Characterizing Internet Worm Infection Structure

Qian Wang
Florida International University
Miami, Florida 33174

Zesheng Chen, Chao Chen
Indiana University - Purdue University Fort Wayne
Fort Wayne, Indiana 46805

Abstract

Internet worm infection continues to be one of top security threats and has been widely used by botnets to recruit new bots. In this work, we attempt to quantify the infection ability of individual hosts and reveal the key characteristics of the underlying topology formed by worm infection, *i.e.*, the number of children and the generation of the worm infection family tree. Specifically, we apply probabilistic modeling methods and a sequential growth model to analyze the infection tree of a wide class of worms. Through both mathematical analysis and simulation, we find that the number of children has asymptotically a geometric distribution with parameter 0.5. As a result, on average half of infected hosts never compromise any vulnerable host, over 98% of infected hosts have no more than five children, and a small portion of infected hosts have a large number of children. We also discover that the generation follows closely a Poisson distribution and the average path length of the worm infection family tree increases approximately logarithmically with the total number of infected hosts.

1 Introduction

Internet worms are malicious software that can compromise vulnerable hosts and use them to attack other victims, and have been one of top security threats since the Code Red and Nimda worms in 2001. Botnets are zombie networks controlled by attackers through Internet relay chat (IRC) systems (*e.g.,* GT Bot) or peer-to-peer (P2P) systems (*e.g.,* Storm) to execute coordinated attacks and have become the number one threat to the Internet in recent years. The main difference between worms and botnets lies in that worms emphasize the procedures of infecting targets and propagating among vulnerable hosts, whereas botnets focus on the mechanisms of organizing the network of compromised computers and setting out coordinated attacks. Most botnets, however, still apply worm-scanning methods to recruit new bots or collect network information [8, 10, 14, 17]. Moreover, although many P2P-based botnets use the existing P2P networks to build a bootstrap procedure, Conficker C forms a P2P botnet through scan-based peer discovery [11, 3]. Specifically, Conficker C searches for new peers by randomly scanning the entire Internet address space. As a result, the way that Conficker C constructs a P2P-based botnet is in principle the same as worm scanning/infection. Therefore, characterizing the structure of worm infection is important and imperative for defending against current and future epidemics such as Internet worms and Conficker C like P2P-based botnets.

Modeling Internet worm infection has been focused on the *macro* level. Most, if not all, mathematical models study the total number of infected hosts over time [12, 24, 4, 15, 8]. The models of the *micro* level of worm infection, however, have been investigated little. The micro-level models can provide more insights into the infection ability of individual compromised hosts and the underlying topologies formed by worm infection. A key micro-level information is "who infects whom" or the worm infection family tree. When a host infects another host, they form a "father-and-son" relationship, which is represented by a directed edge in a graph formed by worm infection. Hence, the procedure of worm propagation constructs a directed tree where patient zero is the root and the infected hosts that do not compromise any vulnerable host are leaves (see Fig. 1). To the best of our knowledge, there is yet no mathematical model for characterizing the structure of such a tree.

The goal of this work is to characterize the Internet worm infection family tree, called the "worm tree" in short. For such a tree, we are particularly interested in two metrics:

- *Number of children:* For a randomly selected node in the tree, how many children does it have? This metric represents the infection ability of individual hosts.
- *Generation:* For a randomly selected node in the tree, which generation (or level) does it belong to? This metric indicates the average path length of the graph formed by worm infection.

These two metrics have important implications and applications for security analysis. First, the distribution of the number of children can be used to answer questions such as what is the probability that an infected host compromises more than 10 vulnerable hosts. Moreover, it provides insights into the robustness of the Conficker C like P2P-based botnets [1, 7]. Second, some schemes

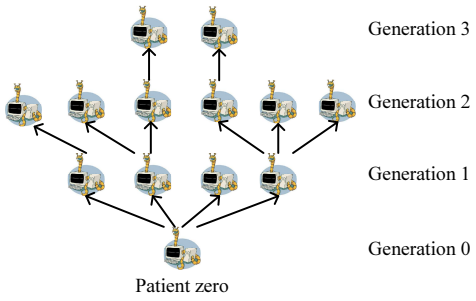


Figure 1: A worm tree.

have been proposed to trace worms back to their origins through the cooperation between infected hosts [21], and the distribution of the generation can provide the information on the number of hosts required to cooperate. Moreover, it sheds light on the delay or the effort for a botmaster to deliver a command to all bots in a P2P-based botnet like Conficker C.

To study these two metrics analytically, we apply probabilistic modeling methods and derive the joint probability distribution of the number of children and the generation through a sequential growth model. From the joint distribution, we analyze the marginal distributions of the number of children and the generation. We also develop closed-form approximations to both marginal distributions and the joint distribution. As a first attempt, we analyze the worm tree formed by a wide class of worms such as random-scanning worms [12], routable-scanning worms [20, 24], importance-scanning worms [5], OPT-STATIC worms [16], and SUBOPT-STATIC worms [16]. For these worms, a new victim is compromised by each existing infected host with equal probability. We then verify the analytical results through simulations.

Through both mathematical analysis and simulation, we make several discoveries from this research as follows. If a worm uses a scanning method for which a new victim is compromised by each existing infected host with equal probability, the number of children is shown to have asymptotically a geometric distribution with parameter 0.5. This means that on average half of infected hosts never compromise any target and over 98% of infected hosts have no more than five children. On the other hand, this also indicates that a small portion of hosts infect a large number of vulnerable hosts. Moreover, the generation is demonstrated to closely follow a Poisson distribution with parameter $H_n - 1$, where n is the number of nodes and H_n is the n-th harmonic number [6]. This means that the average path length of the worm tree increases approximately logarithmically with the number of nodes. To the best of our knowledge, this is the first attempt in understanding and exploiting the topology formed by worm infection quantitatively.

The remainder of this paper is structured as follows. Section 2 presents our sequential growth model and assumptions used in analyzing the worm tree. Section 3 gives our analysis on the worm tree. Section 4 uses simulations to verify the analytical results. Finally, Section 5 concludes this paper.

2 Worm Tree and Sequential Growth Model

In this section, we provide the background on the worm tree, and present the assumptions and the growth model.

An example of a worm tree is given in Fig. 1. Here, patient zero is the root and belongs to generation 0. The tail of an arrow is from the "father" or the infector, whereas the head of an arrow points to the "son" or the infectee. If a father belongs to generation i, then its children lie in generation $i+1$. In a worm tree with n nodes, we use $L_n(i,j)$ $(0 \le i,j \le n-1)$ to denote the number of nodes that have i children and belong to generation j. Note that $\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} L_n(i,j) = n$. We also use $C_n(i)$ $(i = 0,1,2,\cdots,n-1)$ to denote the number of nodes that have i children and $G_n(j)$ $(j = 0,1,2,\cdots,n-1)$ to denote the number of nodes in generation j. Moreover, $L_n(i,j)$, $C_n(i)$, and $G_n(j)$ are random variables. Thus, we define $p_n(i,j) = \frac{\mathrm{E}[L_n(i,j)]}{n}$, representing the joint distribution of the number of children and the generation. Similarly, we define $c_n(i) = \frac{\mathrm{E}[C_n(i)]}{n}$ to represent the marginal distribution of the number of children and $g_n(j) = \frac{\mathrm{E}[G_n(j)]}{n}$ to represent the marginal distribution of the generation. Note that $c_n(i) = \sum_{j=0}^{n-1} p_n(i,j)$ and $g_n(j) = \sum_{i=0}^{n-1} p_n(i,j)$. Although we model worm infection as a tree, different worm trees can show very different structures. In the extended version of this paper [18], we demonstrate that two extreme cases of worm trees (*i.e.*, bus and star topologies) can have very different $c_n(i)$ and $g_n(j)$.

To study the worm tree analytically, in this paper we make several assumptions and considerations. First, to simplify the model, we assume that infected hosts have the same scanning rate. This assumption is removed in Section 4.2, where we use simulations to study the effect of the variation of scanning rates on the worm tree. Second, we consider a wide class of worms for which a new victim is compromised by each existing infected host with equal probability. Such worms include random-scanning worms, routable-scanning worms, importance-scanning worms, OPT-STATIC worms, and SUBOPT-STATIC worms. Random scanning selects targets in the IPv4 address space randomly and has been the main scanning method for both worms and botnets [12, 10]; routable scanning finds victims in the routable IPv4 address space [20, 24]; and importance scanning probes subnets according to the vulnerable-host distribution

[5]. OPT-STATIC and SUBOPT-STATIC are optimal and suboptimal scanning methods that are proposed in [16] to minimize the number of worm scans required to reach a predetermined fraction of vulnerable hosts. Third, we consider the classic susceptible → infected (SI) model, ignoring the cases that an infected host can be cleaned and becomes vulnerable again, or can be patched and becomes invulnerable. The SI model assumes that once infected, a host remains infected. Such a simple model has been widely applied in studying worm infection [12, 24, 16], and presents the worst case scenario. Fourth, we assume that there is no re-infection. That is, if an infected host is hit by a worm scan, this host will not be further re-infected. As a result, every infected host has one and only one father except for patient zero, and the resulting graph formed by worm infection is a tree. Fifth, we assume that the worm starts from one infected host, *i.e.,* patient zero or a hitlist size of 1. When the hitlist size is larger than 1, the underlying infection topology is a worm forest, instead of a worm tree. Our analysis, however, can easily be extended to model the worm forest. Finally, we assume that in our model, nodes are added into the worm tree sequentially and no two nodes are added at the same time. That is, no two vulnerable hosts are infected simultaneously. Rather than attempting to characterize the dynamics of worm propagation (*e.g.,* the total number of infected hosts over time), we make this assumption aim at capturing the main features of the topology formed by worm infection (*e.g.,* the number of children and the generation) without sacrificing much accuracy. Taking the Code Red v2 worm as an example, during its early stage of propagation, there are few infected hosts struggling to find other vulnerable machines out of the IPv4 address space, hence very few hosts get infected at the same time. On the other hand, during its later fast spread period, the number of hosts that get compromised simultaneously is trivial compared with the number of existing infected hosts. In Section 4 where simulations are performed, we relax this assumption.

Based on these considerations and assumptions, the sequential growth model of a worm tree works as follows: We consider a fixed sequence of infected hosts (*i.e.,* nodes) $v_1, v_2, \cdots$ and inductively construct a random worm tree $(T_n)_{n\geq 1}$, where n is the number of nodes and T_1 has only patient zero. Infecting a new host is equivalent to adding a new node into the existing worm tree. Hence, given T_{n-1}, T_n is formed by adding node v_n together with an edge directed from an existing node v_f to v_n. According to the assumption, v_f is randomly chosen among the $n-1$ nodes in the tree, *i.e.,* $\Pr(f=k)=\frac{1}{n-1}$, $k=1,2,\cdots,n-1$. Note that such a sequential growth model and its variations have been widely used in studying topology generators [2]. In this paper, we apply this model to characterize worm infection.

3 Mathematical Analysis

In this section, we study the worm tree through mathematical analysis. Specifically, we first derive the joint distribution of the number of children and the generation, *i.e.,* $p_n(i,j)$, by applying probabilistic methods. We then use $p_n(i,j)$ to analyze two marginal distributions, *i.e.,* $c_n(i)$ and $g_n(j)$, and obtain their closed-form approximations. Finally, we find a closed-form approximation to $p_n(i,j)$.

3.1 Joint Distribution

For a worm tree with only patient zero (*i.e.,* $n=1$), since $L_1(0,0)=1$ with probability 1, $p_1(0,0)=1$. Similarly, for a worm tree with $n=2$, it is evident that $L_2(1,0)=L_2(0,1)=1$. Thus, $p_2(1,0)=p_2(0,1)=\frac{1}{2}$. We now consider $p_n(i,j)$ $(0\leq i,j\leq n-1)$ when $n\geq 3$. Specifically, we study two cases:

(1) $p_n(0,j)$, *i.e.,* the proportion of the number of leaves in generation j in T_n. Assume that T_{n-1} is given, and there are $L_{n-1}(0,j)$ leaves in generation j and totally $G_{n-1}(j-1)=\sum_{i=0}^{n-2}L_{n-1}(i,j-1)$ nodes in generation $j-1$. Note that we have extended the notation so that $G_{n-1}(-1)=L_{n-1}(i,-1)=0$, $0\leq i\leq n-2$. When a new node v_n is added, v_n becomes a leaf of T_n. If v_n is connected to one of existing nodes in generation $j-1$, v_n belongs to generation j; and the probability of such an event is $\frac{G_{n-1}(j-1)}{n-1}$. Moreover, if a leaf in generation j in T_{n-1} connects to v_n, this node is no longer a leaf and now has one child; and the probability of this event is $\frac{L_{n-1}(0,j)}{n-1}$. Therefore, we can obtain the stochastic recurrence of $L_n(0,j)$:

$$L_n(0,j)=\begin{cases} L_{n-1}(0,j)+1, & \text{w.p. } \frac{G_{n-1}(j-1)}{n-1} \\ L_{n-1}(0,j)-1, & \text{w.p. } \frac{L_{n-1}(0,j)}{n-1} \\ L_{n-1}(0,j), & \text{otherwise.} \end{cases} \tag{1}$$

Given T_{n-1} (*i.e.,* $L_{n-1}(0,j)$ and $G_{n-1}(j-1)$), the conditional expected value of $L_n(0,j)$ is $[L_{n-1}(0,j)+1]\cdot\frac{G_{n-1}(j-1)}{n-1}+[L_{n-1}(0,j)-1]\cdot\frac{L_{n-1}(0,j)}{n-1}+L_{n-1}(0,j)\cdot\left[1-\frac{G_{n-1}(j-1)+L_{n-1}(0,j)}{n-1}\right]$, *i.e.,*

$$\mathrm{E}[L_n(0,j)|T_{n-1}]=\tfrac{n-2}{n-1}L_{n-1}(0,j)+\tfrac{1}{n-1}G_{n-1}(j-1). \tag{2}$$

Applying $\mathrm{E}[L_n(0,j)]=\mathrm{E}[\mathrm{E}[L_n(0,j)|T_{n-1}]]$ (*i.e.,* the law of total expectation), we obtain

$$\mathrm{E}[L_n(0,j)]=\tfrac{n-2}{n-1}\mathrm{E}[L_{n-1}(0,j)]+\tfrac{1}{n-1}\mathrm{E}[G_{n-1}(j-1)]. \tag{3}$$

Using the definitions $p_n(0,j)=\frac{\mathrm{E}[L_n(0,j)]}{n}$ and $g_{n-1}(j-1)=\frac{\mathrm{E}[G_{n-1}(j-1)]}{n-1}=\sum_{i=0}^{n-2}p_{n-1}(i,j-1)$, the above

equation leads to

$$p_n(0,j) = \frac{n-2}{n} p_{n-1}(0,j) + \frac{1}{n} g_{n-1}(j-1) \quad (4)$$

$$= \frac{n-2}{n} p_{n-1}(0,j) + \frac{1}{n} \sum_{i=0}^{n-2} p_{n-1}(i,j-1). \quad (5)$$

(2) $p_n(i,j)$, $1 \le i \le n-1$. Given $L_{n-1}(i,j)$ and $L_{n-1}(i-1,j)$ in T_{n-1}, we study $L_n(i,j)$ in T_n. When the new node v_n is added into T_{n-1}, v_n is connected to a node with $i-1$ children and in generation j with probability $\frac{L_{n-1}(i-1,j)}{n-1}$, or is connected to a node with i children and in generation j with probability $\frac{L_{n-1}(i,j)}{n-1}$. Thus, in T_n,

$$L_n(i,j) = \begin{cases} L_{n-1}(i,j)+1, & \text{w.p. } \frac{L_{n-1}(i-1,j)}{n-1} \\ L_{n-1}(i,j)-1, & \text{w.p. } \frac{L_{n-1}(i,j)}{n-1} \\ L_{n-1}(i,j), & \text{otherwise.} \end{cases} \quad (6)$$

This relationship leads to

$$\mathrm{E}[L_n(i,j)|T_{n-1}] = \frac{n-2}{n-1} L_{n-1}(i,j) + \frac{1}{n-1} L_{n-1}(i-1,j). \quad (7)$$

Therefore,

$$\mathrm{E}[L_n(i,j)] = \frac{n-2}{n-1} \mathrm{E}[L_{n-1}(i,j)] + \frac{1}{n-1} \mathrm{E}[L_{n-1}(i-1,j)]. \quad (8)$$

That is,

$$p_n(i,j) = \frac{n-2}{n} p_{n-1}(i,j) + \frac{1}{n} p_{n-1}(i-1,j). \quad (9)$$

Summarizing the above two cases, we have the following theorem:

Theorem 1 *When $n \ge 3$, the joint distribution of the number of children and the generation in a worm tree T_n follows*

$$p_n(i,j) = \begin{cases} \frac{n-2}{n} p_{n-1}(0,j) + \frac{1}{n} g_{n-1}(j-1), & i = 0 \\ \frac{n-2}{n} p_{n-1}(i,j) + \frac{1}{n} p_{n-1}(i-1,j), & \textit{otherwise}, \end{cases} \quad (10)$$

where $0 \le i, j \le n-1$.

Theorem 1 provides a way to calculate $p_n(i,j)$ recursively from $p_2(i,j)$.

3.2 Number of Children

We use $p_n(i,j)$ to derive the marginal distribution of the number of children, *i.e.*, $c_n(i)$. Similarly, we study two cases:

(1) $c_n(0)$, *i.e.*, the proportion of the number of leaves in T_n. Since $c_n(0) = \sum_{j=0}^{n-1} p_n(0,j)$ and $\sum_{j=0}^{n-1} g_{n-1}(j-1) = 1$, we obtain the recursive relationship of $c_n(0)$ from Equation (4):

$$c_n(0) = \frac{n-2}{n} c_{n-1}(0) + \frac{1}{n}. \quad (11)$$

Moreover, note that $c_2(0) = \frac{1}{2}$. If we assume that $c_{n-1}(0) = \frac{1}{2}$, we can obtain by induction that

$$c_n(0) = \frac{1}{2}. \quad (12)$$

This indicates that no matter how many nodes are in the worm tree, on average half of nodes are leaves, *i.e.*, on average 50% of infected hosts never compromise any target.

(2) $c_n(i)$, $1 \le i \le n-1$. From Equation (9) and $c_n(i) = \sum_{j=0}^{n-1} p_n(i,j)$, we find the recurrence of $c_n(i)$ as follows

$$c_n(i) = \frac{n-2}{n} c_{n-1}(i) + \frac{1}{n} c_{n-1}(i-1). \quad (13)$$

Summarizing the above two cases, we have the following theorem on the distribution of the number of children:

Theorem 2 *When $n \ge 3$, the distribution of the number of children in a worm tree T_n follows*

$$c_n(i) = \begin{cases} \frac{1}{2}, & i = 0 \\ \frac{n-2}{n} c_{n-1}(i) + \frac{1}{n} c_{n-1}(i-1), & 1 \le i \le n-1. \end{cases} \quad (14)$$

From Theorem 2, we can derive the statistical properties of the number of children as follows.

Corollary 1 *When $n \ge 1$, the expectation and the variance of the number of children are*

$$E_n[C] = \sum_{i=0}^{n-1} i \cdot c_n(i) = \frac{n-1}{n} \quad (15)$$

$$\mathit{Var}_n[C] = \sum_{i=0}^{n-1} \left(i - E_n[C]\right)^2 \cdot c_n(i) = 2 - \frac{n-1}{n^2} - \frac{2H_n}{n}, \quad (16)$$

where $H_n = \sum_{i=1}^{n} \frac{1}{i}$ is the n-th harmonic number [6].

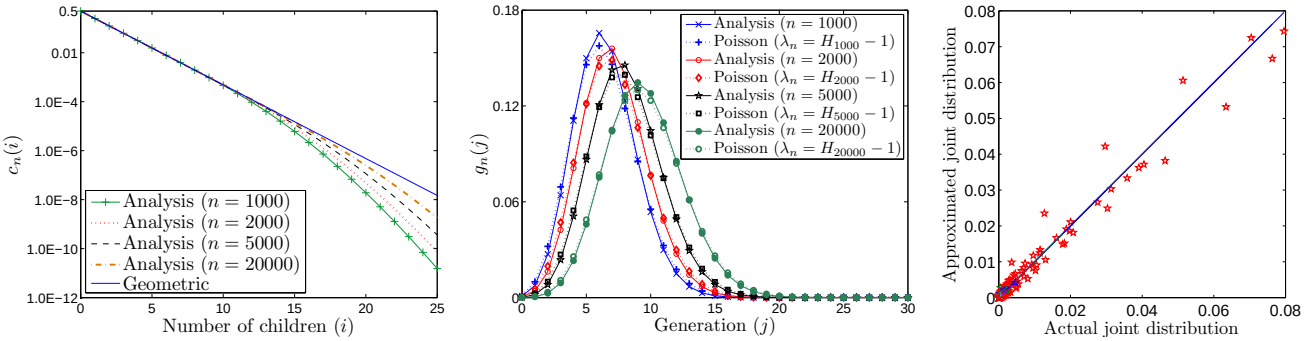
The proof of Corollary 1 is given in the extended version of this paper [18]. One intuitive way to derive $\mathrm{E}_n[C]$ is that in worm tree T_n, there are $n-1$ directed edges and n nodes. Thus, the average number of edges (*i.e*, the average number of children) of a node is $\frac{n-1}{n}$. Moreover, since H_n is $O(1+\ln n)$, $\lim_{n\to\infty} \mathrm{E}_n[C] = 1$, and $\lim_{n\to\infty} \mathrm{Var}_n[C] = 2$.

Theorem 2 also leads to a simple closed-form expression of the distribution of the number of children when n is very large, as shown in the following corollary.

Corollary 2 *When $n \to \infty$, the number of children has a geometric distribution with parameter $\frac{1}{2}$,* i.e.,

$$c(i) = \lim_{n\to\infty} c_n(i) = \left(\frac{1}{2}\right)^{i+1}, \quad i = 0, 1, 2, \cdots. \quad (17)$$

The proof of Corollary 2 is given in [18]. Corollary 2 indicates that when n is very large, $c_n(i)$ decreases approximately exponentially with a decay constant of $\ln 2$ as the number of children increases. We further study when both n and i are finite and large, how $c_n(i)$ varies



(a) Number of children. (b) Generation. (c) Joint distribution ($n = 2000$).

Figure 2: Mathematical analysis of the worm infection structure.

with n, *i.e.*, how the tail of the distribution of the number of children changes with n. First, note that $c_3(0) = \frac{1}{2}$, $c_3(1) = \frac{1}{3}$, and $c_3(2) = \frac{1}{6}$. Thus, from Equation (13), we can prove by induction that $c_n(i)$ ($n \geq 3$) is a decreasing function of i, *i.e.*, $c_n(i) < c_n(i-1)$, for $1 \leq i \leq n-1$. Next, putting this inequality into Equation (13), we have $c_n(i) > \frac{n-1}{n} c_{n-1}(i)$. Hence, when n is very large, $\frac{n-1}{n} \approx 1$, and $c_n(i) > c_{n-1}(i)$, which indicates that the tail of $c_n(i)$ increases with n. Fig. 2(a) verifies this result, showing $c_n(i)$ obtained from Theorem 2 when $n = 1000$, 2000, 5000, and 20000, as well as the geometric distribution with parameter 0.5 obtained from Corollary 2. Note that the y-axis uses log-scale. It can be seen that when n increases from 1000 to 20000, the tail of $c_n(i)$ also increases to approach the tail of the geometric distribution. Moreover, it is shown that the geometric distribution well approximates the distribution of the number of children when n is large.

3.3 Generation

Next, we derive the generation distribution (*i.e.*, $g_n(j)$) in a similar manner to the case of $c_n(i)$. Using Theorem 1 and $g_n(j) = \sum_{i=0}^{n-1} p_n(i,j)$, we obtain the following theorem:

Theorem 3 *When $n \geq 3$, the distribution of the generation in a worm tree T_n follows*

$$g_n(j) = \frac{n-1}{n} g_{n-1}(j) + \frac{1}{n} g_{n-1}(j-1), 0 \leq j \leq n-1, \tag{18}$$

where $g_{n-1}(-1) = 0$.

Theorem 3 gives a method to calculate the distribution of the generation recursively. Moreover, from Theorem 3, we can derive the statistical properties of the generation distribution in the following corollary.

Corollary 3 *When $n \geq 1$, the expectation and the variance of the generation are*

$$E_n[G] = \sum_{j=0}^{n-1} j \cdot g_n(j) = H_n - 1. \tag{19}$$

$$Var_n[G] = \sum_{j=0}^{n-1} (j - E_n[G])^2 \cdot g_n(j) = H_n - H_{n,2}, \tag{20}$$

where $H_n = \sum_{i=1}^{n} \frac{1}{i}$ and $H_{n,2} = \sum_{i=1}^{n} \frac{1}{i^2}$.

The proof of Corollary 3 is given in [18]. From Corollary 3, we have some interesting observations. Since H_n is $O(1 + \ln n)$ and $H_{\infty,2} = \zeta(2) = \frac{\pi^2}{6} \approx 1.645$ is the Riemann zeta function of 2 [13], both $\mathrm{E}_n[G]$ and $\mathrm{Var}_n[G]$ are $O(1 + \ln n)$. This indicates that the average path length of the worm tree (*i.e.*, $\mathrm{E}_n[G]$) increases approximately logarithmically with n. Moreover, when $n \to \infty$, $\lim_{n\to\infty} \mathrm{E}_n[G] - \ln n = \gamma - 1$, and $\lim_{n\to\infty} \mathrm{Var}_n[G] - \ln n = \gamma - \zeta(2)$, where $\gamma \approx 0.577$ is the Euler-Mascheroni constant [9]. Therefore, when n is large, $\mathrm{E}_n[G] \approx \mathrm{Var}_n[G]$. Furthermore, we can use Theorem 3 to obtain a closed-form approximation to $g_n(j)$ as follows.

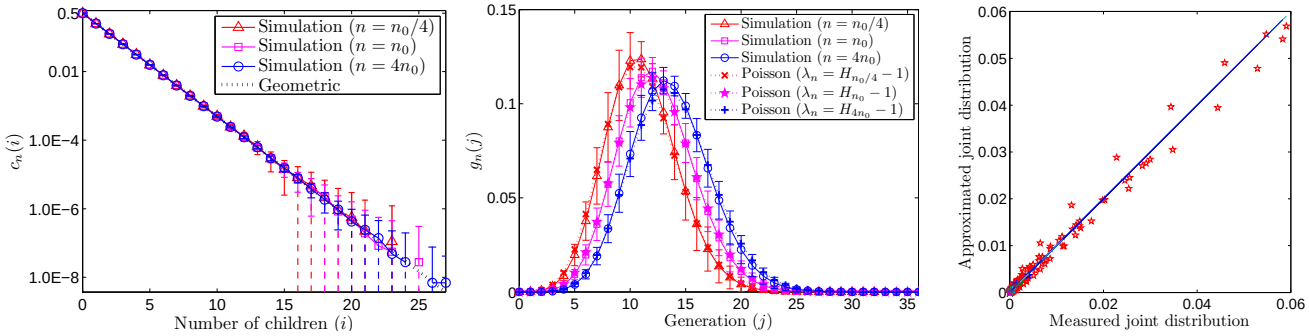
Corollary 4 *When n is very large, the generation distribution $g_n(j)$ can be approximated by a Poisson distribution with parameter $\lambda_n = E_n[G] = H_n - 1$. That is,*

$$g_n(j) \approx \frac{\lambda_n^j}{j!} e^{-\lambda_n}, \quad 0 \leq j \leq n-1. \tag{21}$$

The proof of Corollary 4 is given in [18]. Fig. 2(b) verifies Corollary 4, showing $g_n(j)$ obtained from Theorem 3 when $n = 1000$, 2000, 5000, and 20000, as well as the Poisson distribution with parameter $\mathrm{E}_n[G]$. It can be seen that when n is large, the Poisson distribution fits the generation distribution closely.

3.4 Approximation to the Joint Distribution

Finally, we derive a closed-form approximation to the joint distribution $p_n(i,j)$. From Equation (9), we can see



(a) Number of children. (b) Generation. (c) Joint distribution ($n = n_0$).

Figure 3: Simulating the infection structure of the Code Red v2 worm ($n_0 = 360,000$).

that when $n \to \infty$, $p_n(i,j) = p_{n-1}(i,j)$, which yields

$$p_n(i,j) = \tfrac{1}{2} p_n(i-1,j). \tag{22}$$

Hence, we can obtain

$$p_n(i,j) = \left(\tfrac{1}{2}\right)^i p_n(0,j) \approx \left(\tfrac{1}{2}\right)^{i+1} g_n(j). \tag{23}$$

Since when n is very large, $g_n(j)$ follows closely the Poisson distribution as in Corollary 4,

$$p_n(i,j) \approx \left(\tfrac{1}{2}\right)^{i+1} \cdot \tfrac{\lambda_n^j}{j!} e^{-\lambda_n}, \quad 0 \le i,j \le n-1, \tag{24}$$

where $\lambda_n = H_n - 1$. The above derivation also shows that when n is very large, the number of children and the generation are almost independent random variables.

Fig. 2(c) shows the parity plot of the approximation to the joint distribution when $n = 2000$. In the figure, the x-axis is the actual $p_n(i,j)$ obtained from Theorem 1, and the y-axis is the approximated $p_n(i,j)$ from Equation (24), where $0 \le i,j \le 30$. It can be seen that most points are on or near the diagonal line, indicating that the approximation to the joint distribution is reasonable.

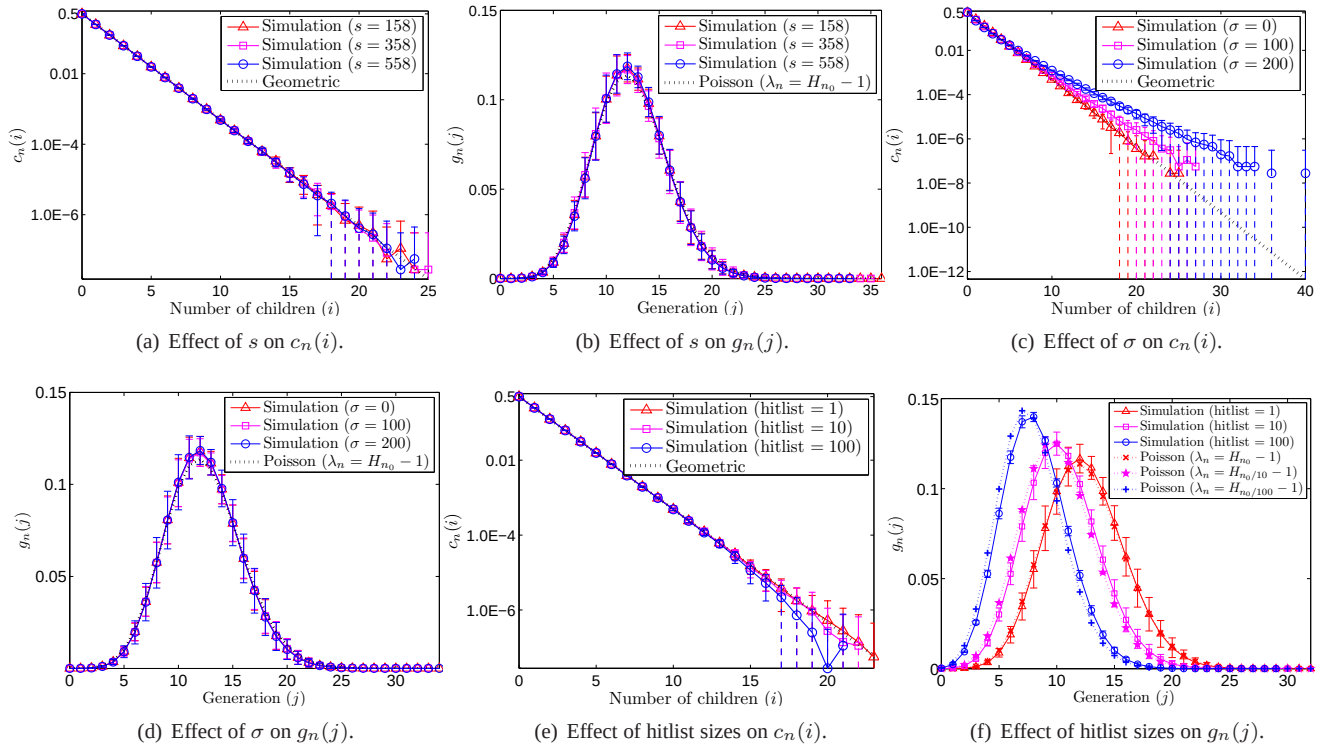
4 Simulations and Verification

In this section, we study the worm infection structure through simulations. As far as we know, there is no publicly available data to show the real worm tree and verify our analytical results. Moreover, real experiments in a controlled environment are impractical for this study, since the closed-form approximations are derived based on the assumption that the number of nodes is very large. Therefore, we apply simulations. Specifically, we first simulate the infection structure of the Code Red v2 worm and then study the effects of important parameters on the worm tree.

4.1 Code Red v2 Worm Verification

We simulate the propagation of the Code Red v2 worm by using and extending the simulator in [22]. Note that the simulator is not based on any mathematical model. Instead, the simulator considers a discrete-time system and mimics the random-scanning behavior of infected hosts in real world scenarios through a random number generator. Moreover, the parameter setting is based on the Code Red v2 worm's characteristics. For example, the vulnerable population is $n_0 = 360,000$, and a newly infected host is assigned with a scanning rate of 358 scans/min. Detailed information about how the parameters are chosen can be found in Section VII of [23]. We then extend the simulator to track the worm infection structure by adding the information of the number of children and the generation to each infected host. Moreover, we set the discrete time unit to 20 seconds and start our simulation at time tick 0 with patient zero. Note that we remove the assumption used in the sequential growth model that no two hosts are compromised at the same time. That is, multiple hosts can be compromised at one time tick. Moreover, all new victims of the current time tick start scanning at the next time tick. The simulation results (mean $\pm$ standard deviation) are obtained from 100 independent runs with different seeds and are presented in Fig. 3.

Fig. 3(a) shows the distribution of the number of children, comparing the simulation results of $c_n(i)$ for $n = n_0/4$, n_0, and $4n_0$ with the geometric distribution obtained from Corollary 2. Note that the y-axis uses the log-scale. The vertical dotted line represents the standard deviation that goes into the negative territory. It can be seen that the distribution of the number of children can be well approximated by the geometric distribution with parameter 0.5. This implies that $c_n(i)$ decreases approximately exponentially with a decay constant of $\ln 2$. Specifically, in all three cases, on average 50.0% of the infected hosts do not have children, about 98.4% of them have no more than five children, and 0.1% of them have no less than ten children. We also calculate the expectation and the variance of the number of children from the simulation and find that they are identical to the analyti-



(a) Effect of s on $c_n(i)$. (b) Effect of s on $g_n(j)$. (c) Effect of σ on $c_n(i)$.

(d) Effect of σ on $g_n(j)$. (e) Effect of hitlist sizes on $c_n(i)$. (f) Effect of hitlist sizes on $g_n(j)$.

Figure 4: Effects of the scanning rate, the scanning rate standard deviation, and the hitlist size on $c_n(i)$ and $g_n(j)$ ($n_0 = 360,000$).

cal results obtained from Corollary 1. Fig. 3(b) demonstrates the generation distribution, comparing the simulation results of $g_n(j)$ for $n = n_0/4$, n_0, and $4n_0$ with the Poisson distributions with parameter $\mathrm{E}_n[G] = H_n - 1$ obtained from Corollary 4. It can be seen that the simulation results of $g_n(j)$ closely follow the Poisson distributions for all three cases. Hence, simulation results verify that the average path length of the worm tree increases approximately logarithmically with the total number of infected hosts. Moreover, we also compute the expectation and the variance of the generation in simulations and verify the analytical results in Corollary 3. Fig. 3(c) compares the measured joint distribution from simulations with the approximated joint distribution from Equation (24) by using the parity plot. It can be seen that most points are on or near the diagonal line, indicating that the approximation works well.

4.2 Effects of Worm Parameters

Next, we extend our simulator to examine the effects of three important parameters of worm propagation on the worm tree: the scanning rate, the scanning rate standard deviation, and the hitlist size. When a parameter is studied and varied, we set other parameters to the parameters of the Code Red v2 worm as used in Section 4.1. The simulation results are obtained from 100 independent simulation runs and are shown in Fig. 4.

Fig.s 4(a) and (b) show the effect of varying the scanning rate s (scans/min) from 158 to 558 on the distributions of the number of children and the generation. Here, the scanning rate is set to a fixed value for every infected host, *i.e.*, the scanning rate standard deviation is 0. The figures also plot the geometric distribution with parameter 0.5 and the Poisson distribution with parameter $H_{n_0} - 1$ for reference. It can be seen that the scanning rate does not affect the worm tree structure.

Fig.s 4(c) and (d) demonstrate the effect of the variation of the scanning rates among different hosts (*i.e.*, σ). In our simulation, a newly infected host is assigned with a scanning rate (scans/min) from a normal distribution $N(358, \sigma^2)$. The figures show the simulation results when $\sigma = 0$, 100, and 200. It can be seen that while the scanning rate standard derivation σ has no effect on the generation distribution, it does affect the distribution of the number of children. Specifically, when σ increases, the tail of $c_n(i)$ moves upward from the geometric distribution with parameter 0.5. This is because when σ becomes larger, the variation of the scanning rate among infected hosts is greater. That is, there are more hosts with high scanning rates and also more hosts with low scanning rates. As a result, those hosts with high scanning rates tend to infect a large number of hosts, making

the tail of $c_n(i)$ move upward. However, it is also observed that when σ is not very large (the case for real worms), the geometric distribution with parameter 0.5 is still a good approximation.

In Fig.s 4(e) and (f), we show the effect of the hitlist size on the worm tree. As pointed out in Section 2, when the hitlist size is greater than 1, the underlying infection topology is a worm forest with the number of trees equal to the hitlist size. Moreover, in a worm forest, it is intuitive that each tree is a smaller version of the single worm tree of hitlist size 1 and has fewer nodes. Hence, it is not surprising to see that in Fig. 4(f), the generation distribution moves leftward when the hitlist size increases. However, the generation distribution can still be well approximated by the Poisson distribution with parameter $H_{n_h} - 1$, where n_h is the average number of nodes in a tree. Moreover, since in each tree the distribution of the number of children can be approximated by the geometric distribution with parameter 0.5, in the worm forest $c_n(i)$ still follows closely the same distribution.

5 Conclusions

In this paper, we attempt to capture the key characteristics of the Internet worm infection family tree. We have analyzed the infection tree formed by a wide class of worms such as random-scanning worms, routable-scanning worms, importance-scanning worms, OPT-STATIC worms, and SUBOPT-STATIC worms. Through both mathematical analysis and simulation, we have shown that the number of children asymptotically has a geometric distribution with parameter 0.5; and the generation closely follows a Poisson distribution with parameter $\mathrm{E}_n[G]$ (*i.e.,* $H_n - 1$).

As part of our ongoing work, we plan to relax our assumptions to include more worm dynamics and apply our observations to botnets. For example, we are studying the infection structure of localized scanning [18], and effect of user defenses and re-infection on the worm tree [19]. Moreover, based on our observations, we are developing methods for detecting bots and studying potential countermeasures for a botnet (*e.g.,* Conficker C) that uses scan-based peer discovery to form a P2P-based botnet [18].

References

[1] ALBERT, R., AND BARABÁSI, A.-L. Statistical Mechanics of Complex Networks. *Review of Modern Physics 74* (2002), 47–97.

[2] BARABÁSI, A.-L., AND ALBERT, R. Emergence of Scaling in Random Networks. *Science 286* (Oct. 1999), 509–512.

[3] CAIDA. Conficker/Conflicker/Downadup as seen from the UCSD Network Telescope. [Online]. Available: `http://www.caida.org/research/security/ms08-067/conficker.xml`.

[4] CHEN, Z., GAO, L., AND KWIAT, K. Modeling the Spread of Active Worms. In *Proc. IEEE INFOCOM* (Apr. 2003).

[5] CHEN, Z., AND JI, C. Optimal Worm-Scanning Method Using Vulnerable-Host Distributions. *International Journal of Security and Networks (IJSN): Special Issue on Computer and Network Security 2*, 1/2 (2007), 71 – 80.

[6] CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., STEIN, C. Introduction to Algorithms (Second Edition). *The MIT Press and McGraw-Hill* (2002).

[7] DAGON, D., GU, G., LEE, C., AND LEE, W. A Taxonomy of Botnet Structures. In *Proc. 23 Annual Computer Security Applications Conference (ACSAC'07)* (Dec. 2007).

[8] DAGON, D., ZOU, C. C., AND LEE, W. Modeling Botnet Propagation Using Time Zones. In *Proc. NDSS* (Feb. 2006).

[9] HAVIL, J. Gamma: Exploring Euler's Constant. *Princeton University Press* (2003).

[10] LI, Z., GOYAL, A., CHEN, Y., AND PAXSON, V. Automating Analysis of Large-Scale Botnet Probing Events. In *Proc. ACM Symposium on Information, Computer and Communication Security (ASIACCS'09)* (Mar. 2009).

[11] PORRAS, P., SAIDI, H., AND YEGNESWARAN, V. Conficker C P2P Protocol and Implementation. *SRI International Technical Report* (Sept. 2009).

[12] STANIFORD, S., PAXSON, V., AND WEAVER, N. How to 0wn the Internet in your spare time. In *Proc. 11th USENIX Security Symposium (Security'02)* (Aug. 2002).

[13] TITCHMARSH, E. C. The Theory of the Riemann Zeta Function. *Oxford University Press* (1986).

[14] VOGT, R., AYCOCK, J., AND M. JACOBSON, J. Army of Botnets. In *Proc. NDSS* (Feb. 2007).

[15] VOJNOVIC, M., AND GANESH, A. J. On the Race of Worms, Alerts and Patches. *IEEE/ACM Transactions on Networking 16*, 5 (Oct. 2008), 1066–1079.

[16] VOJNOVIC, M., GUPTA, V., KARAGIANNIS, T., AND GKANTSIDIS, C. Sampling Strategies for Epidemic-Style Information Dissemination. *to appear in IEEE/ACM Transactions on Networking*.

[17] WANG, P., SPARKS, S., AND ZOU, C. C. An Advanced Hybrid Peer-to-Peer Botnet. *IEEE Transactions on Dependable and Secure Computing 7*, 2 (Apr.-Jun. 2010), 113–127.

[18] WANG, Q., CHEN, Z., AND CHEN, C. Charaterizing Internet Worm Infection Structure. Preprint. [Online]. Available: `http://arxiv.org/abs/1001.1195`.

[19] WANG, Q., CHEN, Z., CHEN, C., AND PISSINOU, N. On the Robustness of the Botnet Topology Formed by Worm Infection. In *Proc. IEEE GLOBECOM* (Dec. 2010).

[20] XIA, J., VANGALA, S., J. WU, L. G., AND KWIAT, K. Effective Worm Detection for Various Scan Techniques. *Journal of Computer Security 14*, 4 (2006), 359 – 387.

[21] XIE, Y., SEKAR, V., MALTZ, D. A., REITER, M. K., AND ZHANG, H. Worm Origin Identification Using Random Walks. In *Proc. IEEE Symposium on Security and Privacy* (May 2005).

[22] ZOU, C. C. Internet Worm Propagation Simulator. [Online]. Available: `http://www.cs.ucf.edu/~czou/research/wormSimulation/simulator-codered-100run.cpp`.

[23] ZOU, C. C., GONG, W., TOWSLEY, D., AND GAO, L. The Monitoring and Early Detection of Internet Worms. *IEEE/ACM Transactions on Networking 13*, 5 (Oct. 2005), 967–974.

[24] ZOU, C. C., TOWSLEY, D., AND GONG, W. On the Performance of Internet Worm Scanning Strategies. *Elsevier Journal of Performance Evaluation 63*, 7 (Jul. 2006), 700–723.