J Comput Virol (2010) 6:31–42
DOI 10.1007/s11416-008-0111-3

ORIGINAL PAPER

Analysis of a scanning model of worm propagation

Ezzat Kirmani · Cynthia S. Hood

Received: 29 May 2008 / Accepted: 7 December 2008 / Published online: 24 December 2008

Abstract The traditional approach to modeling of internet worm propagation is to adopt a mathematical model, usually inspired by modeling of the spread of infectious diseases, describing the expected number of hosts infected as a function of the time since the start of infection. The predictions of such a model are then used to evaluate, improve, or develop defense and containment strategies against worms. However, a proper and complete understanding of worm propagation goes well beyond the mathematical formula given by the chosen model for the expected number of hosts infected at a given time. Thus, questions such as fitting the model, assessing the extent to which a specific realization of a worm spread may differ from the model's predictions, behavior of the time points at which infections occur, and the estimation and effects of misspecification of model's parameters must also be considered. In this paper, we address such questions for the well-known random constant spread (RCS) model of worm propagation. We first generalize the RCS model to our nonhomogeneous random scanning (NHRS) model. The NHRS model allows the worm's contact rate to vary during worm propagation and it thus captures far more situations of interest than the RCS model which assumes a scanning rate constant in time. We consider the problem of fitting these models to empirical data and give a simulation procedure for a RCS epidemic. We also show how to obtain a confidence interval for the unknown contact rate in the RCS model. In addition, the use of prior information about the contact rate is discussed. The results and methodologies of this paper illuminate the structure and application of NHRS and RCS models of worm propagation.

E. Kirmani (✉)
Department of Statistics and Computer Networking,
Saint Cloud State University, Saint Cloud, MN 56301, USA
e-mail: ekirmani@stcloudstate.edu

C. S. Hood
Department of Computer Science, Illinois Institute of Technology,
Chicago, IL 60616, USA
e-mail: hood@iit.edu

1 Introduction

Internet worms such as Code Red, Nimda, Slammer, and Blaster have dramatically exposed the vulnerability of the Internet to malicious programs that self-propagate by exploiting software errors and other security faults. The recent trend in malware towards bots and botnets has not, by any means, eliminated the challenge of worms. In fact, as noted in Lee et al. [6], botnets can provide a platform for simultaneous launching of worms from distributed networks of bots. Such an instantaneous attack necessarily shortens the window of time in which the network administrators must implement the necessary countermeasures. Apparently, the art and science of defense and containment methodologies against such worms is lagging behind. The development of countermeasures depends, among other things, on the worm's function structure, its execution mechanism, scanning strategies, and propagation modeling (see [11] for a brief survey). The traditional approach to modeling of internet worm propagation is to choose a mathematical model describing relationships of interest, such as the expected number of hosts infected as a function of the time since the start of infection. The predictions of such a model are then used in evaluating defense and containment strategies against worms. Thus, mathematical models of worm propagation play an important role and it is necessary to understand their limitations, structure, ramifications, and how they are applied to specific situations. Some important questions in this regard are estimation of the unknown parameters of a worm propagation model, simulation of various realizations to assess deviations from predicted behavior, and sensitivity of the model to misspecifications of parametric values. The objective of this paper is to study selected such aspects of the well-known random constant spread (RCS) model of worm propagation and its generalization to a model allowing for nonhomogeneous contact rates.

In order to evaluate, improve, and develop effective defense and containment strategies against such worms it is necessary to understand how various worms propagate.

Although models for the spread of computer virus were already considered in [4,5], it was the Code Red worm of July, 2001 which led to wide interest in modeling of internet worm propagation [9]. Staniford et al. [13] used empirical data derived from the outbreak of the Code Red worm to develop their RCS model. It describes the cumulative number of hosts infected, as a function of the time since the start of the worm epidemic, in the absence of any measures to counter the epidemic. It is useful in predicting the propagation pattern of new worms. This model has become a benchmark and a source of extensions, generalizations, and refinements. A key assumption in this model is that the rate at which an infected host chooses new victims is constant in time. This, of course, is a simplification which ignores that most hosts have different bandwidths available for scanning and different scanning computing power. Moreover, large scale worm propagation causes network congestion resulting in actually slowing down the worm's contact rate after a certain threshold. The constant scanning rate in a RCS model is thus merely an average which has obvious limitations in describing the evolution of infection in time. We, therefore, develop a generalization of the RCS model in which the worm's scanning rate is not constant but varies with time. This improvement, which we call the nonhomogeneous random scanning model (NHRS), is presented in Sect. 2.

In Sect. 2, we also show how to fit our NHRS model using minimal empirical data. We also give a procedure based on linear regression, to estimate the contact rate of a RCS worm and assess the goodness-of-fit of the RCS model. The general approach to worm propagation models in the literature is to consider the number of infected hosts as a deterministic function of the time since the start of the epidemic. While this approach does give the expected number of infected hosts, it fails to explain the probabilistic structure underlying worm propagation. The importance of fully understanding the probabilistic regime of worm behavior has not been properly appreciated in the literature although Nicol [10] made a strong case for studying the impact of stochastic variance on worm propagation and detection. As noted in Nicol [10], the regime of worm behavior affects simulation-based studies of worm detection/defense mechanisms. The complete probabilistic description of worm propagation requires specification of the probability distribution of the sequence of successive infection times. Indeed, this is our approach in Sect. 3. It enables us to fully explain the probabilistic regime of worm propagation for NHRS and RCS models. Our results in Sect. 3 provide the basis for a sound probabilistic procedure for simulating a RCS epidemic. This procedure, included in Sect. 3 itself, enables us to simulate worm propagation in order to see the extent to which a specific realization may differ from the prediction of the RCS model. Such simulations are important for quantitative assessment of effectiveness of detection mechanisms and countermeasures.

The constant contact rate assumed in the RCS model is generally unknown. In addition to the linear regression based method of Sect. 3, we give a confidence interval for the contact rate in Sect. 4. The data required for this confidence interval is merely the number of hosts infected during an observation period. Since the exact value of the contact rate is unknown, it is important to be able to assess the sensitivity of the predictions made by the RCS model. We address this issue in Sect. 4 by deriving the expected number of hosts infected during any specified time by adopting a Bayesian approach. We conclude with a brief description of future work on an extension of the NHRS model to include possible recovery, patching and immunization of systems.

2 RCS model and its nonhomogeneous improvement NHRS model

Creating workable models of worm propagation is necessary for several reasons. They allow us to learn form previous worm incidents and to possibly predict the behavior of future worms. They help to develop and test containment, disinfection, and patching strategies without actually developing and releasing computer worms. Moreover, worm propagation models appear to be the sole means for predicting the extent of failure and damage a worm may cause to the Internet. This prediction is particularly important for the early phase of worm propagation. In the early phase, the necessary defenses may not be in place which it may be crucial to prevent the worm from spreading into critical parts of networks.

A crucial factor in the propagation of a worm is its spread algorithm. The most popular spread algorithm is random scanning in which the worm picks an IP address at random, attempts to establish contact and infect it. The Code Red worm employs random scanning. Staniford et al. [13] developed their RCS model to describe the propagation of CRv2, the second version of the Code Red worm, which attacked the Internet on July 19, 2001.

2.1 Extending the RCS model to allow time-dependent contact rate

The RCS model assumes that: (a) the total number of vulnerable hosts which can be potentially compromised is a constant

N, (b) the Internet topology can be considered as an unidirected complete graph, (c) the number of vulnerable hosts that an infected host can compromise per unit of time is a constant β, (d) an infected host picks other hosts to attack completely at random, and (e) a host cannot be compromised multiple times. The assumption (a) implies that countermeasures such as patching, disconnecting servers or restricting access are ignored. While unrealistic in general, this may well be reasonable in the early phase of the propagation of a new worm for which the necessary defenses may not be in place. Assumption (b), though not really true, is not a serious limitation and lack of complete connectivity does not fundamentally alter the conclusions of the RCS model. The assumption (c) ignores differences in network connection, bandwidth, and processor speeds.

The RCS model consists of the differential equation

$$\frac{d}{dt}a(t) = \beta a(t)\{1 - a(t)\} \tag{1}$$

where $a(t)$ denotes the proportion of vulnerable hosts infected during the time period $[0, t]$ with $t = 0$ as the instant of infection of the first host(s) infected and β is the constant contact rate. If $a(0) = I_0/N$, where $I_0 \geq 1$, the (1) has the solution

$$a(t) = \frac{1}{1 + \psi \exp(-\beta t)} \tag{2}$$

where $\psi = (N - I_0)/I_0$.

Staniford et al. [13] fitted the RCS model to the total number of inbound scans seen during $[0, t]$, $0 < t \leq 16$, t in hours, on port 80 at the Chemical Abstracts Service during the initial outbreak of CRv2 on July 19, 2001. The contact rate β depends on the worm's probe rate and its target acquisition function. The solution of (1) given by (2) requires the assumption that β is constant in time. However, large scale worm propagation causes network congestion affecting the availability of bandwidth [2,3]. Such bandwidth limitations or human reaction actually slows down the worm's scanning process. To model this phenomenon, a constant value of β is clearly inappropriate; a more realistic assumption, for example, might be to take β as a decreasing function of the number of hosts infected by time t. In any case, the assumption that β is constant in time merely assigns an average value to β which cannot convey the extent and nature of variability in time.

We, therefore, propose a generalization of the RCS model by allowing β to depend on time t. This generalization, given in Proposition 1 below, will be called the NHRS model. We will take the number of hosts infected during $[0, t]$ as a random variable $X(t)$ rather than the deterministic function $Na(t)$. Although the deterministic function $Na(t)$ can itself be interpreted as the expected value of $X(t)$, our approach has two advantages: (1) it makes the underlying probabilistic assumptions fully transparent, and (2) enables the derivation of some useful probability distributions in Sects. 3 and 4. In keeping with the terminology now commonly used in the worm propagation literature, we will refer to hosts vulnerable to be infected by a worm as susceptible hosts. The following proposition describes our NHRS model.

Proposition 1 *Given a set of N hosts, let $I(t) = E\{X(t)\}$ where $X(t)$ is the number of hosts infected during the time period $[0, t]$, $t \geq 0$. Suppose that* (i) *a host once infected becomes infectious (i.e., capable of causing infection) and remains so,* (ii) *a susceptible host becomes infected if and only if it comes into contact with an infectious host,* (iii) *$p(t, t + \Delta t)$ is the probability that a given susceptible host is contacted by a given infectious host during the infinitesimal time period $(t, t + \Delta t]$,* (iv) *this probability is the same for each pairing of susceptible hosts with infected hosts, and* (v) *all contacts between susceptible and infectious hosts are independent.*

If

$$p(t, t + \Delta t) = \{\beta(t)/N\}\Delta t + o(\Delta t) \tag{3}$$

then

(A)

$$\frac{d}{dt}I(t) = \beta(t)I(t)\left\{I - \frac{I(t)}{N}\right\}, \quad t > 0, \tag{4}$$

which has the solution

(B)

$$I(t) = \frac{N}{1 + \psi \exp(-\int_0^t \beta(u)du)}, \quad t \geq 0, \tag{5}$$

where $\psi = \frac{N-I_0}{I_0}$ with $I_0 = I(0)$.

Proof In order to prove (A), we must calculate $I(t + \Delta t) - I(t)$. This is the expected increase in the number of infected hosts during $(t, t+\Delta t]$. It represents the expected number of infections (among the $N - X(t)$ hosts susceptible at time t) during $(t, t+\Delta t]$ generated by the $I(t)$ hosts who are infectious at time t. Therefore,

$$I(t + \Delta t) - I(t) = E\left\{\sum_{i=1}^{N-X(t)} C_i\right\} \tag{6}$$

where

$$C_i = \begin{cases} 1, & \text{if the } i \text{ th susceptible host is contacted by} \\ & \text{at least one infectious host during } (t, t + \Delta t] \\ 0, & \text{otherwise.} \end{cases}$$

In view of the assumptions made above,

$$\begin{aligned} I(t + \Delta t) - I(t) &= \lambda E\{N - X(t)\} \\ &= \lambda\{N - I(t)\} \end{aligned} \tag{7}$$

where λ depends on t and Δt, and, for each i

$$\begin{aligned}\lambda &= E(C_i) = P(C_i = 1)\\ &= \sum_{n=0}^{N} P(C_i = 1|X(t) = n)P(X(t) = n).\end{aligned}$$

Now, for all $n \geq 1$,

$$\begin{aligned} P(C_i = 1|X(t) = n) &= 1 - P(C_i = 0|X(t) = n)\\ &= 1 - P(\text{the } i\text{th susceptible in not contacted by}\\ &\qquad \text{any of the } n \text{ infectious hosts during } (t, t+\Delta t])\\ &= 1 - \prod_{j=1}^{n} P(\text{the } i\text{th susceptible in not contacted by}\\ &\qquad \text{the } j\text{th infectious host during } (t, t+\Delta t])\\ &= 1 - \{1 - p(t, t+\Delta t)\}^n\\ &= 1 - \sum_{k=0}^{n} \binom{n}{k} (-1)^k \{p(t, t+\Delta t)\}^k \end{aligned} \tag{8}$$

Since

$$p(t, t+\Delta t) = \{\beta(t)/N\}\Delta t + o(\Delta t),$$

we have

$$\{p(t, t+\Delta t)\}^k = o(\Delta t) \tag{9}$$

for all $k \geq 2$. Therefore, the last sum above reduces to

$$np(t, t+\Delta t) + o(\Delta t)$$

so that, for all $n \geq 1$,

$$P(C_i = 1|X(t) = n) = np(t, t+\Delta t) + o(\Delta t). \tag{10}$$

Since it is trivial that

$$P(C_i = 1|X(t) = 0) = 0,$$

we get

$$\begin{aligned}\lambda &= \sum_{n=0}^{N}\{np(t, t+\Delta t) + o(\Delta t)\}P(X(t) = n)\\ &= p(t, t+\Delta t)\sum_{n=0}^{N} nP(X(t) = n) + o(\Delta t)\\ &= p(t, t+\Delta t)E\{X(t)\} + o(\Delta t)\\ &= p(t, t+\Delta t)I(t) + o(\Delta t)\\ &= \left\{\left(\frac{\beta(t)}{N}\right)\Delta t + o(\Delta t)\right\} I(t) + o(\Delta t)\\ &= \left\{\frac{\beta(t)}{N}\right\} I(t)\Delta t + o(\Delta t)\end{aligned} \tag{11}$$

Therefore,

$$\begin{aligned}\frac{I(t+\Delta t) - I(t)}{\Delta t} &= \frac{\lambda\{N - I(t)\}}{\Delta t}\\ &= \left\{\frac{\beta(t)}{N}\right\} I(t)\{N - I(t)\} + \frac{o(\Delta t)}{\Delta t}\end{aligned} \tag{12}$$

Taking the limit as $\Delta t \longrightarrow 0$ and noting that

$$\lim_{\Delta t \longrightarrow 0} \frac{o(\Delta t)}{\Delta t} = 0, \tag{13}$$

we get

$$\begin{aligned}\frac{d}{dt}I(t) &= \left\{\frac{\beta(t)}{N}\right\} I(t)\{N - I(t)\}\\ &= \beta(t)I(t)\left\{1 - \frac{I(t)}{N}\right\}\end{aligned}$$

which proves claim (A) of the proposition.

To prove (B), we first observe that the substitution $I(t) = 1/y(t)$ transforms the (4) to

$$\frac{d}{dt}y(t) + \beta(t)y(t) = (1/N)\beta(t). \tag{14}$$

Writing

$$B(t) = \int_0^t \beta(u)du$$

and multiplying both sides of the above equation by $A(t) = \exp\{B(t)\}$, we get

$$\left\{\frac{d}{dt}y(t)\right\} A(t) + \beta(t)A(t)y(t) = (1/N)\beta(t)A(t). \tag{15}$$

Since

$$\frac{d}{dt}A(t) = \beta(t)A(t),$$

we have

$$\left\{\frac{d}{dt}y(t)A(t)\right\} = (1/N)\frac{d}{dt}A(t) \tag{16}$$

Integrating (with respect to t) over the interval $[0, v]$ gives

$$y(v)A(v) - y(0)A(0) = (1/N)\{A(v) - A(0)\}$$

or,

$$y(v)A(v) - \frac{1}{I_0} = (1/N)\{A(v) - 1\}$$

or,

$$y(v)\exp\{B(v)\} = \frac{N + I_0\{\exp(B(v)) - 1\}}{NI_0} \tag{17}$$

Hence,

$$\begin{aligned} I(v) &= \frac{NI_0\exp\{B(v)\}}{N + I_0\{\exp(B(v)) - 1\}}\\ &= \frac{N}{(N/I_0)\exp\{-B(v)\} + 1 - \exp\{-B(v)\}}\\ &= \frac{N}{1 + \{(N/I_0) - 1\}\exp\{-B(v)\}}\\ &= \frac{N}{1 + \psi\exp\{-\int_0^v \beta(t)dt\}}\end{aligned} \tag{18}$$

where $\psi = (N - I_0)/I_0$. This completes the proof of part (B) of the proposition. □

If $I(t)$, the expected number of hosts infected during $[0, t]$, is as given by the above proposition then we will say that we have a NHRS model with contact function $\beta(t)$. Our NHRS model reduces to the RCS model if $\beta(t) \equiv \beta$ for all $t \geq 0$. A general class of non-constant contact functions is defined by

$$\beta(t) = K\beta^K t^{K-1} \tag{19}$$

where β and K are positive constants. The function $\beta(t)$ is decreasing (in $t > 0$) if $0 < K < 1$, constant if $K = 1$, and increasing if $K > 1$. For this choice of $\beta(t)$, the NHRS model reduces to

$$I(t) = \frac{N}{1 + \psi \exp\{-(\beta t)^K\}}. \tag{20}$$

Figure 1 gives a sketch of this $I(t)$ when $N = 360{,}000$, $I_0 = 1{,}780$, $\beta = 0.8$, and $K = 2$. The other curve in Fig. 1 corresponds to the RCS model with the same values of N, I_0, and β.

It may be noted here that the model derived in Proposition 1 does not allow the so-called local preference scanning [15]. Local preference scanning is a strategy in which an infected host scans IP addresses close to its own address with a higher probability than other IP addresses. However, the assumption (iv) of Proposition 1, namely that $p(t, t + \Delta t)$ is the same for each pairing of susceptible hosts with infected hosts, rules out local preference scanning. As seen in the proof of Proposition 1, this assumption, as well as the other assumptions of Proposition 1, are essential for the applicability of the NHRS model. The flexibility of choosing a time-dependent contact function does not translate into spatial preference in scanning.

The rest of this paper provides a rigorous development of further foundational aspects of the NHRS model and its special case of the RCS model. Clearly, the NHRS model is only one of the many possible generalizations of the RCS model. Several generalizations of the RCS model are available in the literature even though their discussions have not been concerned with questions of the kind we deal with in this paper. Motivated by the classical Kermack-Mckendrick model of theory of epidemics, Zou et al. [14] proposed an internet worm model called the two-factor worm model. This model takes into account the measures to counter the spread of randomly scanning worms through removal of infectious and susceptible hosts. Serazzi and Zanero [12] proposed a compartment-based model based on the macro view of the Internet as the interconnection of a number of Autonomous Systems (AS). It models the behavior of the worm in the intra-AS propagation while assuming that the inter-AS spread follow RCS models. Chen et al. [1] developed an active worm propagation (AAWP) model which improves on the classical Kephart–White epidemiological model of computer viruses. This model assumes random scanning, treats time as discrete and employs continuous state deterministic approximation. We refer to [1,12,14] for additional details about these three generalizations of the RCS model. Without going into further details, we note that the classical epidemiological models for disease progression provide many possible ways of modeling worm propagation.

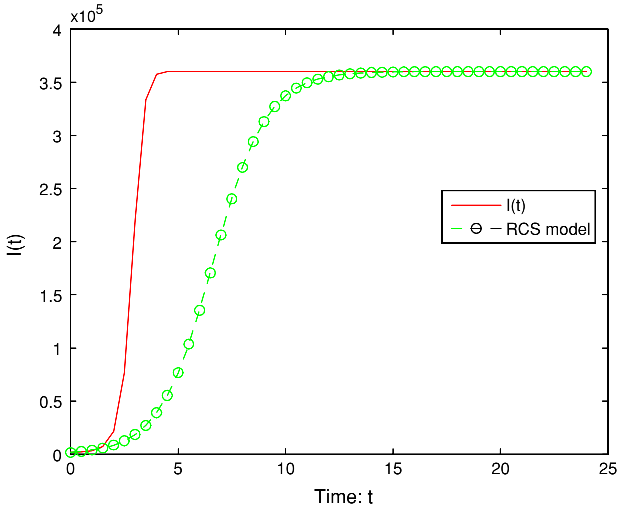

Fig. 1 Expected number of hosts infected against time since start of epidemic

2.2 Fitting the nonhomogeneous random scanning model

If we introduce the function

$$\mu(t) = (1/t) \int_0^t \beta(u) du \quad t > 0,$$

then the NHRS formula (4) can be written as

$$I(t) = \frac{N}{1 + \psi \exp\{-t\mu(t)\}}. \tag{21}$$

The function $\mu(t)$ can be interpreted as the average contact rate over the interval $[0, t]$. If N, the total number of hosts in the network under consideration, and I_0, the number of hosts infected at the beginning of the worm epidemic, are known then fitting the NHRS model amounts to estimating the average contact rate function $\mu(t)$. For this purpose, we propose the approach given below.

It will be assumed that the available data consists of N, I_0, and n pairs $(t_i, X(t_i))$ where $X(t_i)$ denotes the observed number of hosts infected during $[0, t_i]$, $0 \leq t_1 < t_2 < \cdots < t_n$. As we will see in an illustrative example later, n need not be large although a large value of n would give a better estimate of the average contact rate function $\mu(t)$. Since $I(t) = E\{X(t)\}$, we will take $\hat{I}(t_i) = X(t_i)$ as out estimate of $I(t_i)$, $i = 1, 2, \ldots, n$. Our fitting procedure is as follows:

Step 1. For each $i = 1, 2, \dots, n$, compute

$$\mu_i = \left(\frac{1}{t_i}\right) \ln \left(\frac{(N - I_0)\hat{I}(t_i)}{(N - \hat{I}(t_i)) I_0} \right). \tag{22}$$

Step 2. Obtain a smooth function $\hat{\mu}(t)$ which passes through the points (t_i, μ_i), $i = 1, 2, \dots, n$; i.e.,

$$\hat{\mu}(t_i) = \mu_i,$$

and take $\hat{\mu}(t)$ as an estimate of the average contact rate function $\mu(t)$. Estimate $I(t)$

$$\hat{I}(t) = \frac{N}{1 + \psi \exp\{-t\hat{\mu}(t)\}}. \tag{23}$$

The curve $\hat{I}(t)$ is then the NHRS fit to the observed data $(t_i, X(t_i))$, $i = 1, 2, \dots, n$.

The crucial step in the above procedure is, of course, the smoothing part of step2. Piecewise polynomial functions such as splines are popular choices for such smooth functions. The MATLAB function *pchip* finds a piecewise cubic Hermite interpolating polynomial which preserves the shape and monotonicity of the underlying data. The following example illustrates how our procedure can be applied in practice.

Example 1 We will use the data on the CRv2 epidemic given in [7,8]. Figure 2 (taken from [7]) shows the number of distinct IP addresses infected (during the stated 24 h time span) as found on merging three network telescope datasets. According to this figure, more than 359,000 unique IP addresses were detected as victims of CRv2 during the time period $t = 0$ to $t = 24$. Here, t is in hours and $t = 0$ corresponds to midnight (UTC) of July 19, 2001.

For purpose of illustration, we will ignore the detailed information given by Fig. 2 and use instead the approximate numbers given in Table 1. We have obtained these numbers by visual examination of Fig. 2 and make no claims of accuracy.

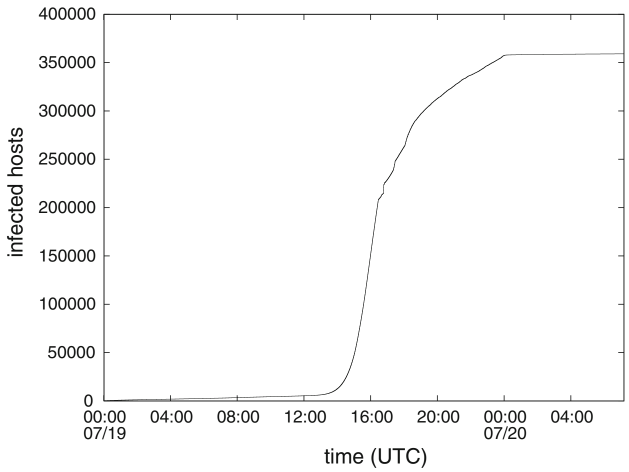

Fig. 2 Number of unique IP addresses infected by CRv2 on July 19, 2001 [From [7]]

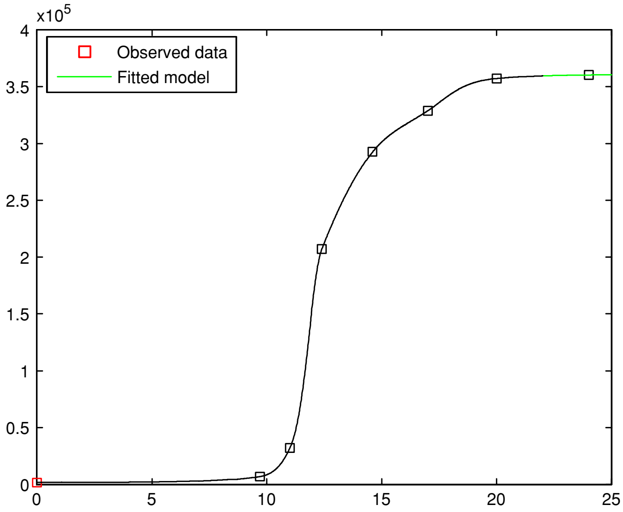

Fig. 3 Fit of NHRS model to the data of Table 1

Table 1 Estimated number of unique IP addresses infected

Row	t	Number of hosts infected during $[0, t]$
1	0.0	1,780
2	9.7	7,140
3	11.0	32,140
4	12.4	207,150
5	14.6	292,850
6	17.0	328,570
7	20.0	357,000
8	24.0	360,000

Our procedure gives Fig. 3 as the fit of the NHRS model to the data of Table 1. We used the *pchip* interpolation function of MATLAB to find the smooth function of step 2 of our procedure. The validity and potential of our approach are clear.

2.3 Fitting the RCS model

The RCS model given in (2) is even easier to fit because it assumes that the contact rate does not change with time. Let

$$Y_t^* = \ln \left(\frac{N}{I(t)} - 1 \right). \tag{24}$$

Then, the RCS model holds with constant contact rate β if, and only if,

$$Y_t^* = \ln \psi - \beta t$$

where $\psi = (N - I_0)/I_0$. Writing

$$Y_t = \ln\left(\frac{N}{X(t)} - 1\right)$$

and noting that $X(t)$ is an estimate of $I(t)$ suggests that

$$Y_t = \ln\psi - \beta t + \epsilon \tag{25}$$

where ϵ measures the random error. It follows that linear regression or al least some parts of it can be applied here. In particular, we can use the principle of least squares and those of its features which do not require random error ϵ to have normal distribution. Assuming that N is known, and that the observed data consists of n pairs $(t_i, X(t_i))$, we propose the following two step procedure for fitting the RCS model and checking its accuracy.

Step 1. Let $Y_i = \ln\left(\frac{N}{X(t_i)} - 1\right)$. Obtain the least-squares regression line of Y on t from the n pairs (t_i, Y_i). Take the slope of this regression line as the estimate of the (constant) contact rate β.

Step 2. Do the usual analysis of residuals to see if the model is appropriate.

Example 2 Most of the infections described in Fig. 2 occurred between 10:00 UTC and 20:00 UTC. It would be interesting to see if this portion of the CRv2 epidemic could possibly be explained by the RCS model. If so, the contact rate β during this period would be estimated by the slope of the regression line of $ln((N/X(t)) - 1)$ on t. Assuming $N = 360{,}000$, we implemented the proposed procedure for the data given in rows 2–7 of Table 1. The fitted regression line has the equation

$$Y = 10.711 - 0.790t \tag{26}$$

However, the plot of residuals against t is found to be curvilinear, indicating that RCS model is not really suitable in this situation.

3 Structure and simulation of NHRS and RCS models

The complete structure of an epidemic cannot be understood by looking at the expected number of infected hosts alone. It is also necessary to understand the behavior of the time points at which the successive infections occur. For this purpose, we need to know the probability distribution of the time until infection for a host which was not infected at the start of the epidemic. This distribution is given below for our NHRS model and its special case the RCS model.

For convenience in writing, let

$$B(t) = \int_0^t \beta(u)du$$

so that the expected number of hosts infected during $[0, t]$ in an NHRS epidemic is given by

$$I(t) = \frac{N}{1 + \psi\exp\{-B(t)\}}.$$

where $\psi = (N - I_0)/I_0$.

Proposition 2 *Let T denote the time until infection (measured from $t = 0$: the start of the epidemic) for a host which was not infected at the start of the epidemic. If the worm propagation in the network of N hosts follows the NHRS model*

$$\frac{d}{dt}I(t) = \beta(t)I(t)\left\{I - \frac{I(t)}{N}\right\}$$

then

$$P(T \le t) = \frac{1 - \exp\{-B(t)\}}{1 + \psi\exp\{-B(t)\}}, \quad t \ge 0, \tag{27}$$

where $\psi = (N - I_0)/I_0$ and $I_0 = I(0)$ is the number of hosts infected at $t = 0$.

Proof For each initially uninfected host i ($i = 1, 2, \ldots, N - I_0$), let

$$C_i(t) = \begin{cases} 1, & \text{if host } i \text{ gets infected during } (0, t] \\ 0, & \text{otherwise.} \end{cases}$$

Then, each $C_i(t)$ has the same probability distribution with

$$\begin{aligned} E\{C_i(t)\} &= P(C_i(t) = 1) \\ &= P(T \le t). \end{aligned} \tag{28}$$

Writing $X(t)$ to denote the number of hosts infected during $[0, t]$, we have

$$X(t) - I_0 = \sum_{i=1}^{N-I_0} C_i(t) \tag{29}$$

so that

$$\begin{aligned} E\{X(t)\} - I_0 &= \sum_{i=1}^{N-I_0} E\{C_i(t)\} \\ &= (N - I_0)P(T \le t). \end{aligned} \tag{30}$$

Therefore,

$$\begin{aligned} P(T \le t) &= \frac{E\{X(t)\} - I_0}{N - I_0} \\ &= \frac{I(t) - I_0}{N - I_0} \\ &= \left(\frac{1}{N - I_0}\right)\left\{\frac{N}{1 + \psi\exp\{-B(t)\}} - I_0\right\} \\ &= \left(\frac{1}{N - I_0}\right)\left\{\frac{N - I_0 - I_0\psi\exp\{-B(t)\}}{1 + \psi\exp\{-B(t)\}}\right\} \\ &= \left(\frac{1}{N - I_0}\right)\left\{\frac{N - I_0 - (N - I_0)\exp\{-B(t)\}}{1 + \psi\exp\{-B(t)\}}\right\} \\ &= \frac{1 - \exp\{-B(t)\}}{1 + \psi\exp\{-B(t)\}}. \end{aligned} \tag{31}$$

□

The probability distribution of T in a RCS epidemic is immediately obtained on putting $B(t) = \beta t$ in the above proposition. The complete structure of the NHRS and RCS epidemics can now be described. Let $L = N - I_0$ denote the number denote the number of hosts which were not infected at the start of the epidemic. Further, let $T_1, T_2, \ldots, T_L$ be L mutually independent random variables such that each of them has the same probability distribution as T. Suppose that $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(L)}$ is the ordered (in increasing magnitude) arrangement of $T_1, T_2, \ldots, T_L$. Then, $T_{(1)}$ is the time of occurrence (measured from the start of the epidemic) of the first infection (excluding the ones which happened at $t = 0$), and $T_{(i)}$ the occurrence time of the ith infection. The last host to be infected is infected at time point $T_{(L)}$. The importance of this description lies in the fact that later it will help us to simulate worm propagation.

In the rest of this section we will focus on the RCS model.

3.1 Simulation of a RCS epidemic

It is crucial to be able to simulate worm propagation in order to see the extent to which a specific realization of an epidemic may differ from the predictions of the model. However, as observed in [10], a comprehensive risk analysis of a detection/defense strategy needs to consider the probability distributions governing worm propagation. The impact of the variability between successive infection times on the variability in worm propagation cannot be captured efficiently unless simulations are explicitly based on the probabilistic regime underlying worm behavior. With this in view, we propose a simulation procedure based on the probability distribution of successive infection times. The following proposition is important for our simulation procedure for the RCS worm propagation.

Proposition 3 *Let U be a random variable having the uniform distribution on the interval* $(0, 1)$. *Then* $\frac{1}{\beta}\ln\left(\frac{1+\psi U}{1-U}\right)$ *has the same probability distribution as* T.

Proof For all $t > 0$,

$$P\left(\frac{1}{\beta}\ln\left(\frac{1+\psi U}{1-U}\right) \leq t\right) = P\left(U \leq \frac{1-\exp(-\beta t)}{1+\psi\exp(-\beta t)}\right)$$
$$= \frac{1-\exp(-\beta t)}{1+\psi\exp(-\beta t)}$$
$$= P(T \leq t) \qquad (32)$$

where the second equality holds because $P(U \leq u) = u$ for all $u \in (0, 1)$. □

We now propose the following procedure for simulating a RCS epidemic in a network of N hosts of which I_0 are infected at the start of the epidemic at $t = 0$. It would be sufficient to simulate the times at which the remaining $N - I_0$ hosts get infected.

Simulation procedure:

Step 1. Draw $L = N - I_0$ numbers randomly (i.e., according to the uniform distribution) from the interval $(0, 1)$. Let $U_1, U_2, \ldots, U_L$ denote the L numbers drawn.

Step 2. Sort $U_1, U_2, \ldots, U_L$ to get $U_{(1)}, U_{(2)}, \ldots, U_{(L)}$ with $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(L)}$.

Step 3. Compute $T_{(i)} = \frac{1}{\beta}\ln(\frac{1+\psi U_{(i)}}{1-U_{(i)}})$, $i = 1, 2, \ldots, L$. Then $T_{(i)}$ is the time point at which infection number i occurs. The epidemic terminates (with all hosts infected) at time point $T_{(L)}$.

Step 4. For all $t > 0$ and $i = 1, 2, \ldots, L$; define

$$K_i(t) = \begin{cases} 1, & T_{(i)} \leq t \\ 0, & T_{(i)} > t \end{cases}$$

Further, let

$$X(0) = I_0$$
$$X(t) = I_0 + \sum_{i=1}^{L} K_i(t), \quad t > 0 \qquad (33)$$

Then, $X(t), t \geq 0$, is the path of the simulated RCS epidemic.

Figures 4, 5, 6, and 7 show some realizations of the RCS epidemic simulated according to the above procedure. In each case, N, I_0, and β are shown for parametric choices indicated under the figure. The solid line in each figure is the expected path predicted by the corresponding RCS model.

To understand the RCS epidemic, we also need the probability distribution (not just the expected value) of the number of hosts infected during the time period $[0, t]$. When the number of hosts infected at $t = 0$ (i.e., at the start of the epidemic) is known, it is sufficient to find the probability distribution of the number of hosts infected during $(0, t]$. This distribution is obtained below.

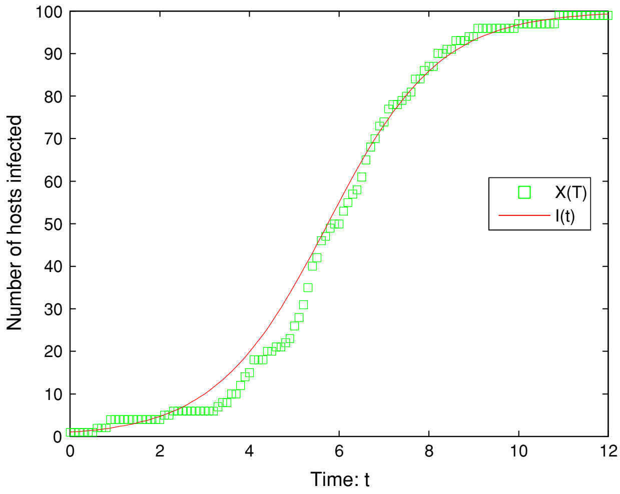

Fig. 4 Simulation of the RCS model for $N = 100$, $I_0 = 1$, $\beta = 0.8$

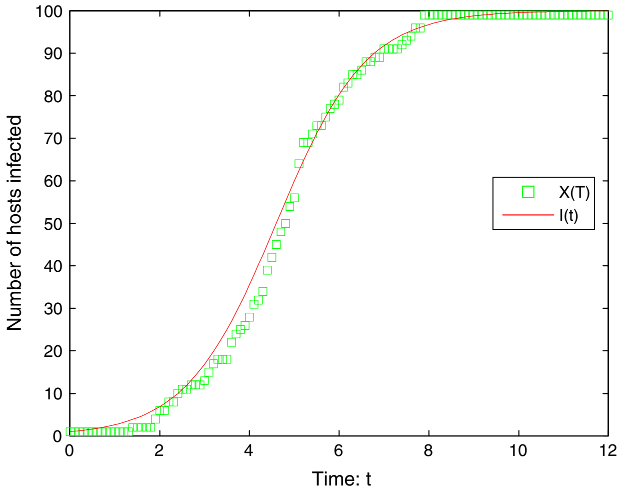

Fig. 5 Simulation of the RCS model for $N = 100$, $I_0 = 1$, $\beta = 1$

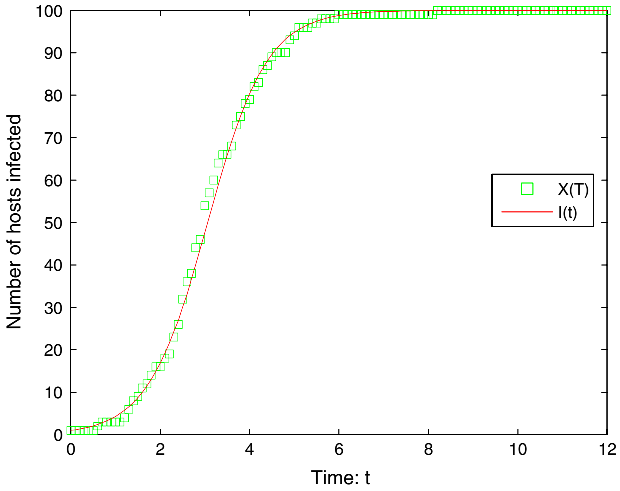

Fig. 6 Simulation of the RCS model for $N = 100$, $I_0 = 1$, $\beta = 1.5$

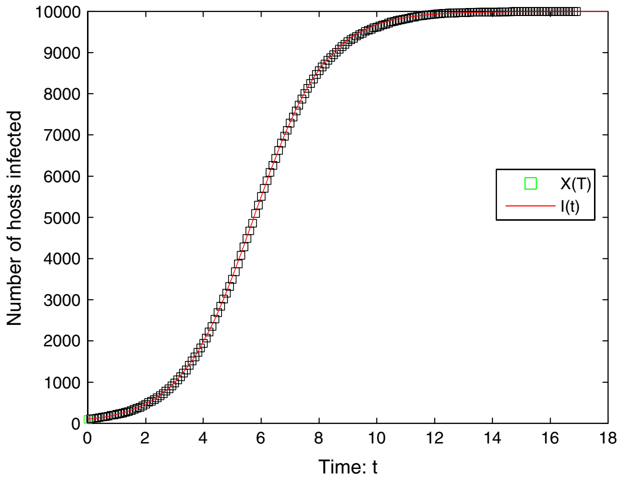

Fig. 7 Simulation of the RCS model for $N = 10{,}000$, $I_0 = 100$, $\beta = 0.8$

Proposition 4 *Let $Y(t) = X(t) - I_0$ so that $Y(t)$ denotes the number of hosts infected during $(0, t]$. If the RCS model of worm propagation is applicable then, given I_0, $Y(t)$ has the binomial distribution*

$$P(Y(t) = y) = \binom{L}{y} \{G(t)\}^y \{1 - G(t)\}^{L-y}$$

where $y = 0, 1, \ldots, L$; $L = N - I_0$, and

$$G(t) = \frac{1 - \exp\{-\beta t\}}{1 + \psi \exp\{-\beta t\}} \quad t \geq 0, \tag{34}$$

with $\psi = (N - I_0)/I_0$.

Proof Let $T_1 < T_2 < \cdots < T_L$ denote the times (measured from $t = 0$: the start of the epidemic) at which the $L = N - I_0$ initially uninfected hosts become infected. Then, as seen above, the L random variables $T_1, T_2, \ldots, T_L$ behave as the order statistics of a random sample of size L from a population with distribution function

$$G(t) \equiv P(T \leq t) = \frac{1 - \exp\{-\beta t\}}{1 + \psi \exp\{-\beta t\}} \quad t \geq 0. \tag{35}$$

It follows that

$$P(T_i \leq t) = \sum_{r=i}^{L} \binom{L}{y} \{G(t)\}^r \{1 - G(t)\}^{L-r}$$

for $i = 1, 2, \ldots, L$. For convenience in writing, let $p = G(t)$. Then for $n = 0, 1, \ldots, L$,

$$\begin{aligned} P(Y(t) = n) &= P(Y(t) \geq n) - P(Y(t) \geq n + 1) \\ &= P(\text{at least } n \text{ hosts are infected during } (0, t]) \\ &\quad - P(\text{at least } n{+}1 \text{ hosts are infected during } (0, t]) \\ &= P(T_n \leq t) - P(T_{n+1} \leq t) \\ &= \sum_{r=n}^{L} \binom{L}{r} p^r (1-p)^{L-r} - \sum_{r=n+1}^{L} \binom{L}{r} p^r (1-p)^{L-r} \\ &= \binom{L}{n} p^n (1 - p)^{L-n}, \quad p = G(t). \end{aligned} \tag{36}$$

□

4 Confidence interval for contact rate and impact of prior distribution on the RCS model

A confidence interval for unknown β can be obtained with the help of the above proposition. If $Y(t_0)$ is the number of hosts infected during an observation period $(0, t_0]$ then an approximately $100(1 - \alpha)\%$ confidence interval for β is

$$\left(\frac{1}{t_0} \ln \left(\frac{1 + \psi \hat{p}_1}{1 - \hat{p}_1} \right), \quad \frac{1}{t_0} \ln \left(\frac{1 + \psi \hat{p}_2}{1 - \hat{p}_2} \right) \right) \tag{37}$$

where

$$\begin{aligned}
\hat{p}_1 &= \hat{p} - z_{\alpha/2}\{\hat{p}(1-\hat{p})/L\}^{1/2},\\
\hat{p}_2 &= \hat{p} + z_{\alpha/2}\{\hat{p}(1-\hat{p})/L\}^{1/2},\\
\hat{p} &= \frac{Y(t_0)}{L},\\
L &= N - I_0 \quad \psi = (N-I_0)/I_0,
\end{aligned} \tag{38}$$

and $z_{\alpha/2}$ is the $100(1-(\alpha/2))$th percentile of the standard normal distribution.

Even though the actual value of β may not be known, there may be some prior information or judgement about the range in which β may lie. In case such information is available, it should be used in the study of worm propagation. We show below the proper approach in such situations.

Suppose the prior information or judgement suggests that the actual value of β is in the interval $[\theta_1, \theta_2]$ where θ_1, θ_2 are specified. If all values in this interval are deemed equally credible then this prior information can be described by the probability density function (pdf)

$$h(\beta) = \begin{cases} \frac{1}{\theta_2-\theta_1}, & \theta_1 \le \beta \le \theta_2 \\ 0, & \text{otherwise.} \end{cases} \tag{39}$$

In effect, we then think of the actual value of β as the realized value of a random variable B whose pdf is the uniform distribution over $[\theta_1, \theta_2]$. This, of course, is the Bayesian approach so common in scientific investigations. In the context of the number of hosts infected during $[0, t]$ under the RCS model, the implication is as follows.

Proposition 5 *Given a set of N hosts, let $X(t)$ be the number of hosts infected during the time period $[0, t]$ with $X(0) = I_0$. Suppose that* (i) *the RCS model is applicable but the exact value of β is unknown, and* (ii) *the prior distribution for β is the uniform distribution on $[\theta_1, \theta_2]$. Then*

$$E\{X(t)\} = N - \left\{\frac{N}{t(\theta_2-\theta_1)}\right\} \ln\left(\frac{1+\psi\exp(-t\theta_1)}{1+\psi\exp(-t\theta_2)}\right) \tag{40}$$

Proof Let B be a random variable having the uniform distribution on $[\theta_1, \theta_2]$. Then, B has pdf

$$h(\beta) = \frac{1}{\theta_2-\theta_1}, \quad \theta_1 \le \beta \le \theta_2$$

and

$$\begin{aligned}
E\{X(t)\} &= E[E\{X(t)|B\}]\\
&= \int_{\theta_1}^{\theta_2} E\{X(t)|B=\beta\}h(\beta)d\beta\\
&= \int_{\theta_1}^{\theta_2} \frac{N}{1+\psi\exp(-\beta t)}\cdot\frac{1}{\theta_2-\theta_1}d\beta\\
&= N\int_{\theta_1}^{\theta_2}\left\{1-\frac{\psi\exp(-\beta t)}{1+\psi\exp(-\beta t)}\right\}\frac{1}{\theta_2-\theta_1}d\beta\\
&= N - \left(\frac{N}{\theta_2-\theta_1}\right)\int_{\theta_1}^{\theta_2}\frac{\psi\exp(-\beta t)}{1+\psi\exp(-\beta t)}d\beta\\
&= N - \left(\frac{N}{\theta_2-\theta_1}\right)\left(\frac{1}{t}\right)\int_a^b \frac{dy}{y}
\end{aligned} \tag{41}$$

where

$a = 1 + \psi\exp(-t\theta_2)$

and

$b = 1 + \psi\exp(-t\theta_1)$.

It follows that

$$E\{X(t)\} = N - \frac{N}{t(\theta_2-\theta_1)}\ln\left(\frac{1+\psi\exp(-t\theta_1)}{1+\psi\exp(-t\theta_2)}\right)$$

□

As an illustration of the importance and implications of incorporating prior information about β, suppose β is assumed to be 1 when, in fact, all that can be justified is that β lies somewhere in an interval containing 1. The dotted curves in Figs. 8 and 9 give the expected number of hosts infected when the prior distribution of β is uniform on the intervals [0.8, 1.2] and [0.5, 1.5], respectively. In both figures, $N = 100$, $I_0 = 1$, and the solid curves gives the expected number of hosts infected when β actually equals1.

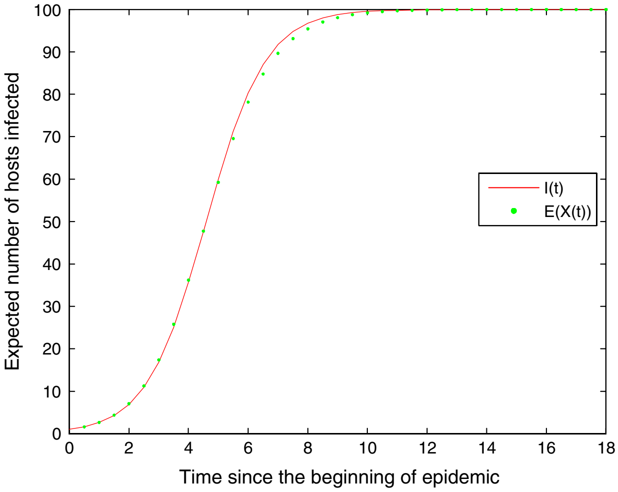

Fig. 8 Prior distribution is uniform on [0.8, 1.2]

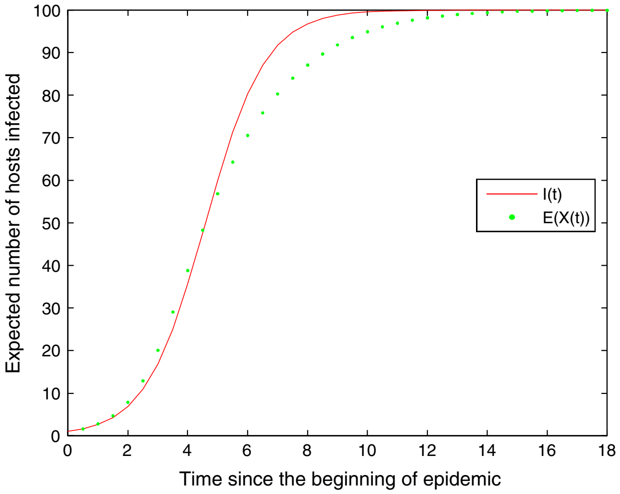

Fig. 9 Prior distribution is uniform on [0.5, 1.5]

5 Conclusion and future work

First of all, we gave a model for taking into account the fact that in practice a worm's scanning rate varies during the duration of the epidemic. Our NHRS model thus captures a much larger class of worm propagation than the well-known RCS model. This generality has been achieved without sacrificing the intrinsic simplicity of the RCS model. The method given by us to estimate the average contact rate function in the NHRS model is practical and convenient for fitting the NHRS model to observed data. We illustrated our method by fitting NHRS model to the number of distinct IP addresses infected during the first twenty four hours of the epidemic caused by version 2 of the Code Red worm. We also gave a linear regression approach to fitting the RCS model. As seen in Sect. 2, the CRv2 considered there is perfectly described by a NHRS model while the RCS model is found to be unsuitable. The probability distribution of the time until infection (for an initially uninfected host) derived for the NHRS model, in Sect. 3, enabled us to describe the structure of the NHRS and RCS epidemics in terms of the times at which successive infections occurred. As a further application, we used this probability distribution to illustrate simulation of RCS epidemic. We also gave a confidence interval for the unknown contact rate of the RCS model. Finally, we demonstrated how the expected number of infected hosts in a RCS epidemic is affected if the contact rate is uniformly distributed over a specified interval. This Bayesian approach enables us to see the effect of uncertainty about the contact rate on the predictions of the RCS model.

The analysis carried out in this paper illuminated a number of aspects of random scanning models of worm propagation and led to several useful procedures. In addition to obtaining the much more widely applicable NHRS model, we laid bare the entire probabilistic regime governing RCS epidemics. As mentioned earlier in Sect. 3.1, it is only through the knowledge of the probability distributions underlying model behavior that a comprehensive risk analysis of defense strategies can be undertaken. It may be added here that the NHRS model can be extended to describe recovery, patching and immunization of systems. Indeed, we have developed an imperfect protection-imperfect recovery (IMP-IMR) model which takes into account possibilities such as (i) preventive steps, not necessarily error proof, may be in place, (ii) immunity obtained by preventive steps may be temporary, and (iii) disinfection, recovery action, or some other strategy, not necessarily error-proof, is being used to contain the epidemic. We are currently investigating the probability distributions associated with our IMP-IMR model. We believe that analysis similar to those in the present paper are required for IMP-IMR as well as other worm propagation models.

Acknowledgments We thank Dr. Eric Filiol, the editor of the Journal in Computer Virology, and the anonymous reviewers for their helpful comments.

References

1. Chen, Z., Gao, L., Kwiat, K.: Modeling the spread of active worms. In: Bauer, F. (ed.) IEEE INFOCOM 2003: The Conference on Computer Communications: 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, Vol. 3, pp. 1890–1900, IEEE Operations Center, New Jersey (2003)
2. Cisco technical notes: Dealing with mallocfail and high CPU utilization resulting from the "Code Red" worm. Cisco Systems, Inc. http://www.cisco.com/warp/public/63/ts_codered_worm.shtml
3. Cisco security advisory: "Code Red" worm—customer impact. Cisco Systems, Inc. http://www.cisco.com/warp/public/707/cisco-code-red-worm-pub.shtml
4. Kephart, J.O., White, S.R.: Measuring and modeling computer virus prevalence. In: Proceedings of the 1993 IEEE Computer Society Symposium on Research in Security and Privacy, pp. 2–15. IEEE Computer Society Press, California (1993)
5. Kephart, J.O., White, S.R., Chess, D.M.: Computers and epidemiology. IEEE Spectr **30**(5), 20–26 (1993)
6. Lee, W., Wang, C., Dagon, D.: Botnet Detection, countering the largest security threat. Springer Science+Business Media, LLC, New York (2008)
7. Moore, D., Shannon, C., Brown, J.: Code-red: a case study on the spread and victims of an internet worm. In: Proceedings of the 2nd Internet Measurement Workshop, pp. 273–284. ACM Press, New York (2002)
8. Moore, D., Shannon, C.: The spread of the code-red worm (CRv2). CAIDA. http://www.caida.org/analysis/security/code-red/coderedv2_analysis.xml
9. Moore, D., Shannon, C., Voelker, G., Savage, S.: Internet quarantine: requirements for containing self-propagating code. In: Bauer, F. (ed.) IEEE INFOCOM 2003: The Conference on Computer Communications: 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, pp. 1901–1910, IEEE Operations Center, New Jersey (2003)
10. Nicol, D.M.: The impact of stochastic variance on worm propagation and detection. In: Proceedings of the 2006 ACM Workshop on Rapid Malcode (WORM'06), pp. 57–63, ACM Press, New York (2006)

11. Qing, S., Wen, W.: A survey and trends on internet worms. Comput Secur **24**, 334–346 (2005)
12. Serrazi, G., Zanero, S.: Computer virus propagation models. In: Calzarossa, M.C., Gelenbe, E. (eds.) MASCOTS 2003: Tutorials of the 11th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. Springer, Heidelberg (2003)
13. Staniford, S., Paxson, V., Weaver, N.: How to own the Internet in your spare time. In: Proceedings of the 11th USENIX Security Symposium, pp. 149–170. USENIX Association, California (2002)
14. Zou, C.C., Gong, W., Towsley, D.: Code red worm propagation modeling and analysis. In: Proceedings of the 9th ACM Conference on Computer and Communications Security, pp. 138–147, ACM Press, New York (2002)
15. Zou, C.C., Towsley, D., Gong, W.: On the performance of internet worm scanning strategies. Perfor Eval **63**, 700–723 (2006)