Statistical Inference of Computer Virus Propagation Using Non-Homogeneous Poisson Processes

Hiroyuki Okamura, Kazuya Tateishi and Tadashi Dohi
Department of Information Engineering
Graduate School of Engineering, Hiroshima University
1–4–1 Kagamiyama, Higashi-Hiroshima, 739–8527, Japan
{okamu, dohi}@rel.hiroshima-u.ac.jp

Abstract

This paper presents statistical inference of computer virus propagation using non-homogeneous Poisson processes (NHPPs). Under some mathematical assumptions, the number of infected hosts can be modeled by an NHPP. In particular, this paper applies a framework of mixed-type NHPPs to the statistical inference of periodic virus propagation. The mixed-type NHPP is defined by a superposition of NHPPs. In numerical experiments, we examine a goodness-of-fit criterion of NHPPs on fitting to real virus infection data, and discuss the effectiveness of the model-based prediction approach for computer virus propagation.

Keywords: computer virus, non-homogeneous Poisson process, mixed-type model, EM algorithm, goodness-of-fit test

1. Introduction

The Internet plays a central role in the progress of information technologies during the last two decades. Through the Internet, we are able to communicate easily with many people all over the world. On the other hand, the Internet explosion causes some serious social problems such as computer viruses and illegal accesses. In particular, activities of the computer viruses tend to get more malicious in recent years.

The computer viruses and their mechanism have been discussed for long time [7], and almost all computer engineers have warned computer users against their dangerousness. However, most people as well as computer engineers tend to underestimate the propagation effect of computer viruses. The most vivid example of damage by a computer virus would be the Code-Red virus [19] which has explosively increased. On July 2001, Code-Red infected more than 359,000 hosts on the Internet within only 14 hours [11]. This experience motivates both computer engineers and users to counter against computer viruses before their outbreaks.

In general, the computer viruses are classified into *viruses*, *worms* and *Trojan horses*. The virus, which often means a computer virus itself, is defined as the program which has self-replication but does not have self-propagation. The worm is often called *the Internet worm*. The Internet worm can make a copy for itself to other computers via the Internet. The Trojan horse is quite different from the virus and the worm. The Trojan horse does not have self-replication, that is, it needs the users' action to propagate itself. The Internet worm generally has the highest propagation ability among them, and Code-Red is categorized to the Internet worm.

To examine the propagation of computer viruses, many researchers have focused on mathematical models for virus propagation. Kephart and White [9, 10], Wierman and Marchette [21] and Sellke et al. [18] proposed deterministic or stochastic models for the Internet worm based on well-known epidemiology models. Allen and Burgin [5] and Okamura et al. [14, 15] also applied the stochastic models based on continuous-time Markov chain (CTMC) to describing the Internet-worm propagation. They considered the birth-and-death processes where each state of CTMC corresponds to the number of infected hosts.

The drawback of using birth-and-death processes is that the model requires a number of states to represent the propagation of real computer virus. This is caused by the fact that the number of infected hosts is quite large in the real world. The CTMC modeling is not always efficient to analyze the virus propagation using the empirical infection data.

In this paper, we consider a virus propagation model from the viewpoint of statistics. That is, we deal with the problem of evaluating the virus propagation based on empirical infection data. More specifically, this paper introduces non-homogeneous Poisson processes (NHPPs) to

1071-9458/07 $25.00 © 2007 IEEE
DOI 10.1109/ISSRE.2007.28

represent the number of virus-infected hosts. The NHPPs are frequently used to assess reliabilities of hardware and software from the failure occurrence data, and thus their mathematical framework is simple enough to perform the empirical data analysis. Although the idea to use the NHPPs for the virus propagation is originally found in [16], this paper enhances the NHPP-based analysis by adding a new modeling framework of NHPPs; the mixed-type NHPPs. Furthermore, we introduce the EM (expectation-maximization) algorithm [12, 13, 17] as an efficient algorithm for estimating the NHPP parameters which can be used in both the mixed and non-mixed NHPPs.

This paper is organized as follows. Section 2 introduces some related works for virus-propagation evaluation. In particular, we present deterministic models based on the differential equations which correspond to the dynamics of virus propagation. In addition, we explain the non-linear regression based on the deterministic models. Section 3 gives some modeling assumptions for the virus propagation, and derive two kinds of NHPP models: the usual NHPP model with one population and the mixed-type NHPP models. In Section 4, we present the techniques of statistical analysis based on the NHPP modeling, and particularly introduce the parameter estimation procedure and a goodness-of-fit evaluation. Section 5 is devoted to the performance evaluation of the proposed NHPP-based models with real virus infection data.

2. Related Work

2.1. Epidemiology models

In the research area of epidemiology, a large number of mathematical models have been developed to describe the propagation of individuals. This paper introduces the SIS (Susceptible-Infected-Susceptible) model. The idea behind the SIS model is to classify each host into two states: susceptible S and infected I.

Let $s(t)$ and $v(t)$ denote the numbers of susceptible and infected hosts at time t, respectively. Note that *susceptible* is regarded as *vulnerable* when we treat the model of computer viruses. The dynamics of $s(t)$ and $v(t)$ can be modeled as follows.

$$\frac{d}{dt}s(t) = -\beta s(t)v(t) + \delta v(t), \quad (1)$$

$$\frac{d}{dt}v(t) = \beta s(t)v(t) - \delta v(t), \quad (2)$$

where β (> 0) and δ (> 0) are a pairwise infection rate and a disinfection rate, respectively. Since the numbers of susceptible and infected hosts $s(t)$ and $v(t)$ are deterministic, and thus this model is called the deterministic SIS model. Assuming that the total number of hosts is finite, i.e., $K = s(t) + v(t)$, the number of infected hosts is described by the following logistic-type differential equation:

$$\frac{d}{dt}v(t) = \{\delta(R_0 - 1) - \beta v(t)\}v(t), \quad (3)$$

where $R_0 = \beta K/\delta$. If R_0 is greater than one, all the hosts are possibly infected with viruses. This implies that the number of infections converges to a certain saturation level in the steady state. This is called the endemic steady state. If R_0 is less than one, the computer virus is eventually extinct. This is called the disease-free steady state. By solving the above differential equation with respect to $v(t)$, we find that the number of infected hosts draws the following logistic curve:

$$v(t) = \frac{\delta(R_0 - 1)v_0}{\beta v_0 + \{\delta(R_0 - 1) - \beta v_0\}\exp\{-\delta(R_0 - 1)t\}}, \quad (4)$$

where $v(0) = v_0$.

The SIS models may lead to some extended models; the SIR (Susceptible-Infected-Removal) model [4, 8], the SIRS (Susceptible-Infected-Removal-Susceptible) model [5], the Predator-Prey model [20] and the Kill-Signal model [9, 10].

The regression models represent the computer virus propagation based on deterministic curves. Here we introduce the representative regression models based on Gompertz and logistic curves. These models are well known as growth models of individuals in epidemiology.

The Gompertz curve, along with the SIS-based logistic curve, can also be used to represent the virus propagation phenomenon. The Gompertz curve was originally developed as a trend curve for the population growth involving death rate of human. Let $\{x(t), t \geq 0\}$ denote the number of virus-infected hosts at time t. Then the Gompertz curve is given by the solution of the following differential equation:

$$\frac{d}{dt}x(t) = \alpha x(t)\exp(-\beta t), \quad \alpha > 0,\ \beta > 0, \quad (5)$$

where α and β indicate an infection rate per virus and a death (disinfection) rate depending on the infection age t, respectively. From Eq. (5), we obtain the well-known Gompertz curve:

$$x(t) = c\exp\{-\alpha\exp(-\beta t)\} = ca^{b^t}, \quad (6)$$

$$c > 0,\ 0 < a < 1,\ 0 < b < 1. \quad (7)$$

2.2. Regression

Before using such deterministic models as the logistic curve and the Gompertz curve for data analysis, we need to fit the models to the observed data. For this purpose, the least squares method is the most commonly used method for

the deterministic models. Let $\mathcal{D} = \{(t_1, y_1), \ldots, (t_k, y_k)\}$ be the observation at time sequence $t_1, \ldots, t_k$, where y_i $(i = 1, \ldots, k)$ denotes the cumulative number of infected hosts. The problem is to estimate model parameters for fitting the curves to the observation. To perform the least squares method, we define the sum of squared errors, or alternatively residual squared sum (RSS), as follows.

$$\mathrm{RSS}(\boldsymbol{\theta}) = \sum_{i=1}^{K} \{y_i - x(t_i; \boldsymbol{\theta})\}^2, \tag{8}$$

where $\boldsymbol{\theta}$ is the parameter vector. The estimates by the least squares method correspond to the parameters minimizing the above RSS, and thus we have to solve a simple non-linear programming problem. Since the non-linear programming is generally solved by Gauss-Newton method, the advantage of regression models is to simplify the numerical procedures for parameter estimation. On the other hand, the least squares method deals with only the information of average tendency of propagation. Therefore the regression models are not appropriate to estimate probabilistic behaviour of computer virus like outbreak and extinction.

3. NHPP Models

3.1. Single population [16]

Consider NHPPs to represent the virus infection process. We first make the following assumptions:

Assumption 1: The total number of susceptible hosts in the network is a Poisson distributed random variable N with mean ω;

$$\Pr\{N = n\} = \frac{\omega^n}{n!} \exp(-\omega), \quad n = 0, 1, \ldots. \tag{9}$$

Assumption 2: Define the indicator function which represents whether a host k is infected or not;

$$I_k = \begin{cases} 1, & \text{!JInfected!K} \\ 0, & \text{!JSusceptible!K} \end{cases} \tag{10}$$

The joint distribution of virus infections for all hosts in the network is given by

$$\Pr\{I_1 = x_1, \ldots, I_N = x_N\} = \prod_{k=1}^{N} \Pr\{I_k = x_k\}. \tag{11}$$

Assumption 1 means that the number of infected hosts is relatively small compared to the number of all hosts in the Internet. Under Assumption 1, one host is possibly connected to all the other hosts, and the infection probabilities to other hosts are uniformly distributed. On the other hand, the expression in Assumption 2 is called a product-form solution, and is well known as a steady-state probability for most practical queueing network systems (see e.g. [6]). This assumption implies that the correlation or interaction between two virus infections is vanishingly small from a macroscopic viewpoint, although the correlation and the interaction exist in the real network. In fact, it may be plausible to assume that, when a host is infected with a virus, its neighborhoods are likely to be infected with the same virus. Then a measurable interaction is observed in such a situation. However, if we deal with the vast number of hosts in the Internet, the virus infection for any two hosts would not interfere from each other. Therefore Assumption 2 could be accepted in many types of viruses.

Let $T_k, k = 1, \ldots, N$, denote the ordered infection times for susceptible hosts. Under Assumption 2, the virus infection times can be regarded as a sequence of independent and identically distributed (i.i.d.) random variables. Thus we define $F(t)$ as the cumulative distribution function (c.d.f.) of their infection times. Then the probability mass function (p.m.f.) of the number of infected hosts, provided that the number of susceptible hosts is n, is given by

$$\Pr\{X(t) = x | N = n\} = \binom{n}{x} F(t)^x \{1 - F(t)\}^{n-x}. \tag{12}$$

Furthermore, Assumption 1 yields the following Poisson p.m.f. of the number of infected hosts:

$$\Pr\{X(t) = x\} = \frac{\{\omega F(t)\}^x}{x!} \exp\{-\omega F(t)\}. \tag{13}$$

Finally the above p.m.f. is equivalent to the p.m.f. of Poisson distribution with mean $\omega F(t)$. In other words, under Assumptions 1 and 2, the number of infected hosts before time t follows an NHPP with a mean value function $\omega F(t)$. It should be noticed that the behavior of virus infection depends on only the infection time distribution $F(t)$ in this framework, and that we can reduce the problem of choosing the best propagation model to the problem of choosing the best infection time distribution.

Although we can apply any probability distribution to the infection time distribution, we suppose the following three probability distributions:

(i) Normal distribution:

$$F(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}. \tag{14}$$

(ii) Logistic distribution:

$$F(t) = \frac{e^{(t-\mu)/\phi}}{1 + e^{(t-\mu)/\phi}}. \tag{15}$$

(iii) Extreme value distribution at maximum:

$$F(t) = \exp\{-e^{-(t-\mu)/\theta}\}. \tag{16}$$

These are typical limiting probability distributions which correspond to the mean, the median and the maximum of i.i.d. random variables, respectively. This paper deals with the macroscopic behavior of virus propagation, and these probability distributions are applied to the infection time distributions. Note that, since these distributions take negative values, we have to use the normal, logistic and extreme value distributions which are truncated at origin. In particular, when we apply the truncated logistic and extreme value distributions, the mean value functions of the resulting NHPPs become the logistic and Gompertz curves, respectively [12, 13]. Also, the analogues but somewhat different approaches by using NHPPs were taken in Alhazmi and Malaiya [2, 3] and Woo et al. [22]. They mainly did not focus on the virus propagation, but on the software faults concerning the system vulnerability.

3.2. Mixed-type NHPP models

The most remarkable characteristic of real virus infection is that the number of infected hosts has two or more peaks. That is, the virus infection is periodic. In general, it is hard to represent the periodic propagation phenomenon by using the simple NHPP modeling framework in Section 3.1. This section proposes another NHPP model with multiple population, where one population indicates a group of hosts which have different infection time distributions. Then the total number of infected hosts at time t forms a mixed-type NHPP.

Recall the model assumptions in Section 3.1; (i) the size of population obeys the Poisson distribution (ii) infection times are i.i.d. random variables. Here we consider the case where the hosts are categorized into two or more groups in terms of the virus infection time. Let N_i and $F_i(\cdot)$ denote the size of the i-th population and its associated infection time distribution, respectively. Similar to Section 3.1, we assume that there is no correlation among populations. Then the p.m.f. of the number of infected hosts, provided that $N_1 = n_1, \ldots, N_m = n_m$, is given by

$$\begin{aligned} &\Pr\{X_1(t) = x_1, \ldots, X_m(t) = x_m \mid \\ &\quad N_1 = n_1, \ldots, N_m = n_m\} \\ &= \frac{\prod_{i=1}^m n_i!}{\prod_{i=1}^m x_i!(n_i - x_i)!} \prod_{i=1}^m F_i(t)^{x_i}\{1 - F_i(t)\}^{n_i - x_i}. \end{aligned} \tag{17}$$

When the size of the i-th population ($i = 1, \ldots, m$) is the Poisson distributed with mean parameter ω_i, the cumulative number of infected hosts $X(t)$ becomes the following NHPP:

$$\begin{aligned} \Pr\{X(t) = x\} = &\frac{\{\sum_{i=1}^m \omega_i F_i(t)\}^x}{x!} \\ &\times \exp\left\{-\sum_{i=1}^m \omega_i F_i(t)\right\}. \end{aligned} \tag{18}$$

Letting $\omega = \sum_{i=1}^m \omega_i$ and $\pi_i = \omega_i/\omega$, we have the mixed-type NHPP:

$$\begin{aligned} \Pr\{X(t) = x\} = &\frac{\{\omega \sum_{i=1}^m \pi_i F_i(t)\}^x}{x!} \\ &\times \exp\left\{-\omega \sum_{i=1}^m \pi_i F_i(t)\right\}. \end{aligned} \tag{19}$$

This equation implies that the NHPP model with multiple population is reduced to an NHPP model with a mixed infection time distribution

$$H(t) = \sum_{i=1}^m \pi_i F_i(t), \tag{20}$$

where π_i is called the mixture ratio. The mixed infection time distribution $H(t)$ is a multimodal distribution, and therefore we can approximately represent the periodic infection of computer viruses.

4. Statistical Analysis

4.1. Parameter estimation

When evaluating the virus propagation based on NHPP models, we need to fit them to observed virus infection data. The commonly used method to estimate parameters is the maximum likelihood (ML) estimation. The advantage of ML estimation over the other estimation methods like the least squares method is that the ML estimates have good properties for statical analysis such as an asymptotic normality. However, compared with the least squares method, the ML estimation requires numerically complex procedures.

The EM algorithms for NHPPs have been developed in [17]. The idea behind the EM algorithm for NHPPs is to regard the total size of population as hidden information. Moreover, in [17], the EM algorithms are developed for the cases where time distributions are exponential, gamma and Weibull distributions. The EM algorithms for the NHPP with logistic and extreme-value distributions are briefly discussed in [12, 13].

For brevity, this paper introduces only the EM algorithm for the NHPP with normal distribution. Let $\mathcal{D} = \{x_1, \ldots, x_K\}$ be the number of infected hosts at time sequence $t_1, \ldots, t_K$, where x_i denotes the number of infected hosts at time interval $[t_{i-1}, t_i)$.

The EM algorithm is developed under the incomplete observation. We define the infection times $T_{-\tilde{N}}, \ldots, T_{-1}, T_1, \ldots, T_N$ as complete data, where the infection time with a negative index indicates the data truncated at $t = 0$. Given $T_{-\tilde{N}}, \ldots, T_{-1}, T_1, \ldots, T_N$, the ML estimates are given by

$$\hat{\omega} = N, \tag{21}$$

$$\hat{\mu} = (\sum_{i=1}^{\tilde{N}} T_{-i} + \sum_{i=1}^{N} T_i)/(\tilde{N} + N), \tag{22}$$

$$\hat{\sigma}^2 = (\sum_{i=1}^{\tilde{N}} T_{-i}^2 + \sum_{i=1}^{N} T_i^2)/(\tilde{N} + N) - \left\{ (\sum_{i=1}^{\tilde{N}} T_{-i} + \sum_{i=1}^{N} T_i)/(\tilde{N} + N) \right\}^2. \tag{23}$$

Since we cannot observe all the information about infection times, the EM algorithm calculates the expectation for unobserved values. Some formulas to compute the expectation are derived in [17]. Finally, by applying the formulas on the expectation, we have the EM algorithm for the NHPP models with normal distribution as Figure 1. In the figure, $\phi(\cdot)$ and $\Phi(\cdot)$ are respective probability density function (p.d.f.) and c.d.f. for the standard normal distribution. The functions $\Phi^{(1)}(\cdot)$ and $\Phi^{(2)}(\cdot)$ are given by

$$\begin{aligned} \overline{\Phi}^{(1)}(z) &= \frac{1}{\sigma} \int_{\sigma z+\mu}^{\infty} u\phi(u)du \\ &= \sigma\phi(z) + \mu\overline{\Phi}(z), \end{aligned} \tag{24}$$

$$\begin{aligned} \overline{\Phi}^{(2)}(z) &= \frac{1}{\sigma} \int_{\sigma z+\mu}^{\infty} u^2\phi(u)du \\ &= (\sigma^2 z + 2\mu\sigma)\phi(z) + (\sigma^2 + \mu^2)\overline{\Phi}(z). \end{aligned} \tag{25}$$

Also, the EM algorithm presented in [17] can directly be applied to estimating the parameters of the mixed-type NHPPs. For example, when we treat the mixed-type NHPP with normal distributions, the EM algorithm can be obtained by combining the algorithm shown in Fig. 1 and the algorithm for the mixture ratios proposed in [17].

EM algorithm for NHPP with normal distribution

1: Determine the starting values; $\tilde{\omega}$, μ and σ.

2: **REPEAT**

3: **E-step:**

3.1: For $i = 1, \ldots, k$, and $u = 1, 2$, compute

$$z_0 = -\mu/\sigma,$$

$$z_i = (t_i - \mu)/\sigma,$$

$$\tau_i^{(u)} = \frac{\overline{\Phi}^{(u)}(z_{i-1}) - \overline{\Phi}^{(u)}(z_i)}{\overline{\Phi}(z_{i-1}) - \overline{\Phi}(z_i)},$$

where $\overline{\Phi}(\cdot) = 1 - \Phi(\cdot)$ and $\overline{\Phi}^{(u)}$, $u = 1, 2$, are defined as Eqs. (24) and (25).

3.2: Compute the expected values

$$\begin{aligned} N &:= \sum_{i=1}^{k}(x_i + b_i) + \tilde{\omega}\{\Phi(z_0) + \overline{\Phi}(z_k)\} \\ T^{(1)} &:= \sum_{i=1}^{k}(x_i\tau_i^{(1)} + b_i t_i) + \tilde{\omega}\{\Phi^{(1)}(z_0) + \overline{\Phi}^{(1)}(z_k)\} \\ T^{(2)} &:= \sum_{i=1}^{k}\{x_i\tau_i^{(2)} + b_i t_i^2\} + \tilde{\omega}\{\Phi^{(2)}(z_0) + \overline{\Phi}^{(2)}(z_k)\}, \end{aligned}$$

where $\Phi^{(u)}(t) = \overline{\Phi}^{(u)}(-\infty) - \overline{\Phi}^{(u)}(t)$.

4: **M-step:** Update the parameters

$$\begin{aligned} \tilde{\omega} &:= N \\ \mu &:= T^{(1)}/N \\ \sigma &:= \sqrt{T^{(2)}/N - (T^{(1)}/N)^2} \end{aligned}$$

5: **UNTIL** satisfying the termination condition.

6: Let $\omega := \tilde{\omega}\overline{\Phi}(z_0)$.

Figure 1. Pseudo-code of EM algorithm for NHPP with normal distribution.

4.2. Model selection

As mentioned before, the model selection of NHPP models is reduced to choosing an appropriate infection time distribution. In general, we utilize information criteria to examine goodness-of-fit of estimated distributions. Almost all the information criteria consist of the maximum log-likelihood in ML estimation and a penalty term concerning the number of free parameters. The well-known information criterion is AIC (Akaike's Information Criterion) [1] which is defined as follows.

$$\mathrm{AIC}(p) = -2(\text{maximum log-likelihood}) + 2p, \tag{26}$$

where p denotes the number of free parameters. Since the information criteria are the estimates of distance between the estimated models and the true model, the model with the least information criterion should be selected as the best model which can fit to the virus infection data.

On the other hand, the Kolmogorov-Smirnov (KS) test is available to statistically decide whether the empirical infection data obey the estimated NHPP or not. Given the virus infection data $\mathcal{D} = \{(t_1, y_1), \ldots, (t_K, y_K)\}$ and the estimated mean value function $\omega F(t)$, the KS test statistic is

then computed as

$$D = \max_{1 \le i \le K-1} \{D_i\}, \quad (27)$$

$$D_i = \max\left\{ \left| \frac{F(t_i)}{F(t_K)} - \frac{y_i}{y_K} \right|, \left| \frac{F(t_i)}{F(t_K)} - \frac{y_{i-1}}{y_K} \right| \right\}, \quad (28)$$

where y_i is the cumulative number of infected hosts at the i-th observation. If the statistic D is greater than a critical value $D_{K;\alpha}$, which depends on the degree of freedom for the infection data K, with given significance level α, we can reject the null hypothesis that the estimated NHPP can fit to the observed data.

5. Numerical Experiments

5.1. Data Fitting

In this section, we perform a statistical test for the proposed NHPP models to the empirical virus infection data. The data sets used in the experiments are collected for two years since October 1st, 2004. The data consist of the number of infected hosts in each day[1].

The data are collected for 116 kinds of viruses, and the viruses are categorized by the following 11 classes:

BKDR: A program which sets a backdoor on PCs. The term backdoor often refers to backdoor programs - applications that open computers for access by remote systems. These programs typically respond to specially-built client programs, but can be designed to respond to legitimate messaging applications.

HTML: A program which is written using HTML and performs unexpected or unauthorized, but always malicious, actions.

JAVA: A program which is written using Java, alternatively Java applet. Java applets allow Web developers to create interactive, dynamic Web pages with broader functionality. They are small, portable Java programs embedded in HTML pages and can run automatically when the pages are viewed.

JS: A program which is written using JavaScript and performs unexpected or unauthorized actions.

PE: Portable Executable (PE) is the standard Win32 executable file format. File infectors that infect 32-bit Windows executables.

TROJ: A Trojan performs a malicious action, but has no replication abilities. The Trojan may arrive as a seemingly harmless file or application, but actually has some hidden malicious intent within its code.

VBS: A program which is written using Visual Basic Script and performs unexpected or unauthorized, but always malicious, actions.

WORM: A computer worm is a self-contained program (or set of programs) that is able to spread functional copies of itself or its segments to other computer systems. The propagation usually takes place via network connections or email attachments.

ADW: Adware is software that displays advertising banners on Web browsers such as Internet Explorer and Mozilla. Adware programs often create unwanted effects on a system. In some instances, the degradation in either network connection or system performance.

HKTL: A program which generally crack or break computer and network security measures (Hacking tool). Hacking tools have different capabilities depending on the systems they have been designed to penetrate.

SPYW: A spyware is a program that monitors and gathers user information for different purposes. Spyware programs usually run in the background, with their activities transparent to most users. Spyware may also cause a general degradation in both network connection and system performance.

In general, different viruses have different ways of infection. This paper further classifies the viruses into three categories in terms of their infection paths.

Worm type: The worm type propagates self-contains via the network. According to the above virus categories, the class WORM belongs to this type in terms of infection paths.

Virus type: The virus type propagates itself by the execution of the infected program. The PE has such a way of infection.

Trojan horse type: The Trojan horse does not make a copy of itself. Instead of propagating itself, it makes some malicious tricks like backdoor on client PCs. Therefore the way of infection relies on user's activities. According to the above classification, the classes except for WORM and PE are Trojan horse type.

To investigate the goodness-of-fit for NHPP models, we assume that the virus infection time distribution is normal distribution (NOR), logistic distribution (LOG), extreme value distribution (EXT) or exponential distribution (EXP).

[1] Trend Micro, Inc.: World Virus Tracking Center http://wtc.trendmicro.com/

Table 1. KS Statistics (BKDR).

NAME	DAYS	NOR	LOG	EXT	EXP
AGENT.AC	5	0.252	0.252	0.252	0.457
BDI.A	91	0.075	0.068	0.056	0.141*
BERBEW.Q	74	0.081	0.082	0.082	0.082
IROFFER.D	19	0.140	0.130	0.127	0.153
JEEMP.C	74	0.066	0.066	0.067	0.075
MAGICON.A	31	0.103	0.102	0.108	0.117
PRORAT.A	86	0.076	0.067	0.078	0.288**
RULEDOR.E	87	0.114	0.116	0.116	0.115
SANDBOX.A	171	0.079	0.075	0.076	0.075
SDBOT.DP	72	0.110	0.109	0.109	0.108
SDBOT.GEN	75	0.118	0.112	0.112	0.111

Table 2. KS Statistics (HTML).

NAME	DAYS	NOR	LOG	EXT	EXP
BAGLE.AC	39	0.072	0.071	0.078	0.110
BYTEVER.A	44	0.068	0.069	0.069	0.069
CITIFRAUD.C	160	0.065	0.065	0.064	0.071
DLOADER.UW	26	0.099	0.100	0.100	0.100
MHTREDIR.A	156	0.120*	0.097	0.148**	0.191**
MHTREDIR.G	52	0.140	0.139	0.146	0.150
MHTREDIR.H	21	0.101	0.105	0.104	0.104
REDIR.A	182	0.164**	0.150**	0.150**	0.149**
STARTPAGE.C	26	0.154	0.160	0.160	0.161
SUNFRAUD.B	109	0.146*	0.137*	0.137*	0.136*
WAMUFRAUD.A	76	0.093	0.094	0.097	0.106
WINSHOW.A	60	0.080	0.081	0.112	0.235**

Table 3. KS Statistics (JAVA).

NAME	DAYS	NOR	LOG	EXT	EXP
BYTEVER.A-1	84	0.059	0.058	0.059	0.169*
BYTEVER.B	142	0.076	0.073	0.079	0.174**
BYTEVER.C	174	0.050	0.050	0.053	0.060
BYTEVER.Q	77	0.089	0.089	0.090	0.096
FEMAD.B	165	0.130**	0.126*	0.126*	0.126*
NOCHEAT.A	142	0.078	0.079	0.081	0.085
OPENSTR.A	123	0.051	0.049	0.049	0.059

Table 4. KS Statistics (JS).

NAME	DAYS	NOR	LOG	EXT	EXP
BAIDU.A	11	0.395*	0.378	0.378	0.378
DIALOGARG.A	88	0.069	0.071	0.074	0.095
DLOADER.J	14	0.142	0.145	0.145	0.147
FORTNIGHT.M	197	0.045	0.042	0.046	0.107*
INOR.M	44	0.090	0.095	0.097	0.103
NOCLOSE.AA	12	0.161	0.161	0.161	0.158
NOCLOSE.I	19	0.131	0.138	0.138	0.139
NOCLOSE.RY	14	0.178	0.172	0.165	0.196
SMALL.D	68	0.076	0.081	0.088	0.103
ZEROLIN.A	49	0.092	0.083	0.083	0.082

Table 5. KS Statistics (PE).

NAME	DAYS	NOR	LOG	EXT	EXP
BAGLE.N-O	44	0.117	0.117	0.118	0.118
BUGBEAR.B	243	0.053	0.050	0.061	0.084
FUNLOVE.4099	268	0.031	0.032	0.030	0.138**
JEEFO.A	188	0.049	0.049	0.048	0.111*
LOVGATE.AC-O	154	0.065	0.063	0.062	0.059
MAGISTR.A	27	0.084	0.093	0.102	0.149
NIMDA.A	125	0.108	0.102	0.102	0.101
NIMDA.A-O	110	0.144*	0.133*	0.133*	0.132*
PARITE.A	233	0.115**	0.112**	0.112**	0.112**
VALLA.A	177	0.061	0.062	0.072	0.081
ZAFI.B	211	0.071	0.071	0.068	0.077

Table 6. KS Statistics (TROJ).

NAME	DAYS	NOR	LOG	EXT	EXP
ALCHEMIC.A	250	0.164**	0.147**	0.147**	0.147**
CLICKER.F	28	0.138	0.126	0.127	0.127
DFC.A	62	0.116	0.115	0.116	0.116
DYFUCA.CN	184	0.048	0.049	0.049	0.049
ISTBAR.AJ	57	0.125	0.110	0.144	0.171
ONECLICK.A	208	0.057	0.056	0.058	0.115**
PORNDIAL.BP	66	0.061	0.062	0.083	0.147
QDOWN.L	92	0.091	0.092	0.091	0.111
QOOLOGIC.P	24	0.114	0.114	0.127	0.127
REALTENS.H	84	0.169*	0.164*	0.180**	0.187**
ROOTKIT.N	15	0.237	0.231	0.240	0.351*
SMALL.SN	34	0.170	0.172	0.172	0.171
UPLOADER.F	105	0.145*	0.132	0.114	0.182**

Table 7. KS Statistics (VBS).

NAME	DAYS	NOR	LOG	EXT	EXP
BAGLE.GEN	128	0.050	0.035	0.038	0.081
BAGLE.X	66	0.051	0.050	0.053	0.067
BAGLE.Z	157	0.070	0.058	0.058	0.058
GEDZA.A	56	0.204*	0.187*	0.187*	0.186*
LOVELETTER.A	15	0.113	0.132	0.131	0.134
PHEL.Q	48	0.084	0.082	0.085	0.110
REDLOF.A-1	247	0.080	0.077	0.078	0.138**
REDLOF.A-2	214	0.042	0.040	0.043	0.097*
SORACI.A	208	0.035	0.035	0.036	0.112*
STARTER.B	62	0.174*	0.172*	0.175*	0.192*
ZIKDOW.A	154	0.062	0.060	0.062	0.428**

Table 8. KS Statistics (WORM).

NAME	DAYS	NOR	LOG	EXT	EXP
ANIG.A	140	0.083	0.078	0.088	0.175**
ANTINNY.A	212	0.068	0.068	0.068	0.070
ANTINNY.B	21	0.085	0.104	0.104	0.105
ANTINNY.G	174	0.069	0.069	0.068	0.071
BUGBEAR.A	159	0.072	0.068	0.068	0.068
MABUTU.A	211	0.072	0.064	0.064	0.064
MSBLAST.A	176	0.043	0.044	0.044	0.044
MYDOOM.A	148	0.158**	0.149**	0.149**	0.149**
MYDOOM.M	256	0.073	0.065	0.065	0.065
NACHI.A	194	0.038	0.037	0.039	0.089
NETSKY.A	291	0.046	0.046	0.045	0.045
NETSKY.B	170	0.050	0.050	0.053	0.236**

Table 9. KS Statistics (ADW).

NAME	DAYS	NOR	LOG	EXT	EXP
BLAZE.B	104	0.051	0.052	0.052	0.324**
FUNWEB.C	34	0.270*	0.272*	0.272*	0.272*
HOTBAR.C	146	0.065	0.061	0.076	0.124*
HOTBAR.E	114	0.073	0.077	0.082	0.099
HOTBAR.G	135	0.047	0.038	0.056	0.207**
ISTBAR.B	157	0.031	0.041	0.048	0.486**
NCASE.A	269	0.044	0.041	0.043	0.276**
NCASE.C	60	0.088	0.088	0.087	0.087
NETPALS.A	205	0.064	0.063	0.063	0.141**
RULEDOR.C	151	0.108	0.096	0.095	0.095
SAVENOW.A	223	0.073	0.069	0.076	0.248**
SOLU180.A	183	0.048	0.047	0.062	0.097
SOLU180.D	157	0.047	0.046	0.046	0.047
WINAD.A	218	0.096*	0.091	0.099*	0.364**

Table 10. KS Statistics (HKTL).

NAME	DAYS	NOR	LOG	EXT	EXP
BRUTFORCE.A	98	0.061	0.060	0.066	0.269**
BRUTUS.A	114	0.069	0.069	0.069	0.070
ENTRY.27	129	0.074	0.076	0.087	0.208**
PSEXEC.A	97	0.042	0.047	0.060	0.557**
RADMIN.A	53	0.148	0.143	0.143	0.237**
RADMIN.B	52	0.095	0.093	0.100	0.177

Table 11. KS Statistics (SPYW).

NAME	DAYS	NOR	LOG	EXT	EXP
AGENT.A	115	0.122	0.115	0.127	0.453**
BISPY.A	204	0.050	0.050	0.050	0.062
DYFUCA.A	169	0.128**	0.128**	0.128**	0.261**
GATOR.B	197	0.067	0.067	0.066	0.120**
IESEARCH.A	156	0.110*	0.111*	0.109*	0.245**
MARKTSCOR.A	178	0.049	0.047	0.048	0.076
PPNETWORK.B	135	0.059	0.054	0.062	0.169**
STINTER.A	53	0.288**	0.270**	0.261**	0.233**
WEBHANCER.A	193	0.055	0.051	0.056	0.187**
WEBSEARCH.A	150	0.155**	0.142**	0.158**	0.351**

Tables 1–11 present the KS statics for 116 viruses when we apply NOR, LOG, EXT and EXP. The column DAYS indicates the total number of days when the infection is reported, and these correspond to the degree of freedom in KS test. The marks '*' and '**' indicate the models which are rejected with significant level 0.90 and 0.95, respectively. In addition, the virus with an underline corresponds to the worm-like infection path, although the virus does not belong to WORM.

From these tables, NOR, LOG and EXT can fit to the number of infected hosts. That is, the NHPPs based on NOR, LOG and EXT are not rejected in the KS test. In particular, compared with the results of EXP, the results of NOR, LOG and EXT are much superior in terms of goodness-of-fit. However, we find that some viruses belonging to HTML, TROJ and SPYW cannot be described well by the NHPP models. This is caused by the fact that HTML, TROJ and SPYW basically belong to Trojan horse. In general, the infection timing of Trojan horse depends on users' activities, and therefore the propagation of Trojan horse often becomes periodic. When we focus on the goodness-of-fit for the viruses with underlines in HTML and TROJ, the goodness-of-fit is higher than the other HTML and TROJ viruses. This is because these viruses behave like worms.

Figures 2, 3 and 4 illustrate the cumulative number of infections and estimated mean value functions for the NHPP models. As shown in these figures, NOR, LOG and EXT can represent exponential and S-shaped curves. However, in Fig. 4, since the number of infections has a periodic propagation, the simple NHPP models cannot represent the behavior of propagation. In fact, NOR, LOG, EXT and EXP are rejected to the data MYDOOM.A in Table 8.

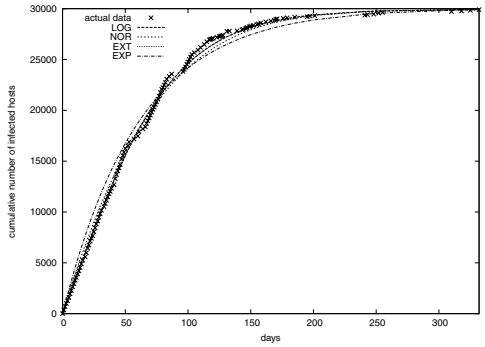

Figure 2. The cumulative number of infections (BAGLE.GEN).

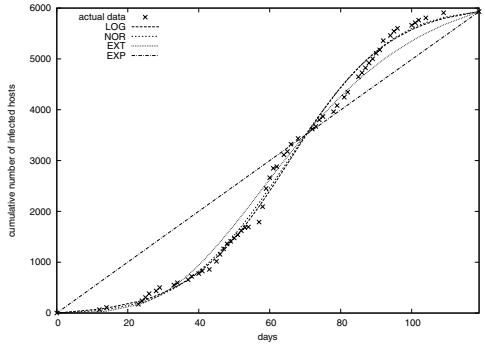

Figure 3. The cumulative number of infections (WINSHOW.A).

Figure 5 shows the fitting result by mixed-type NHPP model. Here we use a mixture of 4 normal distributions. Unlike the simple NHPP model, the mixed-type NHPP model represents periodic propagation of computer virus.

5.2. Prediction Performance

Next we compare prediction abilities of NHPP-based models and the regression for the logistic and Gompertz curves. To examine the quantitative performance of prediction, we define the predictive error variance (PEV) as follows.

$$\mathrm{PEV} = \frac{1}{k}\sum_{i=s}^{s+k}\Big(N(t_i) - \hat{N}(t_i)\Big)^2, \qquad (29)$$

where $N(t)$ and $\hat{N}(t)$ are the observed number of infections at time t and a predicted one by the estimated models, respectively. This indicates a time average of squared errors

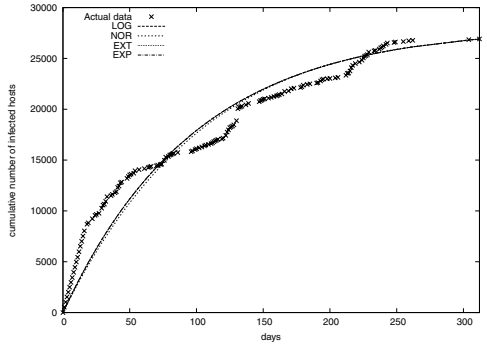

Figure 4. The cumulative number of infections (MYDOOM.A).

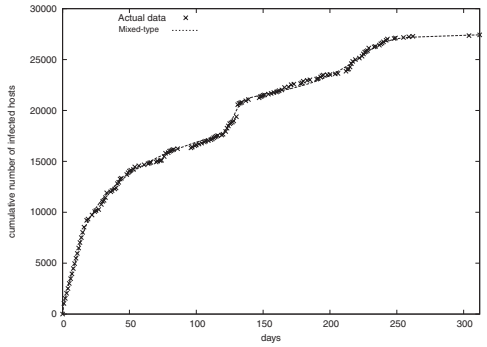

Figure 5. The fitting result of mixed-type NHPP model (MYDOOM.A).

between the actual infection data and the estimated mean value function after k days.

Figure 6 illustrates the prediction results of the mixed-type NHPP with 4 normal distributions (Mixed-type) and the regression models with the logistic curve (LC) and Gompertz curves (GC) when we use the infection data of MYDOOM.A for the first 150 days. Also, Table 12 presents AICs and PEVs for the models, where we compute PEVs for 30, 60 and 100 days. Compared to the results of the regression models, the mixed-type NHPP can drastically reduce the prediction errors. However, we find that the result for a long period, PEV(60) or PEV(100), is not enough to accurately predict the future infections even if we use the mixed-type NHPP. In future researches, we need to improve the prediction ability of this approach.

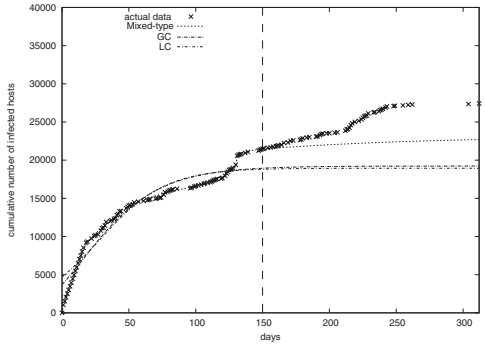

Figure 6. The predicted mean value functions (MYDOOM.A).

Table 12. Predictive performance for NHPPs and regressions (MYDOOM.A).

	AIC	PEV(30)	PEV(60)	PEV(100)
Mix	7025	172404	467023	3392237
GC		5039573	5903384	14098865
LC		5639233	6591395	15362869

6. Conclusions

In this paper, we have developed the statistical models to describe the computer virus propagation based on NHPPs. In particular, when we apply the logistic and extreme value distributions to the infection time distribution, the resulting mean behavior of NHPP models are exactly same as the well-known logistic and Gompertz curves. Thus the framework of NHPP models essentially contains the conventional regression analysis. Moreover, we have introduced the mixed-type NHPP models to represent the propagation of computer virus. Since the mixed-type NHPP models can express periodic infection phenomenon, it is superior to the usual non-mixed NHPPs with unimodal infection time distribution, in terms of goodness-of-fit. For the statistical analysis, we have proposed the EM algorithm so that we could easily estimate model parameters for virus infection data. In numerical experiments, we have performed KS test of the NHPP models for 116 kinds of virus data. As a result, almost all virus infection would be modeled by NHPPs. Also, we examined the prediction abilities for the proposed mixed-type NHPP models, compared to the conventional regression models with the logistic and Gompertz curves. Although we have investigated that the mixed-type NHPP models were capable of fitting to any kind of infection data in numerical experiments, the prediction ability is insufficient to evaluate the future virus infection even when we use the mixed-type NHPP. In practice, the accurate prediction

of virus propagation is a significant challenge. In future, we will develop model-based approaches to provide more accurate prediction of virus propagation by using some statistical techniques such as interval estimation.

Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), Grant Nos. 18510138 (2006-2008), 19510148 (2007-2008) and Young Scientists (B), Grant No. 19700065 (2007-2008).

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proc. 2nd Int'l Sympo. on Information Theory*, pages 267–281. Akademiai Kiado, 1973.

[2] O. H. Alhazmi and Y. K. Malaiya. Modeling the vulnerability discovery process. In *Proceedings of 16th International Symposium on Software Reliability Engineering*, pages 129–138. IEEE CS Press, 2005.

[3] O. H. Alhazmi and Y. K. Malaiya. Measuring and enhancing prediction capabilities of vulnerability discovery models for Apache and IIS HTTP servers. In *Proceedings of the 17th International Symposium on Software Reliability Engineering*, pages 343–352. IEEE CS Press, 2006.

[4] L. J. S. Allen. *An Introduction to Stochastic Processes with Applications to Biology*. Peason Education, Inc., New Jersey, 2003.

[5] L. J. S. Allen and A. M. Burgin. Comparison of deterministic and stochastic SIS and SIR models in discrete time. *Mathematical Biosciences*, 163:1–33, 2000.

[6] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, New York, 2nd edition, 2006.

[7] F. B. Cohen. Computer viruses – theory and experiments. In *Proceedings of the 7th National Computer Security Conference*, pages 240–263, 1984.

[8] J. A. Jacquez and C. P. Simon. The stochastic SI model with recruitment and deaths I. comparison with the closed SIS model. *Mathematical Biosciences*, 1993.

[9] J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. In *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, 1991.

[10] J. O. Kephart and S. R. White. Measuring and modeling computer virus prevalence. In *Proceedings of the 1993 IEEE Computer Society Symposium on Research in Security and Privacy*, 1993.

[11] D. Moore, C. Shannon, and J. Brown. Code-red: a case study on the spread and victims of an Internet worm. In *Proceedings of the Internet Measurement Workshop*, 2002.

[12] H. Okamura and T. Dohi. EM algorithm for extreme-value software reliability models. In *Abstract Book of 4th International Conference on Mathematical Methods in Reliability - Methodology and Practice (CD-ROM)*, 2004.

[13] H. Okamura, T. Dohi, and S. Osaki. EM algorithms for logistic software reliability models. In *Proc. 7th IASTED Int'l Conf. on Software Eng.*, pages 263–268, 2004.

[14] H. Okamura, H. Kobayashi, and T. Dohi. Dependence of computer virus prevalence on network structure - stochastic modeling approach. In *Proceedings of 2004 Asian International Workshop on Advanced Reliability Modeling*, pages 379–386, Singapore, 2004. World Scientific.

[15] H. Okamura, H. Kobayashi, and T. Dohi. Markovian modeling and analysis of Internet worm propagation. In *Proceedings of 16th International Symposium on Software Reliability Engineering*, pages 149–158, 2005.

[16] H. Okamura, K. Tateishi, and T. Dohi. Statistical models for propagation of computer virus based on non-homogeneous Poisson processes (in Japanese). *Transactions of IEICE*, J89-D:1729–1738, 2006.

[17] H. Okamura, T. Watanabe, and T. Dohi. An iterative scheme for maximum likelihood estimation in software reliability modeling. In *Proc. of 14th Int'l Sympo. Software Reliab. Eng.*, pages 246–256. IEEE CS Press, 2003.

[18] S. Sellke, N. B. Shroff, and S. Bagchi. Modeling and automated containment of worms. In *Proceedings of the International Conference on Dependable Systems and Networks*, pages 528–537, 2005.

[19] S. Staniford, V. Paxson, and N. Weaver. How to own the Internet in your spare time. In *Proceedings of the 11th USENIX Security Symposium*, 2002.

[20] H. Toyoizumi and A. Kara. Predators: good will codes combat against computer viruses. In *ACM SIGSAC New Security Paradigms Workshop*, 2002.

[21] J. C. Wierman and D. J. Marchette. Modeling computer virus prevalence with a susceptible-infected-susceptible model with reintroduction. *Computational Statistics & Data Annalysis*, 45:3–23, 2004.

[22] S.-W. Woo, O. H. Alhazmi, and Y. K. Malaiya. Assessing vulnerabilities in Apache and IIS HTTP servers. In *Proceedings of the 2nd IEEE International Symposium on Dependable, Autonomic and Secure Computing*, page IEEE CS Press, 2006.