Distribution in Terms of ASs

Number of ASes	Largest botnet	Smallest botnet
1,0048	2,825,403	1

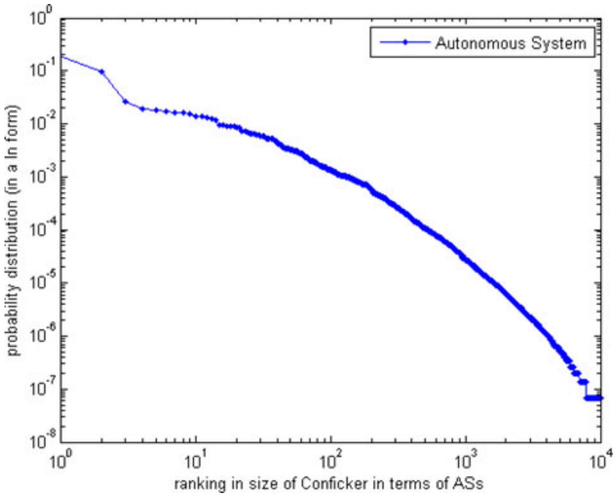


Fig. 5. Power law distribution of Conficker in terms of autonomous networks.

present the data in a loglog format in Fig. 5, which indicates that the distribution does follow the power law.

A unique feature of the power law is the scale free property. In order to examine this feature, we measure the compromised hosts in terms of domain names at three different domain levels: the top level, level 1, and level 2, respectively. Some statistics of this experiment are listed in Table 4.

Once again, we present the data in a loglog format in Figs. 6a, 6b and 6c, respectively. The diagrams show that the main body of the three scale measures are roughly straight lines. In other words, they all fall into power law distributions. We note that the flat head in Fig. 6 can be explained through a Zipf-Mandelbrot distribution. Therefore, Theorem 2 holds.

In order to examine whether the tails are exponential, we take the smallest six data from each tail of the three levels. It is reasonable to say that they are the networks compromised at the last 6 time units, the details are listed in Table 5 (we note that $t = 1$ is the sixth last time point, and $t = 6$ is the last time point).

TABLE 4
Statistics for Conficker Distribution in Terms of Domain Names at the Three Top Levels

	Number of botnets	Largest botnet	Smallest bo
top level	462	2,201,183	1
level 1	20,104	1,718,306	1
level 2	96,756	1,714,283	1

When we present the data of Table 5 into a graph as shown in Fig. 7, we find that they fit an exponential distribution very well, especially for the level 2 and level 3 domain name cases. This experiment confirms our claim in Theorem 3.

8 Further Discussion

In this paper, we have explored the problem of malware distribution in large-scale networks. There are many directions that could be further explored. We list some important ones as follows.

1) The dynamics of the late stage. We have found that the main body of malware distribution follows the power law with a short exponential tail at the late stage. It is very attractive to explore the mathematical mechanism of how the propagation leads to such kinds of mixed distributions.
2) The transition from exponential distribution to power law distribution. It is necessary to investigate when and how a malware distribution moves from an exponential distribution to the power law. In other words, how can we clearly define the transition point between the early stage and the late stage.
3) Multiple layer modelling. We hire the fluid model in both of the two layers in our study as both layers are sufficiently large and meet the conditions for the modelling methods. In order to improve the accuracy of malware propagation, we may extend our work to $n(n > 2)$ layers. In another scenario, we

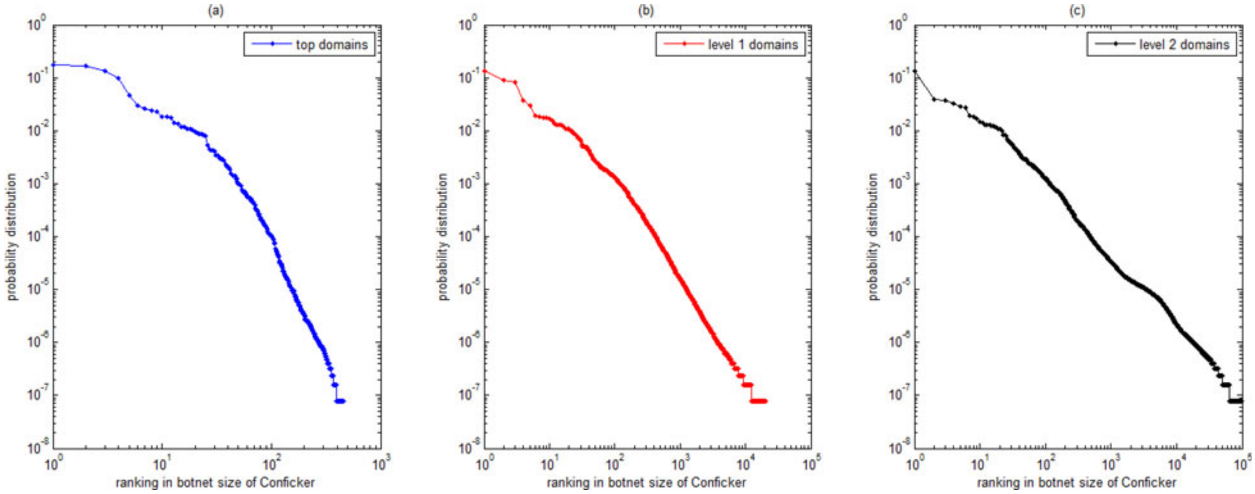


Fig. 6. Power law distribution of Conficker botnet in the top three levels of domain names.

TABLE 5
The Last Six Elements of Conficker Botnet from the Top Three Domain Name Levels

	t=1	t=2	t=3	t=4	t=5	t=6
top level	9	14	18	15	22	68
level 1	543	686	924	1,534	2,972	7,898
level 2	3,461	4,085	5,234	7,451	13,002	33,522

may expect to model a malware distribution for middle size networks, e.g., an ISP network with many subnetworks. In these cases, the conditions for the fluid model may not hold. Therefore, we need to seek suitable models to address the problem.

4) Epidemic model for the proposed two layer model. In this paper, we use the SI model, which is the simplest for epidemic analysis. More practical models, e.g., SIS or SIR, could be chosen to serve the same problem.

5) Distribution of coexist multiple malware in networks. In reality, multiple malware may coexist at the same networks. Due to the fact that different malware focus on different vulnerabilities, the distributions of different malware should not be the same. It is challenging and interesting to establish mathematical models for multiple malware distribution in terms of networks.

9 Summary and Future Work

In this paper, we thoroughly explore the problem of malware distribution at large-scale networks. The solution to this problem is desperately desired by cyber defenders as the network security community does not yet have solid answers. Different from previous modelling methods, we propose a two layer epidemic model: the upper layer focuses on networks of a large scale networks, for example, domains of the Internet; the lower layer focuses on the hosts of a given network. This two layer model improves the accuracy compared with the available single layer epidemic models in malware modelling. Moreover, the proposed two layer model offers us the distribution of malware in terms of the low layer networks.

We perform a restricted analysis based on the proposed model, and obtain three conclusions: The distribution for a given malware in terms of networks follows exponential distribution, power law distribution with a short exponential tail, and power law distribution, at its early, late, and final stage, respectively. In order to examine our theoretical findings, we have conducted extensive experiments based on two real-world large-scale malware, and the results confirm our theoretical claims.

In regards to future work, we will first further investigate the dynamics of the late stage. More details of the findings are expected to be further studied, such as the length of the exponential tail of a power law distribution at the late stage. Second, defenders may care more about their own network, e.g., the distribution of a given malware at their ISP domains, where the conditions for the two layer model may not hold. We need to seek appropriate models to address this problem. Finally, we are interested in studying the distribution of multiple malware on large-scale networks as we only focus on one malware in this paper. We believe it is not a simple linear relationship in the multiple malware case compared to the single malware one.

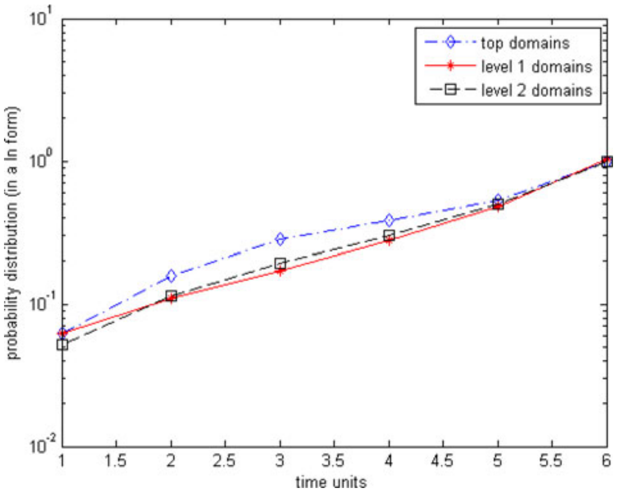


Fig. 7. The three tails from the three domain name levels fit exponential distributions.

Acknowledgments

Dr Yu's work is partially supported by the National Natural Science Foundation of China (grant No. 61379041), Prof. Stojmenovic's work is partially supported by NSERC Canada Discovery grant (grant No. 41801-2010), and KAU Distinguished Scientists Program.

References

[1] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, "Your botnet is my botnet: Analysis of a botnet takeover," in *Proc. ACM Conf. Comput. Commun. Security*, 2009, pp. 635–647.

[2] M. A. Rajab, J. Zarfoss, F. Monrose, and A. Terzis, "My botnet is bigger than yours (maybe, better than yours): Why size estimates remain challenging," in *Proc. 1st Conf. 1st Workshop Hot Topics Understanding Botnets*, 2007, p. 5.

[3] D. Dagon, C. Zou, and W. Lee, "Modeling botnet propagation using time zones," in *Proc. 13th Netw. Distrib. Syst. Security Symp.*, 2006.

[4] P. V. Mieghem, J. Omic, and R. Kooij, "Virus spread in networks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 1, pp. 1–14, Feb. 2009.

[5] Cabir. (2014). [Online]. Available: http://www.f-secure.com/en/web/labs_global/2004-threat-summary

[6] Ikee. (2014). [Online]. Available: http://www.f-secure.com/v-descs/worm_iphoneos_ikee_b.shtml

[7] Brador. (2014). [Online]. Available: http://www.f-secure.com/v-descs/brador.shtml

[8] S. Peng, S. Yu, and A. Yang, "Smartphone malware and its propagation modeling: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 925–941, 2014.

[9] Z. Chen and C. Ji, "An information-theoretic view of network-aware malware attacks," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 3, pp. 530–541, Sep. 2009.

[10] A. M. Jeffrey, X. Xia, and I. K. Craig, "When to initiate HIV therapy: A control theoretic approach," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 11, pp. 1213–1220, Nov. 2003.

[11] R. Dantu, J. W. Cangussu, and S. Patwardhan, "Fast worm containment using feedback control," *IEEE Trans. Dependable Secure Comput.*, vol. 4, no. 2, pp. 119–136, Apr.–Jun. 2007.

[12] S. H. Sellke, N.B. Shroff, and S. Bagchi, "Modeling and automated containment of worms," *IEEE Trans. Dependable Secure Comput.*, vol. 5, no. 2, pp. 71–86, Apr.–Jun. 2008.

[13] P. De, Y. Liu, and S. K. Das, "An epidemic theoretic framework for vulnerability analysis of broadcast protocols in wireless sensor networks," *IEEE Trans. Mobile Comput.*, vol. 8, no. 3, pp. 413–425, Mar. 2009.

[14] G. Yan and S. Eidenbenz, "Modeling propagation dynamics of bluetooth worms (extended version)," *IEEE Trans. Mobile Comput.*, vol. 8, no. 3, pp. 353–368, Mar. 2009.

[15] C. C. Zou, W. Gong, D. Towsley, and L. Gao, "The monitoring and early detection of internet worms," *IEEE/ACM Trans. Netw.*, vol. 13, no. 5, pp. 961–974, Oct. 2005.

[16] C. Gao and J. Liu, "Modeling and restraining mobile virus propagation," *IEEE Trans. Mobile Comput.*, vol. 12, no. 3, pp. 529–541, Mar. 2013.

[17] D. J. Daley and J. Gani, *Epidemic Modelling: An Introduction*. Cambridge, U.K. Cambridge Univ. Press, 1999.

[18] W. Willinger, D. Alderson, and J. C. Doyle, "Mathematics and the internet: A source of enormous confusion and great potential," *Notices Amer. Math. Soc.*, vol. 56, no. 5, pp. 586–599, 2009.

[19] Y. Zhou and X. Jiang, "Dissecting android malware: Characterization and evolution," in *Proc. IEEE Symp. Security Privacy*, 2012, pp. 95–109.

[20] S. Shin, G. Gu, A. L. N. Reddy, and C. P. Lee, "A large-scale empirical study of conficker," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 676–690, Apr. 2012.

[21] M. A. Rajab, J. Zarfoss, F. Monrose, and A. Terzis, "A multifaceted approach to understanding the botnet phenomenon," in *Proc. Internet Meas. Conf.*, 2006, pp. 41–52.

[22] A. J. Ganesh, L. Massoulié, and D. F. Towsley, "The effect of network topology on the spread of epidemics," in *Proc. IEEE Conf. Comput. Commun.*, 2005, pp. 1455–1466.

[23] J. Omic, A. Orda, and P. V. Mieghem, "Protecting against network infections: A game theoretic perspective," in *Proc. IEEE Conf. Comput. Commun.*, 2009, pp. 1485–1493.

[24] R. L. Axtell, "Zipf distribution of U.S. firm sizes," *Science*, vol. 293, pp. 1818–1820, 2001.

[25] M. Mitzenmacher, "A brief history of generative models for power law and lognornal distributions," *Internet Math.*, vol. 1, pp. 129–251, 2004.

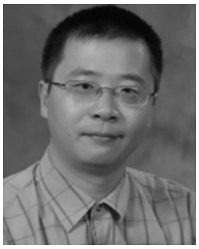
[26] M. Newman, *Networks: An Introduction*. London, U.K. Oxford Univ. Press, 2010.

[27] Z. K. Silagadze, "Citations and the Zipf-Mandelbrot's law," *Complex Syst.*, vol. 11, pp. 487–499, 1997.

[28] M. E. J. Newman, "Power laws, pareto distributions and Zipf's law," *Contemp. Phys.*, vol. 46, pp. 323–351, Dec. 2005.

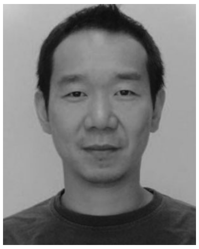
[29] L. Kleinrock, *Queueing Systems*, vol. I Theory, Hoboken, NJ, USA: Wiley-Interscience, 1975.

Shui Yu (M'05-SM'12) received the BEng and MEng degrees from the University of Electronic Science and Technology of China, Chengdu, P. R. China, in 1993 and 1999, respectively, and the PhD degree from Deakin University, Victoria, Australia, in 2004. He is currently a senior lecturer with the School of Information Technology, Deakin University, Victoria, Australia. He has published nearly 100 peer review papers, including top journals and top conferences, such as *IEEE TPDS, IEEE TIFS, IEEE TFS, IEEE TMC*, and IEEE INFOCOM. His research interests include networking theory, network security, and mathematical modeling. His actively servers his research communities in various roles, which include the editorial boards of the *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Communications Surveys and Tutorials*, and *IEEE Access*, IEEE INFOCOM TPC members 2012-2015, symposium co-chairs of IEEE ICC 2014, IEEE ICNC 2013-2015, and many different roles of international conference organizing committees. He is a senior member of the IEEE, and a member of the AAAS.



Guofei Gu (S'06-M'08) received the PhD degree in computer science from the College of Computing, Georgia Institute of Technology. He is an assistant professor in the Department of Computer Science and Engineering, Texas A&M University (TAMU), College Station, TX. His research interests are in network and system security, such as malware analysis, detection, defense, intrusion and anomaly detection, and web and social networking security. He is currently directing the Secure Communication and Computer Systems (SUCCESS) Laboratory at TAMU. He received the 2010 National Science Foundation (NSF) Career Award and a corecipient of the 2010 IEEE Symposium on Security and Privacy (Oakland 10) Best Student Paper Award. He is a member of the IEEE.

Ahmed Barnawi received the PhD degree from the University of Bradford, United Kingdom in 2006. He is an associate professor at the Faculty of Computing and IT, King Abdulaziz University, Jeddah, Saudi Arabia, where he works since 2007. He was visiting professor at the University of Calgary in 2009. His research areas are cellular and mobile communications, mobile ad hoc and sensor networks, cognitive radio networks and security. He received three strategic research grants and registered two patents in the US. He is a member of the IEEE.



Song Guo (M'02-SM'11) received the PhD degree in computer science from the University of Ottawa, Canada in 2005. He is currently a senior associate professor at the School of Computer Science and Engineering, the University of Aizu, Japan. His research interests are mainly in the areas of protocol design and performance analysis for reliable, energy-efficient, and cost effective communications in wireless networks. He is an associate editor of the *IEEE Transactions on Parallel and Distributed Systems* and an editor of *Wireless Communications and Mobile Computing*. He is a senior member of the IEEE and the ACM.

Ivan Stojmenovic was editor-in-chief of the *IEEE Transactions on Parallel and Distributed Systems* (2010-3), and is founder of three journals. He is editor of the *IEEE Transactions on Computers, IEEE Network, IEEE Transactions on Cloud Computing*, and *ACM Wireless Networks* and steering committee member of the *IEEE Transactions on Emergent Topics in Computing*. He is on Thomson Reuters list of Highly Cited Researchers from 2013, has top h-index in Canada for mathematics and statistics, and has more than 15,000 citations. He received five Best Paper Awards. He is a fellow of the IEEE, Canadian Academy of Engineering and Academia Europaea. He has received the Humboldt Research Award.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.