Intelligent Worms: Searching for Preys

By Zesheng Chen and Chuanyi Ji

ABOUT THE AUTHORS. Zesheng Chen is currently a Ph.D. Candidate in the Communication Networks and Machine Learning Group at the School of Electrical and Computer Engineering, Georgia Institute of Technology, advised by Professor Chuanyi Ji. Chen's research interests focus on network security, especially modeling the spread of malware on networks, the performance of distributed detection systems, the effectiveness of defense systems, and the performance evaluation of communication networks.

Chuanyi Ji is an Associate Professor in the School of Electrical and Computer Engineering, Georgia Institute of Technology. Her research lies in both basic and applied areas of networking and machine learning, seeking to understand and investigate fundamental issues of network management and security, to develop engineering solutions for managing and controlling heterogeneous and large networks, to develop and apply algorithmic and analytical approaches in the area of machine learning, statistics and information theory.

Internet worms have been a persistent security threat in recent years since the Morris worm arose in 1988. After the Code Red and Nimda worms were released into the Internet in 2001, the Slammer worm was unleashed with a 376-byte User Datagram Protocol (UDP) packet and infected at least 160,000 computers worldwide on January 25, 2003. Later, the Blaster and Witty worms flooded the Internet in 2003 and 2004, respectively. These active worms caused large parts of the Internet to be temporarily inaccessible, costing both public and private sectors millions of dollars. The frequency and virulence of active-worm outbreaks have been increasing dramatically in the last few years, presenting a significant threat to today's Internet. In this article, we review the prey-searching methods that worms use currently, and may potentially exploit in the future. While reviewing what has been used by worms is doable, predicting what worms may use seems to be prohibitive: There would be million ways for active worms to attack the Internet. We show how mathematics has been playing an important role in providing both a guidance and methodology in studying current and futuristic worm attacks. In particular, we outline how mathematical tools (e.g., epidemic model, statistics, machine learning, and game theory) can be applied in this area.

Self-propagation is a key characteristic of an Internet worm. Using self-propagating malicious code, active worms can spread rapidly by infecting computer systems and by using infected hosts to disseminate the worms in an automated fashion. For example, when a worm is released into the Internet, it begins with a single host and scans randomly for other vulnerable machines (i.e., preys). When a prey is found, the worm sends out a probe to infect the target. After a new host is compromised, the worm transfers a copy of itself to this host. This new host then begins to run the worm and infect other targets. The Sapphire worm combined all these steps into one. That is, the Sapphire worm used a single UDP packet to probe, compromise, and spread the worm to targets.

Worms' designers, attackers and defenders, developed different prey-searching algorithms such as random, localized, topological, and sequential scanning. *Random scanning* is used by such well-known worms as Code Red v2 and Slammer. A worm that employs random scanning selects target IP addresses at random. Therefore, every vulnerable machine is equally likely to be infected by a worm scan. *Localized scanning* is used by Code Red II and Nimda worms, which preferentially scan for hosts in the "local" address space. For example, an infected host can put more emphasis on searching for preys from local hosts, e.g., in the same sub-network (or subnet). An intuition for such a strategy is that vulnerable hosts are clustered, and localized scanning can rapidly compromise all local vulnerable hosts. *Topological scanning* is used by Morris and Secure SHell (SSH) worms. The worm relies on the topological information contained in the victim machine in order to locate new targets. The information may include routing tables, email addresses, a list of peers, and Uniform Resource Locations (URLs). Such a scanning method is analogous to the spread pattern of biological viruses. *Sequential scanning*, where IP addresses are scanned sequentially, is used by the Blaster worm. After the worm compromises a vulnerable host, it will check the host with an IP address near this vulnerable host.

As the above worms have been used by attackers, intelligent worms have been considered by researchers. These worms are futuristic and studied for their virulence as "what-if" scenarios. For example, Nicholas Weaver presented the *hitlist scanning* idea to speed up the spread of worms at an initial stage. There potentially vulnerable machines are gathered into a hitlist beforehand, and targeted first when the worm is released. An extreme example of hitlist-scanning is provided by *flash worms*, where IP addresses of all vulnerable machines are known in advance and put into the list. Flash worms are considered to be the fastest as a worst-case scenario, since every worm scan can hit a vulnerable host. One other scanning method to speed up the spread of worms is to use the information provided by Border Gateway Protocol (BGP) routing tables to reduce the space that worms scan. This scanning method is called *routable scanning*. Zou et al. designed two types of routable-scanning worms (or called routing worms). One is based on class A (x.0.0.0/8) address-allocations, and thus called "Class A routing worm". Such a worm can reduce the scanning space to 45.3% of the entire IPv4 address space. The other is based on BGP routing tables, and thus called "BGP routing worm". This kind of worm can reduce the scanning space to only about 28.6% of the entire IPv4 address space.

One other strategy that a worm can potentially employ is *importance scanning*, where a worm takes advantage of the knowledge of vulnerable-host distribution. Importance scanning is inspired by importance sampling in statistics. When worm scanning methods are considered, random scanning is equivalent to a Monte Carlo method, which samples randomly-chosen targets in IP address space. In contrast, importance scanning samples targets according to an underlying group distribution of vulnerable hosts. The division of groups can follow different criteria, such as Domain Name System (DNS) Top-Level Domains, countries, Autonomous Systems, IP prefixes in Classless Inter-Domain Routing (CIDR), first byte of IP addresses (/8 subnets), or first two bytes of IP addresses (/16 subnets). A key observation is that the vulnerable-host distributions in these groups are highly non-uniform, which can be exploited by importance-scanning strategy. For example, Figure 1 shows the web-server (port 80) distribution. To estimate the distribution of web servers, we exploited a random Uniform Resource Locator (URL) generator from UROULETTE (http://www.uroulette.com/) to collect 13,866 IP addresses of web servers on January 24, 2005. Figure 1 shows the group distribution in /8 subnets for web servers and reflects that web servers are highly-unevenly distributed in the Internet. Importance scanning concentrates on scanning groups that are more likely for worms to find vulnerable hosts. Hence, importance scanning can reduce the number of scans needed for attacking a large number of vulnerable hosts. Importantly, as there would be million ways for attackers to make use of a vulnerable host distribution, importance scanning provides a "best-case scenario" for worm attacks given such a distribution. Importance scanning thus supplies a bench mark to compare with any other scanning schemes that use a vulnerable-host distribution.

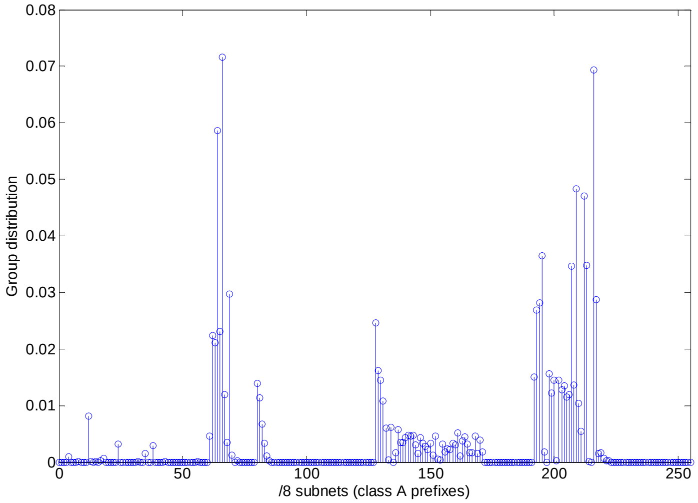

Figure 1: Uneven distribution of web servers.

The above-mentioned advanced scanning mechanisms have been developed based on such a philosophy: *The use of side information by an attacker can help a worm speed up the propagation.* In the Internet, however, it may not be easy for attackers to collect information on vulnerable hosts a priori. For example, Windows SQL database servers do not advertise their addresses. It is therefore difficult for the Slammer worm to obtain a list of vulnerable hosts or an underlying vulnerable-host distribution before the worm is released. Nevertheless, future worms are likely to become more intelligent and potentially learn about a certain knowledge (e.g., on the vulnerable-host distribution) while propagating. A self-learning worm with a simple proportion estimator has been proposed by Chen et al. that learns a vulnerable-host distribution. Such a worm then uses the importance-scanning method to speed up worm propagation using the learned distribution of vulnerable hosts. Advanced machine learning methods can be potentially applied for future intelligent worms.

How can we model the spread of worms that employ different scanning methods? Traditional biological epidemic models have been customized to study the virulence of distinct scanning methods. For example, a homogenous epidemic model is quite suitable for modeling the propagation of random-scanning worms, and has been widely used. Furthermore, epidemic models are applied to study the performance of detection/defense systems and to estimate the speed of worm propagation.

How can we defend against these smart worms? For an intelligent worm, its goal is to find the preys as quickly as possible. While for a defender of the Internet, he/she needs to detain this searching procedure. For example, a defender can potentially deploy the vulnerable hosts in the Internet in some fashion to discourage some advanced scanning methods. Therefore, there is a game between attackers and defenders. Game theory has been exploited to describe such interaction and to provide insights for countering intelligent worms.

In summary, worm-scanning methods have been widely studied. Mathematical tools (e.g., machine learning, epidemic model, and game theory) have been applied to provide guidance and methodology. Weapon race between attackers and defenders has begun, forcing us to understand the future intelligent worms more.

Reference

[1] Z. Chen and C. Ji, “A Self-Learning Worm Using Importance Scanning,” in *ACM CCS Workshop on Rapid Malcode (WORM'05)*, 2005.

[2] Z. Chen and C. Ji, “Importance-Scanning Worm Using Vulnerable-Host Distribution,” in *Proc. of IEEE Globecom 2005*, St. Louis, MO, 2005,

[3] Z. Chen, L. Gao, and K. Kwiat, “Modeling the Spread of Active Worms,” in *Proc. of INFOCOM 2003*, San Francisco, April, 2003.

[4] 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)*, 2002.

[5] N. Weaver, “Warhol Worms, the Potential for Very Fast Internet Plagues,” http://www.cs.berkeley.edu/~nweaver/warhol.html.

[6] J. Wu, S. Vangala, L. Gao, and K. Kwiat, “An Effective Architecture and Algorithm for Detecting Worms with Various Scan Techniques,” in *Network and Distributed System Security Symposium*, 2004.

[7] C. C. Zou, D. Towsley, W. Gong, and S. Cai, “Routing Worm: A Fast, Selective Attack Worm based on IP Address Information,” in *19th ACM/IEEE/SCS Workshop on Principles of Advanced and Distributed Simulation (PADS'05)*, June 1-3, Monterey, USA, 2005.

[8] C. C. Zou, D. Towsley, and W. Gong. “On the Performance of Internet Worm Scanning Strategies,” to appear in Journal of Performance Evaluation.