Fitting a Code-Red Virus Spread Model: An Account of Putting Theory into Practice

Anna Kolesnichenko, Boudewijn R. Haverkort, Anne Remke, Pieter-Tjerk de Boer
Centre for Telematics & Information Technology, University of Twente, Enschede, The Netherlands
{a.v.kolesnichenko, b.r.h.m.haverkort, a.k.i.remke, p.t.deboer}@utwente.nl

***Abstract*—This paper is about fitting a model for the spreading of a computer virus to measured data, contributing not only the fitted model, but equally important, an account of the process of getting there. Over the last years, there has been an increased interest in epidemic models to study the speed of virus spread. But parameterising such models is hard, because due to the unexpected nature of real outbreaks, there is not much solid measurement data available, and the data may often have imperfections. We propose a mean-field model for computer virus spread, and use parameter fitting techniques to set the model's parameter values based on measured data. We discuss a number of steps that had to be taken to make the fitting work, including preprocessing and interpreting the measurement data, and restructuring the model based on the available data. We show that the resulting parameterised model closely mimics real system behaviour, with a relative squared error of 0.7%.**

I. Introduction

Over the last years, there has been an increased interest in epidemic models to study the speed of virus spread (worms) in computer networks [11], [9], [15], [26], [19], [20]. However, although these models are often mathematically elegant, they mostly suffer from the lack of realistic parameters to study real system behavior. This is particular so for internet-scale systems, for which a structured measurement set-up is very difficult to achieve, if not impossible at all. Moreover, when studying attacks launched via the internet, there normally is no adequate measurement infrastructure at all, hence, in such cases, one has to work with whatever has been measured (observed) from the attack. The challenges encountered (as well as some solutions) along that trajectory are described in this paper. Hence, the contribution of the paper not only lies in the product (a parametrised model) but also in the process of parametrisation.

To make this more concrete, this paper aims to obtain a better understanding of the spreading phase of a computer worm, and does so by combining a mean-field model of worm behaviour with parameter fitting techniques, and illustrates this on the case of Code-Red [17]. We explain how to build the mean-field model of the worm [12], and how to estimate the corresponding parameters, so as to find the best fit between the available data and the model prediction. We also present a number of intricate technical issues, ranging from the additional (preprocessing) work to be done on the available measurement data, the interpretation of the data, for instance in relation to performed measurements, as well as a restructuring of the model (based on data unavailability), that has to be performed before applying the parameter fitting algorithms. As proof of the pudding, we show that our approach does provide a set of parameters that, when used in the proposed models, allows us to closely mimic real system behaviour, with a relative squared error of, at most, 0.7%. The presented model and parametric study is, as far as we know, the most detailed study of the spreading phase of Code-Red.

The presented model has certain properties that very well fit the application purpose. First, it does assume a very large number of similar interacting objects cf., [5], and, secondly, it does not assume anything about the underlying network topology. Indeed, we think these two properties are valid at internet-scale, in which potentially millions of computers interact with each other in a fully-connected overlay network (at TCP/IP-level). As third important property we mention that the number of parameters is very small; as we will see, even with such a small number of parameters, the fitting is already quite challenging.

It is important to note that we do *not* claim that our proposed model is the best model (or better than other models), however, we do claim that the model has certain properties that very well fit the application purpose.

To summarise, the aim and contribution of the paper is threefold:

- it shows how a simple behavioural model of a virus can be used as basis for modelling virus spread;
- it illustrates that such a model can be parameterised well, based on measurements performed during the outbreak, using standard parameter estimation techniques, provided the measurements are very carefully dealt with;
- maybe most importantly, it discusses the challenges encountered when performing such a detailed study, in which the measurements show all sorts of artefacts that are easily overlooked, but do have a substantial impact on the fitting.

Finally, one could remark that the Code-Red outbreak is more than ten years ago, hence, that our study is too late to be of value. We do not think that this is the case. Indeed, one could question what, at this point in time, the value is of a good parameter set for a particular Code Red model. However, the learnings from the process of obtaining these parameters under circumstances that are typical for virus spreading, are important, now and in the future. Also, the paper has as implicit message that work on obtaining mathematically more refined models is probably of little use, as long as the model parametrisation process is as challenging as described here.

The paper is further organised as follows. In Section II the background and history of the Code-Red worm is presented.

978-1-4673-8496-4/16/$31.00 ©2016 IEEE

The mean-field model for Code-Red is built in Section III. The available data is described in Section IV, whereupon the proposed mean-field model is re-assessed in Section V. Section VI provides the results of the Code-Red case-study for the July 2001 outbreak; the results for the August 2001 outbreak are provided in Section VII. Related work is discussed in Section VIII, whereas Section IX concludes the paper.

II. Code-Red

On June 18, 2001, information about a buffer-overflow vulnerability in Microsoft's IIS web servers was released by eEye [4], which was followed by a Microsoft patch eight days later [6]. On July 12, 2001, Code-Red version 1 (further referred to as CRv1) started to spread by exploiting this vulnerability. There was no direct damage done by CRv1, except for the phrase "Hacked by Chinese" added to the top level of web pages of some hosts that happened to run web servers. CRv1 did not spread widely due to the static seed in its pseudo-random number generator, which caused each infected host to scan (that is, try to infect) the *same* list of hosts. The only tangible effects were visible in local networks due to the resources consumed on infected hosts (servers); the impact on the global resources was negligible.

Following CRv1, on July 19, 2001, at approximately 10:00, Code-Red version 2 (further referred to as CRv2) started to spread. It appears that unlike CRv1, CRv2 used a random seed. Therefore, each of the infected machines tried to infect a different list of randomly generated IP addresses at an observed rate of, approximately, 11 probes per second. Although the worm did not cause any direct damage (again apart from the "Hacked by Chinese" message), CRv2 had a major impact due to the huge number of infected hosts and probes sent. It is considered to be the most costly malware of 2001, with a total estimated cost of 2.75 billion USD. Moreover, since the lists of IP addresses to infect were drawn randomly, CRv2 was sending the probes not only to vulnerable IIS web-servers, but to all kinds of hosts; although these could not be infected as such, they could crash or reboot under the attack.

Both versions of Code-Red were programmed to take identical actions when infecting a new host. First, the worm checks the system time and date, followed by one of the following actions:

- **Spreading phase**. If the date is between the 1st and the 19th of each month, the worm generates a random list of IP addresses and tries to infect as many machines in this list as possible by trying to connect to them on TCP port 80.
- **Attacking phase**. If the date is between the 20th and the 28th of a month, the worm stops spreading and starts a Denial-of-Service attack against the site `www.whitehouse.gov`. These attacks did overwhelm the corresponding servers with so much useless data that they were unable to function properly [3]. Luckily, the attackers addressed `www.whitehouse.gov` through a fixed IP address and not through the hostname; the problem was solved by moving the website to another IP address.
- **Inactive phase**. The worm is inactive after the 28th of each month.

Note that the employed system clock call returns UTC time [21], therefore, all hosts switch between these three phases simultaneously, unless a host is malfunctioning. When an infected machine is rebooted, it is disinfected; however, it remains vulnerable. The only way to protect a machine is applying a patch.

CRv2 was able to cause major damage during the 14 hours it was spreading; at midnight of July 20 it stopped spreading, as it was programmed to. On August 1, 2001, the worm started to spread again, and by midnight approximately 275 000 unique hosts were infected (according to the CAIDA data set; see details in Section IV). The difference between the first and the second outbreak of CRv2 might be due to the fact that some machines were patched before August 1, which reduces the probability of finding vulnerable hosts.

III. A First Model of CRv2 Spreading

In the following we propose a model for the spreading phase of CRv2, which takes place on infected hosts between the 1st and the 19th of each month. This model is based on the description of the worm behaviour given in the previous section.

Let us first address a model that reflects the behaviour of a single host. From the description of CRv2 we infer that there are three modes a node (or host) can be in while the worm spreads: *Vulnerable, Infected,* and *Patched*. This results in a 3-state model with state space $S^l = \{s_1, s_2, s_3\}$, with $|S^l| = K = 3$ states. The states are labelled as *Vulnerable, Infected*, and *Patched*, as indicated in Figure 1. The transition rates are as follows:

- A *Vulnerable* machine becomes *Infected* with rate k_1^*, which increases if the number of infected hosts grows.
- An *Infected* machine is rebooted and returns back to the *Vulnerable* state with constant rate k_2.
- The patch might be installed on a *Vulnerable* or *Infected* machine, which happens with rates k_4^* and k_3^*, respectively. These rates depend on the awareness of operators on the worm existence. It is to be expected that as the number of infected machine grows, the awareness grows with it, so these rates should increase too.
- A *Patched* machine can not be infected and stays in that state for the remaining time.

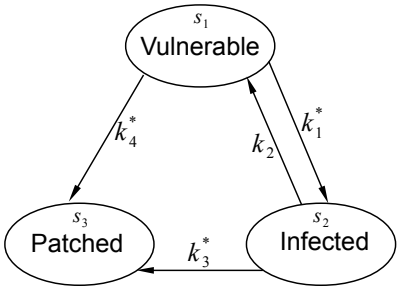

Figure 1. The model of CRv2 propagation for a single host.

Given a network of N nodes (where N is assumed to be large, which is a reasonable assumption in this context), we can model the overall average behaviour via a so-called *mean-field model*. It has the same underlying state space structure as the individual host model, as given in Figure 1, however, each state now should be interpreted as a vector $\overline{m} = (m_1, m_2, m_3)$, where $m_1(t)$ denotes the *fraction* of *Vulnerable* machines at time t, and $m_2(t)$ and $m_3(t)$ correspond to the fraction of *Infected* and *Patched* machines at time t, respectively. That is, a state is given as a triple of non-negative real numbers in $[0, 1]$, together summing up to 1.

After defining the global model the transition rates need to be specified. The rates k_1^*, k_3^*, k_4^* depend on the number of infected hosts and can be expressed as follows:

$$k_1^*(t) = k_1 \cdot m_2(t), \quad k_3^*(t) = k_3 \cdot m_2(t), \quad k_4^*(t) = k_4 \cdot m_2(t),$$

where k_1 is the infection rate of one machine; k_3 and k_4 are the rates of patching for an infected or vulnerable host, respectively. Note that the representation of the infection rate takes into account the fraction $m_2(t)$ of infected computers, which each spread the virus with identical constant rate k_1. The patching rates are difficult to estimate, because human behaviour plays an important role in these. At this point, we assume that the rate of patching is directly proportional to the number of infected machines, as human awareness of the problem presumably is also proportional to this number.

Now that we have defined the rates, we can apply the well-known mean-field theorem, cf. [13], to derive a system of ordinary differential equations (ODEs) that describes the transient behaviour of the model, as follows:

$$\begin{cases} \dot{m}_1(t) &= k_2 m_2(t) - k_1 m_2(t) m_1(t) - k_4 m_1(t) m_2(t), \\ \dot{m}_2(t) &= k_1 m_2(t) m_1(t) - k_2 m_2(t) - k_3 m_2(t) m_2(t), \\ \dot{m}_3(t) &= k_4 m_1(t) m_2(t) + k_3 m_2(t) m_2(t), \end{cases} \tag{1}$$

where $\dot{m}_i(t)$ denotes the time-derivative of $m_i(t)$, and initial condition $\overline{m}(0) = (m_1(0), m_2(0), m_3(0))$. Earlier requirements on the mean-field model state that for all $t \geq 0$ and all $i \in \{1, \cdots, K\}$, we have $m_i(t) \in [0, 1]$, and for all $t \geq 0$, we have $\sum_i m_i(t) = 1$. This is the model we will use as starting point for the data to be fitted. Finally, we let $\overline{M} = (M_1, M_2, M_3)$ represent the vector of actual *numbers* of nodes in each state, rather than *fractions* of the total population; the difference is only a multiplicative factor:

$$\overline{M}(t) = N \cdot \overline{m}(t). \tag{2}$$

IV. Code-Red Measurement Data

The measurement data used in this paper is the so-called "CAIDA Dataset on the Code-Red Worms, July and August 2001", cf. [22]. This is a publicly available set of files, containing summarised information that does not identify infected hosts individually, gathered during the outbreak. For the July outbreak, the data is based on combining measurements done with a /8 "Telescope network" at UCSD (University of California at San Diego) until 16:30 UTC, sampled `netflow` data from a router upstream of this /8 network (after 16:30 UTC), and data from two /16 networks at Lawrence Berkeley Laboratory (LBL) [1]. For the August outbreak, only the USCD Telescope network data has been used.

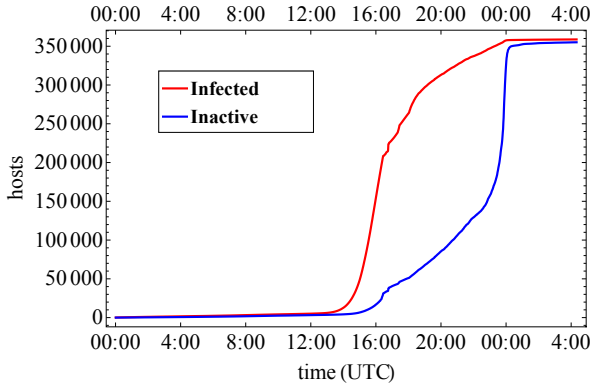

Figure 2. The total number of unique infected (red) and inactive (blue) hosts on July 19-20.

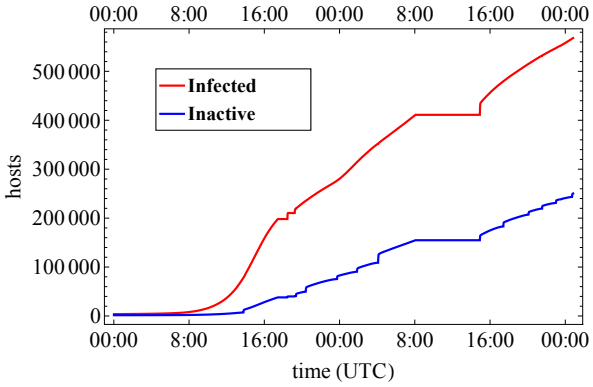

Figure 3. The total number of unique infected (red) and inactive (blue) hosts on August 1-2, 2001.

Two types of traces from this data set have been used in this paper:

(1) The number of **new unique infected hosts** that are starting to spread infection over time. Hosts are considered to be infected if they sent at least two TCP SYN packets to port 80 on non-existent hosts, which helps to eliminate random source denial-of-service attacks from the Code-Red data.

(2) The number of **hosts that have stopped being infected (*inactive*)** over time. A host which was previously infected is considered to be inactive after no further unsollicited traffic has been observed from it.

Note that since the data represents only a sample of all probes sent by infected hosts, it provides a lower bound of the number of hosts infected (and inactive) at any given time.

Figures 2 and 3 depict the total number of unique newly infected hosts (red) and inactive hosts (blue) on July 19–20 and August 1–2, 2001, respectively. Clearly, there are some unnatural jumps in these cumulative curves, indicating moments where apparently data collection was postponed for a while and then resumed; in particular, there has been a large such gap on July 19 between 16:51 and 17:21 [17]. The number of infected (and inactive) hosts stops growing at midnight because the worm was programmed to stop spreading.

The data sets presented here, clearly show the challenges encountered when analysing measurements that have been made in the past (and not for the purpose of this particular modelling study), and that might be partly incomplete due to (non-documented) measurement problems. For instance, the stop in growth in Figure 2 at midnight can be explained from the fact that the worm was programmed to stop spreading at that point in time; there is not much debate about that. However, the reason why the rate of increase (the derivative of the red curve in Figure 2) declines has been attributed in the literature to various reasons, e.g., overloaded networks due to the worm spreading itself [26], or to the un-availability of vulnerable hosts [17]. It is impossible to find ground truth for this now, however, another reason might lie in the fact that according to [17] many of the infected machines were actually office desktops, whose users were not aware that they are running an active web server. Therefore, the slow down might be due to the fact that more and more computers are switched off in each time-zone when the working day is over (starting at 16:00 UTC), hence, do not contribute anymore to the propagation of CRv2.

It is generally assumed that the CRv2 outbreak started at July 19, 10:00 UTC. Before that time point, infected hosts are assumed to have been infected by the CRv1 outbreak. As can be seen in Figure 2, the true outbreak of CRv2 starts around 14:00 UTC. Comparing Figures 2 and 3, we see that the growth of the number of infected hosts is lower in August than it was in July; it appears that some hosts have been patched after the first outbreak, so that there are fewer vulnerable hosts left.

V. CRv2 Spreading Model Reconsideration

In Section III a model for CRv2 was proposed based on the description of the worm. The next step is finding the parameters of this model which provide the best fit to the data as described in the previous section. However, we notice that an extra step has to be taken here, in order to make sure that the model not only reflects the behaviour of the system (network under CRv2 attack), but also matches the available (observable) data. This is the case for the following two reasons.

First of all, we think that the rebooting of infected hosts has to be reconsidered, since rebooting was not captured (measured) in the dataset. A rebooted host either is not re-infected and is therefore added to the set of inactive hosts, or it gets re-infected, however, in that case it will still not be counted since it is *not* a new *unique* infected host (it has been seen before). Therefore, since rebooting cannot be observed from the data, we think that the rebooting transition needs to be eliminated from the model since there is no possibility to fit it. Note, however, that the number of rebooting events is significantly lower compared to actual infections, therefore, the influence of the rebooting is relatively insignificant.

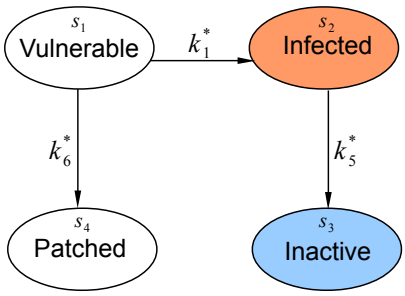

Figure 4. The rethought model for CRv2 virus propagation

Secondly, it appears that the patched hosts have to be split into two groups, that is, (i) hosts which became inactive after being infected, and (ii) hosts which were never infected before getting patched. This distinction is needed because only the first group of hosts can be observed in the available measurement data.

Given the above considerations, a fourth state is added to the model. The states are now labelled as *Vulnerable, Infected, Inactive, Patched*, where the *Inactive* state reflects patching *after* being infected (as also represented in the dataset); hosts patched *before* being infected now are thought to belong to the *Patched* state (see Figure 4). The model again has a finite local state space $S^l = \{s_1, s_2, s_3, s_4\}$ with $|S^l| = K = 4$ states, and the transition rates are as follows:

- A *vulnerable* machine becomes *Infected* with rate $k_1^*(t) = k_1 \cdot m_2(t)$, as discussed in Section III.
- *Infected* machines are patched (and become *Inactive*) with rate $k_5^*(t) = k_5 \cdot m_2(t)$.
- *Vulnerable* machines are patched with rate $k_6^*(t) = k_6 \cdot m_2(t)$.

The expressions for the transition rates are kept the same as in Section III. The infection rate k_1 remains unchanged, the patching rates for infected and vulnerable hosts are now denoted as k_5 and k_6, respectively. Now, given a system of N such hosts, the overall model has state variables $\overline{m} = (m_1, m_2, m_3, m_4)$, with $m_i \in [0, 1]$ and $\sum_i m_i = 1$, and its dynamics are similar to those in Section III:

$$\begin{cases} \dot{m}_1(t) &= -k_1 \cdot m_2(t) \cdot m_1(t) - k_6 \cdot m_1(t) \cdot m_2(t), \\ \dot{m}_2(t) &= k_1 \cdot m_2(t) \cdot m_1(t) - k_5 \cdot m_2(t) \cdot m_2(t), \\ \dot{m}_3(t) &= k_5 \cdot m_2(t) \cdot m_2(t), \\ \dot{m}_4(t) &= k_6 \cdot m_1(t) \cdot m_2(t), \end{cases} \tag{3}$$

with initial conditions $\overline{m}(0) = (m_1(0), m_2(0), m_3(0), m_4(0))$ (using the similar vector notation as before). The actual *number* of hosts in each state is again addressed by $\overline{M} = (M_1, M_2, M_3, M_4)$. Notice that in this extended model, we still do not address issues related to network congestion, nor to the different time-zones from which infections might originate. Although both these issues have been put forward as potential explanations for certain aspects in the spreading behaviour, we do not see how we can sensibly fit our models on the basis of these considerations, simply because we do not have the data available for that purpose. Instead, we make sure that we use the data which fits the proposed model (see Sections VI and VII).

The available measurement data (as shown in Figures 2 and 3) shows the total number of infected and patched hosts, neither of which corresponds directly to state s_2 of the model. To obtain data corresponding to s_2 of the overall model, i.e., the number of hosts still infected at a given time, the number of inactive hosts has to be subtracted from the number of

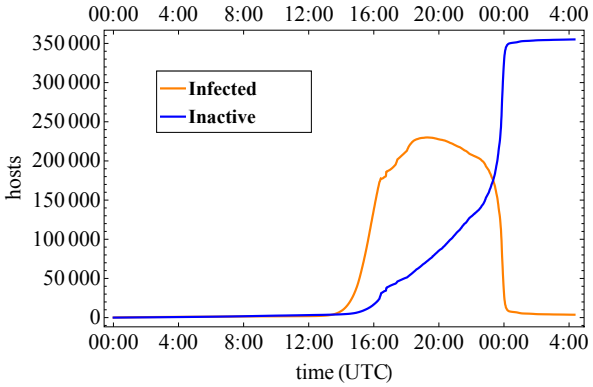

Figure 5. The total number of hosts, still infected at time t (orange); and inactive hosts (blue) on July 19-20, 2001.

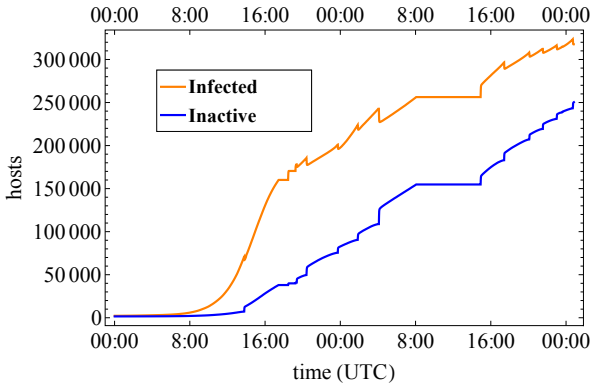

Figure 6. The total number of hosts, still infected at time t (orange); and inactive hosts (blue) on August 1, 2001.

infected hosts. This results in Figures 5 and 6, which depict this modified view on the data, for July and August 2001, respectively. This data corresponds directly to the states of the extended model: (i) the number of still infected hosts (orange solid line) is reflected by $m_2(t)$ (and state s_2), (ii) the number of inactivated hosts (blue solid line) corresponds to $m_3(t)$ (and state s_3).

A final point of notice concerns the estimation of the initial conditions $\overline{m}(0)$. There is no clear evidence to set these four values *a priori* correctly. We see three possible ways to deal with this:

1) We obtain, as good as possible, information about the initial state from the measurements or the measurement set-up. Note that this will be very difficult, as the measurement data we have, has not been specifically set-up for this purpose.
2) We use any circumstantial evidence from the literature describing the Code-Red outbreaks to come up with good estimates.
3) We add the initial conditions as extra parameters to be fitted.

In what follows, we will make use of each of the above three approaches, thus also showing their advantages and disadvantages. Now, with this new view on the model, the data and the initial conditions, the fitting procedure can be started.

VI. CRv2 outbreak in July 2011

In this section we present the best fitting models for the outbreaks of CRv2 in July 2001.

In general terms, there exists a number of well-known parameter estimation techniques, such as, least-squared error [2], maximum likelihood [18], generalised maximum spacing estimates [8], generalized method of moments [10], etc. These methods have been used in a wide variety of application areas, e.g., [16], [25], [14], etc. In this paper, we minimize the *relative squared error*, defined as

$$\mathcal{E}_{rel} = \frac{\sum_{r=1}^{R} ||\mathcal{O}(t_r) - m(t_r)||^2}{\sum_{r=1}^{R} ||\mathcal{O}(t_r) - \overline{\mathcal{O}}||^2}, \tag{4}$$

where $\mathcal{O}$ is the actual data, $\overline{\mathcal{O}}$ its average, and m the data from the mean-field model to be fitted; this measure has the advantage of not depending on the quantities' order of magnitude. In our case, minimizing $\mathcal{E}_{rel}$ can be shown to be equivalent to the least squared error and the maximum likelihood methods. We use the Wolfram Mathematica optimisation function `NMinimize` [24]; this function attempts to find a global minimum subject to given constraints. Accuracy and the number of iterations can be adapted to obtain a good compromise between precision and speed.

As discussed in Section V the measurement data provides information on: (i) the number of infected hosts over time, which corresponds to $M_2(t)$ in our model, and (ii) the number of inactive hosts, which corresponds to $M_3(t)$ in our model. Measurement data directly corresponding to $M_1(t)$ and $M_4(t)$ is not available, so these cannot be included in the fitting procedure. We also discuss different choices of initial conditions.

While estimating the model parameters for the July 2001 outbreak, we limit ourselves to the data collected before 16:20 UTC, in order to avoid mixing data from the two different sources, cf. Section IV, and to avoid the slowing-down phase, which was not included in the model.

We need to set the initial conditions for the model. In [23] it was mentioned that CRv2 infected between 1 and 2 million out of a potential number of 6 million hosts. Therefore, to start with, we set the initial number of vulnerable hosts $M_1(0)$ equal to 6 million (abbreviated as "$6H$"), minus the number of nodes initially infected, patched or inactive. We set the initial number of patched nodes $M_4(0) = 0$, whereas for the initial number of infected and patched nodes we take the corresponding number from the trace at 10:00 o'clock (the generally assumed starting time of the outbreak of CRv2), that is, we set $M_2(0) = 4181$ and $M_3(0) = 2528$. Note that these could be "left-overs" from the CRv1 outbreak, hence, we cannot be sure whether this is a good choice.

Figure 7 depicts the (fitted) model prediction of the number of infected (orange solid line) and inactive (blue solid line) hosts, and the measured number of infected (orange dashed line) and inactive (blue dashed line) hosts. As one can see, the fitted model (solid lines) does not fit the measurement data (dashed lines) very well. It overestimates the number of infected hosts in the beginning by 10 to 20 thousand hosts, and underestimates it after 15:30 UTC. This is also reflected in the relative squared error of the fitting procedure, which is approximately 9.9%. Apparently, there is a factor we did not

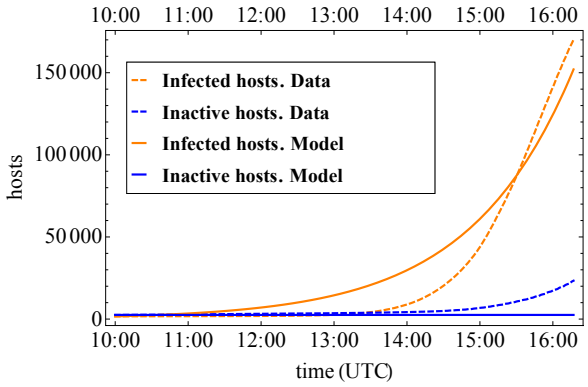

Figure 7. Fitting results for July 19, 2001 with initial condition $\overline{M}(0) = (6H - 6709, 4181, 2528, 0) = (5\,993\,291, 4181, 2528, 0)$.

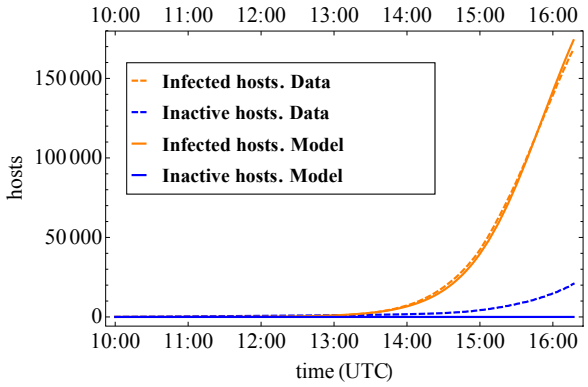

Figure 8. Fitting results for July 19, 2001 with initial condition $\overline{m}(0) = (6H - 3, 3, 0, 0)$.

take in account well enough. Most probably, the parameter k_1 (the speed of the virus propagation) is underestimated (see results after 15:30), whereas the initial overestimation might be due to the incorrect initial settings (the number of the initially infected hosts is too big).

As suggested in [17], another reason for the bad fit might lie in the fact that the activity of CRv1 and other background unsolicited SYN probes were already registered before CRv2 started to spread. Because of this, all infections that took place before 10:00 UTC have to be subtracted; doing so, leaves only 3 infected hosts at 10:00 UTC. In this context, note that the initial number of hosts infected with CRv1 is known to be 3 [17]; it appears consistent that CRv2, as direct "improvement" of CRv1 starts from the same number of initially compromised hosts. Similarly, all the hosts which were registered as inactive at 10:00 UTC have to be eliminated since they were not counted due to the CRv2 activity and are not captured by the model. Hence, we refitted the model, but now with initial conditions $\overline{M}(0) = (6H - 3; 3, 0, 0)$. The results of the new parameter fitting is shown in Figure 8. The fit for the number of infected hosts is quite good: the observed data (orange dashed line) is almost indistinguishable from the number predicted by the model (solid orange line). Also the relative squared error of the fitting procedure has reduced to 1.6%.

To a large extent, the remaining uncertainty is due to the estimation of the number of inactive hosts. This can be explained by the fact that the inactive hosts are difficult to model because it involves modelling human behaviour: as explained in Section III we have assumed for simplicity that the rate of patching hosts is linearly proportional to the number of infected hosts.

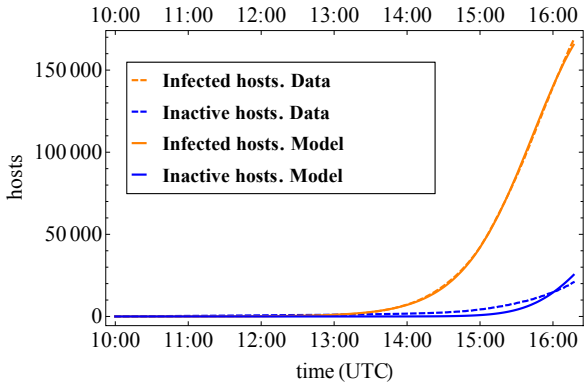

Figure 9. Fitting results for July 19, 2001 with initial condition $\overline{M}(0) = (2H - 3, 3, 0, 0)$.

We now go further into the choice of the initial conditions, in particular, the number of initially vulnerable hosts ($M_1(0)$ in our model). We performed 60 fitting experiments, where we took $M_1(0) = I - M_2(0)$, with $I \in \{500\,000, 600\,000, \cdots, 6\,400\,000\}$, $M_2(0) = 3$ and $M_3(0) = M_4(0) = 0$. We then find that when the number of initially vulnerable hosts is taken in the range from 500 000 to 2 000 000, the relative error is smallest, and almost does not change (and has value close to 0.2%). For the initial number of vulnerable hosts larger than 2 million (denoted as "$2H$"), the relative error increases. Hence, without relying on any textual source, the (purely numerical) optimal choice for the initial number of hosts lies in the smaller range (and is smaller than the 6H assumed previously). Figure 9 presents the fitting results for initial conditions $\overline{M} = (2H - 3, 3, 0, 0)$.

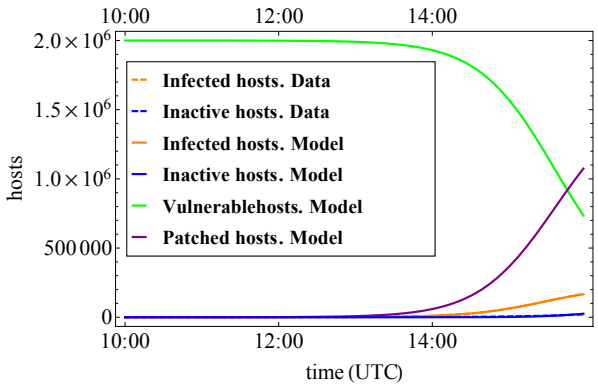

Figure 10. Fitting results for July 19, 2001, including model results for vulnerable and patched hosts for which no measured data is available. Initial condition $\overline{M}(0) = (2H - 3, 3, 0, 0)$.

Finally, Figure 10 depicts the fitted model behaviour for all four states, that is, including the two states for which no data was available. The number of patched hosts is estimated to be quite high, and the number of vulnerable hosts is getting smaller, which can explain the slowing down of the infection.

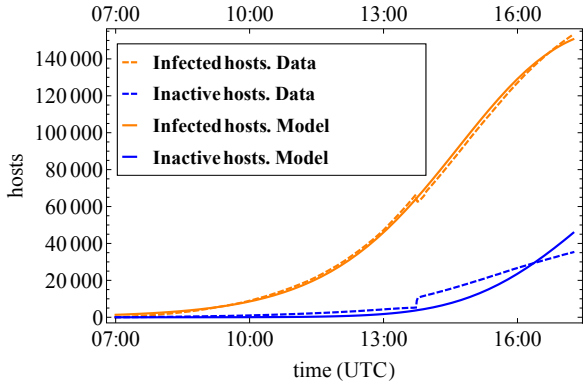

Figure 11. The observed data points and fitting results (dashed lines) for the number of infected hosts at time t (orange); and inactivated hosts (blue) on August 1, with initial conditions $\overline{M}(0) = (1\,498\,669, 1331, 0, 0)$ (rounded values).

VII. CRv2 outbreak in August 2001

We now address the parameter fitting for the measurements from August 1, 2001. As before, also for the August dataset the initial conditions are unknown. It seems reasonable to assume that the number of vulnerable hosts did not change dramatically after the first outbreak of CRv2; only a limited number of hosts was patched during the attacking phase of CRv2, as also suggested in [17]. We therefore set $M_1(0) = 1.5 \cdot 10^6$ (denoted as "$1.5H$"). It is again difficult to find good values for the initial number of infected machines and inactive machines, due to the fact that all the activity of CRv2 before 00:00 UTC has to be taken into account, and, in addition, all the nodes being affected by background activity (as in the case of July) have to be subtracted.

In this section we want to illustrate the third way (cf. Section V) of dealing with unknown initial conditions, that is, we add these as extra parameters to the fitting procedure. Notice that although this way seems to be the most straightforward, adding extra degrees of freedom to the fitting procedure can lead to the worse result. Therefore, using this method while having limited data traces (as in CRv2 case) is a "last resort" rather then obvious solution. We here take the initial numbers of infected nodes ($M_2(0)$) as extra parameters to the parameter fitting procedure; the number of inactive ($M_3(0)$) and patched hosts ($M_4(0)$) can again be set to zero, as the patching before midnight was not due to the activity of CRv2 for this outbreak. Given the above assumptions, we need to find the parameters k_1, k_5, k_6, and initial condition $M_2(0)$ that minimise the relative squared error.

Figure 11 depicts the result of the model fitting procedure for the initial conditions $\overline{M}(0) \approx (1\,498\,669, 1331, 0, 0)$. These initial conditions allowed us to obtain the best fit with a squared error of approximately 0.7%. Note that the number of initially infected hosts is actually not an integer number (but $1\,331.16$) due to the fact that the mean-field model is a model that addresses fractions of objects, but not every object independently. As one can see, the fitted model reflects the number of infected hosts during the second outbreak of CRv2 quite well, despite a still unresolved problem with the data collection around 13:00 UTC.

VIII. Related work

In this section we provide a short overview of the related work on the spreading phase of Code-Red. The measurement data used in our study was collected and studied by CAIDA; the results were presented in [17], providing the background of the worm as well as the details of the data collection. The survey provides insight in the infection and deactivation processes. Moreover, geographical locations and types of the infected hosts were studied. The authors conclude that on July 19, between 11:00 and 16:30 UTC, the infection grows exponentially. However, no formal model justifying this claim has been provided.

Staniford [20] also provides an analysis of worm spreading. A model was proposed to explain the infection rate (only for the number of infected hosts) for both outbreaks of Code-Red. The proposed model is very similar to an epidemiological model:

$$\frac{da}{dt} = K \cdot a \cdot (1 - a),$$

where K is the infection rate of one compromised machine, and a is the fraction of compromised machines. This model is explained as follows: each of the a infected hosts is able to compromise K hosts per unit of time while only $1-a$ machines are not infected yet. Moreover, the author provides a *manually made* fit to the data, obtained from [7]. The parameters were obtained as a guess, and no formal explanation was provided.

Another model of CRv2 worm propagation was proposed by Zou *et al.* in [26]. The proposed model was based on a classical epidemiological model, which was modified in order to provide better accuracy. Their so-called *two-factor worm model* includes two additional aspects:

- human counter-measures against worm spreading (patching, rebooting, etc.);
- slowing down of the worm infection rate due to the worm's impact on internet traffic and infrastructure.

Adding these factors explained the slowing down of the worm spread before midnight of July 19, 2001. This model was compared against the data collected on July 19, 2001 (again for the number of infected machines only), and provided good results for the proposed parameter set. However, no evidence on the source of the chosen parameters was provided. The two-factor model is closely related to the mean-field model proposed by us.

Unlike all the previously proposed models, the mean field model proposed by us is based on insight in the actual operation of the virus. Furthermore, our model is parameterised with well-known estimation techniques and uses the most elaborate data set available so far. Therefore, we think our model is more trustworthy and is better suited for reasoning about the spread of CRv2. Furthermore, the parameters in our model have a physical interpretation, which can help in reasoning about countermeasures.

IX. Conclusions

This paper provides a full account of the fitting process to obtain a fully parameterised model of virus spread. We base our model and fitting procedure on publicly known insight in

978-1-4673-8496-4/16/$31.00 ©2016 IEEE

the operation of CRv2, and on publicly available data sets. The paper foremost shows the challenges encountered when trying to parameterise a simple model of a large-scale distributed system based on publicly available data sets.

Starting from the characterisation of the system itself, i.e., CRv2, we proposed an initial model describing the state of each host in the network, and subsequently discussed why the unavailability of certain measurement data leads to a slightly adapted model. We also presented why the measurement data has to be handled very carefully, e.g., due to the fact that certain measurement intervals are missing or incomplete (due to sampling). Furthermore, the available data only reflects part of the system under study, that is, the data does not provide information on the number of rebooted and patched hosts nor on the total number of vulnerable hosts. For these reasons, the fitting procedure has been a very tedious process, which cannot be easily automated, and from which no final "once-and-for-all recipe" can be given. Moreover, the process described in this paper indicates that the construction of an abstract model for virus spread can best be done "in parallel" with the study of the available measurement data, particularly so for virus outbreaks, when incomplete data sets are (expected to be) more rule than exception.

Despite all the above, we have shown that it is possible to find a model and a set of parameters that closely captures the first part of the virus spreading. Whether these are the "ultimate correct parameters" cannot be concluded, simply because we are missing ground truth. However, the models we fitted, did allow us to obtain parameters which ensure a relative squared error of 0.2% and 0.7% between the model prediction and the measurement, for the July and August outbreaks, respectively.

ACKNOWLEDGEMENTS

The research performed for this paper has been funded by Dutch 3TU Centre of Excellence on Dependable ICT, the NWO VENI project on "Dependability Analysis of Fluid Critical Infrastructures using Stochastic Hybrid Models," the Dutch-German NWO/DFG cooperation grant on "Rigorous Dependability Analysis using Model Checking Techniques for Stochastic Systems", the NWO project "Mean-Field Approximation Techniques for Markov Models", and, finally, the FP7 project Sensation. The authors would like to thank Ramin Sadre (Aalborg University), David Spieler (previously Saarland University) as well as Luuk Hendriks and Rick Hofstede (both University of Twente) and Idilio Drago (formerly University of Twente) for useful discussions on the topic.

REFERENCES

[1] Lawrence Berkeley National Laboratory. http://www.lbl.gov/.

[2] J. Aldrich. Doing Least Squares: Perspectives from Gauss and Yule. *International Statistical Review*, 66(1):61–81, 1998.

[3] H. Berghel. The Code Red Worm. *Communications of the ACM*, 44(12):15–19, 2001.

[4] BeyondTrust, Inc. eEye Digital Security. http://www.eEye.com.

[5] L. Bortolussi, J. Hillston, D. Latella, and M. Massink. Continuous approximation of collective systems behaviour: A tutorial. *Performance Evaluation*, 70(5):317 – 349, 2013.

[6] eEye Digital Security. Advisories and Alerts: AD20010618. https://web.archive.org/web/20010805211728/http://www.eeye.com/html/Research/Advisories/AD20010618.html, 2001.

[7] K. Eichmann. Handlers Diary Blog. https://isc.sans.edu/diary/diary.php.

[8] M. Ekstrom. Consistency of generalized maximum spacing estimates. *Scandinavian Journal of Statistics*, 28(2):343–354, 2001.

[9] É. Gourdin, J. Omic, and P. Van Mieghem. Optimization of network protection against virus spread. In *8th International Workshop on the Design of Reliable Communication Networks, DRCN 2011, Krakow, Poland, 10-12 October, 2011*, pages 86–93. IEEE, 2011.

[10] A. R. Hall. *Generalized Method of Moments*. Advanced Texts in Econometrics. Oxford University Press, 2005.

[11] J.O. Kephart and S.R. White. Directed-graph epidemiological models of computer viruses. In *IEEE Symposium on Security and Privacy*, pages 343–361, 1991.

[12] A. Kolesnichenko, A.K.I. Remke, P.T. de Boer, and B.R. Haverkort. Comparison of the mean-field approach and simulation in a peer-to-peer botnet case study. In *EPEW*, volume 6977 of *LNCS*, pages 133–147. Springer, 2011.

[13] T.G. Kurtz. Solutions of Ordinary Differential Equations as Limits of Pure Jump Markov Processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

[14] A. R. Liddle. Statistical Methods for Cosmological Parameter Selection and Estimation. *Annual Review of Nuclear and Particle Science*, 59(1):95–114, 2009.

[15] P. Van Mieghem, J. Omic, and R.E. Kooij. Virus spread in networks. *IEEE/ACM Transactions on Networking*, 17(1):1–14, 2009.

[16] L. Mikeev and V. Wolf. Parameter Estimation for Stochastic Hybrid Models of Biochemical Reaction Networks. In *Proceedings of the 15th ACM International Conference on Hybrid Systems: Computation and Control*, pages 155–166, New York, NY, USA, 2012. ACM.

[17] D. Moore, C. Shannon, and J. Brown. Code-Red: a case study on the spread and victims of an Internet worm. In *Internet Measurement Workshop (IMW) 2002*, pages 273–284, Marseille, France, Nov 2002. ACM SIGCOMM/USENIX Internet Measurement Workshop.

[18] I. J. Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90 – 100, 2003.

[19] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001.

[20] Silicon Defence. Code Red analysis page. http://web.archive.org/web/20011031043459/http://www.silicondefense.com/cr/, 2011.

[21] R.J. Simon. *Windows NT Win32 API SuperBible*. Waite Group Press, 1997.

[22] The Cooperative Association for Internet Data Analysis. The CAIDA dataset on the Code-Red worms - July and August 2001. http://www.caida.org/data/passive/codered_worms_dataset.xml, 2001.

[23] Wikia. Code Red. http://malware.wikia.com/wiki/CodeRed#, 2006.

[24] Wolfram Research, Inc. NMinimize. http://reference.wolfram.com/mathematica/ref/NMinimize.html, 2014.

[25] Z. Zhang. Parameter estimation techniques: a tutorial with application to conic fitting . *Image and Vision Computing*, 15(1):59 – 76, 1997.

[26] Zou, C.C. and Gong, W. and Towsley, D. Code Red Worm Propagation Modeling and Analysis. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, pages 138–147, New York, NY, USA, 2002. ACM.