Deriving a Closed-Form Expression for Worm-Scanning Strategies

Zesheng Chen*
Department of Electrical & Computer Engineering
Florida International University
Miami, FL 33174
E-mail: zchen@fiu.edu
*Corresponding author

Chao Chen
Department of Engineering
Indiana University - Purdue University Fort Wayne
Fort Wayne, IN 46805
E-mail: chen@engr.ipfw.edu

Yubin Li
Department of Electrical & Computer Engineering
Florida International University
Miami, FL 33174
E-mail: yli004@fiu.edu

Abstract: This work presents a closed-form expression for characterizing the spread of a class of worm-scanning strategies through a mean-field approximation. This expression can both accurately capture the worm propagation speed before the number of infections becomes large and explicitly demonstrate the effects of important parameters such as the vulnerable-host distribution and the worm-scanning strategy. Our approach is based on the mean-field theory that investigates the average number of infected hosts over time. Experiments are carried out based on the parameters chosen from Witty and Code Red worms. Experimental results verify that the closed-form expression can accurately reflect the mean value of infections over time before the infected hosts become saturated for a wide range of scanning methods including static worm-scanning strategies and self-learning worms. Therefore, our model can help defenders design better detection and defense systems and provide a stepping stone towards obtaining closed-form expressions for the propagation of more complex worm-scanning strategies.

Keywords: security; worm-scanning strategy; modeling; closed-form expression; mean-field theory.

1 Introduction

Worm attacks present a significant threat to the Internet. A worm can self-propagate across the Internet in a short time by exploiting security flaws on vulnerable hosts without human intervention. Thus, worms, such as Code Red, Slammer, and Witty, have infected hundreds of thousands of hosts and caused enormous damages. Most worms use a scanning technique that selects a target in an IP address space and then sends out a probe to attempt to compromise this target. Among all scanning methods, *random scanning* is the simplest method that selects a target at random in the IPv4 address space and has been widely used by real worms. Recent studies have shown, however, that worms can potentially apply more advanced scanning strategies, such as *hitlist scanning* [17], *routable scanning*

[21, 23], *importance scanning* [8, 7], and *OPT-STATIC* [19]. These advanced scanning strategies have been demonstrated to be able to spread a worm much faster than the random-scanning method. Therefore, it is imperative that defenders would model the spreading behaviors of advanced worm-scanning strategies accurately.

Vojnovic *et al.* point out that studying worm-scanning methods is also of interest in a wide variety of areas such as streaming broadcasting, database maintenance, and Web-service membership management [19]. These applications potentially exploit epidemic-style information dissemination techniques to spread information among participants quickly. Therefore, modeling epidemic-style information dissemination or worm-scanning strategies can provide further understandings to these areas.

Most advanced worm-scanning strategies take advantage of the non-uniform distribution of vulnerable hosts over groups. For example, *importance scanning* probes the Internet according to an underlying vulnerable-host distribution and forces worm scans on the most relevant parts of an address space [8]. For these advanced scanning strategies, the Internet is partitioned into groups according to such standards as the IP prefix, autonomous systems, and the first byte of IP addresses (/8 subnets). Since the distribution of vulnerable hosts over groups has been observed to be highly uneven [10, 16, 2, 9, 19], a worm would spend more scans on groups that contain many vulnerable hosts to speed up worm propagation. That is, a worm scans different groups with different likelihoods so that a group containing more vulnerable hosts would be hit by a worm scan with a higher probability. In this work, we focus on a class of worm-scanning strategies where the worm group scanning probabilities are independent of the number of infected hosts in groups, including random scanning, static importance scanning [7], OPT-STATIC [19], and self-learning worms [7].

Many approaches have been studied to model the spread of worms using different scanning strategies, including stochastic models [15, 12], deterministic models [17, 24, 5, 18], and optimization methods [19]. The prior work, however, cannot explicitly characterize both the worm propagation speed and the effect of important parameters (*e.g.,* the vulnerable-host distribution and worm-scanning strategies) [3], which are key factors to worm detection and defenses. One obvious solution to overcome the weakness of the prior work is to derive the closed-form expression for worm-scanning strategies. In general, however, it is nearly impossible to derive an exact closed-form expression as the result of the dynamic behavior of worm propagation.

The goal of this work is to characterize both the worm spreading speed and the parameters' effects by deriving a closed-form expression. Specifically, we obtain a closed-form expression from a deterministic dynamic equation through a *mean-field* approximation. As pointed out by [27, 13, 20, 1], the mean-field approach provides a way to gain insight into the behavior of complex systems at a relatively low cost. That is, the mean-field method focuses on the averages of the system, ignoring fluctuations. In this work, we neglect the fluctuation of the number of infected hosts and derive the average of the infections in each group. We further apply the Taylor expansion and focus on the first-order term. In this way, we transform a dynamic equation to a Riccati equation [28], which leads to a closed-form expression for the spread of worm-scanning strategies. Furthermore, based on this closed-form solution, we derive closed-form expressions for both static worm-scanning strategies and self-learning worms. It is shown that our closed-form solution explicitly demonstrates the effects of the vulnerable-host distribution and the worm-scanning strategies. To verify our closed-form expression, we simulate the spread of worms based on the parameters chosen from Witty and Code Red worms. We then compare our expression with the extension of the analytical active worm propagation (AAWP) model [5]. Experimental results show that the closed-form expression can characterize the spread of both static worm-scanning strategies and self-learning worms, before infected hosts become saturated (even beyond the early stage).

Characterizing the worm propagation speed before infections become large is a key element to worm detection and defenses [22]. If a worm can compromise a large number of hosts before it is detected, it is too late for defenders to slow down the worm. Therefore, it is critical that defenders would detect and fight against a worm before it has infected too many hosts. Thus, our closed-form expression provides an accurate picture for defenders to understand the average of the worm spreading speed in the time window of detection and defenses. Moreover, although in this work we assume that the worm group scanning probabilities do not adapt to the number of infected hosts in groups, our approach may provide a stepping stone towards finding closed-form expressions for the propagation of dynamic and adaptive strategies, which are more complex and difficult to obtain.

The remainder of this paper is structured as follows. Section 2 introduces the background of this work. Next, Section 3 derives a closed-form expression for the spread of worm-scanning strategies. Section 4 provides the discussions of our designed closed-form expression. Section 5 further verifies our expression through experiments. Section 6 describes the related work. Finally, Section 7 concludes this paper.

2 Background and Notations

In this section, we provide the background on worm-scanning strategies and deterministic worm-propagation

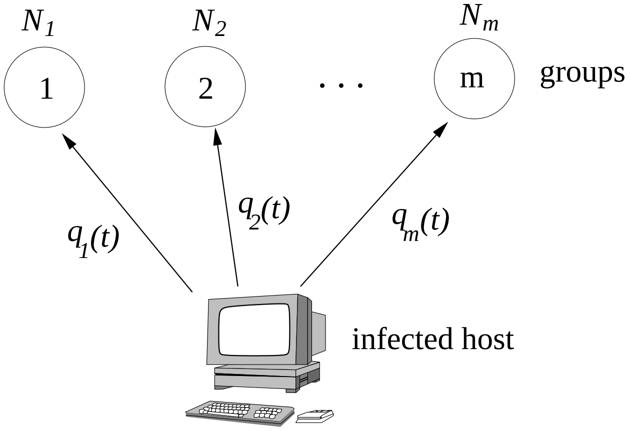


Figure 1: Illustration of worm-scanning strategies.

models, and introduce the notations used in this paper.

2.1 Worm-Scanning Strategies

A fundamental characteristic of Internet worms is self-propagation, *i.e.,* a worm can compromise vulnerable hosts and use them to attack other victims. Moreover, a key component of the epidemic-style attacks is scanning, *i.e.,* how a worm finds vulnerable hosts. In general, worm-scanning strategies can be abstracted and illustrated in Figure 1. Here, the Internet contains Ω IP addresses (*i.e.,* $\Omega = 2^{32}$) and totally N vulnerable hosts, and is partitioned into m groups. Group i ($i = 1, 2, \cdots, m$) contains Ω_i IP addresses and totally N_i vulnerable hosts, where $\sum_{i=1}^{m} \Omega_i = \Omega$ and $\sum_{i=1}^{m} N_i = N$. A worm scans group i with probability $q_i(t)$ at time t, where $\sum_{i=1}^{m} q_i(t) = 1$. Thus, $q_i(t)$'s form a *group scanning distribution* at time t. Once a worm scan hits a group, this scan would select the targets in the group uniformly. The notations used in this paper are summarized in Table 1.

Based on the characteristics of $q_i(t)$'s, the worm-scanning strategies can be classified into two categories: static worm-scanning strategies and dynamic worm-scanning strategies. If $q_i(t)$'s are fixed at all time, *i.e.,* $q_i(t) = q_i$, such strategies are called *static* worm-scanning strategies. Several widely studied strategies belong to this category. For example,

- Random scanning (RS): Such a scanning method chooses target IP addresses uniformly and has been used by Code Red, Slammer, and Witty worms. For RS,
$$q_i(t) = q_i = \frac{\Omega_i}{\Omega}, \quad i = 1, 2, \cdots, m. \qquad (1)$$
- Static importance scanning (STATIC-IS) [7]: Such a strategy exploits the underlying highly uneven distribution of vulnerable hosts, and q_i's depend on N_i's and Ω_i's. Basically, if N_i/Ω_i is large, q_i would be large also.
- OPT-STATIC [19]: Such a method minimizes the number of worm scans required to reach a predetermined faction of vulnerable hosts. This method specifies to scan a set A of initially densest groups. The necessary condition for a set A to be optimal is that the initial density of uninfected vulnerable hosts in every group in A must be larger than the final density of uninfected vulnerable hosts in A (see [19] for the expressions of q_i and A).
- SUBOPT-STATIC: Such a strategy is called *uniform random sampling of a subset of subnets A* in [19]. Here, A is obtained from OPT-STATIC. SUBOPT-STATIC only scans targets in groups A, but samples these targets uniformly.

If $q_i(t)$'s vary with time, such strategies are called *dynamic* worm-scanning strategies. Dynamic strategies can further be classified into two cases: $q_i(t)$'s are independent of the number of infected hosts in groups at time t (*i.e.,* $I_i(t)$'s), such as self-learning worms in [7]; and $q_i(t)$'s depend on $I_i(t)$'s, such as localized scanning [14, 4], dynamic important scanning [7], and K-FAIL [19]. For the self-learning worm in [7], the propagation process can be divided into two stages: the learning stage where the worm uses RS, *i.e.,* $q_i(t) = \Omega_i/\Omega$; and the importance-scanning stage where the worm uses STATIC-IS, *i.e.,* $q_i(t)$ switches from Ω_i/Ω to q_i that depends on estimated N_i/N and Ω_i.

In this paper, we focus on static worm-scanning strategies and dynamic worm-scanning strategies where the group scanning distribution is independent of the number of infected hosts in groups. That is, we assume that $q_i(t)$'s are independent of $I_i(t)$'s.

2.2 Deterministic Worm-Propagation Models

Most analytical models of worm propagation have used deterministic dynamic equations to characterize the spread of Internet worms. Specifically, many dynamic equations are designed to model the infection behavior of worms that is known as the susceptible → infected (SI) model [17, 24, 8]. The SI model assumes that each vulnerable host has two statuses: susceptible or infected. Once infected, the host remains infected.

There are two types of deterministic worm-propagation models. The first type is based on continuous time and differential equations, and is called the epidemic model. For example, assuming that there are $I(t)$ infected hosts at time t and an infected host sends out s scans per unit time, RS can be characterized by the following differential equation [24]

$$\frac{dI(t)}{dt} = sI(t)\frac{N - I(t)}{\Omega}. \qquad (2)$$

Table 1: Notations used in this paper.

Notation	Explanation
N	Total number of vulnerable hosts
Ω	Address space that a worm scans, *i.e.*, $\Omega = 2^{32}$
m	Number of groups in the Internet
N_i	Number of vulnerable hosts in group i and $\sum_{i=1}^{m} N_i = N$
Ω_i	Address space in group i and $\sum_{i=1}^{m} \Omega_i = \Omega$
s	Worm scanning rate: Number of scans that an infected host sends per unit time
$q_i(t)$	Group scanning distribution: Probability of a worm scan hitting group i at time t
$I(t)$	Total number of infected hosts at time t
$I_i(t)$	Number of infected hosts in group i at time t and $\sum_{i=1}^{m} I_i(t) = I(t)$
$S_i(t)$	Number of uninfected vulnerable hosts in group i at time t and $S_i(t) = N_i - I_i(t)$
$u(t)$	Total number of worm scans sent by all infected hosts by time t: $u(t) = s\int_0^t I(x)dx$
$r_i(t)$	Average group scanning distribution by time t: $r_i(t) = \frac{1}{t}\int_0^t q_i(x)dx$

This equation is a *logistic equation* [29] and can lead to a well-known closed-form solution [5]:

$$t = \frac{\Omega}{sN} \ln \frac{I(t)[N - I(0)]}{I(0)[N - I(t)]} \tag{3}$$

or

$$I(t) = \frac{I(0)N}{I(0) + [N - I(0)]e^{-sNt/\Omega}}. \tag{4}$$

The second type of deterministic worm-propagation models is based on discrete time and difference equations. For example, the AAWP model, developed by Chen *et al.* in [5], has been extended to characterize the spread of importance-scanning worms as follows [8]

$$I_i(t+1) = I_i(t) + [N_i - I_i(t)] \left[1 - \left(1 - \frac{1}{\Omega_i}\right)^{sI(t)q_i(t)}\right], \tag{5}$$

where $I_i(t)$ is the number of infected hosts in group i at time t and $I(t) = \sum_{i=1}^{m} I_i(t)$.

The AAWP model and its extensions have been shown to be able to model worm propagation accurately [5, 14, 8]. Based on the dynamic equation, however, it is difficult to understand the effects of important parameters (*e.g.,* N_i's and $q_i(t)$'s) on worm propagation. Therefore, in this work we attempt to derive a closed-form expression from the deterministic dynamic equation to characterize both the worm propagation speed and the parameters' effects explicitly.

3 Deriving a Closed-Form Expression

In this section, we derive a closed-form expression for modeling the spread of worm-scanning strategies through a mean-field approximation.

3.1 Deterministic Dynamic Differential Equation

We assume that at time t the group scanning distribution $q_i(t)$ is independent of the number of infected hosts in group i (*i.e.,* $I_i(t)$). Let s be the worm scanning rate or the rate at which an infected host scans an address space for a vulnerable host. Suppose that there are $S_i(t)$ uninfected vulnerable hosts and $I_i(t)$ infected hosts in group i at time t, where $S_i(t) + I_i(t) = N_i$. At time t, there are $sI(t)q_i(t)$ worm scans hitting group i, where $I(t)$ is the total number of infected hosts in the Internet and $I(t) = \sum_{i=1}^{m} I_i(t)$. If a worm scan hits group i at time t, this scan will hit an uninfected vulnerable host with probability $S_i(t)/\Omega_i$. Thus, $I_i(t)$ can be characterized by the classic SI epidemic model and follows a dynamic differential equation:

$$\frac{dI_i(t)}{dt} = sI(t)\frac{S_i(t)q_i(t)}{\Omega_i}. \tag{6}$$

Summing up $i = 1, 2, \cdots, m$ and using $S_i(t) = N_i - I_i(t)$, we have

$$\frac{dI(t)}{dt} = sI(t)\left(\sum_{i=1}^{m} \frac{N_i q_i(t)}{\Omega_i} - \sum_{i=1}^{m} \frac{I_i(t)q_i(t)}{\Omega_i}\right). \tag{7}$$

Note that if a worm uses RS, *i.e.*, $q_i = \Omega_i/\Omega$, Equation (7) becomes Equation (2) and leads to the closed-form solution, *i.e.,* Equation (3) or Equation (4). In general, however, it is difficult to derive a closed-form expression of $I(t)$ based on Equation (7).

3.2 Mean-Field Approximation

To get a closed-form expression of $I(t)$, we define $u(t)$ as the total number of worm scans sent by all infected hosts by time t, *i.e.*,

$$u(t) = s\int_0^t I(x)dx. \tag{8}$$

We also define $r_i(t)$ as the average group scanning distribution by time t, *i.e.*,

$$r_i(t) = \frac{1}{t}\int_0^t q_i(x)dx. \qquad (9)$$

Note that $\sum_{i=1}^m r_i(t) = 1$. Since there are in average $u(t)r_i(t)$ scans that hit group i among $u(t)$ scans, the mean value of the number of infected hosts in group i can be derived by

$$I_i(t) = S_i(0)\left[1 - \left(1 - \frac{1}{\Omega_i}\right)^{u(t)r_i(t)}\right], \qquad (10)$$

where $S_i(0)$ is the number of uninfected vulnerable hosts in group i at time 0 and $1-(1-1/\Omega_i)^{u(t)r_i(t)}$ is the probability that a vulnerable host in group i is hit by at least one worm scan. For most cases, $S_i(0) \approx N_i$. Note that Equation (10) applies a mean-field approach that neglects the fluctuation of the number of infections in group i and focuses on the average. We then apply the Taylor expansion and get

$$I_i(t) = u(t)\frac{S_i(0)r_i(t)}{\Omega_i} + O\left(\frac{1}{\Omega_i^2}\right). \qquad (11)$$

Assuming that $\Omega_i >> 1$ and $u(t)r_i(t)$ is not very large, we can obtain the approximation of the average of the number of infected hosts in group i or the mean-field approximation:

$$I_i(t) \approx u(t)\frac{S_i(0)r_i(t)}{\Omega_i}. \qquad (12)$$

Summing up $i = 1, 2, \cdots, m$, we have

$$u(t) = \frac{I(t)}{\sum_{i=1}^m S_i(0)r_i(t)/\Omega_i}. \qquad (13)$$

Plugging Equation (12) into Equation (7), we have

$$\frac{dI(t)}{dt} = sI(t)\left(\sum_{i=1}^m \frac{N_i q_i(t)}{\Omega_i} - u(t)\sum_{i=1}^m \frac{S_i(0)q_i(t)r_i(t)}{\Omega_i^2}\right). \qquad (14)$$

Applying Equation (13), Equation (14) becomes

$$\frac{dI(t)}{dt} = sI(t)\left(\sum_{i=1}^m \frac{N_i q_i(t)}{\Omega_i} - \frac{\sum_{i=1}^m S_i(0)q_i(t)r_i(t)/\Omega_i^2}{\sum_{i=1}^m S_i(0)r_i(t)/\Omega_i} I(t)\right). \qquad (15)$$

Setting

$$C(t) = s\sum_{i=1}^m \frac{N_i q_i(t)}{\Omega_i} \qquad (16)$$

$$D(t) = \frac{\sum_{i=1}^m N_i q_i(t)/\Omega_i \cdot \sum_{i=1}^m S_i(0)r_i(t)/\Omega_i}{\sum_{i=1}^m S_i(0)q_i(t)r_i(t)/\Omega_i^2}, \qquad (17)$$

Equation (15) becomes

$$\frac{dI(t)}{dt} = \frac{C(t)}{D(t)} I(t)\left[D(t) - I(t)\right]. \qquad (18)$$

The above differential equation is known as the Riccati equation [28] and can be solved in closed form.

Theorem 1. *The closed-form solution for $I(t)$ is given by*

$$I(t) = \frac{I(0)e^{F(t)}}{1 + I(0)\int_0^t E(x)e^{F(x)}dx}, \qquad (19)$$

where

$$E(t) = \frac{C(t)}{D(t)} \qquad (20)$$

$$F(t) = \int_0^t C(x)dx. \qquad (21)$$

PROOF: Setting $E(t) = C(t)/D(t)$ and $I(t) = 1/W(t)$, we obtain the linear differential equation from Equation (18)

$$\frac{dW(t)}{dt} + C(t)W(t) = E(t). \qquad (22)$$

Multiplying $e^{\int_0^t C(x)dx}$ to both sides of the above equation, we have

$$\frac{d}{dt}\left[W(t)e^{\int_0^t C(x)dx}\right] = E(t)e^{\int_0^t C(x)dx}. \qquad (23)$$

Integrating from 0 to t and setting $F(t) = \int_0^t C(x)dx$, Equation (23) becomes

$$W(t) = e^{-F(t)}\left[W(0) + \int_0^t E(x)e^{F(x)}dx\right]. \qquad (24)$$

Therefore,

$$I(t) = \frac{I(0)e^{F(t)}}{1 + I(0)\int_0^t E(x)e^{F(x)}dx}. \qquad (25)$$

■

4 Discussions

In this section, we apply the closed-form expression to static worm-scanning strategies and self-learning worms.

4.1 Static Worm-Scanning Strategies

For static worm-scanning strategies, $q_i(t) = q_i$, and $r_i(t) = q_i$. Thus, the solution for $I(t)$, *i.e.*, Equation (19), becomes

$$I(t) = \frac{I(0)D}{I(0) + [D - I(0)]e^{-Ct}}, \qquad (26)$$

where

$$C = C(t) = s \sum_{i=1}^{m} \frac{N_i q_i}{\Omega_i} \tag{27}$$

$$D = D(t) = \frac{\sum_{i=1}^{m} N_i q_i/\Omega_i \cdot \sum_{i=1}^{m} S_i(0) q_i/\Omega_i}{\sum_{i=1}^{m} S_i(0) q_i^2/\Omega_i^2}. \tag{28}$$

Equation (26) can also be written as

$$t = \frac{1}{C} \ln \frac{I(t)[D - I(0)]}{I(0)[D - I(t)]}, \tag{29}$$

i.e., the time required for a worm to infect a certain number of vulnerable hosts. Note that if a worm uses RS, *i.e.*, $q_i = \Omega_i/\Omega$, $C = sN/\Omega$ and $D = N$, and Equations (26) and (29) are then reduced to Equations (4) and (3), respectively.

In Equations (26) and (29), C and D are two important factors that control the spreading dynamics of a worm. Meanwhile, C and D are determined by the following parameters: the scanning rate s, the vulnerable-host distribution N_i's, the distribution of initially uninfected vulnerable hosts $S_i(0)$'s, and the worm-scanning strategy q_i's. Thus, Equations (26) and (29) explicitly show how these parameters affect worm spreading. Specifically, when t is small and thus $I(t)$ is small, $[D-I(0)]/[D-I(t)]$ is close to 1, and therefore C dominates the worm propagation speed. It is noted that C is indeed the *infection rate* of a worm that is derived in [8]. As a result, when a worm has a larger infection rate, it can spend much less time to infect the same number of vulnerable hosts at the early stage. Moreover, $\max\{C\} = s \cdot \max_i\{N_i/\Omega_i\}$, *i.e.*, a worm achieves the maximum infection rate when the worm scans only the group containing the largest number of the vulnerable-host density. In this case, the worm uses an extremely non-uniform scanning method. When t and $I(t)$ become larger, D has a greater effect on worm propagation. Since for most cases, $S_i(0) \approx N_i$ or $S_i(0) \propto N_i$,

$$D = \frac{\left(\sum_{i=1}^{m} N_i q_i/\Omega_i\right)^2}{\sum_{i=1}^{m} N_i q_i^2/\Omega_i^2} \leq N \tag{30}$$

by the Cauchy-Schwarz inequality, where the equality holds if and only if $q_i = \Omega_i/\Omega$, *i.e.*, a worm uses RS. Thus, if a worm uses a more uniform scanning method, D becomes larger and gets close to N, and $[D-I(0)]/[D-I(t)]$ becomes smaller, which leads to smaller t in Equation (29). Therefore, C and D affect worm propagation in very different ways.

It has been observed that if a worm uses a non-uniform scanning method, $D < N$ from Inequality (30). Meanwhile, from Equation (26), it can be seen that $I(t) \leq D$ even when t is very large, assuming $D \geq I(0)$. Thus, for the model described by Equation (26), a worm cannot infect more than D vulnerable hosts. This may not be valid, since a worm can infect all N vulnerable hosts under the condition that $q_i > 0$, if $N_i > 0$ for $\forall i$. Therefore, when t is very large, the model may not describe the worm behavior accurately. The reason for this inaccuracy is that in Equation (12), we assume that $u(t)r_i(t)$ is not very large and ignore the higher order terms of the Taylor expansion. Nevertheless, based on the above analysis, if a worm uses a more uniform scanning method, D gets closer to N, and our model is more accurate.

4.2 Self-Learning Worms

For a self-learning worm described in [7], the worm uses RS in the learning stage and STATIC-IS in the importance-scanning stage. That is,

$$q_i(t) = \begin{cases} \frac{\Omega_i}{\Omega}, & t \leq t_0; \\ q_i, & t > t_0, \end{cases} \tag{31}$$

where t_0 is the time when the worm decides to switch from RS to STATIC-IS and q_i's are the group scanning distribution used by STATIC-IS.

One way to derive the closed-form solution for self-learning worms is to use Equation (19) directly. However, we find that it is difficult to derive the closed-form expression for $\int_0^t E(x)e^{F(x)}dx$. Therefore, we follow another method and exploit the closed-form solutions for RS and static worm-scanning strategies, *i.e.,* Equations (4) and (26). Assuming that at time t_0, the number of infected hosts in group i is proportional to the number of vulnerable hosts in group i, *i.e.,* $I_i(t_0) = N_i I(t_0)/N$, we obtain a closed-form expression for self-learning worms

$$I(t) = \begin{cases} \frac{I(0)N}{I(0)+[N-I(0)]e^{-sNt/\Omega}}, & t \leq t_0; \\ \frac{I(t_0)D_1}{I(t_0)+[D_1-I(t_0)]e^{-C(t-t_0)}}, & t > t_0, \end{cases} \tag{32}$$

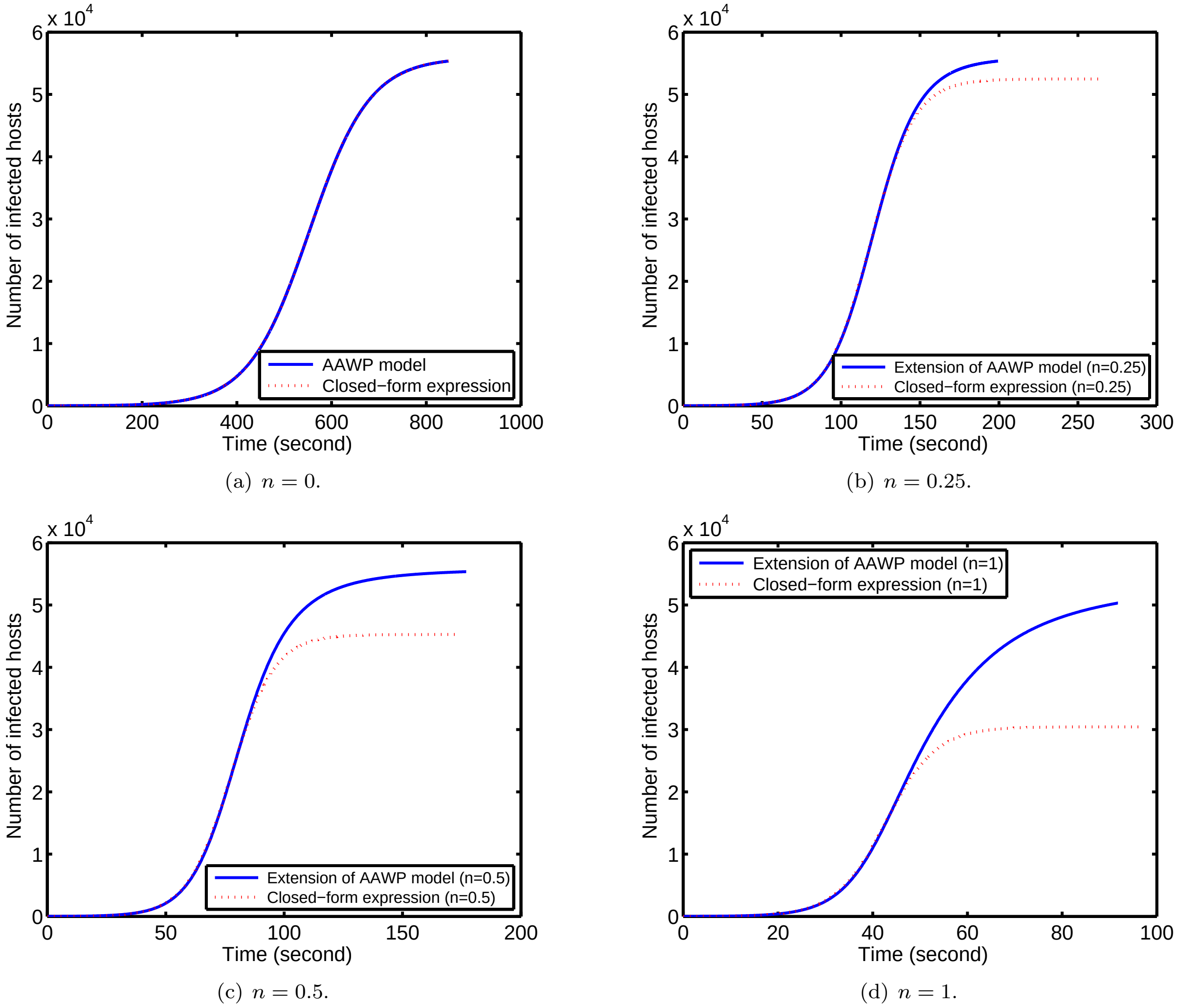
where C follows Equation (27) and

$$D_1 = \frac{\sum_{i=1}^{m} N_i q_i/\Omega_i \cdot \sum_{i=1}^{m} S_i(t_0) q_i/\Omega_i}{\sum_{i=1}^{m} S_i(t_0) q_i^2/\Omega_i^2}, \tag{33}$$

where $S_i(t_0) = N_i - I_i(t_0) = (1 - I(t_0)/N)N_i$. In this way, the solution for self-learning worms can be regarded as the simple extension from static worm-scanning strategies and has the similar properties. That is, we expect the closed-form expression can accurately model the self-learning worm spreading before the number of infected hosts becomes saturated, but may fail to characterize worm propagation at the very late stage.

5 Experimental Results

In this section, we evaluate our designed closed-form expression through experiments. In our experiments, we simulate the propagation of worms based on the parameters



(a) $n = 0$.

(b) $n = 0.25$.

(c) $n = 0.5$.

(d) $n = 1$.

Figure 2: Performance comparison of the closed-form expression and the extension of the AAWP model for Witty-worm propagation ($N = 55,909$, $s = 1,200$ per second, $m = 256$, and $I(0) = 10$).

chosen from Witty and Code Red worms. The Witty worm has a vulnerable population $N = 55,909$ and a scanning rate $s = 1,200$ scans per second [16], whereas the Code Red worm has $N = 360,000$ and $s = 358$ scans per minute [22]. We obtain the distribution of Witty-worm victims from CAIDA [26]. We ignore the effect of disk damages on the Witty worm propagation. Since the Code Red worm attacks Web servers, we assume that the victims of the Code Red worm have the same distribution as DShield data with port 80 [25, 2]. Then, we compare worm propagation characterized by our closed-form expression (*i.e.*, Equation (19)) with worm spreading described by the extension of the AAWP model (*i.e.*, Equation (5)).

5.1 Static Worm-Scanning Strategies

We first study static worm-scanning strategies. In our setting, static worm-scanning strategies exploit the /8 subnet distribution (*i.e.*, $m = 256$ and $\Omega_1 = \Omega_2 = \cdots = \Omega_{256} = 2^{24}$) or the /16 subnet distribution (*i.e.*, $m = 2^{16}$ and $\Omega_1 = \Omega_2 = \cdots = \Omega_{65536} = 2^{16}$). We consider a group of static strategies where q_i's relate to N_i's explicitly, *i.e.*,

$$q_i = \frac{N_i^n}{\sum_{i=1}^{m} N_i^n} \propto \left(\frac{N_i}{N}\right)^n \qquad (34)$$

where $n \geq 0$, representing how strongly q_i depends on N_i/N. If $n = 0$, q_i's are equal and are independent of N_i/N's. In this case, the worm uses RS. When n becomes larger, the worm would focus more scans on the groups with a large number of vulnerable hosts, which represents a more non-uniform scanning strategy. Figure 2 compares

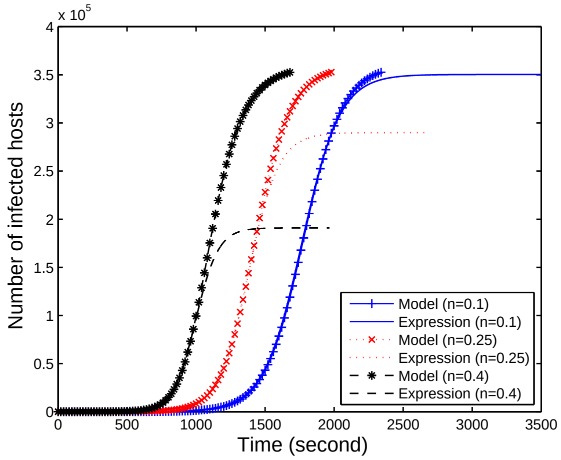


Figure 3: Performance comparison of the closed-form expression and the extension of the AAWP model for Code-Red-worm propagation ($N = 360,000$, $s = 358$ per minute, $m = 2^{16}$, and $I(0) = 1$).

our closed-form expression (*i.e.,* Equation (26)) with the extension of the AAWP model for Witty-worm propagation with $m = 256$, $I(0) = 10$, and $n = 0$, 0.25, 0.5, and 1. It can be seen that when $n = 0$, *i.e.,* the worm uses RS, the curves for both the expression and the model overlap. When n increases, the closed-form expression describes worm behaviors exactly the same as the extension of the AAWP model before $I(t)$ becomes very large. As we expect, when $I(t)$ is very large, our designed closed-form expression cannot characterize worm dynamics as a result of the effect of the parameter D. Before the infected hosts become saturated, however, the closed-form expression characterizes the average of the number of infected hosts accurately. Moreover, the model is more accurate if the scanning strategy is more uniform (*i.e.,* when n is smaller). We perform the same comparisons in Figure 3 for Code-Red-worm spreading with $m = 2^{16}$, $I(0) = 1$, and $n = 0.1$, 0.25, and 0.4. Similarly, we observe that before Code Red worms infect a large number of vulnerable hosts, the closed-form expression can accurately model the average of the number of infected hosts.

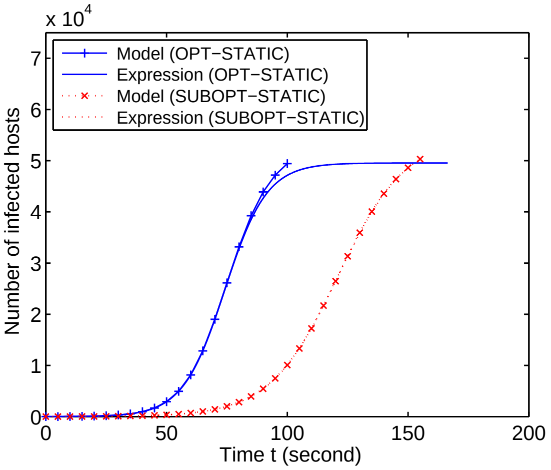


Figure 4: Two static scanning strategies from [19] ($N = 55,909$, $s = 1,200$ per second, $m = 256$, and $I(0) = 10$)).

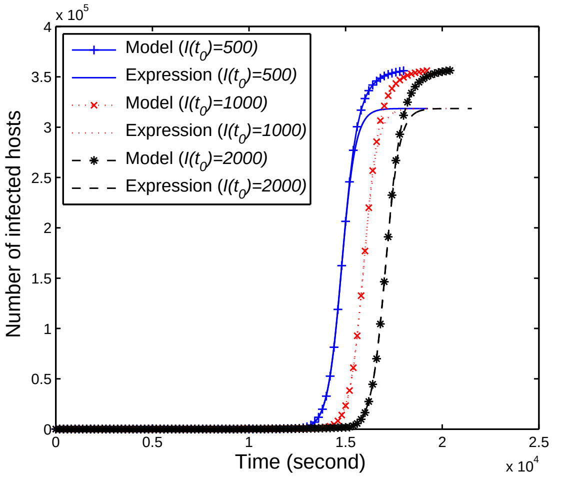


Figure 5: Self-learning worm propagation ($N = 360,000$, $s = 358$ per minute, $m = 256$, $I(0) = 1$, and $n = 0.5$).

We further apply our closed-form expression to describe other static strategies such as OPT-STATIC and SUBOPT-STATIC proposed in [19]. We simulate Witty-worm propagation with $m = 256$ and $I(0) = 10$. It can be seen from Figure 4 that our designed closed-form expression can faithfully capture the dynamic worm behaviors. Furthermore, we observe that the spreading speeds of these two strategies are very different. While the OPT-STATIC strategy takes only 102 seconds to infect 90% vulnerable hosts, the SUBOPT-STATIC strategy requires 155 seconds. The optimization method proposed in [19], however, cannot characterize this difference in the worm propagation speed.

5.2 Self-Learning Worms

We next simulate self-learning worms. We consider a self-learning worm using the parameters of the Code Red worm. We assume that the worm starts from one infected host (*i.e.,* $I(0) = 1$), uses RS to spread and collect the information on the distribution of vulnerable hosts during the learning stage, and then switches to STATIC-IS at time t_0. Here, we assume that the worm can estimate the vulnerable-host distribution accurately at time t_0 and uses $n = 0.5$ in Equation (34) for STATIC-IS as proposed in [7]. Figure 5 compares our closed-form expression (*i.e.,* Equation (32)) with the extension of the AAWP model with $m = 256$ and $I(t_0) = 500$, 1,000, and 2,000. Similarly, it can be seen that our expression can reflect the dynamic behavior of self-learning worms before the infected hosts become saturated.

6 Related Work

Several approaches have been proposed to model the spread of worms. *Stochastic* models have been studied to capture the variance of worm propagation at the early stage [15, 12]. Stochastic models, however, may require extensive computations and focus only on the early stage of worm spreading. Instead, most analytical models of worm propagation have used *deterministic* dynamic equations, ignoring the variance of worm infection [17, 24, 5, 18]. Moreover, the deterministic models have been widely applied to worm detection and defenses [11, 22, 18]. Based on dynamic equations, however, it is difficult to understand the effects of important parameters (*e.g.*, the vulnerable-host distribution and the group scanning distribution) on worm propagation. Furthermore, except for extreme cases (*e.g.*, random scanning), it is nearly impossible to derive an exact closed-form expression from the dynamic equation. An alternate approach to both model worm propagation and capture parameters' effects has been proposed by Vojnovic *et al.* [19]. The authors formulate worm infection as an *optimization* problem and focus on the number of worm scans required to reach a predetermined fraction of vulnerable hosts. It is pointed out, however, that two worm-scanning strategies can use the same number of worm scans to infect the same number of hosts, but differ significantly in the worm propagation speed in [3]. Therefore, to characterize both the worm propagation speed and the parameters' effects, we derive a *closed-form expression* from the deterministic dynamic equation through a mean-field approximation in this work.

Mean-field approaches [27, 13] stem from statistical mechanics and have been applied to networking and network security areas. For example, Wang uses a mean-field theory to analyze Internet router buffer sizing [20]. Baccelli *et al.* study the mean-field model to understand the interaction between HTTP flows using TCP [1]. Chen *et al.* model the spread of topological-scanning malwares by following the spirit of mean-field approximations [6]. It is noted that the topological scanning studied in [6] is a topology-based strategy and is very different from a scan-based method considered in this paper.

7 Conclusions

In this paper, we have presented a closed-form expression for modeling the propagation of a class of worm-scanning strategies. Our expression can both accurately characterize the worm propagation speed in the time window of detection and defenses and explicitly capture the effects of the vulnerable-host distribution and the worm-scanning method. Therefore, our solution can complement with the existing models such as stochastic models [15, 12], deterministic models [17, 24, 5, 18], and optimization methods [19].

As part of our ongoing work, we plan to extend our approach to study the closed-form expressions for modeling the spread of dynamic and adaptive worm-scanning strategies such as localized scanning and dynamic importance scanning.

ACKNOWLEDGMENT

This work was supported in part by the 2008 PRF Summer Faculty Grant at Purdue University.

REFERENCES

[1] F. Baccelli, A. Chaintreau, D. D. Vleeschauwer, and D. R. McDonald, "A mean-field analysis of short lived interacting flows," in *Proc. of ACM SIGMETRICS*, New York, June 2004, pp. 343-354.

[2] P. Barford, R. Nowak, R. Willett, and V. Yegneswaran, "Toward a model for sources of Internet background radiation," in *Proc. of the Passive and Active Measurement Conference (PAM'06)*, Mar. 2006.

[3] Z. Chen and C. Chen, "A closed-form expression for static worm-scanning strategies," in *Proc. of IEEE International Conference on Communications (ICC'08)*, Beijing, China, May 2008.

[4] Z. Chen, C. Chen, and C. Ji, "Understanding localized-scanning worms," in *Proc. of 26th IEEE International Performance Computing and Communications Conference (IPCCC'07)*, New Orleans, LA, Apr. 2007, pp. 186-193.

[5] Z. Chen, L. Gao, and K. Kwiat, "Modeling the spread of active worms," in *Proc. of INFOCOM'03*, vol. 3, San Francisco, CA, Apr. 2003, pp. 1890-1900.

[6] Z. Chen and C. Ji, "Spatial-temporal modeling of malware propagation in networks," *IEEE Transactions on Neural Networks: Special Issue on Adaptive Learning*

Systems in Communication Networks, vol. 16, no. 5, Sept. 2005, pp. 1291-1303.

[7] Z. Chen and C. Ji, "A self-learning worm using importance scanning," in *Proc. ACM/CCS Workshop on Rapid Malcode (WORM'05)*, Fairfax, VA, Nov. 2005, pp. 22-29.

[8] Z. Chen and C. Ji, "Optimal worm-scanning method using vulnerable-host distributions," *International Journal of Security and Networks: Special Issue on Computer and Network Security*, vol. 2, no. 1/2, 2007.

[9] Z. Chen, C. Ji, and P. Barford, "Spatial-temporal characteristics of malicious sources," in *Proc. of INFOCOM'08 Mini-Conference*, Phoenix, AZ, Apr. 2008.

[10] D. Moore, C. Shannon, and J. Brown, "Code-Red: a case study on the spread and victims of an Internet worm," in *ACM SIGCOMM Internet Measurement Workshop*, Marseille, France, Nov. 2002.

[11] D. Moore, C. Shannon, G. Voelker, and S. Savage, "Internet quarantine: requirements for containing self-propagating code," in *Proc. of INFOCOM'03*, vol. 3, San Francisco, CA, Apr. 2003, pp. 1901-1910.

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

[13] M. Opper and D. Saad (Eds.), *Advanced Mean Field Methods, Theory and Practice.* MIT Press, Feb. 2001.

[14] M. A. Rajab, F. Monrose, and A. Terzis, "On the effectiveness of distributed worm monitoring," in *Proc. of the 14th USENIX Security Symposium (Security'05)*, Baltimore, MD, Aug. 2005, pp. 225-237.

[15] K. Rohloff and T. Basar, "Stochastic behavior of random constant scanning worms," in *Proc. of the 14th ICCCN*, 2005.

[16] C. Shannon and D. Moore, "The spread of the Witty worm," *IEEE Security and Privacy*, vol. 2, no. 4, Jul-Aug 2004, pp. 46-50.

[17] S. Staniford, V. Paxson, and N. Weaver, "How to 0wn the Internet in your spare time," in *Proc. of the 11th USENIX Security Symposium (Security'02)*, San Francisco, CA, Aug. 2002.

[18] M. Vojnovic and A. J. Ganesh, "On the race of worms, alerts and patches," to appear in *IEEE/ACM Transactions on Networking*, 2008.

[19] M. Vojnovic, V. Gupta, T. Karagiannis, and C. Gkantsidis, "Sampling strategies for epidemic-style information dissemination," in *Proc. of INFOCOM'08*, Phoenix, AZ, Apr. 2008.

[20] M. Wang, "Mean-field analysis of buffer sizing," in *Proc. of 50th Annual IEEE Global Communications Conference (GLOBECOM'07)*, Washington DC, Nov. 2007.

[21] J. Wu, S. Vangala, L. Gao, and K. Kwiat, "An effective architecture and algorithm for detecting worms with various scan techniques," in *Proc. 11th Ann. Network and Distributed System Security Symposium (NDSS'04)*, San Diego, CA, Feb. 2004.

[22] C. C. Zou, W. Gong, D. Towsley, and L. Gao, "The monitoring and early detection of Internet worms," *IEEE/ACM Transactions on Networking*, vol. 13, no. 5, Oct. 2005, pp. 961-974.

[23] C. C. Zou, D. Towsley, W. Gong, and S. Cai, "Advanced routing worm and its security challenges," *Simulation: Transactions of the Society for Modeling and Simulation International*, vol. 82, no. 1, 2006, pp. 75-85.

[24] C. C. Zou, D. Towsley, and W. Gong, "On the performance of Internet worm scanning strategies," *Elsevier Journal of Performance Evaluation*, vol. 63. no. 7, July 2006, pp. 700-723.

[25] Distributed Intrusion Detection System (DShield), http://www.dshield.org/.

[26] The CAIDA Dataset on the Witty Worm - March 19-24, 2004, Colleen Shannon and David Moore, http://www.caida.org/data/passive/witty_worm_data set.xml. 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, DARPA, Digital Envoy, and CAIDA Members.

[27] Wikipedia, "Mean field theory," http://en.wikipedia.org/wiki/Mean_field_theory.

[28] Wikipedia, "Riccati equation," http://en.wikipedia.org/wiki/Riccati_equation.

[29] Wolfram MathWorld, "Logistic equation," http://mathworld.wolfram.com/LogisticEquation.html.