STOCHASTIC FEATURES OF COMPUTER VIRUSES: TOWARDS THEORETICAL ANALYSIS AND SIMULATION

HIROSHI TOYOIZUMI, YUUZO KOBAYASHI, KENTA KAIWA, JIYUN SHITOZAWA

ABSTRACT. The stochastic features of penetration attempts of computer virus are revealed. By studying real epidemic data obtained form the internet, we verify the validity of Markov model used in the virus spreads. We show some example of theoretical analysis and simulations using this assumption.

1. INTRODUCTION

Past few years, one of the main problems on the internet is malicious mobile codes or computer viruses spreading in the network [1, 2]. Slammer, Nimda, Code-red, Klez, Blaster, Sasser, Netsky... These viruses affect not only client machines infected, but also consume the precious network resources and disturb uninfected machines [3, 4].

Many researchers in both academia and private companies are struggling to provide the protection against these threat of malicious mobile codes. One of the most commonly employed solutions is to install a client-base anti-virus application to PCs. Installing an anti-viruses in your PC may be the mandatory for good citizen in internet world. At the same time, more and more administrators of local network or internet service providers install an anti-virus software to the gateway of their network. However, as we saw persistent outbreaks of viruses on the internet, these existing defenses are not sufficient. Thus, new defense strategies are proposed [5, 6, 7, 8, 9]. The effectivity of defenses should be judged based on the quantitative analysis such as theoretical analysis and simulation analysis.

There has been many such proposed mathematical models of computer viruses. For example, in [7], viruses are discussed in the setting of computer science, whereas in [10], viruses are treated as biological objects in the natural world. In [11], the authors study the real epidemic of computer viruses. Even the spread of Code-red virus is discussed using mathematical models in [12, 13]. Also computer viruses are modeled as Lotka-Voltera equation [9] and birth and death process [14].

Both analytical and simulation analysis requires the statistical features of the existing viruses. In this paper, to evaluate the stochastic features of computer virus spread, we will clarify that (1) the arrival of mails infected by viruses seen by a mail-sever is Poisson Process, and (2) the attempts of to penetrate a machine by scan is also Poisson Process. Thus, Markovian assumption can be validated to model the spread of computer viruses. We also show a couple of preliminary result about the effectiveness of the defense strategies using both analytical method and simulation analysis.

2. MODELING VIRUS SPREAD

Computer viruses are spreading among machines by various methods. There are two major ways to penetrate into your machines nowadays. The first one is by riding on emails,

Date: March 4, 2005.

and we call them mail viruses. Typical mail viruses are Netsky, Klez, Doomsday, and Lovegate, for example. Once a machine is infected by an email virus, the virus scans email addresses saved or cached on the machine, and pick one of them to the next potential victim. The virus runs their own SMTP engine to connect to an SMTP server and send an email with copy of the virus to the potential victim. If the machine is not properly maintained and/or accidentally open the contaminated attachment, the virus will infect the machine. The second type of viruses send port-scan packets to penetrate through an open port, and, we call them scan viruses. Typical scan viruses are Blaster, Nachi, and Sasser. These viruses are quicker and more infectious. Once a machine is infected, the virus on the machine send probe packets randomly to machines on the internet. They check they can exploit a specific security hall to penetrate into the new victims. If the viruses find a machine with this vulnerability, they send a copy of themselves into the new host.

To fight against these viruses effectively, we need to quantify their spreading behavior and their infectious power. Both mail viruses and scan viruses pick the next potential victims randomly and it is uncertain the next victim is going to be infected, so natural choice to analyze the virus behavior is to model them a stochastic process and use stochastic simulations. Thus, we treat both mail and scan viruses simultaneously in the stochastic model.

According to our observation of real viruses in the lab environment, a host infected by computer viruses sends emails or probe packets quite regularly, maybe at the maximum speed of the host or the network interface. Let us assume we are watching penetrating attempts (receiving infected mails or probe packets) from the internet at a machine. Let T_n be the n-th time interval of these attempts. Even though a specific infected host send those attempts regularly, those attempts seen at the observation point can be rare events. Moreover, viruses on the internet act independently each other. Thus, according to weak law of small numbers [15], $\{T_n\}$, the external attempts to penetrate into a machine, can be considered to be a Poisson Process. This assumption towards external penetrating attempts will be checked in the following section.

3. Observed Stochastic Features in Computer Viruses

In order to validate the assumption of the external penetration attempts is Poisson process, we check the inter-arrival time of mail viruses and scan viruses. We observed the arrival of mail viruses at the mail server in University of Aizu. At the same time, we observed a machine which has a global IP address and check the well-known attack of existing scan viruses, such as Sasser and Blaster. We record the penetrating attempts of these viruses and derive their interval of arrivals T_n. Figure 1 and 2 displaying the histogram of interval of virus penetration attemps. Table 1 shows the detail about the observation including the mean and the standard variation of T_n. The date is picked because at that date we observed the most viruses.

TABLE 1. Mail and Scan Viruses

Name	Observation Period	$E[T_n]$	$\sigma[T_n]$
Swen.A	2003/09/19 03:13:31 - 09/22 23:40:52	311.386	581.19
Mimail.R	2004/01/27 08:03:23 - 01/30 14:59:48	98.2059	115.75
Lovegate.F	2003/07/03 13:24:08 - 07/03 13:37:54	1.61004	1.26887
Sasser	2004/11/26 00:00:00 -11/27 00:00:00	54.1465	77.2528
Blaster	2004/11/26 00:00:00 -11/27 00:00:00	41.9356	56.3985

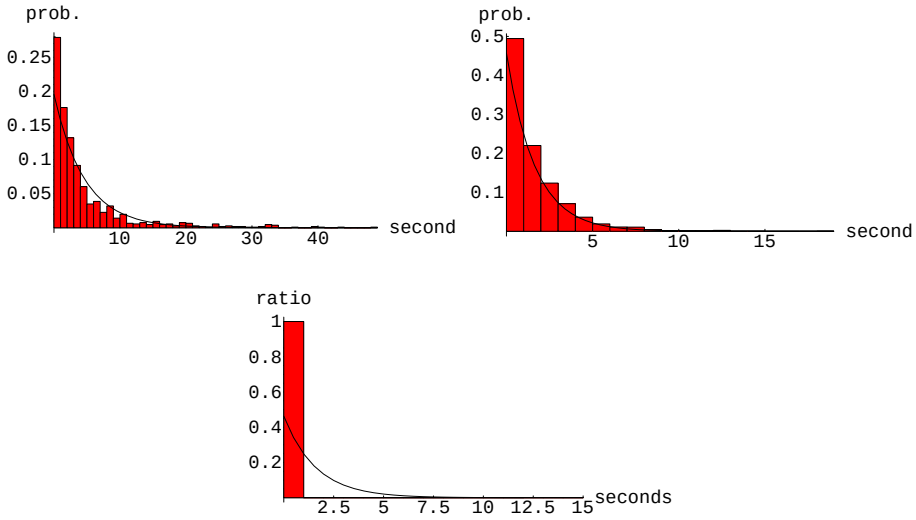

FIGURE 1. Interarrival time distribution of mails infected by mail viruses, Swen, Mimal and Lovegate, respectively. The infected mails are detected at the gateway antivirus software of U. of Aizu. The continuous lines shown in the graphs are the theoretical exponential distribution with the same mean as the observed intervals.

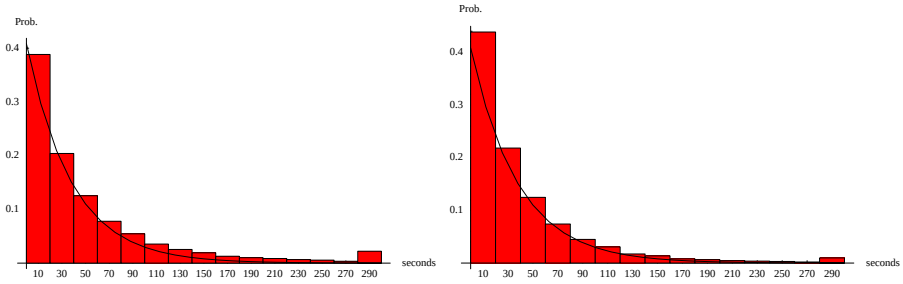

FIGURE 2. Intearrival time distribution of port-scans attempted to a machine on the internet. The left hand side shows the virus called Sasser, while the right one is Blaster. The continuous lines shown in the graphs are the theoretical exponential distribution with the same mean as the observed intervals.

As you can easily seen in the graphs, the inter-arrival times of the mail viruses such as Swen and Mimail, as well as the scan viruses such as Sasser and Blaster, are well-approximated by exponential distribution. On the other hand, the inter-arrival time of the virus Lovegate shows the significant difference from other two viruses. This is because the virus itself comes from inside. The infected machine existed inside U. of Aizu domain[1].

[1]Thanks to the effort of Information Science Technology Center at U. of Aizu, these viruses are eliminated!

Thus, these graphs do not contradict our assumption that the external arrival is Poisson for both mail and scan viruses.

4. Stochastic Model of Virus Spreads

As seen in the previous sections, we can safely model the spread of both mail and scan viruses as a variant of pure birth process. We use probabilistic arguments to model the virus spread in the local network. Since Internet is large enough for us to handle the stream of virus as fluid. However, the local network is consisted by the small number of machines, so the time of the first infection is quite important and we should take into account the stochasticity. We use a special kind of birth and death processes (see for example [16, 17, 18]) for the virus spread in local network. Let us consider the birth and death process satisfying

$$P\{N(t+\Delta t) = N(t)+1 | N(t)\} = (\lambda N(t)+\nu)\Delta t + o(\Delta t).$$
$$P\{N(t+\Delta t) = N(t)-1 | N(t)\} = \mu N(t)\Delta t + o(\Delta t).$$

This process are sometimes called a birth and death process with immigration, where the infection rate is λ, the death rate is μ and the immigration rate is ν. Note that the immigration to the local network is Poisson Process as we checked in Section 3 .

Theorem 1 (Virus Spread as Birth and Death Process)**.** *Let $N(t)$ be the number of machines infected at time t, given $N(0) = 0$. Then, the distribution of $N(t)$ is given by*

$$P\{N(t) = n\} = \binom{n+\nu/\lambda-1}{\nu/\lambda-1} p^{\nu/\lambda}(1-p)^n, \tag{1}$$

where

$$p = \begin{cases} 1/(1+\lambda t), & \lambda = \mu; \\ (\lambda-\mu)/\{\lambda e^{(\lambda-\mu)t} - \mu\}, & \lambda \neq \mu. \end{cases} \tag{2}$$

In addition, the mean of $N(t)$ is given by

$$E[N(t)] = \begin{cases} \nu t, & \lambda = \mu; \\ \nu(e^{(\lambda-\mu)t}-1)/(\lambda-\mu), & \lambda \neq \mu. \end{cases}$$

Remark 1. *Letting $\mu \to \lambda$ in p for $\lambda \neq \mu$ in (2), we have $p \to 1/(1+\lambda t)$, which coincide with p for $\lambda = \mu$. Also, letting $t \to 0$ in (2), we have $p = 1$ and $P\{N(0) = 0\} = 1$, which is consistent with the initial condition.*

Proof. Although the mathematical result itself can be found in classical text book like [19], we give a proof similar to [16]. Define the moment generation function of $N(t)$ by

$$M(\theta, t) = E\left[e^{\theta N(t)}\right]. \tag{3}$$

Then, by using classical arguments of birth and death processes (see [16]), it can be found that $M(\theta,t)$ has to satisfy the following partial differential equation;

$$\frac{\partial M}{\partial t} = \left\{\lambda(e^{\theta}-1) + \mu((e^{-\theta}-1)\right\}\frac{\partial M}{\partial \theta} + \nu(e^{\theta}-1)M, \tag{4}$$

which turned out to be so-called Lagrangian partial differential equation. By solving this equation, we have

$$M(\theta,t) = \frac{(\lambda-\mu)^{\nu/\lambda}}{\left\{(\lambda e^{(\lambda-\mu)t}-\mu) + \lambda(e^{(\lambda-\mu)t}-1)e^{\theta}\right\}^{\nu/\lambda}},$$

when $\lambda \neq \mu$. And it can be rewritten by

$$P(z,t) = \sum_{n=0}^{\infty} z^n \binom{n+\nu/\lambda-1}{\nu/\lambda-1} p^{\nu/\lambda}(1-p)^n. \tag{5}$$

By checking the coefficient of z^n, we obtain the result for $\lambda \neq \mu$. Using the similar arguments when $\lambda = \mu$, we can find the result holds for $\lambda = \mu$. □

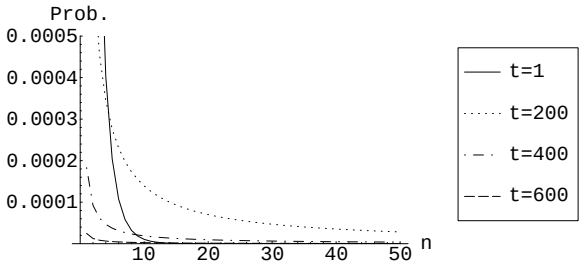

FIGURE 3. The distribution of virus population $P\{N(t) = n\}$ with $\mu = 0$. We can see that the probability mass fades away toward ∞ as $N(t)$ goes to ∞.

Figure 3 illustrates $P\{N(t) = n\}$ with $\mu = 0$. Since the virus population $N(t)$ starts with 0, the probability around $n = 0$ is quite large at small t. Eventually, $N(t)$ increases exponentially, so the probability mass fade away.

5. SIMULATION EXAMPLE

As in Section 4, we can use mathematical analysis for the simple virus spread. However, it is not common we can assume simple model for interaction of virus infection and defense strategies. In those cases, we need to use simulation as a last resort. We will show some examples of stochastic simulation of virus spreads.

Here we analyze the efficiency of IP blacklist approach [20] by using simulation of virus spreads. In IP blacklist approach, you block IP packets from those machine with suspicious behavior, such as sending packets of port scans or bulk emails, so that we avoid the outbreak of viruses. This approach is much faster than the content filtering approach which is widely used in the existing anti-virus softwares. However, even in IP blacklist approach, we have some delay to enforce the blocking from the time when we observed suspicious behavior. Thus, we need to analyze the effect of this delay to the efficiency of IP blacklist approach. As seen in Figure 4, the outbreak is warded off successfully when the delay is limited to 50 seconds.

ACKNOWLEDGEMENTS

We thank Professor Hayashi and ISTC staff at U. of Aizu for their continuous fight against computer virus and providing us the data of their effort.

REFERENCES

[1] R. A. Grimes. *Malicious Mobile Code*. O'Reilly and Associates, 2001.
[2] Nick FitzGerald et al. Virus-l comp.virus frequently asked questions (faq) v2.00. http://nwww.faqs.org/faqs/computer-virus/faq/.

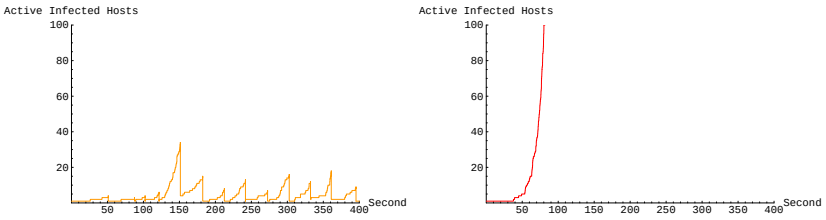

FIGURE 4. The number of active virues, The left hand side shows that the number of active virus is sustained by IP blacklist with 30-second delay, while the right one shows the outbreak of no-defense, where the infection rate $\lambda = 0.0713$

[3] Ido Dubrawsky. Effects of worms on internet routing stability. http://www.securityfocus.com/infocus/1702, 2003.

[4] BJ Premore James Cowie, Andy Ogielski and Yougu Yuan. Global routing instabilities during code red ii and nimda worm propagation.

[5] Carey Nachenberg. Computer virus-antivirus coevolution. *Commun. ACM*, Vol. 40, No. 1, pp. 46–51, 1997.

[6] Prabhat K. Singh and Arun Lakhotia. Analysis and detection of computer viruses and worms: an annotated bibliography. *SIGPLAN Not.*, Vol. 37, No. 2, pp. 29–35, 2002.

[7] Harold Thimbleby, Stuart Anderson, and Paul Cairns. A framework for modelling Trojans and computer virus infection. *The Computer Journal*, Vol. 41, No. 7, pp. 445–458, 1998.

[8] K. G. Anagnostakis, M. B. Greenwald, S. Ioannidis, A. D. Keromytis, and D. Li. A cooperative immunization system for an untrusting internet. In *Proceedings of the 11th IEEE International Conference on Networks (ICON'03)*. IEEE, October 2003.

[9] Hiroshi Toyoizumi and Atsuhi Kara. Predators: Good will mobile codes combat against computer viruses. *New Security Paradigm Workshop 2003*, pp. 11–17, 2003.

[10] S. Jones and C. White. The ipm model of computer virus management. *Computers and Security*, Vol. 9, No. 5, pp. 411–418., 1990.

[11] Jeffrey O. Kephart, Steve R. White, and David M. Chess. Computers and epidemiology. *IEEE Spectrum*, pp. 20–26, MAY 1993.

[12] Security.NL. Code red worm stats. http://www.security.nl/misc/codered-stats/, 2001.

[13] Stuart Staniford. Analysis of spread of july infestation of the code red worm. http://www.silicondefense.com/cr/.

[14] Hiroshi Toyoizumi. Performance evaluation of defense strategies against computer virus. In *Seventh INFORMS TELECOM*, pp. 131–133. INFORMS section on Telecommunications, 2003.

[15] Richard Durrett. *Probability: Theory and Examples*. Thomson Learning, 1991.

[16] Norman T.J. Bailey. *The elements of stochastic processes with applications to the natural*. Wiley Classical Library. J. Wiley, 1990.

[17] S. M. Ross. *Stochastic Processes*. John Wiley and Sons, 1996.

[18] Eric Renshaw. *Modelling Biological Populations in Space and Time*. Cambridge University Press, 1991.

[19] Bharucha-Reid. *Elements of the Theory of Markov Processes and Their Applications Processes and Their Applications*. Dover Pubns, 1960.

[20] David Moore, Colleen Shannon, Geoffrey M. Voelker, and Stefan Savage. Internet quarantine: Requirements for containing self-propagating code. In *INFOCOM*. IEEE, 2003.

GRADUATE SCHOOL OF ACCOUNTING, WASEDA UNIVERSITY, TOKYO, JAPAN 169-8050,
E-mail address: toyoizumi@waseda.jp