Characterising heterogeneity in vulnerable hosts on worm propagation

Zesheng Chen*

Department of Computer Science,
Indiana University – Purdue University Fort Wayne,
Fort Wayne, IN 46805, USA
Email: chenz@ipfw.edu
*Corresponding author

Chao Chen

Department of Electrical and Computer Engineering,
Indiana University – Purdue University Fort Wayne,
Fort Wayne, IN 46805, USA
Email: chenc@ipfw.edu

Abstract: Worm propagation models are important for understanding worm dynamics and designing effective and efficient detection and defence systems. The existing models, however, ignore the heterogeneity in vulnerable hosts and assume that the worm-scanning rate is the same for all infected hosts. In this work, we analytically and empirically study the impact of heterogeneity of vulnerable hosts on worm propagation. Specifically, we first apply the Jensen's inequality to show that the heterogeneity in vulnerable hosts indeed hinders the speed of worm propagation. Next, we propose a novel model to predict and characterise worm dynamics among heterogeneous vulnerable hosts. Finally, applying the real trace from Center for Applied Internet Data Analysis (CAIDA) to simulate the propagation of a Witty worm in the Internet, we verify our analytical results and demonstrate that our proposed model can accurately predict the spread of worms among heterogeneous vulnerable hosts.

Keywords: heterogeneity; vulnerable hosts; worm propagation; witty worm; modelling; simulation.

Reference to this paper should be made as follows: Chen, Z. and Chen, C. (2016) 'Characterising heterogeneity in vulnerable hosts on worm propagation', *Int. J. Security and Networks*, Vol. 11, No. 4, pp.224–234.

Biographical notes: Zesheng Chen is an Assistant Professor with the Department of Computer Science, Indiana University – Purdue University Fort Wayne. He received his MS and PhD from Georgia Institute of Technology in 2005 and 2007, respectively. His current research interests include network security, cognitive radio networks, and performance evaluation of communication networks.

Chao Chen is an Associate Professor with the Department of Electrical and Computer Engineering, Indiana University – Purdue University Fort Wayne. She received her MS and PhD from Georgia Institute of Technology in 2003 and 2005, respectively. Her current research interests include wireless opportunistic networks, wireless ad hoc and sensor networks, cognitive radio networks, network security, modelling and performance evaluation of communication networks.

This paper is a revised and expanded version of a paper entitled 'Heterogeneity in vulnerable hosts slows down worm propagation' presented at *IEEE Global Communications Conference (GLOBECOM 2012)*, Anaheim, CA, 3–7 December, 2012.

1 Introduction

Worms infect vulnerable hosts and use them to compromise other vulnerable hosts. Such a self-propagation attack has been a significant threat to network security since 2001 (Burt et al., 2008; Chen et al., 2009; Faghani and Nguyen, 2013; Sun et al., 2008a, 2008b, 2009; Wen et al., 2013; Zhang et al., 2014; Yun et al., 2015). Internet worms, such as Code Red, Nimda, Slammer, Witty, and Storm, infected a large number of hosts and caused huge damages. In recent years, worms have also been a main tool used by botnets to recruit a certain number of compromised machines and collect the information of infected hosts (Chen et al., 2010;

Dainotti et al., 2015; Han et al., 2012; Li et al., 2011; Liu et al., 2009). Therefore, it is important and imperative to accurately model the spread of worms in the internet.

Worm propagation models can help better understand worm dynamic characteristics. More importantly, such models are fundamental for detecting and defending against internet worms. Mathematical models of worm spreading have been widely studied. For example, differential equations have been used to describe random-scanning worms (Staniford et al., 2002; Vojnovic and Ganesh, 2008; Vojnovic et al., 2010; Zou et al., 2006) and to design a worm detection system (Zou et al., 2005). A discrete-time model has been proposed with the consideration of host recovery and patch, and has been exploited to monitor, detect, and defend against worms (Chen et al., 2003). A stochastic model has been studied to reflect the variation of worm propagation and its impact to worm detection (Nicol, 2006). All existing models, however, assume that vulnerable hosts are homogeneous and as a result, that all infected hosts use the same scanning rate to search for targets. Two related works (Kirmani and Hood, 2010; Zou et al., 2002) consider that the scanning rate of infected hosts can vary with time. But these two works also make the assumption that the worm-scanning rate is the same for all infected hosts. Therefore, the impact of heterogeneity in vulnerable hosts on worm propagation has not been studied yet.

Vulnerable hosts in the internet have been shown to be significantly heterogeneous. The network conditions and the computer performance of end-hosts are *very* different. For example, it has been shown that 70% of the end-hosts in a popular BitTorrent system have an upload capacity between 350 Kbps and 1 Mbps, whereas 10% of them have an upload capacity of 10 Mbps or more (Isdal et al., 2007). Moreover, 64% of the available resources are contributed by only 5% of hosts that have the bandwidth between 55 Mbps and 110 Mbps. A measurement study of the Witty worm also indicates strong heterogeneity in vulnerable hosts (Shannon and Moore, 2004a). For instance, the bit rates of infected hosts span from less than 56 Kbps to more than 100 Mbps. Hence, when studying worm propagation models, we cannot ignore the effect of the heterogeneity in vulnerable hosts.

The goal of this work is to study the impact of heterogeneity in vulnerable hosts on worm propagation (Chen and Chen, 2012). Specifically, we attempt to answer the following questions:

- Does heterogeneity in vulnerable hosts slow down worm propagation?
- If vulnerable hosts have a higher degree of heterogeneity, would this have a greater impact on worm spreading?
- How can we effectively predict and model worm propagation among heterogeneous vulnerable hosts?

To answer these questions, we analytically and empirically study the worm propagation among both homogeneous and heterogeneous vulnerable hosts. Our analysis is based on the probabilistic model, and the inequality and approximation techniques; whereas the simulation uses the scale-down method and mimics the spread of the Witty worm in the internet based on the real trace of scanning rates from CAIDA (Shannon and Moore, 2004b). Specifically, we summarise our discoveries and contributions in the following:

- Through both analysis and simulation, we find that statistically the worm has a smaller spreading speed among heterogeneous vulnerable hosts with distinct scanning rates than among homogeneous vulnerable hosts with the same scanning rate. For instance, we demonstrate that a Witty-like worm can be slowed down almost three times on average in the heterogeneous case than in the homogeneous case. Therefore, heterogeneity in vulnerable hosts can potentially slow down worm spreading significantly.
- We show analytically and conjecture that if the degree of heterogeneity in vulnerable hosts is higher, the worm propagates slower. Our simulation results verify the conjecture. This indicates that the current high degree of heterogeneity among vulnerable hosts in the internet indeed helps defenders to gain some time to respond to worm attacks.
- We then design a novel model to predict the spread of worms among heterogeneous vulnerable hosts. Such a model characterises the worm propagation delay, i.e., the time difference between the homogeneous case and the heterogeneous case. Simulation results show that our model can accurately predict the dynamics of worm propagation among heterogeneous vulnerable hosts.

The remainder of this paper is structured as follows. Section 2 discusses the heterogeneity in vulnerable hosts. Section 3 gives our analysis on worm propagation among heterogeneous vulnerable hosts, whereas Section 4 uses simulations to verify our analytical results. Finally, Section 5 concludes this paper.

2 Heterogeneity in vulnerable hosts

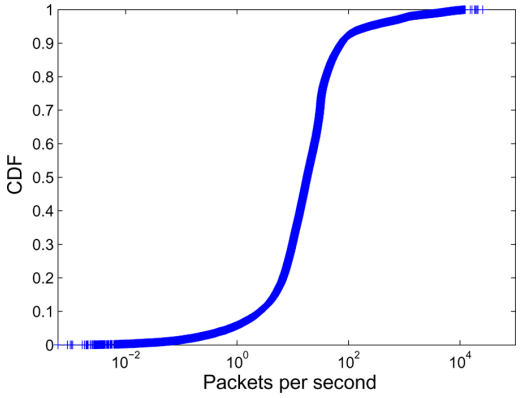
Vulnerable hosts in the internet are heterogeneous. This lies in the fact that end-hosts in the internet have distinct bandwidth and computer performance. A host may connect to the internet through a dial-up connection (e.g., 56 Kbps), a digital subscriber line (DSL) (e.g., 4 Kbps ~ 50 Kbps), a local area network (LAN) (e.g., 10 Mbps, 100 Mbps, or 1 Gbps), or a wireless LAN (e.g., 54 Mbps) (Kurose and Ross, 2008; Shin et al., 2012).

Moreover, many worms such as Slammer and Witty are bandwidth limited and send packets as fast as the infected hosts' internet connection allows (Moore et al., 2003; Shannon and Moore, 2004a). A measurement study on the Witty worm has shown that the infected hosts are heterogeneous (Shannon and Moore, 2004a). Specifically, while the average transmission speed of an infected host is

3 Mbps, 61% of infected hosts transmit with bit rates between 96 Kbps and 512 Kbps. We obtain the real trace of the Witty worm from CAIDA (Shannon and Moore, 2004b) and show how the cumulative distribution function of infected hosts by the Witty worm varies with packages-per-second (PPS) in Figure 1. Here PPS is the number of packages sent by an infected host per second and is indeed the scanning rate. Note that in the figure x-axis uses log scale. It can be seen that the PPS varies from 0.0006 to 25,620.1819. The mean of PPS is 135.9872, whereas the standard deviation is 795.4414. Thus, the scanning rate of the Witty worm varies significantly among infected hosts.

For an individual infected host, the bandwidth mainly determines how many scans per unit time a bandwidth-limited worm can send to find targets, i.e., the worm-scanning rate. If an infected host has a higher bandwidth, the worm-scanning rate is always higher. In this work, therefore, we use the variation of worm-scanning rates to reflect the heterogeneity in vulnerable hosts.

Figure 1 Cumulative distribution function (CDF) of infected hosts by the Witty worm varies with packets per second (PPS) (see online version for colours)



3 Theoretical analysis

Since vulnerable hosts have distinct bandwidth and computer performance, worm-scanning rates from infected hosts can be very different. In this paper, we specifically focus on the impact of the variation of scanning rates on worm propagation and make several simplified assumptions. First, we assume that once a host is infected, it remains in the infection state. Such a susceptible → infected (SI) model has been widely used in studying worm spreading (Chen and Ji, 2009; Nicol, 2006; Staniford et al., 2002; Vojnovic et al., 2010; Zou et al., 2006). Second, we focus on random-scanning worms. Random scanning selects target IPv4 addresses uniformly and has been exploited by many worms such as Code Red (Moore et al., 2002), Slammer (Moore et al., 2003), and Witty (Shannon and Moore, 2004a). The observations found in this paper, however, can be well extended to other scanning methods such as localised scanning (Staniford et al., 2002), importance scanning (Chen and Ji, 2007), and divide-conquer scanning (Chen et al., 2010). Finally, while the scanning rates of infected hosts can be different from each other, we assume that the scanning rate of an individual host does not vary with time. This is a reasonable assumption for two reasons:

- As indicated by our analysis, the time period of worm propagation that we are interested in is at the early stage, i.e., before the worm has infected many hosts and congested networks.
- It has been observed that an infected host always scans for vulnerable hosts at the maximum speed allowed by its network conditions and computing resources (Wei and Mirkovic, 2008).

In this section, we first show theoretically that compared with worm propagation among homogeneous vulnerable hosts, worm spreading is slowed down among heterogeneous vulnerable hosts. We then demonstrate and conjecture that if the degree of heterogeneity in vulnerable hosts is higher, worms spread slower. Finally, we provide a novel worm model that characterises the spread of worms among heterogeneous vulnerable hosts. The notations used in this paper are summarised in Table 1.

Table 1 Notations used in this paper

Notations	*Definition or explanation*
N	Total number of vulnerable hosts
Ω	Address space that a worm scans, i.e., $\Omega = 2^{32}$ for IPv4 address space
I	Number of infected hosts
T	Time to recruit a new victim
S_i	Scanning rate of infected host i
S	Mean of scanning rates
σ^2	Variance of scanning rates
D_H	Delay caused by the variation of scanning rates at the early stage of worm propagation
I_0	Upper bound of the number of infected hosts for calculating D_H

3.1 *Comparing worm propagation with homogeneous vulnerable hosts and with heterogeneous vulnerable hosts*

We use a discrete-time system to analyse the effect of the variation of scanning rates on worm propagation. Specifically, it is assumed that there are totally N vulnerable hosts among Ω scanning space. For IPv4, $\Omega = 2^{32}$, and thus $N << \Omega$. It is also assumed that there are currently I infected hosts, and infected host i ($i = 1, 2, \ldots, I$) uses a scanning rate of s_i, i.e., sends s_i scans per unit time. Then, the total number of scans at the next time step is $\sum_{i=1}^{I} s_i$. Therefore, the probability that an uninfected vulnerable host is hit by a specific worm scan at the next time step is $(N - I)/\Omega$, and the probability that an uninfected vulnerable host is hit by at least one worm scan at the next time step is

$$ph = 1 - \left(1 - \frac{N-I}{\Omega}\right)^{\sum_{i=1}^{I} s_i}. \quad (1)$$

Since $\frac{N-I}{\Omega} << 1$, we have

$$ph \approx \frac{N-I}{\Omega} \cdot \sum_{i=1}^{I} s_i. \quad (2)$$

Thus, the time to recruit a new victim, T, follows the geometric distribution, i.e.,

$$\Pr(T = k) = ph(1-ph)^{k-1}, \quad k = 1, 2, 3, \ldots \quad (3)$$

which leads to

$$E[T \mid s_1, s_2, \ldots s_I] = \frac{1}{ph} = \frac{\Omega}{(N-I)\sum_{i=1}^{I} s_i}. \quad (4)$$

It can be seen that if $E[T]$ is smaller, the worm spreads faster.

If all infected hosts are homogeneous, $s_i = s$, $\forall i$, i.e., the scanning rate for all infected hosts is a constant. Thus, the expected time to recruit a new victim is

$$E[T] = \frac{\Omega}{sI(N-I)}. \quad (5)$$

On the other hand, if infected hosts are heterogeneous, the scanning rate can be very different for distinct infected hosts. Because of the nature of random scanning, each instant of worm propagation can infect vulnerable hosts in totally different orders. Hence, we assume that s_i's are independent and identically-distributed (i.i.d.) random variables with mean s and variance σ^2 ($\sigma^2 \geq 0$). Note that if $\sigma^2 = 0$, vulnerable hosts are homogeneous; otherwise, they are heterogeneous. Therefore, from the law of total expectation, we have

$$E[T] = E[E[T \mid s_1, s_2, \ldots, s_I]] = \frac{\Omega}{N-I} E\left[\frac{1}{\sum_{i=1}^{I} s_i}\right]. \quad (6)$$

According to the Jensen's inequality (Ross, 1996; Cover and Thomas, 1991), if X is a random variable, f is a strictly convex function (i.e., $f''(x) > 0$), and $E[X]$ and $E[f(X)]$ exist, then

$$E[f(X)] \geq f(E[X]), \quad (7)$$

where the equality holds if and only if X is a constant. Next, we apply the Jensen's inequality by setting $f(x) = 1/x$. Since $f'(x) = -1/x^2$ and $f''(x) = 2/x^3 > 0$ when $x > 0$, $1/x$ is a strictly convex function. We then find from equation (6) that

$$E[T] \geq \frac{\Omega}{N-I} \cdot \frac{1}{E[\sum_{i=1}^{I} s_i]} = \frac{\Omega}{sI(N-I)}, \quad (8)$$

where the equality holds if and only if $\sigma^2 = 0$.

Comparing equation (5) and Inequality (8), we have the following theorem.

Theorem 1: *If worm-scanning rates s_i's are i.i.d. random variables with mean s and variance σ^2, then the worm spreads slower when $\sigma^2 > 0$ than when $\sigma^2 = 0$. That is, statistically the worm has a smaller spreading speed among heterogeneous vulnerable hosts with distinct scanning rates than among homogeneous vulnerable hosts with the same scanning rate.*

Theorem 1 indicates that the existing worm propagation models ignore the variation of scanning rates and thus overestimate the worm propagation speed. Moreover, Theorem 1 reflects that the heterogeneity in vulnerable hosts indeed hinders worm propagation and can help defenders gain some time to respond to worm attacks.

Here, we use a very simple example to provide an intuitive explanation to Theorem 1. Assume that a worm has 50% vulnerable hosts using a scanning rate of 1/s and other 50% vulnerable hosts with a scanning rate of 3/s. The average scanning rate is 2/s. When the worm starts from an infected host (i.e., $I = 1$), the expected time to recruit a new victim is

$$E[T] = \frac{\Omega}{N-I}\left(\frac{1}{2} \cdot \frac{1}{1} + \frac{1}{2} \cdot \frac{1}{3}\right) \quad (9)$$

$$E[T] = \frac{4}{3} \cdot \frac{\Omega}{(N-I) \cdot 2} \quad (10)$$

$$E[T] > \frac{\Omega}{(N-I) \cdot 2}. \quad (11)$$

Thus, the worm among heterogeneous vulnerable hosts takes 4/3 times of time to recruit a new victim as that among homogenous vulnerable hosts.

3.2 Conjecturing the impact of the degree of heterogeneity in vulnerable hosts on worm propagation

Since the heterogeneity in vulnerable hosts slows down worm propagation, a question arises: Would the worm spread slower if the degree of the heterogeneity of vulnerable hosts is higher? That is, when σ^2 increases, would $E[T]$ be larger? To answer this question, we apply Taylor expansion and approximation techniques. Specifically, we study the Taylor expansion of function $f(x) = 1/x$, i.e.,

$$f(x) = \frac{1}{a} + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2 + H \quad (12)$$

$$f(x) \approx \frac{1}{a} - \frac{x-a}{a^2} + \frac{(x-a)^2}{a^3}. \quad (13)$$

In the above equation, H contains the higher-order terms and can be ignored. Note that $E[\sum_{i=1}^{I} s_i] = sI$. Then, setting $x = \sum_{i=1}^{I} s_i$ and $a = sI$ in the above equation, we have

$$\frac{1}{\sum_{i=1}^{I} s_i} \approx \frac{1}{sI} - \frac{\sum_{i=1}^{I} s_i - sI}{s^2 I^2} + \frac{(\sum_{i=1}^{I} s_i - sI)^2}{s^3 I^3}. \quad (14)$$

Taking the expectation on both sides of the above equation, we obtain

$$E\left[\frac{1}{\sum_{i=1}^{I} s_i}\right] \approx \frac{1}{sI} + \frac{E[(\sum_{i=1}^{I} s_i - sI)^2]}{s^3 I^3} \tag{15}$$

$$E\left[\frac{1}{\sum_{i=1}^{I} s_i}\right] = \frac{1}{sI} + \frac{Var[\sum_{i=1}^{I} s_i]}{s^3 I^3} \tag{16}$$

$$E\left[\frac{1}{\sum_{i=1}^{I} s_i}\right] = \frac{1}{sI} + \frac{\sigma^2}{s^3 I^2}. \tag{17}$$

Therefore, from equations (6) and (17), the expected time to recruit a new victim is

$$E[T] \approx \frac{\Omega}{sI(N-I)} + \frac{\Omega\sigma^2}{s^3 I^2 (N-I)}. \tag{18}$$

In the above equation, the first term (i.e., $\Omega/sI(N-I)$) is identical to $E[T]$ for the homogeneous case, and the second term is proportional to σ^2. Based on this approximation result, it is obvious that when σ^2 increases, $E[T]$ also increases. Hence, we have the following conjecture.

Conjecture 1: When σ^2 is larger, the worm spreads slower. That is, the worm propagates slower among the vulnerable hosts with a higher degree of heterogeneity.

From equation (18), moreover, we can see when the number of infected hosts (i.e., I) is large, the second term in the equation closes to zero, and the heterogeneous case is similar to the homogeneous case. However, on the other hand when the number of infected hosts is small, the second term can be large, which leads to slower worm propagation for the heterogeneous case. In other words, equation (18) indicates that the main difference between worm propagation with homogenous vulnerable hosts and with heterogeneous vulnerable hosts lies at the early stage of worm spread when the number of infected hosts is small.

3.3 Modelling worm propagation among heterogeneous vulnerable hosts

We apply a novel approach to characterise the spread of random-scanning worms among heterogeneous vulnerable hosts. Instead of obtaining the propagation speed of worms, we attempt to study how much worm propagation delay, compared with the homogeneous case, is caused by the variation of worm-scanning rates. In this way, once we simulate or model the worm spreading among homogeneous vulnerable hosts, we can predict or model the worm propagation among heterogeneous vulnerable hosts.

We first use two worm-scanning rates as an example to demonstrate our modelling procedure. We assume that among N vulnerable hosts, $p{\cdot}N$ hosts have a scanning rate of r_1, and $(1-p){\cdot}N$ hosts have a scanning rate of r_2, where $0 \le p \le 1$ and $r_1 \ne r_2$. That is, a randomly selected infected host has a scanning rate of r_1 with probability p and a scanning rate of r_2 with probability $1-p$. Thus, the average scanning rate is $s = pr_1 + (1-p)r_2$. That is, $p = \frac{r_2 - s}{r_2 - r_1}$. Note that p can be derived, given arbitrary values of r_1, r_2, and s. Moreover, among the I infected hosts, the number of hosts having the scanning rate of r_1 follows the binomial distribution $B(I, p)$. If k infected hosts have a scanning rate of r_1, then $\sum_{i=1}^{I} s_i = kr_1 + (I-k)r_2$. From equation (6), we then obtain

$$E\left[\frac{1}{\sum_{i=1}^{I} s_i}\right] = \sum_{k=0}^{I} \binom{I}{k} p^k (1-p)^{I-k} \frac{1}{kr_1 + (I-k)r_2}. \tag{19}$$

Therefore, based on the above equation and equation (5), we can calculate the time difference to recruit a new victim between the heterogeneous case and the homogeneous case, i.e.,

$$\Delta E[T_I] = \sum_{k=0}^{I} \binom{I}{k} \frac{\Omega p^k (1-p)^{I-k}}{[kr_1 + (I-k)r_2](N-I)} - \frac{\Omega}{sI(N-I)}. \tag{20}$$

According to the feature of the binomial distribution, when I is large, $kr_1 + (I-k)r_2$ approaches sI with a high probability, and thus $\Delta E[T_I]$ is very small and can be ignored. Therefore, we only need to calculate the time difference when I is not large (e.g., $I \le 1\%$ of the total number of vulnerable hosts). In other words, the worm propagation difference between the heterogeneous case and the homogeneous case only occurs at the early stage of worm spreading when the number of infected hosts is small. Statistically, once a worm has recruited a sufficient number of infected hosts, the heterogeneity in vulnerable hosts has little impact on the worm propagation. On the other hand, when a worm has just started spreading from one or a small number of infected hosts, the impact of the heterogeneity in vulnerable hosts on worm dynamics can be significant, which will be shown in the next section.

Specifically, if we assume that a worm starts spreading from one infected host and set I_0 as the upper bound for calculating $\Delta E[T_I]$ in equation (20), then

$$D_H = \sum_{i=1}^{I_0} \Delta E[T_i] \tag{21}$$

represents how much delay is caused by the variation of scanning rates at the early stage of worm propagation. That is, once we obtain the propagation curve for worms among homogeneous vulnerable hosts with the same average scanning rate, we can then shift the curve with the delay D_H to predict the worm spreading among heterogeneous vulnerable hosts.

Note that such a modelling procedure can be easily extended to the case of multiple worm-scanning rates or the case when worm-scanning rates follow an arbitrary distribution. For example, when a worm has multiple scanning rates (i.e., r_1, r_2,…, r_m), an infected host has a scanning rate of r_i with probability p_i, where m is the number of scanning rates and $\sum_{i=1}^{m} p_i = 1$. Let n_i ($n_i \ge 0$)

denote the number of infected hosts among I infected hosts that have the scanning rate of r_i, where $\sum_{i=1}^{m} n_i = I$. Then, n_i's have a multinomial distribution with parameters I and p_i's, and $\sum_{i=1}^{I} s_i = \sum_{i=1}^{m} n_i r_i$. Therefore, equation (19) becomes

$$E\left[\frac{1}{\sum_{i=1}^{I} s_i}\right] = \sum_{\sum ni=I} \frac{(I!)(\prod_{i=1}^{m} p_i^{n_i})}{(\prod_{i=1}^{m} n_i!)(\sum_{i=1}^{m} n_i r_i)}. \quad (22)$$

Moreover, if s_i's are i.i.d. random variables with probability distribution $f_S(s)$. Then,

$$E\left[\frac{1}{\sum_{i=1}^{I} s_i}\right] = \int \cdots \int \frac{\prod_{i=1}^{I} f_S(s_i)}{\sum_{i=1}^{I} s_i} \mathrm{d}s_1 \cdots \mathrm{d}s_I. \quad (23)$$

In a similar way, we can obtain $\Delta E[T_I]$ and D_H for the worm with multiple scanning rates or an arbitrary distribution of scanning rates, and use them to predict the worm propagation among heterogeneous vulnerable hosts.

4 Simulation verification

We verify the analytical results in the previous section by simulating the spread of a worm among vulnerable hosts with both homogeneous and heterogeneous scanning rates. In our simulations, both homogeneous and heterogeneous cases have the same average worm-scanning rate. Moreover, the target of each worm scan is created by a random number generator over the scanning space, so that each host is hit by the worm scan with an equal probability. Once an uninfected vulnerable host is hit by a worm scan, we record the infection time, i.e., when this vulnerable host is compromised. Based on this infection time, we can count the number of infected hosts at each time step and thus obtain the worm propagation curve. Furthermore, the worm starts spreading from one infected host (i.e., hitlist = 1), which is randomly selected from the vulnerable hosts.

To obtain the analytical results for worm propagation in the heterogeneous case, we first obtain the simulation results for worm spreading in the homogeneous case, and use equation (21) to calculate the delay (i.e., D_H) caused by the variation of scanning rates. We then shift the worm propagation curve from the homogeneous case with the delay D_H to predict worm spreading in the heterogeneous case.

We first study random-scanning worms with two scanning rates. That is, we assume that some infected hosts have a scanning rate of scan1, whereas others have a scanning rate of scan2. If scan1 = scan2, it is the homogeneous case; otherwise, it is the heterogeneous case. We then simulate the propagation of the Witty worm using the actual scanning rates obtained from CAIDA (Shannon and Moore, 2004b). Specifically, in this section we start with applying scale-down simulations to obtain the observations of worm propagation in a /16 network with two scanning rates. We then simulate the spread of worms in the IPv4 address space with two scanning rates. Finally, we study Witty-worm spread with multiple scanning rates and the effect of parameters such as the scanning rate of the initially infected host and I_0 in equation (21).

4.1 Scale-down simulations

A scale-down simulation studies worm propagation in a much smaller scanning space, instead of the IPv4 address space that contains 2^{32} IP addresses (Weaver et al., 2004). In such a way, the patterns of worm spreading can be obtained in a much shorter time through simulations. We apply the technique of scale-down simulations and simulate the spread of random-scanning worms in a /16 subnet. Specifically, we assume that the scanning space is 2^{16} (i.e., $\Omega = 65{,}536$), the number of vulnerable hosts is 5000 (i.e., $N = 5000$), and the average scanning rate is 10/s (i.e., $s = 10$/s).

Figure 2 shows the simulation results of worm propagation with four cases of two scanning rates:

- scan1 = scan2 = 10
- scan1 = 5 and scan2 = 15
- scan1 = 1 and scan2 = 19
- scan1 = 1 and scan2 = 91.

The curves in the figure are averages over 10,000 runs. It can be seen that compared with the worm in the homogeneous case (i.e., case (1)), worms spread slower in the heterogeneous cases (i.e., cases (2)–(4)), which verifies Theorem 1. Moreover, if the degree of the heterogeneity in vulnerable hosts is higher, the worm spreads slower, which confirms Conjecture 1. Specifically, the worm takes on average 28.1 s to infect all vulnerable hosts in case (1), whereas the worm uses 28.8, 36.0, and 55.0 s in cases (2), (3), and (4), respectively. Moreover, it can be seen from the figure that after the worm has infected a certain number of hosts (e.g., 1% of vulnerable hosts), the propagation curves for all four cases are identical, which verifies our observations from equation (20).

Figure 2 Impact of scanning-rate variation on worm propagation in scale-down simulations ($\Omega = 65{,}536$, $N = 5000$, $s = 10$/s, and hitlist = 1) (see online version for colours)

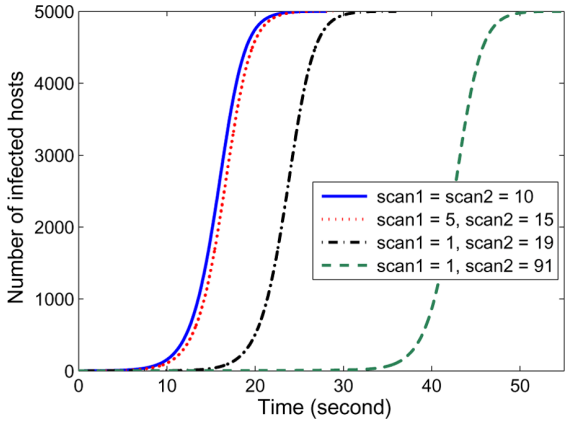
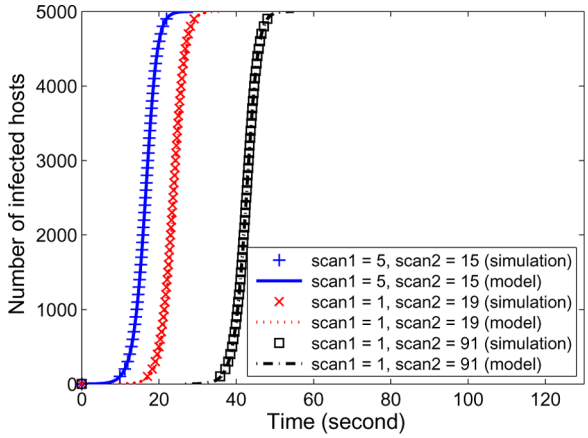


Figure 3 compares the simulation results to our analytical results for the heterogeneous cases. In equation (21), we set 50 as the upper bound (i.e., $I_0 = 50$) to calculate the delay (i.e., D_H). Specifically, we find that $D_H = 0.7$, 8.1, and 26.8 s for cases (2)–(4). From the figure, it can be seen that the curves of analytical results and simulation results overlap, indicating that our prediction is accurate.

Figure 3 Comparisons of worm propagation from scale-down simulations and from the model ($\Omega = 65{,}536$, $N = 5000$, $s = 10$/s, hitlist = 1, and $I_0 = 50$) (see online version for colours)



4.2 Two-scanning-rates worm propagation simulations

Next, we simulate the spread of a worm in the IPv4 address space with two scanning rates, using the parameters from (Shannon and Moore, 2004a). Specifically, the worm scans the entire IPv4 address space (i.e., $\Omega = 2^{32}$), targets 55,909 vulnerable hosts (i.e., $N = 55{,}909$), and uses an average scanning rate of 1200/s (i.e., $s = 1200$/s) (Shannon and Moore, 2004a). We consider three cases of two worm-scanning rates: (1) scan1 = scan2 = 1200; (2) scan1 = 200 and scan2 = 2200; (3) scan1 = 100 and scan2 = 10,000. For case (3), two scanning rates differ 100 times, which is motivated from the observation that the bandwidth capacity of end-hosts can have 100 times difference (Isdal et al., 2007). For each scenario, we simulate 100 runs with different seeds. Since the major difference among three cases occurs in the time period before the worm infect a significant portion of vulnerable hosts, our simulator stops running when the worm has compromised 30,000 hosts.

Figure 4 shows the spread of the worm with three different combinations of two scanning rates. In each sub-figure, the '5%' curve indicates that a worm spreads no faster than this curve in 5 out of 100 simulation runs. The similar definition is applied to the '25%', '50%', '75%', and '95%' curves. Moreover, the 'mean' curve is the average over 100 runs. It can be seen that the worm propagates faster in the homogeneous case than in the heterogeneous cases. Furthermore, when the degree of heterogeneity in vulnerable hosts increases, the worm spreads slower, and the variation of worm propagation is larger. These observations are similar to those in the scale-down simulations and verify our analysis. Specifically, comparing cases (1) and (3), we find that the worm uses on average 756.0 s to infect 30,000 hosts in the homogeneous case, whereas the worm needs 2212.9 s to compromise the same number of hosts in the heterogeneous case. This means that the worm is slowed down about three times due to the variation of scanning rates and indicates that the heterogeneity in vulnerable hosts can potentially impact worm spreading significantly.

We then further evaluate the performance of our prediction to worm propagation among heterogeneous vulnerable hosts in Figure 5. In our predication, we use only 10 as the upper bound in equation (21), i.e., $I_0 = 10$. In this figure, the curves of simulations are the averages over 100 runs, whereas the curves of the model are based on equations (20) and (21). It can also be seen that the curves of simulation and analytical results are very close, indicating that our model well characterises the dynamics of worm propagation among heterogeneous vulnerable hosts.

Figure 4 Impact of scanning-rate variation on worm propagation ($\Omega = 2^{32}$, $N = 55{,}909$, $s = 1200$/s, and hitlist = 1): (a) Case 1: scan1 = scan2 = 1200; (b) Case 2: scan1 = 200 and scan2 = 2200 and (c) Case 3: scan1 = 100 and scan2 = 10,000 (see online version for colours)

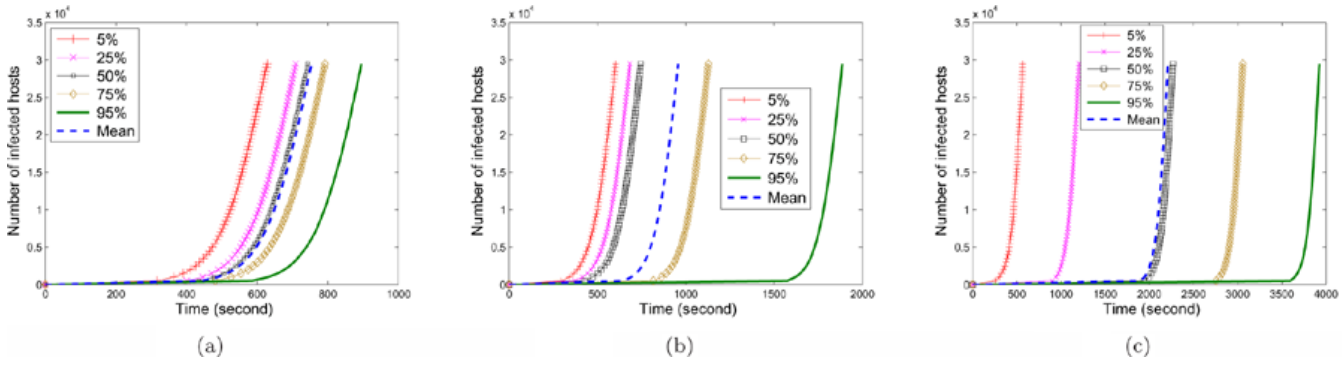
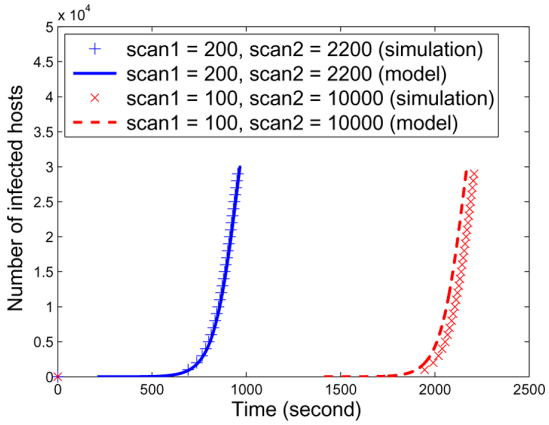


Figure 5 Comparisons of worm propagation from simulations and from the model ($\Omega = 2^{32}$, $N = 55{,}909$, $s = 1200$/s, hitlist = 1, and $I_0 = 10$) (see online version for colours)



4.3 Witty-worm propagation simulations

We further study the propagation of the Witty worm based on the real trace from CAIDA (Shannon and Moore, 2004b). Specifically, we apply the actual scanning rates of the Witty worm from Figure 1. To simplify our analysis, we divide the scanning rates of the Witty worm into four groups:

- scanning rate < 10
- 10 ≤ scanning rate < 100
- 100 ≤ scanning rate < 1000
- 1000 ≤ scanning rate.

Table 2 shows the percentage of infected hosts and the mean of scanning rates for each group.

We then simulate the spread of the Witty worm with four scanning rates as shown in Table 2 and total 55,909 potential targets (i.e., $N = 55{,}909$) over the entire IPv4 address space (i.e., $\Omega = 2^{32}$). To mimic the homogeneous case, we use the average of scanning rates, 135.8537 (i.e., $5 \times 0.3047 + 29 \times 0.6190 + 334 \times 0.0518 + 4044 \times 0.0245$), for all vulnerable hosts. For each case in our simulations, we simulate 200 runs with different seeds and stop simulator when 30,000 hosts have been infected.

Table 2 Summary of the four groups of scanning rates of the Witty worm

Range	*Percentage*	*Mean of scanning rates*
<10	0.3047	5
10~100	0.6190	29
10~1000	0.0518	334
≥1000	0.0245	4044

Figure 6 shows the spread of the Witty worm for both the homogeneous case and the heterogeneous case. Similar to Figure 4, it can be seen that the worm spreads much slower in the heterogeneous case than in the homogeneous case. Specifically, the worm needs on average 18,310 s to infect 30,000 hosts in the heterogeneous case, whereas it uses only 6586 s in the homogeneous case. This indicates that the worm is slowed down almost three times due to the heterogeneity in vulnerable hosts. Moreover, we notice that Figure 6(b) shows a relatively big gap between '75%' curve and '95%' curve. This reflects that in some runs, a worm can spread extremely slow if vulnerable hosts infected at the early stage of worm propagation are with low scanning rates (e.g., 5/s in this study).

Figure 6 Impact of scanning-rate variation on Witty-worm propagation ($\Omega = 2^{32}$, $N = 55{,}909$, and hitlist = 1): (a) homogenous case and (b) heterogeneous case (see online version for colours)

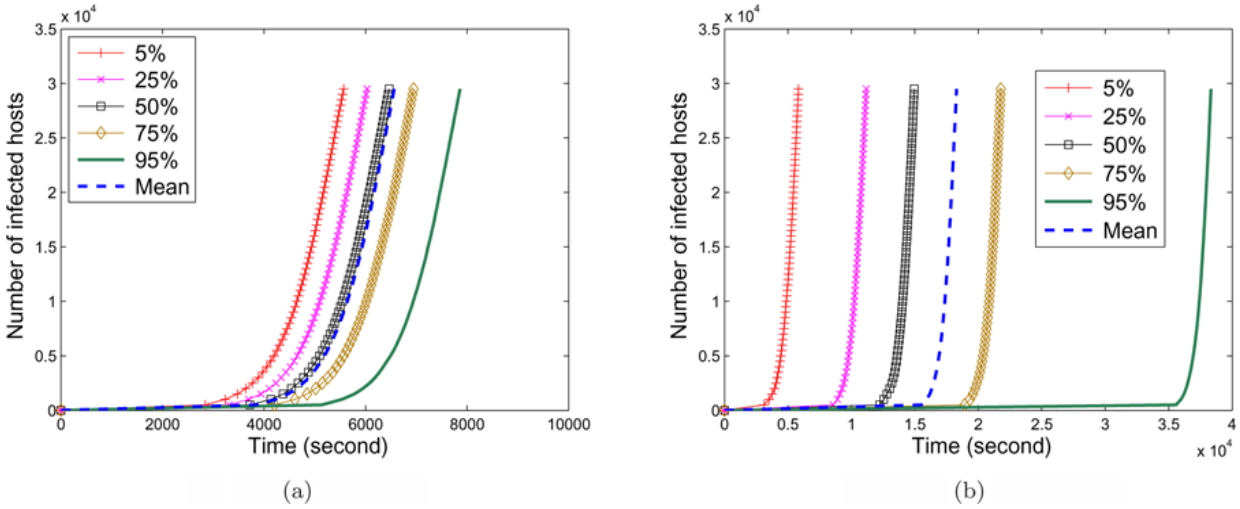


Figure 7 compares the simulation results to our analytical results for the heterogeneous case. In our analysis, we apply equation (22) for four scanning rates and use 20 as the upper bound in equation (21), i.e., $I_0 = 20$. From the figure, it can be seen that the curve of analytical results and simulation results overlap, indicating the performance of our predication is satisfactory.

Figure 7 Comparisons of Witty-worm propagation from simulations and from the model ($\Omega = 2^{32}$, $N = 55{,}909$, hitlist = 1, and $I_0 = 20$) (see online version for colours)

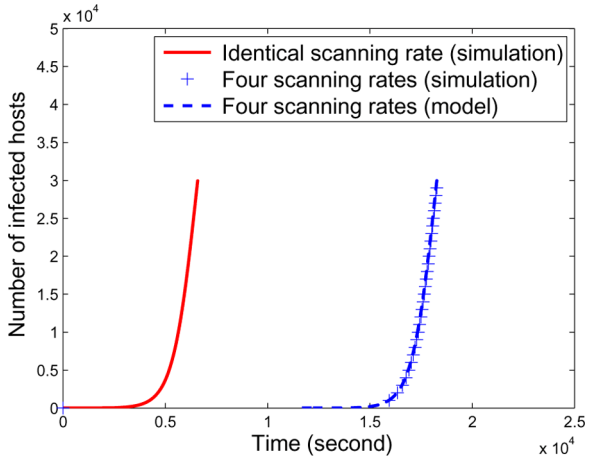
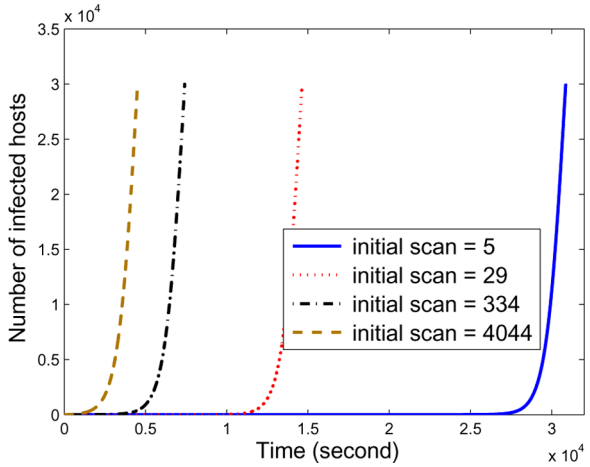
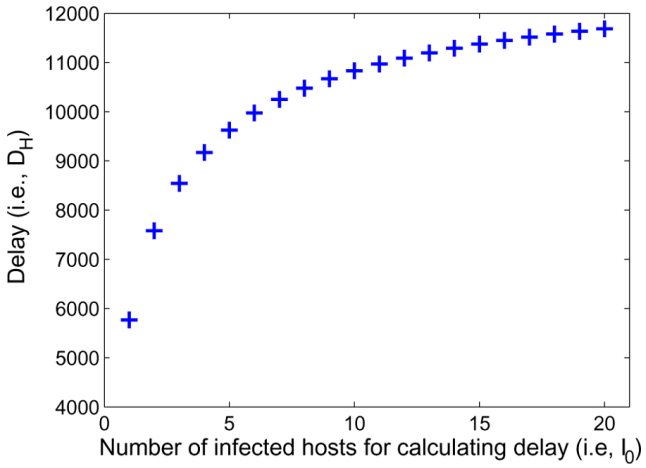


Figure 8 Effect of the scanning rate of the initially infected host (see online version for colours)



We then consider the effect of some parameters on worm propagation and our prediction. We first study the scanning rate of the initially infected host (i.e., hitlist). In Figure 8, we start a worm from a host with a fixed scanning rate and simulate each case with 100 runs. Each curve in the figure is an average over 100 runs. It can be seen that the worm spreads much slower from a host with a scanning rate of 5/s than from a host with a scanning rate of 4044/s. The figure shows that the scanning rate of the initially infected host can affect worm propagation significantly in the heterogeneous case. Next, we study the effect of I_0 in equation (21) on calculating the delay in Figure 9. It can be seen that when I_0 increases, D_H increases also. However, when I_0 is small, the increase rate of D_H is large; whereas when I_0 closes to 20, the increase rate of D_H becomes marginal. Therefore, using 20 as the upper bound seems reasonable in estimating the delay (i.e., D_H) caused by the variation of scanning rates.

Figure 9 Effect of the number of infected hosts used for calculating delay (i.e., I_0) (see online version for colours)



5 Conclusions

In this work, we have shown that heterogeneity in vulnerable hosts slows down worm propagation through both analysis and simulation. Moreover, a higher degree of heterogeneity in vulnerable hosts leads to slower propagation of worms. We have also designed a new model to characterise worm spreading among heterogeneous vulnerable hosts. Our model focuses on the worm propagation time difference between the heterogeneous case and the homogeneous case, and is shown empirically to have a good performance to predict worm dynamics. To the best of our knowledge, this is the first attempt in understanding the impact of the heterogeneity of vulnerable hosts on worm propagation quantitatively.

As our on-going work, we plan to extend the study to other scanning methods such as importance scanning (Chen and Ji, 2007) and divide-conquer scanning (Chen et al., 2010).

References

Burt, A.L., Darschewski, M., Ray, I., Thurimella, R. and Wu, H. (2008) 'Origins: an approach to trace fast spreading worms to their roots', *International Journal of Security and Networks*, Vol. 3, No. 1, pp.36–46.

Chen, C., Chen, Z. and Li, Y. (2010) 'Characterizing and defending against divide-conquer-scanning worms', *Computer Networks*, Vol. 54, No. 18, December, pp.3210–3222.

Chen, Z. and Chen, C. (2012) 'Heterogeneity in vulnerable hosts slows down worm propagation', *IEEE Global Communications Conference (GLOBECOM'12)*, Anaheim, CA, December, pp.923–928.

Chen, Z. and Ji, C. (2007) 'Optimal worm-scanning method using vulnerable-host distributions', *International Journal of Security and Networks: Special Issue on Computer and Network Security*, Vol. 2, Nos. 1–2, pp.71–80.

Chen, Z. and Ji, C. (2009) 'An information-theoretic view of network-aware malware attacks', *IEEE Transactions on Information Forensics and Security*, Vol. 4, No. 3, September, pp.530–541.

Chen, Z., Chen, C. and Li, Y. (2009) 'Deriving a closed-form expression for worm-scanning strategies', *International Journal of Security and Networks*, Vol. 4, No. 3, pp.135–144.

Chen, Z., Chen, C. and Wang, Q. (2010) 'On the scalability of delay-tolerant botnets', *International Journal of Security and Networks*, Vol. 5, No. 4, pp.248–258.

Chen, Z., Gao, L. and Kwiat, K. (2003) 'Modeling the spread of active worms', *22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'03)*, Vol. 3, San Francisco, CA, April, pp.1890–1900.

Cover, T.M. and Thomas, J.A. (1991) *Elements of Information Theory*, Wiley, New York.

Dainotti, A., King, A., Claffy, K., Papale, F. and Pescape, A. (2015) 'Analysis of a '/0' stealth scan from a botnet', *IEEE/ACM Transactions on Networking*, Vol. 23, No. 2, April, pp.341–354.

Faghani, M.R. and Nguyen, U.T. (2013) 'A study of XSS worm propagation and detection mechanisms in online social networks', *IEEE Transactions on Information Forensics and Security*, Vol. 8, No. 11, November, pp.1815–1826.

Han, F., Chen, Z., Xu, H., Wang, H. and Liang, Y. (2012) 'A collaborative botnets suppression system based on overlay network', *International Journal of Security and Networks*, Vol. 7, No. 4, pp.211–219.

Isdal, T., Piatek, M., Krishnamurthy, A. and Anderson, T. (2007) 'Leveraging BitTorrent for end host measurements', *Proc. of the 8th Passive and Active Measurement Conference (PAM '07)*, Louvain-la-neuve, April, Belgium.

Kirmani, E. and Hood, C.S. (2010) 'Analysis of a scanning model of worm propagation', *Journal in Computer Virology*, Vol. 6, No. 1, pp.31–42.

Kurose, J.F. and Ross, K.W. (2008) *Computer Networking: A Top-Down Approach*, 4th ed., Pearson Education, Inc., New York.

Li, Z., Goyal, A., Chen, Y. and Paxson, V. (2011) 'Towards situational awareness of large-scale botnet probing events', *IEEE Transactions on Information Forensics and Security*, Vol. 6, No. 1, March, pp.175–188.

Liu, J., Xiao, Y., Ghaboosi, K., Deng, H. and Zhang, J. (2009) 'Botnet: classification, attacks, detection, tracing, and preventive measures', *EURASIP Journal on Wireless Communications and Networking*, Vol. 2009, Article ID 692654, doi:10.1155/2009/692654.

Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S. and Weaver, N. (2003) 'Inside the slammer worm', *IEEE Security and Privacy*, Vol. 1, No. 4, July, pp.33–39.

Moore, D., Shannon, C. and Brown, J. (2002) 'Code-red: a case study on the spread and victims of an Internet worm', *ACM SIGCOMM/USENIX Internet Measurement Workshop*, Marseille, November, France.

Nicol, D.M. (2006) 'The impact of stochastic variance on worm propagation and detection', *Proc. ACM/CCS Workshop on Rapid Malcode (WORM'06)*, November, Fairfax, VA.

Ross, S.M. (1996) *Stochastic Processes*, 2nd ed., John Wiley & Sons, Inc., Hoboken, New Jersey.

Shannon, C. and Moore, D. (2004a) 'The spread of the Witty worm', *IEEE Security and Privacy*, Vol. 2, No. 4, July–August, pp.46–50.

Shannon, C. and Moore, D. (2004b) *The CAIDA Dataset on the Witty Worm*, 19–24 March, 2004, http://www.caida.org/passive/witty/. Support for the Witty Worm Dataset and the UCSD Network Telescope are provided by Cisco Systems, Limelight Networks, the US Department of Homeland Security, the National Science Foundation, and CAIDA, DARPA, Digital Envoy, and CAIDA Members.

Shin, S., Gu, G., Reddy, N. and Lee, C.P. (2012) 'A large-scale empirical study of conficker', *IEEE Transactions on Information Forensics and Security*, Vol. 7, No. 2, April, pp.676–690.

Staniford, S., Paxson, V. and Weaver, N. (2002) 'How to 0wn the Internet in your spare time', *Proc. of the 11th USENIX Security Symposium (Security'02)*, San Francisco, CA, August, pp.149–167.

Sun, B., Shrestha, D., Yan, G. and Xiao, Y. (2008a) 'Self-propagate mal-packets in wireless sensor networks: dynamics and defense implications', *Proc. of IEEE GLOBECOM (Globecom'08)*, November, New Orleans, LA.

Sun, B., Yan, G. and Xiao, Y. (2008b) 'Worm propagation dynamics in wireless sensor networks', *Proc. of IEEE ICC 2008 (ICC'08)*, May, Beijing, China, pp.1541–1545.

Sun, B., Yan, G., Xiao, Y. and Yang, T.A. (2009) 'Self-propagate Mal-packets in wireless sensor networks: dynamics and defense implications', *(Elsevier) Ad Hoc Networks*, Vol. 7, No. 8, November, pp.1489–1500.

Vojnovic, M. and Ganesh, A. (2008) 'On the race of worms, alerts, and patches', *IEEE/ACM Transactions on Networking*, Vol. 16, No. 5, October, pp.1066–1079.

Vojnovic, M., Gupta, V., Karagiannis, T. and Gkantsidis, C. (2010) 'Sampling strategies for epidemic-style information dissemination', *IEEE/ACM Transactions on Networking*, Vol. 18, No. 4, August, pp.1013–1025.

Weaver, N., Hamadeh, I., Kesidis, G. and Paxson, V. (2004) 'Preliminary results using scale-down to explore worm dynamics', *Proc. of the 2nd ACM Workshop on Rapid Malcode (WORM'04)*, October, Fairfax, VA.

Wei, S. and Mirkovic, J. (2008) 'Correcting congestion-based error in network telescopes observations of worm dynamics', *Proc. of the 8th Internet Measurement Conference (IMC'08)*, October, Vouliagmeni, Greece.

Wen, S., Zhou, W., Zhang, J., Xiang, Y., Zhou, W. and Jia, W. (2013) 'Modeling propagation dynamics of social network worms', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 24, No. 8, August, pp.1633–1643.

Yun, X., Li, S. and Zhang, Y. (2015) 'SMS worm propagation over contact social networks: modeling and validation', *IEEE Transactions on Information Forensics and Security*, Vol. 10, No. 11, September, pp.2365–2380.

Zhang, Y., Juels, A., Reiter, M.K. and Ristenpart, T. (2014) 'An epidemiological study of malware encounters in a large', *Proc. of the 21st ACM Conference on Computer and Communications Security (CCS'14)*, November, Scottsdale, AZ.

Zou, C.C., Gong, W. and Towsley, D. (2002) 'Code red worm propagation modeling and analysis', *Proc. of the 9th ACM Conference on Computer and Communication Security (CCS'02)*, Washington DC, November, pp.138–147.

Zou, C.C., Gong, W., Towsley, D. and Gao, L. (2005) 'The monitoring and early detection of Internet worms', *IEEE/ACM Transactions on Networking*, Vol. 13, No. 5, October, pp.961–974.

Zou, C.C., Towsley, D. and Gong, W. (2006) 'On the performance of Internet worm scanning strategies', *Elsevier Journal of Performance Evaluation*, Vol. 63, No. 7, July, pp.700–723.